Network Working Group
Internet-Draft
Intended Status: Standards Track
Updates: RFC4379
Created: August 11, 2009
Expires: February 11, 2010

A. Farrel (Editor)
Old Dog Consulting
S. Yasukawa
NTT

Detecting Data Plane Failures in Point-to-Multipoint Multiprotocol Label Switching (MPLS) - Extensions to LSP Ping

draft-ietf-mpls-p2mp-lsp-ping-08.txt

Status of this Memo

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.

Copyright Notice

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].

Contents

0. Change Log

This section to be removed before publication as an RFC.

0.1 Changes from 00 to 01

- Update references.
- Fix boilerplate.

0.2 Changes from 01 to 02

- Update entire document so that it is not specific to MPLS-TE, but also includes multicast LDP LSPs.
- Move the egress identifier sub-TLVs from the FEC Stack TLV to a new egress identifier TLV.
- Include Multicast LDP FEC Stack sub-TLV definition from [MCAST-CV].
- Add brief section on use of LSP Ping for bootstrapping.
- Add new references to References section.
- Add details of two new authors.

0.3 Changes from 02 to 03

- Update references.
- Update boilerplate.
- Fix typos.
- Clarify in 3.2.2 that a recipient of an echo request must reply only once it has applied incoming rate limiting.
- Tidy references to bootstrapping for [MCAST-CV] in 1.1.
- Allow multiple sub-TLVs in the P2MP Egress Identifier TLV in sections 3.2.1, 3.2.2, 3.2.4, 3.3.1, and 3.3.4.
- Clarify how to handle a P2MP Egress Identifier TLV with no sub-TLVs in sections 3.2.1 and 3.2.2.

0.4 Changes from 03 to 04

- Revert to previous text in sections 3.2.1, 3.2.2, 3.2.4, 3.3.1, and 3.3.4 with respect to multiple sub-TLVs in the P2MP Egress Identifier TLV.

0.5 Changes from 04 to 05

- Change coordinates for Tom Nadeau. Section 13.
- Fix typos.
- Update references.
- Resolve all acronym expansions.

0.6 Changes from 05 to 06

- New section, 3.2.6, to explain echo response reporting in the Ping case.

- New section, 3.3.7, to explain echo response reporting in the Traceroute case.
- Sections 3.3.2, 3.3.5, and 5. Retire the E-flag for identification of bud nodes. Use the B-flag in a Downstream Mapping TLV with a zero address to provide the necessary indication.
- Section 3.3.4. Note the use of ALLROUTERS address as per RFC 4379
- Section 7. Suggest values for IANA assignment.
- Rename "P2MP Responder Identifier TLV" to "P2MP Responder Identifier TLV", "Egress Identifier sub-TLV" to "Responder Identifier sub-TLV", and "P2MP egresses" multipath type to "P2MP responder". This allows any LSR on the P2MP LSP to be the target of, or responder to, an echo request.

0.7 Changes from 06 to 07

- Sections 3.3.2 and 3.3.3. Delete section 3.3.5. New sections 3.3.2.1 through 3.3.2.3: Retire B-flag from Downstream Mapping TLV. Introduce new Node Properties TLV with Branching Properties and Egress Address sub-TLVs.
- Section 3.3.2.4: Clarify rules on presence of Multipath Information in Downstream Mapping TLVs.
- Section 3.3.5: Clarify padding rules.
- Section 3.3.6: Updated to use Downstream Detailed Mapping TLVs for multiple return conditions reported by a single echo response.
- Section 7: Update IANA values and add new sub-sections.
- Section 11: Add reference draft-ietf-mpls-lsp-ping-enhanced-dsmap.
- Section 13: Update Bill Fenner's coordinates.

0.8 Changes from 07 to 08

- Removed the Node Properties TLV (Section 3.3.2.1 of version 07).
- Removed the New Multipath Type from Multipath Sub-TLV (Section 3.3.5 of version 07).
- Removed the Return Code Sub-TLV from Downstream Detailed TLV (Section 3.3.6.1 of version 07), as it is already included in draft-ietf-mpls-lsp-ping-enhanced-dsmap-02.
- Clarified the behavior of Responder Identifier TLV (Section 3.2.4 of version 07). Two new Sub-TLVs are introduced.
- Downstream Detailed Mapping TLV is now mandatory for implementing P2MP OAM functionality.
- Split Multicast LDP TLV into two TLVs, one for P2MP and other for MP2MP. Also added description to allow MP2MP ping by using this draft.
- Removed Section 4. as it was a duplicate of Section 2.3.

