Abstract

Shared mesh protection allows a set of diversely routed paths with diverse endpoints to collectively oversubscribe protection resources. Under normal conditions no single failure will result in the capacity of the associated protection resources to be exhausted.

When multiple failures occur such that more than one path in the set of paths utilizing shared protection resources is affected, the necessity arises of pre-empting traffic on the basis of business priority rather than application priority.

This memo describes the use of SPMEs and TC marking as a means of indicating business priority for shared mesh protection.

Requirements Language

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

Status of this Memo

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

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

Internet-Drafts are draft documents valid for a maximum of six months and may be updated, replaced, or obsoleted by other documents at any time. It is inappropriate to use Internet-Drafts as reference material or to cite them other than as "work in progress".
The list of current Internet-Drafts can be accessed at http://www.ietf.org/ietf/1id-abstracts.txt.

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

This Internet-Draft will expire on August 2nd 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.

Table of Contents

1. Introduction...................................................3
1.1. Authors......................................................3
2. Conventions used in this document..............................3
2.1. Terminology..................................................3
3. Overview.......................................................4
3.1. Architectural Overview......................................4
4. Signalling Implications........................................5
5. IANA Considerations............................................5
6. Security Considerations.......................................5
7. References.....................................................6
7.1. Normative References........................................6
7.2. Informative References.......................................6
8. Authors’ Addresses.............................................6
1. Introduction

Shared mesh protection is described in [2]. A common interpretation of the behavior of shared mesh protection emerges from the circuit switched world whereby subtending path selectors and selector coordination functions support path preemption to ensure that the highest priority path needing the protection resources is granted ownership of the shared segment, all others being preempted, and such functionality can be successfully delegated to dataplane OAM.

Ultimately this resolves into a business priority decision vs. an application priority decision in how customer traffic is handled. The packet world is different from the circuit world in that there is no guarantee of convenient alignment of resource requirements between preempting and preempted paths. Nor in a packet environment is there the need to completely preempt all the traffic in a lower priority path simply because a higher priority path lays claim to the resources. Finally it is useful to obviate the requirement for preempting and preemptable traffic to be co-routed.

This memo proposes the use of SPMEs with the pipe model of TC copying as an alternative to the use of path pre-emption, path selectors and selector coordination functions for the purposes of implementing business policy.

1.1. Authors

David Allan, Greg Mirsky

2. Conventions used in this document

2.1. Terminology

MPLS-TP: MPLS Transport Profile

MPLS-TP LSP: Uni-directional or Bidirectional Label Switch Path representing a circuit

SMP: Shared Mesh Protection

SPME: Sub-Path Maintenance Entity

TC: Traffic Class

TTL: Time To Live
3. Overview

Shared mesh protection is described in [2]. A common interpretation of the behavior of shared mesh protection emerges from the circuit switched world. In that interpretation subtending path selectors and selector coordination support path preemption functionality to ensure that the highest priority path needing the protection resources is the one granted ownership of the shared segment; all others being preempted. It also assumes that all paths sharing the protection resources conveniently all need exactly the same size pipe.

In packet transport networks there will frequently not be a convenient 1:1 equivalence of the bandwidth requirements of the set of transport paths sharing protection resources such that a simple pre-emption decision can be made. For example 3 paths: A, B, and C sized "n", "n/2" and "n/2" respectively could have a shared segment size "3n/2" such that simultaneous failures necessitating the activation of any two of the protection paths could be accommodated without path preemption. When one ranks A, B and C with a variety of priorities and considers all failure combinations a rather large matrix of possible required behaviors emerges.

If one pursues this line of thinking to its logical conclusion, and envisions a significant set of paths of diverse sizes and diverse priorities, the policy associated with successful path prioritization and preemption becomes quite complex, and ensuring multiple selectors make timely and of necessity common preemption decisions starts to impose network design constraints or require additional coordination protocols that severely impact the utility of SMP.

Further in a packet network there can be a difference in the bandwidth reserved and the bandwidth actually used at any given instant in time. One consequence is that there is no need to completely preempt all the traffic in a lower priority path simply because a higher priority path lays a preferential claim to the bandwidth.

To obviate these problems, this memo proposes an alternative to how business priority can be implemented for shared mesh protection that obviates the need for path preemption and the limitations such an approach imposes.

3.1. Architectural Overview

This memo pre-supposes an operational mode of behavior along the lines of the following:
1) As a matter of network design, specific links (or engineered virtual links) are set aside for the purpose of acting as shared protection resources. The key attribute of these links is that the processing of TC markings will be exclusively for shared protection.

2) The arrangement of the shared protection links can be arbitrary such that contiguous domains can be constructed with an arbitrary number of ingress and egress points. A set of contiguous protection links is known as a protection domain.

3) Either an apriori or on-demand mesh of SPMEs that connect all ingress and egress points in a protection domain is required. These are logically forwarding adjacencies for the purposes of routing protection paths.

4) The instantiation of protection paths requires the mapping of the incoming path at an ingress node for the protection domain to an SPME that connects the ingress to the required egress node from the domain.

5) The pipe model of TC copying is used such that the SPME gets the TC marking associated with the business priority for the path associated with the incoming label value. As the SPME only transits resources where the TC marking has been overloaded in this fashion business priority does not conflict with application requirements.

6) Admission control for the protection paths transiting the protection domain is performed such that the sum of the bandwidth for a given business priority does not oversubscribe any links in the protected domain, but the sum of the bandwidth for all business priorities can. In this way no traffic of the highest business priority using the shared mesh pool will be contended.

4. Signalling Implications

For a future version of this document.

5. IANA Considerations

No IETF protocols were harmed in the publishing of this memo.

6. Security Considerations

For a future version of this document.
7. References

7.1. Normative References


7.2. Informative References


8. Authors’ Addresses

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

Greg Mirsky
Ericsson
Email: Gregory.mirsky@ericsson.com
This Internet-Draft, draft-allan-unified-mpls-ic-req-frmwk-00.txt, has been replaced by another document, draft-allan-mpls-unified-ic-req-frmwk, and has been deleted from the Internet-Drafts directory.

Internet-Drafts are not archival documents, and copies of Internet-Drafts that have been deleted from the directory are not available. The Secretariat does not have any information regarding the future plans of the author(s) or working group, if applicable, with respect to this deleted Internet-Draft. For more information, or to request a copy of the document, please contact the author(s) directly.

Draft Author(s):
D. Allan <david.i.allan@ericsson.com>
Performance-based Path Selection for Explicitly Routed LSPs
draft-atlas-mpls-te-express-path-00

Abstract

In certain networks, it is critical to consider network performance criteria when selecting the path for an explicitly routed RSVP-TE LSP. Such performance criteria can include latency, jitter, and loss or other indications such as the conformance to link SLAs and non-RSVP TE traffic load. This specification uses IGP extension data (which is defined outside the scope of this document) to perform such path selections.

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 26, 2012.

Copyright Notice

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

This document is subject to BCP 78 and the IETF Trust’s Legal
1. Introduction

In certain networks, such as financial information networks, network performance information is becoming as critical to data path selection as other existing metrics. The ability to distribute network performance information in OSPF [I-D.giacalone-ospf-te-express-path] and in ISIS [I-D.previdi-isis-te-metric-extensions] is being defined (outside the scope of this document). This document describes how to use that information for path selection for explicitly routed LSPs signaled via RSVP-TE [RFC3209].

The path selection mechanisms described in this document apply to paths that are fully computed by the head-end of the LSP and then signaled in an ERO where every sub-object is strict. This allows the head-end to consider IGP-distributed performance data without requiring the ability to signal the performance constraints in an object of the RSVP Path message.

When considering performance-based data, it is obvious that there are additional contributors beyond just the links. Clearly end-to-end latency is a combination of router latency, queuing latency, physical link latency and other factors. However, if application traffic requires paths to be selected based upon latency constraints, the same traffic might be in an Expedited Forwarding Per-Hop-Behavior[RFC3246] with minimal queuing delay or another PHB with known maximal per-hop queuing delay. While traversing a router can cause delay, that can be included in the advertised link delay.

This document does not specify how a router determines what values to advertise by the IGP. However, the end-to-end performance that is computed for an LSP path SHOULD be built from the individual link data. Any end-to-end characterization used to determine an LSP’s performance compliance should be fully reflected in the Traffic Engineering Database so that a CSPF calculation can also determine whether a path under consideration would be in compliance.
1.1. Basic Requirements

The following are the requirements that motivate this solution.

1. Select a TE tunnel’s path based upon a combination of existing constraints as well as on link-latency, packet loss, jitter, link SLA conformance, and bandwidth consumed by non-RSVP-TE traffic.

2. Ability to define different end-to-end performance requirements for each TE tunnel regardless of common use of resources.

3. Ability to periodically verify that a TE tunnel’s current LSP complies with its configured end-to-end performance requirements.

4. Ability to move tunnels, using make-before-break, based upon computed end-to-end performance complying with configuration

5. Ability to move tunnels away from any link that is violating an underlying SLA

6. Ability to optionally avoid setting up tunnels using any link that is violating an SLA, regardless of whether end-to-end performance would still meet requirements.

7. Ability to revert back to the best path after a configurable period.

2. Using Performance Data Constraints

2.1. End-to-End Constraints

The per-link performance data available in the IGP [I-D.giacalone-ospf-te-express-path] [I-D.previdi-isis-te-metric-extensions] includes: unidirectional link delay, unidirectional delay variation, and link loss. Each (or all) of these parameters can be used to create the path-level link-based parameter.

While it has been possible to compute a CSPF where the link latency values are used instead of TE metrics, this results in ignoring the TE metrics and causing LSPs to prefer the lowest-latency paths. Instead of this approach to minimize path latency, an end-to-end latency bound merely requires that the path computed be no more than that bound without being the minimum. This bound can be used as a constraint in CSPF to prevent exploring links that would create a path over the end-to-end latency bound.
This is illustrated as follows. Let the LSP have an end-to-end latency bound of 20ms. Assume that the path to node X has been minimized and its latency is 12ms. When X's links are to be explored, the link X<->Y has a link latency of 5ms and the link X<->Z has a link latency of 9ms. The path via X to Y along link X<->Y would have a path latency of 12ms + 5ms = 17ms < 20ms; therefore, the link X<->Y can be explored. In contrast, reaching Z via link X<->Z would result in a path latency of 12ms + 9ms = 21ms > 20ms; therefore the link X<->Z would not be explored in the CSPF.

An end-to-end bound on delay variation can be used similarly as a constraint in the CSPF on what links to explore where the path’s delay variation is the sum of the used links’ delay variations.

For link loss, the path loss is not the sum of the used links’ losses. Instead, the path loss percentage is $(100 - \text{loss}_{L1})(100 - \text{loss}_{L2})\ldots(100 - \text{loss}_{Ln})$, where the links along the path are L1 to Ln. The end-to-end link loss bound, computed in this fashion, can also be used as a constraint in the CSPF on what links to explore.

2.2. Link Constraints

In addition to selecting paths that conform to a bound on performance data, it is also useful to avoid using links that do not meet a necessary constraint. Naturally, if such a parameter were a known fixed value, then resource attribute flags could be used to express this behavior. However, when the parameter associated with a link may vary dynamically, there is not currently a configuration-time mechanism to enforce such behavior. An example of this is described in Section 2.3, where links may move in and out of SLA-conformance with regards to latency, delay variation, and link loss.

When doing path selection for TE tunnels, it has not been possible to know how much actual bandwidth is available that includes the bandwidth used by non-RSVP-TE traffic. In [I-D.giacalone-ospf-te-express-path] [I-D.previdi-isis-te-metric-extensions], the Unidirectional Available Bandwidth is advertised as is the Residual Bandwidth. When computing the path for a TE tunnel, only links with at least a configurable amount of Unidirectional Available Bandwidth might be permitted.

Similarly, only links whose loss is under a configurable value might be acceptable. For these constraints, each link can be tested against the constraint and only explored in the CSPF if the link passes. In essence, a link that fails the constraint test is treated as if it contained a resource attribute in the exclude-any filter.
2.3. Links out of SLA

Link conformance to an SLA can change as a result of rerouting at lower layers. This could be due to optical regrooming or simply rerouting of a FA-LSP. When this occurs, there are three questions to be asked:

a. Should the link be trusted and used for the setup of new LSPs?

b. Should LSPs using this link be immediately verified for continued compliance to their end-to-end constraints?

c. Should LSPs using this link automatically be moved to a secondary path?

2.3.1. Use of Anomalous Links for New Paths

If the answer to (a) is no for latency SLAs, then any link which has the Anomalous bit set in the Unidirectional Link Delay sub-TLV[I-D.giacalone-ospf-te-express-path] [I-D.previdi-isis-te-metric-extensions] should be removed from the topology before a CSPF calculation is used to compute a new path. In essence, the link should be treated exactly as if it fails the exclude-any resource attributes filter.[RFC3209].

Similarly, if the answer to (a) is no for link loss SLAs, then any link which has the Anomalous bit set in the Link Los sub-TLV should be treated as if it fails the exclude-any resource attributes filter.

If the answer to (a) is no for jitter SLAs, then any link that has the Anomalous bit set in the Unidirectional Delay Variation sub-TLV[I-D.previdi-isis-te-metric-extensions] should be treated as if it fails the exclude-any resource attributes filter.

2.3.2. Links entering the Anomalous State

When a link enters the Anomalous state with respect to a parameter, this is an indication that LSPs using that link might also no longer be in compliance with their performance bounds. It can also be considered an indication that something is changing that link and so it might no longer be trustworthy to carry performance-critical traffic. Naturally, which performance criteria are important for a particular LSP is dependent upon the LSP’s configuration and thus the SLA compliance of a link is indicated per performance criterion.

At the ingress of a TE tunnel, a TE tunnel may be configured to be sensitive to the Anomalous state of links in reference to latency, delay variation, and/or loss. Additionally, such a TE tunnel may be configured to either verify continued compliance, to switch
immediately to a standby LSP, or to move to a different path.

When a sub-TLV is received with the Anomalous bit set when previously it was clear, the list of interested TE tunnels must be scanned. Each such TE tunnel should either have its continued compliance verified, be switched to a hot standby, or do a make-before-break to a secondary path.

2.3.3. Links leaving the Anomalous State

When a link leaves the Anomalous state with respect to a parameter, this can serve as an indication that those TE tunnels, whose LSPs were changed when the link entered the Anomalous state, may want to reoptimize to a better path.

3. IANA Considerations

This document includes no request to IANA.

4. Security Considerations

This document is not currently believed to introduce new security concerns.

5. References

5.1. Normative References

[I-D.giacalone-ospf-te-express-path]

[I-D.previdi-isis-te-metric-extensions]

5.2. Informative References


Authors’ Addresses

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

John Drake
Juniper Networks
1194 N. Mathilda Ave.
Sunnyvale, CA  94089
USA
Email: jdrake@juniper.net

Dave Ward
Juniper Networks
1194 N. Mathilda Ave.
Sunnyvale, CA  94089
USA
Email: dward@juniper.net
Spencer Giacalone
Thomson Reuters
195 Broadway
New York, NY 10007
USA

Email: Spencer.giacalone@thomsonreuters.com

Stefano Previdi
Cisco Systems
Via Del Serafico 200
Rome 00142
Italy

Email: sprevidi@cisco.com

Clarence Filsfils
Cisco Systems
Brussels
Belgium

Email: cfilsfil@cisco.com
Abstract

Consider a BGP free core scenario with LDP running in the core. Suppose the edge BGP speakers PE1 and PE2 know about a prefix P/m via the external routers CE1 and CE2. If the edge LSR PE1 crashes or becomes totally disconnected from the core, it is desirable for a core LSR "P", that is carrying traffic to the failed edge LSR PE1, to immediately restore traffic by re-routing packets destined to the prefix P/m from the LSP terminating on PE1 to be forwarded over the LSP terminating on PE2, until BGP reconverges. If the packets originally flowing to the failed edge LSR PE1 are BGP labeled, then the repairing core LSR P must swap the label (corresponding to prefix P/m) advertised by the failed edge LSR PE1 with the label advertised for the same prefix by the edge LSR PE2 before re-routing the packets through an LSP terminating on PE2. To implement BGP edge node protection in a BGP-free LDP core, this document proposes an extension to LDP. This extension allows an LDP speaker running on an Edge LSR Node (e.g. PE1) to inform the LDP speakers running on core LSRs about the "Repair" edge LSR (e.g. PE2), as well as Repair LSR’s label for prefix P/m, if any.

Status of this Memo

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

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

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

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

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

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

This Internet-Draft will expire on September 5, 2012.

Copyright Notice

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

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

Table of Contents

1. Introduction...................................................3
   1.1. Conventions used in this document........................4
   1.2. Terminology.............................................4
   1.3. Problem definition.......................................5
2. The Proposed LDP Extension.....................................5
   2.1. The LDP "BGP Repair Path status" Code....................5
   2.2. The LDP "BGP Repair Path Status" TLV.....................6
   2.3. LDP Capability Negotiation................................8
   2.4. BGP Repair Status in a LDP Notification message.........8
3. BGP Repair Path Signaling Procedures...........................9
   3.1. Signaling a BGP Repair Path.............................9
   3.2. Updating a BGP Repair Path................................10
   3.3. Withdrawing a BGP Repair Path............................10

Bashandy              Expires September 5, 2012                [Page 2]
1. Introduction

In a BGP free core, where traffic is tunneled between edge routers/LSRs, (MP)BGP [2][3] speakers advertise reachability information about prefixes to edge routers only. For labeled address families, namely AFI/SAFI 1/4, 2/4, 1/128, and 2/128, an edge LSR assigns local labels to prefixes and associates the local label with each advertised prefix such as L3VPN [9], 6PE [10], and Softwire [8]. Suppose that a given edge LSR is chosen as the best next-hop for a prefix P/m. An ingress LSR receives a packet destined for the prefix P/m from an external router, and sends the packet to that egress LSR through an LSP terminating on the egress LSR. If the prefix P/m is a BGP labeled prefix, the ingress LSR pushes the BGP label advertised by the egress LSR before sending the packet into the LDP LSP terminating on the egress LSR. Upon receiving the packet from the core, the egress LSR takes the appropriate forwarding decision based on the content of the packet and/or the label(s) pushed on the packet.

In modern networks with redundancy in place, it is not uncommon to have a prefix reachable via multiple edge routers. One example is the best external path [7]. Another more common and widely deployed scenario is L3VPN [9] with multi-homed VPN sites. As an example, consider the L3VPN topology depicted in Figure 1.
As illustrated in Figure 1, the edge router PE0 is the primary NH for both 10.0.0.0/8 and 20.0.0.0/8. At the same time, both 10.0.0.0/8 and 20.0.0.0/8 are reachable through the other edge routers PE1 and PE2, respectively.

1.1. Conventions used in this document

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

In this document, these words will appear with that interpretation only when in ALL CAPS. Lower case uses of these words are not to be interpreted as carrying RFC-2119 significance.

1.2. Terminology

- LSR: Label Switched Router (In the context of this document, this refers to a router doing label switching on LDP and/or BGP labels)

- LSP: Label Switched Path
o Primary LSP: It is the LSP from the ingress PE to the primary egress PE. This is the LDP implementation of the primary tunnel defined in [6].

o Repair LSP: It is the LSP from the repairing P router to the repair egress PE. This is the LDP implementation of the repair tunnel defined in [6].

For the rest of the terms, refer to [6].

1.3. Problem definition

The general problem for the example shown in Section 1. is specified in [6]. The objective of this document is to specify an LDP [4] extension to let the primary egress PE inform repairing core router(s) about the repair path in an LDP core for both labeled and unlabeled protected prefixes. In other words, this is an LDP-based implementation of step 3 in [6]. Other problems, such as determining the repair PE, detecting the protected PE (node/connectivity) failure, and interactions between LDP and BGP on protected edge PE LSR, are beyond the scope of this document.

2. The Proposed LDP Extension

As specified in [4] section 3.5.1, an LDP speaker can use LDP Notification message to send its status or advisory information towards a peer. An LDP Notification message consists of a Status TLV and optional parameters, whereas "Status" TLV holds the status code being signaled. For an egress PE LSR to convey repair path info (for a BGP next-hop) to core LSRs, we propose to convey this information via a LDP Notification message that carries a new status code in "LDP Status" TLV, a new "BGP Repair Path Status" TLV and FEC TLV (corresponding to BGP next-hop) as optional parameters. This information is to be advertised to a peer only if the peer has signaled the support for "Unrecognized Notification" capability as specified in [5].

The proposed extensions are described more in details in following sub-sections.

2.1. The LDP "BGP Repair Path status" Code

A new LDP status code, namely "BGP Repair Path status" is defined that is to be set in the "Status Code" field of the "Status TLV" as defined in [4] section 3.4.6.
2.2. The LDP "BGP Repair Path Status" TLV

A new LDP TLV, namely "BGP Repair Path Status TLV", is defined to be used in an LDP Notification message under optional parameters section only if the Notification message status code is set to "BGP Repair Path status".

This TLV is an implementation of the repair path defined in [6] and is used to convey the information about the Repair edge LSR and its associated BGP label, if any, for traffic destination prefix P/m. This information is conveyed in the context of the protected primary BGP nexthop [6], whose information is carried in the FEC TLV. This document allows only one repair path per BGP nexthop.

The encoding of the "BGP Repair Path Status TLV" is as follows:

```
  0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|U|F| Repair Path TLV Type(TBD) |          Length               |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|A|L|P|     Reserved            |         Address Family        |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|                Repair PE Address (variable size)              |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|             Underlying Repair label (optional)                |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
```

Figure 2 Format of BGP Repair Path Status TLV

The fields are as follows

- **U/F:**
  - Must be set to 1/0 respectively so that this TLV can be ignored if not known or not supported.

- **Repair Path TLV Type:**
  - IANA assigned TLV value

- **A bit**
  - Indicates if Repair Path information is to be "added" or "removed". MUST be set to 1 to signal addition of the information, and set to signal addition of the information, and set to

- **L Bit:**
Indicates whether optional "Underlying Repair Label" [6] field is present or not. Must be set to 1 if the TLV also contains/encodes "Underlying Repair Label", else must be set to 0. This bit MUST be set to 0 when A bit is set to 0.

P Bit:

If set, then the label in the "Underlying Repair label" sub field MUST be pushed instead of swapped. The "P" bit has the same semantics of the "Push" flag in [6]. If the "L" bit is zero then the "P" bit MUST be set to zero.

Reserved:

MUST be zero on transmit and ignored in receipt

Address Family

Identical to the "Address Family" field used in encoding the Prefix FEC element value specified in Section 3.4.1 in [4]. May be different of the address family of the prefix in the FEC

Repair PE Address

The length of this field is dependent on the Address Family field. This is either the 4 octet IPv4 address or 16 octet IPv6 address of a host address belonging to repair PE. The encoding of the Repair PE Address is identical to the encoding of the value of the IPv4 or IPv6 Transport Address TLV specified in Section 3.5.2 in [4]. This field encodes the "Repair Next-hop" defined in [6].

Underlying Repair Label (optional)

If included in the TLV, it is a single label defined in accordance to Generic Label TLV specified in Section 3.4.2.1 in [4]. This field encodes the "Underlying Repair Label" defined in [6].

Length

The length (in octets) of the TLV following the "Length" field. For example, the length MUST be 8 with IPv4 Repair PE address or 20 with IPv6 Repair PE address when L bit is set to 0, and MUST be 12 and 24 octets otherwise for IPv4 and IPv6 address families, respectively.
2.3. LDP Capability Negotiation

This BGP Repair Path information is to be computed by an edge PE LSR under a user configuration control. Once computed, this information can be unsolicitedly sent to core P LSRs for edge PE node protection, and it is up to the receiving P LSR to store and use this information to protect the edge PE LSR.

Given above procedures, no new LDP capability [RFC5561] negotiation is needed between PE and P LSRs to support this feature extensions. However, to ensure backward compatibility and deterministic behavior, it is required that this information be sent to only those P LSRs that support "Unrecognized Notification" capability as specified in [5]. This will ensure that these new status and TLV does not cause any issue at a receiving P LSR if not known or not supported, and be discarded in accordance with "Unrecognized Notification" procedures.

2.4. BGP Repair Status in a LDP Notification message

The general format of an LDP Notification message that carries information regarding BGP repair path is as follows:

```
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
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|0|   Notification (0x0001)     |      Message Length          |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|                       Message ID                              |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|                       Status TLV                             |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|                   BGP Repair Path Status TLV               |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|                        FEC TLV                             |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
Figure 3 : LDP Notification message with BGP Repair Status
```

The "Status TLV" status code is set to "BGP Repair Path status" to indicate that the message is used to convey BGP Repair Path information. When this status code is set, a Notification message MUST contain both "BGP Repair Status TLV" and a "FEC TLV" in the message.

Since this notification does not refer to any particular message, the "Message ID" and "Message Type" fields in the "Status TLV" MUST be set to 0.
The "BGP Repair Path Status TLV" is encoded as described earlier.

The FEC TLV MUST contain a single "Prefix FEC Element" that encodes the BGP nexthop information as host prefix. This field encodes the "Primary Next-hop" defined in [6].

3. BGP Repair Path Signaling Procedures

To describe the signaling procedures clearly, let’s first assume that:

- Protected edge LSR is PE1, Repair edge LSR is PE2 and repairing core LSR is P1 LSR router, and P2 is also a core LSR that does not support BGP Repair Path functionality.
- IPv4 Host addresses for PE1 and PE2 are A1 and A2 respectively.
- Protected BGP IPv4 prefix is P/m.
- BGP label assigned by PE2 for P/m is L2.
- P1 and P2 LSR support LDP "Unrecognized Notification" Capability.

3.1. Signaling a BGP Repair Path

An operator enables this feature on PE1 using a configuration knob. The PE1 computes Repair PE information (PE2 address A2, and PE2 BGP label L2) for a given BGP prefix P/m. PE1 encodes the LDP Notification to advertise the Repair path information to all those core LSR peers (including P1/P2 LSRs) who have advertised "Unrecognized Notification" Capability TLV for given LDP session.

The PE1 LSR encodes the following TLVs:

- "Status" TLV: status code is set to "BGP Repair Path status" and "Message Type" and Message ID fields set to 0.

- "BGP Repair Path Status" TLV: This TLV is encoded with A-bit set to 1, L-bit set to 1, P-bit set to 0, Address Family set to 1 (IPv4), "Repair PE Address" field populated with A2, and "Repair PE’s Label" field set to L2.

- "FEC" TLV: This TLV is encoded with a single "Prefix FEC Element" whose Address Family is set to 1 (IPv4) to indicate IPV4 prefix, and prefix P/m encoded accordingly.

After encoding these TLVs, PE1 LSR bundles them in an LDP Notification message, as shown in Figure 3, and sends them to its upstream core peer P1 and P2 LSRs.
On receipt of this information, P1 stores this information and uses this to fast reroute the BGP destined traffic to PE2 upon PE2 node/connectivity failure detection. P2, on the other hand, does not recognize this new status in the LDP Notification message and hence discards it silently.

In order to be able to protect primary BGP nexthops, it is required that the repairing P LSR (P1) must have a LSP terminating on Repair PE host prefix (as indicated by "Repair PE Address" field in the received "BGP Repair Path Status" TLV).

3.2. Updating a BGP Repair Path

The repair path information is identified by

- the "primary next-hop" encoded in the "FEC TLV" shown in Figure 3 and,
- the "Repair Next-hop" encoded in the "Repair PE Address" shown in Figure 2.

Once a repair path has been signaled to P core LSR, it can be updated by simply sending another LDP Notification message using the procedures described in the previous section.

Upon receipt of a new repair path information, the LDP receiver (P1 LSR) MUST discard any previously learnt Repair information from the sending PE1 LSR, and update it with the most recently received.

3.3. Withdrawing a BGP Repair Path

Once a repair path has been signaled to P core LSR, it can be withdrawn by simply sending another LDP Notification message using the procedures described in the previous section with following changes:

- Set A-bit, L-bit, and P-bit in "BGP Repair Path Status" TLV to 0
- No "Underlying Repair Label" field included

Upon receipt of a withdrawal of a repair path, the LDP receiver (P1 LSR) MUST discard any previously learnt Repair information from the sending PE1 LSR for a given BGP prefix.

4. Security Considerations

No additional security risk is introduced by using the mechanisms proposed in this document

Bashandy Expires September 5, 2012
5. IANA Considerations

This document introduces the following new protocol elements that require code point assignment by IANA:

- "BGP Repair Path status" status code from "LDP Status Code Name Space" registry (requested code point: 0x00000050)
- "BGP Repair Path Status TLV" from "LDP TLV Type Name Space" registry (requested code point: 0x050F)

6. Conclusions

This document proposes a an LDP extension that allows an egress PE to advertise a repair path consisting of a repair egress PE and an underlying label to repairing core router. Advertising this information to core routers allows core routers to provide FRR protection against primary egress PE node failure while keeping the core BGP-free.

7. References

7.1. Normative References


7.2. Informative References


8. Acknowledgments

Special thanks to Keyur Patel and Alton Lo for the valuable help.

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

Authors’ Addresses

Ahmed Bashandy
Cisco Systems
170 West Tasman Dr, San Jose, CA 95134
Email: bashandy@cisco.com

Kamran Raza
Cisco Systems,
2000 Innovation Dr., Kanata, ON K2K 3E8, Canada
EMail: skraza@cisco.com
Abstract

Recent proposals have extended the use of Bidirectional Forwarding Detection to detect data plane failures in multipoint networks. One desirable application of Multipoint BFD [MP-BFD] is to detect data path failures in point-to-multipoint Label Switched Paths (P2MP LSPs). The scope of this document is to discuss the applicability of multipoint BFD for fault detection in the data plane of P2MP LSPs. The document extends the techniques and mechanisms mentioned in [RFC5884] and applies them to MPLS P2MP 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 September 6, 2012.

Copyright Notice

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

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

1.  Introduction .................................................. 3
   1.1.  Terminology .............................................. 3
   1.2.  Conventions ................................................ 3
2.  BFD for P2MP LSPs: Motivation ................................. 4
3.  Notes on bootstrapping Multipoint BFD for P2MP LSPs ............. 5
4.  Proposed Solutions .............................................. 6
   4.1.  Periodic echo requests with "Do not reply" ................ 6
   4.2.  Application driven Ping ................................... 6
5.  Security Considerations ......................................... 8
6.  IANA Considerations ............................................ 9
7.  Acknowledgments ................................................ 10
8.  Normative References ........................................... 11
Authors’ Addresses ................................................ 12
1. Introduction

Application of BFD for fault detection in the data path of point-to-point MPLS LSPs is addressed in [RFC5884]. Since point-to-multipoint LSPs are not less vulnerable to data path failures than their point-to-point counterparts, this document extends the techniques used in [RFC5884] such that they can be applied to P2MP LSPs. This document stresses the reuse of existing LSP ping mechanism to bootstrap the BFD sessions for point-to-point LSPs and its application to point-to-multipoint LSPs and multipoint BFD in order to simplify implementation and network operations.

1.1. Terminology

The terminology used in this document for MPLS OAM can be found in [RFC4379].

The terminology for multipoint BFD can be found in [MP-BFD].

1.2. Conventions

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 [RFC5884].
2. BFD for P2MP LSPs: Motivation

As with P2P LSPs, there is a requirement for faster detection of data plane failures in P2MP LSPs. Existing mechanisms like P2MP LSP ping and P2MP trace route are expensive in terms of the amount of control plane processing required and are catered towards fault isolation or on-demand fault detection at or by the ingress LSR. Multipoint BFD offers a way for the egress LSR to detect a loss of connectivity to a particular ingress along the multicast data path. Some of the advantages of BFD as listed in [RFC5884] which also apply to P2MP LSPs are:

a) Support for fault detection for a greater number of LSPs.

b) Fast detection. Considering how BFD is most suitable to be implemented in the hardware or firmware due to its fixed packet format, BFD can be used to achieve fault detection with sub-second granularity. There are several advantages of sub-second detection of data plane failures in any P2MP LSP at the the egress LSR, a couple of them being:

1) Multicast Live-Live

For applications that make use of multicast live-live where traffic is pulled from two different streams and merged at the egress, fast detection of failures on any one of the streams at the egress of the P2MP LSP might be desirable to provide seamless service by switching over to the backup stream to avoid traffic interruptions. Multipoint BFD is a suitable choice to detect these kinds of data path failures and to notify the applications of these failures.

2) LSPs with discontinuous traffic flow

One way of doing OAM over a P2MP LSP today is for an application to monitor the data traffic coming over the P2MP LSP and correlate that with the health of the data path. For applications that do not have continuous traffic flowing, it is desirable to monitor the connectivity of the data path by some other means. Multipoint BFD can be used for this purpose and the detection can be done at the LHR instead of the receiver.
3. Notes on bootstrapping Multipoint BFD for P2MP LSPs

The procedures of Multipoint BFD are described in [MP-BFD]. Although Multipoint BFD tries to be generic in nature regardless of the technology used to establish the multipoint tunnel, for P2MP LSPs there are certain scenarios where making the association of an incoming BFD packet with its corresponding P2MP LSP becomes a challenge at the egress. For example in case of Penultimate Hop Popping (PHP) for a single label stack P2MP LSP, incoming BFD packets at the egress do not contain enough information to associate them with a particular P2MP LSP.

A similar problem of bootstrapping was encountered when attempting to run BFD over P2P LSPs. These problems have been resolved by following the procedures mentioned in [RFC5884]. The procedures explained in [RFC5884] can be applied to P2MP LSP ping to bootstrap the multipoint BFD session. Although the mechanisms of [RFC5884] when applied to P2MP LSP Ping and Multipoint BFD would work, it may not necessarily be optimal for a P2MP environment, primarily because:

a) Unidirectional Nature.

Multipoint BFD by nature can be unidirectional. Multipoint BFD in its primary form is designed for the egress LSR to detect a loss of connectivity to the ingress along the multipoint path along with providing the ingress with some optional mechanisms to track the connectivity to certain egress LSRs. When running multipoint BFD in this form, the ingress does not have any need to learn the BFD discriminator of the egress to establish the BFD session. Even if the ingress does wish to track the connectivity to certain egress LSRs, this would happen over the unicast path and if the procedures of [MP-BFD] are followed then the unicast BFD sessions can be established without any explicit out of band bootstrapping.

b) Scalability

The procedures of P2MP LSP ping are described in [RFC6425]. Although [RFC6425] talks about various mechanisms to rate limit the echo reply messages coming into the ingress, in a large deployment, it may not be optimal to solicit echo replies from all the egress LSRs. Soliciting replies from all the egress routers may lead to the ingress being flooded by a large number of messages in a short duration.
4. Proposed Solutions

4.1. Periodic echo requests with "Do not reply"

Applying the procedures of [RFC5884] to P2MP LSPs, periodic P2MP LSP Ping echo request messages MUST be sent by the ingress LSR to the egress LSR along the same data path as the P2MP LSP. These echo messages SHOULD contain the local discriminator of the Multipoint BFD session established by the ingress LSR for the P2MP LSP. The rate of generation of echo request messages must be considerably slower than the rate of generation of BFD packets at the ingress LSR. If the primary purpose of running P2MP LSP Ping is to facilitate the establishment of the BFD session and its association with the corresponding P2MP LSP at the egress LSR, then an implementation MAY set the value of the Reply Mode field in the message to 1 (Do Not Reply) thus indicating the egress LSRs to suppress the generation of echo reply messages. By doing so, an implementation avoids the possibility of congestion at the ingress LSR by not soliciting reply messages from all the egress LSRs.

Note that even when P2MP LSP Ping is used for bootstrapping BFD sessions, periodic transmission of the echo request messages are required by the ingress LSR primarily because in a P2MP LSP environment, there are valid scenarios where the ingress LSR may not be aware of the existence of different egress LSRs. In such a case, the ingress LSR may not have sufficient information to decide when to send out the LSP ping with the BFD discriminator to bootstrap the BFD session at the newly attached egress LSRs. By periodically transmitting LSP Ping echo request messages, the ingress LSR ensures that any newly attached egress LSR will be able to bind the BFD session to the P2MP LSP when it receives the next periodic echo request message.

It is highly possible that an egress LSR will start receiving the multipoint BFD packets before the LSP ping from the ingress. How an implementation chooses to handle these incoming BFD packets is outside the scope of this document and is left to the implementation. The egress LSR SHOULDN'T follow the procedures mentioned in [RFC5884] to validate the PING packet and then use the discriminator in the P2MP echo request message to either bind an existing BFD session to the P2MP LSP or setup a new BFD session by accepting BFD packets with the discriminator value it has received in the echo request message from the ingress.

4.2. Application driven Ping

The rate of generation of periodic echo requests by the ingress LSR is considerably slower than the rate of generation of BFD packets.
because the processing of PING packet is a control plane processing extensive operation. For P2MP LSPs, if an implementation of P2MP ping which is used for bootstrapping the BFD session has chosen to solicit replies from all the egress LSRs, then it is crucial that these ping packets are rate limited to avoid congestion at the ingress LSR. Slowing down the rate of generation of Ping packets also in turn slows down the bootstrapping of the BFD sessions at the egress LSRs who might have newly attached themselves to the P2MP LSP. This may not be desirable for certain applications which may want to leverage Multipoint BFD to fast detect failures on the data path of a P2MP LSP at the egress.

Application driven expedited PING is one way to solve this problem. The decision as to when to send the expedited PING is left to the implementation and outside the scope of this document, one use of expedited ping is that if an application running on the ingress LSR can detect a new node attaching itself as the egress to its P2MP LSP, then an implementation MAY choose to send an expedited PING packet over the P2MP data path. An implementation MAY also choose to solicit an echo reply from one or a subset of egress LSRs using the Egress Address P2MP Responder Sub-TLV(as described in [RFC6425]) in the expedited echo request messages, however if an implementation chooses to solicit replies from a subset of egress LSRs using the P2MP LSP Ping, then it MUST ensure to avoid congestion at or near the ingress when the replies arrive.

Periodic echo requests as described in section 4.1 can be used in conjunction with application driven expedited ping to provide an optimal solution to bootstrap BFD sessions. Periodic messages ensures that in a dynamic P2MP LSP environment, where the egress LSRs are constantly attaching or detaching themselves from the P2MP LSPs, the association of BFD packets to a P2MP LSP is possible at the egress LSR. For applications running in the ingress that do track the egress LSRs of a P2MP LSP, expedited PING could be used to quickly bootstrap the BFD session at the egress without having to wait for the next periodic PING, while doing so the application MAY also choose to solicit replies from a subset of the egress LSRs.
5. Security Considerations

No new security issues are introduced beyond those that are described in [RFC6425] and [MP-BFD].
6. IANA Considerations

This draft does not have any request for IANA.
7. Acknowledgments

The authors would like to acknowledge the authors of [RFC6425] and
the authors of [RFC5884] for their work, which is substantially re-
used in this document. The authors would like to thank Santosh PK,
Shivakumar Channalli and Harish Sitaraman for their reviews and
feedback on this material.
8. Normative References


Authors’ Addresses

Vikas Hegde
Juniper Networks
Embassy Business Park
Bangalore, KA  560093
India

Email: vikashegde@juniper.net

Chandrasekar R
Juniper Networks
Embassy Business Park
Bangalore  560093
India

Email: csekar@juniper.net
Extensions to RSVP-TE for P2MP LSP Egress Local Protection
draft-chen-mpls-p2mp-egress-protection-05.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 September 14, 2012.

Copyright Notice

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

This document is subject to BCP 78 and the IETF Trust’s Legal Provisions Relating to IETF Documents
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) ....... 6
   4.4. Detection of Egress Node Failure ............... 6
5. Egress Local Protection with FRR ....................... 7
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 ...... 8
       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
8. IANA Considerations .................................. 11
9. Acknowledgement ...................................... 11
10. References ........................................... 11
    10.1. Normative References ....................... 11
    10.2. Informative References ................... 12
Authors’ Addresses ..................................... 12
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, which is an intermediate node between the ingress node and the egress node of the protected LSP. The second method is a facility bypass backup protection method, where a bypass backup P2P LSP tunnel is created using MPLS label stacking to protect a potential failure point for a set of P2P LSP tunnels. The bypass backup tunnel can protect a set of P2P LSPs having similar backup constraints.

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

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](image)

4.2. Set up of Backup sub LSP

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

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

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

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

4.4. Detection of Egress Node Failure

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

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

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

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

6.1.2. EGRESS_BACKUP_SUB_LSP IPv6 Object

The class of the EGRESS_BACKUP_SUB_LSP IPv6 object is the same as that of the S2L_SUB_LSP IPv6 object defined in RFC 4875. The C-Type of the object is a new number 4, or may be another number assigned by Internet Assigned Numbers Authority (IANA).
EGRESS_BACKUP_SUB_LSP Class = 50,
EGRESS_BACKUP_SUB_LSP_IPv6 C-Type = 4

+-----------------+-----------------+-----------------+-----------------+
|                  |                  |                  |                  |
|                  |                  |                  |                  |
|                  |                  |                  |                  |
|                  |                  |                  |                  |
|                  |                  |                  |                  |
|                  |                  |                  |                  |
+-----------------+-----------------+-----------------+-----------------+

Egress Backup Sub LSP IPv6 destination address
(16 bytes)

....

+-----------------+-----------------+-----------------+-----------------+
|                  |                  |                  |                  |
|                  |                  |                  |                  |
|                  |                  |                  |                  |
|                  |                  |                  |                  |
|                  |                  |                  |                  |
+-----------------+-----------------+-----------------+-----------------+

Egress IPv6 address
(16 bytes)

....

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

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

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

7.2. Processing of Path Message

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

If the previous hop node of the primary egress node receives the Path message with an egress backup sub LSP descriptor list, it generates a new Path message based on the information in the EGRESS_BACKUP_SUB_LSP (and according to EB-SERO if it exists) containing the backup egress node.
The format of this new Path message is the same as that of the Path message defined in RFC 4875. This new Path message is used to signal the segment of a special S2L sub-LSP from the previous hop node to the backup egress node. The new Path message is sent to the next-hop node along the path for the backup sub LSP.

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

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

8. IANA Considerations

TBD

9. Acknowledgement

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

10. References

10.1. Normative References


[ RFC3209 ] Awduche, D., Berger, L., Gan, D., Li, T., Srinivasan, V.,


10.2. Informative References


Authors’ Addresses

Huaimo Chen
Huawei Technologies
Boston, MA
USA

Email: Huaimochen@huawei.com

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

Email: Ning.So@verizonbusiness.com
Autumn Liu
Ericsson
CA
USA

Email: autumn.liu@ericsson.com

Olufemi Komolafe
Cisco Systems
96 Commercial Street
Edinburgh, EH6 6LX
United Kingdom

Email: femi@cisco.com

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

Email: le-liu@kddilabs.jp
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 September 14, 2012.

Copyright Notice

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

This document is subject to BCP 78 and the IETF Trust’s Legal Provisions Relating to IETF Documents
Table of Contents

1. Introduction ......................... 3
2. Terminology ............................. 4
3. Conventions Used in This Document .......... 4
4. Mechanism ................................ 4
   4.1. An Example of Ingress Local Protection ........ 4
   4.2. Set up of Backup P2MP sub Tree .............. 5
   4.3. Forwarding State for Backup P2MP sub Tree ...... 5
   4.4. Detection of Failure around Ingress .......... 6
5. Ingress Local Protection with FRR ........... 7
6. LSP Information Message ................... 7
   6.1. Format of LSP Information Message .......... 8
   6.2. Processing of LSP Information Message ........ 8
   6.3. Discussions on Other Approaches ........... 9
7. PATH Messages for Backup P2MP sub Tree ....... 9
   7.1. Construction of PATH Messages .............. 9
   7.2. Processing of PATH Messages ............. 10
8. IANA Considerations ..................... 10
9. Acknowledgement .......................... 10
10. References ............................. 11
   10.1. Normative References .................... 11
   10.2. Informative References ................... 11
Authors’ Addresses ......................... 12
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, which is an intermediate node between the ingress node and the egress node of the protected LSP. The second method is a facility bypass backup protection method, where a bypass backup P2P LSP tunnel is created using MPLS label stacking to protect a potential failure point for a set of P2P LSP tunnels. The bypass backup tunnel can protect a set of P2P LSPs that have similar backup constraints.

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

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

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

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

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

3. Conventions Used in This Document

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

4. Mechanism

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

4.1. An Example of Ingress Local Protection

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

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

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

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

A failure of the ingress node can be detected through a BFD session between the ingress node and the backup ingress node 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" (FFR 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. LSP Information Message

LSP information messages are used to transfer the information about a P2MP LSP to a backup ingress node from an ingress node. The destination address of the LSP information message is that of the backup ingress node. This section describes the format of an LSP information message and processing of the message. It also discusses
other approaches for transferring the information about a P2MP LSP to a backup ingress from an ingress.

6.1. Format of LSP Information Message

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

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

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).

6.2. Processing of LSP Information Message

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

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

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

6.3. Discussions on Other Approaches

The information about a P2MP LSP may be transferred through other approaches from the ingress node of the LSP to the backup ingress node. One approach is to use OSPF Opaque LSA. The main reason for giving up this option is that more parts need to be changed. Both OSPF and RSVP-TE need to be modified.

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. PATH Messages for Backup P2MP sub Tree

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

7.1. Construction of PATH Messages

When the backup ingress node receives a P2MP LSP information message, it checks to see if anything has been changed. If the message is a new message or the information in the message has been changed, then the PATH messages for the backup P2MP sub tree are to be constructed as follows.
First, a path to the next-hop nodes of the ingress node HAS to be computed. The path MUST satisfy the constraints for the P2MP LSP and not go through the ingress node.

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

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

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

7.2. Processing of PATH Messages

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

8. IANA Considerations

TBD

9. Acknowledgement

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

10.1. Normative References


10.2. Informative References

Authors’ Addresses

Huaimo Chen
Huawei Technologies
Boston, MA
USA

Email: Huaimochen@huawei.com

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

Email: Ning.So@verizonbusiness.com

Autumn Liu
Ericsson
CA
USA

Email: autumn.liu@ericsson.com

Olufemi Komolafe
Cisco Systems
96 Commercial Street
Edinburgh, EH6 6LX
United Kingdom

Email: femi@cisco.com
Lei Liu
KDDI R&D Lab Inc.
2-1-15
Ohara Fujimino-shi, Saitama
Japan

Email: le-liu@kddilabs.jp
Abstract

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

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

Status of this Memo

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

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

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

This Internet-Draft will expire on September 13, 2012.

Copyright Notice

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

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

Table of Contents

1. Introduction .................................................. 3
2. Conventions Used in this Document ........................... 5
   2.1. Acronyms .................................................. 5
   2.2. Definitions and Terminology ................................ 5
3. Shared Mesh Protection Architecture .......................... 6
   3.1. Shared Mesh Protection Group ............................... 7
   3.2. Shared Start and End Nodes ............................... 8
   3.3. SMP protocol communication channel ....................... 10
   3.4. Network planning for SMP ............................... 10
   3.5. Preemption and race conditions ......................... 11
   3.6. SMP Overview ............................................ 12
       3.6.1. LP Protocol extensions for shared protection ....... 12
       3.6.2. Notifying protection switching event ............... 13
       3.6.3. Requesting lockout of protection .................. 13
       3.6.4. Resolving race conditions ......................... 14
4. Protocol ...................................................... 15
   4.1. PDU Format .............................................. 15
   4.2. Message Transmission .................................... 17
5. Operation of Shared Mesh Protection ........................... 17
6. Manageability Considerations .................................. 20
7. IANA Considerations .......................................... 21
8. Security Considerations ...................................... 21
9. References .................................................. 21
   9.1. Normative References .................................... 21
   9.2. Informative References ................................. 21
Authors’ Addresses ............................................. 22
1. Introduction

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

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

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

[RFC5654] also establishes that MPLS-TP should support shared protection (Requirement 68). 1:n end-to-end protection uses one protection path to protect n working paths between the same two endpoints. This improves overall network utilization, but the resource (bandwidth) allocated to a protection path is typically not sufficient to protect multiple simultaneous failures on different working paths. If multiple working paths require concurrent protection switching, the path with the highest priority should be protected as described in [RFC6372].

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

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

In mesh protection, network resources may be shared to provide protection for working paths that do not share the same end nodes at the edge of a protection domain. This type of protection can make very efficient use of network resources, but requires coordination of several segments in order to ensure that only a single traffic flow is switched to the protection resources at any time.

[RFC4428] defines two shared mesh recovery schemes named (1:1)^n and (M:N)^n. The (1:1)^n recovery scheme is a simple case of (M:N)^n recovery scheme. In (1:1)^n protection, n working paths are protected by n dedicated protection paths while sharing the same protection bandwidth. The protection bandwidth can be optimized to allow only one of the n working paths to be protected at any time. In this case, it achieves network utilization similar to 1:n protection.

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

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

The shared mesh protection mechanism defined in this document utilizes the existing MPLS-TP linear protection switching mechanism, and assumes that the protection paths are established and ready to forward data prior to a failure. Upon detection of a failure on a working path, only the two end nodes of the failed working path exchange their linear protection protocol messages to switch data traffic. No explicit activation procedure to switch data traffic to the protection path is needed in the intermediate nodes along the protection path. However, the intermediate nodes that are part of the shared segments need to coordinate the resource allocation on the shared nodes and this coordination will be addressed by the protocol proposed in this document.
2. Conventions Used in this Document

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

2.1. Acronyms

This draft uses the following acronyms:

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

2.2. Definitions and Terminology

This document defines two protection domains as follows:

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

- Shared mesh protection (SMP) domain: A protection domain for protecting a number of P2P or P2MP LSPs. It consists of a number of end-to-end linear protection domains. Each end-to-end linear protection domain shares protection resources with other domains. The shared protection resource may be a node, link, transport path segment or concatenated transport path segment. A shared mesh protection switching protocol runs within the domain.

In addition, we define the following:
Protection segment: A protection segment is a portion of the end-to-end protection path. An end-to-end protection path can be broken down into separate protection segments. A protection segment may either be a shared protection segment or a dedicated protection segment. A shared protection segment is a segment that has a set of resources, e.g. link bandwidth that is reserved as shared amongst several end-to-end protection paths. A dedicated protection segment is a segment of the end-to-end protection path whose resources are reserved for use only by that protection path.

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

3. Shared Mesh Protection Architecture

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

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

Thus, it is possible for the network resources of segment PQR to be shared by the two protection paths. In this way, shared mesh protection can substantially reduce the amount of network resources that need to be reserved to provide protection of the multiple paths.
within the same protection group.

A----B----C----D----E
\       /       /
\       /       /
F----G----H----J----K----L
\       /       /
\       /       M----N----N
V----W----X----Y----Z

Figure 1: A Shared Mesh Protection Topology

3.1. Shared Mesh Protection Group

In Figure 1, two working paths, ABCDE and VWXYZ and their corresponding protection paths, APQRE and VPQRZ, are considered a Shared Mesh Protection Group (SMPG). As pointed out above, each protection path belonging to a SMPG is an individual LSP, but it shares the resources with others in a segment. The resources that are being shared are the nodes, ports, links and bandwidth of the segment.

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

A------B-------C     D------E
\          /     /        \         /
\          /     /          \        /
F---G----H-----J------K-----L
\          /        \     \         /
\          M----------N     \        /
V-------W-------X-------Y-------Z

Figure 2: Shared Mesh Protection Groups

To further clarify, consider the mesh network in Figure 2. In this figure we have the following working paths and corresponding protection paths:
In this network we would define three SMPG – characterized by the three shared segments –

- S1 segment G-H – shared by W1 and W4
- S2 segment J-K – shared by W2, W3, and W4
- S3 segment K-L – shared by W2 and W4

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

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

3.2. Shared Start and End Nodes

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

A SSN is the first node of a unidirectional shared protection segment. For example, in Figure 1, node P is a SSN on unidirectional protection paths A-P-Q-R-E and V-P-Q-R-Z. SSN may act as a Maintenance Entity Group Intermediate Point (MIP) for each protection path sharing the same protection resources.

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

Both SEN and SSN are involved in coordinating the use of the unidirectional sharing protection segment during the shared mesh.
protection operation.

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

<table>
<thead>
<tr>
<th>Protection paths</th>
<th>Shared protection segment</th>
<th>SSN</th>
<th>SEN</th>
</tr>
</thead>
<tbody>
<tr>
<td>A-P-Q-R-E, V-P-Q-R-Z</td>
<td>P-Q-R</td>
<td>P</td>
<td>R</td>
</tr>
<tr>
<td>E-R-Q-P-A, Z-R-Q-P-V</td>
<td>R-Q-P</td>
<td>R</td>
<td>P</td>
</tr>
</tbody>
</table>

Figure 3 shows a more complex example of the shared mesh protection domain. Three working paths ABC, DEF, and GHJ are protected by the protection paths APQC, DRSF, and GPQRSJ, respectively. Table 2 summarizes the relationship between SSN and SEN of the shared protection segments and the related protection paths in the protection domain illustrated in Figure 3.

<table>
<thead>
<tr>
<th>Protection paths</th>
<th>Shared protection segment</th>
<th>SSN</th>
<th>SEN</th>
</tr>
</thead>
<tbody>
<tr>
<td>A-P-Q-C, G-P-Q-R-S-J</td>
<td>P-Q</td>
<td>P</td>
<td>Q</td>
</tr>
<tr>
<td>C-Q-P-A, J-S-R-Q-P-G</td>
<td>Q-P</td>
<td>Q</td>
<td>P</td>
</tr>
<tr>
<td>D-R-S-F, G-P-Q-R-S-J</td>
<td>R-S</td>
<td>R</td>
<td>S</td>
</tr>
<tr>
<td>F-S-R-D, J-S-R-Q-P-G</td>
<td>S-R</td>
<td>S</td>
<td>R</td>
</tr>
</tbody>
</table>

Figure 3: A More Complex Mesh Protection Example

Table 2: SSN/SEN in Figure 3
3.3. SMP protocol communication channel

The MPLS-TP Framework [RFC5921] defines the concept of a Sub-Path Maintenance Entity (SPME) and together with [RFC5586] defines the use of the Generic Associated Channel (G-ACh) for communication of MPLS-TP control protocols between the endpoints of a maintenance entity. While the usual utility of a SPME is to allow tunneling of transport traffic while monitoring the segment with in-band connectivity verification messages, it is possible to use concept of a SPME to describe a LSP that is dedicated to carry a control protocol over the G-ACh between the endpoints of the shared protection segment and the endpoints of the protection paths within the SMPG.

For example, referring to the network in Figure 3, we would configure the following SPME (without identifying the intermediate nodes):

A-P, G-P, P-Q, Q-C and Q-J (for SMPG 1 sharing P-Q), and D-R, G-R, R-S, S-F and S-J (for SMPG 2 sharing R-S).

These SPMEs are bidirectional LSPs that are not used to carry any data traffic, but only the SMP protocol traffic described in Section 4.

The communication channel between the SSN and SEN of the shared protection segment and between themselves and the endpoints of the protection paths within the SMPG is to coordinate the allocation of the shared segment to a single protection path during a protection switching condition. This process is described more fully in Section 3.6

3.4. Network planning for SMP

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

- Preparing the working and protection paths for the different services that require protection.

- Determining which working paths are disjoint and so will not be subject to common failures. It should be clear that working paths within the same SRLG should not be included in the same SMPG.

- Identifying which protection paths share network resources and can constitute a SMPG. Signaling or configuring the proper path information for the shared segment endpoints to allow for communication between the corresponding endpoints of the shared segment and the protection path.
o Assigning Protection Switching Priority and a path identifier for each working path within a shared mesh protection domain.

o Ensuring that working paths of high Protection Switching Priority do not share resources on their protection paths in such a way that would mean that one of them could be unprotected.

o Enabling the necessary shared mesh protection functions at the endpoints of the shared protection segments. This includes preparing the different SPME used for communication between the corresponding endpoints of the shared protection segments and the protection paths, as well as between the endpoints of the shared protection segment.

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

3.5. Preemption and race conditions

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

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

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

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

3.6. SMP Overview

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

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

3.6.1. LP Protocol extensions for shared protection

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

- Function to generate a protection switching event message to the SEN when a switching trigger occurs at the end-to-end LP domain. Switching to the protection path or reverting to the working path should be notified.

- Function to take a Lockout of Protection (LoP) request message from the SEN, and incorporate it as the Lockout of Protection (LoP) command assertion or clearance.

- Function to acknowledge the SEN when the LP domain completes the LoP operation.
3.6.2. Notifying protection switching event

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

At the same time, for the operation of SMP, the protection path endpoint notifies its protection switching event to SENs by sending a "Protection Switching (PS) Event" message.

The PS Event message MUST be transmitted immediately when an endpoint of the end-to-end LP domain changes its selector position either from working to protection or vice versa. The event message SHALL be transmitted over the SPME that is configured between the protection path endpoint and the SEN, using the G-ACh. When bidirectional protection switching is being used by the working path, both endpoints will transmit the event messages to their corresponding SENs using the properly configured SPME. When unidirectional protection switching is being used and a unidirectional failure is detected, only the detecting endpoint will send the messages to its corresponding SENs.

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

3.6.3. Requesting lockout of protection

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

The SEN determines which of the LP domains (within the SMPG) have a lower or equal priority to that of the notifying LP domain. The SEN then sends a "Lockout of Protection (LoP) Request" message to the endpoints of these protection paths that is equivalent to a "Lockout of Protection" operator command. This prevents any protection switching actions in those LP domains. For those LP domains having higher priorities no request is transmitted and those LP domains may continue to perform protection switching actions which they require.

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

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

When the protection path endpoint completes LoP operation, it SHOULD immediately reply with a "LoP Acknowledgement" message to inform the completion of the LoP operation to the SEN.

To minimize potential congestion that may occur when a protection path having higher priority pre-empts other protection paths having lower priorities, the SEN SHOULD block forwarding traffic from the protection paths having lower priorities until it receives the LoP Ack message from the endpoints of those protection paths.

When a SEN receives a PS Event message indicating that the shared protection resources are being released, i.e. the LP domain is reverting to normal state, it sends a LoP Request message to the endpoints of all the protection paths in the SMPG that were previously locked (i.e. those with equal or lower priority) to clear the LoP command. The endpoint of the protection path that receives this message SHALL react as if a Clear command was received.

3.6.4. Resolving race conditions

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

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

The race condition can occur only when more than one protection paths
are configured to have the same Protection Switching Priority within a SMP domain.

When the SSN receives the RA Notification message from the SEN, it first checks whether it has received a PS Event message from an endpoint of the other LP domain having the same Protection Switching Priority that corresponds to the LP domain sending the RA notification message.

If both have the same priority, the SSN compares the working path identifier and sends an RA Ack message to the SEN only when it is determined that the working path identifier contained in the RA Notification message to have been allocated the shared protection resources. Each working path or LP domain has a unique identifier within a SMP domain and rules for deciding the priority will be defined in later.

The SEN does not perform the LoP request procedure described in Section 3.6.3 until it receives an RA Ack from the SSN. This results in the overall protection switching time to be increased. To avoid this, it is RECOMMENDED to configure none of the working paths sharing the protection segment in a SMP domain to have the same Protection Switching Priority.

