Internet Engineering Task Force (IETF)
Request for Comments: 6425
Updates: 4379
Category: Standards Track
ISSN: 2070-1721

S. Saxena, Ed.
G. Swallow
Z. Ali
Cisco Systems, Inc.
A. Farrel
Juniper Networks
S. Yasukawa
NTT Corporation
T. Nadeau
CA Technologies

November 2011

Detecting Data-Plane Failures in Point-to-Multipoint MPLS - Extensions to LSP Ping

Abstract

Recent proposals have extended the scope of Multiprotocol Label Switching (MPLS) Label Switched Paths (LSPs) to encompass point-to-multipoint (P2MP) LSPs.

The requirement for a simple and efficient mechanism that can be used to detect data-plane failures in point-to-point (P2P) MPLS LSPs has been recognized and has led to the development of techniques for fault detection and isolation commonly referred to as "LSP ping".

The scope of this document is fault detection and isolation for P2MP MPLS LSPs. This documents does not replace any of the mechanisms of LSP ping, but clarifies their applicability to MPLS P2MP LSPs, and extends the techniques and mechanisms of LSP ping to the MPLS P2MP environment.

This document updates RFC 4379.

Status of This Memo

This is an Internet Standards Track document.

This document is a product of the Internet Engineering Task Force (IETF). It represents the consensus of the IETF community. It has received public review and has been approved for publication by the Internet Engineering Steering Group (IESG). Further information on Internet Standards is available in Section 2 of RFC 5741.

Information about the current status of this document, any errata, and how to provide feedback on it may be obtained at http://www.rfc-editor.org/info/rfc6425.

Table of Contents

1. Introduction

Simple and efficient mechanisms that can be used to detect data-plane failures in point-to-point (P2P) Multiprotocol Label Switching (MPLS) Label Switched Paths (LSP) are described in [RFC4379]. The techniques involve information carried in MPLS "echo request" and "echo reply" messages, and mechanisms for transporting them. The echo request and reply messages provide sufficient information to check correct operation of the data plane, as well as a mechanism to verify the data plane against the control plane, and thereby localize faults. The use of reliable channels for echo reply messages as described in [RFC4379] enables more robust fault isolation. This collection of mechanisms is commonly referred to as "LSP ping".

The requirements for point-to-multipoint (P2MP) MPLS traffic engineered (TE) LSPs are stated in [RFC4461]. [RFC4875] specifies a signaling solution for establishing P2MP MPLS TE LSPs.

The requirements for P2MP extensions to the Label Distribution Protocol (LDP) are stated in [RFC6348]. [RFC6388] specifies extensions to LDP for P2MP MPLS.

P2MP MPLS LSPs are at least as vulnerable to data-plane faults or to discrepancies between the control and data planes as their P2P counterparts. Therefore, mechanisms are needed to detect such data plane faults in P2MP MPLS LSPs as described in [RFC4687].

This document extends the techniques described in [RFC4379] such that they may be applied to P2MP MPLS LSPs. This document stresses the reuse of existing LSP ping mechanisms used for P2P LSPs, and applies them to P2MP MPLS LSPs in order to simplify implementation and network operation.

1.1. Design Considerations

An important consideration for designing LSP ping for P2MP MPLS LSPs is that every attempt is made to use or extend existing mechanisms rather than invent new mechanisms.

As for P2P LSPs, a critical requirement is that the echo request messages follow the same data path that normal MPLS packets traverse. However, as can be seen, this notion needs to be extended for P2MP MPLS LSPs, as in this case an MPLS packet is replicated so that it arrives at each egress (or leaf) of the P2MP tree.

MPLS echo requests are meant primarily to validate the data plane, and they can then be used to validate data-plane state against the control plane. They may also be used to bootstrap other Operations, Administration, and Maintenance (OAM) procedures such as [RFC5884]. As pointed out in [RFC4379], mechanisms to check the liveness, function, and consistency of the control plane are valuable, but such mechanisms are not a feature of LSP ping and are not covered in this document.

As is described in [RFC4379], to avoid potential denial-of-service attacks, it is RECOMMENDED to regulate the LSP ping traffic passed to the control plane. A rate limiter should be applied to the incoming LSP ping traffic.

1.2. Terminology

The terminology used in this document for P2MP MPLS can be found in [RFC4461]. The terminology for MPLS OAM can be found in [RFC4379]. In particular, the notation <RSC> refers to the Return Subcode as defined in Section 3.1. of [RFC4379].

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. Notes on Motivation

2.1. Basic Motivations for LSP Ping

The motivations listed in [RFC4379] are reproduced here for completeness.

> When an LSP fails to deliver user traffic, the failure cannot always be detected by the MPLS control plane. There is a need to provide a tool that enables users to detect such traffic "black holes" or misrouting within a reasonable period of time. A mechanism to isolate faults is also required.
>
> [RFC4379] describes a mechanism that accomplishes these goals. This mechanism is modeled after the ping/traceroute paradigm: ping (ICMP echo request [RFC792]) is used for connectivity checks, and traceroute is used for hop-by-hop fault localization as well as path tracing. [RFC4379] specifies a "ping mode" and a "traceroute" mode for testing MPLS LSPs.
>
> The basic idea as expressed in [RFC4379] is to test that the packets that belong to a particular Forwarding Equivalence Class (FEC) actually end their MPLS path on an LSR that is an egress for that FEC. [RFC4379] achieves this test by sending a packet (called an "MPLS echo request") along the same data path as other packets belonging to this FEC. An MPLS echo request also carries information about the FEC whose MPLS path is being verified. This echo request is forwarded just like any other packet belonging to that FEC. In "ping" mode (basic connectivity check), the packet should reach the end of the path, at which point it is sent to the control plane of the egress LSR, which then verifies that it is indeed an egress for the FEC. In "traceroute" mode (fault isolation), the packet is sent to the control plane of each transit LSR, which performs various checks that it is indeed a transit LSR for this path; this LSR also returns further information that helps to check the control plane against the data plane, i.e., that forwarding matches what the routing protocols determined as the path.
>
> One way these tools can be used is to periodically ping a FEC to ensure connectivity. If the ping fails, one can then initiate a traceroute to determine where the fault lies. One can also

periodically traceroute FECs to verify that forwarding matches the control plane; however, this places a greater burden on transit LSRs and should be used with caution.

2.2. Motivations for LSP Ping for P2MP LSPs

As stated in [RFC4687], MPLS has been extended to encompass P2MP LSPs. As with P2P MPLS LSPs, the requirement to detect, handle, and diagnose control- and data-plane defects is critical. For operators deploying services based on P2MP MPLS LSPs, the detection and specification of how to handle those defects is important because such defects may affect the fundamentals of an MPLS network, but also because they may impact service-level-specification commitments for customers of their network.