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 an MPLS "echo request" and "echo reply", and mechanisms for transporting the echo reply. 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 point-to-multipoint extensions to the Label Distribution Protocol (LDP) are stated in [P2MP-LDP-REQ]. [P2MP-LDP] 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. Mechanisms are, therefore, desirable 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 and so that they can be used to bootstrap other Operations and Management (OAM) procedures such as [MCAST-CV]. 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, it 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 OAM procedures such as [MPLS-BFD] and [MCAST-CV]. 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 well-known

UDP port defined for use by LSP Ping traffic.

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 [P2MP-LDP] uses the Label Distribution Protocol 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 TE 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 response 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 also returns information on an echo response that helps verify the control plane against the data plane. That is, the information is used by the ingress to check that the data plane forwarding matches what is signaled by the control plane.

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 responses 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. Therefore, the procedures described in this document provide a mechanism that allows the responders to randomly delay (or jitter) their responses so that the chances of swamping the ingress are reduced.

Further, the procedures in this document allow the initiator to limit the scope of an LSP Ping echo request (ping or traceroute mode) to one specific intended egress.

The scalability issues surrounding LSP Ping for P2MP MPLS LSPs may be addressed by other mechanisms such as [MCAST-CV] that utilize the LSP Ping procedures in this document to provide bootstrapping mechanisms as described in Section 2.3.

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.

2.3 Bootstrapping Other OAM Procedures Using LSP Ping

[MPLS-BFD] describes a process where LSP Ping [RFC4379] is used to bootstrap the Bidirectional Forwarding Detection (BFD) mechanism [BFD] for use to track the liveliness of an MPLS LSP. In particular BFD can be used to detect a data plane failure in the forwarding path of an MPLS LSP.

Requirements for MPLS P2MP LSPs extend to hundreds or even thousands of endpoints. If a protocol required explicit acknowledgments to each probe for connectivity verification, the response load at the root would be overwhelming.

A more scalable approach to monitoring P2MP LSP connectivity is described in [MCAST-CV]. It relies on using the MPLS echo request and echo response messages of LSP Ping [RFC4379] to bootstrap the monitoring mechanism in a manner similar to [MPLS-BFD]. The actual monitoring is done using a separate process defined in [MCAST-CV].

Note that while the approach described in [MCAST-CV] was developed in response to the multicast scalability problem, it can be applied to P2P LSPs as well.

3. Operation of LSP Ping for a P2MP LSP

This section describes how LSP Ping is applied to P2MP MPLS LSPs.

It covers the mechanisms and protocol fields applicable to both ping mode and traceroute mode. It explains the responsibilities of the

initiator (ingress), transit nodes, and receivers (egresses).

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.

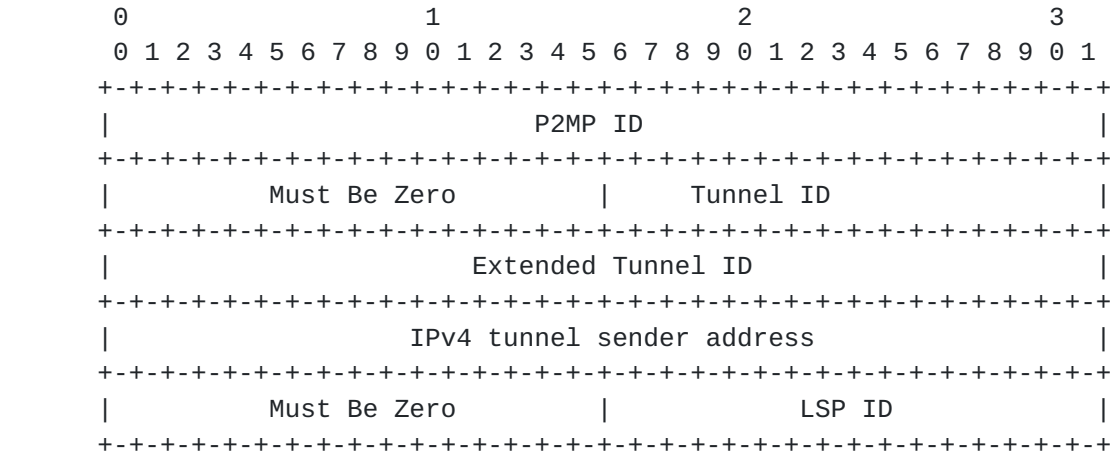
In order to identify the P2MP MPLS TE LSP under test, the echo request message MUST carry a Target FEC Stack TLV, and this MUST carry exactly one of two new sub-TLVs: either an RSVP P2MP IPv4 Session sub-TLV or an RSVP P2MP IPv6 Session sub-TLV. These sub-TLVs carry fields from the RSVP-TE P2MP Session 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
TBD	20	RSVP P2MP IPv4 Session
TBD	56	RSVP P2MP IPv6 Session