4. Protocol

4.1. PDU Format

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

The shared mesh protection protocol messages are as follows:

- Protection Switching (PS) Event message [sent from protection path endpoint to SEN]
- Lockout of Protection (LoP) Request message [sent from SEN to protection path endpoint]
- Lockout of Protection (LoP) Acknowledgement message [sent from protection path endpoint to SEN]
- Resource Allocation (RA) Notification message [sent from SEN to SSN]
- Resource Allocation (RA) Acknowledgement message [sent from SSN to SEN]
The channel type in ACH is used to indicate that the message is a SMP protocol message. The protocol message MUST follow the ACH.

```
0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|0 0 0 1|Version|  Reserved    | Channel Type = Shared Mesh P. |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|                                  Shared Mesh Protection Protocol Message |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
```

Figure 4: Shared mesh protection protocol message header

The SMP protocol message format is defined as follows:

```
0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|Version|  Mode |  Type | ST|     Reserved      |       ID      |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
```

Figure 5: Shared mesh protection protocol message format

Each protocol message includes the following fields:

- **Version** - Version indicator
  - 0x0 - reserved
  - 0x1 - This version
  - 0x2~0xF - reserved

- **Mode** - SMP switching and operation mode
  - 0x0 - Bidirectional
  - 0x1 - Unidirectional
  - 0x2~0xF - reserved

- **Type** - SMP protocol message type indicator
  - 0x0 - reserved
  - 0x1 - PS Event
  - 0x2 - LoP Request
  - 0x3 - LoP Acknowledgement
  - 0x4 - RA Notification
  - 0x5 - RA Acknowledgement
  - 0x6~0xF - reserved
o ST - SMP protocol message sub-type indicator
  0x0     - Switch to working (for PS Event),
           Clear LoP (for LoP Request)
  0x1     - Switch to protection (for PS Event),
           Assert LoP (for LoP Request)
  0x2˜0x3 - reserved

o ID - This is either the identifier of the LP domain that is
       sending the message or the working path that was allocated
       to the resources (dependent upon the message).
  0x0      - reserved
  0x1˜0xFE - ID value
  0xFF     - reserved

o Reserved - reserved fields

All reserved bits SHOULD be zero upon transmission, and MUST be
ignored on reception.

4.2.  Message Transmission

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

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

5.  Operation of Shared Mesh Protection

This section illustrates the operation of the SMP protocol based on
the example illustrated in Figure 3 and the following assumptions:

o The SMP domain consists of the following end-to-end LP domains
  (LPDs):
  * LPD1: Working path ABC (W1) / Protection path APQC (P1)
  * LPD2: Working path GHJ (W2) / Protection path GPQRSJ (P2)
  * LPD3: Working path DEF (W3) / Protection path DRSF (P3)
The SMP domain includes the following SMPG:

* S1: LPD1 & LPD2 (Shared protection segment - PQ)
* S2: LPD3 & LPD2 (Shared protection segment - RS)

Protection Switching Priority is LPD1 > LPD2 > LPD3 (i.e. LPD1 has the highest priority.)

All working paths are protected by 1:1 bidirectional protection switching.

If a unidirectional failure occurs on W2 in the direction from node H to node G as shown in Figure 6, SMP will perform the following:

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

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

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

d. SEN R compares the Protection Switching Priority of LPD2 with those of other members of S2, i.e. LPD3. In this example, as the priority of LPD3 is lower than LPD2, SEN R sends the RA Notification message to the SSN S, blocks forwarding of P3 and sends the LoP Request message requesting the assertion of LoP to node D.

e. SSN S does not process the RA Notification message. (Since in this example, all the LP domains are configured to have different Protection Switching Priorities.)

f. Node D takes the LoP Request message as input to the LP switching, and follows the LP procedure to process the end-to-end LoP command. After completion of the LoP operation, node D sends the LoP Ack message to SEN R.

g. SEN R unblocks forwarding of P3 upon receiving the LoP Ack message from node D.

h. Since LPD2 operates in 1:1 bidirectional protection switching mode, node J performs the switching operations (i.e. switches its
bridge and selector state) to synchronize with node G, and also transmits the PS Event message to node S and Q, which are SENs for G->H->J. Using a parallel procedure to that described in steps c & d, SEN S sends the LoP Request message to node F while the SEN Q does not take an action to node C.

\begin{verbatim}
==A======B======C== ==D======E======F==
 \ /          \ /          \ /          \ /
 \ LPD1 /     \ LPD3 /    \ /          \ /
 \ /          \ /          \ /          \ /      == : Normal traffic
 P======Q================R======S
 //                \//
 //                \//
 //                \//
 ==G------xx---------H------------------J==
 SF on G<-H       W2
\end{verbatim}

Figure 6: Shared Mesh Protection Example 1

Figure 7 shows a progression from Figure 6. While LPD2 is in protecting state with its traffic transported on protection path P2, another unidirectional failure occurs on W1 in the direction from node B to node A.

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

a. Node A detects the failure, and initiates the linear protection switching for the failed W1.

b. At the same time, node A transmits the PS Event message notifying SEN for S1, i.e. node P, that a protection switching event occurred.

c. SEN P compares the Protection Switching Priority of LPD1 with those of the other members in S1, in this case LPD2. In this example, since the priority of LPD2 is lower than LPD1, SEN P sends the RA Notification message requesting the assertion of LoP to node G.

d. SSN Q does not process the RA Notification message. (Since in this example, all the LPDs are configured to have different Protection Switching Priorities.)

e. Node G accepts the LoP Request message as input to linear protection switching, and follows LP procedure to process the LoP command. When LPD2 is forced to lockout its protection path P2,
it may try to find another available path. m:n protection or other recovery mechanism may be used for this, but this discussion is out of scope for this document. After completion of the LoP operation, node G sends the LoP Ack message to SEN P.

f. SEN P unblocks forwarding of P2 upon receiving the LoP Ack message from node G.

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

h. SEN P compares the Protection Switching Priority of LPD2 with the other members of S1, i.e. LPD1, and does not transmit any message to node A, but SEN R sends the LoP Request message to request the clear of LoP to node D, after comparing the Protection Switching Priorities of the members of S2.

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

![Diagram of shared mesh protection example 2]

Figure 7: Shared Mesh Protection Example 2

NOTE: Examples for race condition to be provided in the next version.

6. Manageability Considerations

To be added in future version.
7. IANA Considerations
   To be added in future version.

8. Security Considerations
   To be added in future version.

9. References

9.1. Normative References


9.2. Informative References


[G.808.1] SG15, "Generic Protection Switching - Linear trail and
Authors’ Addresses

Tae-sik Cheung  
ETRI  
161 Gajeong  
Yuseong, Daejeon  305-700  
South Korea  
Phone: +82 42 860 5646  
Email: cts@etri.re.kr

Jeong-dong Ryoo  
ETRI  
161 Gajeong  
Yuseong, Daejeon  305-700  
South Korea  
Phone: +82 42 860 5384  
Email: ryoo@etri.re.kr

Yaacov Weingarten  
Nokia Siemens Networks  
3 Hanagar St. Neve Ne’eman B  
Hod Hasharon,  45241  
Israel  
Phone: +972-54-220 0977  
Email: wyaacov@gmail.com

Nurit Sprecher  
Nokia Siemens Networks  
3 Hanagar St. Neve Ne’eman B  
Hod Hasharon,  45241  
Israel  
Email: nurit.sprecher@nsn.com
Daniel King
Old Dog Consulting
United Kingdom

Email: daniel@olddog.co.uk
Abstract

This document defines how to use ICC-based Multiprotocol Label Switching Transport Profil (MPLS-TP) identifiers for proactive connectivity verification (CV) analogous to RFC 6428. New TLVs are defined to support proactive CV based on identifiers following ITU-T conventions.

Status of this Memo

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

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

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

This Internet-Draft will expire on September 6, 2012.

Copyright Notice

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

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

1. Introduction .................................................. 3
2. Requirements notation ........................................ 3
3. ICC_Operator_ID-based Source MEP-ID TLV Definitions ........ 3
4. Security Considerations ....................................... 4
5. IANA Considerations ......................................... 4
6. Normative References ......................................... 4
Authors’ Addresses ............................................. 5
1. Introduction

Proactive Connectivity Verification, Continuity Check, and Remote Defect Indication (CV, CC and RDI) [RFC6428] is a proactive Operations, Administration, and Maintenance (OAM) mechanism for the MPLS Transport Profile (MPLS-TP). [RFC6428], defines CV message which carry a Global_ID-based Source MEP-ID TLV to permit mis-connectivity detection to be performed by sink MEPs. In transport networks however, the ITU Carrier Code (ICC) is traditionally used to identify a carrier/service provider. Instead of using the Global_ID, which is derived from the AS number of the service provider, this document defines a set of Source MEP-ID TLVs based on the ICC_Operator_ID for sections, LSPs and PWs as specified in [I-D.ietf-mpls-tp-itu-t-identifiers] for use in CV.

2. Requirements notation

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

3. ICC_Operator_ID-based Source MEP-ID TLV Definitions

An MPLS-TP CV message [RFC6428] consists of a BFD control packet prepended by the Associated Channel Header (ACH) and appended by a Source MEP-ID TLV. This section defines an ICC_Operator_ID-based Source MEP-ID TLV object. It is used as a replacement for the Global_ID-based Source MEP-ID TLV as specified in [RFC6428].

The format of ICC_Operator_ID-based Source MEP-ID TLV is shown below.
Figure 1: ICC_Operator_ID-based Source MEP-ID TLV Format

An ICC_Operator_ID-based Source MEP-ID TLV is encoded as a 2-octet field that specifies a type, followed by a 2-octet length field, followed by a value field.

The value field is encoded by appending a 16-bit MEP index to the 15 character long MEG_ID as defined in [I-D.ietf-mpls-tp-itu-t-identifiers]. Padding is used to align the field with a 4 octet boundary. The length of the padding MUST be included in the length field. The type will be 'X', with no distinction between sections, LSPs and Pseudowires [type value to be assigned by IANA].

4. Security Considerations

The security concerns expressed in the security considerations section of [RFC6428] apply. No additional concerns stem from the use of the TLV defined in this document.

5. IANA Considerations

TBD

6. Normative References

[I-D.ietf-mpls-tp-itu-t-identifiers]
Gray, E., Helvoort, H., Betts, M., and R. Winter, "MPLS-TP Identifiers Following ITU-T Conventions",

Cui & Winter Expires September 6, 2012 [Page 4]
draft-ietf-mpls-tp-itu-t-identifiers-02 (work in progress), October 2011.


Authors’ Addresses

Zhenlong Cui
NEC

Email: c-sai@bx.jp.nec.com

Rolf Winter
NEC

Email: Rolf.Winter@neclab.eu
Abstract

This document defines how to use ICC-based MPLS-TP identifiers for on-demand connectivity verification (CV) analogous to RFC 6426. New TLVs are defined to support on-demand CV based on identifiers following ITU-T conventions.

Status of this Memo

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

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

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

This Internet-Draft will expire on August 27, 2012.

Copyright Notice

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

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

1. Introduction .................................................. 3
2. Requirements notation .......................................... 3
3. ICC_Operator_ID-based TLV Definitions ....................... 3
   3.1. ICC_Operator_ID-based Source/Destination Identifier
       TLVs ......................................................... 4
   3.2. ICC_Operator_ID-based Static LSP/PW Sub-TLV ............. 5
      3.2.1. ICC_Operator_ID-based Static LSP Sub-TLV ............ 5
      3.2.2. ICC_Operator_ID-based Static PW Sub-TLV ............ 6
4. Security Considerations ....................................... 6
5. IANA Considerations ........................................... 6
6. Normative References .......................................... 7
Authors’ Addresses ............................................... 7
1. Introduction

MPLS On-Demand Connectivity Verification (CV) and Route Tracing [RFC6426] is an on-demand monitoring mechanism for the MPLS Transport Profile (MPLS-TP). [RFC6426] defines a set of Global_ID-based TLVs to support on-demand CV and route tracing for MPLS-TP LSPs, including PWs and Sections which follow the IP/MPLS conventions.

In transport networks however, the ITU Carrier Code (ICC) is traditionally used to identify a carrier/service provider. Instead of using the Global_ID, which is derived from the AS number of the service provider, this document defines source/destination TLVs and static LSP/PW Sub-TLVs based on the ICC_Operator_ID as specified in [I-D.ietf-mpls-tp-itu-t-identifiers] for use in CV.

2. Requirements notation

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

3. ICC_Operator_ID-based TLV Definitions

In ICC-based transport network, the Global_ID might not be available for on-demand CV and route tracing. In such environments it might be necessary to perform CV and route tracing using the ICC_Operator_ID as specified in [I-D.ietf-mpls-tp-itu-t-identifiers].

The ICC_Operator_ID consists of the Country Code (CC) followed by the ITU carrier code (ICC). The Country Code (alpha-2) is a string of two alphabetic characters, and the ICC itself is a string of one to six left-justified characters, each character being either alphabetic (i.e. A-Z) or numeric (i.e. 0-9).

This section provides the definition for a number of ICC_Operator_ID-based TLV objects. In order to simplify implementations, the length of ICC_Operator_ID field has a fixed length independent of the ICC length. Therefore, zero padding will be used in cases where the ICC length is less than 6 octets long. The total length of the ICC_Operator_ID therefore amounts to 8 octets as shown in Figure 1.
3.1. ICC_Operator_ID-based Source/Destination Identifier TLVs

The Source and Destination Identifier TLVs follow the same format except their only difference being the type. The format is shown below.

```
0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|             Type              |          Length = 16          |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|                                                               |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
| ICC_Operator_ID (8 Octets) +|
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|                     Node_ID   (4 Octets)                      |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
```

Figure 2: ICC_Operator_ID-based Source/Destination Identifier TLV Format

The format of the ICC_Operator_ID is defined in [I-D.ietf-mpls-tp-itu-t-identifiers]. The encoding of the ID is depicted in Figure 1.

The format of the Node_ID is defined in [RFC6370].

Type will be one of either TBD-SRC or TBD-DST. The TLV structure is therefore as follows:

```
<table>
<thead>
<tr>
<th>Type #</th>
<th>Length</th>
<th>Value Field</th>
</tr>
</thead>
<tbody>
<tr>
<td>TBD-SRC</td>
<td>16</td>
<td>ICC_Operator_ID-based Source Identifier TLV</td>
</tr>
<tr>
<td>TBD-DST</td>
<td>16</td>
<td>ICC_Operator_ID-based Destination Identifier TLV</td>
</tr>
</tbody>
</table>
```

Figure 3: ICC_Operator_ID-based Source/Destination Identifier types
3.2. ICC_Operator_ID-based Static LSP/PW Sub-TLV

The new sub-TLVs are assigned sub-type identifiers as follows, and are described in the following sections.

<table>
<thead>
<tr>
<th>Type #</th>
<th>Sub-Type #</th>
<th>Length</th>
<th>Value Field</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>24</td>
<td>28</td>
<td>ICC_Operator_ID-based Static LSP</td>
</tr>
<tr>
<td>1</td>
<td>25</td>
<td>36</td>
<td>ICC_Operator_ID-based Static Pseudowire</td>
</tr>
</tbody>
</table>

Figure 4: ICC_Operator_ID-based Static LSP/PW Sub-types

3.2.1. ICC_Operator_ID-based Static LSP Sub-TLV

The format of the ICC_Operator_ID-based Static LSP Sub-TLV is specified in the following figure. The value fields are taken from [I-D.ietf-mpls-tp-itu-t-identifiers].

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

Figure 5: ICC_Operator_ID-based Static LSP Sub-TLV Format

The ICC_Operator_ID MAY be set to zero. Note however that such use is limited to entities contained within a single operator and MUST NOT be used across an NNI. However, the other fields without the padding field MUST be set to non-zero values.
3.2.2. ICC_Operator_ID-based Static PW Sub-TLV

The format of the ICC_Operator_ID-based Static PW Sub-TLV is specified in the following figure. The value fields are taken from [I-D.ietf-mpls-tp-itu-t-identifiers].

```
+--------------------------+--------------------------+
|                          |                          |
|                          |                          |
|                          |                          |
|                          |                          |
|                          |                          |
|                          |                          |
| Service Identifier       | Source ICC_Operator_ID   |
|                          | Source Node ID           |
|                          | Source AC-ID             |
|                          | Destination ICC_Operator_ID |
|                          | Destination Node ID      |
|                          | Destination AC-ID        |
+--------------------------+--------------------------+
```

Figure 6: ICC_Operator_ID-based Static PW Sub-TLV Format

The ICC_Operator_ID MAY be set to zero. Note that such use is limited to entities contained within a single operator and MUST NOT be used across an NNI. However, the other fields MUST be set to non-zero values.

4. Security Considerations

TBD

5. IANA Considerations

TBD
6. Normative References

[I-D.ietf-mpls-tp-itu-t-identifiers]
Winter, R., Gray, E., Helvoort, H., and M. Betts, "MPLS-TP Identifiers Following ITU-T Conventions",
draft-ietf-mpls-tp-itu-t-identifiers-02 (work in progress), October 2011.


Authors’ Addresses

Zhenlong Cui
NEC

Email: c-sai@bx.jp.nec.com

Rolf Winter
NEC

Email: Rolf.Winter@neclab.eu
Abstract

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

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

Status of this Memo

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

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

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

This Internet-Draft will expire on September 13, 2012.
# Table of Contents

1. Introduction ................................................. 4  
   1.1. 1:n Protection architecture ........................... 4  
   1.2. Locking operation ....................................... 6  
   1.3. Non-Locking ............................................ 7  
   1.4. Path priority .......................................... 7  
   1.5. Preemption ............................................. 8  
   1.6. Contributing authors ................................... 8  
2. Conventions used in this document ............................ 8  
   2.1. Acronyms ................................................ 9  
   2.2. Definitions and Terminology ............................. 9  
3. Use cases and scenarios ..................................... 9  
   3.1. Non-locking use case: Per-node label space ............... 9  
   3.2. Locking use-case: ...................................... 10  
   3.3. PSC Scenarios .......................................... 12  
      3.3.1. Unidirectional failure cases ......................... 13  
      3.3.2. Bidirectional fault scenarios ....................... 15  
      3.3.3. Preemption scenarios ............................... 17  
4. Changes to PSC .............................................. 22  
   4.1. PSC ................................................... 23  
   4.2. Changes to PSC Payload .................................. 23  
      4.2.1. Locking (L) flag ................................... 24  
      4.2.2. Fault path (FPath) field ........................... 24  
      4.2.3. Data path (Path) field ............................. 24  
   4.3. Changes to PSC Operation ................................. 25  
      4.3.1. Basic operation .................................... 25  
      4.3.2. Two-phased operation ............................... 25  
      4.3.3. Acknowledge message ................................. 26  
      4.3.4. Wait for Acknowledge (WFA) timer ................... 27  
      4.3.5. Additional PSC State ............................... 27  
5. IANA Considerations ........................................ 31  
6. Security Considerations .................................... 31  
7. Acknowledgements .......................................... 31  
8. References ................................................ 31  
   8.1. Normative References .................................. 31  
   8.2. Informative References ................................ 32  
Appendix A. PSC state machine tables ............................ 32  
Authors’ Addresses ........................................... 36
1. Introduction

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

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

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

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

1.1. 1:n Protection architecture

Linear protection switching is a fully allocated survivability mechanism. It is fully allocated in the sense that the route and bandwidth of the protection path is reserved for a set of working paths. For 1:n protection the protection path is allocated to protect any one of n working paths between the two endpoints of the protection domain.
Figure 1 shows a protection domain with N working transport paths and a single protection path. In 1:n protection, the protection path may transport the traffic of only a single working path at any particular time. The identity of the working path that is being protected must be communicated between the two endpoints.

Unless otherwise specified, all examples will be based on the network topology in Figure 1, with the working paths referenced as Wi (for 1<=i<=N) and the protection path referenced as P. The end-points of the protection domain will be referred to as LER-A and LER-Z.

The different working paths may be disjoint at the intermediary points on the path between LER-A and LER-Z and may also have different resource requirements. In addition, each of the working paths may be assigned a priority that could be used to decide which working path would be protected in cases of conflict (see more on this topic in Section 1.5). It is usually advised to arrange these protection groups in a way that would minimize any potential conflict situation.

1:n protection in MPLS supports two modes of operation - locking and non-locking. Locking mirrors the behavior that is used by many transport protection mechanisms, and is necessary in some cases but
may incur increased latency (and thus packet loss), as a result of prolonged switching time, in comparison to the non-locking case. Non-locking 1:n can be used in many MPLS networks and has far less packet loss as compared to locking, but must be used with care - since incorrect use of non-locking can lead to misconnectivity.

1.2. Locking operation

The high-level functionality of the locking operation mode of 1:n protection would follow the following basic steps:

- LER-A detects a unidirectional failure of W1 and stops sending traffic on W1.
- LER-A transmits a PSC SF message to LER-Z indicating that W1 has failed and its traffic should be redirected to P. No traffic is sent on P at this point.
- LER-Z receives the PSC message from LER-A and begins transmitting W1 traffic in P, and sends a PSC message to LER-A indicating that W1 is now being protected by P. LER-A receives the normal data traffic intended for W1 from P, LER-Z receives the W1 data traffic from P and also bridges W1 data traffic into P.
- LER-A receives the PSC message from LER-Z and begins transporting W1 traffic in P -- that is, LER-A bridges W1 into P.

It should be clear from this description that no traffic is sent over P until LER-Z processes the PSC message from LER-A, and that traffic is only sent unidirectionally (Z->A) until LER-A processes the "reply" PSC message from LER-Z. As the message processing time is expected to be dwarfed by the propagation delay between LER-A to LER-Z, it can be said that there is complete traffic loss between the endpoints for the duration of the one-way propagation delay from LER-A to LER-Z, and full bidirectional traffic flow is not fully restored until after 1xRTT of the protection path.

This operation mode is referred to as "locking" because the sequence of processing the PSC messages includes periods where the protection path is locked from carrying protected traffic, while the two endpoints verify that both are ready to process the W1 traffic that is received on P. More detailed information on this mode of operation will be supplied later in the document when considering different scenarios.
1.3. Non-Locking

In non-locking protection operation mode, LER-A switches data traffic onto P immediately upon failure detection. This minimizes traffic loss, but at the cost of temporary asymmetry of packet flow. At a high level, it looks like this:

- LER-A detects the failure of W1 and stops sending traffic on W1.
- LER-A immediately begins to transport W1’s data traffic over the protection path P.
- Simultaneously LER-A transmits a PSC message to LER-Z indicating that W1 has failed and is currently being protected in P.
- LER-Z receives the PSC message from LER-A, switches all W1 data traffic to P, and transmits a PSC message to LER-A indicating that W1 is now protected in P.
- LER-A receives the PSC message from LER-Z and needs to take no action, as the protection switch had already been completed.

In the non-locking case, the packet loss between the endpoints is minimized. Packet loss in the A->Z direction is only the failure detection time, which is assumed, for this document, to be negligible. Packet loss in the Z->A direction is almost entirely the result of the one-way propagation delay of the PSC message from LER-A to LER-Z. Assuming the transport path from A->Z has the same delay as that from Z->A, it can be said that the packet loss in the non-locking case is roughly half that of the locking case.

1.4. Path priority

As the 1:n architecture requires the ability for one working path to preempt the traffic of another in the event of multiple failures (see Section 1.5), there must be an indication of priority between the different working paths so that an implementation can decide whether a new failure should be allowed to preempt a protection switch already in place. This priority is purely a local decision, i.e., determined by configuration at both endpoints of the protection domain. It is also possible to assign the same priority to multiple working paths, thus creating a "first come first served" preemption policy. This document provides no means to signal the priority of a given working path, nor a means to detect priority mismatches or misconfigurations. Thus, ensuring that the priorities of all working LSPs in a protection domain is a matter for the operator. Any mismatch or misconfiguration will likely result in unexpected protection behavior.
1.5. Preemption

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

There are two basic scenarios for preemption of traffic -

1. When the protection path is used to transport "extra traffic". While this practice is discouraged by [TPReq], it is still not precluded. When the protection domain triggers a protection switch, the extra traffic should be preempted to allow the transport of the protected traffic from the working path that triggered the switching operation. The subsequent treatment of the interrupted service is out of the scope of this document.

2. When the protection path is transporting traffic from a working path and a second working path triggers a switching condition. This second trigger may either be a trigger with a higher priority (e.g. FS after a SF) or because the operator had assigned a higher priority to the working path of the second trigger. At this point, the traffic for the lower priority working path will be interrupted, and the higher priority traffic will be transmitted on the protection path. The preempted traffic will only renew transmission, when either the working path recovers, or the higher priority traffic relinquishes control of the protection path.

1.6. Contributing authors

Nurit Sprecher (NSN)

2. Conventions used in this document

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

This draft uses the following acronyms:

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

2.2. Definitions and Terminology

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

3. Use cases and scenarios

This section will present some use-cases and scenarios that should illucidate the use of PSC for 1:n protection.

3.1. Non-locking use case: Per-node label space

Non-locking protection can be used when the payload that is received from the protection path is unambiguous and can be properly forwarded without the need to explicitly establish selector and bridge configuration at the time of failure. One example where this applies is when the endpoints of the protection domain are using per-platform label space [RFC3031].

In per-node or per-platform label space, the LIB is established on a node such that it can properly switch any labeled packet regardless of input interface.

Consider, as an example, the protection topology as shown in Figure 1 with four working paths - W1, W2, W3, W4 and a single protection
path, P, that connect between LER-A and LER-Z. Each packet that
transported from LER-A to LER-Z is labelled by LER-A depending upon
the path that it is being transmitted over. From there the packet
will traverse the relevant path and have its label manipulated by the
intermediate LSRs until it arrives at LER-Z, at which point, the LER
will pop the label for the path used within the protection domain and
process the next label down to determine how to forward the packet
payload. The following table gives the label assigned by LER-A and
the one expected by LER-Z for each of the transport paths:

<table>
<thead>
<tr>
<th>Path</th>
<th>Label at LER-A</th>
<th>Label for LER-Z</th>
</tr>
</thead>
<tbody>
<tr>
<td>W1</td>
<td>100</td>
<td>105</td>
</tr>
<tr>
<td>W2</td>
<td>200</td>
<td>205</td>
</tr>
<tr>
<td>W3</td>
<td>300</td>
<td>305</td>
</tr>
<tr>
<td>W4</td>
<td>400</td>
<td>405</td>
</tr>
<tr>
<td>P</td>
<td>500</td>
<td>505</td>
</tr>
</tbody>
</table>

If there is a pseudowire (PW) that needs to be carried over one of
these transport paths between LER-A and LER-Z, whose label is
allocated from the per-platform label space on both LER-A and LER-Z
(e.g. label 888), then when a packet for this PW is transported over
W2, the label stack that will be sent from LER-A will be [200|888|..]
and it will arrive at LER-Z with a label stack [205|888|..]. If W2
were to report a failure that triggers a protection switch and LER-A
would redirect a packet for this PW to P, it would be transported
with a label stack of [500|888|..] and be received by LER-Z with a
label stack [505|888|..]. Since the PW label is drawn from per-node
label space, when LER-Z pops the path label it will be able to
process the PW label regardless of the transport path that was used
between LER-A & LER-Z.

Since the forwarding behavior is preestablished, there is no need to
ensure that LER-A and LER-Z coordinate the bridge/selector functions
as part of the protection protocol. This is true for any underlying
label assigned from per-node space. The label can be allocated by
LDP, MPLS VPNs, PWs, TE tunnels, or any other application. As long
as the label is preprogrammed in the receiving node’s label space,
coordination of the bridge/selection functions is unnecessary.

3.2. Locking use-case:

Locking protection must be used when the payload that is received on
the protection path is ambiguous; that is, the switching behavior for
the payload of the protection path must be established at the time of
failure. One such example where this applies is when the endpoints
of the protection domain are using per-interface label space, where the Working and Protect LSPs are instantiated as interfaces.

In per-interface label space, a node may use the same label value to represent different switching behaviors on different interfaces. For example, the label value 100 when received on LSP W1 may be treated differently than the label value 100 when received on LSP W2. Since either W1 or W2 may be protected in P, LSP P must ensure that it has the proper forwarding behavior defined for label 100. Using the wrong forwarding behavior (e.g., programming P's label space with W1's entry for label 100 when P is protecting W2) is likely to lead to misconnectivity.

Consider, as an example, the protection topology as shown in Figure 1 and in Section 3.1. There are four working paths - W1, W2, W3, W4 - and a single protection path, P, that connect between LER-A and LER-Z. Section 3.1 shows a table with the receive labels [105, 205, 305, 405, 505] at LER-Z, and those do not change. What changes is the payload of those labels. Section 3.1 gives the example of a PW drawn from global label space which uses the label 888 - this label is treated to the same forwarding behavior no matter which LSP is used to carry it from LER-A to LER-Z.

In per-interface label space, each W-LSP has its own label space. For this example, consider a PW switched over W1 with the outgoing label 900. Thus, the label stack when leaving LER-A is [100|900] and when arriving at LER-Z is [105|900]. There is also a PW defined over W2 which also uses label 900, but with a different forwarding behavior. The per-interface label switching tables on LER-Z look like this:

<table>
<thead>
<tr>
<th>Input Interface</th>
<th>Label</th>
<th>Switching behavior</th>
</tr>
</thead>
<tbody>
<tr>
<td>W1</td>
<td>900</td>
<td>Switch to Access Circuit #1</td>
</tr>
<tr>
<td>W2</td>
<td>900</td>
<td>Switch to Access Circuit #2</td>
</tr>
<tr>
<td>W3</td>
<td>900</td>
<td>Switch to Access Circuit #3</td>
</tr>
<tr>
<td>W4</td>
<td>900</td>
<td>Switch to Access Circuit #4</td>
</tr>
<tr>
<td>P</td>
<td>900</td>
<td>none defined (drop, log error)</td>
</tr>
</tbody>
</table>

The label space for P is established at the time of failure, using PSC. When there is no failure, there is no switching behavior defined for the P LSP's contents.

When the protection domain has determined that W2 has failed and needs to be switched, it coordinates this protection, using PSC, between LER-A and LER-Z. Part of the coordination is to establish
the proper receive behavior on LER-Z, i.e. the Switching behavior on
the input interface for Label 900 to be "Switch to Access Circuit
#2". Whereas, if W1 fails and preempts W2, the switching behavior on
LER-Z is changed be "Switch to Access Circuit #1".

Clearly it is imperative that there be no misconnectivity. This
requirement means that there must be a "lock" on P established, such
that there are no packets transmitted on an LSP until both ends agree
on the switching behavior for that LSP. The details of the behavior
in the locking use cases is explored further in Section 3.3. of this
document.

3.3. PSC Scenarios

This section discusses the message exchange necessary to perform both
non-locking and locking PSC options for 1:n protection. There are
several examples presented here that attempt to cover all the
combinations of failure and preemption, unidirectional and
bidirectional protection for the two modes of operation. It should
be noted that this is a non-exhaustive set of scenarios, but were
chosen to highlight the main features of the proposal.

It is not the intent of this document to spell out all the
combinations of preemption, directionality and locking behavior which
can occur. That is not how one builds a robust protocol. This
document spells out a state machine which reacts appropriately in all
possible cases, and as part of that walks through some of the failure
cases as examples. PSC is, at its heart, a simple protocol. A node
is aware of both its local status and the status of the remote node,
and transitions to the appropriate state and takes appropriate action
based on the combination of these two states. Preemption, which as
noted is only relevant in 1:n, does not increase the complexity of
the protocol. The examples are detailed, but the behavior is quite
simple.

All of these examples assume a protection domain consisting of four
working paths [W1, W2, W3, W4] with priority in decreasing order,
i.e. W1 > W2 etc. There is a single protection path, P. These
examples use the notation "B = x" to indicate the protect LSP whose
contents are bridged into the protect LSP. For example, if W3 has
failed and is currently protected, B = 3. If no protection is in
place, B = n/a. All examples end with the REQ(FPath, Path) and B
values for each node in each example.

The non-locking cases assume that both LER-A and LER-Z have
preestablished per-node label spaces, as per the use case above.

All cases assume that the time required to perform on-box operations
such as bridging or selecting is instantaneous. The one-way delay between nodes is abbreviated OWD, and the round trip time is RTT (i.e. RTT = 2 x OWD).

3.3.1. Unidirectional failure cases

The examples in this section provide the message flow between LER-A and LER-Z for the scenario where a unidirectional fault is detected by LER-A on working path W1. The message flow is described as a sequence along a timeline.

3.3.1.1. Non-locking

Considering the scenario of a protection domain operating in non-locking mode the following is the event timeline:

<table>
<thead>
<tr>
<th>Time</th>
<th>Event Description</th>
<th>LER-A PSC Bridge</th>
<th>LER-Z PSC Bridge</th>
</tr>
</thead>
<tbody>
<tr>
<td>t0</td>
<td>Traffic is being transported on W1, P is not carrying any traffic. Both LER-A and LER-Z transmitting PSC NR(0,0) message.</td>
<td>NR(0,0)</td>
<td>NR(0,0)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>B = n/a</td>
<td>B = n/a</td>
</tr>
<tr>
<td>t1</td>
<td>LER-A detects SF on W1, bridges W1 into P and sends SF(1,1). LER-A enters into WFA (Waiting for Acknowledgement) state. LER-A still selects the traffic from W1. This is admittedly of not much use when LER-A sees SF, may be useful when LER-A encounters a partial failure such as SD.</td>
<td>SF(1,1)</td>
<td>NR(0,0)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>B = 1</td>
<td>B = n/a</td>
</tr>
<tr>
<td>t2</td>
<td>LER-Z receives SF(1,1). LER-Z enters PF:W:R state. LER-Z switches W1 onto P and sends SF(1,1). At this point traffic for W1 is protected in both directions</td>
<td>SF(1,1)</td>
<td>NR(0,1)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>B = 1</td>
<td>B = 1</td>
</tr>
<tr>
<td>t3</td>
<td>LER-A receives SF(1,1), which it takes as an ACK from LER-Z. LER-A transits from WFA to PF:W:L state. Switch is complete.</td>
<td>SF(1,1)</td>
<td>NR(0,1)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>B = 1</td>
<td>B = 1</td>
</tr>
</tbody>
</table>

Figure 2: Unidirectional non-locking

Note: Between t1 and t2, LER-A transports the data traffic on P while LER-Z continues transporting it on W1, and there is temporary path asymmetry. After t2, the data traffic is in P in both directions.
In this case, LER-A loses traffic for the OWD time, as it does not receive any traffic from LER-Z on P until LER-Z bridges W1 into P. LER-Z does not lose any traffic due to the immediate bridging on LER-A.

3.3.1.2. Locking

When examining the similar scenario for a protection domain that is using the Locking mode of operation, we have the following time sequence:

<table>
<thead>
<tr>
<th>Time</th>
<th>Event Description</th>
<th>LER-A PSC Bridge</th>
<th>LER-Z PSC Bridge</th>
</tr>
</thead>
<tbody>
<tr>
<td>t0</td>
<td>Traffic is being transported on W1, P is not carrying any traffic. Both LER-A and LER-Z transmitting PSC NR(0,0) message.</td>
<td>NR(0,0) B = n/a</td>
<td>NR(0,0) B = n/a</td>
</tr>
<tr>
<td>t1</td>
<td>LER-A detects SF on W1, LER-A enters into WFA state and sends SF(1,1). LER-A still transports and selects the traffic from W1. This allows traffic to get through if the failure is truly unidirectional.</td>
<td>SF(1,0) B = n/a</td>
<td>NR(0,0) B = n/a</td>
</tr>
<tr>
<td>t2</td>
<td>LER-Z receives SF(1,0). LER-Z enters PF:W:R state. LER-Z bridges W1 into P and sends NR(0,1) but continues to select traffic from W1</td>
<td>SF(1,0) B = 1</td>
<td>NR(0,1) B = 1</td>
</tr>
<tr>
<td>t3</td>
<td>LER-A receives NR(0,1), which it takes as an ACK from LER-Z. LER-A completely switches W1 traffic onto P. LER-A transits from WFA to PF:W:1 state. Switch complete</td>
<td>SF(1,1) B = 1</td>
<td>NR(0,1) B = 1</td>
</tr>
<tr>
<td>t4</td>
<td>LER-Z receives SF(1,1). LER-Z selects W1 traffic from P and sends NR(0,1)</td>
<td>SF(1,1) B = 1</td>
<td>NR(0,1) B = 1</td>
</tr>
</tbody>
</table>

Figure 3: Unidirectional locking

Note: At t1, LER-A stops sending traffic to LER-Z. At t3, it resumes. Since the majority of the time delay at both t1 and t2 is the one-way transmission delay between LER-A and LER-Z, there is a total of 1xRTT traffic loss at both endpoints.
3.3.2. Bidirectional fault scenarios

The examples above focused on unidirectional failures in order to illustrate the basic principles of 1:n protection. However, most failures in carrier networks are bidirectional in nature. Bidirectionality includes not only the failure of both the tx and rx physical path (e.g. a fiber cut) but also a unidirectional failure made bidirectional by mechanisms outside of PSC such as CC-V or LDI.

Both ends of a protection domain may not see the bidirectional failure at the same instant. In the case of a true bidirectional fiber cut, the cut may be physically closer to one end of the domain than the other, and thus the end which is farther away takes longer to notice the failure. This is referred to as "asymmetric notification delay" in this document. Similarly, a unidirectional failure seen by one endpoint which triggers an LDI notification to the far endpoint will not be recognized by this far end until after its has been noticed it at the near endpoint.

There are a number of scenarios that constitute bidirectional failure, and the variety of triggers and notification delays mean that it is impossible to document them all here. The scenario used in this case is of a true bidirectional failure, on working path W1, with asymmetric notification delay, as described above. Both the case of Non-locking and Locking operation modes are presented.

It is perhaps important to understand that a node, when reacting to a failure, simply reacts either to its local LSP status (e.g. SF on the underlying fiber) or the status of the remote node (e.g. the remote node sending SF(x,y)). A node neither knows nor cares whether the failure is bidirectional; it simply reacts to inputs to its local state machine. It can easily be observed that there are no special states needed for unidirectional vs. bidirectional error handling.

3.3.2.1. Non-Locking

First we present the scenario when operating in non-locking mode:
<table>
<thead>
<tr>
<th>Time</th>
<th>Event Description</th>
<th>LER-A PSC Bridge</th>
<th>LER-Z PSC Bridge</th>
</tr>
</thead>
<tbody>
<tr>
<td>t0</td>
<td>Traffic is being transported on W1, P is not carrying any traffic. Both LER-A and LER-Z transmitting PSC NR(0,0) message.</td>
<td>NR(0,0)</td>
<td>NR(0,0)</td>
</tr>
<tr>
<td></td>
<td>Both LER-A and LER-Z transmitting PSC NR(0,0) message.</td>
<td>B = n/a</td>
<td>B = n/a</td>
</tr>
<tr>
<td>t1</td>
<td>LER-A detects SF on W1, bridges W1 into P and sends SF(1,1). LER-A enters into WFA state and continues to select the traffic from W1.</td>
<td>SF(1,1)</td>
<td>SF(1,1)</td>
</tr>
<tr>
<td></td>
<td>SF(1,1)</td>
<td>B = 1</td>
<td>B = n/a</td>
</tr>
<tr>
<td>t2</td>
<td>LER-Z detects the SF on W1. LER-Z enters WFA state and bridges W1 into P and transmitting SF(1,1). At this point traffic for W1 is protected in both directions, however the endpoints are still not coordinated.</td>
<td>SF(1,1)</td>
<td>SF(1,1)</td>
</tr>
<tr>
<td></td>
<td>WFA state and bridges W1 into P and transmitting SF(1,1). At this point traffic for W1 is protected in both directions, however the endpoints are still not coordinated.</td>
<td>B = 1</td>
<td>B = 1</td>
</tr>
<tr>
<td>t3</td>
<td>LER-Z receives the SF(1,1) from LER-A and considers it an Ack and transits from WFA to PF:W:L state.</td>
<td>SF(1,1)</td>
<td>SF(1,1)</td>
</tr>
<tr>
<td></td>
<td>SF(1,1)</td>
<td>B = 1</td>
<td>B = 1</td>
</tr>
<tr>
<td>t4</td>
<td>LER-A receives SF(1,1), which it takes as an Ack from LER-Z and transits from WFA to PF:W:L state. Switch is complete.</td>
<td>SF(1,1)</td>
<td>SF(1,1)</td>
</tr>
<tr>
<td></td>
<td>SF(1,1)</td>
<td>B = 1</td>
<td>B = 1</td>
</tr>
</tbody>
</table>

Figure 4: Bidirectional non-locking

It is perhaps instructive to note that the only differences between the unidirectional non-locking and bidirectional non-locking scenarios are the trigger at t2 which causes Z to send SF(1,1) and the state Z finally enters (PF:W:L rather than PF:W:R). All other actions before and after this point are identical between the two cases.

3.3.2.2. Locking

We now follow the scenario for the locking mode of operation:
<table>
<thead>
<tr>
<th>Time</th>
<th>Event Description</th>
<th>LER-A PSC Bridge</th>
<th>LER-Z PSC Bridge</th>
</tr>
</thead>
<tbody>
<tr>
<td>t0</td>
<td>Traffic is being transported on W1, P is not carrying any traffic. Both LER-A and LER-Z transmitting PSC NR(0,0) message.</td>
<td>NR(0,0) B = n/a</td>
<td>NR(0,0) B = n/a</td>
</tr>
<tr>
<td>t1</td>
<td>LER-A detects SF on W1 and sends SF(1,0). LER-A enters into WFA continues to bridge and select the traffic from W1. This allows traffic to get through if the failure is really unidirectional.</td>
<td>SF(1,0) B = n/a</td>
<td>NR(0,0) B = n/a</td>
</tr>
<tr>
<td>t2</td>
<td>LER-Z detects the SF on W1. LER-Z enters WFA state and continues to bridge and select traffic from W1 while transmitting SF(1,0).</td>
<td>SF(1,0) B = n/a</td>
<td>SF(1,0) B = n/a</td>
</tr>
<tr>
<td>t3</td>
<td>LER-Z receives the SF(1,0) from LER-A and bridges traffic from W1 to P remaining in WFA state now transmitting a SF(1,1)</td>
<td>SF(1,0) B = n/a</td>
<td>SF(1,1) B = 1</td>
</tr>
<tr>
<td>t4</td>
<td>LER-A receives the SF(1,0) from LER-Z and bridges traffic from W1 to P remaining in WFA state now transmitting a SF(1,1)</td>
<td>SF(1,1) B = 1</td>
<td>SF(1,1) B = 1</td>
</tr>
<tr>
<td>t5</td>
<td>LER-A receives the SF(1,1) from LER-Z and considers it an Ack and transits from WFA to PF:W:L state</td>
<td>SF(1,1) B = 1</td>
<td>SF(1,1) B = 1</td>
</tr>
<tr>
<td>t6</td>
<td>LER-Z receives SF(1,1), which it takes as an Ack from LER-A and transits from WFA to PF:W:L state. Switch is complete.</td>
<td>SF(1,1) B = 1</td>
<td>SF(1,1) B = 1</td>
</tr>
</tbody>
</table>

Figure 5: Bidirectional locking

As with non-locking, the major difference between the unidirectional and bidirectional scenarios of this failure are the alarm which causes LER-Z to take action and the final state LER-Z enters as a result.

3.3.3. Preemption scenarios

In addition to a bidirectional failure, it is also necessary to consider preemption. When protecting n entities e.g [W1, W2, W3] it
is possible for multiple working LSPs to simultaneously fail. Consider the case where LSP W1 fails and starts to use the protection LSP. After this failure, LSP W2 fails before W1 has been restored. If W2 is of a lower relative priority than W1, there is no preemption. However, if W2 has a higher priority than W1, when W2 fails it preempts W1 from the protection LSP. Preemption is not an issue in 1:1 or 1+1, as with only a single working LSP there’s nothing to preempt.

There are multiple scenarios of preemption depending on where the failures were detected. In addition to the combinations of failure directionality and preemption, it is also necessary to consider how these combinations behave in both the locking and non-locking modes of operation.

First consider, the two flavors of preemption due to multiple unidirectional failures.

The difference between Locking and Non-Locking is that in Non-Locking a node can continue to send traffic on the P-LSP during the preemption process. The P-LSP contents may momentarily disagree (A may send W1 on P, Z may send W2 on P) but in the non-locking case there is no risk of misconnectivity as explained in the previous discussion. For this reason, the identity of the path that the endpoints are selecting incoming traffic from are irrelevant. In a sense there is no selector; each node is able to properly process arbitrary data on the P-LSP.

However, WFA state is still necessary in order to ensure that the endpoints converge on the identity of the working path whose traffic is being transported on the P-LSP. Failure to converge is a problem that should be flagged to the operator.

The scenarios start after the two endpoints have converged on protecting a unidirectional SF condition that was detected on W2, when a new SF condition is detected on W1 (with higher priority):

3.3.3.1. Unidirectional non-locking

First, consider the event sequence for unidirectional faults in a domain in non-locking mode:
Figure 6: Preemption unidirectional non-locking

As mentioned, in steady state LER-A is sending SF(2,2) and LER-Z is sending NR(0,2). If LER-A detects an SF on W1, W1 must preempt W2 in its use of the protection LSP. What the network subsequently does with W2 is outside the scope of PSC, but likely recovery actions may include rerouting W2, alerting W2’s clients as to the unprotected failure status of W2, and so forth.

3.3.3.2. Unidirectional locking

In locking operation mode, when A detects an SF on W1, it needs to alert the far-end, LER-Z, that the W2 traffic must be preempted. LER-A does this by indicating an SF on the higher priority LSP and by emptying the protection LSP. The following table presents the sequence for this scenario (we include the indication of the working path that is expected by each endpoint to be on the protection path, shown as "S = n")
## 3.3.3.3. Bidirectional non-locking

Looking, similarly, at the implications of preemption on the basic scenarios of bidirectional faults in multiple working paths. Both of the operating modes, i.e. non-locking and locking, are presented. The scenarios begin at the point where W2 traffic is being transported on the protection path in a coordinated fashion, when a SF is detected by both endpoints of the 1:n protection domain. W1 traffic has a higher priority than that of W2 traffic and, therefore, will preempt the current protected traffic.

The following presents the scenario in non-locking operation:

<table>
<thead>
<tr>
<th>Time</th>
<th>Event Description</th>
<th>LER-A PSC Bridge Selector</th>
<th>LER-Z PSC Bridge Selector</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>SF(2,2)</td>
<td>NR(0,2)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>B = 2</td>
<td>B = 2</td>
</tr>
<tr>
<td></td>
<td></td>
<td>S = 2</td>
<td>S = 2</td>
</tr>
<tr>
<td>t0</td>
<td>Traffic from W2 is being transported on P and both endpoints are coordinated</td>
<td>SF(1,0)</td>
<td>NR(0,1)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>B = n/a</td>
<td>B = 1</td>
</tr>
<tr>
<td></td>
<td></td>
<td>S = n/a</td>
<td>S = 1</td>
</tr>
<tr>
<td>t1</td>
<td>LER-A detects SF on W1 and sends SF(1,0). LER-A enters into WFA blocks all traffic on the protection path</td>
<td>SF(1,0)</td>
<td>NR(0,1)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>B = n/a</td>
<td>B = 1</td>
</tr>
<tr>
<td></td>
<td></td>
<td>S = n/a</td>
<td>S = 1</td>
</tr>
<tr>
<td>t2</td>
<td>LER-Z receives the SF(1,0) from LER-A and bridges traffic from W1 to P (higher priority), and begins transmitting NR(0,1)</td>
<td>SF(1,0)</td>
<td>NR(0,1)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>B = n/a</td>
<td>B = 1</td>
</tr>
<tr>
<td></td>
<td></td>
<td>S = n/a</td>
<td>S = 1</td>
</tr>
<tr>
<td>t3</td>
<td>LER-A receives NR(0,1) from LER-Z and considers it an Ack and transits from WFA to PF:W:L state and transmits SF(1,1)</td>
<td>SF(1,1)</td>
<td>NR(0,1)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>B = 1</td>
<td>B = 1</td>
</tr>
<tr>
<td></td>
<td></td>
<td>S = 1</td>
<td>S = 1</td>
</tr>
<tr>
<td>t4</td>
<td>LER-Z receives SF(1,1), and begins selecting the protected traffic as W1 data Switch is complete.</td>
<td>SF(1,1)</td>
<td>NR(0,1)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>B = 1</td>
<td>B = 1</td>
</tr>
<tr>
<td></td>
<td></td>
<td>S = 1</td>
<td>S = 1</td>
</tr>
</tbody>
</table>

Figure 7: Preemption unidirectional locking

Traffic loss is asymmetric. Loss A→Z starts at t1 and ends at t4, roughly 1.5xRTT. Loss Z→A starts at t1 and ends at t3, roughly 0.5xRTT.
3.3.3.4. Bidirectional locking

When considering the locking mode of operation, we must consider that the protection path, P, must be cleared of all traffic during the transition of traffic caused by preemption. The bidirectional case will be similar to the scenario for a unidirectional fault with the major difference being the final state of the two endpoints. The following would be the sequence of events:

- **t0**: Traffic from W2 is being transported on P and both endpoints are coordinated.
  - LER-A PSC Bridge: SF(2,2), B = 2
  - LER-Z PSC Bridge: NR(0,2)

- **t1**: LER-A detects SF on W1, bridges W1 into P and sends SF(1,1). LER-A enters into WFA state and continues to select the protected traffic from P that is for W2.
  - LER-A PSC Bridge: SF(1,1), B = 1
  - LER-Z PSC Bridge: NR(0,2)

- **t2**: LER-Z detects the SF on W1. LER-Z enters WFA state and bridges W1 into P and transmitting SF(1,1). At this point traffic for W1 is protected in both directions, however the endpoints are still not coordinated.
  - LER-A PSC Bridge: SF(1,1), B = 1
  - LER-Z PSC Bridge: SF(1,1)

- **t3**: LER-Z receives the SF(1,1) from LER-A and considers it an Ack and transits from WFA to PF:W:L state.
  - LER-A PSC Bridge: SF(1,1), B = 1
  - LER-Z PSC Bridge: SF(1,1)

- **t4**: LER-A receives SF(1,1), which it takes as an Ack from LER-Z and transits from WFA to PF:W:L state. Switch is complete.
  - LER-A PSC Bridge: SF(1,1), B = 1
  - LER-Z PSC Bridge: SF(1,1)

Figure 8: Preemption bidirectional non-locking
<table>
<thead>
<tr>
<th>Time</th>
<th>Event Description</th>
<th>LER-A PSC</th>
<th>LER-Z PSC</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Bridge Selector</td>
<td>Bridge Selector</td>
</tr>
<tr>
<td>t0</td>
<td>Traffic from W2 is being transported on P and both endpoints are coordinated</td>
<td>SF(2,2)</td>
<td>NR(0,2)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>B = 2</td>
<td>B = 2</td>
</tr>
<tr>
<td></td>
<td></td>
<td>S = 2</td>
<td>S = 2</td>
</tr>
<tr>
<td>t1</td>
<td>LER-A detects SF on W1 and sends SF(1,0). LER-A enters into WFA blocks all traffic on the protection path</td>
<td>SF(1,0)</td>
<td>NR(0,2)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>B = n/a</td>
<td>B = 2</td>
</tr>
<tr>
<td></td>
<td></td>
<td>S = n/a</td>
<td>S = 2</td>
</tr>
<tr>
<td>t2</td>
<td>LER-Z detects the SF on W1. LER-Z enters WFA state and blocks all traffic on the protection path while transmitting SF(1,0)</td>
<td>SF(1,0)</td>
<td>SF(1,0)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>B = n/a</td>
<td>B = n/a</td>
</tr>
<tr>
<td></td>
<td></td>
<td>S = n/a</td>
<td>S = n/a</td>
</tr>
<tr>
<td>t3</td>
<td>LER-Z receives the SF(1,0) from LER-A and bridges traffic from W1 to P (higher priority) At this point W1 traffic is flowing Z-&gt;A but not A-&gt;Z</td>
<td>SF(1,0)</td>
<td>SF(1,1)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>B = n/a</td>
<td>B = 1</td>
</tr>
<tr>
<td></td>
<td></td>
<td>S = n/a</td>
<td>S = n/a</td>
</tr>
<tr>
<td>t4</td>
<td>LER-A receives NR(0,1) from LER-Z and considers it an Ack and transits from WFA to PF:W:L state</td>
<td>SF(1,1)</td>
<td>SF(1,1)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>B = 1</td>
<td>B = 1</td>
</tr>
<tr>
<td></td>
<td></td>
<td>S = 1</td>
<td>S = n/a</td>
</tr>
<tr>
<td>t5</td>
<td>LER-Z receives SF(1,1), and begins selecting the protected traffic as W1 data and Switch is complete.</td>
<td>SF(1,1)</td>
<td>SF(1,1)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>B = 1</td>
<td>B = 1</td>
</tr>
<tr>
<td></td>
<td></td>
<td>S = 1</td>
<td>S = 1</td>
</tr>
</tbody>
</table>

Figure 9: Preemption bidirectional locking

4. Changes to PSC

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

The changes to the message structure, include both the addition of new information and extension of the semantics of some of the existing fields of the message. These changes will be described in Section 4.2.
The changes relative to the behavior of the base PSC protocol will be described in Section 4.3.

4.1. PSC

Base PSC (as defined in [LinProt]) is a single-phased protocol, i.e. the endpoints perform protection switching without waiting for acknowledgement from the far end LER. The protocol messages are transmitted using the G-ACh and the format is described in Figure 10.

```
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
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|0 0 0 1|Version|  Reserved     |       PSC-CT = 0x0024         |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|Ver|Request|PT |R|  Reserved1  |     FPath     |     Path      |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|         TLV Length            |         Reserved2             |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
˜                         Optional TLVs                         ˜
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
```

Figure 10: Format of basic PSC packet with a G-ACh header

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

4.2. Changes to PSC Payload

In order to support 1:n protection there is a need to make one small change to the format of the PSC payload (see Figure 11). In particular, we have added a new flag (L), taken from the Reserved1 space, to whether the protection domain is locking or non-locking. In addition, the semantics of the FPath and Path field are adjusted to indicate an index of the multiple working paths. The details of these changes are supplied in the following subsections.

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

The Locking flag is used to indicate that the end-point is configured for Locking mode (see Section 1.2).

If the value is 1 then the protection-domain is using the locking mode.

The Locking flag must be the same on both ends; if the two endpoints of a protection domain have different L-flag settings, this MUST raise an error to the network operator.

4.2.2. Fault path (FPath) field

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

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

4.2.3. Data path (Path) field

The Path field indicates which data is being transmitted on the protection path. Under normal conditions, the protection path does not need to carry any user data traffic, but may carry extra traffic. If there is a failure/degrade condition on one of the working paths, then that working path’s data traffic will be transmitted over the protection path. The following are the possible values:
o 0: indicates that the protection path is not transporting user
data traffic.
o 1-128: indicates that the protection path is transmitting user
traffic replacing the use of the working path indexed.
o 129-255: for future extensions or experimental use.

4.3. Changes to PSC Operation

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

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

4.3.1. Basic operation

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

4.3.2. Two-phased operation

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

In general, after the LER has detected a trigger for protection
switching, e.g. a FS operator command, or a SF indication for one of
the working paths, the LER SHALL transmit the appropriate PSC message
as described in [LinProt] with the following changes:
If the protection domain is currently in either Protecting administrative or Protecting failure state, then the endpoint SHALL verify that the new trigger has a higher priority than the currently protected traffic. If the new trigger has a lower priority then it MUST be ignored.

The PSC message SHALL set the FPath value to the index of the working path that generated the trigger. The Path value SHOULD be set to 0, unless the protection path was previously transporting traffic from another working path (as indicated by the value of the Path field.)

If the protection path is currently transporting protected traffic and the protection domain is operating in locking mode, then the endpoint SHALL block all traffic of the protected working path.

The endpoint SHALL transit to WFA state (see below).

Upon reception of the switching PSC message, the far end LER SHALL verify that the received request is of higher priority than the known current traffic on the protection path, and if so SHALL interrupt the current traffic on the protection path, perform the switch to the requested protected traffic, and send a PSC message with the Path field set to the index of the current protected working path.

Upon reception of the PSC message, the initiating LER SHALL verify that the Path field is set to the index of the working path of the highest priority. If the Path field matches the highest priority path the LER SHALL perform the protection switch and transmit the appropriate PSC message, with the FPath field indicating the index of the working path that triggered the protection switch and the Path field set to the index of the working path whose traffic is being transported on the protection path.

4.3.3. Acknowledge message

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

1. A reply message with the Request field reflecting the state of the far end, and the Path field set to the index of the working path that triggered the switching condition. For example, if there is a Forced Switch command detected by LER-Z on working path W4, then LER-Z will have sent an FS(4,0) message to LER-A. Then when LER-Z receives a message such as NR(0,4)Ack this should...
be considered acknowledgement of the switching and that the protection path is available to switch the traffic from working path W4.

2. A remote message with the same Request field and FPath field as that transmitted by the LER in the WFA state. For example, if there is a bidirectional Signal fault detected by LER-A on working path W4, then LER-A will enter WFA state and transmit a SF(4,0) message. When it receives the SF(4,0) message from LER-Z, that has also detected the SF condition, it should be considered an acknowledgement of the switching and that the protection path is available to switch the traffic from working path W2.

4.3.4. Wait for Acknowledge (WFA) timer

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

4.3.5. Additional PSC State

As described above and demonstrated in the scenarios in Section 3.3, there is a need, in some scenarios, for the endpoint that is reporting on a trigger for protection-switching to delay the actual switchover until an acknowledge is received from the far end LER. In order to facilitate this wait period it is necessary to define a new PSC State - Wait for Acknowledge (WFA) state. WFA is used in both the Locking and Non-Locking cases. It is more essential to the Locking mode of operation, as agreement is the mechanism to establish and release the lock on the protection LSP. However, it is necessary for the Non-Locking mode as a persistent disagreement on the contents of the protection LSP indicates an error in the network devices and WFA is the method used to detect this error.

In the locking mode, WFA comes into play when a failed LSP preempts another LSP. This is highlighted in the scenarios presented in Figure 7 & Figure 9.

When a working path is preempted, the protection domain must transition the contents of the protecting path from the preempted working path to the preempting working path. In the locking case, the protecting path must temporarily be blocked (that is, nothing is
being protected) in order to ensure that there is no misconnectivity. In the case where W1 preempts W2, the contents of the protection path transitions from transporting the W2 to not carrying any traffic before beginning to transport W1 traffic.

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

4.3.5.1. Wait for Acknowledge (WFA) State

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

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

- A local LO SHALL cause the LER to go into Unavailable state and begin to transmit LO(x, 0) [where x indicates the index of the working path that triggered the WFA state].

- A local FS SHALL cause the LER to remain in WFA state and transmit the FS(x, 0) message [where x indicates the index of the protected working path]. If the LER is in WFA state due to a FS from a different working path, then the working path with the higher priority SHALL be the protected working path. If the LER is in WFA state due to any other switching trigger, then the working path that is identified in this FS will be the protected working path.

- A local SF SHALL cause the LER to remain in WFA state. If the LER is in WFA state due to an existing FS trigger, then ignore the local SF and continue to transmit the FS(x, 0) PSC message. If the LER is in WFA state due to an existing SF trigger then transmit the SF(x, 0) PSC message [where x indicates the index of protected working path, i.e. the highest priority working path indicating an SF condition]. If the LER is in WFA state due to any other trigger, then begin transmitting a SF(x, 0) PSC message [where x indicates the index of the working path that is generating the SF condition].
o A local ClearSF indication where the working path is the same as the path that triggered the LER into WFA state SHALL cause the LER to go into WTR state (note: 1:N protection is always revertive) and to transmit the WTR(0, 0) message. If the ClearSF indicates a different index from the protected working path or incates the protection path then the indication SHALL be ignored.

o A local MS operator command SHALL cause the LER to remain in WFA state. If the LER is in WFA state due an existing MS trigger, then the node continues to transmit MS(x, 0) messages [where x indicates the index of the protected working path, i.e. the highest priority working path indicating the MS condition]. If the LER is in WFA state due to any other trigger, ignore the MS command and continue transmitting the current message.

o If the WFA timer expires, i.e. the LER did not receive the Acknowledge message from the far end in a timely manner, then the LER SHALL go to Unavailable state, i.e. it assumes that there is a problem on the protection path (where all PSC traffic is transmitted) and send an error notification to the management system. The LER SHALL continue transmitting the current PSC message with Path field set to 0.

o All other local indications SHALL be ignored.

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

o Any remote message with the Acknowledge flag set to 1 and the Path field set to the index of the protected working path SHALL cause the LER to change state. If the trigger was either FS or MS command, the LER enters Protecting administrative state. The LER transmits the appropriate message according to the trigger (i.e. FS(x,x) for FS command and MS(x,x) for the MS command). If the trigger was a SF condition, then the LER enters the Protecting failure state and begins to transmit the appropriate SF(x, x) message. A remote message with the Acknowledge flag set to 1 but where the Path field does not match, according to the description above, SHALL be ignored.

o A remote LO message SHALL cause the LER to go into Unavailable state and transmit the appropriate message for the trigger that caused the WFA state.

