Internet Engineering Task Force
Internet-Draft
Updates: 4379,6424 (if approved)
Intended status: Standards Track
Expires: January 24, 2016

N. Akiya
Big Switch Networks
G. Swallow
Cisco Systems
S. Litkowski
B. Decraene
Orange
J. Drake
Juniper Networks
July 23, 2015

Label Switched Path (LSP) Ping/Trace Multipath Support for Link Aggregation Group (LAG) Interfaces draft-ietf-mpls-lsp-ping-lag-multipath-01

Abstract

This document defines an extension to the MPLS Label Switched Path (LSP) Ping and Traceroute as specified in RFC 4379. The extension allows the MPLS LSP Ping and Traceroute to discover and exercise specific paths of Layer 2 (L2) Equal-Cost Multipath (ECMP) over Link Aggregation Group (LAG) interfaces.

This document updates RFC4379 and RFC6424.

Requirements Language

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

Status of This Memo

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

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

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

This Internet-Draft will expire on January 24, 2016.

Copyright Notice

Copyright (c) 2015 IETF Trust and the persons identified as the document authors. All rights reserved.

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

Table of Contents

1. Introduction

1.1. Terminology

The following acronyms/terms are used in this document:

- MPLS - Multiprotocol Label Switching.
- LSP - Label Switched Path.
- LSR - Label Switching Router.
- ECMP - Equal-Cost Multipath.
- LAG - Link Aggregation Group.
- Initiator LSR - LSR which sends MPLS echo request.
- Responder LSR - LSR which receives MPLS echo request and sends MPLS echo reply.

1.2. Background

The MPLS Label Switched Path (LSP) Ping and Traceroute as specified in [RFC4379] are powerful tools designed to diagnose all available layer 3 (L3) paths of LSPs, i.e., provides diagnostic coverage of L3 Equal-Cost Multipath (ECMP). In many MPLS networks, Link Aggregation Group (LAG) as defined in [IEEE802.1AX], which provide Layer 2 (L2) ECMP, are often used for various reasons. MPLS LSP Ping and Traceroute tools were not designed to discover and exercise specific paths of L2 ECMP. The result raises a limitation for following scenario when LSP X traverses over LAG Y:

- Label switching of LSP X over one or more member links of LAG Y have succeeded.

- Label switching of LSP X over one or more member links of LAG Y have failed.

- MPLS echo request for LSP X over LAG Y is load balanced over a member link which is label switching successfully.

With the above scenario, MPLS LSP Ping and Traceroute will not be able to detect the label switching failure of problematic member link(s) of the LAG. In other words, lack of L2 ECMP diagnostic coverage can produce an outcome where MPLS LSP Ping and Traceroute can be blind to label switching failures over problematic LAG interface. It is, thus, desirable to extend the MPLS LSP Ping and Traceroute to have deterministic diagnostic coverage of LAG interfaces.

Creation of this document was motivated by issues encountered in live networks.

2. Overview

This document defines an extension to the MPLS LSP Ping and Traceroute to describe Multipath Information for LAG member links separately, thus allowing MPLS LSP Ping and Traceroute to discover and exercise specific paths of L2 ECMP over LAG interfaces. Reader is expected to be familiar with mechanics of the MPLS LSP Ping and Traceroute described in Section 3.3 of [RFC4379] and Downstream Detailed Mapping TLV (DDMAP) described in Section 3.3 of [RFC6424].

MPLS echo request carries a DDMAP and an optional TLV to indicate that separate load balancing information for each L2 nexthop over LAG is desired in MPLS echo reply. Responder LSR places the same optional TLV in the MPLS echo reply to provide acknowledgement back to the initiator. It also adds, for each downstream LAG member, a load balance information (i.e. multipath information and interface index). The following figure and the texts provides an example using an LDP network. However the problem and the mechanism is applicable to all types of LSPs which can traverse over LAG interfaces.