3.1.1.1. RSVP P2MP IPv4 Session Sub-TLV

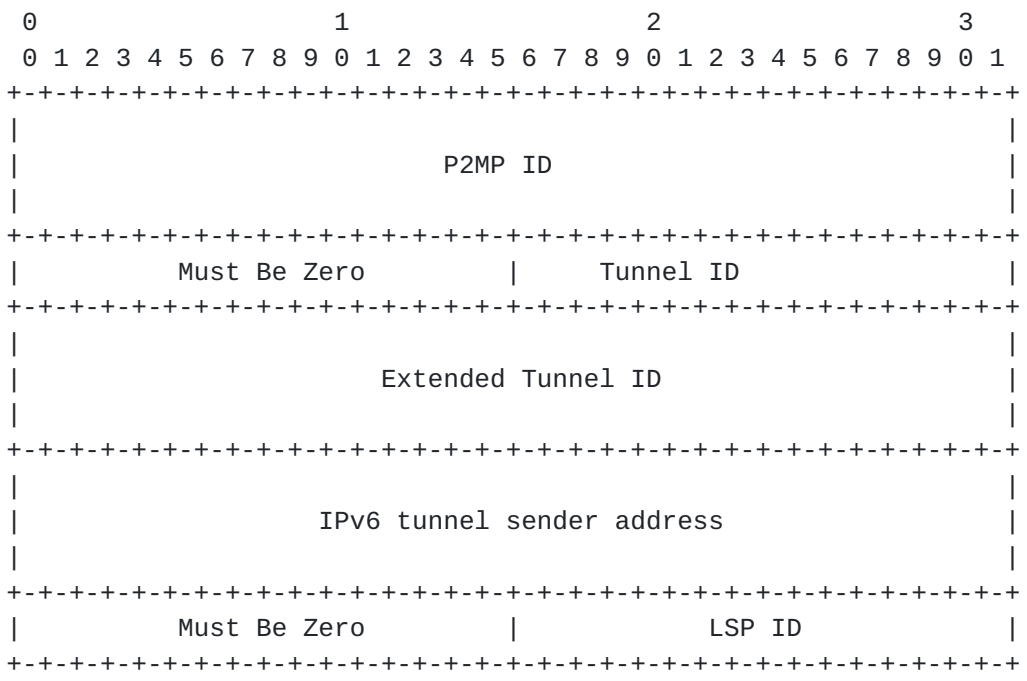
The format of the RSVP P2MP IPv4 Session sub-TLV value field is specified in the following figure. The value fields are taken from the definitions of the P2MP IPv4 LSP Session Object and the P2MP IPv4 Sender-Template Object in [RFC4875]. Note that the Sub-Group ID of the Sender-Template is not required.



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 [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.

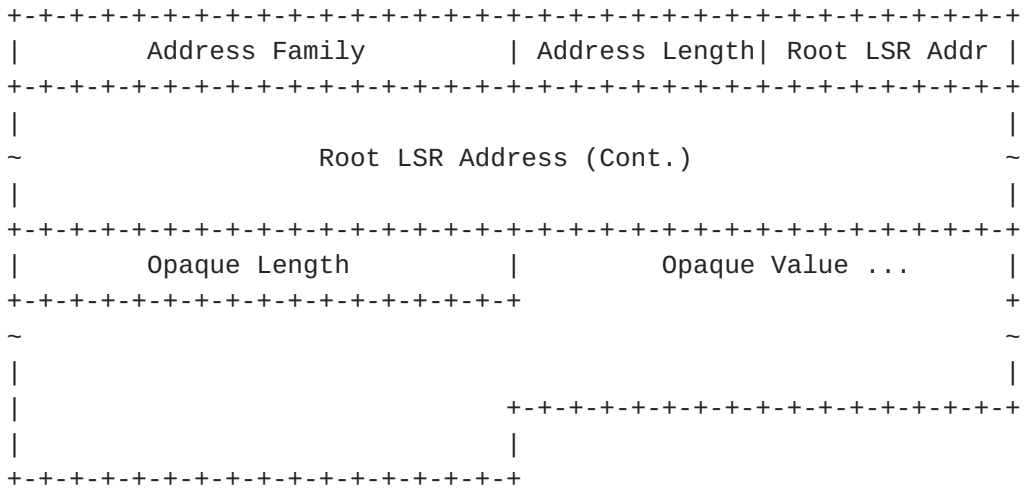
In order to identify a multicast LDP LSP under test, the echo request message MUST carry a Target FEC Stack TLV, and this MUST carry exactly one new sub-TLV: the Multicast LDP FEC Stack sub-TLV. This sub-TLV uses fields from the multicast LDP messages [P2MP-LDP] and so provides sufficient information to uniquely identify the LSP.