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

- A remote SF message indicating an error on the protection path SHALL cause the LER to go into Unavailable state and transmit the appropriate message for the trigger that caused to WFA state.

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

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

- All other remote messages SHOULD be ignored.

5. IANA Considerations

This document does not include any required IANA considerations.

6. Security Considerations

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

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

7. Acknowledgements

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

8. References

8.1. Normative References

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

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

[LinProt] Bryant, S., Sprecher, N., Osborne, E., Fulignoli, A., and

8.2. Informative References


Appendix A. PSC state machine tables

Note/Disclaimer: This state machine is not currently in sync with the text of the document and will be updated in a future revision.

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

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

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

<table>
<thead>
<tr>
<th>State</th>
<th>REQ(FP,P)</th>
</tr>
</thead>
<tbody>
<tr>
<td>N</td>
<td>NR(0,0)</td>
</tr>
<tr>
<td>UA:LO:L</td>
<td>LO(0,0)</td>
</tr>
<tr>
<td>UA:P:L</td>
<td>SF(0,0)</td>
</tr>
<tr>
<td>UA:LO:R</td>
<td>NR(0,0)</td>
</tr>
<tr>
<td>UA:P:R</td>
<td>NR(0,0)</td>
</tr>
<tr>
<td>PF:W:L</td>
<td>SF(1,1)</td>
</tr>
<tr>
<td>PF:W:R</td>
<td>NR(0,1)</td>
</tr>
<tr>
<td>PA:F:L</td>
<td>FS(1,1)</td>
</tr>
<tr>
<td>PA:M:L</td>
<td>MS(1,1)</td>
</tr>
<tr>
<td>PA:F:R</td>
<td>NR(0,1)</td>
</tr>
<tr>
<td>PA:M:R</td>
<td>NR(0,1)</td>
</tr>
<tr>
<td>WTR</td>
<td>WTR(0,1)</td>
</tr>
<tr>
<td>DNR</td>
<td>DNR(0,1)</td>
</tr>
</tbody>
</table>

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

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

and the remote inputs are:

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

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

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

**Part 1: Local input state machine**

<table>
<thead>
<tr>
<th>OC</th>
<th>LO</th>
<th>SF-P</th>
<th>FS</th>
<th>SF-W</th>
<th>SFc</th>
<th>MS</th>
<th>WTExp</th>
</tr>
</thead>
<tbody>
<tr>
<td>UA:LO:L</td>
<td>N</td>
<td>i</td>
<td>i</td>
<td>i</td>
<td>i</td>
<td>i</td>
<td>i</td>
</tr>
<tr>
<td>UA:P:L</td>
<td>i</td>
<td>UA:LO:L</td>
<td>i</td>
<td>i</td>
<td>[5]</td>
<td>i</td>
<td>i</td>
</tr>
<tr>
<td>PA:F:L</td>
<td>N</td>
<td>UA:LO:L</td>
<td>UA:P:L</td>
<td>i</td>
<td>i</td>
<td>i</td>
<td>i</td>
</tr>
</tbody>
</table>
### Part 2: Remote messages state machine

<table>
<thead>
<tr>
<th>LO</th>
<th>SF-P</th>
<th>FS</th>
<th>SF-W</th>
<th>MS</th>
<th>WTR</th>
<th>DNR</th>
<th>NR</th>
</tr>
</thead>
<tbody>
<tr>
<td>UA:LO:L</td>
<td>i</td>
<td>i</td>
<td>i</td>
<td>i</td>
<td>i</td>
<td>i</td>
<td>i</td>
</tr>
<tr>
<td>UA:LO:R</td>
<td>i</td>
<td>i</td>
<td>i</td>
<td>i</td>
<td>i</td>
<td>i</td>
<td>[16]</td>
</tr>
<tr>
<td>UA:P:R</td>
<td>i</td>
<td>i</td>
<td>i</td>
<td>i</td>
<td>i</td>
<td>i</td>
<td>i</td>
</tr>
<tr>
<td>PA:F:L</td>
<td>UA:LO:R</td>
<td>UA:P:R</td>
<td>i</td>
<td>i</td>
<td>i</td>
<td>i</td>
<td>[17]</td>
</tr>
</tbody>
</table>

The following are the footnotes for the table:

1. Remain in the current state (UA:LO:R) and transmit SF(0,0)
2. Remain in the current state (UA:LO:R) and transmit SF(1,0)
3. Remain in the current state (UA:P:R) and transmit SF(1,0)
4. Remain in the current state (PA:F:R) and transmit SF(1,1)
5. If the SF being cleared is SF-P, Transition to N. If it’s SF-W, ignore the clear.
6. Remain in current state (UA:x:R), if the SFc corresponds to a previous SF then begin transmitting NR(0,0).
7. If domain configured for revertive behavior transition to WTR, else transition to DNR
8. Remain in PA:F:R and transmit NR(0,1)
9. Remain in WTR, send NR(0,1)
10. Transition to UA:LO:R continue sending SF(0,0)
11. Transition to UA:LO:R and send SF(1,0)
12. Transition to UA and send SF(1,0)
13. Transition to PF:W:R and send NR(0,1)
[14] Transition to WTR state and continue to send the current message.

[15] Transition to DNR state and continue to send the current message.

[16] If the local input is SF-P then transition to UA:P:L. If the local input is SF-W then transition to PF:W:L. Else – transition to N state and continue to send the current message.

[17] If the local input is SF-W then transition to PF:W:L. Else – transition to N state and continue to send the current message.

[18] If the receiving LER’s WTR timer is running, maintain current state and message. If the WTR timer is stopped, transition to N.

Authors’ Addresses

Eric Osborne
Cisco
United States

Email: eosborne@cisco.com

Fei Zhang
ZTE
China

Email: zhang.fe3@zte.com.cn

Yaacov Weingarten
Nokia Siemens Networks
34 Hagefen St
Karniel Shomron, 44853
Israel

Email: wyaacov@gmail.com

Abstract

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

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

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

Requirements Language

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

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

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

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

This Internet-Draft will expire on September 10, 2012.

Copyright Notice

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

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

1. Introduction ............................................. 4
2. Architecture requirements overview ..................... 4
   2.1. Communicate Latency and Loss as TE Metric .......... 4
2.2. Requirement for Composite Link ....................... 5
2.3. Requirement for Hierarchy LSP ......................... 5
2.4. Latency Accumulation and Verification ................. 5
2.5. Restoration, Protection and Rerouting ................. 6
3. End-to-End Latency ....................................... 7
4. End-to-End Jitter ........................................ 8
5. End-to-End Loss .......................................... 8
6. Protocol Considerations ................................... 9
7. Control Plane Implication ................................ 9
   7.1. Implications for Routing ............................. 9
   7.2. Implications for Signaling ........................... 11
8. IANA Considerations ...................................... 12
9. Security Considerations .................................. 12
10. Acknowledgements ........................................ 12
11. References .............................................. 12
   11.1. Normative References ............................... 12
   11.2. Informative References .............................. 13
Authors’ Addresses ........................................... 13
1. Introduction

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

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

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

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

2. Architecture requirements overview

2.1. Communicate Latency and Loss as TE Metric

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

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

In the case where it is known that either the changes are too frequent or there is a backup which is preferred, the solution shall put the node or the link in unusable state for services requiring a particular service capability. This unusable state is on a capability basis and not a global basis. The condition to get into the state is locally configured and all routers in a domain should have this criteria synchronized.
Path computation entity MUST have the capability to compute one end-to-end path with latency and packet loss constraint. For example, it has the capability to compute a route with X amount of bandwidth with less than Y ms of latency and less than Z% packet loss limit based on the latency and packet loss traffic engineering database. It MUST also support the path computation with routing constraints combination with pre-defined priorities, e.g., SRLG diversity, latency, loss, jitter and cost. If the performance of link exceeds its configured maximum threshold, path computation entity may not select this kind of link although end-to-end performance is still met.

2.2. Requirement for Composite Link

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

The solution MUST provide a means to indicate that a traffic flow should select a component link with minimum latency and/or packet loss, maximum acceptable latency and/or packet loss value and maximum acceptable delay variation value as specified by protocol. The endpoints of Composite Link will take these parameters into account for component link selection or creation. Details of how transient response is taken is specified in Section 4.1 [CL-REQ]. The exact details for component links will be taken up separately and are not part of this document.

2.3. Requirement for Hierarchy LSP

Hierarchical LSP’s may traverse server layer LSP’s. For such LSP’s there may be some latency and packet loss constraint requirement for the segment in server layer.

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

2.4. Latency Accumulation and Verification

The solution SHOULD provide a means to accumulate (e.g., sum) latency information of links and nodes along that an LSP traverses, (e.g., Inter-AS, Inter-Area or Multi-Layer) so that the source node can validate if the desired maximum latency constraint can be satisfied
for a packet traversing the LSP. [Y.1541] provides details of how the latency value is accumulated.

Both One-way and Round-trip latency collection along the LSP by signaling protocol and latency verification at the end of LSP should be supported.

The accumulation of the delay is "simple" for the static component i.e. its a linear addition, the dynamic/network loading component is more interesting and would involve some estimate of the "worst case". However, method of deriving this worst case appears to be more in the scope of Network Operator policy than standards i.e. the operator needs to decide, based on the SLAs offered, the required confidence level.

2.5. Restoration, Protection and Rerouting

Some customers may insist on having the ability to re-route if the latency and loss SLA is not being met. If a "provisioned" end-to-end LSP latency and/or loss could not meet the latency and loss agreement between operator and his user, the solution SHOULD support pre-defined or dynamic re-routing (e.g., make-before-break) to handle this case based on the local policy. In revertive behaviour is supported, the original LSP must not be released and is monitored by control plane. When the end-to-end performance is repaired, the service is restored to the original LSP.

The solution SHOULD support to move an end-to-end LSP away from any link whose performance violates the configured threshold.

End-to-end measurements of the LSP also need to be performed in addition to the link-by-link measurements. A threshold violation of the End-to-End criteria as measured by the head end node should cause rerouting of the LSP.

The anomalous path can be switch to protection path or rerouted to new path because of end-to-end performance couldn’t meet any more.

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

The latency performance of pre-defined protection or dynamic re-routing LSP MUST meet the latency SLA parameter.

Due to some flapping conditions the latency and loss of an LSP may change, this may cause the LSP to be frequently switched to a new
path. In order to avoid churn, the solution SHOULD specify the switchover of the LSP according to maximum acceptable change rate.

3. End-to-End Latency

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

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

Latency or one-way delay is the time it takes for a packet within a stream going from measurement point 1 to measurement point 2, as defined in [Y.1540].

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

The total measured latency of an LSP consists of the sum of the latency of the LSP hop, as well as the average latency of switching on a device, which may vary based on queuing and buffering.

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

The switching latency on a device, can be measured internally, and multiple mechanisms and data structures to do the same have been defined. [Add references to papers by Verghese, Kompella, Duffield].

We also looked at other measurement granularities before deciding on an interface based measurement. An approximation of the Flow based measurement is the per DSCP value, measurement from the ingress of one port to the egress of every other port in the device.

Another approximation that can be used is per interface DSCP based measurement, which can be an aggregate of the average measurements per interface. The average can itself be calculated in ways, so as to provide closer approximation.
For the purpose of this draft it is assumed that the node latency is a small factor of the total latency in the networks where this solution is deployed. The node latency is hence ignored for the benefit of simplicity in this solution.

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

4. End-to-End Jitter

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

This architecture uses the assumptions of [Y.1540] to calculate the accumulated jitter from the individual components approximately. Though using this may not be perfect, it however gives a good approximation that can be used for Traffic Engineering (TE) purposes.

The buffering and queuing within a device will lead to the jitter. Just like latency measurements, jitter measurements can be approximated as either per DSCP per port pair (Ingress and Egress) or as per DSCP per egress port, however such measurements have been left out for the sake of simplicity of the solution.

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

The jitter is measured in micro-seconds.

5. End-to-End Loss

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

This architecture uses the assumptions of [Y.1540] to calculate the accumulated loss from the individual components approximately. Though using the accumulated metrics may not be perfect, it however gives a good approximation that can be used for Traffic Engineering (TE) purposes.

The buffering and queuing mechanisms within a device will decide which packet is to be dropped. Just like latency and jitter
measurements, the loss can best be approximated as either per DSCP per port pair (Ingress and Egress) or as per DSCP per egress port. However such mechanisms are not used in this solution to keep the solution simple.

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

6. Protocol Considerations

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

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

A separate measurement should be done for an LSP when it is UP. Also LSP's path should only be recalculated when the end-to-end metrics changes in a way it becomes more than desired.

7. Control Plane Implication

7.1. Implications for Routing

The latency and packet loss performance metric MUST be advertised into path computation entity by IGP (OSPF-TE, OSPFv3-TE or IS-IS-TE) to perform route computation and network planning based on latency and packet loss SLA target.

Latency, latency variation and packet loss value MUST be reported as a average value which is calculated by data plane measurements.

Latency and packet loss characteristics of these links and nodes may change dynamically. In order to control IGP messaging and avoid being unstable when the latency, latency variation and packet loss value changes, a threshold and a limit on rate of change MUST be configured in the IGP control plane.

Latency and packet loss values changes need to be updated and flooded in the IGP control messages only when there is significant changes in the value. When the head end-node determines the IGP update affects the LSP for which it is ingress, it recalculates the LSP.
A target value MUST be configured to control plane for each link. If the link performance improves beyond a configurable target value, it must be re-advertised. The receiving node determines whether a "provisioned" end-to-end LSP latency and/or loss performance is improved.

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

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

- One option is that the latency and packet loss of composite link may be the range (e.g., at least minimum and maximum) latency value of all component links. It may also be the maximum or average latency value of all component links. In both cases, only partial information is transmitted in the IGP. So the path computation entity has insufficient information to determine whether a particular path can support its latency and packet loss requirements. This leads to signaling crankback.

- Another option is that latency and packet loss of each component link within one Composite Link could be advertised but having only one IGP adjacency.

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

If the FA-LSP is able to form a routing adjacency and/or as a TE link in the client network, the total latency and packet loss value of the FA-LSP can be as an input to a transformation that results in a FA traffic engineering metric and advertised into the client layer routing instances. Note that this metric will include the latency and packet loss of the links and nodes that the trail traverses.

If total latency and packet loss information of the FA-LSP changes (e.g., due to a maintenance action or failure in OTN rings), the boundary node of the FA-LSP will receive the TE link information.
advertisement including the latency and packet value which is already changed and if it is over than the threshold and a limit on rate of change, then it will compute the total latency and packet value of the FA-LSP again. If the total latency and packet loss value of FA-LSP changes, the client layer MUST also be notified about the latest value of FA. The client layer can then decide if it will accept the increased latency and packet loss or request a new path that meets the latency and packet loss requirement.

7.2. Implications for Signaling

In order to assign the LSP to one of component links with different latency and loss characteristics, RSVP-TE message needs to carry a indication of request minimum latency and/or packet loss, maximum acceptable latency and/or packet loss value and maximum acceptable delay variation value for the component link selection or creation. The composite link will take these parameters into account when assigning traffic of LSP to a component link.

One end-to-end LSP (e.g., in IP/MPLS or MPLS-TP network) may traverse a FA-LSP of server layer (e.g., OTN rings). There will be some latency and packet loss constraint requirement for the segment route in server layer. So RSVP-TE message needs to carry a indication of request minimum latency and/or packet loss, maximum acceptable latency and/or packet loss value and maximum acceptable delay variation value. The boundary nodes of FA-LSP will take these parameters into account for FA selection or FA-LSP creation.

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

Restoration, protection and equipment variations can impact "provisioned" latency and packet loss (e.g., latency and packet loss increase). For example, restoration/provisioning action in transport network that increases latency seen by packet network observable by customers, possibly violating SLAs. The change of one end-to-end LSP latency and packet loss performance MUST be known by source and/or sink node. So it can inform the higher layer network of a latency and packet loss change. The latency or packet loss change of links and nodes will affect one end-to-end LSPs total amount of latency or
packet loss. Applications can fail beyond an application-specific threshold. Some remedy mechanism could be used.

Pre-defined protection or dynamic re-routing could be triggered to handle this case. In the case of predefined protection, large amounts of redundant capacity may have a significant negative impact on the overall network cost. Service provider may have many layers of pre-defined restoration for this transfer, but they have to duplicate restoration resources at significant cost. Solution should provides some mechanisms to avoid the duplicate restoration and reduce the network cost. Dynamic re-routing also has to face the risk of resource limitation. So the choice of mechanism MUST be based on SLA or policy. In the case where the latency SLA can not be met after a re-route is attempted, control plane should report an alarm to management plane. It could also try restoration for several times which could be configured.

8. IANA Considerations

No new IANA consideration are raised by this document.

9. Security Considerations

This document raises no new security issues.

10. Acknowledgements

TBD.

11. References

11.1. Normative References


11.2. Informative References


Authors' Addresses

Xihua Fu
ZTE

Email: fu.xihua@zte.com.cn
John Drake
Juniper Networks

Email: jdrake@juniper.net
The Use of Entropy Labels in MPLS Forwarding

draft-ietf-mpls-entropy-label-01

Abstract

Load balancing is a powerful tool for engineering traffic across a network. This memo suggests ways of improving load balancing across MPLS networks using the concept of "entropy labels". It defines the concept, describes why entropy labels are useful, enumerates properties of entropy labels that allow maximal benefit, and shows how they can be signaled and used for various applications.

Status of this Memo

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

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

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

This Internet-Draft will expire on May 3, 2012.

Copyright Notice

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

Table of Contents

1. Introduction ........................................... 3
  1.1. Conventions used ................................... 4
  1.2. Motivation .......................................... 5
2. Approaches ............................................. 6
3. Entropy Labels and Their Structure ....................... 7
4. Data Plane Processing of Entropy Labels ................. 8
  4.1. Ingress LSR .......................................... 8
  4.2. Transit LSR ......................................... 9
  4.3. Egress LSR .......................................... 9
5. Signaling for Entropy Labels ........................... 10
  5.1. LDP Signaling ...................................... 10
  5.2. BGP Signaling ....................................... 11
  5.3. RSVP-TE Signaling .................................. 12
6. Operations, Administration, and Maintenance (OAM) and
   Entropy Labels .......................................... 13
7. MPLS-TP and Entropy Labels ............................. 14
8. Point-to-Multipoint LSPs and Entropy Labels ............. 15
9. Entropy Labels and Applications ......................... 15
  9.1. Tunnels ............................................. 15
  9.2. LDP Pseudowires .................................... 17
  9.3. BGP Applications .................................... 18
    9.3.1. Inter-AS BGP VPNs ............................... 19
  9.4. Multiple Applications ............................... 20
10. Security Considerations ................................ 21
11. IANA Considerations ................................... 22
  11.1. LDP Entropy Label TLV .............................. 22
  11.2. BGP Entropy Label Attribute ........................ 22
  11.3. Attribute Flags for LSP_Attributes Object ........... 22
  11.4. Attributes TLV for LSP Attributes Object .......... 22
12. Acknowledgments ....................................... 23
13. References ............................................ 23
  13.1. Normative References ............................... 23
  13.2. Informative References ............................. 23
Appendix A. Applicability of LDP Entropy Label sub-TLV ..... 24
Authors’ Addresses ......................................... 25
1. Introduction

Load balancing, or multi-pathing, is an attempt to balance traffic across a network by allowing the traffic to use multiple paths. Load balancing has several benefits: it eases capacity planning; it can help absorb traffic surges by spreading them across multiple paths; it allows better resilience by offering alternate paths in the event of a link or node failure.

As providers scale their networks, they use several techniques to achieve greater bandwidth between nodes. Two widely used techniques are: Link Aggregation Group (LAG) and Equal-Cost Multi-Path (ECMP). LAG is used to bond together several physical circuits between two adjacent nodes so they appear to higher-layer protocols as a single, higher bandwidth 'virtual' pipe. ECMP is used between two nodes separated by one or more hops, to allow load balancing over several shortest paths in the network. This is typically obtained by arranging IGP metrics such that there are several equal cost paths between source-destination pairs. Both of these techniques may, and often do, co-exist in various parts of a given provider’s network, depending on various choices made by the provider.

A very important requirement when load balancing is that packets belonging to a given 'flow' must be mapped to the same path, i.e., the same exact sequence of links across the network. This is to avoid jitter, latency and re-ordering issues for the flow. What constitutes a flow varies considerably. A common example of a flow is a TCP session. Other examples are an L2TP session corresponding to a given broadband user, or traffic within an ATM virtual circuit.

To meet this requirement, a node uses certain fields, termed 'keys', within a packet’s header as input to a load balancing function (typically a hash function) that selects the path for all packets in a given flow. The keys chosen for the load balancing function depend on the packet type; a typical set (for IP packets) is the IP source and destination addresses, the protocol type, and (for TCP and UDP traffic) the source and destination port numbers. An overly conservative choice of fields may lead to many flows mapping to the same hash value (and consequently poorer load balancing); an overly aggressive choice may map a flow to multiple values, potentially violating the above requirement.

For MPLS networks, most of the same principles (and benefits) apply. However, finding useful keys in a packet for the purpose of load balancing can be more of a challenge. In many cases, MPLS encapsulation may require fairly deep inspection of packets to find these keys at transit LSRs.
One way to eliminate the need for this deep inspection is to have the ingress LSR of an MPLS Label Switched Path extract the appropriate keys from a given packet, input them to its load balancing function, and place the result in an additional label, termed the ‘entropy label’, as part of the MPLS label stack it pushes onto that packet.

The packet’s MPLS entire label stack can then be used by transit LSRs to perform load balancing, as the entropy label introduces the right level of "entropy" into the label stack.

There are four key reasons why this is beneficial:

1. at the ingress LSR, MPLS encapsulation hasn’t yet occurred, so deep inspection is not necessary;
2. the ingress LSR has more context and information about incoming packets than transit LSRs;
3. ingress LSRs usually operate at lower bandwidths than transit LSRs, allowing them to do more work per packet, and
4. transit LSRs do not need to perform deep packet inspection and can load balance effectively using only a packet’s MPLS label stack.

This memo describes why entropy labels are needed and defines the properties of entropy labels; in particular how they are generated and received, and the expected behavior of transit LSRs. Finally, it describes in general how signaling works and what needs to be signaled, as well as specifics for the signaling of entropy labels for LDP ([RFC5036]), BGP ([RFC3107], [RFC4364]), and RSVP-TE ([RFC3209]).

1.1. Conventions used

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

The following acronyms are used:

- LSR: Label Switching Router;
- LER: Label Edge Router;
- PE: Provider Edge router;
CE: Customer Edge device; and
FEC: Forwarding Equivalence Class.

The term ingress (or egress) LSR is used interchangeably with ingress (or egress) LER. The term application throughout the text refers to an MPLS application (such as a VPN or VPLS).

A label stack (say of three labels) is denoted by \(<L1, L2, L3>\), where \(L1\) is the "outermost" label and \(L3\) the innermost (closest to the payload). Packet flows are depicted left to right, and signaling is shown right to left (unless otherwise indicated).

The term ‘label’ is used both for the entire 32-bit label and the 20-bit label field within a label. It should be clear from the context which is meant.

1.2. Motivation

MPLS is very successful generic forwarding substrate that transports several dozen types of protocols, most notably: IP, PWE3, VPLS and IP VPNs. Within each type of protocol, there typically exist several variants, each with a different set of load balancing keys, e.g., for IP: IPv4, IPv6, IPv6 in IPv4, etc.; for PWE3: Ethernet, ATM, Frame-Relay, etc. There are also several different types of Ethernet over PW encapsulation, ATM over PW encapsulation, etc. as well. Finally, given the popularity of MPLS, it is likely that it will continue to be extended to transport new protocols.

Currently, each transit LSR along the path of a given LSP has to try to infer the underlying protocol within an MPLS packet in order to extract appropriate keys for load balancing. Unfortunately, if the transit LSR is unable to infer the MPLS packet’s protocol (as is often the case), it will typically use the topmost (or all) MPLS labels in the label stack as keys for the load balancing function. The result may be an extremely inequitable distribution of traffic across equal-cost paths exiting that LSR. This is because MPLS labels are generally fairly coarse-grained forwarding labels that typically describe a next-hop, or provide some of demultiplexing and/or forwarding function, and do not describe the packet’s underlying protocol.

On the other hand, an ingress LSR (e.g., a PE router) has detailed knowledge of an packet’s contents, typically through a priori configuration of the encapsulation(s) that are expected at a given PE-CE interface, (e.g., IPv4, IPv6, VPLS, etc.). They also have more flexible forwarding hardware. PE routers need this information and these capabilities to:
a) apply the required services for the CE;
b) discern the packet’s CoS forwarding treatment;
c) apply filters to forward or block traffic to/from the CE;
d) to forward routing/control traffic to an onboard management processor; and,
e) load-balance the traffic on its uplinks to transit LSRs (e.g., P routers).

By knowing the expected encapsulation types, an ingress LSR router can apply a more specific set of payload parsing routines to extract the keys appropriate for a given protocol. This allows for significantly improved accuracy in determining the appropriate load balancing behavior for each protocol.

If the ingress LSR were to capture the flow information so gathered in a convenient form for downstream transit LSRs, transit LSRs could remain completely oblivious to the contents of each MPLS packet, and use only the captured flow information to perform load balancing. In particular, there will be no reason to duplicate an ingress LSR’s complex packet/payload parsing functionality in a transit LSR. This will result in less complex transit LSRs, enabling them to more easily scale to higher forwarding rates, larger port density, lower power consumption, etc. The idea in this memo is to capture this flow information as a label, the so-called entropy label.

Ingress LSRs can also adapt more readily to new protocols and extract the appropriate keys to use for load balancing packets of those protocols. This means that deploying new protocols or services in edge devices requires fewer concommitant changes in the core, resulting in higher edge service velocity and at the same time more stable core networks.

2. Approaches

There are two main approaches to encoding load balancing information in the label stack. The first allocates multiple labels for a particular Forwarding Equivalence Class (FEC). These labels are equivalent in terms of forwarding semantics, but having multiple labels allows flexibility in assigning labels to flows belonging to the same FEC. This approach has the advantage that the label stack has the same depth whether or not one uses label-based load balancing; and so, consequently, there is no change to forwarding operations on transit and egress LSRs. However, it has a major
The drawback in that there is a significant increase in both signaling and forwarding state.

The other approach encodes the load balancing information as an additional label in the label stack, thus increasing the depth of the label stack by one. With this approach, there is minimal change to signaling state for a FEC; also, there is no change in forwarding operations in transit LSRs, and no increase of forwarding state in any LSR. The only purpose of the additional label is to increase the entropy in the label stack, so this is called an "entropy label". This memo focuses solely on this approach.

3. Entropy Labels and Their Structure

An entropy label (as used here) is a label:

1. that is not used for forwarding;
2. that is not signaled; and
3. whose only purpose in the label stack is to provide 'entropy' to improve load balancing.

Entropy labels are generated by an ingress LSR, based entirely on load balancing information. However, they MUST NOT have values in the reserved label space (0-15). To ensure that they are not used inadvertently for forwarding, entropy labels SHOULD have a TTL of 0. The CoS field of an entropy label can be set to any value deemed appropriate.

Since entropy labels are generated by an ingress LSR, an egress LSR MUST be able to tell unambiguously that a given label is an entropy label. If any ambiguity is possible, the label above the entropy label MUST be an ‘entropy label indicator’ (ELI), which indicates that the following Label is an entropy label. An ELI is typically signaled by an egress LSR and is added to the MPLS label stack along with an entropy label by an ingress LSR. For many applications, the use of entropy labels is unambiguous, and an ELI is not needed. An ELI MUST have ‘Bottom of Stack’ (S) bit = 0 ([RFC3032]). The TTL SHOULD be set to whatever value the label above it in the stack has. The CoS field can be set to any value deemed appropriate; typically, this will be the value in the label above it in the stack.

Applications for MPLS entropy labels include pseudowires ([RFC4447]), Layer 3 VPNs ([RFC4364]), VPLS ([RFC4761], [RFC4762]) and Tunnel LSPs carrying, say, IP traffic. [I-D.ietf-pwe3-fat-pw] explains how entropy labels can be used for RFC 4447-style pseudowires, and thus...
is complementary to this memo, which focuses on several other applications of entropy labels.

4. Data Plane Processing of Entropy Labels

4.1. Ingress LSR

Suppose that for a particular application (or service or FEC), an ingress LSR X is to push label stack <TL, AL>, where TL is the ‘tunnel label’ and AL is the ‘application label’. (Note the use of the convention for label stacks described in Section 1.1. The use of a two-label stack is just for illustrative purposes.) Suppose furthermore that the egress LSR Y has told X that it is capable of processing entropy labels for this application. If X cannot insert entropy labels, it simply uses a label stack of <TL, AL> for this application. If X can insert entropy labels, it does the following for an incoming packet:

1. X identifies the application to which the packet belongs, identifies the egress LSR as Y, and thereby picks the outgoing label stack <TL, AL> to push onto the packet to send to Y.

2. X determines which keys that it will use for load balancing.

3. X, having kept state that Y can process entropy labels for this application, generates an entropy label EL (based on the output of the load balancing function).

4. If Y does not need an ELI, X pushes <TL, AL, EL> onto the packet before forwarding it to the next hop to Y.

5. If Y requires an ELI, X pushes <TL, AL, E, EL> onto the packet before forwarding it to the next hop to Y, where E is a label whose 20-bit label field is the ELI that Y signaled, and whose other fields are set as per Section 3.

Note that ingress LSR X MUST NOT include an entropy label unless the egress LSR Y for this application has indicated that it is ready to receive entropy labels. Furthermore, if Y has signaled that an ELI is needed, then X MUST include the ELI before the entropy label.

Note that the signaling and use of entropy labels in one direction (signaling from Y to X, and data path from X to Y) has no bearing on the behavior in the opposite direction (signaling from X to Y, and data path from Y to X).
4.2. Transit LSR

Transit LSRs have virtually no change in forwarding behavior. For load balancing, transit LSRs SHOULD use the whole label stack as keys for the load balancing function. Transit LSRs MUST NOT include reserved labels as input to its load balancing function. Transit LSRs MAY choose to look beyond the label stack for further keys; however, if entropy labels are being used, this may not be very useful. Looking beyond the label stack may be the simplest approach in an environment where some ingress LSRs use entropy labels and others don’t, or for backward compatibility. Thus, other than using the full label stack as input to the load balancing function, transit LSRs are almost unaffected by the use of entropy labels.

4.3. Egress LSR

Suppose egress LSR Y signals that it is capable of processing entropy labels for a tunnel or an application with label L. There are three cases of interest: (a) L is the implicit NULL label, in which case an ELI is mandatory; (b) L is not the implicit NULL label and an ELI is not required (L’s S bit will be used to determine whether or not there is an EL); and (c) L is not the implicit NULL label but an ELI is required.

a1) Y receives an unlabeled packet. There is obviously no EL; Y processes the packet as usual.

a2) Y receives a packet whose top label is the ELI. Y processes the TTL and CoS fields of the ELI label, ensures that the S bit is 0, then pops it, and pops the next label as well (which must be the EL), then pops it. Y processes the remaining payload as usual.

b) Y receives a packet with top label L, and an ELI is not required. Y processes L as usual; if L’s S bit is 1, the label stack is done. If L’s S bit is 0, the following label is the EL. Y pops the EL. Y processes the payload as usual.

c) Y receives a packet with top label L. Y processes L as usual; if L’s S bit is 1, the label stack is done. If L’s S bit is 0, Y checks the following label. If it is the ELI label, Y processes the TTL and CoS fields of the ELI, ensures that the S bit is 0, pops the ELI label and the following label (which is the EL), and processes the remaining payload as usual.

If there is an ELI with S bit = 1, there is an error in the label stack. Note that the TTL field of the EL (if present) will be 0; Y MUST NOT react to this.
5. Signaling for Entropy Labels

An egress LSR Y may signal to ingress LSR(s) its ability to process entropy labels on a per-application (or per-FEC) basis. As part of this signaling, Y also signals the ELI to use, if any.

In cases where an application label is used and must be the bottommost label in the label stack, Y MAY signal that no ELI is needed for that application.

In cases where no application label exists, or where the application label may not be the bottommost label in the label stack, Y MUST signal a valid ELI to be used in conjunction with the entropy label for this FEC. In this case, an ingress LSR will either not add an entropy label, or push the ELI before the entropy label. This makes the use or non-use of an entropy label by the ingress LSR unambiguous. Valid ELI label values are strictly greater than 15.

It should be noted that egress LSR Y may use the same ELI value for all applications for which an ELI is needed. The ELI MUST be a label that does not conflict with any other labels that Y has advertised to other LSRs for other applications. Furthermore, it should be noted that the ability to process entropy labels (and the corresponding ELI) may be asymmetric: an LSR X may be willing to process entropy labels, whereas LSR Y may not be willing to process entropy labels. The signaling extensions below allow for this asymmetry.

For an illustration of signaling and forwarding with entropy labels, see Figure 9.

5.1. LDP Signaling

When using LDP for signaling tunnel labels ([RFC5036]), a Label Mapping Message sub-TLV (Entropy Label sub-TLV) is used to signal an egress LSR’s ability to process entropy labels.

The presence of the Entropy Label sub-TLV in the Label Mapping Message indicates to ingress LSRs that the egress LSR can process an entropy label. In addition, the Entropy Label sub-TLV contains a label value for the ELI. If the ELI is zero, this indicates the egress doesn’t need an ELI for the signaled application; if not, the egress requires the given ELI with entropy labels. An example where an ELI is needed is when the signaled application is an LSP that can carry IP traffic.
The structure of the Entropy Label sub-TLV is shown below.

```
+---+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|U|F|        Type (TBD)         |           Length (8)          |
+---+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|               Value                   |     Must Be Zero      |
+---+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
```

Figure 1: Entropy Label sub-TLV

where:

U: Unknown bit. This bit MUST be set to 1. If the Entropy Label sub-TLV is not understood, then the TLV is not known to the receiver and MUST be ignored.

F: Forward bit. This bit MUST be set to 1. Since this sub-TLV is going to be propagated hop-by-hop, the sub-TLV should be forwarded even by nodes that may not understand it.

Type: sub-TLV Type field, as specified by IANA.

Length: sub-TLV Length field. This field specifies the total length in octets of the Entropy Label sub-TLV.

Value: value of the Entropy Label Indicator Label.

5.2. BGP Signaling

When BGP [RFC4271] is used for distributing Network Layer Reachability Information (NLRI) as described in, for example, [RFC3107], [RFC4364] and [RFC4761], the BGP UPDATE message may include the Entropy Label attribute. This is an optional, transitive BGP attribute of type TBD. The inclusion of this attribute with an NLRI indicates that the advertising BGP router can process entropy labels as an egress LSR for that NLRI. If the attribute length is less than three octets, this indicates that the egress doesn’t need an ELI for the signaled application. If the attribute length is at least three octets, the first three octets encode an ELI label value as the high order 20 bits; the egress requires this ELI with entropy labels. An example where an ELI is needed is when the NLRI contains unlabeled IP prefixes.

A BGP speaker S that originates an UPDATE should only include the Entropy Label attribute if both of the following are true:
A1: S sets the BGP NEXT_HOP attribute to itself; AND

A2: S can process entropy labels for the given application.

If both A1 and A2 are true, and S needs an ELI to recognize entropy labels, then S MUST include the ELI label value as part of the Entropy Label attribute. An UPDATE SHOULD contain at most one Entropy Label attribute.

Suppose a BGP speaker T receives an UPDATE U with the Entropy Label attribute ELA. T has two choices. T can simply re-advertise U with the same ELA if either of the following is true:

B1: T does not change the NEXT_HOP attribute; OR

B2: T simply swaps labels without popping the entire label stack and processing the payload below.

An example of the use of B1 is Route Reflectors; an example of the use of B2 is illustrated in Section 9.3.1.2.

However, if T changes the NEXT_HOP attribute for U and in the data plane pops the entire label stack to process the payload, T MUST remove ELA. T MAY include a new Entropy Label attribute ELA’ for UPDATE U’ if both of the following are true:

C1: T sets the NEXT_HOP attribute of U’ to itself; AND

C2: T can process entropy labels for the given application.

Again, if both C1 and C2 are true, and T needs an ELI to recognize entropy labels, then T MUST include the ELI label value as part of the Entropy Label attribute.

5.3. RSVP-TE Signaling

Entropy Label support is signaled in RSVP-TE [RFC3209] using an Entropy Label Attribute TLV (Type TBD) of the LSP_ATTRIBUTES object [RFC5420]. The presence of this attribute indicates that the signaler (the egress in the downstream direction using Resv messages; the ingress in the upstream direction using Path messages) can process entropy labels. The Entropy Label Attribute contains a value for the ELI. If the ELI is zero, this indicates that the signaler doesn’t need an ELI for this application; if not, then the signaler requires the given ELI with entropy labels. An example where an ELI is needed is when the signaled LSP can carry IP traffic.

The format of the Entropy Label Attribute is as follows:
An egress LSR includes the Entropy Label Attribute in a Resv message to indicate that it can process entropy labels in the downstream direction of the signaled LSP.

An ingress LSR includes the Entropy Label Attribute in a Path message for a bi-directional LSP to indicate that it can process entropy labels in the upstream direction of the signaled LSP. If the signaled LSP is not bidirectional, the Entropy Label Attribute SHOULD NOT be included in the Path message, and egress LSR(s) SHOULD ignore the attribute, if any.

As described in Section 8, there is also the need to distribute an ELI from the ingress (upstream label allocation). In the case of RSVP-TE, this is accomplished using the Upstream ELI Attribute TLV of the LSP_ATTRIBUTES object, as shown below:

```
0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-
| Upstream ELI Attribute | Length (4) |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-
| ELI Label | MBZ |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-
```

6. Operations, Administration, and Maintenance (OAM) and Entropy Labels

Generally OAM comprises a set of functions operating in the data plane to allow a network operator to monitor its network infrastructure and to implement mechanisms in order to enhance the general behavior and the level of performance of its network, e.g., the efficient and automatic detection, localization, diagnosis and handling of defects.

Currently defined OAM mechanisms for MPLS include LSP Ping/Traceroute [RFC4379] and Bidirectional Failure Detection (BFD) for MPLS [RFC5884]. The latter provides connectivity verification between the endpoints of an LSP, and recommends establishing a separate BFD session for every path between the endpoints.
The LSP traceroute procedures of [RFC4379] allow an ingress LSR to obtain label ranges that can be used to send packets on every path to the egress LSR. It works by having ingress LSR sequentially ask the transit LSRs along a particular path to a given egress LSR to return a label range such that the inclusion of a label in that range in a packet will cause the replying transit LSR to send that packet out the egress interface for that path. The ingress provides the label range returned by transit LSR N to transit LSR N + 1, which returns a label range which is less than or equal in span to the range provided to it. This process iterates until the penultimate transit LSR replies to the ingress LSR with a label range that is acceptable to it and to all LSRs along path preceding it for forwarding a packet along the path.

However, the LSP traceroute procedures do not specify where in the label stack the value from the label range is to be placed, whether deep packet inspection is allowed and if so, which keys and key values are to be used.

This memo updates LSP traceroute by specifying that the value from the label range is to be placed in the entropy label. Deep packet inspection is thus not necessary, although an LSR may use it, provided it do so consistently, i.e., if the label range to go to a given downstream LSR is computed with deep packet inspection, then the data path should use the same approach and the same keys.

In order to have a BFD session on a given path, a value from the label range for that path should be used as the EL value for BFD packets sent on that path.

As part of the MPLS-TP work, an in-band OAM channel is defined in [RFC5586]. Packets sent in this channel are identified with a reserved label, the Generic Associated Channel Label (GAL) placed at the bottom of the MPLS label stack. In order to use the inband OAM channel with entropy labels, this memo relaxes the restriction that the GAL must be at the bottom of the MPLS label stack. Rather, the GAL is placed in the MPLS label stack above the entropy label so that it effectively functions as an application label.

7. MPLS-TP and Entropy Labels

Since MPLS-TP does not use ECMP, entropy labels are not applicable to an MPLS-TP deployment.
8. Point-to-Multipoint LSPs and Entropy Labels

Point-to-Multipoint (P2MP) LSPs [RFC4875] typically do not use ECMP for load balancing, as the combination of replication and multipathing can lead to duplicate traffic delivery. However, P2MP LSPs can traverse Bundled Links [RFC4201] and LAGs. In both these cases, load balancing is useful, and hence entropy labels can be of some value for P2MP LSPs.

There are two potential complications with the use of entropy labels in the context of P2MP LSPs, both a consequence of the fact that the entire label stack below the P2MP label must be the same for all egress LSRs. First, all egress LSRs must be willing to receive entropy labels; if even one egress LSR is not willing, then entropy labels MUST NOT be used for this P2MP LSP. Second, if an ELI is required, all egress LSRs must agree to the same value of ELI. This can be achieved by upstream allocation of the ELI; in particular, for RSVP-TE P2MP LSPs, the ingress LSR distributes the ELI value using the Upstream ELI Attribute TLV of the LSP_ATTRIBUTES object, defined in Section 5.3.

With regard to the first issue, the ingress LSR MUST keep track of the ability of each egress LSR to process entropy labels, especially since the set of egress LSRs of a given P2MP LSP may change over time. Whenever an existing egress LSR leaves, or a new egress LSR joins the P2MP LSP, the ingress MUST re-evaluate whether or not to include entropy labels for the P2MP LSP.

In some cases, it may be feasible to deploy two P2MP LSPs, one to entropy label capable egress LSRs, and the other to the remaining egress LSRs. However, this requires more state in the network, more bandwidth, and more operational overhead (tracking EL-capable LSRs, and provisioning P2MP LSPs accordingly). Furthermore, this approach may not work for some applications (such as mVPNs and VPLS) which automatically create and/or use P2MP LSPs for their multicast requirements.

9. Entropy Labels and Applications

This section describes the usage of entropy labels in various scenarios with different applications.

9.1. Tunnels

Tunnel LSPs, signaled with either LDP or RSVP-TE, typically carry other MPLS applications such as VPNs or pseudowires. This being the case, if the egress LSR of a tunnel LSP is willing to process entropy
labels, it would signal the need for an Entropy Label Indicator to
distinguish between entropy labels and other application labels.

In the figures below, the following convention is used to depict
information signaled between X and Y:

```
 X ---------- ... ---------- Y
 app:   <-- [label L, ELI value]
```

This means Y signals to X label L for application app. The ELI value
can be one of:

-: meaning entropy labels are NOT accepted;
0: meaning entropy labels are accepted, no ELI is needed; or
E: entropy labels are accepted, ELI label E is required.

The following illustrates a simple intra-AS tunnel LSP.

```
 X -------- A --- ... --- B -------- Y
 tunnel LSP L:   [TL, E] <--- ... <--- [TL0, E]
 IP pkt:           push <TL, E, EL> --------------->
```

Figure 2: Tunnel LSPs and Entropy Labels

Tunnel LSPs may cross Autonomous System (AS) boundaries, usually
using BGP ([RFC3107]). In this case, the AS Border Routers (ASBRs)
MAY simply propagate the egress LSR’s ability to process entropy
labels, or they MAY declare that entropy labels may not be used. If
an ASBR (say A2 below) chooses to propagate the egress LSR Y’s
ability to process entropy labels, A2 MUST also propagate Y’s choice
of ELI.

```
 X ---- ... ---- A1 ------- A2 ---- ... ---- Y
 intra-AS LSP A2-Y:  <-- [TL0, E]
 inter-AS LSP A1-A2:  [AL, E]
 intra-AS LSP X-A1:  <--- [TL1, E]
 IP pkt:           push <TL1, E, EL>
```

Here, ASBR A2 chooses to propagate Y’s ability to process entropy
labels, by "translating" Y’s signaling of entropy label capability
(say using LDP) to BGP; and A1 translate A2’s BGP signaling to (say)
RSVP-TE. The end-to-end tunnel (X to Y) will have entropy labels if
X chooses to insert them.

Figure 3: Inter-AS Tunnel LSP with Entropy Labels

X ---- ... ---- A1 ------- A2 ---- ... ---- Y
intra-AS LSP A2-Y: <<< [TL0, E]
inter-AS LSP A1-A2: [AL, E]
intra-AS LSP X-A1: <<< [TL1, -]

IP pkt: push <TL1> -->

Here, ASBR A1 decided that entropy labels are not to be used; thus, the end-to-end tunnel cannot have entropy labels, even though both X and Y may be capable of inserting and processing entropy labels.

Figure 4: Inter-AS Tunnel LSP with no Entropy Labels

9.2. LDP Pseudowires

[I-D.ietf-pwe3-fat-pw] describes the signaling and use of entropy labels in the context of RFC 4447 pseudowires, so this will not be described further here.

[RFC4762] specifies the use of LDP for signaling VPLS pseudowires. An egress VPLS PE that can process entropy labels can indicate this by adding the Entropy Label sub-TLV in the LDP message it sends to other PEs. An ELI is not required. An ingress PE must maintain state per egress PE as to whether it can process entropy labels.

X -------- A --- ... --- B -------- Y

tunnel LSP L: [TL, E] <<< ... <<< [TL0, E]
VPLS label: <------------------------ [VL, 0]

VPLS pkt: push <TL, VL, EL> -------------->

Figure 5: Entropy Labels with LDP VPLS

Note that although the underlying tunnel LSP signaling indicated the need for an ELI, VPLS packets don’t need an ELI, and thus the label stack pushed by X do not have one.

[RFC4762] also describes the notion of "hierarchical VPLS" (H-VPLS). In H-VPLS, ‘hub PEs’ remove the label stack and process VPLS packets; thus, they must make their own decisions on the use of entropy labels, independent of other hub PEs or spoke PEs with which they exchange signaling. In the example below, spoke PEs X and Y and hub
PE B can process entropy labels, but hub PE A cannot.

X ---- ... ---- A ---- ... ---- B ---- ... ---- Y
spoke PW1:                             <--- [SL1, 0]
hub-hub PW:                     <---- [HL, 0]
spoke PW2:      <--- [SL2, -]
SPW2 pkt:       push <TL1, SL2>
H-H PW pkt:                     push <TL2,HL,EL>
SPW1 pkt:                                       push <TL3,SL1,EL>

Figure 6: Entropy Labels with H-VPLS

9.3. BGP Applications

Section 9.1 described a BGP application for the creation of inter-AS tunnel LSPs. This section describes two other BGP applications, IP VPNs ([RFC4364]) and BGP VPLS ([RFC4761]). An egress PE for either of these applications indicates its ability to process entropy labels by adding the Entropy Label attribute to its BGP UPDATE message. Again, ingress PEs must maintain per-egress PE state regarding its ability to process entropy labels. In this section, both of these applications will be referred to as VPNs.

In the intra-AS case, PEs signal application labels and entropy label capability to each other, either directly, or via Route Reflectors (RRs). If RRs are used, they must not change the BGP NEXT_HOP attribute in the UPDATE messages; furthermore, they can simply pass on the Entropy Label attribute as is.

X -------- A --- ... --- B -------- Y
tunnel LSP L:   [TL, E] <--- ... <--- [TL0, E]
BGP VPN label: <------------------------ [VL, 0]

BGP VPN pkt:    push <TL, VL, EL> -------------->

Figure 7: Entropy Labels with Intra-AS BGP apps

For BGP VPLS, the application label is at the bottom of stack, so no ELI is needed. For BGP IP VPNs, the application label is usually at the bottom of stack, so again no ELI is needed. However, in the case of Carrier's Carrier (CsC) VPNs, the BGP VPN label may not be at the bottom of stack. In this case, an ELI is necessary for CsC VPN packets with entropy labels to distinguish them from nested VPN packets. In the example below, the nested VPN signaling is not shown; the egress PE for the nested VPN (not shown) must signal
whether or not it can process egress labels, and the ingress nested VPN PE may insert an entropy label if so.

Three cases are shown: a plain BGP VPN packet, a CsC VPN packet originating from X, and a transit nested VPN packet originating from a nested VPN ingress PE (conceptually to the left of X). It is assumed that the nested VPN packet arrives at X with label stack <ZL, CVL> where ZL is the tunnel label (to be swapped with <TL, CL>) and CVL is the nested VPN label. Note that Y can use the same ELI for the tunnel LSP and the CsC VPN (and any other application that needs an ELI).

```
X -------- A --- ... --- B -------- Y
```

```
tunnel LSP L:       [TL, E] <--- ... <--- [TL0, E]
BGP VPN label:     <------------------------ [VL, 0]
BGP CsC VPN label: <------------------------ [CL, E]
BGP VPN pkt:       push <TL, VL, EL> ---------->
CsC VPN pkt:       push <TL, CL, E, EL> --------->
nested VPN pkt:    swap <ZL> with <TL, CL> -------->
```

Figure 8: Entropy Labels with CoC VPN

9.3.1. Inter-AS BGP VPNs

There are three commonly used options for inter-AS IP VPNs and BGP VPLS, known informally as "Option A", "Option B" and "Option C". This section describes how entropy labels can be used in these options.

9.3.1.1. Option A Inter-AS VPNs

In option A, an ASBR pops the full label stack of a VPN packet exiting an AS, processes the payload header (IP or Ethernet), and forwards the packet natively (i.e., as IP or Ethernet, but not as MPLS) to the peer ASBR. Thus, entropy label signaling and insertion are completely local to each AS. The inter-AS paths do not use entropy labels, as they do not use a label stack.

9.3.1.2. Option B Inter-AS VPNs

The ASBRs in option B inter-AS VPNs have a choice (usually determined by configuration) of whether to just swap labels (from within the AS to the neighbor AS or vice versa), or to pop the full label stack and process the packet natively. This choice occurs at each ASBR in each direction. In the case of native packet processing at an ASBR, entropy label signaling and insertion is local to each AS and to the...
inter-AS paths (which, unlike option A, do have labeled packets).

In the case of simple label swapping at an ASBR, the ASBR can propagate received entropy label signaling onward. That is, if a PE signals to its ASBR that it can process entropy labels (via an Entropy Label attribute), the ASBR can propagate that attribute to its peer ASBR; if a peer ASBR signals that it can process entropy labels, the ASBR can propagate that to all PEs within its AS). Note that this is the case even though ASBRs change the BGP NEXT_HOP attribute to "self", because of clause B2 in Section 5.2.

9.3.1.3. Option C Inter-AS VPNs

In Option C inter-AS VPNs, the ASBRs are not involved in signaling; they do not have VPN state; they simply swap labels of inter-AS tunnels. Signaling is PE to PE, usually via Route Reflectors; however, if RRs are used, the RRs do not change the BGP NEXT_HOP attribute. Thus, entropy label signaling and insertion are on a PE-pair basis, and the intermediate routers, ASBRs and RRs do not play a role.

9.4. Multiple Applications

It has been mentioned earlier that an ingress PE must keep state per egress PE with regard to its ability to process entropy labels. An ingress PE must also keep state per application, as entropy label processing must be based on the application context in which a packet is received (and of course, the corresponding entropy label signaling).

In the example below, an egress LSR Y signals a tunnel LSP L, and is prepared to receive entropy labels on L, but requires an ELI. Furthermore, Y signals two pseudowires PW1 and PW2 with labels PL1 and PL2, respectively, and indicates that it can receive entropy labels for both pseudowires without the need of an ELI; and finally, Y signals a L3 VPN with label VL, but Y does not indicate that it can receive entropy labels for the L3 VPN. Ingress LSR X chooses to send native IP packets to Y over L with entropy labels, thus X must include the given ELI (yielding a label stack of <TL, ELI, EL>). X chooses to add entropy labels on PW1 packets to Y, with a label stack of <TL, PL1, EL>, but chooses not to do so for PW2 packets. X must not send entropy labels on L3 VPN packets to Y, i.e., the label stack must be <TL, VL>.

10. Security Considerations

This document describes advertisement of the capability to support receipt of entropy-labels and an Entropy Label Indicator that an ingress LSR may apply to MPLS packets in order to allow transit LSRs to attain better load-balancing across LAG and/or ECMP paths in the network.

This document does not introduce new security vulnerabilities to LDP. Please refer to the Security Considerations section of LDP ([RFC5036]) for security mechanisms applicable to LDP.

Given that there is no end-user control over the values used for entropy labels, there is little risk of Entropy Label forgery which could cause uneven load-balancing in the network.

If Entropy Label Capability is not signaled from an egress PE to an ingress PE, due to, for example, malicious configuration activity on the egress PE, then the PE’s will fall back to not using entropy labels for load-balancing traffic over LAG or ECMP paths which, in some cases, is no worse than the behavior observed in current production networks. That said, operators are recommended to monitor changes to PE configurations and, more importantly, the fairness of load distribution over equal-cost LAG or ECMP paths. If the fairness of load distribution over a set of paths changes that could indicate a misconfiguration, bug or other non-optimal behavior on their PE’s and they should take corrective action.

Given that most applications already signal an Application Label, e.g.: IPVPNs, LDP VPLS, BGP VPLS, whose Bottom of Stack bit is being re-used to signal entropy label capability, there is little to no additional risk that traffic could be misdirected into an inappropriate IPVPN VRF or VPLS VSI at the egress PE.
In the context of downstream-signaled entropy labels that require the use of an Entropy Label Indicator (ELI), there should be little to no additional risk because the egress PE is solely responsible for allocating an ELI value and ensuring that ELI label value DOES NOT conflict with other MPLS labels it has previously allocated. On the other hand, for upstream-signaled entropy labels, e.g.: RSVP-TE point-to-point or point-to-multipoint LSP’s or Multicast LDP (mLDP) point-to-multipoint or multipoint-to-multipoint LSP’s, there is a risk that the head-end MPLS LER may choose an ELI value that is already in use by a downstream LSR or LER. In this case, it is the responsibility of the downstream LSR or LER to ensure that it MUST NOT accept signaling for an ELI value that conflicts with MPLS label(s) that are already in use.

11. IANA Considerations

11.1. LDP Entropy Label TLV

IANA is requested to allocate the next available value from the IETF Consensus range in the LDP TLV Type Name Space Registry as the "Entropy Label TLV".

11.2. BGP Entropy Label Attribute

IANA is requested to allocate the next available Path Attribute Type Code from the "BGP Path Attributes" registry as the "BGP Entropy Label Attribute".

11.3. Attribute Flags for LSP_Attributes Object

IANA is requested to allocate a new bit from the "Attribute Flags" sub-registry of the "RSVP TE Parameters" registry.

<table>
<thead>
<tr>
<th>Bit No</th>
<th>Name</th>
<th>Attribute Flags Path</th>
<th>Attribute Flags Resv</th>
<th>RRO</th>
</tr>
</thead>
<tbody>
<tr>
<td>TBD</td>
<td>Entropy Label LSP</td>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
</tr>
</tbody>
</table>

11.4. Attributes TLV for LSP_Attributes Object

IANA is requested to allocate the next available value from the "Attributes TLV" sub-registry of the "RSVP TE Parameters" registry.
12. Acknowledgments

We wish to thank Ulrich Drafz for his contributions, as well as the entire 'hash label' team for their valuable comments and discussion.

13. References

13.1. Normative References


13.2. Informative References


Appendix A. Applicability of LDP Entropy Label sub-TLV

In the case of unlabeled IPv4 (Internet) traffic, the Best Current Practice is for an egress LSR to propagate eBGP learned routes within a SP’s Autonomous System after resetting the BGP next-hop attribute to one of its Loopback IP addresses. That Loopback IP address is injected into the Service Provider’s IGP and, concurrently, a label assigned to it via LDP. Thus, when an ingress LSR is performing a forwarding lookup for a BGP destination it recursively resolves the associated next-hop to a Loopback IP address and associated LDP label of the egress LSR.

Thus, in the context of unlabeled IPv4 traffic, the LDP Entropy Label sub-TLV will typically be applied only to the FEC for the Loopback IP address of the egress LSR and the egress LSR will not announce an entropy label capability for the eBGP learned route.
Authors’ Addresses

Kireeti Kompella
Juniper Networks
1194 N. Mathilda Ave.
Sunnyvale, CA 94089
US

Email: kireeti@juniper.net

John Drake
Juniper Networks
1194 N. Mathilda Ave.
Sunnyvale, CA 94089
US

Email: jdrake@juniper.net

Shane Amante
Level 3 Communications, LLC
1025 Eldorado Blvd
Broomfield, CO 80021
US

Email: shane@level3.net

Wim Henderickx
Alcatel-Lucent
Copernicuslaan 50
2018 Antwerp
Belgium

Email: wim.henderickx@alcatel-lucent.com

Lucy Yong
Huawei USA
1700 Alma Dr. Suite 500
Plano, TX 75075
US

Email: lucyyong@huawei.com
Abstract

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

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

Requirements Language

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

Status of this Memo

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

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

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

# Table of Contents

1. Introduction .............................................. 4  
2. Reference Topologies ................................. 5  
   2.1. Access Topologies with Static Routing ........... 6  
   2.2. Access Topologies with Access IGP ............. 9  
3. LDP DoD Use Cases ................................. 11  
   3.1. Initial Network Setup ......................... 11  
      3.1.1. AN with Static Routing ................. 11  
      3.1.2. AN with Access IGP ................. 13  
   3.2. Service Provisioning and Activation .......... 13  
   3.3. Service Changes and Decommissioning ....... 16  
   3.4. Service Failure ................................ 16  
   3.5. Network Transport Failure .................. 17  
      3.5.1. General Notes ......................... 17  
      3.5.2. AN Node Failure ....................... 17  
      3.5.3. AN/AGN Link Failure .................. 18  
      3.5.4. AGN Node Failure ...................... 19  
      3.5.5. AGN Network-side Reachability Failure .... 19  
4. LDP DoD Procedures ................................. 20  
   4.1. LDP Label Distribution Control and Retention Modes .... 20  
   4.2. IPv6 Support .................................. 21  
   4.3. LDP DoD Session Negotiation ................ 22  
   4.4. Label Request Procedures .................... 22  
      4.4.1. Access LSR/ABR Label Request .......... 22  
      4.4.2. Label Request Retry .................... 23  
      4.4.3. Label Request with Fast-Up Convergence .... 24  
   4.5. Label Withdraw ................................ 26  
   4.6. Label Release ................................ 27  
   4.7. Local Repair ................................ 27  
5. IANA Considerations ................................... 27  
   5.1. LDP TLV TYPE .................................. 28  
6. Security Considerations ........................... 28  
   6.1. Security and LDP DoD ......................... 28  
      6.1.1. Access to network packet flow direction .... 28  
      6.1.2. Network to access packet flow direction .... 29  
   6.2. Data Plane Security .......................... 30  
   6.3. Control Plane Security ....................... 30  
7. Acknowledgements .................................... 31  
8. References ............................................ 31  
   8.1. Normative References ......................... 31  
   8.2. Informative References ....................... 31  
Authors’ Addresses ...................................... 32
1. Introduction

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

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

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

a. A single AN homed to a single AGN.

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

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

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

e. Two ANs dual-homed to two AGNs.

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

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

The amount of MPLS forwarding state however requires additional consideration. In general MPLS routers implement LDP Downstream Unsolicited (LDP DU) label advertisement [RFC5036] and advertise MPLS labels for all valid routes in their RIB. This is seen as a very insufficient approach for ANs, as they only require a small subset of
the total routes (and associated labels) based on the required
connectivity for the provisioned services. And although filters can
be applied to those LDP DU labels advertisements, it is not seen as a
suitable tool to facilitate any-to-any AN-driven connectivity between
access and the rest of the MPLS network.

This document describes an access node driven "subscription model"
for label distribution in the access. The approach relies on the
standard LDP Downstream-on-Demand (LDP DoD) label advertisements as
specified in [RFC5036]. LDP DoD enables on-demand label distribution
ensuring that only required labels are requested, provided and
installed.

Note that LDP DoD implementation is not widely available in today’s
IP/MPLS devices despite the fact that it has been described in the
LDP specification [RFC5036]. This is due to the fact that the
originally LDP DoD advertisement mode was aimed mainly at ATM and
Frame Relay MPLS implementations, where conserving label space used
on the links was essential for compatibility with ATM and Frame Relay
LSRs.

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

2. Reference Topologies

LDP DoD use cases are described in the context of a generic reference
end-to-end network topology based on Seamless MPLS design
[I-D.ietf-mpls-seamless-mpls] shown in Figure 1
The access network is either single or dual homed to AGN1x, with either a single or multiple parallel links to AGN1x.