P2MP LDP [RFC6388] uses LDP to establish multicast LSPs. These LSPs distribute data from a single source to one or more destinations across the network according to the next hops indicated by the routing protocols. Each LSP is identified by an MPLS multicast FEC.

P2MP MPLS TE LSPs [RFC4875] may be viewed as MPLS tunnels with a single ingress and multiple egresses. The tunnels, built on P2MP LSPs, are explicitly routed through the network. There is no concept or applicability of a FEC in the context of a P2MP MPLS TE LSP.

MPLS packets inserted at the ingress of a P2MP LSP are delivered equally (barring faults) to all egresses. In consequence, the basic idea of LSP ping for P2MP MPLS LSPs may be expressed as an intention to test that packets that enter (at the ingress) a particular P2MP LSP actually end their MPLS path on the LSRs that are the (intended) egresses for that LSP. The idea may be extended to check selectively that such packets reach specific egresses.

The technique in this document makes this test by sending an LSP ping echo request message along the same data path as the MPLS packets. An echo request also carries the identification of the P2MP MPLS LSP (multicast LSP or P2MP TE LSP) that it is testing. The echo request is forwarded just as any other packet using that LSP, and so is replicated at branch points of the LSP and should be delivered to all egresses.

In "ping" mode (basic connectivity check), the echo request should reach the end of the path, at which point it is sent to the control plane of the egress LSRs, which verify that they are indeed an egress (leaf) of the P2MP LSP. An echo reply message is sent by an egress to the ingress to confirm the successful receipt (or announce the erroneous arrival) of the echo request.

In "traceroute" mode (fault isolation), the echo request is sent to the control plane at each transit LSR, and the control plane checks that it is indeed a transit LSR for this P2MP MPLS LSP. The transit LSR returns information about the outgoing paths. This information can be used by ingress LSRs to build topology or by downstream LSRs to do extra label verification.

P2MP MPLS LSPs may have many egresses, and it is not necessarily the intention of the initiator of the ping or traceroute operation to collect information about the connectivity or path to all egresses. Indeed, in the event of pinging all egresses of a large P2MP MPLS LSP, it might be expected that a large number of echo replies would arrive at the ingress independently but at approximately the same time. Under some circumstances this might cause congestion at or around the ingress LSR. The procedures described in this document provide two mechanisms to control echo replies.

The first procedure allows the responders to randomly delay (or jitter) their replies so that the chances of swamping the ingress are reduced. The second procedure allows the initiator to limit the scope of an LSP ping echo request (ping or traceroute mode) to one specific intended egress.

LSP ping can be used to periodically ping a P2MP MPLS LSP to ensure connectivity to any or all of the egresses. If the ping fails, the operator or an automated process can then initiate a traceroute to determine where the fault is located within the network. A traceroute may also be used periodically to verify that data-plane forwarding matches the control-plane state; however, this places an increased burden on transit LSRs and should be used infrequently and with caution.

3. Packet Format

The basic structure of the LSP ping packet remains the same as described in [RFC4379]. Some new TLVs and sub-TLVs are required to support the new functionality. They are described in the following sections.

3.1. Identifying the LSP Under Test

3.1.1. Identifying a P2MP MPLS TE LSP

[RFC4379] defines how an MPLS TE LSP under test may be identified in an echo request. A Target FEC Stack TLV is used to carry either an RSVP IPv4 Session or an RSVP IPv6 Session sub-TLV.

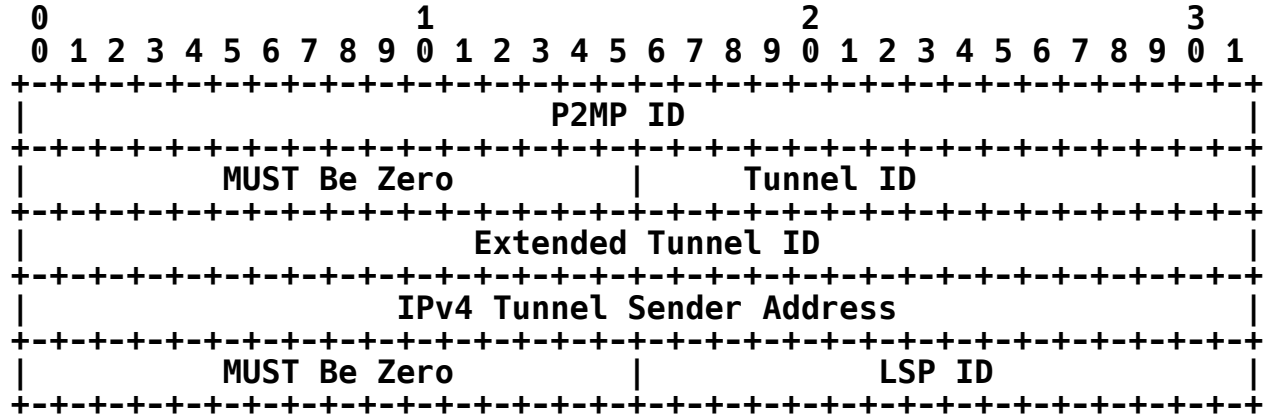
In order to identify the P2MP MPLS TE LSP under test, the echo request message MUST carry a Target FEC Stack TLV, and this MUST carry exactly one of two new sub-TLVs: either an RSVP P2MP IPv4 Session sub-TLV or an RSVP P2MP IPv6 Session sub-TLV. These sub-TLVs carry fields from the RSVP-TE P2MP SESSION and SENDER_TEMPLATE objects [RFC4875] and so provide sufficient information to uniquely identify the LSP.

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

Sub-Type #	Length	Value Field
17	20	RSVP P2MP IPv4 Session
18	56	RSVP P2MP IPv6 Session