The new sub-TLV is assigned a sub-type identifier as follows, and is described in the following section.

Sub-Type #	Length	Value Field
TBD	Variable	Multicast P2MP LDP FEC Stack
TBD	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-PORT] 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.

Opaque Value

An opaque value element which 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 nodes.

3.2. Ping Mode Operation

3.2.1. Controlling Responses to LSP Pings

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

However, the intended egress under test can be identified by the inclusion of a P2MP Responder Identifier TLV. The details of this TLV and its Sub-TLVs are in section 3.2.4. The initiator may choose whether only the node identified in the TLV responds or any node on the path to the node identified in the TLV may respond.

An initiator may indicate that it wishes all egresses to respond to an echo request by omitting the P2MP Responder Identifier TLV.

Note that the ingress of a multicast LDP LSP will not know the identities of the egresses of the LSP except by some external means such as running P2MP LSP Ping to all egresses.

3.2.2. Ping Mode Egress Procedures

An egress node is RECOMMENDED to rate limit its receipt of echo request messages as described in [RFC4379]. After rate limiting, an egress node that receives an echo request carrying an RSVP P2MP IPv4 Session sub-TLV, an RSVP P2MP IPv6 Session sub-TLV, or a Multicast LDP FEC Stack sub-TLV MUST determine whether it is an egress of the P2MP LSP in question by checking with the control plane.

- If the node is not an egress, it MUST respond according to the setting of the Response Type field in the echo message following the rules defined in [RFC4379].

- If the node is an egress of the P2MP LSP, the node must check whether it is a receipient of the echo request.
  - If a P2MP Responder Identifier TLV is present, then the node must follow the procedures defined in section 3.2.4 to determine whether it should respond to the reqeust or not.
  - If the P2MP Responder Identifier TLV is not present (or, in the error case, is present, but does not contain any sub-TLVs), and the egress node that received the echo request is an intended egress of the LSP, the node MUST respond according to the setting of the Response Type field in the echo message following the rules defined in [RFC4379].

3.2.3. Jittered Responses

The initiator (ingress) of a ping request MAY request the responding egress to introduce a random delay (or jitter) before sending the response. The randomness of the delay allows the responses from multiple egresses to be spread over a time period. Thus this technique is particularly relevant when the entire LSP tree is being pinged since it helps prevent the ingress (or nearby routers) from being swamped by responses, or from discarding responses due to rate limits that have been applied.

It is desirable for the ingress to be able to control the bounds within which the egress delays the response. 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 ingress informs the egresses of the jitter bound by supplying a value in a new TLV (the Echo Jitter TLV) carried on the echo request message. If this TLV is present, the responding egress MUST delay sending a response 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.

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

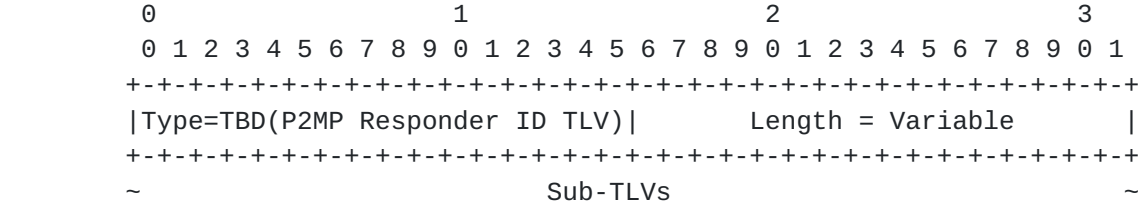
The use of echo jittering does not change the processes for gaining information, but note that the responding egress MUST set the value in the Timestamp Received fields before applying any delay.