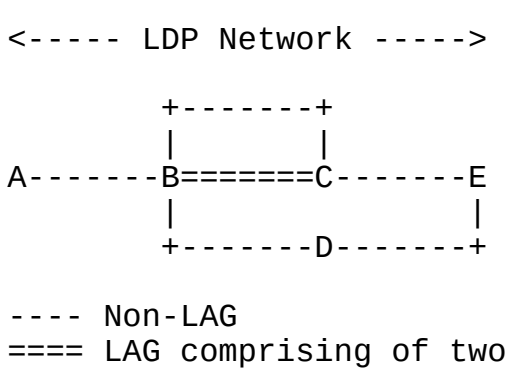


Figure 1: Example LDP Network

When node A is initiating LSP Traceroute to node E, node B will return to node A load balance information for following entries.

1. Downstream C over Non-LAG (upper path).

2. First Downstream C over LAG (middle path).

3. Second Downstream C over LAG (middle path).

4. Downstream D over Non-LAG (lower path).

This document defines:

- In Section 3, a mechanism discover capabilities of responder LSRs;
- In Section 4, a mechanism to discover L2 ECMP multipath information;
- In Section 5, a mechanism to validate L2 ECMP traversal in some LAG provisioning models;
- In Section 6, the LSR Capability TLV;
- In Section 7, the LAG Description Indicator flag;
- In Section 8, the Local Interface Index Sub-TLV;
- In Section 9, the Remote Interface Index Sub-TLV;
- In Section 10, the Detailed Interface and Label Stack TLV;
- In Appendix A, issues with LAG having an L2 Switch.

Note that the mechanism described in this document does not impose any changes to scenarios where an LSP is pinned down to a particular

LAG member (i.e. the LAG is not treated as one logical interface by the LSP).

Also note that many LAGs are built from p2p links, and thus router X and router X+1 have the same number of LAG members. It is possible to build LAGs asymmetrically by using Ethernet switches in the middle. Appendix A lists some cases which this document does not address; if an operator deploys LAGs in a manner similar to what's shown in Appendix A, the mechanisms in this document may not suit them.

3. LSR Capability Discovery

The MPLS Ping operates by an initiator LSR sending an MPLS echo request message and receiving back a corresponding MPLS echo reply message from a responder LSR. The MPLS Traceroute operates in a similar way except the initiator LSR potentially sends multiple MPLS echo request messages with incrementing TTL values.

There has been many extensions to the MPLS Ping and Traceroute mechanism over the years. Thus it is often useful, and sometimes necessary, for the initiator LSR to deterministically disambiguate the difference between:

- The responder LSR sent the MPLS echo reply message with contents C because it has feature X, Y and Z implemented.

- The responder LSR sent the MPLS echo reply message with contents C because it has subset of features X, Y and Z implemented but not all.

- The responder LSR sent the MPLS echo reply message with contents C because it does not have features X, Y and Z implemented.

To allow the initiator LSR to disambiguate the above differences, this document defines the LSR Capability TLV (described in Section 6). When the initiator LSR wishes to discover the capabilities of the responder LSR, the initiator LSR includes the LSR Capability TLV in the MPLS echo request message. When the responder LSR receives an MPLS echo reply message with the LSR Capability TLV included, then the responder LSR MUST include the LSR Capability TLV in the MPLS echo reply message with the LSR Capability TLV describing features and extensions supported by the local LSR.

It is RECOMMENDED that implementations supporting the LAG Multipath extensions defined in this document include the LSR Capability TLV in MPLS echo request messages.

4. Mechanism to Discover L2 ECMP Multipath

4.1. Initiator LSR Procedures

The MPLS echo request carries a DDMAP with the "LAG Description Indicator flag" (G) set in the DS Flags to indicate that separate load balancing information for each L2 nexthop over LAG is desired in MPLS echo reply. The new "LAG Description Indicator flag" is described in Section 7.

4.2. Responder LSR Procedures

This section describes the handling of the new TLVs by nodes which understand the "LAG Description Indicator flag". There are two cases - nodes which understand the "LAG Description Indicator flag" but which for some reason cannot describe LAG members separately, and nodes which both understand the "LAG Description Indicator flag" and are able to describe LAG members separately. Note that Section 6, Section 8 and Section 9 describe the new TLVs referenced by this section , and looking over the definition of the new TLVs first may make it easier to read this section.