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

2.1. Access Topologies with Static Routing

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

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

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

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

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

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

The reference static routing and LDP configuration for [V] access topology is shown in Figure 2. The same static routing and LDP configuration also applies to [I1] topology.
In line with the Seamless MPLS design, static routes configured on AGN1x and pointing towards the access network are redistributed in either ISIS or BGP labeled unicast (BGP-LU) [RFC3107].

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

The reference static routing and LDP configuration for [Y] access topology is shown in Figure 4. The same static routing and LDP configuration also applies to [I] topology.
ANs support Inter-area LDP [RFC5283] in order to use the IP default route to match the LDP FEC advertised by AGN1x and other ANs.

2.2. Access Topologies with Access IGP

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

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

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

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

The reference access IGP and LDP configuration for [U] access
topology is shown in Figure 5.

```
+-------+   +-----+   +----+     |       +---/
| AN3 |---| AN2 |---|AN1 +-----+ AGN11 |
+-----+   +-----+   +----+     |       +---
   .
   .
+-----+   +-----+   +----+     |       +---/
|ANn-2|---|ANn-1|---|ANn +-----+ AGN12 |
+-----+   +-----+   +----+     |       +---
   +-------+

<-------- access IGP ------------> <--- ISIS --->
<-- LDP DU -->
<-------- LDP DoD --------------> <--- BGP LU -->
```

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

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

```
+-------+   +-----+   +----+     |       +---/
|       |---/ AGN11 |2
+-----+   +-----+   +----+     |       +---
|ANn +...|AN2 +---|AN1 +-----/ +-----+ AGN12 |
+-----+   +-----+   +----+     |       +---
   \-----+ AGN12 |
   +-------+

<-------- access IGP ------------> <--- ISIS --->
<-- LDP DU -->
<-------- LDP DoD --------------> <--- BGP LU -->
```

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

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

In both of the above topologies, ANs (ANn ... AN1) and AGN1x share the access IGP and advertise their IPv4 and IPv6 loopbacks and link addresses. AGN1x advertise a default route into the access IGP.
ANs support Inter-area LDP [RFC5283] in order to use the IP default route for matching the LDP FECs advertised by AGN1x or other ANs.

3. LDP DoD Use Cases

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

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

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

This document is focusing on IPv4 LDP DoD procedures. Similar procedures are required for IPv6 LDP DoD, however some extension specific to IPv6 are likely to apply including LSP mapping, peer discovery, transport connection establishment. These will be added in this document once LDP IPv6 standardization is advanced as per [I-D.ietf-mpls-ldp-ipv6].

3.1. Initial Network Setup

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

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

3.1.1. AN with Static Routing

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

a. [I1, V, U2] - Static default route 0/0 pointing to links connected to AGN1x. Requires support for Inter-area LDP [RFC5283].
b. [U2] - Static /32 or /128 routes pointing to the other AN. Lower preference static default route 0/0 pointing to links connected to the other AN. Requires support for Inter-area LDP [RFC5283].

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

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

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

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

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

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

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

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

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

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

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

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

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

3.2. Service Provisioning and Activation

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

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

MPLS PWE3 service - for description simplicity it is assumed that a single segment pseudowire is signaled using targeted LDP FEC128 (0x80), and it is provisioned with the pseudowire ID and the loopback IPv4 address of the destination node. The following IP/MPLS
operations need to be completed on the AN* to successfully establish such PWE3 service:

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

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

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

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

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

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

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

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

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

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

e. Upon successful completion of above operations, AN* programs its RIB/LIB and LFIB tables, and activates the BGP/MPLS IPVPN service.
Note - only minimum operations applicable to service connectivity have been listed. Other non IP/MPLS connectivity operations that may be required for successful service provisioning are out of scope in this document.

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

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

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

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

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

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

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

3.3. Service Changes and Decommissioning

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

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

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

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

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

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

3.4. Service Failure

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

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

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

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

3.5.1. General Notes

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

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

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

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

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

3.5.2. AN Node Failure

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

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

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

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

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

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

3.5.3. AN/AGN Link Failure

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

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

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

c. [U2, U, V, Y] - link failed, but at least one ECMP or alternate path remains - AN/AGN1x node must stop using the failed link and immediately switchover (local-repair) to the remaining ECMP path or alternate path. AN/AGN1x must remove affected next-hops and
labels from its tables and invoke LDP label withdraw as per point (a) above. If there is an AGN1x node terminating the failed link, it must remove routes pointing to the failed link immediately from the RIB, remove associated labels from their LIB and LFIB tables, and must propagate the failure on the network side using BGP-LU and/or core IGP procedures.

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

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

3.5.4. AGN Node Failure

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

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

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

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

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

3.5.5. AGN Network-side Reachability Failure

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

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

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

4. LDP DoD Procedures

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

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

```
+----------+      +------------+
| upstream |      | downstream |
------|   LSR    |      | LSR      |----
traffic |          |      |            |  address
source   +----------+      +------------+  (/32 for IPv4)
label distribution for IPv4 FEC destination
<-------------------------
                     traffic flow
                     ------------------------->
```

Figure 7: LDP label assignment direction

4.1. LDP Label Distribution Control and Retention Modes

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

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

- Ordered mode - an LSR binds a label to a particular FEC if it is the egress router for that FEC or if it has already received a label binding for that FEC from its next-hop LSR for that FEC.
Using independent label distribution control with LDP DoD and access static routing will prevent the access LSRs from propagating label binding failure along the access topology, making it impossible to switchover to an alternate path, even if such a path exists.

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

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

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

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

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

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

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

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

### 4.2. IPv6 Support

Current LDP protocol specification [RFC5036] defines procedures and messages for exchanging FEC-label bindings over IPv4 and/or IPv6 networks. However number of IPv6 usage areas are not clearly...
specified including: packet to LSP mapping for IPv6 destination router, no IPv6 specific LSP identifier, no LDP discovery using IPv6 multicast address, separate LSPs for IPv4 and IPv6, and others.

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

4.3. LDP DoD Session Negotiation

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

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

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

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

4.4. Label Request Procedures

4.4.1. Access LSR/ABR Label Request

Upstream access LSR/ABR will request label bindings from adjacent downstream access LSR/ABR based on the following trigger events:
a. Access LSR/ABR is configured with /32 static route with LDP DoD label request policy in line with initial network setup use case described in Section 3.1.

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

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

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

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

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

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

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

4.4.2. Label Request Retry

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

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

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

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

4.4.3. Label Request with Fast-Up Convergence

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

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

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

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

<table>
<thead>
<tr>
<th>Optional Parameter</th>
<th>Length</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Queue Request TLV</td>
<td>0</td>
<td>see below</td>
</tr>
</tbody>
</table>
U-bit = 1
Unknown TLV bit is set to 1. If this optional TLV is unknown, it should be ignored without sending "no route" notification. Ensures backward compatibility.

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

Type
Queue Request Type value to be allocated by IANA.

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

The operation is as follows.

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

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

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

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

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

4.5. Label Withdraw

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

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

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

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

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

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

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

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

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

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

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

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

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

4.6. Label Release

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

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

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

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

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

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

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

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

4.7. Local Repair

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

5. IANA Considerations
5.1. LDP TLV TYPE

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

<table>
<thead>
<tr>
<th>TLV type</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>0x0971</td>
<td>Queue Request TLV</td>
</tr>
</tbody>
</table>

6. Security Considerations

MPLS LDP Downstream on Demand deployment in the access network is subject to similar security threats as any MPLS LDP deployment. It is recommended that baseline security measures are considered as described in the LDP specification [RFC5036] including ensuring authenticity and integrity of LDP messages, as well as protection against spoofing and Denial of Service attacks.

Some deployments may require increased measures of network security if a subset of Access Nodes are placed in locations with lower levels of physical security e.g. street cabinets (common practice for VDSL access). In such cases it is the responsibility of the system designer to take into account the physical security measures (environmental design, mechanical or electronic access control, intrusion detection), as well as monitoring and auditing measures (configuration and Operating System changes, reloads, routes advertisements).

But even with all this in mind, the designer still should consider network security risks and adequate measures arising from the lower level of physical security of those locations.

6.1. Security and LDP DoD

6.1.1. Access to network packet flow direction

An important property of MPLS LDP Downstream on Demand operation is that the upstream LSR (requesting LSR) accepts only mappings it sent a request for (in other words the ones it is interested in), and does not accept any unsolicited label mappings by design.

This limits the potential of an unauthorized third party fiddling with label mappings operations on the wire. It also enables ABR LSR to monitor behaviour of any Access LSR in case the latter gets compromised and attempts to get access to an unauthorized FEC or remote LSR. Note that ABR LSR is effectively acting as a gateway to
the MPLS network, and any label mapping requests made by any Access LSR are processed and can be monitored on this ABR LSR.

6.1.2. Network to access packet flow direction

Another important property of MPLS LDP DoD operation in the access is that the number of access nodes and associated MPLS FECs per ABR LSR is not large in number, and they are all known at the deployment time. Hence any changes of the access MPLS FECs can be easily controlled and monitored on the ABR LSR.

And then, even in the event when Access LSR manages to advertise a FEC that belongs to another LSR (e.g. in order to ‘steal’ third party data flows, or breach a privacy of VPN), such Access LSR will have to influence the routing decision for affected FEC on the ABR LSR.

Following measures SHOULD be considered to prevent such event from occurring:

a. ABR LSR - access side with static routes - this is not possible for Access LSR. Access LSR has no way to influence ABR LSR routing decisions due to static nature of routing configuration here.

b. ABR LSR - access side with IGP - this is still not possible if the compromised Access LSR is a leaf in the access topology (leaf node in topologies I1, I, V, Y described earlier in this document), due to the leaf metrics being configured on the ABR LSR. If the compromised Access LSR is a transit LSR in the access topology (transit node in topologies I, Y, U), it is possible for this Access LSR to attract to itself traffic destined to the nodes upstream from it. However elaborate such ‘man in the middle attack’ is possible, but can be quickly detected by upstream Access LSRs not receiving traffic, and legitimate traffic from them getting dropped.

c. ABR LSR - network side - designer SHOULD consider giving a higher administrative preference to the labeled unicast BGP routes vs. access IGP routes.

In summary MPLS in access design with LDP DoD has number of native properties that prevent number of security attacks and make their detection quick and straightforward.

Following two sections describe other security considerations applicable to general MPLS deployments in the access.
6.2. Data Plane Security

Data plane security risks applicable to the access MPLS network are listed below (a non-exhaustive list):

a. packets from a specific access node flow to an altered transport layer or service layer destination.

b. packets belonging to undefined services flow to and from the access network.

c. unlabelled packets destined to remote network nodes.

Following mechanisms should be considered to address listed data plane security risks:

1. addressing (a) - Access and ABR LSRs SHOULD NOT accept labeled packets over a particular data link, unless from the Access or ABR LSR perspective this data link is known to attach to a trusted system based on employed authentication mechanism(s), and the top label has been distributed to the upstream neighbour by the receiving Access or ABR LSR.

2. addressing (a) - ABR LSR MAY restrict network reachability for access devices to a subset of remote network LSR, based on authentication or other network security technologies employed towards Access LSRs. Restricted reachability can be enforced on the ABR LSR using local routing policies, and can be distributed towards the core MPLS network using routing policies associated with access MPLS FECs.

3. addressing (b) - labeled service routes (e.g. MPLS/VPN, tLDP) are not accepted from unreliable routing peers. Detection of unreliable routing peers is achieved by engaging routing protocol detection and alarm mechanisms, and is out of scope of this document.

4. addressing (a) and (b) - no successful attacks have been mounted on the control plane and has been detected.

5. addressing (c) - ABR LSR MAY restrict IP network reachability to and from the access LSR.

6.3. Control Plane Security

Similarly to Inter-AS MPLS/VPN deployments [RFC4364], the data plane security depends on the security of the control plane.
To ensure control plane security access LDP DoD connections MUST only be made with LDP peers that are considered trusted from the local LSR perspective, meaning they are reachable over a data link that is known to attach to a trusted system based on employed authentication mechanism(s) on the local LSR.

The TCP/IP MD5 authentication option [RFC5925] should be used with LDP as described in LDP specification [RFC5036]. If TCP/IP MD5 authentication is considered not secure enough, the designer may consider using a more elaborate and advanced TCP Authentication Option (TCP-AO RFC 5925) for LDP session authentication.

Access IGP (if used) and any routing protocols used in access network for signalling service routes SHOULD also be secured in a similar manner.

For increased level of authentication in the control plane security for a subset of access locations with lower physical security, designer could also consider using:

- different crypto keys for use in authentication procedures for these locations.
- stricter network protection mechanisms including DoS protection, interface and session flap dampening.

7. Acknowledgements

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

8. References

8.1. Normative References


8.2. Informative References

(work in progress), January 2012.

[I-D.ietf-mpls-seamless-mpls]


Authors’ Addresses

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

Phone: +49 6151 58 12825
Fax:
Email: thomas.beckhaus@telekom.de
URI:
Updates to LDP for IPv6
draft-ietf-mpls-ldp-ipv6-06

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 July 23, 2012.

Copyright Notice

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

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

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

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

However, RFC5036 specification has the following deficiencies in regards to IPv6 usage:

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

2) LDP Identifier: No details specific to IPv6 usage

3) LDP Discovery: No details for using a particular IPv6 destination (multicast) address or the source address (with or without IPv4 co-existence)

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

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

6) Next Hop Address & LDP Identifier: No rule for accommodating the usage of duplicate link-local IPv6 addresses

7) LDP TTL Security: No rule for built-in Generalized TTL Security Mechanism (GTSM) in LDP

This document addresses the above deficiencies by specifying the desired behavior/rules/details for using LDP in IPv6 enabled networks. It also clarifies the scope (section 1.1).

Note that this document updates RFC5036.
1.1. Scope

1.1.1. Topology Scenarios

The following scenarios in which the LSRs may be inter-connected via one or more dual-stack interfaces (figure 1), or two or more single-stack interfaces (figure 2 and figure 3) are addressed by this document:

R1------------------R2
IPv4+IPv6

Figure 1 LSRs connected via a Dual-stack Interface

IPv4
R1------------------R2
IPv6

Figure 2 LSRs connected via two single-stack Interfaces

IPv4
R1------------------R2------------------R3
IPv4                 IPv6

Figure 3 LSRs connected via a single-stack Interface

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

Note that the topology scenario illustrated in figure 2 also covers the case of two routers getting connected via an additional single-stack interface (IPv6, say), even though the IPv4 LDP session may already be established between the LSRs over the existing interface.
1.1.2. LDP TTL Security

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

As a consequence, any LDP feature that relies on multi-hop LDP peering session would not work with GTSM and will warrant (statically or dynamically) disabling GTSM. Please see section 8.

2. Specification Language

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

Abbreviations:

LDP      - Label Distribution Protocol
LDPv4    - LDP for enabling IPv4 MPLS forwarding
LDPv6    - LDP for enabling IPv6 MPLS forwarding
LDPoIPv4 - LDP over IPv4 transport session
LDPoIPv6 - LDP over IPv6 transport session
FEC      - Forwarding Equivalence Class
TLV      - Type Length Value
LSR      - Label Switch Router
LSP      - Label Switched Path
LSPv4    - IPv4-signaled Label Switched Path [RFC4798]
LSPv6    - IPv6-signaled Label Switched Path [RFC4798]
3. LSP Mapping

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

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

Suffice to say, this rule is correct for IPv4, but not for IPv6, since an IPv6 router may not have any /32 address.

This document proposes to modify this rule by also including a /128 address (for IPv6). In fact, it should be reasonable to just say IPv4 or IPv6 address instead of /32 or /128 addresses as shown below in the updated rule:

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

Additionally, it is desirable that a packet is forwarded to an LSP of an egress router, only if LSP’s address-family (e.g. LSPv4 or LSPv6) matches with that of the LDP hello adjacency on the next-hop interface.

4. LDP Identifiers

Section 2.2.2 of [RFC5036] specifies formulating at least one LDP Identifier, however, it doesn’t provide any consideration in case of IPv6 (with or without dual-stacking). Additionally, section 2.5.2 of [RFC5036] implicitly prohibits using the same label space for both IPv4 and IPv6 FEC-label bindings.

The first four octets of the LDP identifier, the 32-bit LSR Id, identify the LSR and is a globally unique value. This is regardless of the address family used for the LDP session. Hence, this document preserves the usage of 32-bit LSR Id on an IPv6 only LSR.
Please note that 32-bit LSR Id value would not map to any IPv4-address in an IPv6 only LSR (i.e., single stack), nor would there be an expectation of it being DNS-resolvable. In IPv4 deployments, the LSR Id is typically derived from an IPv4 address, generally assigned to a loopback interface. In IPv6 only deployments, this 32-bit LSR Id must be derived by some other means that guarantees global uniqueness.

This document qualifies the first sentence of last paragraph of Section 2.5.2 of [RFC5036] to be per address family and therefore updates that sentence to the following: "For a given address family over which a Hello is sent, and a given label space, an LSR MUST advertise the same transport address." This rightly enables the per-platform label space to be shared between IPv4 and IPv6.

In summary, this document not only allows the usage of a common LDP identifier i.e. same LSR-Id, but also the common Label space id for both IPv4 and IPv6 on a dual-stack LSR.

This document reserves 0.0.0.0 as the LSR-Id, and prohibits its usage.

5. Peer Discovery

5.1. Basic Discovery Mechanism

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

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

- FF01:0:0:0:0:0:0:2 = Interface-local scope
- FF02:0:0:0:0:0:0:2 = Link-local scope
- FF05:0:0:0:0:0:0:2 = Site-local scope

[RFC5036] does not specify which particular IPv6 ‘all routers on this subnet’ group multicast IP address should be used by LDP Link Hellos.
This document specifies the usage of link-local scope e.g. FF02:0:0:0:0:0:0:2 as the destination multicast IP address in IPv6 LDP Link Hellos. An LDP Hello packet received on any of the other destination addresses must be dropped. Additionally, the link-local IPv6 address MUST be used as the source IP address in IPv6 LDP Link Hellos.

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

More importantly, if an interface is a dual-stack LDP interface (e.g. enabled with both IPv4 and IPv6 LDP), then the LSR must periodically send both IPv4 and IPv6 LDP Link Hellos (using the same LDP Identifier per section 4) and must separately maintain the Hello adjacency for IPv4 and IPv6 on that interface.

In summary, the IPv4 and IPv6 LDP Link Hellos must carry the same LDP identifier (assuming per-platform label space usage).

5.2. Extended Discovery Mechanism

Suffice to say, the extended discovery mechanism (defined in section 2.4.2 of [RFC5036]) doesn’t require any additional IPv6 specific consideration, since the targeted LDP Hellos are sent to a pre-configured (unicast) destination IPv6 address.

The link-local IP addresses MUST NOT be used as the source or destination IPv6 addresses in extended discovery.

6. LDP Session Establishment and Maintenance

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

1. Transport connection establishment
2. Session initialization
The forthcoming sub-sections discuss the LDP consideration for IPv6 and/or dual-stacking in the context of session establishment and maintenance.

6.1. Transport connection establishment

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

This document specifies that:

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

2. An LSR SHOULD accept the Hello message that contains both IPv4 and IPv6 transport address optional objects, but MUST use only the transport address whose address family is the same as that of the IP packet carrying Hello.

3. An LSR MUST send separate Hellos (each containing either IPv4 or IPv6 transport address optional object) for each IP address-family, if LDP was enabled for both IP address-families.

4. An LSR MUST use a global unicast IPv6 address in IPv6 transport address optional object of outgoing targeted hellos, and check for the same in incoming targeted hellos.

5. An LSR MUST prefer using global unicast IPv6 address for an LDP session with a remote LSR, if it had to choose between global unicast IPv6 address and link-local IPv6 address (pertaining to the same LDP Identifier) for the transport connection.

6. An LSR SHOULD NOT create (or honor the request for creating) a TCP connection for a new LDP session with a remote LSR, if they
already have an LDP session (for the same LDP Identifier) established over whatever IP version transport.

This means that only one transport connection is established, even if there are two Hello adjacencies (one for IPv4 and another for IPv6). This is independent of whether the Hello Adjacencies are created over a single interface (scenarios 1 in section 1.1) or multiple interfaces (scenario 2 in section 1.1) between two LSRs.

7. An LSR SHOULD prefer the LDP/TCP connection over IPv6 for a new LDP session with a remote LSR, if it has both IPv4 and IPv6 hello adjacencies for the same LDP Identifier (over a dual-stack interface, or two or more single-stack IPv4 and IPv6 interfaces). This applies to the section 2.5.2 of RFC5036.

8. An LSR SHOULD prefer the LDP/TCP connection over IPv6 for a new LDP session with a remote LSR, if they attempted two TCP connections using IPv4 and IPv6 transport addresses simultaneously.

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

6.2. Maintaining Hello Adjacencies

As outlined in section 2.5.5 of RFC5036, this draft describes that if an LSR has a dual-stack interface, which is enabled with both IPv4 and IPv6 LDP, then the LSR must periodically send both IPv4 and IPv6 LDP Link Hellos and must separately maintain the Hello adjacency for IPv4 and IPv6 on that interface.

This ensures successful labeled IPv4 and labeled IPv6 traffic forwarding on a dual-stacked interface, as well as successful LDP peering using the appropriate transport on a multi-access interface (even if there are IPv4-only, IPv6-only and dual-stack LSRs connected to that multi-access interface).
6.3. Maintaining LDP Sessions

Two LSRs maintain a single LDP session between them, as described in section 6.1, whether they are connected via a dual-stack LDP enabled interface or via two single-stack LDP enabled interfaces. This is also true when a single-stack interface is converted to a dual-stack interface (e.g. figure 1), or when another interface is added between two LSRs (e.g. figure 2).

Needless to say that the procedures defined in section 6.1 would always result in preferring LDPoIPv6 session after the loss of an existing LDP session (because of link failure, node failure, reboot etc.).

On the other hand, if a dual-stack interface is converted to a single-stack interface (by disabling IPv4 or IPv6 routing), then the LDP session should be torn down ONLY if the disabled IP version was the same as that of the transport connection. Otherwise, the LDP session should stay intact.

If the LDP session is torn down for whatever reason (LDP disabled for the corresponding transport, hello adjacency expiry etc.), then the LSRs should initiate establishing a new LDP session as per the procedures described in section 6.1 of this document along with RFC5036.

7. Label Distribution

An LSR MAY NOT advertise both IPv4 and IPv6 FEC-label bindings (as well as interface addresses via ADDRESS message) from/to the peer over an LDP session (using whatever transport), unless it has valid IPv4 and IPv6 Hello Adjacencies for that peer, as specified in section 6.2.

Another solution for getting the same result as above is by negotiating the IP Capability for a given AFI, as specified in [IPPWCap].

An LSR MUST NOT allocate and advertise FEC-Label bindings for link-local IPv6 address, and ignore such bindings, if ever received. An LSR MUST treat the IPv4-mapped IPv6 address, defined in section 2.5.1 of [RFC4291], the same as that of a global IPv6 address and not mix it with the ‘corresponding’ IPv4 address.
Additionally, to ensure backward compatibility (and interoperability with IPv4-only LDP implementations), this document specifies that -

1. An LSR MUST NOT send a label mapping message with a FEC TLV containing FEC Elements of different address-family. In other words, a FEC TLV in the label mapping message MUST contain the FEC Elements belonging to the same address-family.

2. An LSR MUST NOT send an Address message (or Address Withdraw message) with an Address List TLV containing IP addresses of different address-family. In other words, an Address List TLV in the Address (or Address Withdraw) message MUST contain the addresses belonging to the same address-family.

8. LDP Identifiers and Next Hop Addresses

RFC5036 section 2.7 specifies logic for mapping between a peer LDP Identifier and the peer’s addresses to find the correct LIB entry for any prefix by using a database populated by the Address message. However, this logic is insufficient to deal with overlapping IPv6 (link-local) addresses used by two or more peers. One may note that all interior IP routing protocols specify using link-local IPv6 addresses as the next-hops.

This document specifies that the logic is enhanced with the usage of (Hello Adjacency) database populated by the Hello messages. This additional database lookup is useful only if/when two or more peers use the same link-local IPv6 address as the IP routing next-hops (causing duplicate next-hop entries).

Specifically, this document specifies that an LSR should (continue to) use the machinery described in RFC5036 section 2.7 to map between a peer LDP Identifier and the peer’s addresses (learned via ADDRESS message) for any prefix. However, if this mapping fails (for reasons such as the one described earlier), then an LSR can find the peer LDP Identifier by checking for the particular link-local IPv6 address in the hello adjacency database.

If an LSR can’t find such a mapping in either database, then LSR should follow procedures specified in RFC5036 (e.g. not resolve the label).

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

This document also specifies that the LDP/TCP transport connection over IPv6 (i.e. LDPoIPv6) must follow the Generalized TTL Security Mechanism (GTSM) procedures (Section 3 of [RFC5082]) for an LDP session peering established between the adjacent LSRs using Basic Discovery, by default.

In other words, GTSM is enabled by default for an IPv6 LDP peering session using Basic Discovery. This means that the ‘IP Hop Limit’ in IPv6 packet is set to 255 upon sending, and checked to be 255 upon receipt. The IPv6 packet must be dropped failing such a check upon receipt.

The reason GTSM is enabled for Basic Discovery by default, but not for Extended Discovery is that the usage of Basic Discovery typically results in a single-hop LDP peering session, whereas the usage of Extended Discovery typically results in a multi-hop LDP peering session. While the latter is deemed out of scope (section 1.2), in line with GTSM [RFC5082], it is worth clarifying the following exceptions that may occur with Basic or Extended Discovery usage:

a) Two adjacent LSRs (i.e. back-to-back PE routers) forming a single-hop LDP peering session after doing an Extended Discovery (for Pseudowire, say)

b) Two adjacent LSRs forming a multi-hop LDP peering session after doing a Basic Discovery, due to the way IP routing changes between them (temporarily (e.g. session protection) or permanently)

c) Two adjacent LSRs (i.e. back-to-back PE routers) forming a single-hop LDP peering session after doing both Basic and Extended Discovery

In (a), GTSM is not enabled for the LDP peering session by default, hence, it would not do any harm or good.

In (b), GTSM is enabled by default for the LDP peering session by default and enforced, hence, it would prohibit the LDP peering session from getting established.

In (c), GTSM is enabled by default for Basic Discovery and enforced on the subsequent LDP peering. However, if each LSR uses the same IPv6 transport address object value in both Basic and Extended discoveries, then it would result in a single LDP peering session

and that would be enabled with GTSM. Otherwise, GTSM would not be enforced on the 2nd LDP peering session corresponding to the Extended Discovery.

This document allows for the implementation to provide an option to statically (configuration) and/or dynamically override the default behavior (enable/disable GTSM) on a per-peer basis. This would also address the exception (b) above. Suffice to say that such an option could be set on either LSR (since GTSM negotiation would ultimately disable GTSM between LSR and its peer(s)).

The built-in GTSM inclusion is intended to automatically protect IPv6 LDP peering session from off-link attacks.

10. IANA Considerations

None.

11. Security Considerations

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

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

Moreover, this document allows the use of IPsec [RFC4301] for IPv6 protection, hence, LDP can benefit from the additional security as specified in [RFC4835] as well as [RFC5920].

12. Acknowledgments

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

Thanks to Bob Thomas for providing critical feedback to improve this document early on. Thanks to Eric Rosen, Lizhong Jin, Bin Mo, Mach
Chen, and Kishore Tiruveedhula for reviewing this document. The authors also acknowledge the help of Manoj Dutta and Vividh Siddha.

Also, thanks to Andre Pelletier (who brought up the issue about active/passive determination, and helped us craft the appropriate solutions.

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

13. Additional Contributors

The following individuals contributed to this document:

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

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

14.1. Normative References


14.2. Informative References


Author’s Addresses

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

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

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

Carlos Pignataro
Cisco Systems, Inc.
7200 Kit Creek Road
Research Triangle Park, NC 27709-4987
Email: cpignata@cisco.com
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 describes the LDP protocol extensions required to support MT routing in an MPLS environment.

Copyright Notice

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

1. Terminology ............................................. 4  
2. Introduction .............................................. 4  
3. Requirements ............................................... 4  
4. Signaling Extensions ....................................... 5  
   4.1. Topology-Scoped FEC .................................... 5  
   4.2. New Address Families: MT IP .......................... 5  
   4.3. LDP FEC Elements with MT IP AF ........................ 7  
   4.4. IGP MT-ID Mapping and Translation ....................... 8  
   4.5. LDP MT Capability Advertisement ....................... 8  
   4.6. Procedures .............................................. 9  
   4.7. LDP Sessions ............................................. 10  
   4.8. Reserved MT ID Values ................................... 11  
5. MT Applicability on FEC-based features ..................... 11  
   5.1. Typed Wildcard FEC Element ............................ 11  
   5.2. End-of-LIB .............................................. 12  
6. Error Handling ............................................. 12  
   6.1. MT Error Notifications .................................. 12  
   6.2. MT Advisory Notifications .............................. 12  
7. Backwards Compatibility ................................... 12  
8. MPLS Forwarding in MT ..................................... 13  
9. Security Consideration ..................................... 13  
10. IANA Considerations ....................................... 13  
11. Acknowledgement .......................................... 14  
12. References ................................................. 14  
   12.1. Normative References ................................... 14  
   12.2. Informative References ................................ 15  
Appendix A. Appendix ........................................ 15  
   A.1. Application Scenarios ................................... 15  
      A.1.1. Simplified Data-plane .............................. 15  
      A.1.2. Using MT for P2P Protection ....................... 15  
      A.1.3. Using MT for mLDP Protection ..................... 16  
      A.1.4. Service Separation .................................. 16  
      A.1.5. An Alternative inter-AS VPN Solution ............ 16  
Authors’ Addresses ............................................. 16
1. Terminology

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

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

It would be advantageous for Communications Service Providers (CSP) to support Multiple Topologies (MT) within MPLS environments (MPLS-MT). Beneficial MPLS-MT deployment scenarios include:

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

3. 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.

4. Signaling Extensions

4.1. Topology-Scoped FEC

LDP assigns and binds a label to a FEC, where a FEC is a list of one of 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 info. 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 (4.2 New Address Families: MT IP) defines the extension required to bind "prefix-related" FEC to a topology.

4.2. New Address Families: MT IP

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

```
0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|  IP Address                                 |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|       Reserved              |        MT-ID                  |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
```

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.

4.3. LDP FEC Elements with MT IP AF

When introducing the new Address Family, it will make the extension to all the prefix-related FEC Elements by nature. This 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 will be encoded as follows:

```
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
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|  Prefix (2)   | Address Family (MT IP/MT IPv6)| PreLen |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
   Prefix
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
| Reserved             |        MT-ID                  |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+

Figure 3: MT Prefix FEC Element Format
```

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

```
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
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
| mLDP FEC Type | Address Family (MT IP/MT IPv6) | Address Length |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
   Root Node Address
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
| Reserved             |        MT-ID                  |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
| Opaque Length              | Opaque Value ...           |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+

Figure 4: MT mLDP FEC Element Format
And the MT Typed Wildcard FEC element encoding is as follows:

<table>
<thead>
<tr>
<th>0</th>
<th>1</th>
<th>2</th>
<th>3</th>
</tr>
</thead>
<tbody>
<tr>
<td>0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
| +----------------------------------+
| Typed Wcard (5) | FEC Type | Len = 6 | AF = MT IP .. |
| +----------------------------------+
| ... or MT IPv6 | MT ID |
| +----------------------------------+

Figure 5: MT Typed Wildcard FEC Element

4.4. 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.

4.5. 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:
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
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|U|F| Multi-Topology Cap.(IANA) |            Length             |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|S| Reserved    |                                               |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
˜                Typed Wildcard FEC element(s)                  ˜
|                                                               |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+

Figure 6: Multi-Topology Capability TLV Format

Where:

- U- and F-bits: MUST be 1 and 0, respectively, as per Section 3 of LDP Capabilities [RFC5561].
- Multi-Topology Capability: Capability TLV type (IANA assigned)
- 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.
- Typed Wildcard FEC element(s): One or more elements specified as the "Capability data".
- Length: The length (in octets) of TLV.

The encoding of Typed Wcard FEC element, as defined in [RFC5561], is defined in the section 4.3 of this document.

4.6. Procedures

To announce its MT capability for given IP address family, given LDP FEC type, and given Multi Topology, an LDP speaker MAY send "MT Capability" including the exact Typed Wildcard FEC element with corresponding "Address Family" 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.

- The capability for supporting multi-topology in LDP can be advertised during LDP session initialization stage by including the LDP MT capability TLV in LDP Initialization message. After
LDP session is established, the MT capability can also be advertised or withdrawn using Capability message (only if "Dynamic Announcement" capability [RFC5561] has already been successfully negotiated).

- If an LSR has not advertised MT capability, its peer must not send messages that include MT identifier to this LSR.

- If an LSR receives a Label Mapping message with MT parameter from downstream LSR-D and its upstream LSR-U has not advertised MT capability, an LSP for the MT will not be established.

- We propose to add a new notification event to signal the upstream that the downstream is not capable.

- If an LSR is changed from non-MT capable to MT capable, it sets the S bit in MT capability TLV and advertises via the Capability message. The existing LSP is treated as LSP for default MT (ID 0).

- If an LSR is changed from 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.

- 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.

- There will be case where IGP is MT capable but MPLS is not and the handling procedure for this case is TBD.

4.7. LDP Sessions

Depending on the number of label spaces supported, if a single global label space is supported, there will be one session supported for each pair of peer, even there are multiple topologies supported between these two peers. If there are different label spaces supported for different topologies, which means that label spaces overlap with each other for different MTs, then it is suggested to establish multiple sessions for multiple topologies between these two peers. In this case, multiple LSR-IDs need to be allocated beforehand so that each multiple topology can have its own label space ID.

[Editors Note - This section requires further discussion]
4.8. Reserved MT ID Values

Certain MT topologies are assigned to serve pre-determined 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.


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

We propose a new IANA registry "LDP Multi-Topology ID Name Space" under IANA "LDP Parameter" namespace to keep 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.

[Editors Note - This section requires further discussion].

5. MT Applicability on FEC-based features

5.1. 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 4.3 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 "Wildcard Topology" MT-ID as defined in section 4.8. For instance, upon local un-configuration of topology "x", an LSR may send wildcard label withdraw with MT-ID "x" to withdraw all its labels from peer that were advertised under the scope of topology "x". On the other hand, upon some global configuration change, an LSR may send wildcard
label withdraw with MT-ID set to "Wildcard Topology" to withdraw all its labels under all topologies from the peer.

5.2. End-of-LIB

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

6. Error Handling

The extensions defined in this document utilise the existing LDP error handling defined in [RFC5036]. Errors and events are signaled to MPLS-MT peers using LDP notification messages. There are two kinds of MPLS-MT notification messages:

1. Error Notifications.

These are used to signal fatal errors. 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.

2. Advisory Notifications.

These are used to pass an LSR information about the MT-ID LDP session and the status of some previous message received from the peer.

6.1. MT Error Notifications

Multi-Topology Capability not supported.

Invalid Topology ID

6.2. MT Advisory Notifications

Unknown Address Family ("MT IP" and "MT IPv6")

7. 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.
8. 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.

9. Security Consideration

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

10. IANA Considerations

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

- New LDP Capability TLV: "Multi-Topology Capability" TLV (requested code point: 0x510 from LDP registry "TLV Type Name Space").

- New Status Code: "Multi-Topology Capability not supported" (requested code point: 0x50 from LDP registry "Status Code Name Space").

- New Status Code: "Invalid Topology ID" (requested code point: 0x51 from LDP registry "Status Code Name Space").

- New Status Code: "Unknown Address Family" (requested code point: 0x52 from LDP registry "Status Code Name Space").

<table>
<thead>
<tr>
<th>Registry:</th>
<th>E</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Range/Value</td>
<td></td>
<td>-----------------------------------------</td>
</tr>
<tr>
<td>0x000000050</td>
<td>1</td>
<td>Multi-Topology Capability not supported</td>
</tr>
<tr>
<td>0x000000051</td>
<td>1</td>
<td>Invalid Topology ID</td>
</tr>
<tr>
<td>0x000000052</td>
<td>0</td>
<td>Unknown Address Family</td>
</tr>
</tbody>
</table>

Figure 7: New Status Codes for LDP Multi Topology Extensions
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 8: Address Family Numbers

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

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

Figure 9: LDP Multi-Topology (MT) ID Name Space

11. Acknowledgement

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

12. References

12.1. Normative References


(FEC)”, RFC 5918, August 2010.


12.2. Informative References


Appendix A. Appendix

A.1. Application Scenarios

A.1.1. Simplified Data-plane

IGP-MT requires additional data-plane resources maintain multiple forwarding for each configured MT. On the other hand, MPLS-MT does not change the data-plane system architecture, if an IGP-MT is mapped to an MPLS-MT. In case MPLS-MT, incoming label value itself can determine an MT, and hence it requires a single NHLFE space. MPLS-MT requires only MT-RIBs in the control-plane, no need to have MT-FIBs. Forwarding IP packets over a particular MT requires either configuration or some external means at every node, to maps an attribute of incoming IP packet header to IGP-MT, which is additional overhead for network management. Whereas, MPLS-MT mapping is required only at the ingress-PE of an MPLS-MT LSP, because of each node identifies MPLS-MT LSP switching based on incoming label, hence no additional configuration is required at every node.

A.1.2. Using MT for P2P Protection

We know that [IP-FRR-MT] can be used for configuring alternate path via backup-mt, such that if primary link fails, then backup-MT can be used for forwarding. However, such techniques require special marking of IP packets that needs to be forwarded using backup-MT. MPLS-LDP-MT procedures simplify the forwarding of the MPLS packets over backup-MT, as MPLS-LDP-MT procedure distribute separate labels for each MT. How backup paths are computed depends on the implementation, and the algorithm. The MPLS-LDP-MT in conjunction with IGP-MT could be used to separate the primary traffic and backup traffic. For example, service providers can create a backup MT that consists of links that are meant only for backup traffic. Service providers can then establish bypass LSPs, standby LSPs, using backup MT, thus keeping undeterministic backup traffic away from the primary
traffic.

A.1.3. 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.1.4. Service Separation

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

A.1.5. An Alternative inter-AS VPN Solution

When the LSP is crossing multiple domains for the inter-as VPN scenarios, the LSP setup process can be done by configuring a set of routers which are in different domains into a new single domain with a new topology ID using the LDP multiple topology. All the routers belong this new topology will be used to carry the traffic across multiple domains and since they are in a single domain with the new topology ID, so the LDP LSP set up can be done without propagating VPN routes across AS boundaries.

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

Liyanuan Li  
China Mobile  
53A, Xibianmennei Ave.  
Xunwu District, Beijing  01719  
China  
Email: liliyanuan@chinamobile.com

Ning So  
Verizon Business  
2400 North Glenville Drive  
Richardson, TX  75082  
USA  
Email: Ning.So@verizonbusiness.com

Kamran Raza  
Cisco Systems  
2000 Innovation Drive  
Kanata, ON K2K-3E8, MA  
Canada  
Email: E-mail: skraza@cisco.com
Raveendra Torvi  
Juniper Networks  
10, Technology 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  
Huawei Technology  
2330 Central Expressway  
Santa Clara, CA  95050  
US  
Email: emily.cheny@huawei.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
Abstract

This specification describes the configuration of pro-active MPLS-TP Operations, Administration, and Maintenance (OAM) Functions for a given LSP using a set of TLVs that are carried by the LSP Ping protocol.

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

Status of this Memo

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

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

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

This Internet-Draft will expire on May 3, 2012.

Copyright Notice

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

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

Table of Contents

1. Introduction .............................................. 3
  1.1. Requirements Language ................................. 4
2. Overview of MPLS OAM for Transport Applications ............ 4
3. Theory of Operations ....................................... 5
  3.1. MPLS OAM Configuration Operation Overview ............. 5
    3.1.1. Configuration of BFD sessions ....................... 5
    3.1.2. Configuration of Performance Monitoring ............ 6
    3.1.3. Configuration of Fault Management Signals .......... 6
  3.2. OAM Functions TLV ....................................... 6
    3.2.1. BFD Configuration sub-TLV ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... 8
      3.2.1.1. Local Discriminator sub-TLV ....................... 10
      3.2.1.2. Negotiation Timer Parameters sub-TLV ....... 10
      3.2.1.3. BFD Authentication sub-TLV ....................... 11
    3.2.2. MPLS OAM Source MEP-ID sub-TLV ................... 12
    3.2.3. Performance Monitoring sub-TLV ..................... 13
      3.2.3.1. MPLS OAM PM Loss sub-TLV ....................... 14
      3.2.3.2. MPLS OAM PM Delay sub-TLV ...................... 15
  3.2.4. MPLS OAM FMS sub-TLV ................................ 16
  3.3. IANA Considerations ................................... 17
4. OAM configuration errors ................................... 17
5. Security Considerations ................................... 18
6. References ................................................ 18
  6.1. Normative References .................................. 18
  6.2. Informative References ................................ 19
Authors’ Addresses ........................................ 20
1. Introduction

This document describes the configuration of pro-active MPLS-TP Operations, Administration, and Maintenance (OAM) Functions for a given LSP using TLVs carried in LSP Ping [BFD-Ping]. In particular it specifies the mechanisms necessary to establish MPLS-TP OAM entities at the end points for monitoring and performing measurements on an LSP, as well as defining information elements and procedures to configure pro-active MPLS OAM functions running between LERs. Initialization and control of on-demand MPLS OAM functions are expected to be carried out by directly accessing network nodes via a management interface; hence configuration and control of on-demand OAM functions are out-of-scope for this document.

The Transport Profile of MPLS must, by definition [RFC5654], be capable of operating without a control plane. Therefore there are three options for configuring MPLS-TP OAM, without a control plane by either using an NMS or LSP Ping, or with a control plane using GMPLS (specifically RSVP-TE).

Pro-active MPLS OAM is performed by three different protocols, Bidirectional Forwarding Detection (BFD) [RFC5880] for Continuity Check/Connectivity Verification, the delay measurement protocol (DM) [RFC6374] for delay and delay variation (jitter) measurements, and the loss measurement protocol (LM) [RFC6374] for packet loss and throughput measurements. Additionally there is a number of Fault Management Signals that can be configured.

BFD is a protocol that provides low-overhead, fast detection of failures in the path between two forwarding engines, including the interfaces, data link(s), and to the extent possible the forwarding engines themselves. BFD can be used to track the liveliness and detect data plane failures of MPLS-TP point-to-point and might also be extended to support point-to-multipoint connections.

The delay and loss measurements protocols [RFC6374] use a simple query/response model for performing bidirectional measurements that allows the originating node to measure packet loss and delay in both directions. By timestamping and/or writing current packet counters to the measurement packets at four times (Tx and Rx in both directions) current delays and packet losses can be calculated. By performing successive delay measurements the delay variation (jitter) can be calculated. Current throughput can be calculated from the packet loss measurements by dividing the number of packets sent/received with the time it took to perform the measurement, given by the timestamp in LM header. Combined with a packet generator the throughput measurement can be used to measure the maximum capacity of a particular LSP.
MPLS Transport Profile (MPLS-TP) describes a profile of MPLS that enables operational models typical in transport networks, while providing additional OAM, survivability and other maintenance functions not currently supported by MPLS. [RFC5860] defines the requirements for the OAM functionality of MPLS-TP.

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

1.1. Requirements Language

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

2. Overview of MPLS OAM for Transport Applications

[MPLS-TP-OAM-FWK] describes how MPLS OAM mechanisms are operated to meet transport requirements outlined in [RFC5860].

[BFD-CCCV] specifies two BFD operation modes: 1) "CC mode", which uses periodic BFD message exchanges with symmetric timer settings, supporting Continuity Check, 2) "CV/CC mode" which sends unique maintenance entity identifiers in the periodic BFD messages supporting Connectivity Verification as well as Continuity Check.

[RFC6374] specifies mechanisms for performance monitoring of LSPs, in particular it specifies loss and delay measurement OAM functions.

[MPLS-FMS] specifies fault management signals with which a server LSP can notify client LSPs about various fault conditions to suppress alarms or to be used as triggers for actions in the client LSPs. The following signals are defined: Alarm Indication Signal (AIS), Link Down Indication (LDI) and Locked Report (LKR). To indicate client faults associated with the attachment circuits Client Signal Failure Indication (CSF) can be used. CSF is described in [MPLS-TP-OAM-FWK] and in the context of this document is for further study.

[MPLS-TP-OAM-FWK] describes the mapping of fault conditions to consequent actions. Some of these mappings may be configured by the operator, depending on the application of the LSP. The following defects are identified: Loss Of Continuity (LOC), Misconnectivity, MEP Misconfiguration and Period Misconfiguration. Out of these defect conditions, the following consequent actions may be
configurable: 1) whether or not the LOC defect should result in blocking the outgoing data traffic; 2) whether or not the "Period Misconfiguration defect" should result in a signal fail condition.

3. Theory of Operations

3.1. MPLS OAM Configuration Operation Overview

LSP Ping, or alternatively RSVP-TE [RSVP-TE CONF], can be used to simply enable the different OAM functions, by setting the corresponding flags in the "OAM Functions TLV". Additionally one may include sub-TLVs for the different OAM functions in order to specify different parameters in detail.

The presence of OAM configuration TLVs at intermediate nodes is justified because the intermediate nodes need to forward the LSP-ping message to the end point. No TLV processing or modification or following OAM actions need to be taken at the intermediate points.

3.1.1. Configuration of BFD sessions

For this specification, BFD MUST be run in either one of the two modes:

- Asynchronous mode, where both sides should be in active mode
- Unidirectional mode

In the simplest scenario LSP Ping, or alternatively RSVP-TE [RSVP-TE CONF], is used only to bootstrap a BFD session for an LSP, without any timer negotiation.

Timer negotiation can be performed either in subsequent BFD control messages (in this case the operation is similar to LSP Ping based bootstrapping described in [RFC5884]) or directly in the LSP ping configuration messages.

When BFD Control packets are transported in the G-ACh they are not protected by any end-to-end checksum, only lower-layers are providing error detection/correction. A single bit error, e.g. a flipped bit in the BFD State field could cause the receiving end to wrongly conclude that the link is down and in turn trigger protection switching. To prevent this from happening the "BFD Configuration sub-TLV" has an Integrity flag that when set enables BFD Authentication using Keyed SHA1 with an empty key (all 0s) [RFC5880]. This would make every BFD Control packet carry an SHA1 hash of itself that can be used to detect errors.
If BFD Authentication using a pre-shared key / password is desired (i.e. actual authentication not only error detection) the "BFD Authentication sub-TLV" MUST be included in the "BFD Configuration sub-TLV". The "BFD Authentication sub-TLV" is used to specify which authentication method that should be used and which pre-shared key / password that should be used for this particular session. How the key exchange is performed is out of scope of this document.

3.1.2. Configuration of Performance Monitoring

It is possible to configure Performance Monitoring functionalities such as Loss, Delay and Throughput as described in [RFC6374].

When configuring Performance monitoring functionalities it can be chosen either the default configuration (by only setting the respective flags in the "OAM functions TLV") or a customized configuration (by including the respective Loss and/or Delay sub-TLVs).

3.1.3. Configuration of Fault Management Signals

Additional OAM functions may be configured by setting the appropriate flags in the "OAM Functions TLV", these include Performance Measurements (packet loss, throughput, delay, and delay variation) and Fault Management Signal handling.

By setting the PM Loss flag in the "OAM Functions TLV" and including the "MPLS OAM PM Loss sub-TLV" one can configure the measurement interval and loss threshold values for triggering protection.

Delay measurements are configured by setting PM Delay flag in the "OAM Functions TLV" and including the "MPLS OAM PM Loss sub-TLV" one can configure the measurement interval and the delay threshold values for triggering protection.

To configure Fault Monitoring Signals and their refresh time the FMS flag in the "OAM Functions TLV" MUST be set and the "MPLS OAM FMS sub-TLV" included. If an intermediate point is meant to originate fault management signal messages this mean that such intermediate point is associated to a server MEP through a co-located MPLS-TP client/server adaptation function and such server MEP needs to be configured by its own LSP-ping session (or, in alternative, NMS or RSVP-TE).

3.2. OAM Functions TLV

The "OAM Functions TLV" depicted below is carried as a TLV of the LSP Echo request/response messages.
The "OAM Functions TLV" contains a number of flags indicating which OAM functions should be activated as well as OAM function specific sub-TLVs with configuration parameters for the particular function.

Type: indicates a new type, the "OAM Functions TLV" (IANA to define, suggested value 16).

Length: the length of the OAM Function Flags field including the total length of the sub-TLVs in octets.

OAM Function Flags: a bitmap numbered from left to right as shown in the figure.

These flags are defined in this document:

<table>
<thead>
<tr>
<th>OAM Function Flag bit#</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>0 (C)</td>
<td>Continuity Check (CC)</td>
</tr>
<tr>
<td>1 (V)</td>
<td>Connectivity Verification (CV)</td>
</tr>
<tr>
<td>2 (F)</td>
<td>Fault Management Signals (FMS)</td>
</tr>
<tr>
<td>3 (L)</td>
<td>Performance Monitoring/Loss (PM/Loss)</td>
</tr>
<tr>
<td>4 (D)</td>
<td>Performance Monitoring/Delay (PM/Delay)</td>
</tr>
<tr>
<td>5 (T)</td>
<td>Throughput Measurement</td>
</tr>
<tr>
<td>6-31</td>
<td>Reserved (set all to 0s)</td>
</tr>
</tbody>
</table>

Sub-TLVs corresponding to the different flags are as follows:

- "BFD Configuration sub-TLV", which MUST be included if the CC and/or the CV OAM Function flag is set. This sub-TLV MUST carry a "BFD Local Discriminator sub-TLV" and a "Timer Negotiation Parameters sub-TLV" if the N flag is cleared. "MPLS OAM Source MEP-ID sub-TLV" MUST also be included. If the I flag is set, the "BFD Authentication sub-TLV" may be included.
- "MPLS OAM PM Loss sub-TLV" within the "Performance Monitoring sub-TLV", which MAY be included if the PM/Loss OAM Function flag is set. If the "MPLS OAM PM Loss sub-TLV" is not included, default configuration values are used. Such sub-TLV MAY also be included in case the Throughput function flag is set and there is the need to specify measurement interval different from the default ones. In fact the throughput measurement make use of the same tool as the loss measurement, hence the same TLV is used.

- "MPLS OAM PM Delay sub-TLV" within the "Performance Monitoring sub-TLV", which MAY be included if the PM/Delay OAM Function flag is set. If the "MPLS OAM PM Delay sub-TLV" is not included, default configuration values are used.

- "MPLS OAM FMS sub-TLV", which MAY be included if the FMS OAM Function flag is set. If the "MPLS OAM FMS sub-TLV" is not included, default configuration values are used.

3.2.1. BFD Configuration sub-TLV

The "BFD Configuration sub-TLV" (depicted below) is defined for BFD OAM specific configuration parameters. The "BFD Configuration sub-TLV" is carried as a sub-TLV of the "OAM Functions TLV".

This TLV accommodates generic BFD OAM information and carries sub-TLVs.

```
| Type: indicates a new type, the "BFD Configuration sub-TLV" (IANA to define, suggested value 1).
| Length: indicates the length of the TLV including sub-TLVs but excluding the Type and Length field, in octets.
| Version: identifies the BFD protocol version. If a node does not support a specific BFD version an error must be generated: "OAM Bellagamba, et al. Expires May 3, 2012 [Page 8]"
Problem/Unsupported OAM Version”.

PHB: Identifies the Per-Hop Behavior (PHB) to be used for periodic continuity monitoring messages.

BFD Negotiation (N): If set timer negotiation/re-negotiation via BFD Control Messages is enabled, when cleared it is disabled.

Symmetric session (S): If set the BFD session MUST use symmetric timing values.

Integrity (I): If set BFD Authentication MUST be enabled. If the "BFD Configuration sub-TLV" does not include a "BFD Authentication sub-TLV" the authentication MUST use Keyed SHA1 with an empty pre-shared key (all 0s).

Encapsulation Capability (G): if set, it shows the capability of encapsulating BFD messages into G-Ach channel. If both the G bit and U bit are set, configuration gives precedence to the G bit.

Encapsulation Capability (U): if set, it shows the capability of encapsulating BFD messages into UDP packets. If both the G bit and U bit are set, configuration gives precedence to the G bit.

Bidirectional (B): if set, it configures BFD in the Bidirectional mode. If it is not set it configures BFD in unidirectional mode. In the second case, the source node does not expect any Discriminator values back from the destination node.

The "BFD Configuration sub-TLV" MUST include the following sub-TLVs in the LSP Echo request message:

- "Local Discriminator sub-TLV";
- "Negotiation Timer Parameters sub-TLV" if the N flag is cleared.

The "BFD Configuration sub-TLV" MUST include the following sub-TLVs in the LSP Echo reply message:

- "Local Discriminator sub-TLV;"
- "Negotiation Timer Parameters sub-TLV" if:
  - the N and S flags are cleared
  - the N flag is cleared and the S flag is set and a timing interval larger than the one received needs to be used
Reserved: Reserved for future specification and set to 0.

3.2.1.1. Local Discriminator sub-TLV

The "Local Discriminator sub-TLV" is carried as a sub-TLV of the "BFD Configuration sub-TLV" and is depicted below.

[Author’s note: This should be aligned with RFC5884, exactly how to do that is under discussion.]

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
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|  Lcl. Discr. Type (1) (IANA)  |         Length (4)            |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|                       Local Discriminator                     |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+

Type: indicates a new type, the "Local Discriminator sub-TLV" (IANA to define, suggested value 1).

Length: indicates the TLV total length in octets.

Local Discriminator: A unique, nonzero discriminator value generated by the transmitting system and referring to itself, used to demultiplex multiple BFD sessions between the same pair of systems.

3.2.1.2. Negotiation Timer Parameters sub-TLV

The "Negotiation Timer Parameters sub-TLV" is carried as a sub-TLV of the "BFD Configuration sub-TLV" and is depicted below.

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
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|  Timer Neg. Type (2) (IANA)  |          Length (16)          |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|         Acceptable Min. Asynchronous TX interval              |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|         Acceptable Min. Asynchronous RX interval              |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|               Required Echo TX Interval                       |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+

Type: indicates a new type, the "Negotiation Timer Parameters sub-TLV" (IANA to define, suggested value 2).
Length: indicates the length of the parameters in octets (16).

Acceptable Min. Asynchronous TX interval: in case of S (symmetric) flag set in the "BFD Configuration" TLV, it expresses the desired time interval (in microseconds) at which the LER initiating the signaling intends to both transmit and receive BFD periodic control packets. If the receiving edge LSR cannot support such value, it is allowed to reply back with an interval greater than the one proposed.

In case of S (symmetric) flag cleared in the "BFD Configuration sub-TLV", this field expresses the desired time interval (in microseconds) at which a edge LSR intends to transmit BFD periodic control packets in its transmitting direction.

Acceptable Min. Asynchronous RX interval: in case of S (symmetric) flag set in the "BFD Configuration sub-TLV", this field MUST be equal to "Acceptable Min. Asynchronous TX interval" and has no additional meaning respect to the one described for "Acceptable Min. Asynchronous TX interval".

In case of S (symmetric) flag cleared in the "BFD Configuration sub-TLV", it expresses the minimum time interval (in microseconds) at which edge LSRs can receive BFD periodic control packets. In case this value is greater than the "Acceptable Min. Asynchronous TX interval" received from the other edge LSR, such edge LSR MUST adopt the interval expressed in this "Acceptable Min. Asynchronous RX interval".

Required Echo TX Interval: the minimum interval (in microseconds) between received BFD Echo packets that this system is capable of supporting, less any jitter applied by the sender as described in [RFC5880] sect. 6.8.9. This value is also an indication for the receiving system of the minimum interval between transmitted BFD Echo packets. If this value is zero, the transmitting system does not support the receipt of BFD Echo packets. If the receiving system cannot support this value an error MUST be generated "Unsupported BFD TX Echo rate interval". By default the value is set to 0.

3.2.1.3. BFD Authentication sub-TLV

The "BFD Authentication sub-TLV" is carried as a sub-TLV of the "BFD Configuration sub-TLV" and is depicted below.
Type: indicates a new type, the "BFD Authentication sub-TLV" (IANA to define).

Length: indicates the TLV total length in octets. (8)

Auth Type: indicates which type of authentication to use. The same values as are defined in section 4.1 of [RFC5880] are used.

Auth Key ID: indicates which authentication key or password (depending on Auth Type) should be used. How the key exchange is performed is out of scope of this document.

Reserved: Reserved for future specification and set to 0.

3.2.2. MPLS OAM Source MEP-ID sub-TLV

The "MPLS OAM Source MEP-ID sub-TLV" depicted below is carried as a sub-TLV of the "OAM Functions TLV".

Type: indicates a new type, the "MPLS OAM Source MEP-ID sub-TLV" (IANA to define, suggested value 3).

Length: indicates the length of the parameters in octets (8).

Source Node ID: 32-bit node identifier as defined in [MPLS-TP-IDENTIF].

Tunnel ID: a 16-bit unsigned integer unique to the node as defined in [MPLS-TP-IDENTIF].
LSP ID: a 16-bit unsigned integer unique within the Tunnel_ID as defined in [MPLS-TP-IDENTIF].

3.2.3. Performance Monitoring sub-TLV

If the "OAM functions TLV" has either the L (Loss), D (Delay) or T (Throughput) flag set, the "Performance Monitoring sub-TLV" MUST be present.

In case the values need to be different than the default ones the "Performance Monitoring sub-TLV", "MPLS OAM PM Loss sub-TLV" MAY include the following sub-TLVs:

- "MPLS OAM PM Loss sub-TLV" if the L flag is set in the "OAM functions TLV";
- "MPLS OAM PM Delay sub-TLV" if the D flag is set in the "OAM functions TLV";

The "Performance Monitoring sub-TLV" depicted below is carried as a sub-TLV of the "OAM Functions TLV".

```
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
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|    Perf Monitoring Type (IANA) |          Length               |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|D|L|J|Y|K|C|            Reserved (set to all 0s)               |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|                                                               |
˜                           sub-TLVs                            ˜
|                                                               |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
Configuration Flags, for the specific function description please refer to [RFC6374]:

- D: Delay inferred/direct (0=INFERRED, 1=DIRECT)
- L: Loss inferred/direct (0=INFERRED, 1=DIRECT)
- J: Delay variation/jitter (1=ACTIVE, 0=NOT ACTIVE)
- Y: Dyadic (1=ACTIVE, 0=NOT ACTIVE)
- K: Loopback (1=ACTIVE, 0=NOT ACTIVE)
C: Combined (1=ACTIVE, 0=NOT ACTIVE)

3.2.3.1. MPLS OAM PM Loss sub-TLV

The "MPLS OAM PM Loss sub-TLV" depicted below is carried as a sub-TLV of the "Performance Monitoring sub-TLV".

```
0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
| PM Loss Type (1) (IANA) |          Length               |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
| OTF |T|B|                    RESERVED                         |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|                    Measurement Interval                       |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|                       Test Interval                           |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|                      Loss Threshold                           |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
```

Type: indicates a new type, the "MPLS OAM PM Loss sub-TLV" (IANA to define, suggested value 1).

Length: indicates the length of the parameters in octets (12).

OTF: Origin Timestamp Format of the Origin Timestamp field described in [RFC6374]. By default it is set to IEEE 1588 version 1.

Configuration Flags, please refer to [RFC6374] for further details:

- T: Traffic-class-specific measurement indicator. Set to 1 when the measurement operation is scoped to packets of a particular traffic class (DSCP value), and 0 otherwise. When set to 1, the DS field of the message indicates the measured traffic class. By default it is set to 1.

- B: Octet (byte) count. When set to 1, indicates that the Counter 1-4 fields represent octet counts. When set to 0, indicates that the Counter 1-4 fields represent packet counts. By default it is set to 0.

Measurement Interval: the time interval (in microseconds) at which Loss Measurement query messages MUST be sent on both directions. If the edge LSR receiving the Path message can not support such value, it can reply back with a higher interval. By default it is set to (TBD).
Test Interval: test messages interval as described in [RFC6374]. By default it is set to (TBD).

Loss Threshold: the threshold value of lost packets over which protections MUST be triggered. By default it is set to (TBD).

3.2.3.2. MPLS OAM PM Delay sub-TLV

The "MPLS OAM PM Delay sub-TLV" depicted below is carried as a sub-TLV of the "OAM Functions TLV".