It is RECOMMENDED that echo response jittering is not used except in the case of P2MP LSPs. If the Echo Jitter TLV is present in an echo request for any other type of TLV, the responding egress MAY apply the jitter behavior described here.

3.2.4. P2MP Responder Identifier TLV and Sub-TLVs

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

The P2MP Responder Identifier TLV is assigned the TLV type value TBD 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 MUST be processed as described in this document, and subsequent sub-TLVs SHOULD be ignored.

The P2MP Responder Identifier TLV only has meaning on an echo request message. If present on an echo response message, it SHOULD 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 Identifier
2	16	IPv6 Egress Address P2MP Responder Identifier
3	4	IPv4 Node Address P2MP Responder Identifier
4	16	IPv6 Node Address P2MP Responder Identifier

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. Please note that the echo response is always controlled by Response Type field in the echo message as defined in [RFC4379] and whether or not the responding node is part for the P2MP tree being identified in the Target FEC Stack TLV. The Sub-TLVs defined in this section provide additional constraints to those requirements and are not a replacement for those requirements.

3.2.4.1. Egress Address P2MP Responder Identifier Sub-TLVs

The IPv4 or IPv6 Egress Address P2MP Responder Identifier Sub-TLVs MAY be used in an echo request carrying RSVP P2MP Session Sub-TLV. They SHOULD NOT be used with an echo request carrying Multicast LDP FEC Stack Sub-TLV.

A node that receives an echo request with this Sub-TLV present MUST respond only if the node lies on the path to the address in the Sub-TLV.

The address in this Sub-TLV SHOULD be of an egress or bud node and SHOULD NOT be of a transit or branch node. This address MUST be known to the nodes upstream of the target node, possibly via control plane signaling, such as RSVP. This Sub-TLV may be used to trace a specific egress or bud node in the P2MP tree.

3.2.4.2. Node Address P2MP Responder Identifier Sub-TLVs

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

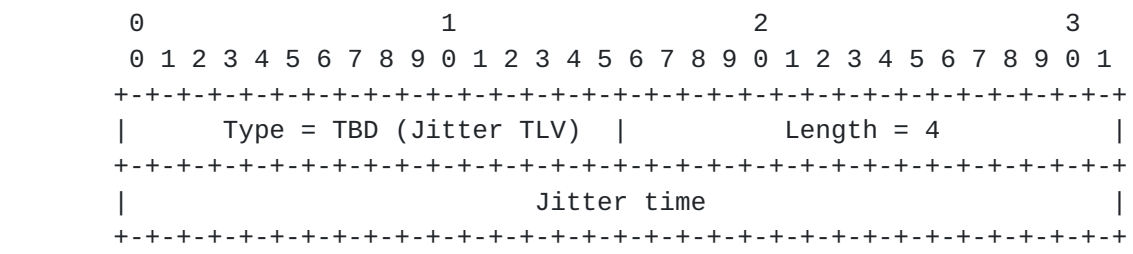
A node that receives an echo request with this Sub-TLV present MUST respond only if the address in the Sub-TLV corresponds to any address that is local to the node. This address in the Sub-TLV may be of any physical interface or may be the router id of the node itself.

The address in this Sub-TLV SHOULD be of any transit, branch, bud or egress node for that P2MP tree. This Sub-TLV may be used to ping any specific node in the P2MP tree.

3.2.5. Echo Jitter TLV

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

The Echo Jitter TLV is assigned the TLV type value TBD 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 response. A responding node SHOULD 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 response message, it SHOULD be ignored.

3.2.6. Echo Response Reporting

Echo response 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 response applies only to the reporting node. Similarly,

when a node reports that it does not form part of the LSP described by the FEC (i.e. there is a misconnection) then the echo response

applies to the reporting node.

However, it should be noted that an echo response 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 shape of the P2MP tree is already known, and it may be necessary to use the Traceroute mode of operation (see Section 3.3) to further diagnose the LSP.

Note also that a transit node may discover an error but 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 response.

3.2.6.1 Ping Responses at Transit and Branch Nodes

If the TTL of the MPLS packet carrying an echo request expires at a transit or branch node, the packet MUST be passed to the control plane as specified in [RFC4379].