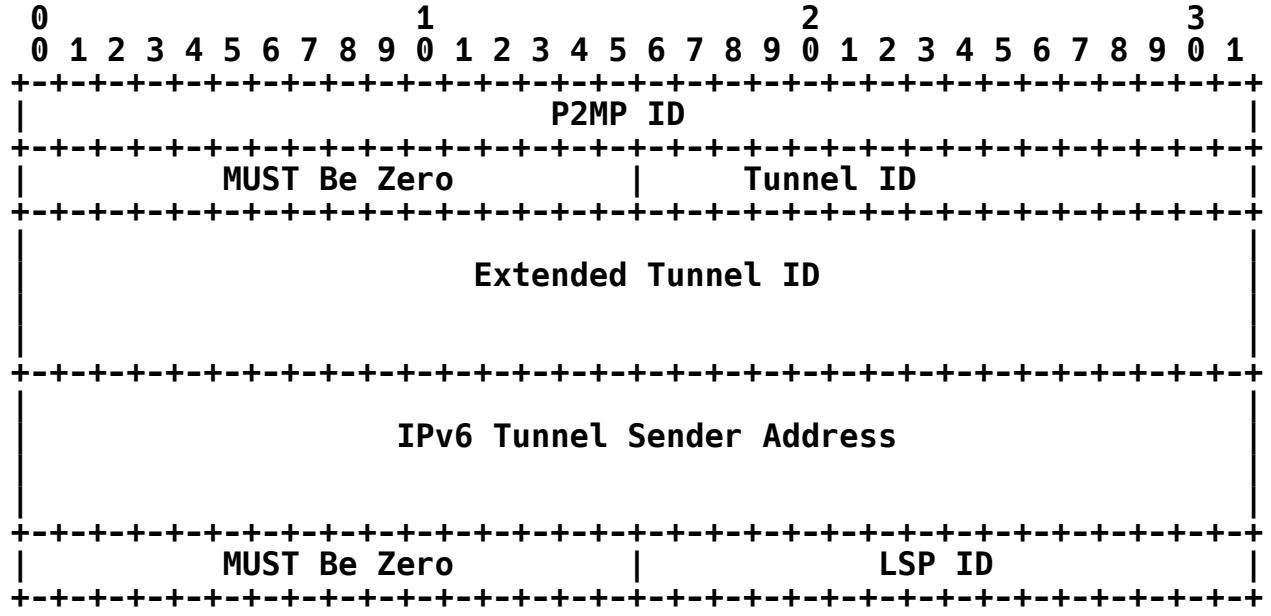
3.1.1.1. RSVP P2MP IPv4 Session Sub-TLV

The format of the RSVP P2MP IPv4 Session sub-TLV value field is specified in the following figure. The value fields are taken from the definitions of the P2MP IPv4 LSP SESSION Object and the P2MP IPv4 SENDER_TEMPLATE Object in Sections 19.1.1 and 19.2.1 of [RFC4875]. Note that the Sub-Group ID of the SENDER_TEMPLATE is not required.



3.1.1.2. RSVP P2MP IPv6 Session Sub-TLV

The format of the RSVP P2MP IPv6 Session sub-TLV value field is specified in the following figure. The value fields are taken from the definitions of the P2MP IPv6 LSP SESSION Object and the P2MP IPv6 SENDER_TEMPLATE Object in Sections 19.1.2 and 19.2.2 of [RFC4875]. Note that the Sub-Group ID of the SENDER_TEMPLATE is not required.



3.1.2. Identifying a Multicast LDP LSP

[RFC4379] defines how a P2P LDP LSP under test may be identified in an echo request. A Target FEC Stack TLV is used to carry one or more sub-TLVs (for example, an IPv4 Prefix FEC sub-TLV) that identify the LSP.

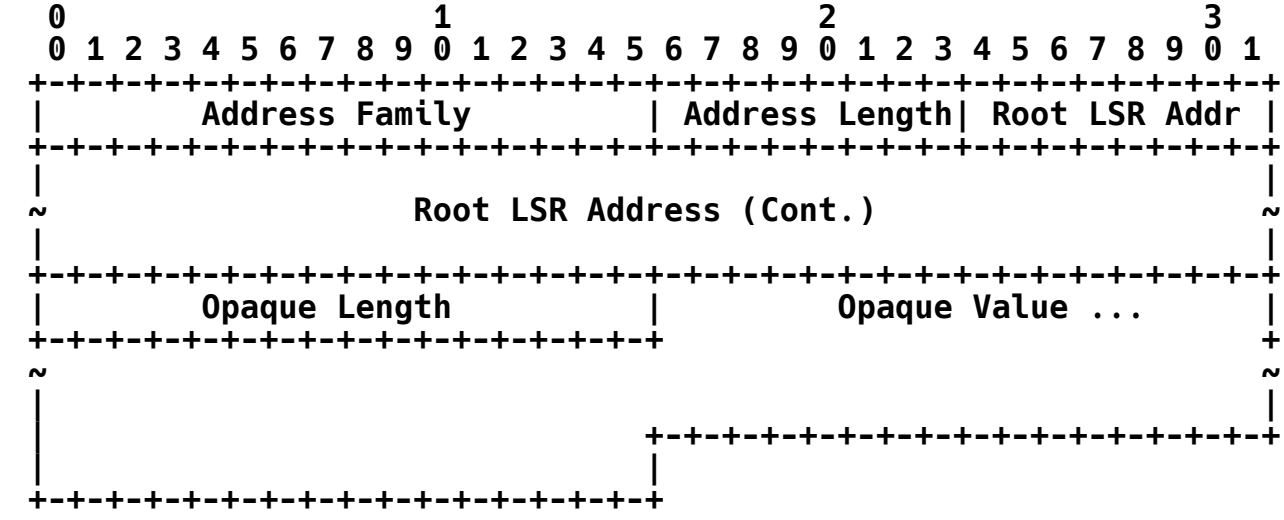
In order to identify a multicast LDP LSP under test, the echo request message MUST carry a Target FEC Stack TLV, and this MUST carry exactly one of two new sub-TLVs: either a Multicast P2MP LDP FEC Stack sub-TLV or a Multicast MP2MP LDP FEC Stack sub-TLV. These sub-TLVs use fields from the multicast LDP messages [RFC6388] and so provide sufficient information to uniquely identify the LSP.

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

Sub-Type #	Length	Value Field
19	Variable	Multicast P2MP LDP FEC Stack
20	Variable	Multicast MP2MP LDP FEC Stack

3.1.2.1. Multicast LDP FEC Stack Sub-TLVs

Both Multicast P2MP and MP2MP LDP FEC Stack have the same format, as specified in the following figure.



Address Family

Two-octet quantity containing a value from ADDRESS FAMILY NUMBERS in [IANA-AF] that encodes the address family for the Root LSR Address.

Address Length

Length of the Root LSR Address in octets.

Root LSR Address

Address of the LSR at the root of the P2MP LSP encoded according to the Address Family field.

Opaque Length

The length of the opaque value, in octets. Depending on the length of the Root LSR Address, this field may not be aligned to a word boundary.

Opaque Value

An opaque value element that uniquely identifies the P2MP LSP in the context of the Root LSR.

If the Address Family is IPv4, the Address Length MUST be 4. If the Address Family is IPv6, the Address Length MUST be 16. No other Address Family values are defined at present.