```
0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|  PM Delay Type (2) (IANA)     |          Length               |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
| OTF |T|B|                    RESERVED                         |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|                    Measurement Interval                       |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|                       Test Interval                           |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|                      Delay Threshold                          |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
```

Type: indicates a new type, the "MPLS OAM PM Loss sub-TLV" (IANA to define, suggested value 1).

Length: indicates the length of the parameters in octets (12).

OTF: Origin Timestamp Format of the Origin Timestamp field described in [RFC6374]. By default it is set to IEEE 1588 version 1.

Configuration Flags, please refer to [RFC6374] for further details:

- T: Traffic-class-specific measurement indicator. Set to 1 when the measurement operation is scoped to packets of a particular traffic class (DSCP value), and 0 otherwise. When set to 1, the DS field of the message indicates the measured traffic class. By default it is set to 1.

- B: Octet (byte) count. When set to 1, indicates that the Counter 1-4 fields represent octet counts. When set to 0, indicates that the Counter 1-4 fields represent packet counts. By default it is set to 0.

Measurement Interval: the time interval (in microseconds) at which
Delay Measurement query messages MUST be sent on both directions. If the edge LSR receiving the Path message can not support such value, it can reply back with a higher interval. By default it is set to (TBD).

Test Interval: test messages interval as described in [RFC6374]. By default it is set to (TBD).

Delay Threshold: the threshold value of measured delay (in microseconds) over which protections MUST be triggered. By default it is set to (TBD).

3.2.4. MPLS OAM FMS sub-TLV

The "MPLS OAM FMS sub-TLV" depicted below is carried as a sub-TLV of the "OAM Functions TLV".

```
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
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
| Fault mgmt Type (4) (IANA) |        Length (8)             |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|A|D|L|C|                Reserved (set to all 0s)         | PHB |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|                      Refresh Timer                            |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
```

Type: indicates a new type, the "MPLS OAM FMS sub-TLV" (IANA to define, suggested value 4).

Length: indicates the length of the parameters in octets (8).

Signal Flags should not be processed at intermediate nodes as they only have an end-point significance. They are used to enable the following signals at end points:

- A: Alarm Indication Signal (AIS) as described in [MPLS-FMS]
- D: Link Down Indication (LDI) as described in [MPLS-FMS]
- L: Locked Report (LKR) as described in [MPLS-FMS]
- C: Client Signal Failure (CSF) as described in [MPLS-CSF]
- Remaining bits: Reserved for future specification and set to 0.

Configuration Flags:
- PHB: identifies the per-hop behavior of packets with fault management information. It is significant only when C flag is set.

Refresh Timer: indicates the refresh timer (in microseconds) of fault indication messages. If the edge LSR receiving the Path message cannot support such value, it can reply back with a higher interval.

3.3. IANA Considerations

This document specifies the following new TLV types:

- "OAM Functions" type: 16;

sub-TLV types to be carried in the "OAM Functions TLV":

- "BFD Configuration" type: 1;
- "MPLS OAM PM Loss" type: 2;
- "MPLS OAM PM Delay" type: 3;
- "MPLS OAM FMS" type: 4.

sub-TLV types to be carried in the "BFD Configuration sub-TLV":

- "Local Discriminator" type: 1;
- "Negotiation Timer Parameters" type: 2;
- "MPLS OAM Source MEP-ID" type: 3.
- "BFD Authentication" sub-TLV type: 4.

4. OAM configuration errors

This document specifies additional Return Codes to LSP Ping:

- "MPLS OAM Unsupported Functionality" (IANA to assign, suggested value 16);
- "OAM Problem/Unsupported TX rate interval" (IANA to assign, suggested value 17).
5. Security Considerations

The signaling of OAM related parameters and the automatic establishment of OAM entities introduces additional security considerations to those discussed in [RFC3473]. In particular, a network element could be overloaded if an attacker were to request high frequency liveliness monitoring of a large number of LSPs, targeting a single network element.

Security aspects will be covered in more detailed in subsequent versions of this document.

6. References

6.1. Normative References

[MPLS-FMS]

[MPLS-TP-IDENTIF]

[OAM-CONF-FWK]


6.2. Informative References

[BFD-CCCV]

[BFD-Ping]

[ETH-OAM]

[MPLS-TP OAM Analysis]

[MPLS-TP-OAM-FWK]

[RFC3479]

[RFC4447]
Martini, L., Rosen, E., El-Aawar, N., Smith, T., and G.

Heron, "Pseudowire Setup and Maintenance Using the Label Distribution Protocol (LDP)", RFC 4447, April 2006.


Authors’ Addresses

Elisa Bellagamba (editor)
Ericsson
Torshamnsgatan 48
Kista, 164 40
Sweden

Email: elisa.bellagamba@ericsson.com

Loa Andersson
Ericsson
Torshamnsgatan 48
Kista, 164 40
Sweden

Phone:
Email: loa.andersson@ericsson.com

Pontus Skoldstrom (editor)
Acreo AB
Electrum 236
Kista, 164 40
Sweden

Phone: +46 8 6327731
Email: pontus.skoldstrom@acreo.se

Dave Ward
Juniper

Phone:
Email: dward@juniper.net

Abstract

This document provides a security framework for Multiprotocol Label Switching Transport Profile (MPLS-TP). MPLS-TP extends MPLS technologies and introduces new OAM capabilities, a transport-oriented path protection mechanism, and strong emphasis on static provisioning supported by network management systems. This document addresses the security aspects relevant in the context of MPLS-TP specifically. It describes potential security threats, security requirements for MPLS-TP, and mitigation procedures for MPLS-TP networks and MPLS-TP interconnection to other MPLS and GMPLS networks.

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

This Informational Internet-Draft is aimed at achieving IETF Consensus before publication as an RFC and will be subject to an IETF Last Call.

[RFC Editor, please remove this note before publication as an RFC and insert the correct Streams Boilerplate to indicate that the published RFC has IETF Consensus.]
# Table of Contents

1.  Introduction .................................................. 4
   1.1.  Background and Motivation .............................. 4
   1.2.  Scope .................................................... 4
   1.3.  Requirement Language .................................... 5
   1.4.  Terminology .............................................. 6
   1.5.  Structure of the document ............................. 7
2.  Security Reference Models .................................... 7
   2.1.  Security Reference Model 1 ............................. 7
   2.2.  Security Reference Model 2 ............................. 9
   2.3.  Security Reference Model 3 ............................. 12
   2.4.  Trusted Zone Boundaries ............................... 13
3.  Security Threats ............................................... 14
   3.1.  Attacks on the Control Plane .......................... 16
   3.2.  Attacks on the Data Plane ............................. 17
4.  Security Requirements for MPLS-TP ........................... 17
5.  Defensive Techniques for MPLS-TP Networks .................. 19
   5.1.  Authentication ......................................... 19
       5.1.1.  Management System Authentication .................. 19
       5.1.2.  Peer-to-Peer Authentication ....................... 20
       5.1.3.  Cryptographic Techniques for Authenticating
               Identity ........................................... 20
   5.2.  Access Control Techniques ............................ 20
   5.3.  Use of Isolated Infrastructure ....................... 21
   5.4.  Use of Aggregated Infrastructure ..................... 21
   5.5.  Protection against Denial of Service ................ 21
   5.6.  Verification of Connectivity .......................... 21
6.  Monitoring, Detection, and Reporting of Security Attacks .. 21
7.  Security Considerations ...................................... 22
8.  IANA Considerations .......................................... 22
9.  References ................................................... 23
   9.1.  Normative References ................................... 23
   9.2.  Informative References ............................... 23
Authors’ Addresses ................................................. 24
1. Introduction

1.1. Background and Motivation

This document provides a security framework for Multiprotocol Label Switching Transport Profile (MPLS-TP).

The MPLS-TP Requirements and MPLS-TP Framework are defined in [RFC5654] and [RFC5921], respectively. The intent of MPLS-TP development is to address the needs for transport evolution and the fast-growing bandwidth demand accelerated by new packet-based services and multimedia applications, from Ethernet Services, Layer 2 and Layer 3 VPNs, and triple play to Mobile Access Network (RAN) backhaul, etc. MPLS-TP is based on MPLS technologies to take advantage of this technology’s maturity, and maintaining the transport characteristics of MPLS is an MPLS-TP requirement.

To focus on meeting transport requirements, MPLS-TP uses a subset of MPLS features and introduces extensions to reflect the characteristics of the transport technology. The added functionalities include in-band OAM, transport-oriented path protection and recovery mechanisms, etc. There is strong emphasis on static provisioning supported by network management systems (NMS) or Operation Support Systems (OSS). MPLS-TP and MPLS without TP also need to interwork.

The security aspects of the extensions particularly designed for MPLS-TP need to be addressed. The security models, threats, requirements, and defense techniques previously defined in [RFC5920] can be applied to reuse existing functionality in MPLS and GMPLS but are not sufficient to cover the TP extensions.

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

1.2. Scope

This document addresses the security aspects specific to MPLS-TP. It defines security models that apply to various MPLS-TP deployment scenarios, identifies potential security threats and mitigation procedures for MPLS-TP networks and MPLS-TP interconnection to GMPLS or MPLS networks without TP, and provides security requirements for MPLS-TP. Inter-AS and Inter-provider security for MPLS-TP to MPLS-TP connections or MPLS-TP to MPLS connections without TP are discussed, because these connections present higher security risks than connections for Intra-AS MPLS-TP.

The general security analysis and guidelines for MPLS and GMPLS are addressed in [RFC5920], and the content of [RFC5920] that has no new impact on MPLS-TP is not repeated in this document. Other general security issues regarding transport networks that are not specific to MPLS-TP are also found elsewhere. Readers may also refer to the "Security Best Practices Efforts and Documents" Opsec Effort [opsec-efforts] and "Security Mechanisms for the Internet" [RFC3631] (if there are linkages to the Internet in the applications) for general network operations security considerations. This document does not define the specific mechanisms or methods that must be implemented to satisfy the security requirements.

The issues and areas addressed with respect to MPLS-TP security are:

- Attacks against G-Ach integrity, availability, or confidentiality
- Misuse of G-Ach to attack data plane resources
- ID Spoofing attacks
- Attacks against the loopback mechanism and Authentication TLV
- Attacks against the network management system (NMS)
- NMS and CP interaction vulnerabilities
- MIP and MEP assignment and attacks on these mechanisms
- Topology discovery vulnerabilities
- Data plane authentication (using G-Ach or by other means)
- Label authentication
- DoS attacks on the data plane
- Performance monitoring vulnerabilities

1.3. Requirement Language

The key words "MUST", "MUST NOT", "REQUIRED", "SHALL", "SHALL NOT", "SHOULD", "SHOULD NOT", "RECOMMENDED", "MAY", and "OPTIONAL" in this document are to be interpreted as described in [RFC2119]. Although this document is not a protocol specification, the use of this language clarifies the instructions to protocol designers producing solutions that satisfy the requirements set out in this document.
1.4. Terminology

This document uses MPLS, MPLS-TP, and security terminology. Detailed definitions and additional terminology for MPLS-TP may be found in [RFC5654] and [RFC5921]. MPLS/GMPLS security-related terminology can be found in [RFC5920].

- AC: Attachment Circuit
- BFD: Bidirectional Forwarding Detection
- CE: Customer-Edge device
- DoS: Denial of Service
- DDoS: Distributed Denial of Service
- GAL: Generic Alert Label
- G-ACh: Generic Associated Channel
- GMPLS: Generalized Multi-Protocol Label Switching
- LDP: Label Distribution Protocol
- LSP: Label Switched Path
- MCC: Management Communication Channel
- MEP: Maintenance End Point
- MIP: Maintenance Intermediate Point
- MPLS: MultiProtocol Label Switching
- OAM: Operations, Administration, and Management
- PE: Provider-Edge device
- PSN: Packet-Switched Network
- PW: Pseudowire
- RSVP: Resource Reservation Protocol
- RSVP-TE: Resource Reservation Protocol with Traffic Engineering Extensions
- S-PE: Switching Provider Edge
2. Security Reference Models

This section defines reference models for security in MPLS-TP networks.

The models are built on the architecture of MPLS-TP defined in [RFC5921]. The Service Provider (SP) boundaries play an important role in determining the security models for any particular deployment.

This document defines a trusted zone as being where a single SP has total operational control over that part of the network. A primary concern is about security aspects that relate to breaches of security from the "outside" of a trusted zone to the "inside" of this zone.

2.1. Security Reference Model 1

In reference model 1, a single SP has total control of the PE/T-PE to PE/T-PE part of the MPLS-TP network.
Security reference model 1(a)

An MPLS-TP network with Single Segment Pseudowire (SS-PW) from PE to PE. The trusted zone is PE1 to PE2 as illustrated in MPLS-TP Security Model 1 (a) (Figure 1).

```
<----- Emulated Service ----------->
|                               |
|<------ Pseudo Wire ------->   |<-- PSN Tunnel -->|
|       V                      |       V        |
| AC +----+                   | AC +----+     |
V PE1 |-------------------------| PE2 |-------------------------|
V     |                PW1        |    |                PW2        |
      |                      V    |      |                      V    |
      |    +----+            +----+  |    +----+            +----+  |
      | CE1 |-----------------| CE2 | | CE1 |-----------------| CE2 |
      |     |                |     |     |                |     |
      ^    |-----------------|    ^ | ^    |-----------------|    ^ |
      |  Provider Edge 1 |          |  Provider Edge 2 |          |
      Customer  Edge 1  |          | Customer  Edge 2  |
```

Native service

----Untrusted---->|<------ Trusted Zone ------>|<---Untrusted----

MPLS-TP Security Model 1 (a)

Figure 1

Security reference model 1(b)

An MPLS-TP network with Multi-Segment Pseudowire (MS-PW) from T-PE to T-PE. The trusted zone is T-PE1 to T-PE2 in this model as illustrated in MPLS-TP Security Model 1 (b) (Figure 2).
Figure 2

2.2. Security Reference Model 2

In reference model 2, a single SP does not have total control of the PE/T-PE to PE/T-PE part of the MPLS-TP network. S-PE and T-PE may be under the control of different SPs, or their customers or may not be trusted for some other reason. The MPLS-TP network is not contained within a single trusted zone.

Security Reference Model 2(a)

An MPLS-TP network with Multi-Segment Pseudowire (MS-PW) from T-PE to T-PE. The trusted zone is T-PE1 to S-PE, as illustrated in MPLS-TP Security Model 2 (a) (Figure 3).
An MPLS-TP network with Multi-Segment Pseudowire (MS-PW) from T-PE to T-PE. The trusted zone is the S-PE, as illustrated in MPLS-TP Security Model 2 (b) (Figure 4).
An MPLS-TP network with Multi-Segment Pseudowire (MS-PW) from different Service Providers with inter-provider PW connections. The trusted zone is T-PE1 to S-PE3, as illustrated in MPLS-TP Security Model 2 (c) (Figure 5).
An MPLS-TP network with a Transport LSP from PE1 to PE2. The trusted zone is PE1 to PE2 as illustrated in MPLS-TP Security Model 3 (a) (Figure 6), where the two PEs and the devices in between them are under control of a single service operator.
2.4. Trusted-Zone Boundaries

The boundaries of a trusted zone should be carefully defined when analyzing the security properties of each individual network. As illustrated above, the security boundaries determine which reference model should be applied to analyze use cases.

A key requirement of MPLS-TP networks is that the security of a trusted zone MUST NOT be compromised by interconnecting one SP’s MPLS-TP or MPLS infrastructure with another SP’s core devices, T-PE devices, or end users.

In addition, neighboring nodes in the network may be trusted or untrusted. Neighbors may also be authorized or unauthorized. Even though a neighbor may be authorized for communication, it may not be trusted. For example, when connecting with another provider’s S-PE to set up Inter-AS LSPs, the other provider is considered to be untrusted but may be authorized for communication.
3. Security Threats

This section lists various network security threats that may endanger MPLS-TP networks. It emphasizes threats that are new to MPLS-TP networks or affect MPLS-TP networks in new ways.

A successful attack on a particular MPLS-TP network or on a SP’s MPLS-TP infrastructure may cause one or more of the following adverse effects:

1. Observation (including traffic pattern analysis), modification, or deletion of a provider’s or user’s data, as well as replay or insertion of inauthentic data into a provider’s or user’s data stream. These types of attacks apply to MPLS-TP traffic regardless of how the LSP or PW is set up in a similar way to how they apply to MPLS traffic regardless how the LSP is set up.

2. Compromised GAL label or BFD messaging:
   a. GAL label or BFD label manipulation, which includes insertion of false labels or messages and modification, deletion, or replay of GAL labels or messages.
   b. DoS attack through in-band OAM G-ACh/GAL and BFD messages.
   c. Attacks via G-ACh to cause protection switchover, restoration, or locking of a transport connection.

3. Disruption of a provider’s or user’s connectivity, or degradation of a provider’s service quality.
a. Attacks against a SP’s connectivity:

+ In the case in which an NMS is used for LSP setup, the
attacks occur through attacks on the NMS.

+ In the case in which dynamic provisioning is used, the
attacks occur on the dynamic control plane. Most aspects
of these are addressed in [RFC5920].

b. Attacks against user’s connectivity. These are similar to
PE/CE attacks against access in typical MPLS networks and
are addressed in [RFC5920].

4. Probes of a provider’s network to determine its configuration,
capacity, or usage. This can occur through attacks against an
NMS in the case of static provisioning or attacks against the
control plane in dynamic MPLS-TP networks. It can also result
from combined attacks.

It is helpful to consider that threats, whether resulting from
malicious behavior or accidental errors, may come from different
sources or categories of attackers. For example, they may come from:

- Users of the MPLS-TP network itself, who may attack the network or
  other users. These other users’ services may be provided by the
  same or a different MPLS-TP core.

- The MPLS-TP SP its employees.

- Other persons who obtain physical access to a MPLS-TP SP’s site.

- Other persons who use social engineering to influence the behavior
  of a SP’s personnel.

- Outsiders, e.g., attackers from the other sources, including the
  Internet (if connectivity can be obtained).

- Other SPs in the case of MPLS-TP inter-provider connection. The
  other provider may or may not be using MPLS-TP.

- Those who create, deliver, install, and maintain hardware or software
  for network equipment.

Security is a tradeoff between cost and risk, so it is useful to
consider the likelihood of different attacks, the cost of preventing
them, and the possible damage resulting from their occurrence. There is
at least a perceived difference in the likelihood of most types of
attacks being successfully mounted in different environments, such as:
o A MPLS-TP network inter-connected with another provider’s core

o A MPLS-TP configuration inter-connected with the public Internet, e.g., for control or management functions

Most types of attacks become easier to mount and hence more likely as the shared infrastructure via which service is provided expands from a single SP to multiple cooperating SPs to the global Internet. Attacks that may not be of sufficient likeliness to warrant concern in a closely controlled environment often merit defensive measures in broader, more open environments. Even though surveys show that 40% to 60% of attacks originate from insiders, in closed communities, it is often practical to identify and to deal with misbehavior after the fact: an employee can be disciplined, for example.

The following sections list specific types of exploits that threaten MPLS-TP networks.

3.1. Attacks on the Control Plane

This category includes attacks that may compromise the availability of control plane capabilities, the integrity of these operations, and, potentially, the confidentiality of these operations for either in-band (G-ACh) or out-of-band (GMPLS) configurations. Attacks against GMPLS include attacks against its constituent protocols (i.e., RSVP-TE, LDP, OSPFv2, or PCEP). Attacks against G-ACh may be directed against the label mechanism (GAL) or any of the encapsulated signaling or management protocols (SCC, MCC, or OAM, or protection). The following attacks may target the provisioning, management, or survivability functions of the control plane:

o Improper MPLS-TP LSP or PW creation or deletion. This may result from a failure of control plane authentication or authorization mechanisms or compromise of control plane traffic. One result might be improper cross-connection of different users’ traffic.

o Improper use of MPLS-TP protection and restoration capabilities. This also may result from a failure of control plane authentication or authorization mechanisms or compromise of control plane traffic.

o Unauthorized observation of control plane traffic, which includes information about a SP’s MPLS-TP configuration, equipment, or users.

o Denial of service attacks on the control plane or use of the control plane to carry out denial of service attacks against the data plane.

o Attacks on the SP’s MPLS-TP equipment or software. These may occur during the normal lifecycle of the equipment and software or via management interfaces or other points of entry. These include social engineering attacks on the SP’s infrastructure.
3.2. Attacks on the Data Plane

This category encompasses attacks on the SP’s or end user’s data. Note that from the MPLS-TP network end user’s point of view, some of this may be control plane traffic, e.g. a routing protocol running from user’s site A to site B via IP or non-IP connections, e.g., a VPN.

- Denial of service, misconnection, loss of bandwidth, or other service disruptions.
- Unauthorized observation of data traffic, includes LSP or PW message interception and traffic pattern analysis.
- Modification or deletion of data traffic, which may include insertion of inauthentic data traffic (spoofing or replay).

4. Security Requirements for MPLS-TP

This section covers requirements for securing an MPLS-TP network infrastructure. The MPLS-TP network can be operated without a control plane or via dynamic control plane protocols. The security requirements related to MPLS-TP OAM, recovery mechanisms, MPLS-TP interconnection with other technologies, and operations specific to MPLS-TP are addressed in this section.

A service provider may deploy the security options best fitting its network operations. This document does not mandate that MPLS-TP network operators must configure and use technical mechanisms to satisfy all of the security requirements listed in this document.

These requirements are focused on: 1) how to protect the MPLS-TP network from various attacks originating outside the trusted zone, including those from network users, both accidental and malicious; 2) prevention of operational errors resulting from misconfiguration within the trusted zone.

R01: MPLS-TP MUST support the physical and logical separation of the data plane from the control plane and the management plane. That is, if the control plane, management plane, or both are attacked and cannot function normally, the data plane should continue to forward packets without being impacted.

R02: MPLS-TP MUST support static provisioning of MPLS-TP LSPs and PWs without using control protocols (with or without a NMS). This is particularly important in cases where components of the provisioning process are not in the trusted zone (security model 2(a) and security model 2(b), where some or all T-PEs are not in the trusted zone and the inter-provider cases in security model 2(c), where the connecting S-PE is not in the trusted zone; see Figures 3, 4, and 5).
R03: MPLS-TP MUST support non-IP path options in addition to the IP loopback option.

R04: MPLS-TP MUST support authentication, integrity, and replay protection for any control protocol used in an MPLS-TP network.

R05: MPLS-TP SHOULD support confidentiality, algorithm agility, and key management for any control protocol used in an MPLS-TP network.

R06: MPLS-TP MUST support authentication, integrity, and replay protection for dynamic MPLS network inter-connection protocols.

R07: MPLS-TP SHOULD support confidentiality, algorithm agility, and key management for dynamic MPLS network inter-connection protocols.

R08: MPLS-TP MUST support mechanisms to prevent denial of service (DoS) attacks carried out over any control plane protocol or management protocol, including OAM and G-ACh, whether in-band or out-of-band. This applies to denial-of-service attacks against the control or management protocol itself or against the data channel.

R09: MPLS-TP MUST support hiding of the Service Provider’s infrastructure for all reference models whether using static configuration or a dynamic control plane.

R10: MPLS-TP MUST provide protection from operational errors. The extensive use of static provisioning with or without a NMS increases the likelihood of operational errors that result in misconfigurations that may compromise user’s data, system security, or network security is greater.

R11: MPLS-TP MUST support event logging and auditing. Logging and auditing capabilities provide critical resources for tracking down problems and repairing the damage after a security incident.

Management security requirements are covered in [RFC5951]. This document mandates protocol security, access controls, and protection against denial of service attacks for all management protocols. [RFC3871] contains guidelines on appropriately strong and open cryptography.

R12: MPLS-TP MUST support authentication, integrity, and replay protection for the management communication channel (MCC) and all network traffic and protocols used to support management functions. This includes protocols used for configuration, monitoring, Configuration backup, logging, time synchronization, authentication, and routing.

R13: MPLS-TP SHOULD support confidentiality, algorithm agility, and key management for the management communication channel (MCC) and all
network traffic and protocols used to support management functions. This includes protocols used for configuration, monitoring, Configuration backup, logging, time synchronization, authentication, and routing.

R14: The MCC MUST support access controls by protocol and port number.

5. Defensive Techniques for MPLS-TP Networks

The defensive techniques presented in this document are intended to describe methods by which some security threats can be addressed. They are not intended as requirements for all MPLS-TP deployments. The specific operational environment determines the security requirements for any instance of MPLS-TP. Therefore, protocol designers should provide a full set of security capabilities, which can be selected and used where appropriate. The MPLS-TP provider should determine the applicability of these techniques to the provider's specific service offerings, and the end user may wish to assess the value of these techniques to the user's service requirements.

The techniques discussed here include entity authentication for identity verification, encryption for confidentiality, message integrity and replay detection to ensure the validity of message streams, network-based access controls such as packet filtering and firewalls, host-based access controls, isolation, aggregation, protection against denial of service, and event logging. Where these techniques apply to MPLS and GMPLS in general, they are described in Section 5.2 of [RFC5920]. The remainder of this section covers aspects that apply particularly to MPLS-TP.

5.1. Authentication

To prevent security issues arising from impersonation, masquerade, some denial-of-service attacks, or from malicious or accidental misconfiguration, it is critical that MPLS-TP devices should accept connections or control messages only from known sources. Authentication refers to methods for ensuring that the identities of message sources are properly verified by the devices with which they communicate. This section focuses on scenarios in which sender authentication is required and recommends authentication mechanisms for these scenarios.

5.1.1. Management System Authentication

Management system authentication includes the authentication of a PE to a centrally-managed network management or directory server when directory-based auto-discovery is used. It also includes authentication of a CE to the configuration server when a configuration server system is used. This type of authentication should be bi-directional. The PE or CE needs to be certain it is communicating with the right server.
5.1.2. Peer-to-Peer Authentication

Peer-to-peer authentication includes peer authentication for network control protocols and other peer authentication (e.g., authentication of one IPsec security gateway by another).

Authentication should be bi-directional, including S-PE, T-PE, PE, or CE to authentication to a configuration server so that a PE or CE can be certain it is communicating with the right server.

5.1.3. Cryptographic Techniques for Authenticating Identity

Cryptographic techniques offer several mechanisms for authenticating the identity of devices or individuals. These include the use of shared secret keys, one-time keys generated by accessory devices or software, user-ID and password pairs, and a variety of public-private key systems. Some of these use digital certificates binding a user’s name and public key. One method of using digital certificates is within a hierarchical Certification Authority system.

5.2. Access Control Techniques

Many of the security issues related to management interfaces can be addressed through the use of authentication as described in Section 5.1. However, additional security may be provided by controlling access to management interfaces or to specific resources with an access control model. In addition to identification and authentication, access control deals with authorization.

SNMP security efforts have focused on access control models. For the Most recent version of SNMP security, see the work of the ISMS WG.

The Optical Internetworking Forum has worked on protecting interfaces to management systems with TLS, SSH, IPsec, WSS, etc. See Security for Management Interfaces to Network Elements [OIF-SMI-03.0].

Management interfaces, especially console ports on MPLS-TP devices, may be configured so they are only accessible out-of-band, through a system which is physically or logically separated from the rest of the MPLS-TP infrastructure.

Where management interfaces are accessible in-band within the MPLS-TP domain, filtering or firewalling techniques can be used to restrict unauthorized in-band traffic from having access to management interfaces. Depending on device capabilities, these filtering or firewalling techniques can be configured either on other devices through which the traffic might pass, or on the individual MPLS-TP devices themselves.
5.3. Use of Isolated Infrastructure

One way to protect the infrastructure used for support of MPLS-TP is to separate the resources for support of MPLS-TP services from the resources used for other purposes. For example, in security model 2 (Section 2.2), the potential risk of attacks on the S-PE or T-PE in the trusted zone may be reduced by using non-IP-based communication paths.

5.4. Use of Aggregated Infrastructure

In general, it is not feasible to use a completely separate set of resources for support of each service. In fact, one of the main reasons for MPLS-TP enabled services is to allow sharing of resources between multiple services and multiple users. Thus, even if certain services use a separate network from Internet services, nonetheless there will still be multiple MPLS-TP users sharing the same network resources.

In general, the use of aggregated infrastructure allows the service provider to benefit from stochastic multiplexing of multiple bursty flows, and also may in some cases thwart traffic pattern analysis by combining the data from multiple users. However, service providers must minimize security risks introduced from any individual service or individual users.

5.5. Protection against Denial of Service

It is possible to lessen the potential and impact of denial-of-service attacks by using secure protocols, turning off unnecessary processes, logging and monitoring, and using ingress filtering. See [RFC4732] for background on denial-of-service attacks in the context of the Internet.

5.6. Verification of Connectivity

To protect against deliberate or accidental misconnection, mechanisms can be put in place to verify both end-to-end connectivity and hop-by-hop resources. These mechanisms can trace the routes of LSPs in both the control plane and the data plane.

6. Monitoring, Detection, and Reporting of Security Attacks

MPLS-TP networks and services may be subject to attacks from a variety of security threats. Many types of threats are described in the Security Requirements (Section 4) section of this document. The defensive techniques described in this document and elsewhere provide significant levels of protection from many of these threats. However, in addition to employing defensive techniques silently to protect against attacks, MPLS-TP services can also add value for...
both providers and customers by implementing security monitoring systems to detect and report on any security attacks, regardless of whether the attacks are effective.

Attackers often begin by probing and analyzing defenses, so systems that can detect and properly report these early stages of attacks can provide significant benefits.

Information concerning attack incidents, especially if available quickly, can be useful in defending against further attacks. It can be used to help identify attackers or their specific targets at an early stage. This knowledge about attackers and targets can be used to strengthen defenses against specific attacks or attackers, or to improve the defenses for specific targets on an as-needed basis. Information collected on attacks may also be useful in identifying and developing defenses against novel attack types.

Also, extensive logging of normal processing, error conditions, and security events can be an invaluable source of information for tracking down attacks, recovering from them, and determining how to prevent future attacks. Different methods may be appropriate from case to case, and in fact comparing the same or similar information obtained in different ways (e.g., with syslog and SNMP) sometimes reveals subtle security flaws or actual intrusions. Implementations should also pay attention to the security of the logs themselves.

7. Security Considerations

Security considerations constitute the sole subject of this document and hence are discussed throughout.

The document describes a variety of defensive techniques that may be used to counter the potential threats. All of the techniques presented involve mature and widely implemented technologies that are practical to implement.

The document evaluates MPLS-TP security requirements from a customer’s perspective as well as from a service provider’s perspective. These sections re-evaluate the identified threats from the perspectives of the various stakeholders and are meant to assist equipment vendors and service providers, who must ultimately decide what threats to protect against in any given configuration or service offering.

8. IANA Considerations

This document contains no new IANA considerations.
9. References

9.1. Normative References


9.2. Informative References


Authors’ Addresses

Luyuan Fang (editor)
Cisco Systems
111 Wood Ave. South
Iselin, NJ 08830
US

Email: lufang@cisco.com

Ben Niven-Jenkins (editor)
Velocix
326 Cambridge Science Park
Milton Road
Cambridge CB4 0WG
UK

Email: ben@niven-jenkins.co.uk

Scott Mansfield (editor)
Ericsson
300 Holger Way
San Jose, CA 95134
US

Email: scott.mansfield@ericsson.com

Richard F. Graveman (editor)
RFG Security, LLC
15 Park Avenue
Morristown, NJ 07960 US

Email: rfg@acm.org

Raymond Zhang
British Telecom
BT Center
81 Newgate Street
London EC1A 7AJ
UK

Email: raymond.zhang@bt.com
Abstract

This document defines the method for transporting PTP messages (PDUs) over an MPLS network. The method allows for the easy identification of these PDUs at the port level to allow for port level processing of these PDUs in both LERs and LSRs.

The basic idea is to transport PTP messages inside dedicated MPLS LSPs. These LSPs only carry PTP messages and possibly Control and Management packets, but they do not carry customer traffic.

Two methods for transporting 1588 over MPLS are defined. The first method is to transport PTP messages directly over the dedicated MPLS LSP via UDP/IP encapsulation, which is suitable for IP/MPLS networks. The second method is to transport PTP messages inside a PW via Ethernet encapsulation, which is more suitable for MPLS-TP networks.

Status of this Memo

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

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

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

This Internet-Draft will expire on April 9, 2012.
Table of Contents

1.  Introduction ................................................. 6
2.  Terminology ............................................... 7
3.  Problem Statement ......................................... 8
4.  1588 over MPLS Architecture ............................... 9
5.  Dedicated LSPs for PTP messages ............................ 10
6.  1588 over MPLS Encapsulation .............................. 11
   6.1.  1588 over LSP Encapsulation ........................... 11
   6.2.  1588 over PW Encapsulation ............................ 11
7.  1588 Message Transport ..................................... 14
8.  Protection and Redundancy .................................. 16
9.  ECMP ......................................................... 17
10. OAM, Control and Management ............................... 18
11. QoS Considerations ......................................... 19
12. FCS Recalculation ......................................... 20
13. UDP Checksum Correction .................................... 21
14. Routing extensions for 1588aware LSRs ................. 22
   14.1. 1588aware Link Capability for OSPF .................. 22
   14.2. 1588aware Link Capability for IS-IS .................. 23
15. RSVP-TE Extensions for support of 1588 .................. 25
16. Behavior of LER/LSR ....................................... 26
   16.1. Behavior of 1588-aware LER ........................... 26
   16.2. Behavior of 1588-aware LSR ........................... 26
   16.3. Behavior of non-1588-aware LSR ...................... 26
17. Other considerations ...................................... 28
18. Security Considerations ................................... 29
19. Acknowledgements ......................................... 30
20. IANA Considerations ...................................... 31
The key words "MUST", "MUST NOT", "REQUIRED", "SHALL", "SHALL NOT", "SHOULD", "SHOULD NOT", "RECOMMENDED", "MAY", and "OPTIONAL" in this document are to be interpreted as described in RFC2119 [RFC2119].

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

The objective of Precision Time Protocol (PTP) is to synchronize independent clocks running on separate nodes of a distributed system. [IEEE] defines PTP messages for clock and time synchronization. The PTP messages include PTP PDUs over UDP/IP (Annex D and E of [IEEE]) and PTP PDUs over Ethernet (Annex F of [IEEE]). This document defines mapping and transport of the PTP messages defined in [IEEE] over MPLS networks.

PTP defines several clock types: ordinary clocks, boundary clocks, end-to-end transparent clocks, and peer-to-peer transparent clocks. One key attribute of all of these clocks is the recommendation for PTP messages processing to occur as close as possible to the actual transmission and reception at the physical port interface. This targets optimal time and/or frequency recovery by avoiding variable delay introduced by queues internal to the clocks. To facilitate the fast and efficient recognition of PTP messages at the port level when the PTP messages are carried over MPLS LSPs, this document defines the specific encapsulations that should be used. In addition, it can be expected that there will exist LSR/LEs where only a subset of the physical ports will have the port based PTP message processing capabilities. In order to ensure that the PTP carrying LSPs always enter and exit ports with this capability, routing extensions are defined to advertise this capability on a port basis and to allow for the establishment of LSPs that only transit such ports. While this path establishment restriction may be applied only at the LER ingress/egress ports, it becomes more important when using Transparent Clock capable LSRs in the path.

The port based PTP message processing involves PTP event message recognition. Once the PTP event messages are recognized they can be modified based on the reception or transmission timestamp. An alternative technique to actual packet modification could include the enforcement of a fixed delay time across the LSR to remove variability in the transit delay. This latter would be applicable in a LSR which does not contain a PTP transparent Clock function.

This document provides two methods for transporting PTP messages over MPLS. One is principally focused on an IP/MPLS environment and the second is focused on the MPLS-TP environment.

While the techniques included herein allow for the establishment of paths optimized to include PTP Timestamping capable links, the performance of the Slave clocks is outside the scope of this document.
2. Terminology

1588: The timing and synchronization as defined by IEEE 1588

PTP: The timing and synchronization protocol used by 1588

Master Clock: The source of 1588 timing to a set of slave clocks.

Master Port: A port on a ordinary or boundary clock that is in Master state. This is the source of timing toward slave ports.

Slave Clock: A receiver of 1588 timing from a master clock

Slave Port: A port on a boundary clock or ordinary clock that is receiving timing from a master clock.

Ordinary Clock: A device with a single PTP port.

Transparent Clock. A device that measures the time taken for a PTP event message to transit the device and then updates the correctionField of the message with this transit time.

Boundary Clock: A device with more than one PTP port. Generally boundary clocks will have one port in slave state to receive timing and then other ports in master state to re-distribute the timing.

PTP LSP: An LSP dedicated to carry PTP messages

PTP PW: A PW within a PTP LSP that is dedicated to carry PTP messages.

CW: Pseudowire Control Word

LAG: Link Aggregation

ECMP: Equal Cost Multipath

CF: Correction Field, a field inside certain PTP messages (message type 0-3) that holds the accumulative transit time inside intermediate switches
3. Problem Statement

When PTP messages are transported over MPLS networks, there is a need for PTP message processing at the physical port level. This requirement exists to minimize uncertainty in the transit delays. If PTP message processing occurs interior to the MPLS routers, then the variable delay introduced by queuing between the physical port and the PTP processing will add noise to the timing distribution. Port based processing applies at both the originating and terminating LERs and also at the intermediate LSRs if they support transparent clock functionality.

PTP messages over Ethernet or IP can always be tunneled over MPLS. However, there is a requirement to limit the possible encapsulation options to simplify the PTP message processing required at the port level. This applies to all 1588 clock types implemented in MPLS routers. But this is particularly important in LSRs that provide transparent clock functionality.

When 1588-awareness is needed, PTP messages should not be transported over LSPs or PWs that are carrying customer traffic because LSRs perform Label switching based on the top label in the stack. To detect PTP messages inside such LSPs requires special hardware to do deep packet inspection at line rate. Even if such hardware exists, the payload can’t be deterministically identified by LSRs because the payload type is a context of the PW label and the PW label and its context are only known to the Edge routers (PEs); LSRs don’t know what is a PW’s payload (Ethernet, ATM, FR, CES, etc.). Even if one restricts an LSP to only carry Ethernet PWs, the LSRs don’t have the knowledge of whether PW Control Word (CW) is present or not and therefore can’t deterministically identify the payload.

Therefore a generic method is defined in this document that does not require deep packet inspection at line rate, and can deterministically identify PTP messages. The defined method is applicable to both MPLS and MPLS-TP networks.
4. 1588 over MPLS Architecture

1588 communication flows map onto MPLS nodes as follows: 1588 messages are exchange between PTP ports on Ordinary and boundary clocks. Transparent clocks do not terminate the PTP messages but they do modify the contents of the PTP messages as they transit across the Transparent clock. SO Ordinary and boundary clocks would exist within LERs as they are the termination points for the PTP messages carried in MPLS. Transparent clocks would exist within LSRs as they do not terminate the PTP message exchange.

Perhaps a picture would be good here.
5. Dedicated LSPs for PTP messages

Many methods were considered for identifying the 1588 messages when they are encapsulated in MPLS such as by using GAL/ACH or a new reserved label. These methods were not attractive since they either required deep packet inspection and snooping at line rate or they required use of a scarce new reserved label. Also one of the goals was to reuse existing OAM and protection mechanisms.

The method defined in this document can be used by LER/LSRs to identify PTP messages in MPLS tunnels by using dedicated LSPs to carry PTP messages.

Compliant implementations MUST use dedicated LSPs to carry PTP messages over MPLS. These LSPs are herein referred to as "PTP LSPs" and the labels associated with these LSPs as "PTP labels". These LSPs could be P2P or P2MP LSPs. The PTP LSP between Master Clocks and Slave Clocks MAY be P2MP or P2P LSP while the PTP LSP between each Slave Clock and Master Clock SHOULD be P2P LSP. The PTP LSP between a Master Clock and a Slave Clock and the PTP LSP between the same Slave Clock and Master Clock MUST be co-routed. Alternatively, a single bidirectional co-routed LSP can be used. The PTP LSP MAY be MPLS LSP or MPLS-TP LSP. This co-routing is required to limit differences in the delays in the Master clock to Slave clock direction compared to the Slave clock to Master clock direction.

The PTP LSPs could be configured or signaled via RSVP-TE/GMPLS. New RSVP-TE/GMPLS TLVs and objects are defined in this document to indicate that these LSPs are PTP LSPs.

The PTP LSPs MAY carry essential MPLS/MPLS-TP control plane traffic such as BFD and LSP Ping but the LSP user plane traffic MUST be PTP only.
6. 1588 over MPLS Encapsulation

This document defines two methods for carrying PTP messages over MPLS. The first method is carrying IP encapsulated PTP messages over PTP LSPs and the second method is to carry PTP messages over dedicated Ethernet PWs (called PTP PWs) inside PTP LSPs.

6.1. 1588 over LSP Encapsulation

The simplest method of transporting PTP messages over MPLS is to encapsulate PTP PDUs in UDP/IP and then encapsulate them in PTP LSP. The 1588 over LSP format is shown in Figure 1.

```
+----------------------+
|   PTP Tunnel Label   |
+----------------------+
    | IPv4/6             |
+----------------------+
    | UDP                |
+----------------------+
    | PTP PDU            |
+----------------------+
```

Figure 1 - 1588 over LSP Encapsulation

This encapsulation is very simple and is useful when the networks between 1588 Master Clock and Slave Clock are IP/MPLS networks.

In order for an LSR to process PTP messages, the PTP Label must be the top label of the label stack.

The UDP/IP encapsulation of PTP MUST follow Annex D and E of [IEEE].

6.2. 1588 over PW Encapsulation

Another method of transporting 1588 over MPLS networks is by encapsulating PTP PDUs in Ethernet and then transporting them over Ethernet PW (PTP PW) as defined in [RFC4448], which in turn is transported over PTP LSPs. Alternatively PTP PDUs MAY be encapsulated in UDP/IP/Ethernet and then transported over Ethernet PW.

Both Raw and Tagged modes for Ethernet PW are permitted. The 1588 over PW format is shown in Figure 2.
The Control Word (CW) as specified in [RFC4448] SHOULD be used to ensure a more robust detection of PTP messages inside the MPLS packet. If CW is used, the use of Sequence number is optional.

The use of VLAN and UDP/IP are optional. Note that 1 or 2 VLANs MAY exist in the PW payload.

In order for an LSR to process PTP messages, the top label of the label stack (the Tunnel Label) MUST be from PTP label range. However in some applications the PW label may be the top label in the stack, such as cases where there is only one-hop between PEs or in case of PHP. In such cases, the PW label SHOULD be chosen from the PTP Label range.

In order to ensure congruency between the two directions of PTP message flow, ECMP should not be used for the PTP LSPs. Therefore, no Entropy label [I-D.ietf-pwe3-fat-pw] is necessary and it SHOULD NOT be present in the stack.

The Ethernet encapsulation of PTP MUST follow Annex F of [IEEE] and the UDP/IP encapsulation of PTP MUST follow Annex D and E of [IEEE].

For 1588 over MPLS encapsulations that are PW based, there are some cases in which the PTP LSP label may not be present:
o When PHP is applied to the PTP LSP, and the packet is received without PTP LSP label at PW termination point.

o When the PW is established between two routers directly connected to each other and no PTP LSP is needed.

In such cases it is required for a router to identify these packets as PTP packets. This would require the PW label to also be a label that is distributed specifically for carrying PTP traffic (aka PTP PW label). Therefore there is a need to add extension to LDP/BGP PW label distribution protocol to indicate that a PW label is a PTP PW labels.
7. 1588 Message Transport

1588 protocol comprises of the following message types:

- Announce
- SYNC
- FOLLOW_UP
- DELAY_REQ (Delay Request)
- DELAY_RESP (Delay Response)
- PDELAY_REQ (Peer Delay Request)
- PDELAY_RESP (Peer Delay Response)
- PDELAY_RESP_FOLLOW_UP (Peer Delay Response Follow up)
- Management
- Signaling

A subset of PTP message types that require timestamp processing are called Event messages:

- SYNC
- DELAY_REQ (Delay Request)
- PDELAY_REQ (Peer Delay Request)
- PDELAY_RESP (Peer Delay Response)

SYNC and DELAY_REQ are exchanged between Master Clock and Slave Clock and MUST be transported over PTP LSPs. PDELAY_REQ and PDELAY_RESP are exchanged between adjacent PTP clocks (i.e. Master, Slave, Boundary, or Transparent) and MAY be transported over single hop PTP LSPs. If Two Step PTP clocks are present, then the FOLLOW_UP, DELAY_RESP, and PDELAY_RESP_FOLLOW_UP messages must also be transported over the PTP LSPs.

For a given instance of 1588 protocol, SYNC and DELAY_REQ MUST be transported over two PTP LSPs that are in opposite directions. These PTP LSPs, which are in opposite directions MUST be congruent and co-routed. Alternatively, a single bidirectional co-routed LSP can be used.
Except as indicated above for the two-step PTP clocks, Non-Event PTP message types don’t need to be processed by intermediate routers. These message types MAY be carried in PTP Tunnel LSPs.
8. Protection and Redundancy

In order to ensure continuous uninterrupted operation of 1588 Slaves, usually as a general practice, Redundant Masters are tracked by each Slave. It is the responsibility of the network operator to ensure that physically disjoint PTP tunnels that don’t share any link are used between the redundant Masters and a Slave.

When redundant Masters are tracked by a Slave, any prolonged PTP LSP or PTP PW outage will trigger the Slave Clock to switch to the Redundant Master Clock. However LSP/PW protection such as Linear Protection Switching (1:1, 1+1), Ring protection switching or MPLS Fast Reroute (FRR) SHOULD still be used to provide resiliency to individual network segment failures.

Note that any protection or reroute mechanism that adds additional label to the label stack, such as Facility Backup Fast Reroute, MUST ensure that the pushed label is a PTP Label to ensure recognition of the MPLS frame as containing PTP messages as it transits the backup path.
9. ECMP

To ensure the optimal operation of 1588 Slave clocks and avoid errors introduced by forward and reverse path delay asymmetry, the physical path for PTP messages from Master Clock to Slave Clock and vice versa must be the same for all PTP messages listed in section 7 and must not change even in the presence of ECMP in the MPLS network.

To ensure the forward and reverse paths are the same PTP LSPs and PWs MUST NOT be subject to ECMP.
10. OAM, Control and Management

In order to manage PTP LSPs and PTP PWs, they MAY carry OAM, Control and Management messages. These control and management messages can be differentiated from PTP messages via already defined IETF methods.

In particular BFD [RFC5880], [RFC5884] and LSP-Ping [RFC4389] MAY run over PTP LSPs via UDP/IP encapsulation or via GAL/G-ACH. These Management protocols are easily identified by the UDP Destination Port number or by GAL/ACH respectively.

Also BFD, LSP-Ping and other Management messages MAY run over PTP PW via one of the defined VCCVs (Type 1, 2 or 3) [RFC5085]. In this case G-ACH, Router Alert Label (RAL), or PW label (TTL=1) are used to identify such management messages.
11. QoS Considerations

In network deployments where not every LSR/LER is PTP-aware, then it is important to reduce the impact of the non-PTP-aware LSR/LERs on the timing recovery in the slave clock. The PTP messages are time critical and must be treated with the highest priority. Therefore 1588 over MPLS messages must be treated with the highest priority in the routers. This can be achieved by proper setup of PTP tunnels. It is recommended that the PTP LSPs are setup and marked properly to indicate EF-PHB for the CoS and Green for drop eligibility.

In network deployments where every LSR/LER supports PTP LSPs, then it MAY NOT be required to apply the same level of prioritization as specified above.
12. FCS Recalculation

Ethernet FCS of the outer encapsulation MUST be recalculated at every LSR that performs the Transparent Clock processing and FCS retention for the payload Ethernet described in [RFC4720] MUST NOT be used.
13. UDP Checksum Correction

For UDP/IP encapsulation mode of 1588 over MPLS, the UDP checksum is optional when used for IPv4 encapsulation and mandatory in case of IPv6. When IPv4/v6 UDP checksum is used each 1588-aware LSR must either incrementally update the UDP checksum after the CF update or should verify the UDP checksum on reception from upstream and recalculate the checksum completely on transmission after CF update to downstream node.
14. Routing extensions for 1588-aware LSRs

MPLS-TE routing relies on extensions to OSPF [RFC2328] [RFC5340] and IS-IS [ISO] [RFC1195] in order to advertise Traffic Engineering (TE) link information used for constraint-based routing.

Indeed, it is useful to advertise data plane TE router link capabilities, such as the capability for a router to be 1588-aware. This capability MUST then be taken into account during path computation to prefer or even require links that advertise themselves as 1588-aware. In this way the path can ensure the entry and exit points into the LERs and, if desired, the links into the LSRs are able to perform port based timestamping thus minimizing their impact on the performance of the slave clock.

For this purpose, the following sections specify extensions to OSPF and IS-IS in order to advertise 1588 aware capabilities of a link.

14.1. 1588-aware Link Capability for OSPF

OSPF uses the Link TLV (Type 2) that is itself carried within either the Traffic Engineering LSA specified in [RFC3630] or the OSPFv3 Intra-Area-TE LSA (function code 10) defined in [RFC5329] to advertise the TE related information for the locally attached router links. For an LSA Type 10, one LSA can contain one Link TLV information for a single link. This extension defines a new 1588-aware capability sub-TLV that can be carried as part of the Link TLV.

The 1588-aware capability sub-TLV is OPTIONAL and MUST NOT appear more than once within the Link TLV. If a second instance of the 1588-aware capability sub-TLV is present, the receiving system MUST only process the first instance of the sub-TLV. It is defined as follows:

```
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
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|     Flags     |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|     Type      |     Length    |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|     Flags     |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
```

Figure 3: 1588-aware Capability TLV

Where:

Type, 16 bits: 1588-aware Capability TLV where the value is TBD
Length, 16 bits: Gives the length of the flags field in octets, and is currently set to 1.

Flags, 8 bits: The bits are defined least-significant-bit (LSB) first, so bit 7 is the least significant bit of the flags octet.

```
0 1 2 3 4 5 6 7
+---------------+
|   Reserved    |
|               |
|  Correction   |
+---------------+
```

Figure 4: Flags Format

Correction (C) field Update field, 1 bit: Setting the C bit to 1 indicates that the link is capable of recognizing the PTP event packets and can compensate for residence time by updating the PTP packet Correction Field. When this is set to 0, it means that this link cannot perform the residence time correction but is capable of performing MPLS frame forwarding of the frames with PTP labels using a method that support the end to end delivery of accurate timing. The exact method is not defined herein.

Reserved, 7 bits: Reserved for future use. The reserved bits must be ignored by the receiver.

The 1588-aware Capability sub-TLV is applicable to both OSPFv2 and OSPFv3.

14.2. 1588-aware Link Capability for IS-IS

The IS-IS Traffic Engineering [RFC3784] defines the intra-area traffic engineering enhancements and uses the Extended IS Reachability TLV (Type 22) [RFC5305] to carry the per link TE-related information. This extension defines a new 1588-aware capability sub-TLV that can be carried as part of the Extended IS Reachability TLV.

The 1588-aware capability sub-TLV is OPTIONAL and MUST NOT appear more than once within the Extended IS Reachability TLV or the Multi-Topology (MT) Intermediate Systems TLV (type 222) specified in [RFC5120]. If a second instance of the 1588-aware capability sub-TLV is present, the receiving system MUST only process the first instance of the sub-TLV.

The format of the IS-IS 1588-aware sub-TLV is identical to the TLV format used by the Traffic Engineering Extensions to IS-IS [RFC3784]. That is, the TLV is comprised of 1 octet for the type, 1 octet
specifying the TLV length, and a value field. The Length field defines the length of the value portion in octets.

```
 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1 2 3
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|     Type      |     Length    |    Flags      |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
Figure 5: 1588-aware Capability sub-TLV
```

Where:

- **Type**, 8 bits: 1588-aware Capability sub-TLV where the value is TBD
- **Length**, 8 bits: Gives the length of the flags field in octets, and is currently set to 1
- **Flags**, 8 bits: The bits are defined least-significant-bit (LSB) first, so bit 7 is the least significant bit of the flags octet.