If the P2MP Responder Identifier is not present or does not contain any Sub-TLV, then the node MUST respond. If the P2MP Responder Identifier Sub-TLV is present, then the node MUST respond as per section 3.2.4.

If the echo response being sent is not indicating an error condition, such as Malformed request, then the Return Code in the echo response header may be set to value 8 ('Label switched at stack-depth <RSC>') or any other error value as needed.

3.2.6.2 Ping Responses at Egress and Bud Nodes

The echo request packet MUST be sent to the control plane at egress and bud nodes.

If the P2MP Responder Identifier is not present or does not contain any Sub-TLV, then the node MUST respond. If the P2MP Responder Identifier Sub-TLV is present, then the node MUST respond as per section 3.2.4.

If the echo response being sent is not indicating an error condition, such as Malformed request, then the Return Code in the echo response header may be set to value 3 ('Replying router is an egress for the FEC at stack-depth <RSC>') or any other error value as needed.

3.3. Traceroute Mode Operation

The traceroute mode of operation is described in [RFC4379]. Like other traceroute operations, it relies on the expiration of the TTL

of the packet that carries the echo request. When the TTL expires the echo request is passed to the control plane on the transit node which

responds according to the Response Type in the message (and any Responder Identifier TLV that may be present).

Echo requests MAY include a Downstream Detailed Mapping TLV, and a responding node fills in the fields of the Downstream Detailed Mapping TLV to indicate the downstream interfaces and labels used by the reported LSP from the responding node. In this way, by successively sending out echo requests with increasing TTLs, the ingress may gain a picture of the path and resources used by an LSP. This process continues either to the point of failure when no response is received, or an error response is generated by a node where the control plane does not expect to be handling the LSP.

For P2MP Traceroute, a node MUST support Downstream Detailed Mapping TLV [DDMT]. Downstream Mapping TLV [RFC4379] SHOULD NOT be used for P2MP traceroute functionality. As per Section 4.3 of [DDMT], Downstream Mapping TLV is being deprecated. A node MUST ignore any Downstream Mapping TLV it receives in the echo request.

If there are nodes in the P2MP tree that do not support Downstream Detailed Mapping TLV, they will send an echo reply with Return Code set to 2. The ingress node upon receiving such a value SHOULD send subsequent echo requests with a larger TTL.

The traceroute mode of operation is equally applicable to P2MP MPLS TE LSP and P2MP Multicast LDP LSP and is described in the following sections.

The traceroute mode can be applied to all destinations of the P2MP tree just as in the ping mode. In the case of P2MP MPLS TE LSPs, the traceroute mode can also be applied to individual traceroute targets identified by the presence of a P2MP Responder Identifier TLV. In this case, the responding node must follow the behavior specified in 3.2.4. These targets SHOULD be egresses or bud nodes. However, since a transit node of a multicast LDP LSP is unable to determine whether it lies on the path to any one destination or any other transit node, the traceroute mode limited to specific nodes of such an LSP MUST NOT be used.

In the absence of a P2MP Responder Identifier TLV, the echo request is asking for traceroute information applicable to all egresses.

The echo response jitter technique described for the ping mode is equally applicable to the traceroute mode and is not additionally described in the procedures below.

3.3.1. Correlating Traceroute Responses

When traceroute is simultaneously applied to multiple responders (e.g. egresses), it is important that the ingress is able to

correlate the echo responses with the nodes in the P2MP tree. Without this information the ingress will be unable to determine the correct

ordering of transit nodes. One possibility is for the ingress to poll the path to each responder in turn, but this may be inefficient, undesirable, or (in the case of multicast LDP LSPs) illegal.

The Downstream Detailed Mapping TLV MUST be included in the echo response from transit, bud, or branch nodes. The information from Downstream Detailed Mapping TLV can be pieced together by the ingress to reconstruct the P2MP tree although it may be necessary to refer to the routing information distributed by the IGP to correlate next hop addresses and node reporting addresses in subsequent echo responses.

The following sections describe the Return Code used in the echo response header and in the Downstream Detailed Mapping TLV. It is possible to identify the type of node (transit, branch, bud and egress) by using various values in the Return Code and presence of Downstream Detailed Mapping TLV.

3.3.2. Traceroute Responses at Transit Nodes

When the TTL of the MPLS packet carrying an echo request expires the packet MUST be passed to the control plane as specified in [RFC4379].