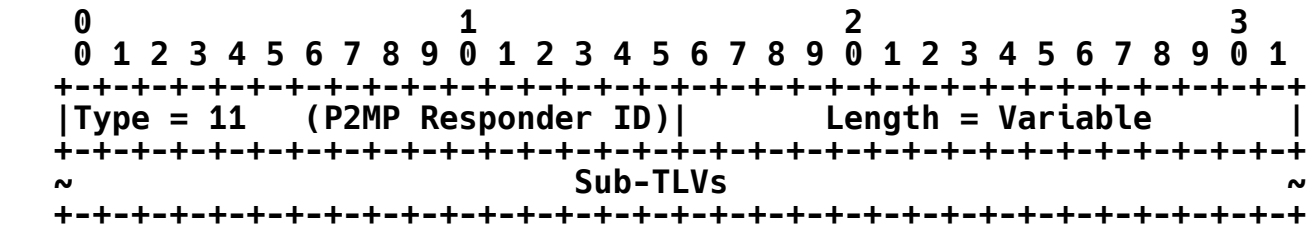
3.1.2.2. Applicability to Multipoint-to-Multipoint LSPs

The mechanisms defined in this document can be extended to include Multipoint-to-Multipoint (MP2MP) Multicast LSPs. In an MP2MP LSP tree, any leaf node can be treated like a head node of a P2MP tree. In other words, for MPLS OAM purposes, the MP2MP tree can be treated like a collection of P2MP trees, with each MP2MP leaf node acting like a P2MP head-end node. When a leaf node is acting like a P2MP head-end node, the remaining leaf nodes act like egress or bud nodes.

3.2. Limiting the Scope of Responses

A new TLV is defined for inclusion in the echo request message.

The P2MP Responder Identifier TLV is assigned the TLV type value 11 and is encoded as follows.



Sub-TLVs:

Zero, one, or more sub-TLVs as defined below.

If no sub-TLVs are present, the TLV MUST be processed as if it were absent. If more than one sub-TLV is present, the first TLV MUST be processed as described in this document, and subsequent sub-TLVs SHOULD be ignored. Interpretation of additional sub-TLVs may be defined in future documents.

The P2MP Responder Identifier TLV only has meaning on an echo request message. If present on an echo reply message, it MUST be ignored.

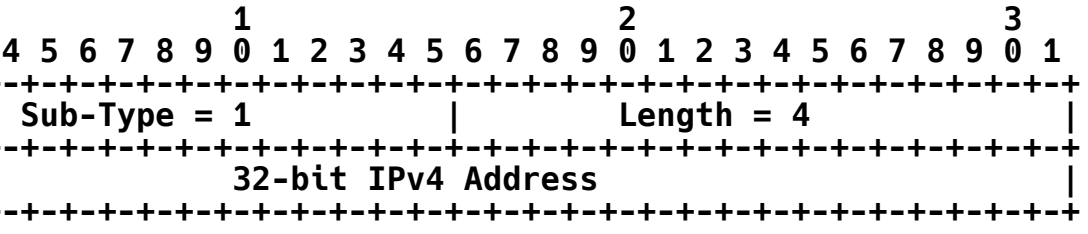
Four sub-TLVs are defined for inclusion in the P2MP Responder Identifier TLV carried on the echo request message. These are:

Sub-Type #	Length	Value Field
1	4	IPv4 Egress Address P2MP Responder
2	16	IPv6 Egress Address P2MP Responder
3	4	IPv4 Node Address P2MP Responder
4	16	IPv6 Node Address P2MP Responder

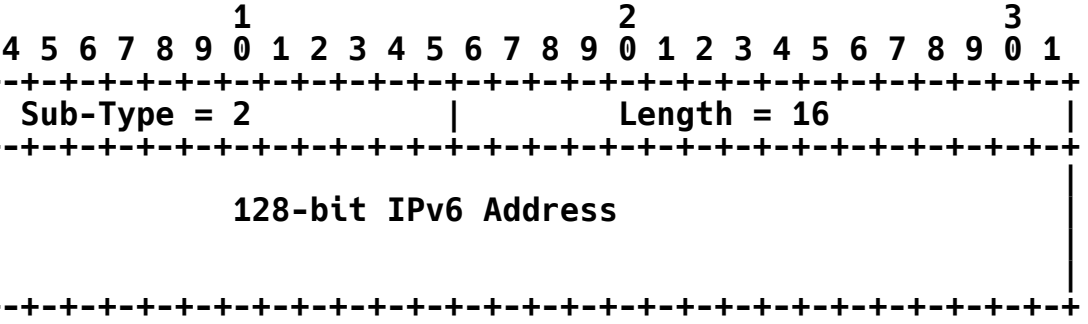
The content of these sub-TLVs are defined in the following sections. Also defined is the intended behavior of the responding node upon receiving any of these sub-TLVs.

3.2.1. Egress Address P2MP Responder Sub-TLVs

The encoding of the IPv4 Egress Address P2MP Responder sub-TLV is as follows:



The encoding of the IPv6 Egress Address P2MP Responder sub-TLV is as follows:



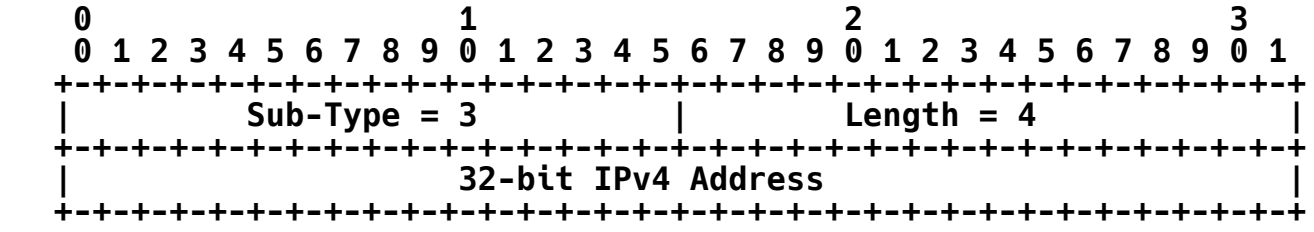
A node that receives an echo request with this sub-TLV present MUST respond if the node lies on the path to the address in the sub-TLV and MUST NOT respond if it does not lie on the path to the address in the sub-TLV. For this to be possible, the address in the sub-TLV must be known to the nodes that lie upstream in the LSP. This can be the case if RSVP-TE is used to signal the P2MP LSP, in which case this address will be the address used in the Destination Address

field of the S2L_SUB_LSP object, when corresponding egress or bud node is signaled. Thus, the IPv4 or IPv6 Egress Address P2MP Responder sub-TLV MAY be used in an echo request carrying RSVP P2MP Session sub-TLV.

However, in Multicast LDP, there is no way for upstream LSRs to know the identity of the downstream leaf nodes. Hence, these TLVs cannot be used to perform traceroute to a single node when Multicast LDP FEC is used, and the IPv4 or IPv6 Egress Address P2MP Responder sub-TLV SHOULD NOT be used with an echo request carrying a Multicast LDP FEC Stack sub-TLV. If a node receives these TLVs in an echo request carrying Multicast LDP, then it will not respond since it is unaware of whether it lies on the path to the address in the sub-TLV.