```
 0 1 2 3 4 5 6 7
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|   Reserved  |C|   
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
Figure 6: Flags Format
```

Correction (C) field Update field, 1 bit: Setting the C bit to 1 indicates that the link is capable of recognizing the PTP event packets and can compensate for residence time by updating the PTP packet Correction Field. When this is set to 0, it means that this link cannot perform the residence time correction but is capable of performing MPLS frame forwarding of the frames with PTP labels using a method that support the end to end delivery of accurate timing. The exact method is not defined herein.

- **Reserved**, 7 bits: Reserved for future use. The reserved bits must be ignored by the receiver.
15. RSVP-TE Extensions for support of 1588

RSVP-TE signaling MAY be used to setup the PTP LSPs. A new RSVP object is defined to signal that this is a PTP LSP. The OFFSET to the start of the PTP message header MAY also be signaled. Implementations can trivially locate the correctionField (CF) location given this information. The OFFSET points to the start of the PTP header as a node may want to check the PTP messageType before it touches the correctionField (CF). The OFFSET is counted from TBD.

The LSRs that receive and process the RSVP-TE/GMPLS messages MAY use the OFFSET to locate the start of the PTP message header.

Note that the new object/TLV Must be ignored by LSRs that are not compliant to this specification.

The new RSVP 1588_PTP_LSP object should be included in signaling PTP LSPs and is defined as follows:

```
+-------------+-------------+-------------+-------------+
|       Length (bytes)      |  Class-Num  |   C-Type    |
+-------------+-------------+-------------+-------------+
| Offset to locate the start of the PTP message header  |
+-------------+-------------+-------------+-------------+
```

Figure 7: RSVP 1588_PTP_LSP object

The ingress LSR MUST include this object in the RSVP PATH Message. It is just a normal RSVP path that is exclusively set up for PTP messages.
16. Behavior of LER/LSR

16.1. Behavior of 1588-aware LER

A 1588-aware LER advertises its 1588-awareness via the OSPF procedure explained in earlier section of this specification. The 1588-aware LER then signals PTP LSPs by including the 1588_PTP_LSP object in the RSVP-TE signaling.

When a 1588 message is received from a non-MPLS interface, the LER MUST redirect them to a previously established PTP LSP. When a 1588 over MPLS message is received from an MPLS interface, the processing is similar to 1588-aware LSR processing.

16.2. Behavior of 1588-aware LSR

1588-aware LSRs are LSRs that understand the 1588_PTP_LSP RSVP object and can perform 1588 processing (e.g. Transparent Clock processing).

A 1588-aware LSR advertises its 1588-awareness via the OSPF procedure explained in earlier section of this specification.

When a 1588-aware LSR distributes a label for PTP LSP, it maintains this information. When the 1588-aware LSR receives an MPLS packet, it performs a label lookup and if the label lookup indicates it is a PTP label then further parsing must be done to positively identify that the payload is 1588 and not OAM, BFD or control and management. Ruling out non-1588 messages can easily be done when parsing indicates the presence of GAL, ACH or VCCV (Type 1, 2, 3) or when the UDP port number does not match one of the 1588 UDP port numbers.

After a 1588 message is positively identified in a PTP LSP, the PTP message type indicates whether any timestamp processing is required. After 1588 processing the packet is forwarded as a normal MPLS packet to downstream node.

16.3. Behavior of non-1588-aware LSR

It is most beneficial that all LSRs in the path of a PTP LSP be 1588-aware LSRs. This would ensure the highest quality time and clock synchronization by 1588 Slave Clocks. However, this specification does not mandate that all LSRs in path of a PTP LSP be 1588-aware.

Non-1588-aware LSRs are LSRs that either don’t have the capability to process 1588 packets (e.g. perform Transparent Clock processing) or don’t understand the 1588_PTP_LSP RSVP object.

Non-1588-aware LSRs ignore the RSVP 1588_PTP_LSP object and just
switch the MPLS packets carrying 1588 messages as data packets and
don't perform any timestamp related processing. However as explained
in QoS section the 1588 over MPLS packets MUST be still be treated
with the highest priority.
17. Other considerations

The use of Explicit Null (Label= 0 or 2) is acceptable as long as either the Explicit Null label is the bottom of stack label (applicable only to UDP/IP encapsulation) or the label below the Explicit Null label is a PTP label.

The use of Penultimate Hop Pop (PHP) is acceptable as long as either the PHP label is the bottom of stack label (applicable only to UDP/IP encapsulation) or the label below the PHP label is a PTP label.
18. Security Considerations

MPLS PW security considerations in general are discussed in [RFC3985] and [RFC4447], and those considerations also apply to this document.

An experimental security protocol is defined in [IEEE]. The PTP security extension and protocol provides group source authentication, message integrity, and replay attack protection for PTP messages.
19. Acknowledgements

The authors would like to thank Luca Martini, Ron Cohen, Yaakov Stein, Tal Mizrahi and other members of the TICTOC WG for reviewing and providing feedback on this draft.
20. IANA Considerations

20.1. IANA Considerations for OSPF

IANA has defined a sub-registry for the sub-TLVs carried in an OSPF TE Link TLV (type 2). IANA is requested to assign a new sub-TLV codepoint for the 1588aware capability sub-TLV carried within the Router Link TLV.

<table>
<thead>
<tr>
<th>Value</th>
<th>Sub-TLV</th>
<th>References</th>
</tr>
</thead>
<tbody>
<tr>
<td>TBD</td>
<td>1588aware node sub-TLV</td>
<td>(this document)</td>
</tr>
</tbody>
</table>

20.2. IANA Considerations for IS-IS

IANA has defined a sub-registry for the sub-TLVs carried in the IS-IS Extended IS Reacability TLV. IANA is requested to assign a new sub-TLV code-point for the 1588aware capability sub-TLV carried within the Extended IS Reacability TLV.

<table>
<thead>
<tr>
<th>Value</th>
<th>Sub-TLV</th>
<th>References</th>
</tr>
</thead>
<tbody>
<tr>
<td>TBD</td>
<td>1588aware node sub-TLV</td>
<td>(this document)</td>
</tr>
</tbody>
</table>

20.3. IANA Considerations for RSVP

IANA is requested to assign a new Class Number for 1588 PTP LSP object that is used to signal PTP LSPs.

1588 PTP LSP Object

Class-Num of type 11bbbbbb

Suggested value TBD

Defined CType: 1 (1588 PTP LSP)
21. References

21.1. Normative References


21.2. Informative References


Authors’ Addresses

Shahram Davari
Broadcom Corp.
San Jose, CA 95134
USA

Email: davari@broadcom.com

Amit Oren
Broadcom Corp.
San Jose, CA 95134
USA

Email: amito@broadcom.com

Manav Bhatia
Alcatel-Lucent
Bangalore,
India

Email: manav.bhatia@alcatel-lucent.com

Peter Roberts
Alcatel-Lucent
Kanata,
Canada

Email: peter.roberts@alcatel-lucent.com

Laurent Montini
Cisco Systems
San Jose CA
USA

Email: lmontini@cisco.com
Abstract

This document presents a set of requirements for the establishment
and maintenance of Receiver-Driven Point-to-Multipoint (P2MP) and
Multipoint-to-Multipoint (MP2MP) Traffic-Engineered (TE)
Multiprotocol Label Switching (MPLS) Label Switched Paths (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 September 13, 2012.

Copyright Notice

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

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

1. Introduction ...................................................... 3
   1.1. Motivation .................................................. 3
   1.2. Terminology ............................................... 4
   1.3. What is not covered in this doc? ......................... 4
2. Basic Requirements .............................................. 5
3. Leaf Driven mRSVP-TE LSP Examples ............................. 8
4. Detailed Requirements ........................................... 10
   4.1. Leaf Driven mRSVP-TE LSP ............................... 10
   4.2. P2MP TE LSP Establishment, Teardown, and Modification
        Mechanisms ................................................. 10
   4.3. Re-Optimization of Leaf Driven mRSVP-TE LSPs .......... 11
   4.4. Support for LAN interfaces ............................. 11
   4.5. P2MP/MP2MP MPLS Label ................................... 12
   4.6. Advertisement of Leaf Driven mRSVP-TE Capability ..... 12
   4.7. Scalability ............................................... 12
   4.8. Variation of LSP Parameters ............................. 14
5. In-Band Signalling .............................................. 14
6. Backwards Compatibility ........................................ 14
7. Acknowledgements ............................................... 14
8. IANA Considerations ............................................ 14
9. Security Considerations ........................................ 15
10. References ..................................................... 15
   10.1. Normative References .................................... 15
   10.2. Informative References .................................. 17
Authors’ Addresses ................................................. 17
1. Introduction

For content delivery services that rely upon the IP multicast transmission scheme, the distribution trees are receiver-initiated. The delivery of such services over MPLS networking infrastructures may rely upon P2MP LSP tree structures that are currently source-initiated, with the root of the P2MP tree being at the LSR router directly connected to the sender ([RFC4875]).

Multiparty multimedia applications are getting greater attention in the telecom world. Such applications are QoS-demanding and can therefore benefit from the activation of MPLS traffic engineering capabilities that lead to the dynamic computation and establishment of MPLS LSPs whose characteristics comply with application-specific QoS requirements.

P2MP-RE-REQ[RFC4461] specifies the signalling requirements to setup multipoint LSPs from source to receiver. P2MP-TE [RFC4875] defines the procedure to setup multipoint LSPs from source to receiver.

This document presents a set of requirements specific for the dynamic establishment and maintenance of receiver-driven P2MP-TE and MP2MP-TE LSPs.

1.1. Motivation

IP multicast distribution trees are receiver-initiated and dynamic by nature. IP multicast-enabled applications are also bandwidth savvy, especially in the area of residential Live TV broadcasting services, where several hundreds of thousands of IPTV receivers need to be served with the appropriate level of quality. Current source-driven P2MP LSP establishment assumes a prior knowledge of receiver(s) location for the sake of the P2MP LSP tree structure’s forwarding efficiency. But the receiver’s location information is not available a priori for the root MPLS router to compute and establish the relevant P2MP tree structure. In addition, with the source-driven P2MP-TE solution, full mesh P2MP LSP tree structures need to be established to setup a MP2MP LSP, possibly raising scalability issues for the delivery of some multicast services, such as videoconferencing at large scale.

From this perspective, it is believed that receiver-driven MPLS P2MP tree structures represent the best of breed that combines the benefits of IP multicast, receiver-driven dynamics with the benefits of MPLS-based, QoS-guaranteed LSP path computation.

Thus, receiver-driven MPLS P2MP/MP2MP tree structure do not require that the root of the MPLS P2MP tree structure dynamically discovers
and maintains receiver information a priori. In addition, such a receiver-driven approach encourages computation schemes that can take into account the network access conditions of the receivers (bandwidth capabilities, IPTV customer profile, etc.) for the sake of customer’s Quality of Experience (time to access an IPTV channel, time to zap from one channel to another, etc.).

1.2. Terminology

The following terms are used in this document:

- **Sender**: Sender refers to the source of the content/payload. As in [RFC2205].
- **Receiver**: Receiver refers to the Receiver of the content/payload. As in [RFC2205].
- **Upstream**: The direction of a flow from a content Receiver towards a content Sender. As defined in [RFC2205].
- **Downstream**: The direction of a flow from a content Sender towards a content Receiver. As defined in [RFC2205].
- **Path-Sender**: The sender of the RSVP_PATH message, with NO correlation to the directionality of the corresponding content flow.
- **Path-Receiver**: The receiver of the RSVP_PATH message, with NO correlation to the directionality of the corresponding content flow.
- **Path-Initiator**: The Path-Sender that originated the RSVP_PATH message. The Path-Initiator is a notion that differs from the Path-Sender because an intermediate node can be a Path-Sender, and therefore different from the node that created and initiated the RSVP_PATH message, which is precisely the Path-Initiator.
- **Path-Terminator**: The Path-Receiver that does NOT propagate the Path message. This is a notion that differs from the Path-Receiver because an intermediate node can be a Path-Receiver.

1.3. What is not covered in this doc?

This document does not specify any requirements for the following functions.
2. Basic Requirements

[RFC4461] specifies requirements for P2MP traffic engineering over MPLS. It describes sender-driven P2MP traffic engineering in detail. Those definitions and requirements are equally applicable to receiver-driven traffic engineering in a point-to-multipoint service environment and are therefore not repeated here.

Following are key requirements that are specific to the receiver-driven paradigm:

REQ1: The receiver-driven P2MP/MP2MP approach must be scalable, i.e., the dynamic computation and maintenance of P2MP/MP2MP tree structures must be adapted to typical live TV broadcasting services that connect several hundreds of thousands of receivers.

REQ2: Leaves of receiver-driven P2MP/MP2MP tree structures must be aware of the corresponding P2MP/MP2MP LSP identifiers for tree computation and maintenance purposes.

REQ3: The receiver-driven mechanism MUST allow the dynamic addition and removal of leaves to and from a P2MP/MP2MP tree structure, without any restriction (provided there is network connectivity). It is RECOMMENDED that the corresponding operations be leaf-initiated.
REQ4: The dynamic addition and removal of leaves to and from a receiver-driven P2MP/MP2MP LSP MUST not impact the performances of the data forwarding (packet loss, replication, delay) towards other leaves.

REQ5: The dynamic addition and removal of leaves to and from a receiver-driven P2MP/MP2MP LSP SHOULD NOT infer any additional processing than that on the path from the added/removed Leaf LSR to the Branch LSR.

REQ6: A Leaf router MUST initiate RSVP PATH message towards the root and also signal the type of the LSP.

REQ7: The destination of a L2S (Leaf-to-Source) sub-path MUST be one of the ingress routers where multicast data sent by a source enter the P2MP/MP2MP LSP tree structure.

REQ8: RSVP P2MP PATH messages MUST be forwarded along the path from a given leaf to the root of the P2MP tree structure.

REQ9: RSVP P2MP RESV messages MUST be forwarded along the path from the root to the receiver if the leaf is the MPLS router that directly connects receivers to the corresponding receiver-driven P2MP tree structure. Otherwise, RSVP P2MP RESV messages will be forwarded from the PATH Terminator to the PATH Initiator.

REQ10: A node receiving a RSVP RESV message SHOULD interpret it as a successful resource reservation from the upstream node for the establishment of the P2MP tree structure.

REQ11: A node receiving a RSVP RESV message SHOULD interpret it as a successful resource reservation from the downstream node for the establishment of the MP2MP tree structure.

REQ12: Label allocation on incoming interface MUST be done prior to sending RSVP PATH messages upstream for P2MP tree structures.

REQ13: Label allocation on incoming interface MUST be done prior to sending RSVP RESV messages upstream for MP2MP tree structures.

REQ14: For P2MP LSP tree structures, a node receiving a RSVP PATH message MUST first decide if this RSVP PATH message will make itself a branch LSR or not. In the case that it will become a transit LSR because of this PATH message, then it will allocate required resources on the interface through which the RSVP PATH message is received, before sending the RSVP
PATH message upstream. So that the upstream node can send traffic soon after successfully reserving resources on the downstream link, on which the RSVP PATH message SHOULD be received. In the case that the node is a branch or transit node already before it receives the PATH message, then it will allocate required resources (provided they are available) on the interface through which the RSVP PATH message is received, and then send the RESV message back to the node that sent the PATH message without sending the PATH message upstream.

REQ15: For MP2MP LSP tree structures, a node will allocate required resources on the interface through which the RSVP PATH message is sent, before sending the RSVP PATH message upstream. A node receiving a RSVP PATH message MUST first decide if this RSVP PATH message will make itself a branch LSR or not. In the case that it will become a transit LSR because of this PATH message, and then allocate required resources (provided they are available) on the interface through which the RSVP PATH message was received and will allocate required resources on the interface through which the RSVP PATH message is sent, before sending it upstream. So that the downstream node can send traffic soon after successfully reserving resources on the upstream link, on which the RSVP PATH message SHOULD be sent. The upstream node can then send traffic soon after successfully reserving resources on the downstream link, on which the RSVP PATH message SHOULD be received. In the case that the node is a branch or transit node already before it receives the PATH message, then it will allocate required resources on the interface through which the RSVP PATH message is received, and send the RESV message to the node which sent the PATH message without sending the PATH message upstream.
3. Leaf Driven mRSVP-TE LSP Examples

Figure 1: P2MP Example

Figure 1 shows that R5 is added as the first leaf of the RD P2MP TE LSP, the message flow goes from R5->msg1->R3->msg2->R1->msg3->R3->msg4->R5. When the leaf R4 is added, the message flow goes from R4->msg5->R3->msg6->R4. In this case, when R3 receives msg5, R3 finds out that the RD P2MP LSP has already been set up for R5: therefore, R3 finds itself a branch node for leaf R4 and R5, so it will terminate the PATH message and build the corresponding RESV message and send it back to R4. The association of the LSP initiated by R4 to the existing RD P2MP LSP is determined based on the processing of the session object from the mRSVP-TE message. This session object is further documented in Section 2 of this draft.
Figure 2: MP2MP Example

Figure 2 shows that R5 is added as the first leaf (as both a sender and a receiver) of the RD MP2MP TE LSP, the message flow goes from R5->msg1->R3->msg2->R1->msg3->R3->msg4->R5. When the leaf R4 (both receiver and sender) is added, the message flow goes from R4->msg5->R3->msg6. In this case, when R3 receives msg5, R3 finds out that the RD MP2MP LSP has already been set up for R5, and R3 will become the branch LSR for the leaf R4 and R5, so it will terminate the PATH message, build a RESV message and send the RESV message back to R4. The association of the LSP initiated by R4 to the existing MP2MP LSP is determined based on the processing of the session object from the mRSVP-TE message. This session object is further documented in Section 2 of this draft.
4. Detailed Requirements

4.1. Leaf Driven mRSVP-TE LSP

The Receiver-Driven RSVP-TE extensions MUST be applicable to the signaling of P2MP LSPs for different switching types. For example, it MUST be possible to signal a P2MP/MP2MP TE LSP in any switching medium, whether it is packet or non-packet based (including frame, cell, TDM, lambda, etc.).

As with P2P MPLS technology [RFC3031], a given traffic is associated with a FEC in this extension. All packets that belong to a particular FEC and that travel from a particular node MUST follow the same Receiver-Driven P2MP/MP2MP tree structure.

In order to scale to a large number of branches, P2MP/MP2MP TE LSPs SHOULD be identified by a unique identifier (the P2MP/MP2MP ID) that is constant for the whole LSP regardless of the number of branches and/or leaves.

4.2. P2MP TE LSP Establishment, Teardown, and Modification Mechanisms

The Receiver-Driven P2MP LSP approach MUST support the dynamic establishment, maintenance, and teardown of Receiver-Driven P2MP/MP2MP LSPs in a manner that the scalability is not affected by the number of leaves, and it is scalable in a linear way to the number of branches for a node. This MUST imply the ability to compute multiple P2MP tree structures at once, and to compute P2MP/MP2MP LSP tree structures that are composed of many leaves, i.e., within the magnitude of typical Live TV broadcasting designs that accommodate several hundreds of thousands of receivers.

In addition to signaling capabilities for P2MP/MP2MP LSP establishment and teardown, the solution SHOULD support capabilities that allow the dynamic modification of part of a given P2MP/MP2MP tree structure without altering the whole structure:

For the purpose of adding sub-P2MP TE LSPs to an existing P2MP/MP2MP TE LSP, the extensions SHOULD support a grafting mechanism. For the purpose of deleting a sub-P2MP TE LSP from an existing P2MP/MP2MP TE LSP, the extensions SHOULD support a pruning mechanism.

It is RECOMMENDED that these grafting and pruning operations cause no additional processing in nodes that are not along the path from the grafting or pruning node to the upstream branch node. Moreover, both grafting and pruning operations MUST NOT disrupt the forwarding of traffic along the P2MP/MP2MP tree at any given time.
There is no assumption that the explicitly routed P2MP/MP2MP LSP remains on an optimal path after several grafts and prunes have occurred. In this context, scalability considerations refer to the signaling process for the P2MP/MP2MP TE LSP. The TE nature of the LSP allows that re-optimization may take place from time to time to restore the optimality of the LSP.

4.3. Re-Optimization of Leaf Driven mRSVP-TE LSPs

The detection of a more optimal path (for example, one with a lower overall cost) is an example of a situation where re-routing of the Receiver-Driven P2MP/MP2MP LSP may be required. While re-routing is in progress, an important requirement is to avoid double bandwidth reservation (over the common parts between the old and new LSP) through the use of resource sharing.

A Make-before-break design MUST be supported for Receiver-Driven P2MP/MP2MP LSPs to ensure that there is minimal traffic disruption during the re-routing operations.

Make-before-break that only applies to a sub-P2MP tree without impacting the data on all the other parts of the P2MP tree MUST be supported.

The solution SHOULD allow for make-before-break re-optimization of any subdivision of the P2MP LSP. It SHOULD do so by not having any signaling impact on the rest of the P2MP LSP, and without affecting the ability of the management plane to manage the LSP.

The solution SHOULD also provide the ability for any downstream LSR to have control over the re-optimization process.

Where sub-LSP re-optimization is allowed by the ingress LSR, such re-optimization MAY be initiated by a downstream LSR that is the root of the sub-LSP to be re-optimized. Sub-LSP re-optimization initiated by a downstream LSR MUST be carried out with the same regard to minimizing the impact on active traffic as mentioned above for other re-optimization purposes.

4.4. Support for LAN interfaces

Receiver-Driven P2MP/MP2MP LSPs may be used to traverse network segments that are provided by multi-access media such as Ethernet. In these contexts, it is also possible that the entry point to the network segment is a branch LSR of the Receiver-Driven P2MP/MP2MP LSP.

To avoid all replicated data are sent through the same port and
carried on the same segment, the solution SHOULD provide a mechanism for a branch LSR to send a single copy of the data onto a multi-access network to reach multiple (adjacent) downstream nodes.

The Receiver-Driven P2MP/MP2MP computation mechanism SHOULD provide a means for a Branch LSR to send a single copy of the data onto an Ethernet LAN interface to reach multiple adjacent downstream nodes. This requires that the same label be negotiated with all downstream LSRs for the P2MP/MP2MP LSP.

When there are several candidate upstream LSRs on a LAN interface, the RD P2MP TE mechanism SHOULD provide a means for all downstream LSRs of a given P2MP LSP to select the same upstream LSR, so as to avoid traffic replication. In addition, the RD P2MP TE mechanism SHOULD allow for an efficient balancing of a set of P2MP LSPs among a set of candidate upstream LSRs on a LAN interface.

4.5. P2MP/MP2MP MPLS Label

A solution for the dynamic establishment and maintenance of Receiver-Driven P2MP LSPs MUST allow the continued use of existing techniques to establish P2P and legacy P2MP LSPs (TE and otherwise) within the same network, and MUST allow the coexistence of Receiver-Driven P2MP/MP2MP LSPs with P2P and legacy P2MP LSPs within the same network.

A solution for the dynamic establishment and maintenance of Receiver-Driven P2MP LSPs MUST be specified in such a way that it allows legacy P2MP and P2P TE LSPs to be signaled on the same interface.

4.6. Advertisement of Leaf Driven mRSVP-TE Capability

To facilitate the computation of Receiver-Driven P2MP trees using TE constraints within a network whose LSRs do not all support the same capability level with respect to Receiver-Driven P2MP RSVP-TE signaling and data forwarding, the capability of an LSR to support the RSVP-TE signaling and forwarding for Receiver-Driven P2MP LSPs MUST be advertised to its neighbor LSRs.

4.7. Scalability

Scalability is a key requirement in mRSVP-TE systems. Solutions MUST be designed to scale well as a function of the number of any of the following:

- the number of receivers
- the number of sources
o the number of egress LSRs

o the number of branch LSRs

o the number of branches

Furthermore, scalability of control plane operation (setup, maintenance, modification, and teardown) MUST be considered.

Key considerations MUST include:

o The amount of refresh processing associated with maintaining a P2MP/MP2MP TE LSP.

o The amount of protocol state that must be maintained by ingress and transit LSRs along a P2MP/MP2MP tree.

o The number of protocol messages required to set up or tear down a P2MP/MP2MP LSP as a function of the number of egress LSRs.

o The number of protocol messages required to repair a P2MP/MP2MP LSP after failure or to perform make-before-break.

o The amount of protocol information transmitted to manage a P2MP/MP2MP TE LSP (i.e., the message size).

o The amount of additional data distributed in potential routing extensions.

o The amount of additional control plane processing required in the network to detect whether an add/delete of a new branch is required, and in particular, the amount of processing in steady state when no add/delete is requested.

o The amount of control plane processing required by the ingress, transit, and egress LSRs to add/delete a branch LSP to/from an existing P2MP LSP.

It is expected that the applicability of each solution will be evaluated with regards to the aforementioned scalability criteria.

In order to accommodate a growing number of leaves, it is RECOMMENDED that the amount of a P2MP LSP/MP2MP state on a LSR, for one particular LSP, depends only on the number of adjacent LSRs on the LSP.
4.8. Variation of LSP Parameters

Certain parameters (such as priority and bandwidth) are associated with an LSP. The parameters are installed by the signaling exchanges associated with the establishment and the maintenance of the P2MP/MP2MP LSP.

Any solution MUST NOT allow for variance of these parameters within a single P2MP/MP2MP LSP. That is:

- Downstream or upstream LSRs MUST NOT alter the attributes set and signaled by a first leaf router of a P2MP/MP2MP tree structure.
- A consistent QoS policy SHOULD be enforced from the root to all leaves of a single P2MP/MP2MP LSP.
- Some leaves of a given tree may yield the enforcement of a different QoS policy, depending on the various access capabilities of the receivers. Still, content will be delivered to these receivers by using the same (core) P2MP/MP2MP tree structure.
- Changing the parameters for the whole tree MAY be supported, but the change MUST apply to the whole tree from ingress LSR to all egress LSRs.

5. In-Band Signalling

TBD.

6. Backwards Compatibility

Any P2MP/MP2MP TE LSP solution SHOULD offer as much backward compatibility as possible. Also, RD P2MP TE LSPs MUST be able to coexist with IP unicast and IP multicast networks.

7. Acknowledgements

We would like to thank authors of [RFC4461] and the authors of [RFC6348] from which some text of this document has been inspired.

8. IANA Considerations

This memo includes no request to IANA.
9. Security Considerations

This document does not define any protocol extensions and does not, therefore, make any changes to any security models. It is a requirement that any RD P2MP/MP2MP solution developed to meet some or all of the requirements expressed in this document MUST include mechanisms to enable the secure establishment and management of P2MP/MP2MP RSVP-TE LSPs. This includes, but is not limited to:

- A receiver MUST be authenticated before it is allowed to trigger the establishment of an additional leaf of a RD P2MP LSP tree structure, in addition to hop-by-hop security issues identified in RFC 3209 and RFC 4206.
- mechanisms that provide some guarantees about the identity of an ingress LSR of a P2MP/MP2MP LSP;
- mechanisms to ensure that communicating signaling entities can verify each other’s identities;
- mechanisms to ensure that control plane messages are protected against spoofing and tampering;
- mechanisms to ensure that unauthorized leaves or branches are not added to the P2MP/MP2MP LSP; and mechanisms to protect signaling messages from snooping.

Note that RD P2MP/MP2MP signaling mechanisms built on P2P RSVP-TE signaling and RSVP-TE P2MP signalling are likely to inherit all the security techniques and problems associated with RSVP-TE. These problems may be exacerbated in P2MP/MP2MP situations where security relationships may need to be maintained between an ingress LSR and multiple egress LSRs. Such issues are similar to security issues for IP multicast.

10. References

10.1. Normative References


10.2. Informative References


Authors’ Addresses

Christian Jacquenet
France Telecom Orange
4 rue du Clos Courtel
35512 Cesson Sevigne
France

Phone: +33 2 99 12 43 82
Email: christian.jacquenet@orange-ftgroup.com

Quintin Zhao
Huawei Technology
125 Nagog Park
Acton, MA  01919
US

Phone:
Email: qzhao@huawei.com
Abstract

This document describes extensions to Resource Reservation Protocol - Traffic Engineering (RSVP-TE) for the setup of Receiver-Driven Traffic Engineered (TE) point-to-multipoint (P2MP) and multipoint-to-multipoint (MP2MP) Label Switched Paths (LSPs) in Multi-Protocol Label Switching (MPLS) and Generalized MPLS (GMPLS) networks.

Status of this Memo

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

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

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

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

Copyright Notice

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

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

1.  Introduction .................................................. 4
  1.1.  Motivation .................................................. 4
  1.2.  Terminology ................................................ 4
  1.3.  Overview .................................................... 5
  1.4.  Receiver-Driven mRSVP-TE LSP Examples ........................ 8
2.  Signaling Protocol Extensions ................................ 10
  2.1.  L2S Sub-LSPs ............................................... 11
  2.2.  Explicit Routing ............................................ 12
  2.3.  L2S Sub-LSPs and Path Messages ............................. 12
  2.4.  Resv Message Extensions ................................... 13
  2.5.  PathErr Message Extensions ................................ 14
  2.6.  ResvErr Message Extensions ................................ 14
  2.7.  PathTear Message Extensions ................................ 14
  2.8.  SESSION Object .............................................. 15
  2.8.1.  mRSVP-TE LSP Tunnel IPv4 SESSION Object ................. 16
  2.8.2.  mRSVP-TE LSP Tunnel IPv6 SESSION Object ............... 16
  2.9.  SENDER_TEMPLATE Object .................................... 17
  2.9.1.  mRSVP-TE LSP Tunnel IPv4 SENDER_TEMPLATE Object ........ 17
  2.9.2.  mRSVP-TE LSP Tunnel IPv6 SENDER_TEMPLATE Object ........ 17
  2.10. L2S_SUB_LSP Object ........................................ 18
   2.10.1. L2S_SUB_LSP IPv4 Object ............................... 18
   2.10.2. L2S_SUB_LSP IPv6 Object ............................... 19
  2.11. FILTER_SPEC Object ......................................... 19
   2.11.1. P2MP LSP_IPv4 FILTER_SPEC Object .................... 19
   2.11.2. mRSVP-TE LSP_IPv6 FILTER_SPEC Object ................ 19
   2.11.3. mRSVP-TE SECONDARY_EXPLICIT_ROUTE Object (SERO) ... 20
   2.11.4. mRSVP-TE SECONDARY_RECORD_ROUTE Object (SRRO) ... 20
3.  In-Band and Out-Band Signalling ................................ 20
  3.1.  In-Band Signalling ........................................ 20
  3.2.  Out-Band Signalling ....................................... 20
4.  Broadcast Interfaces ........................................... 21
5.  Fast Re-Route Considerations .................................. 21
6.  Backward Compatibility ........................................ 21
7.  Acknowledgements ............................................. 21
8.  IANA Considerations ........................................... 21
9.  Security Considerations ....................................... 21
10. References .................................................... 22
  10.1.  Normative References ..................................... 22
  10.2.  Informative References ................................... 23
Authors’ Addresses .................................................. 24

1. Introduction

Multiparty multimedia applications are getting greater attention in the telecom world. Such applications are QoS-demanding and can therefore benefit from the activation of MPLS traffic engineering capabilities that lead to the dynamic computation and establishment of MPLS LSPs whose characteristics comply with application-specific QoS requirements. P2MP-TE [RFC4875] defines a procedure to set up point-to-multipoint LSPs from sender to receivers. This document extends RSVP-TE for the dynamic computation of receiver-driven P2MP and MP2MP LSP tree structures.

1.1. Motivation

IP multicast distribution trees are receiver-initiated and dynamic by nature. IP multicast-enabled applications are also bandwidth savvy, especially in the area of residential IPTV services, where the delivery of multicast contents to several hundreds of thousands of IPTV receivers assumes the appropriate level of quality. Current source-driven P2MP LSP establishment, as defined as in [RFC4875], assumes a priori knowledge of receiver locations, and the LSP signalling is initiated and driven by the data sender(headend).

Receiver-driven MPLS P2MP/MP2MP tree structures do not require sender to maintain/discover receiver information a priori, and their design should better accommodate the receiver-specific QoS conditions, such as network access capabilities.

1.2. Terminology

The following terms are used in this document:

- **Sender**: Sender refers to the Originator (and hence) Sender of the content/payload, as defined in [RFC2205].
- **Receiver**: Receiver refers to the Receiver of the content/payload, as defined in [RFC2205].
- **Upstream**: The direction of flow from content Receiver toward content Sender, as defined in [RFC2205].
- **Downstream**: The direction of flow from content Sender toward content Receiver, as defined in [RFC2205].
Path-Sender: The sender of RSVP PATH messages, with NO correlation to the direction of content/payload flows. Its flow direction is irrelevant to that of Sender defined above. All other control messages discussed in this document will use this as the reference.

Path-Receiver: The receiver of RSVP PATH messages, with NO correlation to the direction of content/payload flows.

Path-Initiator: The Path-Sender that originated a RSVP PATH message. This is different from Path-Sender in that an intermediate node can be a Path-Sender, but such an intermediate node cannot create and initiate the RSVP PATH message.

Path-Terminator: The Path-Receiver that does NOT propagate the Path message any further. This is different from Path-Receiver in that an intermediate node can be a Path-Receiver, but such an intermediate node will propagate the Path message to the next hop.

Root: A router where a multicast LSP is rooted at. Data enters the root and then is distributed to leaves along the P2MP/MP2MP LSP.

1.3. Overview

Although receiver-driven P2MP LSPs as defined in this document use existing sender-driven syntax, there are important semantic differences that need to be defined for correct interpretation and interoperability. In the receiver-driven approach, we inverted the semantics of P2MP-TE RSVP [RFC4875] messages, while keeping the syntax unchanged.

Following are some key differences that are specific to the receiver-driven paradigm:

- The leaf router the receiver is connected to: that is, upon receipt of an IGMP/MLD Report message, the router that embeds the IGMP/MLD Querier would typically trigger a RSVP_PATH message towards the source. We keep the same convention with respect to data flows, which are opposite to control flows.

- L2S (Leaf-to-Source) Destinations (the sender or root) are routers where user data payload traffic enters the LSP.

- RSVP P2MP PATH messages traverse from receivers to the root.

- RSVP P2MP RESV messages traverse from the root to the leaf routers of the P2MP tree structure.
A RSVP RESV message received by a router is interpreted as a successful resource reservation made by the upstream node for the establishment of the P2MP tree structure.

A RSVP RESV message received by a router is interpreted as successful resource reservation made by the downstream node for the establishment of a MP2MP tree structure.

Label allocation on incoming interfaces is done prior to sending RSVP PATH messages upstream for P2MP tree structures.

Label allocation on incoming interfaces is done prior to sending RSVP RESV messages upstream for MP2MP tree structures.

For P2MP LSP tree structures, a node receiving a RSVP PATH message first decides if this RSVP PATH message will make the said node a branch LSR or not. If it is not a branch LSR, it is a transit LSR. In the case that it will become a transit LSR because of this PATH message, then it will, before sending the RSVP PATH message upstream, allocate required bandwidth on the interface on which the RSVP PATH message is received. The upstream node can send traffic soon after successfully reserving resources on the downstream link, on which the RSVP PATH message SHOULD be received. In the case that the node is already a branch or a transit node before it receives the PATH message, then it will allocate required bandwidth on the interface on which the RSVP PATH message is received, and send the RESV message to the node which sends the PATH message without propagating the PATH message further to the upstream node. For P2MP LSPs, a label is carried by the PATH message and should be used by the upstream node when distributing the data from upstream to downstream.

For MP2MP LSP tree structures, a node will allocate required bandwidth on the interface through which the RSVP PATH message is sent before sending the RSVP PATH message upstream. A node receiving a RSVP PATH message MUST first decide if this RSVP PATH message will make the said node a branch LSR or not. In the case it will become a transit LSR because of this PATH message, then it will allocate required bandwidth on the interface on which the RSVP PATH message is received and will allocate required bandwidth on the interface through which the RSVP PATH message is sent, before sending the RSVP PATH message upstream. The downstream node can send traffic soon after successfully reserving bandwidth on the upstream link through which the RSVP PATH message SHOULD be sent. The upstream node can send traffic soon after successfully reserving bandwidth on the downstream link on which the RSVP PATH message SHOULD be received. In the case that the node is already a branch or a transit node before it receives the PATH message, then it will allocate required resources on the interface on which the RSVP PATH message is received, and send the RESV message to the node.
which sends the PATH message without propagating the PATH message further to the upstream node. The label carried by the PATH message should be used by the Path-Receiver node to forward data from the Path-Receiver node to the Path-Sender node, and the label carried by RESV messages should be used by its corresponding Path-Sender node to deliver data from the Path-Sender node to the Path-Receiver node.

o For the sake of readability, from now on, all mRSVP-TE messages will be used to represent all the RSVP-TE messages which are used for the computation of receiver-driven P2MP/MP2MP tree structures.

o For the sake of readability, from now on all mRSVP-TE LSPs will be used to represent all P2MP and/or MP2MP LSPs in receiver-driven (RD) multicast P2MP/MP2MP MPLS environments. We will sometimes use RD P2MP TE LSP or RD MP2MP TE LSP to represent such receiver-driven multicast LSPs.
1.4. Receiver-Driven mRSVP-TE LSP Examples

Figure 1: P2MP EXAMPLE

Figure 1 shows that R5 is added as the first leaf of the RD P2MP TE LSP, the message flow goes from R5->msg1->R3->msg2->R1->msg3->R3->msg4->R5. When the leaf R4 is added, the message flow goes from R4->msg5->R3->msg6->R4. In this case, when R3 receives msg5, R3 finds out that the RD P2MP LSP has already been set up for R5: therefore, R3 finds itself a branch node for leaf R4 and R5, so it will terminate the PATH message and build the corresponding RESV message and send it back to R4. The association of the LSP initiated by R4 to the existing RD P2MP LSP is determined based on the processing of the session object from the mRSVP-TE message. This session object is further documented in Section 2 of this draft.
Figure 2 shows that R5 is added as the first leaf
(as both a sender and a receiver) of the RD MP2MP TE LSP, the message
flow goes from R5->msg1->R3->msg2->R1->msg3->R3->msg4->R5. When the
leaf R4 (both receiver and sender) is added, the message flow goes
from R4->msg5->R3->msg6. In this case, when R3 receives msg5, R3
finds out that the RD MP2MP LSP has already been set up for R5, and R3
will become the branch LSR for the leaf R4 and R5, so it will
terminate the PATH message, build a RESV message and send the RESV
message back to R4. The association of the LSP initiated by R4 to
the existing MP2MP LSP is determined based on the processing of the
session object from the mRSVP-TE message. This session object is further
documented in Section 2 of this draft.

2. Signaling Protocol Extensions

mRSVP-TE is similar to the RSVP-TE protocol as specified in [RFC4875],
[RFC3473] and [RFC3209], but differs in that the receivers (or the leaf
routers they are connected to) initiate the RSVP PATH messages toward the sender. Compared with
[RFC4875], mRSVP-TE to be specified in this document can also be used
to set up MP2MP LSPs.

Within RD P2MP/MP2MP LSP tree structure environments, the Receiver is the
Path-Originator. The RSVP RESV messages flow in the opposite direction as
compared to the RSVP PATH messages, i.e. RSVP RESV messages are generated
by the Sender or a branch LSR.

Within this receiver-driven context, the processing of receiver-
initiated mRSVP-TE P2MP messages is different from that of the
other RSVP messages. Following the method used by RSVP-TE and P2MP
RSVP-TE, this draft documents the use of new SESSION C-Type as
follows:

Class Name = SESSION

C-Type
XX+0   mRSVP_TE_P2MP_LSP_TUNNEL_IPv4 C-Type
XX+1   mRSVP_TE_P2MP_LSP_TUNNEL_IPv6 C-Type
XX+2   mRSVP_TE_MP2MP_LSP_TUNNEL_IPv4 C-Type
XX+3   mRSVP_TE_MP2MP_LSP_TUNNEL_IPv6 C-Type

Where XX is a number to be allocated by IANA.

The new SESSION C-Type MUST be used in all receiver-driven P2MP
RSVP-TE messages.

The following sections describe the receiver-driven P2MP RSVP-TE
extensions to the P2MP RSVP-TE protocol. When there is no difference
in the protocol, usage of [RFC4875] is assumed.
2.1. L2S Sub-LSPs

A RD P2MP or MP2MP LSP is composed of one or more L2S sub-LSPs, which are merged together at the branch nodes.

An L2S sub-LSP exists within the context of a RD P2MP or MP2MP LSP. There are two ways to identify each sub-LSP:

- From the Sender’s perspective, each sub-LSP is identified by the SESSION object, the SENDER_TEMPLATE object and S2L_SUB_LSP object, as specified in [RFC 4875]. The SESSION object encodes P2MP ID, which is unique within the scope of the sender (ingress LSR) and remains constant throughout the lifetime of the P2MP tree structure. The Extended Tunnel ID, which remains constant throughout the lifetime of the P2MP tree structure, and which should contain the sender’s address to make sure the identifier is globally unique identifier. Finally, the Tunnel ID, also remains constant throughout the lifetime of the P2MP tree structure. The SENDER_TEMPLATE object contains the ingress LSR source address. The S2L_SUB_LSP contains the destination address of the sub-LSP.

- From the Receiver’s perspective, each sub-LSP is identified by a new SESSION object specified in this document, the SENDER_TEMPLATE object and L2S_SUB_LSP. The SESSION object, different from the one used in typical sender-driven environments, contains some opaque values to be used as the key to associate different PATH messages originated from different leaves. The SENDER_TEMPLATE object contains the Path-Sender’s address, which is actually the Data Receiver. The L2S_SUB_LSP contains the source address of the sub-LSP, i.e. the data Sender’s address.

This document takes the approach from the Receiver’s perspective. The approach from the Sender’s perspective is documented in [RFC 4875]. In this draft, the SESSION object will encode multicast information such as multicast source and group addresses as the opaque value.

Once either the Sender LSR, a transit LSR, or a branch LSR receives a PATH message containing SESSION object, the LSR should be able to use the opaque values encoded in the SESSION object to determine whether the sub-LSP signaled by this PATH message should be merged with existing LSPs.

An EXPLICIT_ROUTE Object (ERO) or mRSVP-TE_SECONDARY_EXPLICIT_ROUTE Object (SERO) is used to optionally specify the explicit route of a L2S sub-LSP. Each ERO or SERO that is signaled corresponds to a particular L2S_SUB_LSP object. Details of explicit route encoding are specified in section 4.5 of [RFC4875]. The
SECONDARY_EXPLICIT_ROUTE Object is defined in [RFC4873], the mRSVP-TE SECONDARY_EXPLICIT_ROUTE Object C-type and the matching mRSVP-TE_SECONDARY_RECORD_ROUTE Object C-type is defined in this document.

2.2. Explicit Routing

When a Path message signals a L2S sub-LSP, the EXPLICIT_ROUTE object encodes the path from the leaf to the root LSR. The Path message also includes the L2S_SUB_LSP object for the L2S sub-LSP being signaled. The \<(EXPLICIT_ROUTE), (L2S_SUB_LSP)\> tuple represents the L2S sub-LSP and is referred to as the sub-LSP descriptor. The absence of the ERO should be interpreted as requiring hop-by-hop routing for the sub-LSP based on the L2S sub-LSP root address field of the L2S_SUB_LSP object.

2.3. L2S Sub-LSPs and Path Messages

The mechanism specified in this document allows a RD P2MP/MP2MP LSP to be signaled using one or more Path messages. Each Path message may signal one L2S sub-LSPs.

Receiver-driven P2MP MPLS-TE LSP uses the Path message to carry the LABEL object upstream towards the Sender. With a receiver-driven usage of the RSVP PATH message, the LABEL_REQUEST object carried by the PATH message is no longer mandatory, it becomes optional for receiver-driven PATH messages, as mentioned in Figure 4:

\[
\text{<Path Message> ::= \quad <Common Header> \quad [\quad <INTEGRITY> \quad ]
\quad [\quad [\quad <MESSAGE_ID_ACK> \quad |\quad <MESSAGE_ID_NACK>\quad ] \quad ...\quad ]
\quad [\quad <MESSAGE_ID> \quad ]
\quad <SESSION> \quad <RSVP_HOP>
\quad <TIME_VALUES>
\quad [\quad <EXPLICIT_ROUTE> \quad ]
\quad [\quad <LABEL_REQUEST> \quad ]
\quad [\quad <PROTECTION> \quad ]
\quad [\quad <LABEL_SET> \quad ...\quad ]
\quad [\quad <SESSION_ATTRIBUTE> \quad ]
\quad [\quad <NOTIFY_REQUEST> \quad ]
\quad [\quad <ADMIN_STATUS> \quad ]
\quad [\quad <POLICY_DATA> \quad ...\quad ]
\quad <sender descriptor>
\quad [<L2S sub-LSP descriptor list>]
\]

Figure 4: PATH Message Extensions.

The SESSION object encodes an opaque value which is used as a key to associate different PATHs to the same RD P2MP/MP2MP tree structure.
Using [RFC4875] as the base specification, with the LABEL object being added to the SENDER DESCRIPTOR object (Figure 5):

\[
<\text{sender descriptor}> ::=  <\text{SENDER\_TEMPLATE}> <\text{SENDER\_TSPEC}>
\]
\[
\text{[} <\text{ADSPEC} \text{]} \]
\[
\text{[} <\text{RECORD\_ROUTE} \text{]} \]
\[
\text{[} <\text{SUGGESTED\_LABEL} \text{]} \]
\[
\text{[} <\text{RECOVERY\_LABEL} \text{]} \]
\[
<\text{LABEL}> \]

Figure 5: SENDER DESCRIPTOR Object.

The LABEL object is defined in section 4.1 of [RFC3209].

Please note that receiver-driven PATH messages convey the LABEL_REQUEST as an optional object. When the receiver-driven P2MP LSP is uni-directional, the LABEL_REQUEST in the PATH message is not used. When bi-directional receiver-driven P2MP LSP (MP2MP) are used, the LABEL_REQUEST will operate as described in [RFC4875], providing the label allocation operation in the other direction.

2.4. Resv Message Extensions

Receiver-driven P2MP RSVP-TE does not need any change to the basic RESV message, as illustrated in section 6.1 of [RFC4875], as long as the SESSION object is using one of the new C-Types to be assigned by IANA. In this document, we specify four new C-Types for RD P2MP/MP2MP tree structures: IPv4 P2MP tunnels, IPv4 MF2MP tunnels, IPv6 P2MP tunnels, IPv6 MF2MP tunnels.

For receiver-driven P2MP tree structures, the PATH message carries the LABEL object, and thus the RESV message doesn’t have to carry the LABEL object anymore.

But for MP2MP LSP tree structures, as indicated by the new C-Type of the SESSION object, both PATH and RESV messages will carry LABEL objects for sending and receiving purposes, respectively. Within the context of MP2MP tree structures, one of the directions is established as per [RFC3209]. Thus, this document is changing the use of the LABEL object in the FF Flow Descriptor and SE Filter Spec from mandatory to optional. As indicated in Figure 6:
2.5. PathErr Message Extensions

The receiver-driven PathErr messages have the same syntax and utilization as the PathErr message described in [RFC4875], with the difference in the SENDER DESCRIPTOR object carried by the PathErr message. The receiver-driven PathErr message will use the SENDER DESCRIPTOR object defined in Section 2.1 of this document, the same SENDER DESCRIPTOR object carried by the Path message the PathErr message corresponds to, allowing the indication of the LABEL object in the SENDER DESCRIPTOR. With the ERROR_SPEC object being able to indicate Label errors with Error Code 24 for Routing Problems and Error Value sub-code of 9 for MPLS label allocation failure as defined in section 7.3 of [RFC3209].

2.6. ResvErr Message Extensions

The receiver-driven ResvErr messages have the same syntax and utilization as the ResvErr message described in [RFC4875]. But this ResvErr message will be processed as per Section 2.2 of this document, given that the FF FLOW DESCRIPTOR object and the SE FILTER SPEC object can optionally contain the LABEL object (instead of mandating the use of the LABEL object). The optional use of the LABEL object is conditioned by the nature of the P2MP tree structure, either uni-directional (P2MP) or bi-directional (MP2MP).

2.7. PathTear Message Extensions

The receiver-driven PathTear message have the same syntax and utilization as the PathTear message described in [RFC4875], assuming a difference in the SENDER DESCRIPTOR object carried by the PathTear message. The receiver-driven PathTear message will use the SENDER DESCRIPTOR object defined in Section 2.1 of this document, the same SENDER DESCRIPTOR object carried by the Path message the PathTear message corresponds to, allowing the indication of the LABEL object.
in the SENDER DESCRIPTOR.

2.8. SESSION Object

An mRSVP-TE LSP SESSION object is used to implicitly represent a RD P2MP or an MP2MP tree structure. The opaque values encoded in such SESSION objects are used by Path- Receivers to determine whether or not to associate different PATH messages from different leaves to the same P2MP/MP2MP tree structure. The SESSION object has the following general format, whose length is determined by the C-type (Figure 7):

```
+-----------------------------------------------+-----------------------------------------------+
|                                                               |
|                                                               |
|                                                               |
+-----------------------------------------------+-----------------------------------------------+
```

Figure 7: Format of the mRSVP-TE-LSP SESSION-Object.

The SESSION object uses the existing SESSION C-Num. New C-Types are defined to accommodate different lengths of opaque values. For native IPv4/IPv6 multicast, the opaque values should encode IPv4/IPv6 (S, G) or (*. G, RP) for P2MP or MP2MP LSPs. For IPv4/IPv6 multicast VPNs, the opaque values should encode a VPN ID. For tunnel aggregation where multiple multicast streams are sharing a same LSP tunnel, the opaque values are TBD.

The combination of the SESSION object, the SENDER TEMPLATE object and the L2S_SUB_LSP object identifies each L2S sub-LSP. This follows the existing P2P RSVP-TE notion of using the SESSION object for identifying a P2P Tunnel, which in turn can contain multiple LSPs, each distinguished by a unique SENDER TEMPLATE object.

The mechanism for propagating the values defined in the SESSION object is outside the scope of this document.

The mechanism for setting the values defined in the SESSION object is TBD.
2.8.1. mRSVP-TE LSP Tunnel IPv4 SESSION Object

Class = SESSION, mRSVP_TE_P2MP_LSP_TUNNEL_IPv4 C-Type = TBD,
mRSVP_TE_MP2MP_LSP_TUNNEL_IPv4 C-Type = TBD.

Figure 8: mRSVP-TE-LSP-Tunnel-IPv4-SESSION-Object.

2.8.2. mRSVP-TE LSP Tunnel IPv6 SESSION Object

This is the same as the mRSVP-TE IPv4 LSP SESSION object with the
difference that the multicast source and group addresses are 16-byte
long (Figure 8).

Class = SESSION, mRSVP_TE_P2MP_LSP_TUNNEL_IPv6 C-Type = TBD,
mRSVP_TE_MP2MP_LSP_TUNNEL_IPv6 C-Type = TBD.

Figure 9: mRSVP-TE-LSP-Tunnel-IPv6-SESSION-Object.
2.9. SENDER_TEMPLATE Object

The SENDER_TEMPLATE object contains the Path-Initiator LSR address. In this document, the Path-Initiator is the same as the Leaf Router. The LSP ID can be changed to allow a sender to do a certain level of aggregation. Thus, multiple instances of the P2MP tunnel can be created, each with a different LSP ID. The instances can share resources with each other. The L2S sub-LSPs corresponding to a particular instance use the same LSP ID.

2.9.1. mRSVP-TE LSP Tunnel IPv4 SENDER_TEMPLATE Object

Class = SENDER_TEMPLATE, mRSVP-TE_LSP_TUNNEL_IPv4 C-Type = TBD.

IPv4 tunnel receiver address: the IPv4 address of the Leaf Router IPv4 address.

LSP ID

See [RFC3209].

2.9.2. mRSVP-TE LSP Tunnel IPv6 SENDER_TEMPLATE Object

Class = SENDER_TEMPLATE, mRSVP-TE_LSP_TUNNEL_IPv6 C-Type = TBD.
IPv6 tunnel receiver address: the IPv6 Address of the Leaf Router.

LSP ID

See [RFC3209].

2.10. L2S_SUB_LSP Object

An L2S_SUB_LSP object identifies a particular L2S sub-LSP belonging to the mRSVP-TE LSP.

2.10.1. L2S_SUB_LSP IPv4 Object

L2S_SUB_LSP Class = 50, L2S_SUB_LSP_IPv4 C-Type = TBD.
IPv4 L2S Sub-LSP Root Address: IPv4 address of the L2S sub-LSP sender.

2.10.2. L2S_SUB_LSP IPv6 Object

L2S_SUB_LSP Class = TBD, L2S_SUB_LSP_IPv6 C-Type = TBD

It is the same as the IPv4 L2S Sub-LSP object, with the difference that the root address is a 16-byte IPv6 address.