If the echo request packet contains an IPv4 or IPv6 Egress Address P2MP Responder Identifier TLV, and the FEC is IPv4 or IPv6 P2MP TE LSP, then the node MUST respond only if the node lies on the path to the egress specified in the Sub-TLV.

If the LSP under test is a multicast LDP LSP and echo request has an IPv4 or IPv6 Egress Address P2MP Responder Identifier TLV, then the node MUST treat the echo request as malformed and MUST process it according to the rules specified in [RFC4379].

If the echo response being sent is not indicating an error condition, such as Malformed request, it MUST identify the next hop of the path of the LSP in the data plane by including a Downstream Detailed Mapping TLV as described in [DDMT].

The Return Code in echo response header will be value TBD ('See DDM TLV for Return Code and Return SubCode') as defined in [DDMT]. The Return Code for the Downstream Detailed Mapping TLV will depend on the state of the output interface.

3.3.3. Traceroute Responses at Branch Nodes

A branch node MUST follow the procedures described in Section 3.3.2 to determine whether it should respond to an echo request.

If the P2MP Responder Identifier is not present or does not contain any Sub-TLV (that is, if all egresses are being traced), then the

branch node MUST add a Downstream Detailed Mapping TLV to the echo response for each outgoing branch that it reports.

If an IPv4 or IPv6 Egress Address P2MP Responder Identifier is present, it MUST report only the branch that is on the path to the specified egress node and it MUST NOT report the other branches.

The Return Code in echo response header will be value TBD ('See DDM TLV for Return Code and Return SubCode') as defined in [DDMT]. The Return Code for each of the Downstream Detailed Mapping TLV will depend on the state of the output interface being reported in this TLV.

3.3.4. Traceroute Responses at Egress Nodes

If P2MP Responder Identifier is not present or does not contain any Sub-TLV (that is, if all egresses are being traced), then the egress node MUST respond to the echo request.

If an IPv4 or IPv6 Egress Address P2MP Responder Identifier is present, it MUST respond only if the specified address belongs the egress node.

Egress node MUST NOT return a Downstream Detailed Mapping TLV.

The Return Code in the echo response header will be value 3 ('Replying router is an egress for the FEC at stack-depth <RSC>') as defined in [RFC4379].

3.3.5. Traceroute Responses at Bud Nodes

Some nodes on a P2MP MPLS LSP may be an egress as well as a branch (i.e. have one or more downstream nodes). Such nodes are known as bud nodes [RFC4461]. A bud node's response is a combination of branch node and egress node behavior.

If P2MP Responder Identifier is not present or does not contain any Sub-TLV (that is, if all egresses are being traced), then the bud node MUST respond to the echo request. It MUST add a Downstream Detailed Mapping TLV to the echo response for each outgoing branch that it reports. The Return Code in the echo response header will be value 3 ('Replying router is an egress for the FEC at stack-depth <RSC>') as defined in [RFC4379]. The Return Code for each of the Downstream Detailed Mapping TLV will depend on the state of the output interface being reported in this TLV.

If an IPv4 or IPv6 Egress Address P2MP Responder Identifier is present, and the specified address belongs the bud node, then it MUST respond as if it were an egress node. The Return Code in the echo response header will be value 3 ('Replying router is an egress for the FEC at stack-depth <RSC>') as defined in [RFC4379]. It MUST NOT report any Downstream Detailed Mapping TLV.

If an IPv4 or IPv6 Egress Address P2MP Responder Identifier is

present, and the bud node lies on the path to the specified egress address, then it MUST respond as if it was a branch node. The Return Code in the echo response header will be value TBD ('See DDM TLV for Return Code and Return SubCode') as defined in [DDMT]. The Return Code for each of the Downstream Detailed Mapping TLV will depend on the state of the output interface being reported in this TLV.

3.3.6. Non-Response to Traceroute Echo Requests

There are multiple reasons for which an ingress node may not receive a response to its echo request. For example, perhaps because the transit node has failed, or perhaps because the transit node does not support LSP Ping, or the Responder Identifier TLV failed to match a valid node.

When no response to an echo request is received by the ingress, then as per [RFC4379] the subsequent echo request with a larger TTL SHOULD be sent.

3.3.7 Use of Downstream Detailed Mapping TLV in Echo Request

If no Responder Identifier TLV is being used, then in the Echo Request packet, the "Downstream IP Address" field, of the Downstream Detailed Mapping TLV, MUST be set to the ALLROUTERs multicast address.