3.2.2. Node Address P2MP Responder Sub-TLVs

The encoding of the IPv4 Node Address P2MP Responder sub-TLV is as follows:



The encoding of the IPv6 Node Address P2MP Responder sub-TLV is as follows:

The IPv4 or IPv6 Node Address P2MP Responder sub-TLVs MAY be used in an echo request carrying either RSVP P2MP Session or Multicast LDP FEC Stack sub-TLVs.

A node that receives an echo request with one of these sub-TLVs present MUST respond if the address in the sub-TLV matches any address that is local to the node and MUST NOT respond if the address

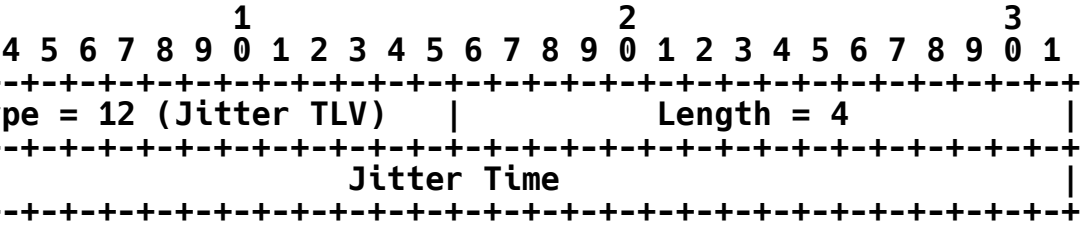
in the sub-TLV does not match any address that is local to the node. The address in the sub-TLV may be of any physical interface or may be the router ID of the node itself.

The address in this sub-TLV SHOULD be of any transit, branch, bud, or egress node for that P2MP LSP. The address of a node that is not on the P2MP LSP MAY be used as a check for that no reply is received.

3.3. Preventing Congestion of Echo Replies

A new TLV is defined for inclusion in the Echo request message.

The Echo Jitter TLV is assigned the TLV type value 12 and is encoded as follows.



Jitter Time:

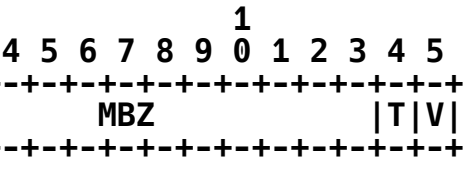
This field specifies the upper bound of the jitter period that should be applied by a responding node to determine how long to wait before sending an echo reply. A responding node MUST wait a random amount of time between zero milliseconds and the value specified in this field.

Jitter time is specified in milliseconds.

The Echo Jitter TLV only has meaning on an echo request message. If present on an echo reply message, it MUST be ignored.

3.4. Respond Only If TTL Expired Flag

A new flag is being introduced in the Global Flags field defined in [RFC4379]. The new format of the Global Flags field is:



The V flag is described in [RFC4379].

The T (Respond Only If TTL Expired) flag MUST be set only in the echo request packet by the sender. This flag MUST NOT be set in the echo reply packet. If this flag is set in an echo reply packet, then it MUST be ignored.

If the T flag is set to 0, then the receiving node MUST process the incoming echo request.

If the T flag is set to 1 and the TTL of the incoming MPLS label is equal to 1, then the receiving node MUST process the incoming echo request.

If the T flag is set to 1 and the TTL of the incoming MPLS label is more than 1, then the receiving node MUST drop the incoming echo request and MUST NOT send any echo reply to the sender.

If the T flag is set to 1 and there are no incoming MPLS labels in the echo request packet, then a bud node with PHP configured MAY choose to not respond to this echo request. All other nodes MUST ignore this bit and respond as per regular processing.

3.5. Downstream Detailed Mapping TLV

The Downstream Detailed Mapping TLV is described in [RFC6424]. A transit, branch or bud node can use the Downstream Detailed Mapping TLV to return multiple Return Codes for different downstream paths. This functionality can not be achieved via the Downstream Mapping TLV. As per Section 3.4 of [RFC6424], the Downstream Mapping TLV as described in [RFC4379] is being deprecated.

Therefore, for P2MP, a node MUST support the Downstream Detailed Mapping TLV. The Downstream Mapping TLV [RFC4379] is not appropriate for P2MP traceroute functionality and MUST NOT be included in an Echo Request message. When responding to an RSVP IPv4/IPv6 P2MP Session FEC type or a Multicast P2MP/MP2MP LDP FEC type, a node MUST ignore any Downstream Mapping TLV it receives in the echo request and MUST continue processing as if the Downstream Mapping TLV is not present.

The details of the Return Codes to be used in the Downstream Detailed Mapping TLV are provided in Section 4.

4. Operation of LSP Ping for a P2MP LSP

This section describes how LSP ping is applied to P2MP MPLS LSPs. As mentioned previously, an important design consideration has been to extend the existing LSP ping mechanism in [RFC4379] rather than invent new mechanisms.

As specified in [RFC4379], MPLS LSPs can be tested via a "ping" mode or a "traceroute" mode. The ping mode is also known as "connectivity verification" and traceroute mode is also known as "fault isolation". Further details can be obtained from [RFC4379].

This section specifies processing of echo requests for both ping and traceroute mode at various nodes (ingress, transit, etc.) of the P2MP LSP.

4.1. Initiating LSR Operations

The LSR initiating the echo request will follow the procedures in [RFC4379]. The echo request will contain a Target FEC Stack TLV. To identify the P2MP LSP under test, this TLV will contain one of the new sub-TLVs defined in Section 3.1. Additionally, there may be other optional TLVs present.

4.1.1. Limiting Responses to Echo Requests

As described in Section 2.2, it may be desirable to restrict the operation of P2MP ping or traceroute to a single egress. Since echo requests are forwarded through the data plane without interception by the control plane, there is no facility to limit the propagation of echo requests, and they will automatically be forwarded to all reachable egresses.

However, a single egress may be identified by the inclusion of a P2MP Responder Identifier TLV. The details of this TLV and its sub-TLVs are in Section 3.2. There are two main types of sub-TLVs in the P2MP Responder Identifier TLV: Node Address sub-TLV and Egress Address sub-TLV.

These sub-TLVs limit the replies either to the specified LSR only or to any LSR on the path to the specified LSR. The former capability is generally useful for ping mode, while the latter is more suited to traceroute mode. An initiating LSR may indicate that it wishes all egresses to respond to an echo request by omitting the P2MP Responder Identifier TLV.

4.1.2. Jittered Responses to Echo Requests

The initiating LSR MAY request that the responding LSRs introduce a random delay (or jitter) before sending the reply. The randomness of the delay allows the replies from multiple egresses to be spread over a time period. Thus, this technique is particularly relevant when the entire P2MP LSP is being pinged or traced since it helps prevent the initiating (or nearby) LSRs from being swamped by replies, or from discarding replies due to rate limits that have been applied.