```
0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|        IPv6 L2S Sub-LSP Root Address (16 bytes)               |
|                        ....                                   |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
```

Figure 13: L2S_SUB_LSP_Ipv6_Object.

2.11. FILTER_SPEC Object

The FILTER_SPEC object is canonical to the P2MP SENDER_TEMPLATE object.

2.11.1. P2MP LSP_IPv4 FILTER_SPEC Object

Class = FILTER_SPEC, P2MP LSP_IPv4 C-Type = TBD.

The format of the P2MP LSP_IPv4 FILTER_SPEC object is identical to the P2MP LSP_IPv4 SENDER_TEMPLATE object.

2.11.2. mRSVP-TE LSP_IPv6 FILTER_SPEC Object

The mRSVP-TE LSP_IPv6 FILTER_SPEC Object uses the P2MP LSP_IPv6 SENDER_TEMPLATE object defined in [RFC4875].
2.11.3. mRSVP-TE SECONDARY_EXPLICIT_ROUTE Object (SERO)

The mRSVP-TE SECONDARY_EXPLICIT_ROUTE Object (SERO) used the P2MP SECONDARY_EXPLICIT_ROUTE Object (SERO) defined in [RFC4875].

2.11.4. mRSVP-TE SECONDARY_RECORD_ROUTE Object (SRRO)

The mRSVP-TE SECONDARY_RECORD_ROUTE Object (SRRO) uses the P2MP SECONDARY_RECORD_ROUTE Object (SRRO) defined in [RFC4875].

3. In-Band and Out-Band Signalling

3.1. In-Band Signalling

In typical Live TV broadcasting environments where the corresponding IP multicast distribution trees usually computed and established by means of the Protocol Independent Multicast (PIM), need to be deployed over an MPLS domain, Receiver-Driven P2MP LSP tree structures can be created. The part of the IP multicast tree that traverses the MPLS domain can be instantiated as a RD P2MP LSP. When a PIM Join message is received at the border of the MPLS domain, information from that message will be encoded into RSVP-TE P2MP messages. When the RSVP-TE P2MP messages reach the border of the next IP multicast domain, the encoded information will be able to be used to generate PIM messages that can be sent through this IP multicast domain. The result will be an IP multicast tree consisting of a set of IP multicast sub-trees that can be spliced together with a RD P2MP RSVP-TE based LSP.

The detailed protocol extensions and procedures are TBD.

3.2. Out-Band Signalling

In the case that In-Band Signalling is not used, whenever a PIM Join message is received at the border of the MPLS domain, information from that message can also be encoded into BGP messages. When the BGP messages reach the border of the next IP multicast domain, the encoded information should be used to generate PIM messages that can be sent through the said IP multicast domain. The result should be an IP multicast tree consisting of a set of IP multicast sub-trees that can be spliced together with a RD P2MP LSP tree structure.
4. Broadcast Interfaces

The receiver-driven approach interoperates with the RSVP upstream label allocation mechanism [I-D.ietf-mpls-rsvp-upstream]. Path-Senders SHOULD detect the presence of such P2MP-LSP over broadcast interfaces for each Path-Receiver. Instead of labels carried by PATH messages, the upstream-assigned labels are carried by RESV messages.

The considerations for MP2MP is TBD.

5. Fast Re-Route Considerations

TBD.

6. Backward Compatibility

A receiver-driven P2MP LSP mechanism uses a unique C-Type. LSRs that do not support receiver-driven P2MP-TE LSP, send Path Error [TBD] back to the Path Initiator.

The complete discussion on the Backward Compatibility will be provided in the Next version of the document.

7. Acknowledgements

We would like to thank authors of [RFC4461], [RFC4875] and the authors of draft-ietf-mpls-mp-ldp-reqs from which some text of this document has been inspired.

8. IANA Considerations

This section is TBD.

9. Security Considerations

How a receiver is authenticated is outside the scope of this document. But we briefly summarize the requirements which are detailed in the requirements draft.

It is a requirement that any mRSVP-TE solution developed to meet some or all of the requirements expressed in this document MUST include mechanisms to enable the secure establishment and management of mRSVP-TE MPLS-TE LSPs. This includes, but is not limited to:
A receiver MUST be authenticated before it is allowed to establish mRSVP-TE LSP with source, in addition to hop-by-hop security issues identified by in RFC 3209 and RFC 4206.

mechanisms to ensure that the ingress LSR of a P2MP LSP is identified;

mechanisms to ensure that communicating signaling entities can verify each other's identities;

mechanisms to ensure that control plane messages are protected against spoofing and tampering;

mechanisms to ensure that unauthorized leaves or branches are not added to the mRSVP-TE LSP; and

mechanisms to protect signaling messages from snooping.

Note that mRSVP-TE signaling mechanisms built on P2P RSVP-TE signaling are likely to inherit all the security techniques and problems associated with RSVP-TE. These problems may be exacerbated in mRSVP-TE situations where security relationships may need to maintained between an ingress LSR and multiple egress LSRs. Such issues are similar to security issues for IP multicast.

It is a requirement that documents offering solutions for P2MP LSPs MUST have detailed security sections.

10. References

10.1. Normative References


10.2. Informative References


Authors’ Addresses

Renwei Li
Huawei Technologies
2330 Central Expressway
Santa Clara, CA 95050
USA

Phone:
Email: renwei.li@huawei.com

Quintin Zhao
Huawei Technologies
Boston, MA
USA

Phone:
Email: quintin.zhao@huawei.com

Christian Jacquenet
France Telecom Orange
4 rue du Clos Courtel
35512 Cesson Sevigne, France

Phone: +33 2 99 12 43 82
Email: christian.jacquenet@orange.com
Abstract

This document is framework for Unified MPLS.

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 July 28, 2012.

Copyright Notice

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

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

## Table of Contents

1. Introduction .................................................. 3  
   1.1. Motivation and Background ............................... 3  
   1.2. Scope ................................................... 4  
   1.3. Terminology .............................................. 5  
      1.3.1. MPLS Protocols and Packages ...................... 5  
      1.3.2. MPLS Technology ..................................... 6  
      1.3.3. Topology Driven MPLS ............................... 6  
      1.3.4. MPLS Traffic Engineering ......................... 6  
      1.3.5. MPLS Transport Profile ............................. 7  
      1.3.6. Additional Terms and Acronyms .................... 7  
      1.3.7. MPLS Terminology structure ....................... 8  
2. Unified MPLS Requirements .................................. 13  
   2.1. Full Interoperability and Interworking ................ 13  
   2.2. Common Data Plane ...................................... 13  
   2.3. Common OAM ............................................. 13  
   2.4. Data Plane Agnostic ................................... 13  
   2.5. Interworking ........................................... 13  
3. Unified MPLS Overview ...................................... 15  
   3.1. Unified MPLS Control Plane .............................. 15  
   3.2. Unified MPLS Data Plane ................................ 15  
   3.3. Unified MPLS Management ............................... 16  
   3.4. Unified MPLS OAM ..................................... 17  
4. Interfaces and Interworking ............................... 18  
   4.1. Interfaces ............................................. 18  
   4.2. Network Layer Interworking ............................ 18  
   4.3. Service Layer Interworking ............................. 18  
   4.4. Network Overlay ....................................... 19  
5. MPLS Resiliency ............................................. 20  
   5.1. MPLS Recovery .......................................... 20  
   5.2. MPLS Protection ........................................ 20  
   5.3. MPLS Survivability ..................................... 20  
6. Other Unified MPLS Considerations ......................... 21  
7. Security Considerations ................................... 22  
8. IANA considerations ....................................... 23  
9. Acknowledgments ........................................... 24  
10. References .................................................. 25  
   10.1. Normative References ................................ 25  
   10.2. Informative references ............................... 25  
Authors’ Addresses ........................................... 26
1. Introduction

The term "Unified MPLS" indicates several different things:

The ambition to continue to keep MPLS together as one single technology, based on a common architecture.

The ambition to maximize interworking between different flavors of MPLS (MPLS Packages) within the MPLS technology.

The ambition that functionality developed for one of the MPLS Aggregate Packages should be re-useable within other MPLS Aggregate Packages.

MPLS is a mature Internet technology that has been used over the last 18 years to give e.g. QoS, resiliency, scalability, virtualization, etc. to IP networks.

Through e.g. Pseudowires and L2VPN development it has also been adapted to emulate legacy protocols such as Frame Relay, TDM, ATM and Ethernet.

The development of the MPLS technology has been very rapid since the start, the last 4 to 5 years have not been an exception. In particular we have seen developments in OAM in order to meet transport environments requirements, specification of multicast techniques and over the last years new resiliency strategies have been deployed.

Over the years it has become possible to distinguish different MPLS Packages, e.g. MPLS traffic Engineering (MPLS-TE), Topology Driven MPLS (MPLS-TD) and the Transport Profile of MPLS (MPLS-TP).

Revision note: Version -01 has been updated after comments from several sources, especially we are grateful for comments that have helped improve the structure of the document. We have also received comment to the effect "in the list in (now) section 1.3 you have missed" we have tried adding this information as best as we can.

There is still some glaring holes that we need to cover and is open to input and participation driving the draft forward.

1.1. Motivation and Background

Needless to say this very dynamic environment has put stress on the MPLS architecture. Another factor that contributed is that MPLS technology over the years has evolved to addresses almost all networking market segments.
Vendors that are only interested to incorporate one of the MPLS packages often tend to prioritize the development of one of the packages at the expense of the others, one typical example is the T-MPLS standardization that was started within ITU-T and rapidly developed into a separate technology. This in itself put stress on the MPLS architecture.

The MPLS architecture has stood up surprisingly well, but there are a few things that need to be documented as part of the MPLS architecture, primarily on the OAM side, but also on MPLS multicast and traffic engineering.

The time has come to take steps to bring the MPLS architecture together and create a unified MPLS architecture - an architecture and protocol suite that can operate end to end, with pieces that may be combined in a fashion that can meet the needs of the service provider according to their requirements and environment and not the limitations of the profiles.

1.2. Scope

The term "Unified MPLS" has come to indicate that the environment where MPLS is used has been gradually changing over the years, it is highly likely that architectural updates are needed. This framework document addresses how the environment of RFC 3031 [RFC3031] has changed. MPLS started out as a set of tools to support IP networks, but over time developed to have much broader application than that. We have seen the development of IP-VPNs, L2VPNs, PWs for carrying non-IP payload, deployment of MPLS in networks where we don’t even require or have IP in the forwarding plane.

Starting with existing MPLS protocols, i.e. data plane, signaling (label distribution) protocols, routing, TE-extensions, OAM and the different management tools it is possible to build a surprisingly large number of viable MPLS deployments.

For lack of a better name we have chosen to call the deployable MPLS constructs "packages". In Terminology (Section 1.3) we will elaborate on this terminology.

This document discusses interworking between MPLS Packages on LSP level. Discussion of interworking between PWs is for further study.

A number of standardization and technology development projects have extended MPLS in very different directions. In Unified MPLS we make two assumptions:
1. The Interworking potential between different MPLS Packages and MPLS Aggregate packages shall be maximized and that this is a design goal in all development and standardization of MPLS.

2. Functionality developed for one MPLS Aggregate Package, e.g. the MPLS-TP OAM tools developed in the joint project with ITU-T, shall be designed in such a way that it is re-useable within other MPLS (Aggregate) Packages.

1.3. Terminology

In the "ontology" section (Section 1.3.7) an outline of this aspect of the MPLS terminology is given.

1.3.1. MPLS Protocols and Packages

For the purpose of understanding the MPLS terminology structure we operate with four concepts; MPLS Protocols, MPLS Packages, MPLS Aggregate Packages and MPLS Technology.

1.3.1.1. MPLS Protocols

We have a rather wide definition of "protocol", i.e. all the IETF protocols that specify MPLS behavior of the data, control and management plane. OAM is considered to be part of the data plane.

1.3.1.2. MPLS Packages

An MPLS Package is a ordered collection of MPLS protocols put together because it can give you a desired functionality. One example would be running OSPF and LDP in your network to create a topology driven MPLS see Topology Driven MPLS (Section 1.3.3).

1.3.1.3. MPLS Aggregate Packages

An MPLS Aggregate Package is a grouping of MPLS Packages. Packages that can be used to build MPLS network of identical functionality can be grouped into MPLS Aggregate Package. There is for example nothing that stops you from running Topology Driven MPLS (MPLS-TD) based on ISIS instead of OSPF; those two MPLS packages form together the MPLS-TD Aggregate package.

Note: We need to revisit "identical functionality" in the paragraph above and see if there is a better way do describe this.

We see that it is possible to form three different MPLS Aggregate Packages. First we have MPLS Traffic Engineering (MPLS-TE), Topology Driven MPLS (MPLS-TD) and the MPLS Transport Profile (MPLS-TP).
1.3.2. MPLS Technology

The "MPLS" or "MPLS Technology" is the sum of the MPLS protocols and the method of grouping them to MPLS Packages and Aggregate MPLS Packages as illustrated in the Appendix.

The MPLS technology is also scoped, described, and explained in a number of documents such as applicability statements, requirements, and frameworks. Though those documents are not part of the Standards Track they are important as they give a comprehensive view of the MPLS Technology.

1.3.3. Topology Driven MPLS

We talk about Topology Driven MPLS (MPLS-TD) when the protocols establishing LSPs are limited to use the topology information created by IP routing protocols; i.e., packets sent over an MPLS LSP will traverse the network over the same path as it would have taken if the Shortest Path IP forwarding had been used.

MPLS-TD consists of at least two different MPLS Packages, depending if ISIS or OSPF is used as the routing protocol. LDP is almost the only signaling protocol that is used in this type of network.

Note: It is technically possible to use other routing protocols, e.g., Routing Information Protocol (RIP) RFC 2453 [RFC2453] or even static configuration or routes. However, use of RIP or static routes in MPLS is very limited, and for all practical purposes these cases can be left aside.

1.3.4. MPLS Traffic Engineering

We talk about MPLS Traffic Engineering (MPLS-TE) when other parameters than just the shortest path is taken into consideration when deciding on how and where an LSP should be set up. By far the most common criteria are available BW and explicit routing, this explains why the RSVP-TE protocol was well suited to be adapted as MPLS-TE signaling protocol.

Deciding how many MPLS-TE Packages there are is slightly harder than for the MPLS-TD.

There are two and only two routing protocols - OSPF-TE and ISIS-TE.

For the signaling protocols the situation is a bit ambivalent; the
MPLS developed a RSVP-TE for MPLS traffic engineering RFC 3209 [RFC3209]. When the RSVP-TE was "generalized" for Generalized MPLS (GMPLS) the intention was that the more specific objects should be possible to use, but for the Label object this was never achieved. One could therefore say that we have two different signaling protocols.

1.3.5. MPLS Transport Profile

The MPLS Transport Profile (MPLS-TP) is the latest addition of the MPLS Aggregate Packages. It is an extension of MPLS to meet the requirements from transport networks.

There are five different MPLS Packages for MPLS-TP

Note: We have a comment that we also have a hybrid package, when both an NMS and an CP is used.

- NMS Driven MPLS-TP (ND MPLS-TP), LSPs, PWs and segments is set up and configured from an NMS
- Control Plane driven MPLS (CD MPLS-TP), LSPs, PWs and segments is set up and configured from the control plane.

There are two variants of CD MPLS-TP depending on which routing protocol that is used; for CD MPLS-TP there is a decision to use the GMPLS version of RSVP-TE.

* based on RSVP-TE and OSPF-TE

* based on RSVP-TE and ISIS-TE

The fourth variant is MPLS-TP P2MP configured from an NMS

The fifth variant is when the MPLS-TP P2MP is configured from a control plane using the MPLS signaling and routing protocols.

1.3.6. Additional Terms and Acronyms

This section lists terms and acronyms used in this document.

1.3.6.1. New terms

Control Plane Driven -
LSPs are set up and configured from the control plane. This is true for almost all MPLS, but the Control Plane Driven terminology originates from the MPLS-TP project.
NMS Driven -
LSPs are set up and configured from an NMS. This is the mode Transport Networks has been operated in traditionally.

1.3.6.2. Acronyms

CD MPLS-TP - Control Plane Driven MPLS-TP
CLI - Command Line Interface
EM - Element Manager
LCT - (correct expansion??)
MPLS-TD - Topology Driven MPLS
MPLS-TE - Traffic Engineered MPLS
MPLS-TP - MPLS for Transport Networks (MPLS Transport Profile)
ND MPLS-TP - NMS Driven MPLS-TP
NE - Network Element
NMS - Network Management System
TE P2MP - Traffic Engineered P2MP

1.3.7. MPLS Terminology structure

In the abbreviated table below a way of thinking about the relationship between the MPLS RFCs, MPLS protocols, MPLS packages and MPLS Aggregate Packages is outlined.

We take the approach that RFCs can be grouped into protocols, and that protocols may form MPLS packages, which in turn can be grouped into Aggregate Packages.

In talking about "MPLS RFCs" and "MPLS protocols" we do so in a wide sense, i.e. RFCs and protocols that might be used in MPLS networks. Most of those are developed by working groups that we listed as "MPLS working groups" in RFC 4929 [RFC4929], but there are also other protocols that are used and necessary in some MPLS networks, e.g. the IGP that were developed in other working groups.

We have added NMS Configuration and MPLS Data Plane, even though they they are not "protocols" of the same flavor as the other protocols
MPLS Terminology Structure
- a strawman ontology

MPLS RFCs (examples)

- RFC5036  RFC3478  RFC3479  RFC3209  RFC3477  RFC6428
- RFC3031
- RFC3032  RFC3033  RFC3034  RFC3035  RFC3038  RFC3107
- RFC3209  RFC3270  RFC3473  RFC3477  RFC3478  RFC3479
- RFC3811  RFC3812  RFC3813  RFC3814  RFC3815  RFC4023
- RFC4090  RFC4182  RFC4201  RFC4206  RFC4220  RFC4368
- RFC4379  RFC5420  RFC4561  RFC4781  RFC4817  RFC4875
- RFC4950  RFC5283  RFC5330  RFC5331  RFC5332  RFC5462
- RFC%61  RFC5586  RFC5710  RFC5711  RFC5712  RFC5718
- RFC5918  RFC5919  RFC5960  RFC6178  RFC6370  RFC6372
- RFC6374  RFC6375  RFC6378  RFC6388  RFC6424  (mLDP)
- RFC5316  RFC5392  etc
- RFC2328  RFC2370
- RFC1195  RFC3784  RFC4205
- RFC5880  RFC5884

(depending on how one counts the number of RFCs may vary)

MPLS Protocols

+----------+----------+----------+----------+----------+
<table>
<thead>
<tr>
<th>LDP</th>
<th>OSPF</th>
<th>ISIS</th>
<th>RSVP-TE</th>
<th>LSP-PING</th>
</tr>
</thead>
<tbody>
<tr>
<td>RFC5036</td>
<td>RFC2328</td>
<td>RFC1195</td>
<td>RFC5712</td>
<td>RFC6424</td>
</tr>
<tr>
<td>RFC6388</td>
<td></td>
<td></td>
<td>RFC5711</td>
<td>RFC4379</td>
</tr>
<tr>
<td>RFC5919</td>
<td></td>
<td></td>
<td>RFC5710</td>
<td></td>
</tr>
<tr>
<td>RFC5918</td>
<td></td>
<td></td>
<td>RFC5330</td>
<td></td>
</tr>
<tr>
<td>RFC5561</td>
<td></td>
<td></td>
<td>RFC4875</td>
<td></td>
</tr>
<tr>
<td>(mLDP)</td>
<td></td>
<td></td>
<td>RFC3209</td>
<td></td>
</tr>
<tr>
<td>etc</td>
<td></td>
<td></td>
<td>etc</td>
<td></td>
</tr>
</tbody>
</table>


Andersson & Martinotti Expires July 28, 2012

[Page 9]
<table>
<thead>
<tr>
<th>MPLS-OAM</th>
<th>OSPF-TE</th>
<th>ISIS-TE</th>
<th>BFD</th>
<th>NMS Conf</th>
</tr>
</thead>
<tbody>
<tr>
<td>RFC6378</td>
<td>RFC2370</td>
<td>RFC3784</td>
<td>RFC5880</td>
<td>nil</td>
</tr>
<tr>
<td>RFC6375</td>
<td>RFC3471</td>
<td>RFC5316</td>
<td>RFC5884</td>
<td></td>
</tr>
<tr>
<td>RFC6374</td>
<td>RFC3473</td>
<td>RFC4205</td>
<td></td>
<td></td>
</tr>
<tr>
<td>RFC6372</td>
<td>RFC4203</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>RFC6370</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>RFC5586</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>etc</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>MPLS-DP</th>
<th>MPLS-NM</th>
<th>BGP</th>
</tr>
</thead>
<tbody>
<tr>
<td>RFC6178</td>
<td>RFC5718</td>
<td>RFC3107</td>
</tr>
<tr>
<td>RFC5960</td>
<td>RFC4368</td>
<td></td>
</tr>
<tr>
<td>RFC5462</td>
<td>RFC4220</td>
<td></td>
</tr>
<tr>
<td>RFC4182</td>
<td>RFC3815</td>
<td></td>
</tr>
<tr>
<td>RFC4023</td>
<td>RFC3814</td>
<td></td>
</tr>
<tr>
<td>RFC3473</td>
<td>RFC3813</td>
<td></td>
</tr>
<tr>
<td>etc</td>
<td>etc</td>
<td></td>
</tr>
</tbody>
</table>

MPLS Packages

<table>
<thead>
<tr>
<th>MPLS-TD 1</th>
<th>MPLS-TD 2</th>
<th>MPLS-TD P2MP 1</th>
<th>MPLS-TD P2MP 2</th>
<th>GMPLS 1</th>
</tr>
</thead>
<tbody>
<tr>
<td>ISIS</td>
<td>OSPF</td>
<td>ISIS</td>
<td>OSPF</td>
<td>OSPF-TE</td>
</tr>
<tr>
<td>LDP</td>
<td>LDP</td>
<td>LDP</td>
<td>LDP</td>
<td>RSVP-TE</td>
</tr>
</tbody>
</table>
### MPLS Aggregate Packages

<table>
<thead>
<tr>
<th>MPLS-TD</th>
<th>MPLS-TE</th>
<th>MPLS-TP</th>
</tr>
</thead>
<tbody>
<tr>
<td>MPLS-TD 1</td>
<td>MPLS-TE 1</td>
<td>MPLS-TP 1</td>
</tr>
<tr>
<td>MPLS-TD 2</td>
<td>MPLS-TE 2</td>
<td>MPLS-TP 2</td>
</tr>
<tr>
<td>----------</td>
<td>-----------</td>
<td>-----------</td>
</tr>
<tr>
<td></td>
<td>CD</td>
<td>MPLS-TP</td>
</tr>
<tr>
<td>MPLS-TD</td>
<td>MPLS-TE</td>
<td>CD</td>
</tr>
<tr>
<td>P2MP 1</td>
<td>P2MP 1</td>
<td>MPLS-TP</td>
</tr>
<tr>
<td>MPLS-TD</td>
<td>MPLS-TE</td>
<td>MPLS-TP</td>
</tr>
<tr>
<td>P2MP 2</td>
<td>P2MP 2</td>
<td>P2MP TP</td>
</tr>
</tbody>
</table>

---

MPLS Technology
2. Unified MPLS Requirements

This section lists the high-level requirements for Unified MPLS.

2.1. Full Interoperability and Interworking

One important aspect of Unified MPLS is to promote interoperability and interworking between different flavors of MPLS. Unified MPLS addresses the problem of cross-MPLS interoperation and interworking.

Requirement: Unified MPLS SHALL guarantee full interoperability between all MPLS packages, i.e. SHALL e.g. be possible to start an LSP in an MPLS-TP network, cross MPLS-TE network and terminate it in a MPLS-TD network.

2.2. Common Data Plane

Requirement: The actions taken on a MPLS encapsulated packet relative to the label on top of the label stack, SHALL be the same regardless if the LSP has been set up using any of the MPLS routing and signaling protocols or if it has been established manually or from an NMS.

2.3. Common OAM

OAM is defined as part of the data plane. One goal with Unified MPLS is that it shall be possible to use MPLS-TP OAM for all MPLS packages.

Requirement: Given that MPLS-TP OAM is used there SHALL NOT, from the point of view of an MPLS encapsulated packet traversing the data plane of a MPLS Network, be any differences depending on whether the LSP has been set up using any of the MPLS routing and signaling protocols or if it has been established manually or from an NMS.

2.4. Data Plane Agnostic

Requirement: The relationship between LSPs and the payload transported on those LSPs, i.e. PWs, LSPs and IP, SHALL be the same regardless if the payload is transported by one or the other of the MPLS packages.

2.5. Interworking

Requirement: There SHALL be no limitations in the possibilities to hand over payload from one LSP to another regardless of how the LSPs has been established, i.e. it SHALL be possible to carry a PW on a MPLS-TP or MPLS-TD segment and then hand it over to a MPLS-TE LSP and
vice versa.
3. Unified MPLS Overview

This section should contain a high-level overview of Unified MPLS. Here we discuss interfacing different mpls types!

QUESTION: Do we have to consider also co-existence of different packages (e.g. CD-MPLS-TP and MPLS-TE)? Do we want to consider a network being physically partitioned or logically too?

QUESTION: Do we want to describe the possibility of reusing tools/protocols defined into one package into other ones?

3.1. Unified MPLS Control Plane

Unified MPLS comprises a set of protocols for implementing a Control Plane for Network Layer, for Routing and LSP label distribution. We consider the following routing protocols: OSPF and ISIS and their versions with Traffic Engineering extensions. We also consider the following LSP label distribution protocols: LDP and RSVP-TE (and its extensions for MPLS-TP). Each package having a control plane uses a combination of them.

We also consider the possibility where PW level is setup and maintained via tLDP.

One aim of Unified MPLS is to consider the Control Plane implications of putting two network domains in relationship, where one or both of them implements a control plane; not all the combinations may interwork in principle. Further considerations are done in Section 4.4.

3.2. Unified MPLS Data Plane

Unified MPLS comprises a standard MPLS data plane [RFC3031]. The following information is relevant to Unified MPLS.

The forwarding functions comprise the mechanisms required for forwarding the encapsulated native service traffic over an MPLS network, for example, LSP and PW labels.

MPLS label switching operations and Time-to-Live (TTL) processing procedures are used, as defined in [RFC3031], [RFC3032], [RFC3443] and [RFC5921].

SS-PW and MS-PW forwarding operations are used, as defined in [RFC3985] and [RFC5659].

Per-platform label space is used for PWs. Either per-platform, per-
interface, or other context-specific label space [RFC5331] may be used for LSPs.

MPLS forwarding is based on the label that identifies the path (LSP or PW). The label value specifies the processing operation to be performed by the next hop at that level of encapsulation. A swap of this label is an atomic operation in which the contents of the packet after the swapped label are opaque to the forwarder. The only event that interrupts a swap operation is TTL expiry. TTL expiry, which is a fundamental architectural construct of MPLS to be taken into account when designing protocol extensions (such as those for OAM) that require packets to be sent to an intermediate LSR.

Further processing to determine the context of a packet occurs when a swap operation is interrupted in this manner, or a pop operation exposes a specific reserved label at the top of the stack (including the GAL). Otherwise, the packet is forwarded according to the procedures in [RFC3032].

MPLS Differentiated Services (Diffserv) architecture [RFC3270] is the suggested mode of supporting Quality of Services in Unified MPLS. Both E-LSP and L-LSP MPLS Diffserv modes are supported.

Multicast MPLS is left for further study within Unified MPLS.

3.3. Unified MPLS Management

Here we want to describe the possible methods for Management of Unified MPLS. Commonly used methods, such as CLI, may be used in conjunction with LCT, EM, NMS, as described in [RFC5950]. Most important functions related to Management are: provisioning of network layer and service layer, fault and performance monitoring, inventory.

In several profiles of MPLS, provisioning of network layer and service layer is done by means of Control Plane. Where Management plane only is used for provisioning (e.g. in ND-MPLS-TP), routing, Path Computation Element, label distribution, configuration of the parameters for network and service layer are done via network management tools.

Where several subnetworks are deployed, it may happen that only a subset of them is operated via Management, while another is operated via Control Plane. The dependencies of the two has to be analysed.
3.4. Unified MPLS OAM

OAM is used to perform functions of operation, administration and maintenance (a comprehensive description of OAM functions is described in [RFC6291]). These functions are applicable both to LSP and PW. Among all the functions provided by OAM, fault management and performance monitoring are used to have a common method to verify the behavior of a network domain and a single mean to monitor the different kinds of service being provided.

OAM for MPLS has been defined in [LIST-OF-RFC-OF-OAM-FOR-LSP], especially because of the requirements of MPLS-TP. OAM for PW has been defined in [LIST-OF-RFC-OF-OAM-FOR-PW].

One critical objective of Unified MPLS is to describe how these tools can be used in a multi-domain MPLS network, where different packages can be deployed.
4. Interfaces and Interworking

Where two or more subnetworks (or profiles instances) are deployed within an operator network, it is important to consider the interface between each pair of them. Peering relationship is where the two subnetworks are at the same level with respect to the end-to-end service being provided. Peering relationship may happen at LSP, PW or client service level. Overlay relationship is where one subnetwork is client of the other one, so that the server subnetwork provides connectivity to part of the client one. In this section we discuss the most important interfaces and subnetwork interworking.

4.1. Interfaces

Several packages are considered within MPLS Technology. Anyhow there is not yet a guideline on how to use a combination of different packages in a single operator network. With Unified MPLS we want to describe the interfaces between the different packages. The number of the possible interfaces is quite high, so a brute force approach for their description is not beneficial to the reader. An objective of Unified MPLS is to start the analysis with the most common interworking scenarios.

The interface between two domains may be a node or a link. The two cases may lead to different requirements and behaviors at the interface. It may happen that one or several interfaces (of the same kind) happen between two domains.

4.2. Network Layer Interworking

This document uses the term Network Layer in the same sense as it is used in [RFC3031] and [RFC3032].

Two peering network domains (each one implementing one instance of any MPLS package) interwork at Network Layer when they are configured to interconnect at LSP or PW level. LSP stitching and MS-PW are used, depending on the level chosen. LSP stitching may be used for any client signal encapsulated into the end-to-end LSP (MPLS, PW, IP), while MS-PW can only be used for PW encapsulated signals.

4.3. Service Layer Interworking

Service interface (UNI and NNI) reference models are described in [RFC5921] and further analyzed in [RFC6125].

Two peering network domains (each one implementing one instance of any MPLS package) interwork at Service Layer when they are configured to interconnect at client service level. In case of border link
interconnecting the two domains, it can be named UNI in [RFC6125] terminology.

In case of border node interconnecting the two domains, Termination mode applies.

4.4. Network Overlay

When two network domains are in Overlay relationship, in principle there isn’t any interworking between them. Anyhow there are few aspects to be taken into account, at least from CP and Fault Management point of view.

CP aspects has to be considered when the client network domain is CD (it implements a CP). If the server network domain is CD, there is the possibility to interact between the two CPs. If the server network is ND, then the resources for the client domain must be pre-allocated.

In case OAM tool set is used in both client and server network domain, there is the possibility to propagate fault indication, happened within the server domain, into the client domain.

One Unified MPLS aim is also to specific constraints given by the used packages (E.g. if a domain using MPLS-TE is client of a domain using MPLS-TD, bandwidth requirements of the client layer may not be guaranteed.)
5. MPLS Resiliency

This section will discuss different aspects of MPLS resiliency, like protection, survivability and recovery.

5.1. MPLS Recovery

Note: Text needed!

5.2. MPLS Protection

Note: Text needed!

5.3. MPLS Survivability

Note: Text needed!
6. Other Unified MPLS Considerations

It is quite possible that Unified MPLS should also address a number of other aspects of MPLS.
7. Security Considerations

Security considerations to be added.
8. IANA considerations

There are no IANA considerations for this document (at least currently) an IANA section will be added as necessary.
9. Acknowledgments

We have received valuable feedback from several people working with adapting and utilizing MPLS in their networks.
10. References

10.1. Normative References


10.2. Informative references

[I-D.bryant-mpls-tp-jwt-report]

Authors’ Addresses

Loa Andersson
Ericsson

Email: loa@pi.nu

Riccardo Martinotti
Ericsson

Email: riccardo.martinotti@ericsson.com
Abstract

Shared mesh protection is a common protection and recovery mechanism in transport networks, where multiple paths can share the same set of network resources for protection purposes.

In the context of MPLS-TP, it has been explicitly requested as a part of the overall solution (Req. 67, 68 and 69 in RFC5654 [RFC5654]).

It’s important to note that each MPLS-TP LSP may be associated with transport network resources. In event of network failure, it may require explicit activation on the protecting paths before switching user traffic over.

In this memo, we define a lightweight signaling mechanism for protecting path activation in shared mesh protection-enabled MPLS-TP networks.

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 September 12, 2012.

Copyright Notice

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

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

1. Introduction ................................................. 4
2. Acronyms ..................................................... 4
3. Solution Overview ........................................... 5
   3.1. Protection Switching ................................... 7
   3.2. Operation Overview ................................... 8
4. SMP Message and Action Definition ............................ 9
   4.1. Protection Switching Control (PSC) Logic ................ 9
   4.2. SMP Action Types ..................................... 11
   4.3. PSC Signal to SMP Action Mapping .................... 11
5. Protocol Definition ........................................... 12
   5.1. Message Encapsulation .................................. 13
   5.2. Reliable Messaging ..................................... 14
   5.3. Message Scoping ....................................... 14
6. Processing Rules ............................................... 15
   6.1. Enable a Protecting Path ............................... 15
   6.2. Disable a Protecting Path ............................. 15
   6.3. Get Protecting Path Status ............................ 16
   6.4. Preemption ............................................ 16
7. Security Consideration ....................................... 16
8. IANA Considerations .......................................... 17
9. Acknowledgments .............................................. 17
10. References .................................................. 17
    10.1. Normative References ................................ 17
    10.2. Informative References ............................... 17
Authors’ Addresses .............................................. 18
1. Introduction

Shared mesh protection (SMP) is a common traffic protection mechanism in transport networks, where multiple paths can share the same set of network resources for protection purposes.

In the context of MPLS-TP, it has been explicitly requested as a part of the overall solution (Req. 67, 68 and 69 in RFC5654 [RFC5654]). Its operation has been further outlined in Section 4.7.6 of MPLS-TP Survivability Framework.

It’s important to note that each MPLS-TP LSP may be associated with transport network resources. In event of network failure, it may require explicit activation on the protecting paths before switching user traffic over.

In this memo, we define a lightweight signaling mechanism for protecting path activation in shared mesh protection-enabled MPLS-TP networks.

Here are the key design goals:

1. Fast: The protocol is to activate the previously configured protecting paths in a timely fashion, with minimal transport and processing overhead. The goal is to support 50msec end-to-end traffic switch-over in large transport networks.

2. Reliable message delivery: Activation and deactivation operation have serious impact on user traffic. This requires the protocol to adapt a low-overhead reliable messaging mechanism. The activation messages may either traverse through a "trusted" transport channel, or require some level of built-in reliability mechanism.

3. Modular: Depending on deployment scenarios, the signaling may need to support functions such as preemption, resource re-allocation and bi-directional activation in a modular fashion.

2. Acronyms

This draft uses the following acronyms:

- SMP: Shared Mesh Protection
- LO: Lockout of protection
o DNR: Do not revert
o FS: Forced Switch
o SF: Signal Fail
o SD: Signal Degrade
o MS: Manual Switch
o NR: No Request
o WTR: Wait-to-Restore
o EXER: Exercise
o RR: Reverse Request
o ACK: Acknowledgement
o NACK: Negative Acknowledgement
o G-ACh: Generic Associated Channel
o MPLS-TP Transport Profile for MPLS

3. Solution Overview

In this section, we describe the SMP operation in the context of MPLS-TP networks, and outline some of the relevant definitions.

We refer to the figure below for illustration:
Working paths: \( X = \{A, B, C, D\} \), \( Y = \{H, I, J, K\} \)

Protecting paths: \( X' = \{A, E, F, G, D\} \), \( Y' = \{H, E, F, G, K\} \)

The links between E, F and G are shared by both protecting paths. All paths are established via MPLS-TP control plane prior to network failure.

All paths are assumed to be bi-directional. An edge node is denoted as a headend or tailend for a particular path in accordance to the path setup direction.

Initially, the operators setup both working and protecting paths. During setup, the operators specify the network resources for each path. The working path \( X \) and \( Y \) will configure the appropriate resources on the intermediate nodes, however, the protecting paths, \( X' \) and \( Y' \), will reserve the resources on the nodes, but won’t occupy them.

Depending on network planning requirements (such as SRLG), \( X' \) and \( Y' \) may share the same set of resources on node E, F and G. The resource assignment is a part of the control-plane CAC operation taking place on each node.

At some time, link B-C is cut. Node A will detect the outage, and initiate activation messages to bring up the protecting path \( X' \). The intermediate nodes, E, F and G will program the switch fabric and configure the appropriate resources. Upon the completion of the activation, A will switch the user traffic to \( X' \).

The operation may have extra caveat:

1. Preemption: Protecting paths \( X' \) and \( Y' \) may share the same resources on node E, F or G due to resource constraints. \( Y' \) has higher priority than that of \( X' \). In the previous example, \( X' \) is up and running. When there is a link outage on I-J, H can activate its protecting path \( Y' \). On E, F or G, \( Y' \) can take over the resources from \( X' \) for its own traffic. The behavior is acceptable with the condition that A should be notified about the preemption action.

2. Over-subscription (1:N): A unit of network resource may be reserved by one or multiple protecting paths. In the example, the network resources on E-F and F-G are shared by two protecting paths, \( X' \) and \( Y' \). In deployment, the over-subscription ratio is an important factor on network resource utilization.
3.1. Protection Switching

The entire activation and switch-over operation need to be within the range of milliseconds to meet customer’s expectation. This section illustrates how this may be achieved on MPLS-TP-enabled transport switches. Note that this is for illustration of protection switching operation, not mandating the implementation itself.

The diagram below illustrates the operation:

```
+---------------+   +---------------+   +---------------+
| Control       |---| MPLS-TP       |---| Control       |
| <--- Signaling---|--| Control Plane|---|--| Signaling--->
+---------------+   +---------------+   +---------------+
     /                 / (MPLS label assignment)
    /                 /
   /                 /
+-------+   +-------+   +-------+   +-------+
| Activation|---| Line Module|---| Switch Fabric|---| Line Module|---| Activation|
| <--- Messages---|--|-- Module|---|-- Module|---|-- Messages--->
```

Typical MPLS-TP user flows (or, LSP’s) are bi-directional, and setup as co-routed or associated tunnels, with a MPLS label for each of the upstream and downstream traffic. On this particular type of transport switch, the control-plane can download the labels to the line modules. Subsequently, the line module will maintain a label lookup table on all working and protecting paths.

Upon the detection of network failure, the headend nodes will transmit activation messages along the MPLS LSP’s. When receiving the messages, the line modules can locate the associated protecting path from the label lookup table, and perform activation procedure by programming the switching fabric directly. Upon its success, the line module will swap the label, and forward the activation messages to the next hop.

In summary, the activation procedure involves efficient path lookup and switch fabric re-programming.

To achieve the tight end-to-end switch-over budget, it’s possible to implement the entire activation procedure with hardware-assistance (such as in FPGA or ASIC). The activation messages are encapsulated with a MPLS-TP Generic Associated Channel Header (GACH) [RFC5586]. Detailed message encoding is explained in later sections.
3.2. Operation Overview

To achieve high performance, the activation procedure is designed to be simple and straightforward on the network nodes.

In this section, we describe the activation procedure using the same figure shown before:

```
----- B -------- C ----
    /                    \
   /                      \
  A                        D
  /                        /\ 
 /                         / \
== E == F == G ==
   /               \
  /                 \
 H                  K
  /                    \
  ----- I -------- J ----
```

Working paths: $X = \{A, B, C, D\}$, $Y = \{H, I, J, K\}$

Protecting paths: $X' = \{A, E, F, G, D\}$, $Y' = \{H, E, F, G, K\}$

Upon the detection of working path failure, the edge nodes, $A$, $D$, $H$ and $K$ may trigger the activation messages to activate the protecting paths, and redirect user traffic immediately after.

We assume that there is a consistent definition of priority levels among the paths throughout the network. At activation time, each node may rely on the priority levels to potentially preempt other paths.

When the nodes detect path preemption on a particular node, they should inform all relevant nodes to free the resources by sending out notification messages. Upon the reception of notification messages, the relevant nodes will send out de-activation messages.

To optimize traffic protection and resource management, each headend may periodically poll the protecting paths about resource availability. The intermediate nodes have the option to inform the current resource utilization.

Note that, upon the detection of a working path failure, both headend
and tailend may initiate the activation simultaneously (known as bi-directional activation). This may expedite the activation time. However, both headend and tailend nodes need to coordinate the order of protecting paths for activation, since there may be multiple protecting paths for each working path (i.e., 1:N protection). For clarity, we will describe the operation from headend in the memo. The tailend operation will be available in the subsequent revisions.

4. SMP Message and Action Definition

4.1. Protection Switching Control (PSC) Logic

Protection switching processes the local triggers described in requirements 74-79 of [RFC5654] together with inputs received from the tailend node. Based on these inputs the headend will take SMP actions, and transmit different protocol messages.

Here, we reuse the switching control logic described in MPLS Linear Protection, with the following logical decomposition at headend node:
The Local Request logic unit accepts the triggers from the OAM, external operator commands, from the local control plane (when present), and the Wait-to-Restore timer. By considering all of these local request sources it determines the highest priority local request. This high-priority request is passed to the PSC Control logic, that will cross-check this local request with the information received from the tailend node. The PSC Control logic uses this input to determine what actions need to be taken, e.g. local actions at the headend, or what message should be sent to the tailend node.

Specifically, the signals could be the following:

- **Clear** - if the operator cancels an active local administrative command, i.e. LO/FS/MS.

- **Lockout of Protection (LO)** - if the operator requested to prevent switching data traffic to the protection path, for any purpose.

- **Signal Fail (SF)** - if any of the Server Layer, Control plane, or OAM indications signaled a failure condition on either the protection path or one of the working paths.

- **Signal Degraded (SD)** - if any of the Server Layer, Control plane, or OAM indications signaled a degraded transmission condition on either the protection path or one of the working paths.

- **Forced Switch (FS)** - if the operator requested that traffic be switched from one of the working paths to the protection path.

- **Manual Switch (MS)** - if the operator requested that traffic be switched from the working path to the protection path. This is only relevant if there is no currently active fault condition or Operator command.

- **WTR Expires** - generated by the WTR timer completing its period. If none of the input sources have generated any input then the request logic should generate a No Request (NR) request.

In addition to the local requests, the PSC Control Logic SHALL accept PSC messages from the tailend node of the transport path. Remote messages indicate the status of the transport path from the viewpoint of the tailend nodes. The remote requests may include remote LO, SF, SD, FS, MS, WTR and NR.

Much of the signal definition is further described in ITU G.709 and G.873.1.
4.2. SMP Action Types

As shown in the previous section, SMP requires four action types throughout the operation:

- **ACTIVATION**: This action is triggered by the head-end (or tail-end) to activate a protecting connection. The intermediate nodes need to propagate this towards the other end of the protecting connection.

- **DE-ACTIVATION**: This action is used to deactivate a particular protecting connection. This can be originated by one end of a protecting connection (i.e. head-end, or tail-end). The intermediate nodes need to propagate this towards the other end of the protecting connection.

- **QUERY**: This action is used when an operator decides to query a particular protecting connection.

- **NOTIFICATION**: SMP operation requires the coordination between nodes. The coordination takes place in two occasions:
  1. The activation/de-activation is initiated at the headend (tailend) nodes. To avoid potential mis-connection, the user traffic cannot be switched on to the protecting connection until the reception of an acknowledgement from the tailend (headend) nodes.
  2. If an intermediate node cannot process the activation requests, due to lack of resources or preemption levels, it needs to report the failure to the request originators.

It is conceivable that this message can be used to report the location of the fault, with respect to a protecting connection so that the head-end may use this information as one of the criteria for restoring the working transport entity. The fault location could be used by the head-end node to select among a list of possible protecting connections associated with the working connection (i.e. avoid the faulty location), or to determine that none of the provisioned protecting connections is usable at the time the failure is reported and then fallback to restoring the working connection.

4.3. PSC Signal to SMP Action Mapping

In SMP operation, there is the action-signal mapping:

- Activation action: FS, SF, SD, MS
5. Protocol Definition

Each SMP message has the following format:

<table>
<thead>
<tr>
<th>Ver</th>
<th>Request</th>
<th>Rsv</th>
<th>Reserved</th>
<th>Status</th>
<th>Seq</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

- Version: 1
- Request:
  * 1111b: Lockout of Protection (LO)
  * 1110b: Forced Switch (FS). This triggers activation
  * 1100b: Signal Fail (SF). This triggers activation
  * 1011b: Acknowledgement (ACK). This is to acknowledge a successful activation/de-activation request
  * 1010b: Signal Degrade (SD). This triggers activation
  * 1001b: Negative Acknowledgement (NACK). This is to report failure in activation/de-activation process.
  * 1000b: Manual Switch (MS). This triggers activation
  * 0110b: Wait-to-Restore (WTR). Used for revertive switching
  * 0100b: Exercise (EXER). Triggers SMP query
  * 0001b: Do Not Revert (DNR). Used for revertive switching
  * 0000b: No Request (NR). This triggers de-activation
0  R: Revertive field
   * 0: non-revertive mode
   * 1: revertive mode

0  Rsv/Reserved: This field is reserved for future use

0  Status: this informs the status of the AMP activation. This field is only relevant with ACK and NACK requests. Specifically, the Status Code has the following encoding value and definition:
   * 1: end-to-end ack
   * 2: hop-to-hop ack
   * 3: no such path
   * 4: no more resource for the path
   * 5: preempted by another path
   * 6: system failure
   * 7: shared resource has been taken by other paths

0  Seq: This uniquely identifies a particular message. This field is defined to support reliable message delivery

Note that the message format and naming convention are very similar to that of MPLS linear protection [6] and ITU G.873.1.

5.1. Message Encapsulation

SMP messages use MPLS labels to identify the paths. Further, the messages are encapsulated in GAL/GACH:

```
0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|                      MPLS Label stack                         |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|                          GAL                                  |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
| 0 0 0 1|Version|    Reserved    | Activation Channel Type     |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|                      Activation Message Payload               |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
```
GAL is described in [RFC5586]. Activation Channel Type is the GACH channel number assigned to the protocol. This uniquely identifies the activation messages.

Specifically, the messages have the following message format:

```
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
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|                Label                  |  Exp |S|    TTL        |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|                Label (13)             |  Exp |S|      TTL      |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|0 0 0 1|Version|   Reserved    | Activation Channel Type |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|Ver|Request|Rsv|R|   Reserved   |     Status    |     Seq       |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
```

5.2. Reliable Messaging

The activation procedure adapts a simple two-way handshake reliable messaging. Each node maintains a sequence number generator. Each new sending message will have a new sequence number. After sending a message, the node will wait for a response with the same sequence number.

Specifically, upon the generation of activation, de-activation, query and notification messages, the message sender expects to receive acknowledgement in reply with same sequence number.

If a sender is not getting the reply within a time interval, it will retransmit the same message with a new sequence number, and starts to wait again. After multiple retries (by default, 3), the sender will declare activation failure, and alarm the operators for further service.

5.3. Message Scoping

Activation signaling uses MPLS label TTL to control how far the message would traverse. Here are the processing rules on each intermediate node:

- On receive, if the message has label TTL = 0, the node must drop the packet without further processing.
o The receiving node must always decrement the label TTL value by one. If TTL = 0 after the decrement, the node must process the message. Otherwise, the node must forward the message without further processing (unless, of course, the node is headend or tailend)

o On transmission, the node will adjust the TTL value. For hop-by-hop messages, TTL = 1. Otherwise, TTL = 0xFF, by default.

6. Processing Rules

6.1. Enable a Protecting Path

Upon the detection of network failure (SF/SD/FS) on a working path, the headend node identifies the corresponding MPLS-TP label and initiates the protection switching by sending an activation message.

The activation messages always use MPLS label TTL = 1 to force hop-by-hop process. Upon reception, a next-hop node will locate the corresponding path and activate the path.

If the activation message is received on an intermediate node, due to label TTL expiry, the message is processed and then propagated to the next hop of the MPLS TP LSP, by setting the MPLS TP label TTL = 1.

The headend node will declare the success of the activation only when it gets a positive reply from the tailend node. This requires that the tailend nodes must reply the messages with ACK to the headend nodes in all cases.

If the headend node is not receiving the acknowledgement within a time internal, it will retransmit another activation message with a different Seq number.

If the headend node is not receiving a positive reply within a longer time interval, it will declare activation failure.

If an intermediate node cannot activate a protecting path, it will reply a message with NACK to report failure. When the headend node receives the message for failure, it must initiate the de-activation messages to clean up networks resources on all the relevant nodes on the path.

6.2. Disable a Protecting Path

The headend removes the network resources on a path by sending the de-activation messages.
In the message, the MPLS label represents the path to be de-activated. The MPLS TTL is one to force hop-by-hop processing.

Upon reception, a node will de-activate the path, by freeing the resources from the data-plane.

As a part of the clean-up procedure, each de-activation message must traverse through and be processed on all the nodes of the corresponding path. When the de-activation message reaches to the tailend node, the tailend is required to reply with an acknowledgement message to the headend.

The de-activation process is complete when the headend receives the corresponding acknowledgement message from the tailend.

6.3. Get Protecting Path Status

The operators have the option to trigger the query messages from the headend to check on the protecting path periodically or on-demand. The process procedure on each node is very similar to that of the activation messages on the intermediate nodes, except the query messages should not trigger any network resource re-programming. Upon reception, the node will check the availability of resources.

If the resource is no longer available, the node will reply an NACK with error conditions.

6.4. Preemption

The preemption operation typically takes place when processing an activation message. If the activating network resources have been used by another path and carrying user traffic, the node needs to compare the priority levels.

If the existing path has higher priority, the node needs to reject the activation request by sending an ACK to the corresponding headend to inform the unavailability of network resources.

If the new path has higher priority, the node will reallocate the resource to the new path, and send an ACK to old path’s headend node to inform about the preemption.

7. Security Consideration

The protection activation takes place in a controlled networking environment. Nevertheless, it is expected that the edge nodes will encapsulate and transport external traffic into separated tunnels,
and the intermediate nodes will never have to process them.

8. IANA Considerations

Activation messages are encapsulated in MPLS-TP with a specific GACH channel type that needs to be assigned by IANA.

9. Acknowledgments

Authors like to thank Eric Osborne, Lou Berger, Nabil Bitar and Deborah Brungard for detailed feedback on the earlier work, and the guidance and recommendation for this proposal.

We also thank Maneesh Jain, Mohit Misra, Yalin Wang, Ted Sprague, Ann Gui and Tony Jorgenson for discussion on network operation, feasibility and implementation methodology.

During ITU-T SG15 Interim meeting in May 2011, we have had long discussion with the G.SMP contributors, in particular Fatai Zhang, Bin Lu, Maarten Vissers and Jeong-dong Ryoo. We thank their feedback and corrections.

10. References

10.1. Normative References


10.2. Informative References


Authors’ Addresses

Ping Pan
(Infinera)

Email: ppan@infinera.com

Rajan Rao
(Infinera)

Email: rrao@infinera.com

Biao Lu
(Infinera)

Email: blu@infinera.com

Luyuan Fang
(Cisco)

Email: lufang@cisco.com

Andrew G. Malis
(Verizon)

Email: andrew.g.malis@verizon.com

Fatai Zhang
(Huawei)

Email: zhangfatai@huawei.com
Sam Aldrin
(Huawei)

Email: sam.aldrin@huawei.com

Fei Zhang
(ZTE)

Email: zhang.fei3@zte.com.cn

Sri Mohana Satya Srinivas Singamsetty
(Tellabs)

Email: SriMohanS@Tellabs.com
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 a 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 September 6, 2012.

Copyright Notice

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

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

Internet Engineering Task Force Y. Shen
Internet-Draft Juniper Networks
Intended status: Standards Track Y. Kamite
Expires: September 6, 2012 NTT Communications Corporation
March 5, 2012

RSVP Setup Protection
draft-shen-mpls-rsvp-setup-protection-00

Shen & Kamite Expires September 6, 2012 [Page 1]
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 Attributes TLVs .................................. 5
       3.1.1. Protected LSP Sender IPv4 Address TLV .............. 5
       3.1.2. Protected LSP Sender IPv6 Address TLV .............. 6
3.2. PLR behavior ................................................ 6
3.3. MP behavior .................................................. 8
3.4. Local Revertive Mode ....................................... 9
4. IANA Considerations ........................................... 9
5. Security Considerations ....................................... 9
6. Acknowledgements ............................................. 9
7. References .................................................... 9
   7.1. Normative References ..................................... 9
   7.2. Informative References ................................... 10
Authors’ Addresses ............................................... 10
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 onto a bypass LSP upon a failure of the immediate downstream link or node. The establishment of such protection is normally triggered by receiving a Resv message. In link protection, the PLR must learn the label and address of the nexthop 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-nexthop router. The information is carried in Resv message.

Imagine a scenario where an LSP is being signaled, but its Path message carries an EXPLICIT_ROUTE object (ERO) that is pre-computed, 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 on this path has already failed, the signaling will halt at the router immediate upstream of the failure. This will still be the case even if there is an existing bypass LSP protecting the link or node for some other P2P or P2MP LSP. In other words, the LSP is not protected during 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 originate a PathErr message, so that the ingress router of the LSP can compute and signal a new path to avoid the failed link or node. However, this approach may not always be possible or desirable, as can be seen in the scenarios described below.

1. Pre-computed or explicitly defined paths. If the path is pre-computed or fixed, or the ingress router is incapable of re-computing the path, an alternative path will not be set up.

2. Requirements for LSP setup time. Control protocol convergence and path computation may introduce a significant delay, which may impact signaling performance for services that have a specific requirement for LSP setup time.

3. Sibling sub-LSPs of a P2MP LSP sharing the failed link. In this case, the existing bypass LSP will not be used, even if it is the preferred alternative path. For example, the LSP being signaled is a sub-LSP of a P2MP LSP, and it is expected to share the same downstream link with an existing sibling sub-LSP (sub-LSP of the same P2MP LSP). If the new sub-LSP is rerouted through another path, unnecessary traffic will be generated on 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 on the desired path, and 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 its 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 allows 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, and optionally the "node protection desired" and "bandwidth protection desired" flags in the SESSION_ATTRIBUTE object.

On a given 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, and hence reliably detect its status. For link protection, the strict next hop also indicates the merge point (MP), i.e. the egress router of the bypass LSP to be used for rerouting the LSP. For node protection, the strict next-next ERO hop indicates the MP.

During setup protection operation, the PLR signals a backup LSP by tunneling a Path message through the bypass LSP. Unlike the normal facility-backup FRR, this Path message carries additional information of the protected LSP (Section 3.1). When the MP receives the Path message, it terminates the backup LSP, and also 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 is established end to end, with the backup LSP being tunneled through the bypass LSP from the PLR to the MP. The RSVP states on the PLR and the MP and the RSVP messages generated by these routers are the same as those in a normal facility-backup FRR scenario.

After the failed link or node is restored, the PLR MAY revert the LSP to the primary path. This is referred to as local revertive mode, as described 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 3.3.

3.1. 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.

   o Protected LSP Sender IPv4 Address TLV
   o Protected LSP Sender IPv6 Address TLV

Both TLVs are carried by the LSP_REQUIRED_ATTRIBUTES object in the Path message of a backup LSP.

3.1.1. Protected LSP Sender IPv4 Address TLV

The Protected LSP Sender IPv4 Address TLV is defined with type 2. It is allowed on LSP_REQUIRED_ATTRIBUTES object, and not allowed on LSP_ATTRIBUTES object. It is encoded as the following.

```
  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
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|             Type (TBD)          |           Length (8)          |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|                             Value                             |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
```

Figure 1

Type
3.1.2. Protected LSP Sender IPv6 Address TLV

The Protected LSP Sender IPv6 Address TLV is defined with type 3. It is allowed on LSP_REQUIRED_ATTRIBUTES object, and not allowed on LSP_ATTRIBUTES object. It is encoded as the following.