If a Responder Identifier TLV is being used, then the Echo Request packet MAY reuse a received Downstream Detailed Mapping TLV.

4. Non-compliant Routers

If an egress for a P2MP LSP does not support MPLS LSP ping, then no reply will be sent, resulting in a "false negative" result. There is no protection for this situation, and operators may wish to ensure that end points for P2MP LSPs are all equally capable of supporting this function. Alternatively, the traceroute option can be used to verify the LSP nearly all the way to the egress, leaving the final hop to be verified manually.

If, in "traceroute" mode, a transit node does not support LSP ping, then no reply will be forthcoming from that node for some TTL, say n. The node originating the echo request SHOULD continue to send echo request with TTL=n+1, n+2, ..., n+k to probe nodes further down the path. In such a case, the echo request for TTL > n SHOULD be sent with Downstream Detailed Mapping TLV "Downstream IP Address" field set to the ALLROUTERs multicast address as described in Section 3.3.4 until a reply is received with a Downstream Detailed Mapping TLV.

5. OAM 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 must be made.

- 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 responses 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, it SHOULD provide the ability to limit the scope of an LSP Ping echo request for a P2MP MPLS LSP to a single egress.

  Such an interface 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.

6. IANA Considerations

6.1. New Sub-TLV Types

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

IANA is requested to assign sub-type values to the following sub-TLVs from the "Multiprotocol Label Switching Architecture (MPLS) Label Switched Paths (LSPs) Parameters - TLVs" registry, "TLVs and sub-TLVs" sub-registry.

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

6.2. New TLVs

Two new LSP Ping TLV types are defined for inclusion in LSP Ping

messages.

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

P2MP Responder Identifier TLV (see Section 3.2.4) is a mandatory TLV. Suggested value 11. Four sub-TLVs are defined.
- Type 1: IPv4 Egress Address P2MP Responder Identifier
- Type 2: IPv6 Egress Address P2MP Responder Identifier
- Type 3: IPv4 Node Address P2MP Responder Identifier
- Type 4: IPv6 Node Address P2MP Responder Identifier

Echo Jitter TLV (see Section 3.2.5) is a mandatory TLV. Suggested value 12.

7. 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 to regulate 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. Note that this rate limiting might lead to false positives.

8. 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 review, and to Yakov Rekhter for useful discussions.

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

9. References

9.1 Normative References

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

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

[DDMT] Bahadur, N., Kompella, K., and Swallow, G., "Mechanism for Performing LSP-Ping over MPLS Tunnels", draft-ietf-mpls-lsp-ping-enhanced-dsmap, work in progress.

9.2 Informative References

[RFC792] Postel, J., "Internet Control Message Protocol", RFC 792.

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

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

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

[P2MP-LDP-REQ] J.-L. Le Roux, et al., "Requirements for point-to-multipoint extensions to the Label Distribution Protocol", draft-ietf-mpls-mp-ldp-reqs, work in progress.

[P2MP-LDP] Minei, I., and Wijnands, I., "Label Distribution Protocol

Extensions for Point-to-Multipoint and
Multipoint-to-Multipoint Label Switched Paths",
draft-ietf-mpls-ldp-p2mp, work in progress.

[MCAST-CV] Swallow, G., and Nadeau, T., "Connectivity Verification for Multicast Label Switched Paths", draft-swallow-mpls-mcast-cv, work in progress.

[BFD] Katz, D., and Ward, D., "Bidirectional Forwarding Detection", draft-ietf-bfd-base, work in progress.

[MPLS-BFD] Aggarwal, R., Kompella, K., Nadeau, T., and Swallow, G., "BFD For MPLS LSPs", draft-ietf-bfd-mpls, work in progress.

[IANA-PORT] IANA Assigned Port Numbers, http://www.iana.org

10. Authors' Addresses

Seisho Yasukawa
NTT Corporation
(R&D Strategy Department)
3-1, Otemachi 2-Chome Chiyodaku, Tokyo 100-8116 Japan
Phone: +81 3 5205 5341
Email: s.yasukawa@hco.ntt.co.jp

Adrian Farrel
Old Dog Consulting
EMail: adrian@olddog.co.uk

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

Bill Fenner
Arastra, Inc.
275 Middlefield Rd.
Suite 50
Menlo Park, CA 94025
Email: fenner@fenron.com

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

British Telecom
BT Centre
81 Newgate Street
EC1A 7AJ
London
Email: tom.nadeau@bt.com

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

11. Full Copyright Statement