It is desirable for the initiating LSR to be able to control the bounds of the jitter. If the tree size is small, only a small amount of jitter is required, but if the tree is large, greater jitter is needed.

The initiating LSR can supply the desired value of the jitter in the Echo Jitter TLV as defined in Section 3.3. If this TLV is present, the responding LSR MUST delay sending a reply for a random amount of time between zero milliseconds and the value indicated in the TLV. If the TLV is absent, the responding egress SHOULD NOT introduce any additional delay in responding to the echo request, but MAY delay according to local policy.

LSP ping MUST NOT be used to attempt to measure the round-trip time for data delivery. This is because the P2MP LSPs are unidirectional, and the echo reply is often sent back through the control plane. The timestamp fields in the echo request and echo reply packets MAY be used to deduce some information about delivery times, for example the variance in delivery times.

The use of echo jittering does not change the processes for gaining information, but note that the responding node MUST set the value in the Timestamp Received fields before applying any delay.

Echo reply jittering SHOULD be used for P2MP LSPs, although it MAY be omitted for simple P2MP LSPs or when the Node Address P2MP Responder sub-TLVs are used. If the Echo Jitter TLV is present in an echo request for any other type of LSPs, the responding egress MAY apply the jitter behavior as described here.

4.2. Responding LSR Operations

Usually the echo request packet will reach the egress and bud nodes. In case of TTL Expiry, i.e., traceroute mode, the echo request packet may stop at branch or transit nodes. In both scenarios, the echo request will be passed on to the control plane for reply processing.

The operations at the receiving node are an extension to the existing processing as specified in [RFC4379]. As described in that document, a responding LSR SHOULD rate-limit the receipt of echo request messages. After rate-limiting, the responding LSR must verify the general sanity of the packet. If the packet is malformed or certain TLVs are not understood, the [RFC4379] procedures must be followed for echo reply. Similarly, the Reply Mode field determines if the reply is required or not (and the mechanism to send it back).

For P2MP LSP ping and traceroute, i.e., if the echo request is carrying an RSVP P2MP FEC or a Multicast LDP FEC, the responding LSR MUST determine whether it is part of the P2MP LSP in question by checking with the control plane.

- If the node is not part of the P2MP LSP, it MUST respond according to [RFC4379] processing rules.
- If the node is part of the P2MP LSP, the node must check whether or not the echo request is directed to it.
  - If a P2MP Responder Identifier TLV is present, then the node must follow the procedures defined in Section 3.2 to determine whether or not it should respond to the request. The presence of a P2MP Responder Identifier TLV or a Downstream Detailed Mapping TLV might affect the Return Code. This is discussed in more detail later.
  - If the P2MP Responder Identifier TLV is not present (or, in the error case, is present, but does not contain any sub-TLVs), then the node MUST respond according to [RFC4379] processing rules.

4.2.1. Echo Reply Reporting

Echo reply messages carry Return Codes and Subcodes to indicate the result of the LSP ping (when the ping mode is being used) as described in [RFC4379].

When the responding node reports that it is an egress, it is clear that the echo reply applies only to that reporting node. Similarly, when a node reports that it does not form part of the LSP described by the FEC, then it is clear that the echo reply applies only to that reporting node. However, an echo reply message that reports an error from a transit node may apply to multiple egress nodes (i.e., leaves) downstream of the reporting node. In the case of the ping mode of operation, it is not possible to correlate the reporting node to the affected egresses unless the topology of the P2MP tree is already known, and it may be necessary to use the traceroute mode of operation to further diagnose the LSP.

Note that a transit node may discover an error, but it may also determine that while it does lie on the path of the LSP under test, it does not lie on the path to the specific egress being tested. In this case, the node SHOULD NOT generate an echo reply unless there is a specific error condition that needs to be communicated.

The following sections describe the expected values of Return Codes for various nodes in a P2MP LSP. It is assumed that the sanity and other checks have been performed and an echo reply is being sent back. As mentioned in Section 4.2, the Return Code might change based on the presence of a Responder Identifier TLV or Downstream Detailed Mapping TLV.

4.2.1.1. Responses from Transit and Branch Nodes

The presence of a Responder Identifier TLV does not influence the choice of the Return Code. For a success response, the Return Code MAY be set to value 8 ('Label switched at stack-depth <RSC>'). The notation <RSC> refers to the Return Subcode as defined in Section 3.1. of [RFC4379]. For error conditions, use appropriate values defined in [RFC4379].

The presence of a Downstream Detailed Mapping TLV will influence the choice of Return Code. As per [RFC6424], the Return Code in the echo reply header MAY be set to 'See DDM TLV for Return Code and Return Subcode' as defined in [RFC6424]. The Return Code for each Downstream Detailed Mapping TLV will depend on the downstream path as described in [RFC6424].

There will be a Downstream Detailed Mapping TLV for each downstream path being reported in the echo reply. Hence, for transit nodes, there will be only one such TLV, and for branch nodes, there will be more than one. If there is an Egress Address Responder sub-TLV, then the branch node will include only one Downstream Detailed Mapping TLV corresponding to the downstream path required to reach the address specified in the Egress Address sub-TLV.

4.2.1.2. Responses from Egress Nodes

The presence of a Responder Identifier TLV does not influence the choice of the Return Code. For a success response, the Return Code MAY be set to value 3 ('Replying router is an egress for the FEC at stack-depth <RSC>'). For error conditions, use appropriate values defined in [RFC4379].

The presence of the Downstream Detailed Mapping TLV does not influence the choice of Return Code. Egress nodes do not put in any Downstream Detailed Mapping TLV in the echo reply [RFC6424].

4.2.1.3. Responses from Bud Nodes

The case of bud nodes is more complex than other types of nodes. The node might behave as either an egress node or a transit node, or a combination of an egress and branch node. This behavior is

determined by the presence of any Responder Identifier TLV and the type of sub-TLV in it. Similarly, the Downstream Detailed Mapping TLV can influence the Return Code values.

To determine the behavior of the bud node, use the following rules. The intent of these rules is to figure out if the echo request is meant for all nodes, or just this node, or for another node reachable through this node or for a different section of the tree. In the first case, the node will behave like a combination of egress and branch node; in the second case, the node will behave like pure egress node; in the third case, the node will behave like a transit node; and in the last case, no reply will be sent back.

Node behavior rules:

- If the Responder Identifier TLV is not present, then the node will behave as a combination of egress and branch node.

- If the Responder Identifier TLV containing a Node Address sub-TLV is present, and:

  - If the address specified in the sub-TLV matches to an address in the node, then the node will behave like a combination of egress and branch node.