A responder LSR that understand the "LAG Description Indicator flag" but is not capable of describing outgoing LAG member links separately uses the following procedures:

- If the received MPLS echo request message had the LSR Capability TLV, the responder LSR MUST include the LSR Capability TLV in the MPLS echo reply message.

- The responder LSR MUST clear the "Downstream LAG Info Accommodation flag" in the LSR Capability Flags field of the LSR Capability TLV. This will allow the initiator LSR to understand that the responder LSR cannot describe outgoing LAG member links separately in the DDMAP.

A responder LSR that understands the "LAG Description Indicator flag" and is capable of describing outgoing LAG member links separately uses the follow procedures, regardless of whether or not outgoing interfaces include LAG interfaces:

- If the received MPLS echo request message had the LSR Capability TLV, the responder LSR MUST include the LSR Capability TLV in the MPLS echo reply message.

- The responder LSR MUST set the "Downstream LAG Info Accommodation flag" in the LSR Capability Flags field of the LSR Capability TLV.

- For each downstream that is a LAG interface:
  - The responder LSR MUST add DDMAP in the MPLS echo reply.
  - The responder LSR MUST set the "LAG Description Indicator flag" in the DS Flags field of the DDMAP.
  - In the DDMAP, Local Interface Index Sub-TLV, Remote Interface Index Sub-TLV and Multipath Data Sub-TLV are to describe each LAG member link. All other fields of the DDMAP are to describe the LAG interface.
  - For each LAG member link of this LAG interface:
    - The responder LSR MUST add a Local Interface Index Sub-TLV (described in Section 8) with the "LAG Member Link Indicator flag" set in the Interface Index Flags field, describing the interface index of this outgoing LAG member link (the local interface index is assigned by the local LSR).
    - The responder LSR MAY add a Remote Interface Index Sub-TLV (described in Section 9) with the "LAG Member Link Indicator flag" set in the Interface Index Flags field, describing the interface index of the incoming LAG member link on the downstream LSR (this interface index is assigned by the downstream LSR). How the local LSR obtains the interface index of the LAG member link on the downstream LSR is outside the scope of this document.
    - The responder LSR MUST add an Multipath Data Sub-TLV for this LAG member link, if received DDMAP requested multipath information.

Based on the procedures described above, every LAG member link will have a Local Interface Index Sub-TLV and a Multipath Data Sub-TLV entries in the DDMAP. The order of the Sub-TLVs in the DDMAP for a LAG member link MUST be Local Interface Index Sub-TLV immediately followed by Multipath Data Sub-TLV. A LAG member link may also have a corresponding Remote Interface Index Sub-TLV. When a Local Interface Index Sub-TLV, a Remote Interface Index-Sub-TLV and a Multipath Data Sub-TLV are placed in the DDMAP to describe a LAG member link, they MUST be placed in the order of Local Interface Index Sub-TLV, Remote Interface Index-Sub-TLV and Multipath Data Sub-TLV.

A responder LSR possessing a LAG interface with two member links would send the following DDMAP for this LAG interface:

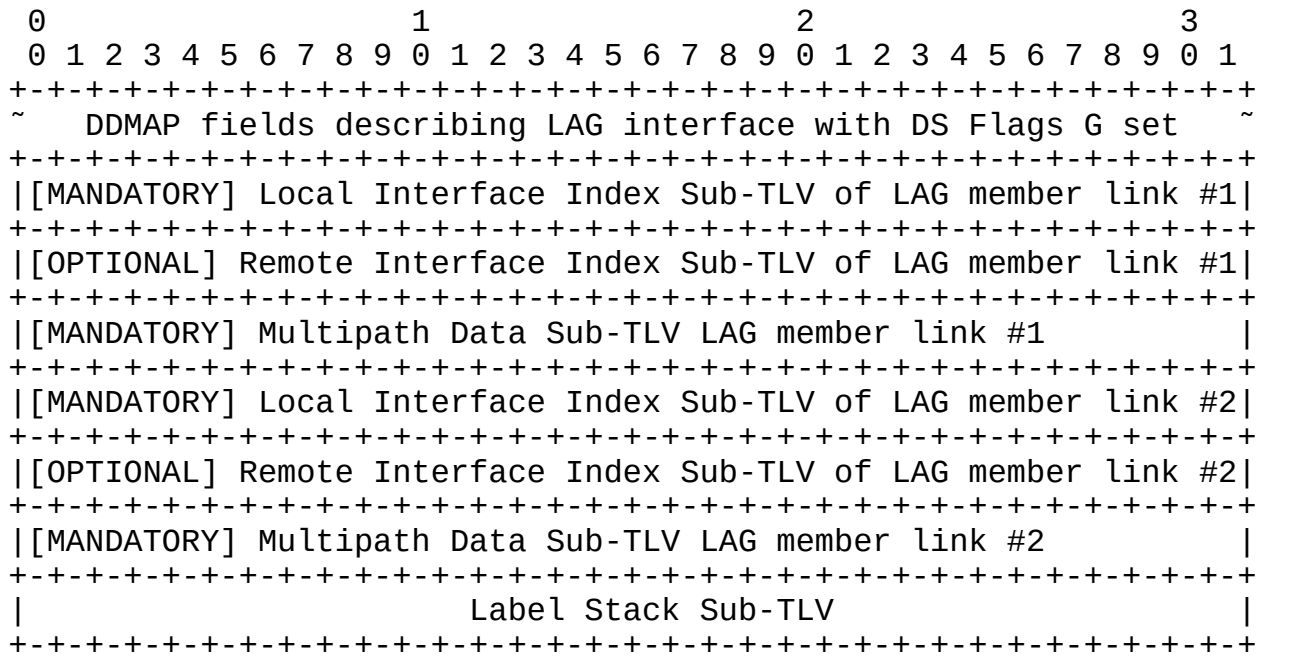


Figure 2: Example of DDMAP in MPLS Echo Reply

When none of the received multipath information maps to a particular LAG member link, then the responder LSR MUST still place the Local Interface Index Sub-TLV and the Multipath Data Sub-TLV for that LAG member link in the DDMAP, with the Multipath Length field of the Multipath Data Sub-TLV being zero.

4.3. Additional Initiator LSR Procedures

The procedures above allow an initiator LSR to:

- Identify whether or not the responder LSR can describe outgoing LAG member links separately, by looking at the LSR Capability TLV.

- Utilize the value of the "LAG Description Indicator flag" in DS Flags to identify whether each received DDMAP describes a LAG interface or a non-LAG interface.

- Obtain multipath information which is expected to traverse the specific LAG member link described by corresponding interface index.

When an initiator LSR receives a DDMAP containing LAG member information from a downstream LSR with TTL=n, then the subsequent DDMAP sent by the initiator LSR to the downstream LSR with TTL=n+1 through a particular LAG member link MUST be updated with following procedures:

- The Local Interface Index Sub-TLVs MUST be removed in the sending DDMAP.

- If the Remote Interface Index Sub-TLVs were present and the initiator LSR is traversing over a specific LAG member link, then the Remote Interface Index Sub-TLV corresponding to the LAG member link being traversed SHOULD be included in the sending DDMAP. All other Remote Interface Index Sub-TLVs MUST be removed from the sending DDMAP.

- The Multipath Data Sub-TLVs MUST be updated to include just one Multipath Data Sub-TLV. The initiator MAY keep just the Multipath Data Sub-TLV corresponding to the LAG member link being traversed, or combine the Multipath Data Sub-TLVs for all LAG member links into a single Multipath Data Sub-TLV when diagnosing further downstream LSRs.

- All other fields of the DDMAP are to comply with procedures described in [RFC6424].