```
0                   1                   2                   3
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|             Type (TBD)        |           Length (20)          |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|                                                               |
| //                            Value                            |
|                                                               |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
```

Figure 2

Type

TBD

Length

20

Value

Original sender address in the IPv6 SENDER_TEMPLATE object of the protected LSP.

3.2. PLR behavior

When a router has a Path message to send out and the next ERO hop is a strict IPv4 or IPv6 prefix, the router validates the hop against the routing table, traffic engineering (TE) database, and/or a
topology database. If the hop is reachable and one hop away from the router, the Path message is sent as it is. Otherwise, there is a possibility that the hop has experienced a link or node failure.

The router determines this by searching for an existing bypass LSP that is protecting the hop. If the LSP being signaled desires link protection, the egress router of the bypass LSP (i.e. MP) must be the router that owns the IP prefix of the hop. If the LSP desires node protection, the next-next ERO hop of the LSP must also be a strict IP prefix, and the MP must be the router that owns this IP 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 LSP via the bypass LSP. If multiple such bypass LSPs exist, the PLR MAY select one based on bandwidth constraints or policy. 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. This helps against generating duplicate traffic on two separate bypass LSPs.

The PLR SHOULD NOT send the Path message of the protected LSP. Instead, it MUST create a backup LSP, and send a Path message for the backup LSP 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 SENDER_TEMPLATE object set to an address of the PLR. It MUST also carried a LSP_REQUIRED_ATTRIBUTES object containing a Protected LSP Sender IPv4 Address TLV or Protected LSP Sender IPv6 Address TLV.

Upon receiving a Resv message from the MP, the PLR brings 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, 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 MUST originate a PathErr message with code = 25 (notify error) and sub-code = 3 (tunnel locally repaired).

The PLR also installs a forwarding entry for the LSP. The nexthop 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 propagate the PathErr message upstream. 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 propagate the PathErr message upstream.

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.

In any case where the PLR tears down the protected LSP due to receipt of a PathTear message, state time-out, configuration, etc, the PLR MUST also tear down the backup LSP by sending a PathTear message through the bypass LSP.

3.3. MP behavior

When a MP receives the Path message of a backup LSP, it detects 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 the setup protection mode is disabled on the router, it MUST reject the Path message. In this case, the router recognizes the TLV, but does not support it. Therefore, it MUST originate a PathErr with code = 2 (policy control failure).

The MP then terminates the backup LSP, and re-creates the protected LSP. If the MP is the egress router of the LSP, it MUST also terminate the protected LSP. Otherwise, it MUST send a Path message of the protected LSP downstream. 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 a Protected LSP Sender IPv4 Address TLV or Protected LSP Sender IPv6 Address TLV.

The MP MUST allocate a label for the 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, the MP MAY allocate the same label as an existing sibling sub-LSP, in order to avoid traffic duplication.

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 case where the MP receives or originates a PathErr or ResvTear message of the protected LSP, it SHOULD translate it into a message of the backup LSP and send it to the PLR.

3.4. Local Revertive Mode

When the failed link or node is restored, the PLR MAY revert the protected LSP to its primary path, following the procedure of local revertive mode described in [RFC 4090].

4. IANA Considerations

This document defines two new RSVP Attributes TLVs. New type values need to 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


7.2. Informative References


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
Abstract

This memo defines a portion of the Management Information Base (MIB) for use with network management protocols. In particular it defines objects for managing MPLS Transport Profile (MPLS-TP) Linear Protection.

Status of This Memo

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

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

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

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

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

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

Copyright and License Notice

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

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

This memo defines a portion of the Management Information Base (MIB) for use with network management protocols. In particular, it defines objects for managing MPLS Transport Profile (MPLS-TP) Linear Protection.

This MIB module should be used for configuring and managing the MPLS TP linear protection for MPLS TP LSPs.

2. The Internet-Standard Management Framework

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

Managed objects are accessed via a virtual information store, termed the Management Information Base or MIB. MIB objects are generally accessed through the Simple Network Management Protocol (SNMP). Objects in the MIB are defined using the mechanisms defined in the Structure of Management Information (SMI). This memo specifies a MIB module that is compliant to the SMIv2, which is described in STD 58, RFC 2578 [RFC2578], STD 58, RFC 2579 [RFC2579] and STD 58, RFC 2580 [RFC2580].

3. Conventions

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].

4. Overview

RFC6378 defines the protocol to provide a linear protection switching mechanism for MPLS transport profile with protection domain as point-to-point LSP. The detailed protocol specification of MPLS transport profile linear protection is described in RFC
6378. This document specifies a MIB module for the LER that supports MPLS TP Linear protection (which includes 1:n protection architecture) and a MIB module that defines textual conventions.

5. Structure of the MIB Module

5.1. Textual Conventions

The following new textual conventions are defined in a separate MIB module in this document:
   - MplsLpsReq
   - MplsLpsFpathPath
   - MplsLpsCommand

5.2. The MPLS TP Linear Protection Subtree

MPLS-TP-LPS-MIB is the MIB module defined in this document, and it is put under mplsStdMIB.

5.3. The Notifications Subtree

Notifications are defined to inform the management station about switchover and mode mismatch of linear protection switching group. Two notifications are defined for this purpose. The notification, mplsLpsEventSwitchover is to inform the management station about the switchover of the active path and the notification mplsLpsEventModeMismatch is to inform the management station about the mismatch in the revertive mode across the end point of the protection domain.

5.4. The Table Structures

The MPLS TP Linear protection MIB module has 4 tables. The tables are as follows:

- mplsLpsConfigTable

  This table is used to configure MPLS TP Linear protection switching Group. The protection switching group is identified by mplsLpsConfigGroupName. The other attributes in this table can be used to configure properties that are specific to the protection switching group.

- mplsLpsStatusTable

  This table provides the current status information of mpls linear protection groups that have been configured on the system. When a protection group is configured or deleted in the mplsLpsConfigTable, then the corresponding row of that session in the mplsLpsStatusTable is, respectively, automatically created or deleted.

- mplsLpsMeConfigTable

  This table is used to associate the Maintenance Entities (MEs) to the protection switching group. The ME is identified by mplsOamIdMegIndex, mplsOamIdMeIndex and mplsOamIdMeMpIndex.

- mplsLpsMeStatusTable
This table provides the current information about the protection state of MEs that have been configured on the system. When a ME configured or deleted in the mplsLpsMeConfigTable, then the corresponding row of that session in the mplsLpsMeStatusTable is, respectively, automatically created or deleted.

6. Relationship to Other MIB Modules

6.1. Relationship to the MPLS OAM maintenance identifiers MIB module

There is a dependency between the MPLS-TP-LPS-MIB module and MPLS-OAM-ID-STD-MIB defined in draft-vkst-mpls-tp-oam-id-mib. The mplsOamIdMegIndex, mplsOamIdMeIndex and mplsOamIdMeMpIndex defined in mplsOamIdMeTable of MPLS-OAM-ID-STD-MIB is used as the index of the mplsLpsMeConfigTable defined in the MPLS-TP-LPS-MIB module. Each time that an entry is created in the mplsOamIdMeTable for which the LER supports MPLS TP Linear protection a row is created automatically in the mplsLpsMeConfigTable.

6.2. MIB modules required for IMPORTS

The MPLS-TP-LPS-MIB module requires following MIB modules for IMPORTS:

- SNMPv2-SMI defined in [RFC2578]
- RMON2-MIB defined in [RFC4502]
- SNMPv2-CONF defined in [RFC2580]
- SNMPv2-TC defined in [RFC2579]
- MPLS-OAM-ID-STD-MIB defined in [draft-vkst-mpls-tp-oam-id-mib]

7. Definitions

MPLS-TP-LPS-MIB DEFINITIONS ::= BEGIN

IMPORTS
   MODULE-IDENTITY, NOTIFICATION-TYPE, OBJECT-TYPE,
   Gauge32, Counter32, Integer32, transmission
   FROM SNMPv2-SMI

   TEXTUAL-CONVENTION, RowStatus,
   TimeStamp, StorageType
   FROM SNMPv2-TC

   SnmpAdminString
   FROM SNMP-FRAMEWORK-MIB

   MODULE-COMPLIANCE, OBJECT-GROUP, NOTIFICATION-GROUP
   FROM SNMPv2-CONF;

mplsLpsMIB MODULE-IDENTITY
   LAST-UPDATED "201201020000Z" -- January 24, 2012
   ORGANIZATION "Multiprotocol Label Switching (MPLS) Working Group"
   CONTACT-INFO
   "Editor:
   Kingston Smiler Selvaraj"
DESCRIPTION
"This management information module supports the configuration and management of MPLS TP linear protection groups."

REVISION
"201201020000Z" -- January 02, 2012
DESCRIPTION
"MPLS Protection Switching Group objects for LSP MEPs"

::= { mplsStdMIB xxx } -- xxx to be replaced with correct value

-- Top level components of this MIB module.
-- traps
mplsLpsMIBNotifications
OBJECT IDENTIFIER ::= { mplsLpsMIB 0 }

-- tables, scalars
mplsLpsMIBObjects
OBJECT IDENTIFIER ::= { mplsLpsMIB 1 }

-- conformance
mplsLpsMIBConformance
OBJECT IDENTIFIER ::= { mplsLpsMIB 2 }

MplsLpsReq ::= TEXTUAL-CONVENTION
STATUS current
DESCRIPTION
"This Textual Convention describes an object that stores the PSC Request field of the PSC control packet. The values are as follows

1110  Lockout of Protection
1100  Forced Switch
1010  Signal Fail (SF)
0111  Signal Degrade (SD)
0101  Manual Switch
0100  Wait-to-Restore
0001  Do Not Revert
0000  No Request"

REFERENCE
"Section 4.2.2 of RFC 6378"
SYNTAX      OCTET STRING (SIZE (2))

MplsLpsFPathPath ::= TEXTUAL-CONVENTION
STATUS current
DESCRIPTION
"This Textual Convention describes an object that stores the Fault Path (FPath) field and Data Path (Path) field of the PSC control packet.

FPath is located in the first octet and Path is located in the second octet. Bits are numbered from left to right.

The value and the interpretation of FPath field is as follows

2-255 for future extensions
1 the anomaly condition is on the working path
0 the anomaly condition is on the protection path

The value and the interpretation of Path field is as follows

2-255 for future extensions
1 protection path is transporting user data traffic
0 protection path is not transporting user data traffic

REFERENCE
"Section 4.2.5 and 4.2.6 of RFC 6378"

SYNTAX OCTET STRING (SIZE (2))

MplsLpsCommand ::= TEXTUAL-CONVENTION
STATUS current
DESCRIPTION "This command allows a user to perform any action over ME. If the protection command cannot be executed because an equal or higher priority request is in effect, an inconsistentValue error is returned.

The command values are:

noCmd

This value should be returned by a read request when no command has been written to the object in question since initialization. This value may not be used in a write operation. If noCmd is used in a write operation a wrongValue error is returned.

clear

Clears all of the commands listed below for the specified ME.

lockoutOfProtection

Prevents any of the working ME from switching to the protection ME. The specified ME should be the protection ME, otherwise an inconsistentValue error is returned.

forcedSwitchWorkToProtect

Switches the specified working ME to the protection path. If the protection ME is specified an inconsistentValue error is returned.

manualSwitchWorkToProtect

Switches the specified working ME to the protection ME."
If the protection ME is specified an inconsistentValue error is returned.

SYNTAX INTEGER {
  noCmd(1),
  clear(2),
  lockoutOfProtection(3),
  forcedSwitchWorkToProtect(4),
  manualSwitchWorkToProtect(5)
}

-- Start of MPLS Transport Profile Protection Switching
-- Table
--
-- MPLS TP Protection Switching Configuration Table
--
-- This table supports the addition, configuration and deletion
-- of MPLS TP Protection Switching groups.
--
mplsLpsConfig OBJECT IDENTIFIER ::= { mplsLpsMIBObjects 1 }

mplsLpsConfigGroups OBJECT-TYPE
  SYNTAX Gauge32
  MAX-ACCESS read-only
  STATUS current
  DESCRIPTION "The object hold the count of MPLS Protection Switching
groups. This count includes all rows in mplsLpsConfigTable,
regardless of the value of mplsLpsConfigRowStatus."
  ::= { mplsLpsConfig 1 }

mplsLpsConfigTable OBJECT-TYPE
  SYNTAX SEQUENCE OF MplsLpsConfigEntry
  MAX-ACCESS not-accessible
  STATUS current
  DESCRIPTION "This table lists the mpls linear protection groups that
 have been configured on the system."
  ::= { mplsLpsConfig 2 }

mplsLpsConfigEntry OBJECT-TYPE
  SYNTAX MplsLpsConfigEntry
  MAX-ACCESS not-accessible
  STATUS current
  DESCRIPTION "A conceptual row in the mplsLpsConfigTable."
  INDEX { mplsLpsConfigGroupName }
  ::= { mplsLpsConfigTable 1 }

MplsLpsConfigEntry ::= SEQUENCE {
  mplsLpsConfigGroupName         SnmpAdminString,
  mplsLpsConfigRowStatus         RowStatus,
  mplsLpsConfigMode              INTEGER,
  mplsLpsConfigRevertive         INTEGER,
  mplsLpsConfigProtectionScheme  INTEGER,
  mplsLpsConfigSdThreshold       Integer32,
  mplsLpsConfigWaitToRestore     Integer32,
  mplsLpsConfigCreationTime      TimeStamp,
  mplsLpsConfigStorageType       StorageType
}
mplsLpsConfigGroupName OBJECT-TYPE
SYNTAX      SnmpAdminString (SIZE (1..32))
MAX-ACCESS  not-accessible
STATUS      current
DESCRIPTION
   "Textual name represents the mpls tp protection group. Each Protection Group is identified by a unique protection group name."
::= { mplsLpsConfigEntry 1 }

mplsLpsConfigRowStatus OBJECT-TYPE
SYNTAX      RowStatus
MAX-ACCESS  read-create
STATUS      current
DESCRIPTION
   "This represents the status of the MPLS TP Linear Protection group Entry. This variable is used to create, modify, and/or delete a row in this table. An entry may not exist in the active state unless all objects in the entry have an appropriate value."
::= { mplsLpsConfigEntry 2 }

mplsLpsConfigMode OBJECT-TYPE
SYNTAX INTEGER {
   1+1(1),
   1:1(2),
   1:n(3)
}
MAX-ACCESS read-create
STATUS    current
DESCRIPTION
   "The architectural mode of the Protection group. This can either be 1+1, 1:1, 1:n.

1+1
   In the 1+1 protection scheme, a fully dedicated protection entity is allocated. Data traffic is copied and fed at the source to both the working and the protection entities. The traffic on the working and the protection entities is transmitted simultaneously to the sink of the protection domain, where selection between the working and protection entities is performed.

1:1
   In the 1:1 scheme, a protection path is allocated to protect against a defect, failure, or a degradation in a working path. In normal conditions, data traffic is transmitted over the working entity, while the protection entity functions in the idle state. If there is a defect on the working entity or a specific administrative request, traffic is switched to the protection entity.

1:n
   In case of 1:n linear protection, one protection entity is allocated to protect n working entities. The protection entity might not have sufficient resources to protect all the working entities that may be affected by fault conditions at a
specific time. In this case, in order to guaranteed protection, the protection entity should support enough capacity and bandwidth to protect any of the n working entities."

DEFVAL {1+1}
::= { mplsLpsConfigEntry 3 }

mplsLpsConfigRevert OBJECT-TYPE
SYNTAX INTEGER { nonrevertive(1), revertive(2) }
MAX-ACCESS read-create
STATUS current
DESCRIPTION
"This object represents the reversion mode of the Linear Protection Switching group. The reversion mode of protection mechanism may be either revertive or non-revertive.

nonrevertive

In non-revertive mode, after a service has been recovered, traffic will be forwarded on the recovery path

revertive

In revertive mode, after a service has been recovered, traffic will be redirected back onto the original working path."

DEFVAL { nonrevertive }
::= { mplsLpsConfigEntry 4 }

mplsLpsConfigProtectionScheme OBJECT-TYPE
SYNTAX INTEGER { bidirectional(1), unidirectional(2) }
MAX-ACCESS read-create
STATUS current
DESCRIPTION
"The object represents the operational scheme of protection switching group. The protection scheme may either be unidirectional or bidirectional.

bidirectional

In bidirectional protection scheme, both the directions will be switched simultaneously even if the fault applies to only one direction of the path.

unidirectional

In unidirectional protection scheme protection switching will be performed independently for each direction of a bidirectional transport path

This object may not be modified if the associated mplsLpsConfigRowStatus object is equal to active(1). "

DEFVAL {bidirectional}
::= { mplsLpsConfigEntry 5 }

mplsLpsConfigSdThreshold OBJECT-TYPE
SYNTAX Integer32 (1..9)
MAX-ACCESS read-create
STATUS current
DESCRIPTION
"This object holds the threshold value of the Signal Degrad.

When the MPLS DM OAM reaches this threshold value, the
Signal Degrad event will be given to this
protection domain.

This object may be modified if the associated
mplsLpsConfigRowStatus object is equal to active(1)."
DEFVAL { }
::= { mplsLpsConfigEntry 6 }

mplsLpsConfigWaitToRestore OBJECT-TYPE
SYNTAX Integer32 (0..720)
UNITS "seconds"
MAX-ACCESS read-create
STATUS current
DESCRIPTION
"This object hold the Wait To Restore timer value in
seconds.
The WTR timer is used to delay reversion of PSC state
to Normal state when recovering from a failure
condition on the working path when the protection
domain is configured for revertive behavior

This object may not be modified if the associated
mplsLpsConfigRowStatus object is equal to active(1)."
DEFVAL { 300 }
::= { mplsLpsConfigEntry 7 }

mplsLpsConfigCreationTime OBJECT-TYPE
SYNTAX TimeStamp
MAX-ACCESS read-only
STATUS current
DESCRIPTION
"The value of sysUpTime at the time the row was
created"
::= { mplsLpsConfigEntry 8 }

mplsLpsConfigStorageType OBJECT-TYPE
SYNTAX StorageType
MAX-ACCESS read-create
STATUS current
DESCRIPTION
"The storage type for this conceptual row.
Conceptual rows having the value 'permanent' need not
allow write-access to any columnar objects in the row."
DEFVAL { nonVolatile }
::= { mplsLpsConfigEntry 9 }

--
-- MPLS TP Linear Protection Switching Status Table
--
-- This table provides Protection Switching group statistics.
--

mplsLpsStatusTable OBJECT-TYPE
SYNTAX SEQUENCE OF MplsLpsStatusEntry
MAX-ACCESS not-accessible
STATUS current
DESCRIPTION
"This table provides status information about mpls linear protection groups that have been configured on the system."
::= { mplsLpsMIBObjects 2 }

mplsLpsStatusEntry OBJECT-TYPE
SYNTAX MplsLpsStatusEntry
MAX-ACCESS not-accessible
STATUS current
DESCRIPTION
"A conceptual row in the mplsLpsStatusTable."
AUGMENTS { mplsLpsConfigEntry }
::= { mplsLpsStatusTable 1 }

MplsLpsStatusEntry ::= SEQUENCE {
  mplsLpsStatusReqRcv                 MplsLpsReq,
  mplsLpsStatusReqSent               MplsLpsReq,
  mplsLpsStatusFpathPathRcv          MplsLpsFpathPath,
  mplsLpsStatusFpathPathSent         MplsLpsFpathPath,
  mplsLpsStatusModeMismatches        Counter32
}

mplsLpsStatusReqRcv OBJECT-TYPE
SYNTAX MplsLpsReq
MAX-ACCESS read-only
STATUS current
DESCRIPTION
"The current value of the PSC Request field received on more recent PSC packet"
::= { mplsLpsStatusEntry 1 }

mplsLpsStatusSent OBJECT-TYPE
SYNTAX MplsLpsReq
MAX-ACCESS read-only
STATUS current
DESCRIPTION
"The current value of the PSC Request field sent on the more recent PSC packet"
::= { mplsLpsStatusEntry 2 }

mplsLpsStatusPathFpathRcv OBJECT-TYPE
SYNTAX MplsLpsFpathPath
MAX-ACCESS read-only
STATUS current
DESCRIPTION
"The current value of the FPath and Path fields received on more recent PSC packet"
::= { mplsLpsStatusEntry 3 }

mplsLpsStatusPathFpathSent OBJECT-TYPE
SYNTAX MplsLpsFpathPath
MAX-ACCESS read-only
STATUS current
DESCRIPTION
"The current value of the FPath and Path fields sent on more recent PSC packet"
::= { mplsLpsStatusEntry 4 }
mplsLpsStatusModeMismatches OBJECT-TYPE
SYNTAX     Counter32
MAX-ACCESS read-only
STATUS     current
DESCRIPTION
    "This object hold the count of Mode Mismatch
    conditions."
::= { mplsLpsStatusEntry 5 }

--
-- MPLS Linear Protection ME Association Configuration Table
--
-- This table supports the addition, configuration and deletion
-- of MPLS Linear Protection Maintenance Entities in Protection
-- Switching groups.
--
mplsLpsMeConfigTable OBJECT-TYPE
SYNTAX      SEQUENCE OF MplsLpsMeConfigEntry
MAX-ACCESS  not-accessible
STATUS      current
DESCRIPTION
    "This table lists Maintenance Association that have been
    configured in Protection groups."
::= { mplsLpsMIBObjects 3 }

mplsLpsMeConfigEntry OBJECT-TYPE
SYNTAX      MplsLpsMeConfigEntry
MAX-ACCESS  not-accessible
STATUS      current
DESCRIPTION
    "A conceptual row in the mplsLpsMeConfigTable."
INDEX {mplsOamIdMegIndex, mplsOamIdMeIndex, mplsOamIdMeMpIndex}
::= { mplsLpsMeConfigTable 1 }

MplsLpsMeConfigEntry ::= SEQUENCE {
    mplsLpsMeConfigGroupName             SnmpAdminString,
    mplsLpsMeConfigRowStatus             RowStatus,
    mplsLpsMeConfigState                 INTEGER,
    mplsLpsMeConfigCommand               MplsLpsCommand,
    mplsLpsMeConfigStorageType           StorageType
}

mplsLpsMeConfigGroupName OBJECT-TYPE
SYNTAX     SnmpAdminString (SIZE (1..32))
MAX-ACCESS not-accessible
STATUS     current
DESCRIPTION
    "This object holds the Protection group name wherein
    this ME included in. If this ME is not part of a protection
    group this value is set to a string of size 0."
::= { mplsLpsMeConfigEntry 1 }

mplsLpsMeConfigRowStatus OBJECT-TYPE
SYNTAX     RowStatus
MAX-ACCESS read-create
STATUS     current
DESCRIPTION
    "The status of this Protection Switching ME entry."
An entry may not exist in the active state unless all objects in the entry have an appropriate value.

::= { mplsLpsMeConfigEntry 2 }

mplsLpsMeConfigState OBJECT-TYPE
SYNTAX INTEGER { primary(1), backup(2) }
MAX-ACCESS read-create
STATUS current
DESCRIPTION
"This object represents the operational state of the ME as either primary or backup"
::= { mplsLpsMeConfigEntry 3 }

mplsLpsMeConfigCommand OBJECT-TYPE
SYNTAX MplsLpsCommand
MAX-ACCESS read-write
STATUS current
DESCRIPTION
"Allows the initiation of an MPLS Linear protection command on the protection group and the ME specified by the index values. When read this object returns the last command written or noCmd if no command has been written to this ME since initialization. The return of the last command written does not imply that this command is currently in effect. This request may have been preempted by a higher priority local or remote request.

The value lockoutOfProtection should only be applied to the protection path / ME since that switch command prevents any of the working path / ME from switching to the protection path. Following the same logic, forcedSwitchProtectToWork and manualSwitchProtectToWork should only be applied to the protection.

forcedSwitchWorkToProtect and manualSwitchWorkToProtect should only be applied to a working ME."
::= { mplsLpsMeConfigEntry 4 }

mplsLpsMeConfigStorageType OBJECT-TYPE
SYNTAX StorageType
MAX-ACCESS read-create
STATUS current
DESCRIPTION
"The storage type for this conceptual row. Conceptual rows having the value 'permanent' need not allow write-access to any columnar objects in the row."
DEFVAL { nonVolatile }
::= { mplsLpsMeConfigEntry 5 }

--
-- MPLS Linear Protection ME Status Table
--
-- This table provides Protection Switching ME statistics.
--

mplsLpsMeStatusTable OBJECT-TYPE
SYNTAX SEQUENCE OF MplsLpsMeStatusEntry
MAX-ACCESS not-accessible
STATUS current
DESCRIPTION
"This table contains status information of all the ME
that are included in MPLS Protection groups.

::= { mplsLpsMIBObjects 4 }

mplsLpsMeStatusEntry OBJECT-TYPE
SYNTAX     MplsLpsMeStatusEntry
MAX-ACCESS not-accessible
STATUS      current
DESCRIPTION
"A conceptual row in the mplsLpsMeStatusTable."
AUGMENTS { mplsLpsMeConfigEntry }
::= { mplsLpsMeStatusTable 1 }

MplsLpsMeStatusEntry ::= SEQUENCE {
  mplsLpsMeStatusCurrent               BITS,
  mplsLpsMeStatusSignalDegrades        Counter32,
  mplsLpsMeStatusSignalFailures        Counter32,
  mplsLpsMeStatusSwitchovers           Counter32,
  mplsLpsMeStatusLastSwitchover        TimeStamp,
  mplsLpsMeStatusSwitchoverSeconds     Counter32,
  mplsLpsMeStatusDiscontinuityTime     TimeStamp
}

mplsLpsMeStatusCurrent OBJECT-TYPE
SYNTAX     BITS {
  lockedOut(0),
  sd(1),
  sf(2),
  switched(3),
  wtr(4)
}
MAX-ACCESS read-only
STATUS      current
DESCRIPTION
"Indicates the current state of the MA.

lockedOut

This bit, when it is set on a working ME or working path indicates that the working path is prevented from switching to the protection path.  When it is set on protection / backup path, this bit indicates that none of the working path (in case of 1:n) can switch to the protection path.

sd

This bit implies that signal degrade condition is in effect on this ME / path.

sf

This bit implies that signal failure condition is in effect on this ME / path.

switched

This bit is only applicable to the working ME / path.  It implies that the working path is currently switched to the protection path.

wtr
This bit implies that Wait-to-Restore state is in effect.

::= { mplsLpsMeStatusEntry 1 }

mplsLpsMeStatusSignalDegrades OBJECT-TYPE
SYNTAX Counter32
MAX-ACCESS read-only
STATUS current
DESCRIPTION "Represents the count of Signal Degrade conditions. This condition occurs when the DM exceeds the currently configured value of the relevant instance of mplsLpsConfigSdThreshold."
::= { mplsLpsMeStatusEntry 2 }

mplsLpsMeStatusSignalFailures OBJECT-TYPE
SYNTAX Counter32
MAX-ACCESS read-only
STATUS current
DESCRIPTION "Represents the count of Signal failure conditions. This condition occurs when the OAM running on this MA detects the Signal Fail event."
::= { mplsLpsMeStatusEntry 3 }

mplsLpsMeStatusSwitchovers OBJECT-TYPE
SYNTAX Counter32
MAX-ACCESS read-only
STATUS current
DESCRIPTION "Represents the count of SwitchOvers happened in this MA.

When the mplsLpsMeConfigState is primary, this object will return the number of times this path has switched to the protection path.

When the mplsLpsMeConfigState is backup, this object will return the number of times that any working paths has been switched back to the working path from this protection path.
::= { mplsLpsMeStatusEntry 4 }

mplsLpsMeStatusLastSwitchover OBJECT-TYPE
SYNTAX TimeStamp
MAX-ACCESS read-only
STATUS current
DESCRIPTION "This object holds the value of sysUpTime wherein the last switchover happened.

When the mplsLpsMeConfigState is primary, this object will return the value of sysUpTime when this path last completed a switchover. If this path has never switched to the protection line, the value 0 will be returned.

When the mplsLpsMeConfigState is backup, this object will return the value of sysUpTime the last time that a working path was switched back to the working path from this protection path. If no working path has ever
switched back to the working path from this protection path, the value 0 will be returned."

::= { mplsLpsMeStatusEntry 5 }

mplsLpsMeStatusSwitchoverSeconds OBJECT-TYPE
SYNTAX Counter32
MAX-ACCESS read-only
STATUS current
DESCRIPTION
"The cumulative Protection Switching Duration (PSD) time in seconds.

For a working path, this is the cumulative number of seconds that traffic was carried on the protection path.

For the protection path, this is the cumulative number of seconds that the protection path has been used to carry any working path traffic."

::= { mplsLpsMeStatusEntry 6 }

mplsLpsNotificationEnable OBJECT-TYPE
SYNTAX BITS {
  switchover(0),
  modeMismatch(1)
}
MAX-ACCESS read-write
STATUS current
DESCRIPTION
"Provides the ability to enable and disable notifications defined in this MIB.

switchover

Indicates mplsLpsEventSwitchover notifications should be generated.

modeMismatch

Indicates mplsLpsEventModeMismatch notifications should be generated."

DEFVAL { { } }
::= { mplsLpsMIBObjects 5 }

--
-- MPLS Linear Protection EVENTS
--

mplsLpsNotificationsPrefix OBJECT IDENTIFIER
 ::= { mplsLpsMIBNotifications 0 }

mplsLpsEventSwitchover NOTIFICATION-TYPE
OBJECTS { mplsLpsMeStatusSwitchovers, mplsLpsMeStatusCurrent }
STATUS current
DESCRIPTION
"An mplsLpsEventSwitchover notification is sent when the value of an instance of mplsLpsMeStatusSwitchovers increments."
::= { mplsLpsNotificationsPrefix 1 }
mplsLpsEventModeMismatch NOTIFICATION-TYPE
OBJECTS { mplsLpsStatusModeMismatches }
STATUS current
DESCRIPTION
   "An mplsLpsEventModeMismatch notification is sent when the
    value of an instance of mplsLpsStatusModeMismatches increments."
 ::= { mplsLpsNotificationsPrefix 2 }

8. Security Considerations

There are a number of management objects defined in this MIB module
with a MAX-ACCESS clause of read-write and/or read-create. Such
objects may be considered sensitive or vulnerable in some networks
in a non-secure environment without proper protection can have a
negative effect on network operations. These are the tables and
objects and their sensitivity/vulnerability:

Some of the readable objects in this MIB module (i.e., objects with a
MAX-ACCESS other than not-accessible) may be considered sensitive or
vulnerable in some network environments. It is thus important to
control even GET and/or NOTIFY access to these objects and possibly
to even encrypt the values of these objects when sending them over
the network via SNMP. These are the tables and objects and their
sensitivity/vulnerability:

SNMP versions prior to SNMPv3 did not include adequate security.
Even if the network itself is secure (for example by using IPsec),
even then, there is no control as to who on the secure network is
allowed to access and GET/SET (read/change/create/delete) the objects
in this MIB module.

It is RECOMMENDED that implementers consider the security features as
provided by the SNMPv3 framework (see [RFC3410], section 8),
including full support for the SNMPv3 cryptographic mechanisms (for
authentication and privacy).

Further, deployment of SNMP versions prior to SNMPv3 is NOT
RECOMMENDED. Instead, it is RECOMMENDED to deploy SNMPv3 and to
enable cryptographic security. It is then a customer/operator
responsibility to ensure that the SNMP entity giving access to an
instance of this MIB module is properly configured to give access to
the objects only to those principals (users) that have legitimate
rights to indeed GET or SET (change/create/delete) them.

9. IANA Considerations

To be added in a later version of this document.

10. Contributors

11. References

11.1. Normative References
11.2. Informative References


[MPLS-TP-OAM-ID-MIB] Sam Aldrin, M.Venkatesan, Kannan KV Sampath, Thomas D. Nadeau, Sami Boutros, Ping Pan, "MPLS-TP Operations, Administration, and Management (OAM) Identifiers Management Information Base (MIB)", Work in Progress, October 2011.

11.3. URL References

[idguidelines] IETF Internet Drafts editor, 
"http://www.ietf.org/ietf/1id-guidelines.txt".

[idnits] IETF Internet Drafts editor, 
"http://www.ietf.org/ID-Checklist.html".

[xml2rfc] XML2RFC tools and documentation, 
"http://xml.resource.org".
mLDP Node Protection
draft-wijnands-mpls-mldp-node-protection-00

Abstract

This document describes procedures to support node protection for Point-to-Multipoint and Multipoint-to-Multipoint Label Switched Paths (MP LSPs) built by LDP ("Label Distribution Protocol"), or simply mLDP. In order to protect a node N, the Point of Local Repair (PLR) LSR of N must learn the Merge Point (MPT) LSR(s) of node N such that traffic can be redirected to them in case node N fails. Redirecting the traffic around the failed node N depends on existing P2P LSPs originated from the PLR LSR to the MPT LSRs while bypassing LSR node N.

Status of this Memo

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

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

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

This Internet-Draft will expire on August 27, 2012.

Copyright Notice

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

This document describes procedures to support node protection for Point-to-Multipoint and Multipoint-to-Multipoint Label Switched Paths (MP-LSPs) built by LDP ("Label Distribution Protocol"), or simply mLDP. In order to protect a node N, the Point of Local Repair (PLR) of N must learn the Merge Point (MPT) LSR(s) of node N such that traffic can be redirected to them in case node N fails. Redirecting the traffic around the failed node N depends on existing P2P LSPs originating from the PLR LSR to the MPT LSR(s) while bypassing node N. The procedures to setup these P2P LSPs are outside the scope of this document, but one can imagine using RSVP-TE or LDP LFA based techniques to accomplish this.

There are different solutions for a PLR LSR to learn the downstream MPT LSR(s). One solution is documented in [I-D.zhao-mpls-mldp-protections]. This solution is based on ‘tunneling’ the MPT LSR(s) through node N via the existing LDP session towards the PLR, like as ships-in-the-night. The downside of that approach is that as soon as node N fails, no signaling is possible between the MPT LSR(s) and PLR LSR. A direct consequence of this is that the MPT LSR(s) have no mechanism to signal a withdraw to the PLR to stop forwarding packets after the MPT LSR(s) have re-converged. The PLR has to associate a timer with the forwarding state towards the MPT LSR(s) to stop forwarding. Determining a good timer value is challenging since it depends on many variables which could change over time.

After a PLR decides to stop forwarding towards a MPT LSR, another problem is releasing the label that PLR was using. The PLR has no mechanism to send a label release to the MPT LSR such that it can release the label and return it to the free pool. This more or less breaks the LDP design.

The solution described in this document does not ‘tunnel’ the MPT LSR(s) information but explicitly signals it from the MPT LSR(s) to the PLR LSR(s) via a Targeted LDP (T-LDP) session [RFC5036]. By using a T-LDP session to signal between the MPT LSR(s) and the PLR LSR(s), we don’t suffer from the above problems faced by [I-D.zhao-mpls-mldp-protections]. By having a T-LDP session with the PLR, most of the (m)LDP features currently defined should just work, like Make-Before-Break (MBB), Graceful Restart (GR), Typed Wildcard FEC support, etc. All this is achieved at the expense of having an additional T-LDP session between an MPT and PLR LSR.
1.1. Conventions used in this document

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

The terms "node" is used to refer to an LSR and used interchangeably. The terms "PLR" and "MPT" are used as shorthand to refer to "PLR LSR" and "MPT LSR" respectively.

1.2. Terminology

mLDP: Multipoint extensions to LDP.

PLR: Point of Local Repair (the LSR that redirects the traffic to one or more Merge Point LSRs).

MPT: Merge Point (the LSR that merges the backup LSP with primary LSP. Note, there can be multiple MPT LSRs for a single MP-LSP node protection).

T-LDP: Targeted LDP session.

MP LSP: Multi-Point LSP (either a P2MP or MP2MP LSP).

2. PLR Determination

In order for a MPT to establish a T-LDP session with the PLR, it first has to learn the PLR for a particular MP LSP. It is the responsibility of the protected node N to advertise the PLR address to the MPT. The PLR address for a MP LSP on node N is the address of the upstream LDP peer, but only when node N is NOT the root node of the MP2MP LSP. If node N is the root node, the procedures are slightly different as described in Section 2.2. The procedures that follow assume that all the participating nodes (N, PLRs, MPTs) are enabled (e.g. by a user configuration) to support and implement this feature.

2.1. Transit node procedure

Below we are describing the procedures when the protected node is a transit node along the path to the root.
N: The node being protected,
...: Backup LSPs from LSR1 to the LSR2 and LSR3.

Node N uses the root address of the MP LSP to determine the upstream LSR for a given MP LSP following the procedures as documented in [RFC6388] section 2.4.1.1. The upstream LSR in figure 1 is LSR1 because it is the first hop along the shortest path to reach the root address. After determining the upstream LSR, node N (which is feature enabled), MUST advertise the address of LSR1 as the PLR address to the downstream members of the MP LSP (i.e. LSR2 and LSR3) if the given downstream member has announced support for node protection (see Section 6) for Capability negotiation). For the format and encoding of PLR address information, see Section 2.3.

2.2. MP2MP root node procedure

In this section we are describing the procedures for when the protected node is the root of a MP2MP LSP. Consider figure 2 below;

N: The MP2MP root node being protected.
...: Backup LSPs between LSR1, LSR2 and LSR3.

Assume that LSR1, LSR2 and LSR3 are all members of a MP2MP LSP for
which N is the root node. Since N is the root of the MP2MP LSP, there is no upstream LSR and no ‘single’ PLR LSR for protecting node N’s. In order to protect node N, all the members of the MP2MP must participate in protecting node N by acting both as PLR and MPT LSR. An LSR will act as MPT for traffic coming from the other LSR(s) and it will act as PLR for traffic it is sending to the other LSR(s). Since node N knows the members of the MP2MP LSP, it will advertise the member list to its directly connected members, excluding the member it is sending to. For example, node N will advertise (LSR3, LSR1) list to LSR2 excluding LSR2 from it. Instead of advertising a single PLR when node N is not the root, a list of PLRs is advertised using the procedures documented in Section 2.3.

It should be noted that the MP2MP root node protection mechanism don’t replace the Root Node Redundancy (RNR) procedures as described in [RFC6388] section 7. The node protection procedures in this draft will help restoring traffic for the existing MP2MP LSPs after node failure, but a new root node has to be elected eventually in order to allow new MP2MP LSPs to be created.

2.3. PLR information encoding

The upstream LSR address is conveyed via an LDP Notification message with MP Status, where the MP status contains a new "PLR Status Value Element" that specifies the address of the PLR.

The new "PLR Status Value Element" is encoded as follows;

PLR Status Element:

```
0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|   Type        |           Length              |    Address    |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|   Family      | Num PLR entry |                               |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+                               +
|                                                               |
|                         PLR entry (0 or more)                 |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
```

Where

Type: PLR (Type=3 to be assigned by IANA)

Length: The Length field encodes the length of the Status Value following the Length field. The encoded Length varies based on
the Address Family and the number of PLR entries.

Address Family: Two octet quantity containing a value from IANA’s "Address Family Numbers" registry that encodes the address family for the PLR Address encoded in the PLR entry.

Num PLR entry: Number of "PLR entries" encoded in the Status Value Element, followed by "Num PLR entry" field (please see format of a PLR entry below).

The format of a "PLR Entry" is as follows:

```
0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|A|        Reserved             |       PLR address             ~
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
~                                                               ~
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
```

Where

A bit: 0 = Withdraw, 1 = Add.

Reserved: 15 bits, must be zero on transmit and ignored on receipt

PLR address: PLR Address encoded according to Address Family field encoded in the PLR Status Value Element.

The size of a "PLR Entry" is the 2 octets ("A bit + Reserved") + PLR address length. The length of the PLR address is depending on the Address Family as encoded in the PLR Status Value Element. The size of a "PLR entry" is 6 octets and 18 octets respectively for an IPv4 PLR address and an IPv6 PLR address.

If the PLR address on N changes for a give MP LSP, N needs to trigger a new PLR Status to update the MPT(s). A node N can advertise or withdraw a given PLR from its PLR set by setting "A bit" to 1 or 0 respectively in corresponding PLR entry. Removing a PLR address is likely due to a link failure, see the procedures as documented in Section 5. To remove all PLR addresses belonging to the encoded Address Family, an LSR N MUST encode PLR Status Value Element with no PLR entry and "Num PLR entry" field MUST be set to zero.

Along with the PLR MP Status a MP FEC TLV MUST be included in the LDP Notification message so that a receiver is able to associate the PLR Status with the MP LSP.
3. Using the T-LDP session

The receipt of a PLR MP Status (with PLR addresses) for a MP LSP on a receiving LSR makes it an MPT for node protection. If not already established, the MPT LSR MUST establish a T-LDP session with all of the learned PLR addresses using the procedures as documented in [I-D.napierala-mpls-targeted-mldp].

Using Figure 1 as the reference topology, let us assume that both LSR2 and LSR3 are MPTs and have established a T-LDP session with the PLR being LSR1. Assume that both LSR2 and LSR3 have a FEC <R,X> with an upstream LSR N and label Ln assigned to FEC towards N. The MPTs will create a secondary upstream LSR (using the received PLR address) and assign a Label Lpx to FEC <R,X> towards PLR for it. The MPTs will do that for each PLR address that was learned for the MP LSP. In this example, the MPTs will have a FEC <R,X> with two local labels associated with it. Ln that was assigned to N via the normal mLDP procedures, and Label Lpx that was assigned for PLR (LSR1) for the purpose of node protecting MP LSP via node N. Note, when the protected node is a MP2MP root node, there will be an upstream LSR for each PLR address that was advertised along with a unique Label Lpx.

It is not preferable that a PLR is always sending traffic to an MPT over the backup P2P LSP. The PLR should only send traffic over the backup P2P LSP if node N fails. The receipt of a FEC Label Mapping alone over the T-LDP session from MPT on a PLR conveys the label information but does not convey the node being protected. The information about a protected node is known to the MPT LSR and needs to be communicated to the PLR as well. For this reason, the FEC Label Mapping (FEC <R,X> : Lpx) sent by the MPT over the T-LDP session to the PLR MUST include a Status TLV with MP Status including a new LDP MP status Value Element called the "Protected Node Status Value Element". This new value element is used to specify the address of the node being protected. The "Protected Node Status Value Element" has the following format:

```
0                   1                   2                   3
0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|    Type = 4   |           Length              | Address Family
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|        Node address                           ˜
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
˜               |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
```

Type: Protected Node (Type = 4 to be assigned by IANA)

Length: The Length field encodes the length of the Status Value following the Length field. The encoded Length varies based on the Address Family and is 4 octets and 16 octets respectively for an IPv4 address and an IPv6 address.

Address Family: Two octet quantity containing a value from IANA’s "Address Family Numbers" registry that encodes the address family for the Node Address.

Node address: Protected node address encoded according to Address Family field.

When a PLR receives a Label Mapping for FEC <R,X> that includes a Protected Node Status, it will only use that label binding once the Node advertised in the Status value becomes unreachable. If the LSP is a MP2MP LSP, the PLR would have assign a Label Mapping for the upstream MP2MP FEC Element to the MPT ([RFC6388] section 3) for FEC <R,X>. This label binding on the MPT MUST only be used once node N becomes unreachable.

The procedures to determine if a node is unreachable is a local decision and not spelled out in this draft. Typical link failure or Bidirectional Forwarding Detection (BFD) can be used to determine and detect node unreachability.

4. Link or node failure

Consider the following topology;

```
root
  ^
  . (LSR1)
  . / .
  . (M) \ .
  . /  \ .
  . (N)  .
  /   \
(LSR2) (LSR3)
```

Figure 3.

N: The node being protected
M: The backup node to protect link LSR1 - N
...; Backup LSPs from LSR1 to LSR2 and LSR3.

Assume that LSR1 is the PLR for protected node N, LSR2 and LSR3 are MPTs for node N. When LSR1 discovered that node N is unreachable, it can’t determine whether it is the ‘LSR1 - N’ link or node N that failed. In Figure 3, the link between LSR1 and N is also protected using Fast ReRoute (FRR) [RFC4090] link protection via node M. LSR1 MAY potentially invoke 2 protection mechanisms at the same time, redirection the traffic due to link protection via node M to N, and for node protection directly to LSR1 and LSR2. If only the link failed, LSR2 and LSR3 will receive duplicate packets due to the two protection mechanisms. To prevent duplicate packets to be forwarded to LSR2 and LSR3, either the primary upstream LSRs or the secondary upstream LSRs should be forwarding MPLS packets, but never both at the same time. The selection between the primary upstream LSR or (one or more) secondary upstream LSRs is based on the reachability of N. As long as N is reachable, N is the primary upstream LSR by which the MPLS packets are forwarded. Once N becomes unreachable, the secondary upstream LSRs that were installed for node protection are activated. Note that detecting if N is unreachable is a local decision and not spelled out in this draft. Typical link failure or Bidirectional Forwarding Detection (BFD) can be used to determine and detect node unreachability.

5. Re-convergence after node/link failure

Consider the following topology;

```
    root
     /|_
    /  
   (-LSR1)
  /  
 /   |
(M)  
 / 
(P) \|
 \ /   
| /   
\ /   
 N  
 /   
/   
(LSR2) (LSR3)

Figure 4.
```

N: The node being protected.
M: The backup node to protect link ‘LSR1 - N’.
P and Q: The nodes on the new primary path after N failure.
Assume that LSR1 has detected that Node N is unreachable and invoked both the Link Protection and Node Protection procedures as described in this draft. LSR1 is acting as PLR and sending traffic over both the backup P2P LSP to node N (via M) and the P2P LSPs directly to LSR2 and LSR3, acting as MPT LSRs. The procedures following are depending on whether the link ‘LSR1 - N’ has failed or node N itself.

5.1. Node failure

If node N failed, both LSR2 and LSR3 will have changed the primary upstream LSR to the secondary upstream LSR (LSR1) due to node N being unreachable. With that, the label bindings previously assigned to LSR1 will be activated on the MPTs (LSR2 and LSR3) and the label binding to N will be disabled. Traffic is now switched over the label bindings that where installed for node protection.

5.2. Link failure

If the link ‘LSR1 - N’ has failed, both LSR2 and LSR3 will not change the primary upstream LSR because node N is still reachable. LSR2 and LSR3 will receive traffic over two different bindings, the primary label binding assigned to node N (due to link protection via node M) as well as over the binding assigned to LSR1 for the node protection. Since the secondary upstream LSRs have not been activated, the traffic received due to node protection will be dropped. Node N will re-converge and update LSR2 and LSR3 (Section 2.3) with the information that the PLR address (LSR1) is no longer applicable and must be removed. In repose, LSR2 and LSR3 MUST sent a Label Withdraw to LSR1 to withdraw the label binding. This will stop the traffic being forwarded over the backup P2P LSPs for node protection. LSR1 will respond back with a Label Release as soon as the binding has been removed.

5.3. Switching to new primary path

The network will eventually re-converge and a new best path to the root will be found by LSR2 and LSR3. LSR1 will find that M is its new primary upstream LSR to reach the Root and LSR3 will find Q. Note that although the current active upstream LSR can either be node N or LSR1 (depending on link or node failure), it does not matter for the following procedures. Both LSR2 and LSR3 SHOULD use the Make-Before-Break (MBB) procedures as described in [RFC6388] section 8 to switch to the new primary upstream node. As soon as the new primary upstream LSRs M and Q are activated, a Label Withdraw message MUST be sent to the old upstream LSR. Note that an upstream LSR switchover from a T-LDP neighbor to a directly connected LDP neighbor is no
different compared to switching between two directly connected neighbors. After the Label Withdraw message has been received by LSR1 or node N, forwarding will stop and a Label Release will be sent.

When it is determined that after re-convergence there is no more interest in the T-LDP session between the MPT and the PLR, the T-LDP session MAY be taken down. It is possible that having no more interest in the T-LDP session is temporarily due to link flapping. In order to avoid the T-LDP session from flapping, it is RECOMMENDED to apply a delay before tearing down the session. Determining the delay is a local implementation matter.

6. mLDP Capabilities for Node Protection

In order to describe the capabilities of the participating LSRs, we are organizing it per role in the network i.e., Point of Local Repair (PLR), Merge Point (MPT), and Protected Node (as depicted in Fig 1).

6.1. PLR capability

A PLR node should handle the following conditions;

1. Accept an incoming T-LDP session from the MPT LSR.

2. Support the receipt of a "Protected Node Status Value Element" status in a MP Status TLV over T-LDP session.

3. Upon node failure detection, capable of switching traffic towards one or more MPT(s) over P2P LSP (bypassing N) using the labels previously advertised for MP LSPs over the T-LDP session.

An LSR capable of performing these actions will advertise itself as PLR capable in the Node Protection capability (see Section 6.4). This is a unidirectional capability announced from PLR to the protected LSR.

6.2. MPT capability

An MPT node should handle the following conditions;

1. Support the receipt of "PLR Status Value Element" in a MP Status TLV from a protected node N.

2. Support to transmit "Protected Node Status Value Element" in a MP Status TLV to a PLR.
A LSR capable of performing these actions will advertise itself as the MPT capable in the Node Protection capability (see Section 6.4). This is a unidirectional capability from MPT to the protected LSR.

6.3. The Protected LSR

A protected node should handle the following conditions;

1. Determine the PLR and MPT capability for directly connected upstream and downstream LSRs for a given MP FEC.

2. Support transmitting of "PLR Status Value Element" in a MP Status TLV to one or more downstream MPT LSRs.

The protected LSR does not advertise any capability for mLDP Node Protection because it does not need to receive any of the defined MP Status values as described above. However, the protected node does play an important role in the signaling and setup of the node protection. For a given FEC, the protected node can only send PLR information to a downstream LSR if the PLR has signaled the PLR capability and the downstream LSR has signaled the MPT capability. When the downstream LSR (acting as MPT) receives the PLR status, it can implicitly infer that the advertised LSR(s) are PLR capable. The MPT LSR can now proceed with setting up a T-LDP session with the PLR(s) and MP LSP node protection signaling.

6.4. The Node Protection Capability

We define a single capability "MP Node Protection Capability" to announce the PLR and MPT capability.

The format of the capability parameter TLV is as follows:

```
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
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|U|F| MP Node Prot Cap. (IANA)  |           Length = 2          |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|S| Reserved    |P|M| Reserved  |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
```

Where

U/F bits: MUST be set to 1 and 0 respectively (as per [RF5561])

MP Node Protection Capability: TLV type (value to be assigned by IANA)
Length: MUST be set to 2.

S bit: Set to 1 to announce and 0 to withdraw the capability (as per [RFC5561])

P bit: PLR capable for MP LSP node protection

M bit: MPT capable for MP LSP node protection

Reserved: Must be zero on transmit and ignored on receipt

The above capability can be sent in an LDP Initialization message to announce capability at the session establishment time, or it can be sent in LDP Capability message to dynamically update (announce or withdraw) its capability towards its peer using procedures specified in [RFC5561].

An LSR that supports the PLR functionality LSR MAY send this capability to its downstream MP peers with "P" bit set; whereas, an LSR that supports an the MPT functionality MAY send this capability to its upstream peer with "M" bit set. Moreover, an LSR that supports both the PLR and MPT functionality MAY sent this capability to its peers with both "P" and "M" bit set.

7. Security Considerations

The same security considerations apply as those for the base mLDP specification, as described in [RFC6388].

8. IANA considerations

IANA is requested to allocate two new code points from the "LDP MP Status Value Element type" registry;

- PLR Status Value Element - 3
- Protected Node Status Value Element - 4

IANA is requested to assign one new code points for a new Capability Parameter TLVs from the LDP registry "TLV Type Name Space", corresponding to the advertisement of the new MP Status values. The values is:

- MP Node Protection Capability - TBD
9. Acknowledgments

The authors like to thank Nagendra Kumar for his input on this draft.

10. References

10.1. Normative References


10.2. Informative References


Authors’ Addresses

IJsbrand Wijnands (editor)
Cisco Systems, Inc.
De kleetlaan 6a
Diegem 1831
Belgium

Email: ice@cisco.com

Eric Rosen
Cisco Systems, Inc.
1414 Massachusetts Avenue
Boxborough MA 01719
USA

Email: erosen@cisco.com

Kamran Raza
Cisco Systems, Inc.
2000 Innovation Drive
Ottawa Ontario K2K-3E8
Canada

Email: skraza@cisco.com

Jeff Tantsura
Ericsson
300 Holger Way
San Jose CA 95134
USA

Email: jeff.tantsura@ericsson.com

Alia Atlas
Juniper Networks
10 Technology Park Drive
Westford MA 01886
USA

Email: akatlas@juniper.net
Protection Mechanisms for Label Distribution Protocol P2MP/MP2MP Label Switched Paths
draft-zhao-mpls-mldp-protections-02.txt

Abstract

Service providers continue to deploy real-time multicast applications using Multicast LDP (mLDP) across MPLS networks. There is a clear need to protect these real-time applications and to minimize switching times in the event of failure.

This document outlines the requirements, procedures and extensions to facilitate mLDP Point-to-Multipoint (P2MP) and Multipoint-to-Multipoint (MP-to-MP) LSP protection within an MPLS network.

Status of this Memo

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

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

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

This Internet-Draft will expire on September 14, 2012.

Copyright Notice

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

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

Table of Contents

1. Terminology ......................................................... 4
2. Requirement Language .............................................. 4
3. Introduction ........................................................ 4
   3.1. Requirements .................................................. 6
   3.2. Scope .......................................................... 6
4. mLDP Node Protection using P2P LSPs ............................... 6
   4.1. Signaling Procedures for P2P Based Node Protection .......... 8
       4.1.1. P2P Based Node Protection Procedure Example ........... 8
       4.1.2. PLR Switching Over Considerations ........................ 10
       4.1.3. Backup Path Cleanup Considerations ....................... 11
   4.2. Protocol Extensions for P2P Based Node Protection ........... 12
       4.2.1. P2P Based MP Protection Capability Parameter TLV ...... 12
       4.2.2. P2P Based MP Node Protection Status Element ........... 13
5. mLDP Node Protection using P2MP LSPs ............................. 14
   5.1. Signaling Procedures for P2MP Based Node Protection .......... 16
       5.1.1. P2MP Based Node Protection Procedure Example ........... 16
       5.1.2. PLR Switching Over Considerations ........................ 17
   5.2. Protocol Extensions for P2MP Based Node Protection .......... 18
       5.2.1. P2MP Based MP Protection Capability Parameter TLV ...... 18
       5.2.2. P2MP Based MP Node Protection Status Elements .......... 19
6. mLDP End-to-End Protection using LDP/mLDP Multiple Topology . 20
   6.1. Signaling Procedures for MT Based End-to-end Protection . 23
   6.2. Protocol extensions for MT Based End-to-end Protection . 24
7. IANA Considerations .................................................. 24
8. Manageability Considerations ....................................... 24
   8.1. Control of Function and Policy ................................ 24
   8.2. Information and Data Models ................................... 24
   8.3. Liveness Detection and Monitoring .............................. 24
   8.4. Verifying Correct Operation .................................. 24
   8.5. Requirements on Other Protocols and Functional Component .... 24
   8.6. Impact on Network Operation ................................... 24
   8.7. Policy Control ................................................ 24
9. Security Considerations .............................................. 24
10. Acknowledgements .................................................... 25
11. References .......................................................... 25
   11.1. Normative References ......................................... 25
   11.2. Informative References ....................................... 26
Authors’ Addresses ...................................................... 26
1. Terminology

This document uses terminology discussed in [RFC5036] and [MT-LDP]. Additionally the following section provides further explanation for key terms and terminology:

- **PLR**: The node where the traffic is logically redirected onto the preset backup path is called Point of Local Repair (PLR).
- **N**: The node being protected.
- **Pn**: The node(s) on the backup path for protecting node N.
- **MP**: The node where the backup path merges with the primary path is called Merge Point (MP).
- **FD**: The node that detects the failure on primary path, and then triggers the necessary action for traffic protection is called Failure Detector (FD). Either traffic sender or receiver can be the FD, depending on which protection mode has been deployed. Further specification is provided in later sections.
- **SP**: The node where the traffic is physically switched/duplicated onto the backup path is called Switchover Point (SP). In multicast scenarios, PLR and SP can be two different nodes. Further specification is provided in later sections.
- **T-LDP**: Targeted LDP session.

2. Requirement Language

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

3. Introduction

In order to meet user demands, operators and service providers continue to deploy multicast applications using Multicast LDP (mLDP) across MPLS networks. In certain key scenarios, conventional IGP-mLDP convergence mechanisms fail to meet protection switching times required to minimise, or negate entirely, application interruptions for real-time applications, including stock trading, on-line games, and multimedia teleconferencing.

Current best practice for protecting services, and 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 existing 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:

- The use of an RSVP-TE P2P tunnel as a logical out-going interface, consequently utilize the mature high availability technologies of RSVP-TE.
- The use of an LDP P2P LSP as a packet encapsulation, so that the complex configuration of P2P RSVP-TE can be skipped.
- Creating a P2MP/MP2MP backup LSP according to IGP’s loop-free alternative route. Comparing to using P2P LSPs, this solution can prevent unnecessary packet duplication on common links.
- 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:

- 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.

- 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 minimise high-layer application impact.

Ideally if switching time performance for Option 2 can be closer to the Option 1, it is reasonable to choose it to avoid bandwidth wastage. The recommendations of this document are based on this point of view.
This document provides several ways to setup the backup path for mLDP LSP, including P2P based mLDP node protection, P2MP based mLDP node protection, and MT based end-to-end protection. The goal is to build a reliable umbrella to against traffic black hole.

Note that the backup path computation is out of the scope of this draft, the algorithm can be either LFA or any other algorithms available including the 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:

- Computation of a disjointed (link and node) backup path within the multicast tree;
- Minimization of protection convergence time;
- Minimization of operation and maintenance cost;
- Optimization of bandwidth usage;
- Minimization the impact on the existing network deployment.

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 P2P LSPs

By encapsulating mLDP packets within an P2P TE tunnel or P2P LDP backup LSP, the LDP P2MP/MP2MP LSP can be protected by the P2P protection mechanisms. However, this protection mechanism is not capable of recovering the failure on the destination node of the P2P backup LSP. Thus, this section provides an extra method to protect node using an P2P LSP.
In Figure 1 (P2P Based mLDP Node Protection Example) above, the preferential path from R1 to R4/R5 is through R2, and the secondary path is through R3. In this case, the mLDP LSP <X,Y>, will be established according to the IGP preferential path as R1--R2--R4/R5. R4’s backup P2P path is R1--R3--R5--R4. R5’s backup P2P path is R1--R3--R5. (We assume that all the nodes in Figure 1 support the P2P Based mLDP Node Protection capability.)

The procedure for P2P Based mLDP Node Protection is as follows:

The MP(s) (in this example R4 or R5) sends label mapping message with protection information to the the protected node N(in this example R2). Node N notifies all the downstream information to the node PLR(in this example R1). PLR selects P2P LSP toward each MP(s) as backup path encapsulation. When PLR detects N failed, PLR switches traffic to MP(s) over corresponding backup path(s). PLR will stop the backup traffic forwarding after relative MP(s) finish convergence procedure.

There are two methods for PLR to switch over the traffic:
Option 1:
If PLR can differentiate the link and node failure, such as binding two BFD sessions on link and node, then it can feed the traffic to the node protection path only when the node failure is detected. Only N is required to maintain BFD session with the PLR, it does not need to maintain the large number of BFD sessions with MPs.

Option 2:
If the PLR can not differentiate between link and node failure, it should feed the traffic to both link protection path and node protection path at the same time, and MP must take the responsibility to drop one stream. In this case, N must maintain BFD sessions with PLR and all the MPs.

[Editors Note - The authors recommend the first method to save BFD resource usage, the details are specified in section 4.1.2. This Editors note and remaining options will be adjusted once we get more feedbacks from users.]

Additionally there are two methods for the backup label cleanup between the MP and PLR:

Option 1:
The PLR can delete the backup label by receiving a withdraw notification message from MP through a T-LDP between PLR and MP. This method needs PLR maintains a huge number of T-LDP sessions with MPs.

Option 2:
The second method requires cooperation between PLR and MP by a synchronous timer. The MP will synchronize the timers to the PLR and hold the backup label resource until its local timer expires. In addition the PLR will delete the backup label when its local timer expires.

[Editors Note - The authors recommend the second method in order to save T-LDP resource usage, the details are specified in section 4.1.3. This Editors note and remaining options will be adjusted once we get more feedbacks from users.]

4.1. Signaling Procedures for P2P Based Node Protection

4.1.1. P2P Based Node Protection Procedure Example

[Editors Note - This section introduces the procedures for P2P Based Node Protection based on the two options recommended above.]
The following in this section demonstrates the signaling and procedures for P2P Based Node Protection. Note that STEP5 and STEP6 should be acted at the same time.

STEP1  MP’s procedures for setting up backup path:
   Take R4 for example, which acts as a node MP. R4 determines N (in this example R2) as its upstream LSR, and then sends label mapping message to R2 including label L4 for <X,Y>. This label mapping message MUST include a new LDP MP Status Value Element, which includes a backup label and a reserve-time. The backup label L4’ is assigned by MP for PLR’s backup path. In order to avoid upgrading the MP’s hardware to support filtrating traffic, this backup label can be same with L4. This reserve-time is the value of the reserve-timer on MP, which can be configured, and it is recommend to be longer than the IGP convergence time and LDP MBB procedure time.

STEP2  N’s procedures for setting up backup path:
   When R2 receives the mapping message from MP, it determines R1 as its upstream LSR and send a label mapping message with label L2 to R1. Besides, R2 transfers all the new LDP MP Status Value Elements, received from its downstreams, to R1 by a notification message.

STEP3  PLR’s procedures for setting up backup path:
   When R1 receives this notification message and label mapping message from R2, R1 creates two forwarding paths: the primary forwarding state is <X,Y,L2> , and the secondary forwarding states is <X,Y,L4’>. Note that there might be more than one secondary forwarding states, such as <X,Y,L5’>.

STEP4  PLR’s procedures when node N fails:
   Once the node PLR(R1) detects the node N(R2) failure, before protocol converging, R1 will switch the traffic to MP(R4) over P2P LSP, with inner label as L4’, and outer label as the P2P backup LSP R1--R3-- R5--R4. Note that the PLR MUST NOT switch traffic to backup path until it detects N failure. R1 will also duplicate the traffic to other MP(s)(for example R5) over relative P2P LSPs. Meanwhile, the node PLR will create a reserve-timer with the reserve-time value in the notification message for each N’s MP respectively. Note that once the primary forwarding state is removed, the secondary forwarding state MUST be deleted too.

STEP5  MP’s procedures after convergence:
   MP will also create the reserve-timer when it detects the network convergence. MP(s) will remove the previous forwarding state after a new path to root is created or MBB
procedure finishes. Note that MP MUST hold the old backup label resource (L4') until its local reserve-timer expires.

STEP6 PLR’s procedures after convergence:
When R1’s local reserve-timers expire, it will stop the traffic on the backup paths and remove the secondary forwarding states.

Note that the mLDP Local Protection mechanism can be used in any part of the mLDP LSP other than the ingress and egress nodes. In other words, R1 can be either Ingress or Transit node, R4/R5 can be either Transit or Egress node.

4.1.2. PLR Switching Over Considerations

PLR switching over method depends on its failure detection mode. If PLR switches traffic when a link failure happens, MP may receive reduplicate traffic because the node N still feeds the traffic to MP. This is a problem if MP accepts these reduplicate traffic. There are two optional methods to solve this problem.

4.1.2.1. Single Feed Mode When Node Failure Detection is Supported

In this method, the node PLR MUST be capable of differentiating link and node failure, and MP will not receive reduplicate traffic. It does not need MP’s cooperation in such case, and the new label assigned to PLR has no need to be different with the one assigned to N.

MP also needs to send label mapping message with a new LDP MP Status Value Element to the node N. Especially, the MP Status TLV’s Node Failure Required Flag need to be set as ‘Y’. The node N transfers this MP Status TLV to the node PLR by notification message.

In such case, the node PLR MUST NOT switch the traffic to the backup path until it detects the node N failure.

4.1.2.2. Duel Feed Mode When Node Failure Detection is Not Supported

In this method, MP must support traffic filtering. Thus, PLR doesn’t need to differentiate link and node failure because it can cooperate with the node MP. The node MP may receive reduplicate traffic and MUST drop the backup traffic when the node N doesn’t fail. To support this manner, MP MUST assign PLR a new label Lp, which is different from the label Ln assigned for N.

The node MP send label mapping message with a new LDP MP Status Value Element to the node N. The label Ln is encoded in the label TLV and
the Backup Label Lp is encoded in the MP Status TLV. Besides, the MP Status TLV's Node Failure Required Flag need be set as 'N'. The node N transfers this MP Status TLV to the node PLR by notification message.

If the MP Status TLV's Node Failure Required Flag is 'N', before failure occurs, the label Ln is activated and Lp is deactivated on the node MP. The PLR will switch traffic to the backup path when it detects the failure, no matter whether it is link failure or node failure.

The node of MP needs to detect whether the node N is failure or not. This is why N needs to maintain BFD sessions with all the MP nodes. If N fails, MP should deactivate the primary forwarding state of label Ln, and active the secondary forwarding state of label Lp. So the MP will receive the traffic on the backup path, and drop the traffic on the primary path. Otherwise, if it is link failure, MP will keep the backup label Lp deactivate and the traffic PLR switching to the backup path is dropped by MP.

4.1.3. Backup Path Cleanup Considerations

In order to prevent traffic duplication and unnecessary traffic lost, PLR and MP should delete the label at the same time. There are two methods to ensure this.

4.1.3.1. Timer Synchronization Mode

In this mode, a reserve-timer is needed on both MP and PLR. The value of this timer can be configured on MP, and it is recommend to be longer than the IGP convergence time and mLDP MBB procedure time.

Once the reserve time is configured, the new LDP MP Status Value Element in label mapping message MUST include this time value and the Delete Flag MUST be set as 'implicit-delete', mentioned in Section 4.4.

The Node N transfers this time value and flag to the node PLR by the notification message. When the node PLR receives this notification message, PLR checks the Delete Flag and uses this time value as PLR's reserve timer if Delete Flag equals 'implicit-delete'.

PLR will create the reserve-timer when it detects the failure and switch traffic to the backup path. PLR will remove the secondary forwarding state when this reserve-timer expire.

MP will create the reserve-timer when it detects the network convergence. MP will remove the old forwarding state after a new
4.1.3.2. T-LDP Mode

In this mode, an automatic setting up T-LDP is needed between MP and PLR. It requires the node MP and PLR MUST be capable of setting up T-LDP as documented in [RFC5036].

If MP need a T-LDP, the new LDP MP Status Value Element TLV in label mapping message MUST set its Delete Flag as ‘explicit-delete’.

The Node N transfers this flag to the node PLR by the notification message. When the node PLR receives this notification message, PLR checks the Delete Flag and trigger a T-LDP toward to MP if flag equals ‘explicit-delete’.

MP will remove the old forwarding state after a new path to root is created or MBB procedure is end. Meanwhile, MP MUST send a notification message to PLR through T-LDP. This message include a new LDP MP Status Value Element TLV, whose status code is ‘withdraw’ defined.

PLR will not remove the secondary forwarding state until it receive the notification message with ‘withdraw’ status code.

Noted that if the primary forwarding state is removed, the secondary forwarding state need be deleted no matter which delete method is used.

4.2. Protocol Extensions for P2P Based Node Protection

4.2.1. P2P Based MP Protection Capability Parameter TLV

A new Capability Parameter TLV is defined as P2P Based MP Protection Capability. Following is the format of this new Capability Parameter TLV:

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
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|1|0| P2P Based MP Prot. (IANA) |          Length (= 1)         |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|S|   Reserved  |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|                  |                  |                  |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
S: As specified in [RFC5561]
This is an unidirectional capability announced.

An LSR, which supports the P2P based protection procedures, should advertise this P2P Based MP Protection Capability TLV to its LDP speakers.

An LSR can consider that has no P2P Based MP Protection Capability if it does not announce its capability. An LSR MUST NOT send any message including the new LDP MP Status Value Element TLV to its peer, which does not have the P2P Based MP Protection Capability.

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.

4.2.2. P2P Based MP Node Protection Status Element

A new type of LDP MP Status Value Element is introduced, for notifying downstream LSR information, including respective labels and other parameters. It is encoded as follows:

```
+------------------  +------------------  +------------------  +------------------+
| mLDP P2P Type=2  |      Length     |    Status Code   |
|  +------------------  +------------------  +------------------|
| Downstream Element 1 |  +------------------  +------------------|
|  |  +------------------  +------------------|
|  | Downstream Element N |  +------------------  +------------------|
|  +------------------  +------------------  +------------------+
```

Status Code: 1 = Advertise the existing downstream LSRs
2 = Withdraw the deleted downstream LSRs

The Downstream Element is encoded as follows:
Backup Label: The label assigned by MP for PLR
D Bit: Delete Flag, The type of deleting backup label:
1 = 'explicit-delete', delete by MP's notification message through T-LDP
0 = 'implicit-delete', delete by reserve-timer expire
N Bit: Node Failure Required Flag, the occasion of switching traffic's on PLR
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
Downstream Node Address: Downstream node's LSR-ID address
Res-time: The time of MP's reserve-timer, synchronizing to PLR.
It is effective when D bit set as 'implicit-delete' and MUST be ignored when D bit set as 'explicit-delete'.

5. mLDP Node Protection using P2MP LSPs

By using IGP-FRR, LDP can build the backup mLDP LSP among PLR, the protected node, 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 protect the network from traffic congestion risk.
In Figure 2 (P2MP Based mLDP Local Protection Example), the preferential path from R1 to R4/R5 is through R2, and the secondary path is through R3. In this case, the mLDP LSP will be established according to the IGP preferential path as R1--R2--R4/R5. This section will take the Protected Node as R2 for example, actually the Protected Node can be any Transit node of the mLDP LSP. (We assume that all the nodes in Figure 2 support this P2MP based node protection method, including Pn.)

The procedure for P2P Based mLDP Node Protection is as follows:

As the Protected Node, R2 will announce its selected upstream node R1 to all its downstream nodes, which are R4 and R5 in this example, when it receives these label mapping messages.

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.

The primary LSP and backup LSP are differentiated by the signaling procedure, so normally PLR can only feed traffic on the primary path. When R2 node fails, R1 will 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. If R2 is protected by mLDP LSP instead, R3 will only receive one stream, and the packet duplication will be done on R3.

5.1. Signaling Procedures for P2MP Based Node Protection

5.1.1. P2MP Based Node Protection Procedure Example

[Editors Note - This section introduces the procedures for P2MP Based Node Protection based on the PLR being capable for node detection.]

We assume all the involved nodes have advertised their corresponding capabilities. And the following in this section demonstrates the signaling and procedures for P2MP Based Node Protection.

STEP1 N’s procedures for setting up backup path:
MP determines N as its upstream and sends label mapping for to N, following the procedures as documented in [RFC6388] without any extention. When the Protected Node (R2) receives these label mapping messages and determines its upstream LSR (R1), it will notify to all its downstream nodes immediately. If there are other LSR(s) becoming its downstream node(s) later, it will do the announcement for the new downstream node(s).

STEP2 MP’s procedures for setting up backup path:
When the Merge Point (R4/R5) receive the notification, they individually determine the primary and secondary paths toward R1 according to the IGP-FRR results. Then they will send out label mapping messages including an LDP MP Status TLV that carries a FRR Status Code to indicate the primary path and secondary path. The backup path is uniquely identified by root address, opaque value, PLR Node address, and Protected Node address. 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.
STEP3  Pn’s procedures for setting up backup path:
When the transit nodes of the secondary LSP receive the FRR label mapping message, they can easily consider it as a new mLDP LSP establishment, and follow the existing protocol procedures. The modification for these nodes is dealing with the FRR FEC, which is identified by root address, opaque value, PLR address, and Protected Node address. To avoid the backup LSP going through the Protected Node, additional path selection rule(s) can be applied. A simple method is that the transit nodes can not choose the specified Protected Node as its upstream LSR on the secondary LSP. Other methods, such as not-via policy, are under study, and will be added in the future.

STEP4  PLR’s procedures for setting up backup path:
When the Point of Local Repair (R1) receives the FRR label mapping message, it will generate the backup forwarding entry for the specific LSP, which is identified by the root address and opaque value in the message, 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.

STEP5  PLR’s procedures when node N fails:
When failure is detected by PLR, it will switch the traffic to the secondary path. MP will also locally merge the traffic back to the primary LSP. The switchover manner on PLR is specified in the later section.

STEP6  Procedures after network re-converges:
When Merge Point(s) see the next hop to Root changed, it/they will advertise the new mapping, 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 as the process of normally P2MP LSP. And the entire backup P2MP LSP will be deleted when all the node MP leave the backup P2MP LSP.

5.1.2. 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.

- Option 1:
  If PLR can not differentiate link and node failure, MP must take the responsibility to drop one of the two reduplicate traffic 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.]

5.2. Protocol Extensions for P2MP Based Node Protection

5.2.1. P2MP Based MP Protection Capability Parameter TLV

A new Capability Parameter TLV is defined as P2MP Based MP Protection Capability for node protection. Following is the format of this new Capability Parameter TLV:

```
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
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|1|0| P2MP Based MP Prot.(IANA) |          Length (= 2)         |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|S| Reserved    |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
```

S: As specified in [RFC5561]

This is an unidirectional capability announced.

An LSR, which supports the P2MP based protection procedures, should advertise this P2MP Based MP Protection Capability TLV to its LDP speakers. Without receiving this capability announcement, an LSR MUST NOT send any message including the P2MP Based MP Node Protection Status 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.

5.2.2. P2MP 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:

```
+-----------------+-----------------+-----------------+-----------------+
| mLDP FRR Type=3 |      Length      |   Reserved      |
+-----------------+-----------------+-----------------+
\ | PLR Node Address |
+-----------------+-----------------+-----------------+
```

- **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.

Besides, another new type of LDP MP Status Value Element is introduced, for setting up secondary mLDP LSP. It is encoded as follows:

```
+-----------------+-----------------+-----------------+-----------------+
| mLDP FRR Type=4 |      Length      |   Status code   |
+-----------------+-----------------+-----------------+
\ | PLR Node Address |
+-----------------+-----------------+-----------------+
\ | Protected Node Address |
+-----------------+-----------------+-----------------+
```

- **mLDP FRR Type**: Type 4 (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.
- **Protected Node Address**: The host address of the protected node.
mLDP FRR Type: Type 4 (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

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, which indicates the switchover timing on PLR.
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.

6. mLDP End-to-End Protection using LDP/mLDP Multiple Topology

[I-D.ietf-mpls-ldp-multi-topology] also provides the mechanism to setup disjointed LSPs within different topologies. So that applications can use these redundant LSPs for end-to-end protection.
In Figure 3 (mLDP End-to-end Protection Example), there are two separated topologies from Root node to Leaf 1 and Leaf 2. For the same root address and opaque value, the Leaf node can trigger mLDP LSPs in each topology. Root node can setup 1:1 or 1+1 end-to-end protection, using these two mLDP LSPs.
In Figure 4 (mLDP End-to-end Protection with Shared Upstream Node Example), there are two separated topologies from Root node to Leaf 1 and Leaf 2 except the link between R1 and Root node. For the same root address and opaque value, the Leaf node can trigger mLDP LSPs in each topology. Root node can setup 1:1 or 1+1 end-to-end protection, using these two mLDP LSPs. The difference in this example comparing to the last example where the primary and backup topology are totally disjoint, if there is a link failure between the Root and R1 or node R1 fails, there is no protection available.
6.1. Signaling Procedures for MT Based End-to-end Protection

Using the protocol extensions and signaling procedure provided by [I-D.ietf-mpls-ldp-multi-topology], Leaf 1 and Leaf 2 in figure 8 or figure 9 are able to trigger mLDP LSPs in different topologies, sending label mapping messages with same root address, same opaque value, different MT-ID and different label. Based on the two new Address Families named "MT IP" and "MT IPv6" introduced in [I-D.ietf-mpls-ldp-multi-topology] that can be used to specify IP prefixes within a topology scope, the mLDP FEC elements for leaf1 and leaf2 will be encoded as follows:

```
0          1          2          3
+--------------------------------------------------+
| mLDP FEC Type | Address Family (MT IP/MT IPv6) | Address Length |
+--------------------------------------------------+
| Root Node Address                                    |
+--------------------------------------------------+
| Reserved | MT-ID of Primary Topology       |
+--------------------------------------------------+
| Opaque Length | Opaque Value ... |
+--------------------------------------------------+
|                                                         |
+--------------------------------------------------+
```

```
0          1          2          3
+--------------------------------------------------+
| mLDP FEC Type | Address Family (MT IP/MT IPv6) | Address Length |
+--------------------------------------------------+
| Root Node Address                                    |
+--------------------------------------------------+
| Reserved | MT-ID of Secondary Topology     |
+--------------------------------------------------+
| Opaque Length | Opaque Value ... |
+--------------------------------------------------+
|                                                         |
+--------------------------------------------------+
```
When the Root node receives the label mapping messages from different topologies, it will set up two mLDP LSPs for application as end-to-end protection. Failure detection for the primary mLDP LSP is outside the scope of this document. Either Root node or Leaf node can be the Failure Detector.

6.2. Protocol extensions for MT Based End-to-end Protection

The protocol extensions required to build mLDP LSPs in different topologies are defined in [I-D.ietf-mpls-ldp-multi-topology].

7. IANA Considerations

This memo includes the following requests to IANA:

- P2P Based MP Protection Capability.
- P2MP Based MP Protection Capability.
- mLDP P2P Encapsulation type for LDP MP Status Value Element.
- mLDP FRR types for LDP MP Status Value Element.

8. Manageability Considerations

[Editors Note - This section requires further discussion]

8.1. Control of Function and Policy

8.2. Information and Data Models

8.3. Liveness Detection and Monitoring

8.4. Verifying Correct Operation

8.5. Requirements on Other Protocols and Functional Component

8.6. Impact on Network Operation

8.7. Policy Control

9. 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.

10. Acknowledgements

We would like to thank authors of draft-ietf-mpls-mp-ldp-reqs and the authors of draft-ietf-mpls-ldp-multi-topology from which some text of this document has been inspired. We also would like to thank Robin Li, Lujun Wan, Alia Atlas and IJsbrand Wijnands for their comments and suggestions to the draft.

11. References

11.1. Normative References


11.2. 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-03 (work in progress),
March 2012.

Authors’ Addresses

Quintin Zhao
Huawei Technology
125 Nagog Technology Park
Acton, MA 01719
US
Email: quintin.zhao@huawei.com

Emily Chen
Huawei Technology
2330 Central Expressway
Santa Clara, CA 95050
US
Email: emily.chenying@huawei.com

Tao Chou
Huawei Technology
156 Beiqing Rd
Beijing, P.R. 100095
China
Email: tao.chou@huawei.com

Daniel King
Old Dog Consulting
Email: E-mail: daniel@olddog.co.uk