  - If the address specified in the sub-TLV does not match any address in the node, then no reply will be sent.

- If the Responder Identifier TLV containing an Egress Address sub-TLV is present, and:

  - If the address specified in the sub-TLV matches to an address in the node, then the node will behave like an egress node only.

  - If the node lies on the path to the address specified in the sub-TLV, then the node will behave like a transit node.

  - If the node does not lie on the path to the address specified in the sub-TLV, then no reply will be sent.

Once the node behavior has been determined, the possible values for Return Codes are as follows:

- If the node is behaving as an egress node only, then for a success response, the Return Code MAY be set to value 3 ('Replying router is an egress for the FEC at stack-depth <RSC>'). For error conditions, use appropriate values defined

in [RFC4379]. The echo reply MUST NOT contain any Downstream Detailed Mapping TLV, even if one is present in the echo request.

- If the node is behaving as a transit node, and:
  - If a Downstream Detailed Mapping TLV is not present, then for a success response, the Return Code MAY be set to value 8 ('Label switched at stack-depth <RSC>'). For error conditions, use appropriate values defined in [RFC4379].
  - If a Downstream Detailed Mapping TLV is present, then the Return Code MAY be set to 'See DDM TLV for Return Code and Return Subcode' as defined in [RFC6424]. The Return Code for the Downstream Detailed Mapping TLV will depend on the downstream path as described in [RFC6424]. There will be only one Downstream Detailed Mapping corresponding to the downstream path to the address specified in the Egress Address sub-TLV.
- If the node is behaving as a combination of egress and branch node, and:
  - If a Downstream Detailed Mapping TLV is not present, then for a success response, the Return Code MAY be set to value 3 ('Replying router is an egress for the FEC at stack-depth <RSC>'). For error conditions, use appropriate values defined in [RFC4379].
  - If a Downstream Detailed Mapping TLV is present, then for a success response, the Return Code MAY be set to value 3 ('Replying router is an egress for the FEC at stack-depth <RSC>'). For error conditions, use appropriate values defined in [RFC4379]. The Return Code for the each Downstream Detailed Mapping TLV will depend on the downstream path as described in [RFC6424]. There will be a Downstream Detailed Mapping for each downstream path from the node.

4.3. Special Considerations for Traceroute

4.3.1. End of Processing for Traceroutes

As specified in [RFC4379], the traceroute mode operates by sending a series of echo requests with sequentially increasing TTL values. For regular P2P targets, this processing stops when a valid reply is received from the intended egress or when some errored return code is received.

For P2MP targets, there may not be an easy way to figure out the end of the traceroute processing, as there are multiple egress nodes. Receiving a valid reply from an egress will not signal the end of processing.

For P2MP TE LSP, the initiating LSR has a priori knowledge about the number of egress nodes and their addresses. Hence, it is possible to continue processing until a valid reply has been received from each end point, provided that the replies can be matched correctly to the egress nodes.

However, for Multicast LDP LSP, the initiating LSR might not always know about all of the egress nodes. Hence, there might not be a definitive way to estimate the end of processing for traceroute.

Therefore, it is RECOMMENDED that traceroute operations provide for a configurable upper limit on TTL values. Hence, the user can choose the depth to which the tree will be probed.

4.3.2. Multiple Responses from Bud and Egress Nodes

The P2MP traceroute may continue even after it has received a valid reply from a bud or egress node, as there may be more nodes at deeper levels. Hence, for subsequent TTL values, a bud or egress node that has previously replied would continue to get new echo requests. Since each echo request is handled independently from previous requests, these bud and egress nodes will keep on responding to the traceroute echo requests. This can cause an extra processing burden for the initiating LSR and these bud or egress LSRs.

To prevent a bud or egress node from sending multiple replies in the same traceroute operation, a new "Respond Only If TTL Expired" flag is being introduced. This flag is described in Section 3.4.

It is RECOMMENDED that this flag be used for P2MP traceroute mode only. By using this flag, extraneous replies from bud and egress nodes can be reduced. If PHP is being used in the P2MP tree, then bud and egress nodes will not get any labels with the echo request packet. Hence, this mechanism will not be effective for PHP scenarios.

4.3.3. Non-Response to Traceroute Echo Requests

There are multiple reasons for which an ingress node may not receive a reply to its echo request. For example, the transit node has failed or the transit node does not support LSP ping.

When no reply to an echo request is received by the ingress, then (as per [RFC4379]) the subsequent echo request with a larger TTL SHOULD be sent in order to trace further toward the egress, although the ingress MAY halt the procedure at this point. The time that an ingress waits before sending the subsequent echo request is an implementation choice.

4.3.4. Use of Downstream Detailed Mapping TLV in Echo Requests

As described in Section 4.6 of [RFC4379], an initiating LSR, during traceroute, SHOULD copy the Downstream Mapping(s) into its next echo request(s). However, for P2MP LSPs, the initiating LSR will receive multiple sets of Downstream Detailed Mapping TLVs from different nodes. It is not practical to copy all of them into the next echo request. Hence, this behavior is being modified for P2MP LSPs. If the echo request is destined for more than one node, then the Downstream IP Address field of the Downstream Detailed Mapping TLV MUST be set to the ALLROUTERS multicast address, and the Address Type field MUST be set to either IPv4 Unnumbered or IPv6 Unnumbered depending on the Target FEC Stack TLV.

If an Egress Address Responder sub-TLV is being used, then the traceroute is limited to only one egress. Therefore this traceroute is effectively behaving like a P2P traceroute. In this scenario, as per Section 4.2, the echo replies from intermediate nodes will contain only one Downstream Detailed Mapping TLV corresponding to the downstream path required to reach the address specified in the Egress Address sub-TLV. For this case, the echo request packet MAY reuse a received Downstream Detailed Mapping TLV. This will allow interface validation to be performed as per [RFC4379].

4.3.5. Cross-Over Node Processing

A cross-over node will require slightly different processing for traceroute mode. The following definition of cross-over is taken from [RFC4875].

> The term "cross-over" refers to the case of an ingress or transit node that creates a branch of a P2MP LSP, a cross-over branch, that intersects the P2MP LSP at another node farther down the tree. It is unlike re-merge in that, at the intersecting node, the cross-over branch has a different outgoing interface as well as a different incoming interface.

During traceroute, a cross-over node will receive the echo requests via each of its input interfaces. Therefore, the Downstream Detailed Mapping TLV in the echo reply MUST carry information only about the outgoing interface corresponding to the input interface.

If this restriction is applied, the cross-over node will not duplicate the outgoing interface information in each of the echo request it receives via the different input interfaces. This will reflect the actual packet replication in the data plane.

5. Non-Compliant Routers

If a node for a P2MP LSP does not support MPLS LSP ping, then no reply will be sent, causing an incorrect result on the initiating LSR. There is no protection for this situation, and operators may wish to ensure that all nodes for P2MP LSPs are all equally capable of supporting this function.

If the non-compliant node is an egress, then the traceroute mode can be used to verify the LSP nearly all the way to the egress, leaving the final hop to be verified manually.

If the non-compliant node is a branch or transit node, then it should not impact ping mode. However the node will not respond during traceroute mode.

6. OAM and Management Considerations

The procedures in this document provide OAM functions for P2MP MPLS LSPs and may be used to enable bootstrapping of other OAM procedures.

In order to be fully operational, several considerations apply.

- Scaling concerns dictate that only cautious use of LSP ping should be made. In particular, sending an LSP ping to all egresses of a P2MP MPLS LSP could result in congestion at or near the ingress when the replies arrive.

  Further, incautious use of timers to generate LSP ping echo requests either in ping mode or especially in traceroute may lead to significant degradation of network performance.

- Management interfaces should allow an operator full control over the operation of LSP ping. In particular, such interfaces should provide the ability to limit the scope of an LSP ping echo request for a P2MP MPLS LSP to a single egress.

  Such interfaces should also provide the ability to disable all active LSP ping operations, to provide a quick escape if the network becomes congested.

- A MIB module is required for the control and management of LSP ping operations, and to enable the reported information to be inspected.

  There is no reason to believe this should not be a simple extension of the LSP ping MIB module used for P2P LSPs.

7. IANA Considerations

7.1. New Sub-TLV Types

Four new sub-TLV types are defined for inclusion within the LSP ping [RFC4379] Target FEC Stack TLV (TLV type 1).

IANA has assigned sub-type values to the following sub-TLVs under TLV type 1 (Target FEC Stack) from the "Multi-Protocol Label Switching (MPLS) Label Switched Paths (LSPs) Ping Parameters" registry, "TLVs and sub-TLVs" sub-registry.

17 RSVP P2MP IPv4 Session (Section 3.1.1)
18 RSVP P2MP IPv6 Session (Section 3.1.1)
19 Multicast P2MP LDP FEC Stack (Section 3.1.2)
20 Multicast MP2MP LDP FEC Stack (Section 3.1.2)

7.2. New TLVs

Two new LSP ping TLV types are defined for inclusion in LSP ping messages.

IANA has assigned a new value from the "Multi-Protocol Label Switching Architecture (MPLS) Label Switched Paths (LSPs) Ping Parameters" registry, "TLVs and sub-TLVs" sub-registry as follows using a Standards Action value.

11 P2MP Responder Identifier TLV (see Section 3.2) is a mandatory TLV.

Four sub-TLVs are defined.
- Sub-Type 1: IPv4 Egress Address P2MP Responder
- Sub-Type 2: IPv6 Egress Address P2MP Responder
- Sub-Type 3: IPv4 Node Address P2MP Responder
- Sub-Type 4: IPv6 Node Address P2MP Responder

12 Echo Jitter TLV (see Section 3.3) is a mandatory TLV.

7.3. New Global Flags Registry

IANA has created a new sub-registry of the "Multi-Protocol Label Switching (MPLS) Label Switched Paths (LSPs) Ping Parameters" registry. The sub-registry is called the "Global Flags" registry.

This registry tracks the assignment of 16 flags in the Global Flags field of the MPLS LSP ping echo request message. The flags are numbered from 0 (most significant bit, transmitted first) to 15.

New entries are assigned by Standards Action.

Initial entries in the registry are as follows:

Bit number	Name	Reference
15	V Flag	[RFC4379]
14	T Flag	[RFC6425]
13-0	Unassigned	

8. Security Considerations

This document does not introduce security concerns over and above those described in [RFC4379]. Note that because of the scalability implications of many egresses to P2MP MPLS LSPs, there is a stronger concern about regulating the LSP ping traffic passed to the control plane by the use of a rate limiter applied to the LSP ping well-known UDP port. This rate limiting might lead to false indications of LSP failure.

9. Acknowledgements

The authors would like to acknowledge the authors of [RFC4379] for their work, which is substantially re-used in this document. Also, thanks to the members of the MBONED working group for their review of this material, to Daniel King and Mustapha Aissaoui for their reviews, and to Yakov Rekhter for useful discussions.

The authors would like to thank Bill Fenner, Vanson Lim, Danny Prairie, Reshad Rahman, Ben Niven-Jenkins, Hannes Gredler, Nitin Bahadur, Tetsuya Murakami, Michael Hua, Michael Wildt, Dipa Thakkar, Sam Aldrin, and IJsbrand Wijnands for their comments and suggestions.

10. References

10.1. Normative References

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

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

[RFC6424] Bahadur, N., Kompella, K., and G. Swallow, "Mechanism for Performing LSP-Ping over MPLS Tunnels", RFC 6424, November 2011.

10.2. Informative References

[IANA-AF] IANA Assigned Port Numbers, <http://www.iana.org/assignments/address-family-numbers>.

[RFC792] Postel, J., "Internet Control Message Protocol", STD 5, RFC 792, September 1981.

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

[RFC4687] Yasukawa, S., Farrel, A., King, D., and T. Nadeau, "Operations and Management (OAM) Requirements for Point-to-Multipoint MPLS Networks", RFC 4687, September 2006.

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

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

[RFC6348] Le Roux, JL., Ed., and T. Morin, Ed., "Requirements for Point-to-Multipoint Extensions to the Label Distribution Protocol", RFC 6348, September 2011.

[RFC6388] Wijnands, IJ., Ed., Minei, I., Ed., Kompella, K., and B. Thomas, "Label Distribution Protocol Extensions for Point-to-Multipoint and Multipoint-to-Multipoint Label Switched Paths", RFC 6388, November 2011.

Authors' Addresses

Shaleen Saxena
Cisco Systems, Inc.
1414 Massachusetts Ave
Boxborough, MA 01719
EMail: ssaxena@cisco.com

George Swallow
Cisco Systems, Inc.
1414 Massachusetts Ave
Boxborough, MA 01719
EMail: swallow@cisco.com

Zafar Ali
Cisco Systems Inc.
2000 Innovation Drive
Kanata, ON, K2K 3E8, Canada.
Phone: 613-889-6158
EMail: zali@cisco.com

Adrian Farrel
Juniper Networks
EMail: adrian@olddog.co.uk

Seisho Yasukawa
NTT Corporation
3-9-11, Midori-Cho Musashino-Shi
Tokyo 180-8585 Japan
Phone: +81 422 59 2684
EMail: yasukawa.seisho@lab.ntt.co.jp

Thomas D. Nadeau
CA Technologies, Inc.
273 Corporate Drive
Portsmouth, NH 03801

EMail: thomas.nadeau@ca.com