Using the DDMAP example described in the Figure 2, the DDMAP being sent by the initiator LSR through LAG member link #1 to the next downstream LSR should be:

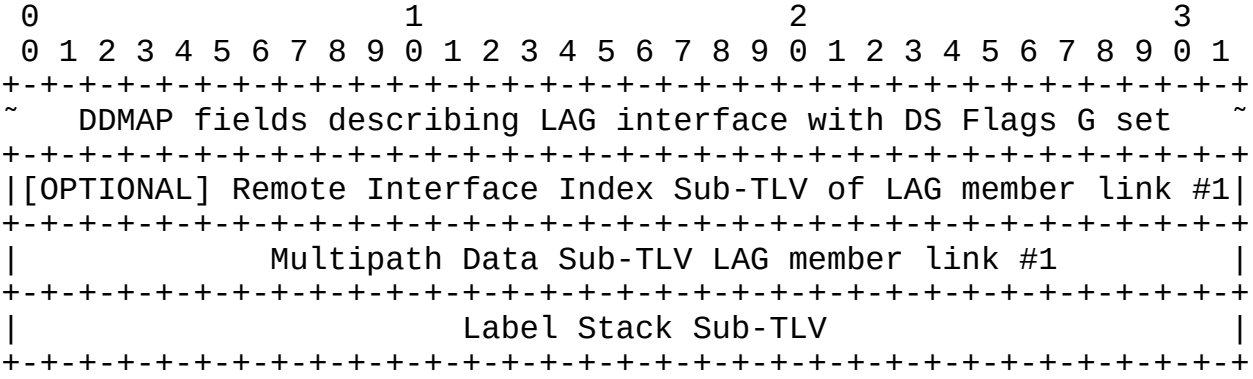


Figure 3: Example of DDMAP in MPLS Echo Request

5. Mechanism to Validate L2 ECMP Traversal

Section 4 defines the responder LSR procedures to constructs a DDMAP for a downstream LAG, and also defines that inclusion of the Remote Interface Index Sub-TLVs describing the incoming LAG member links of the downstream LSR is optional. The reason why it is optional for the responder LSR to include the Remote Interface Index Sub-TLVs is that this information from the downstream LSR is often not available on the responder LSR. In such case, the traversal of LAG member links can be validated with procedures described in Section 5.1. If

LSRs can provide the Remote Interface Index Sub-TLVs in DDMAP objects, then the validation procedures described in Section 5.2 can be used.

5.1. Incoming LAG Member Links Verification

Without downstream LSRs returning remote Interface Index Sub-TLVs in the DDMAP, validation of the LAG member link traversal requires that initiator LSR traverses all available LAG member links and taking the results through a logic. This section provides the mechanism for the initiator LSR to obtain additional information from the downstream LSRs and describes the additional logic in the initiator LSR to validate the L2 ECMP traversal.

5.1.1. Initiator LSR Procedures

The MPLS echo request is sent with a DDMAP with the "Interface and Label Stack Object Request flag" and "LAG Description Indicator flag" set in the DS Flags to indicate the request for Detailed Interface and Label Stack TLV with additional LAG member link information (i.e. interface index) in the MPLS echo reply.

5.1.2. Responder LSR Procedures

A responder LSR that understands the "LAG Description Indicator flag" but is not capable of describing incoming LAG member link is to use following procedures:

- If the received MPLS echo request message had the LSR Capability TLV, the responder LSR MUST include the LSR Capability TLV in the MPLS echo reply message.

- The responder LSR MUST clear the "Upstream LAG Info Accommodation flag" in the LSR Capability Flags field of the LSR Capability TLV. This will allow the initiator LSR to understand that the responder LSR cannot describe incoming LAG member link.

A responder LSR that understands the "LAG Description Indicator flag" and is capable of describing incoming LAG member link MUST use the following procedures, regardless of whether or not incoming interface was a LAG interface:

- If the received MPLS echo request message had the LSR Capability TLV, the responder LSR MUST include the LSR Capability TLV in the MPLS echo reply message.

- The responder LSR MUST set the "Upstream LAG Info Accommodation flag" in the LSR Capability Flags field of the LSR Capability TLV.

- When the received DDMAP had "Interface and Label Stack Object Request flag" set in the DS Flags field, the responder LSR MUST add the Detailed Interface and Label Stack TLV (described in Section 10) in the MPLS echo reply.

- When the received DDMAP had "Interface and Label Stack Object Request flag" set in the DS Flags field and the incoming interface was a LAG, the responder LSR MUST add the Incoming Interface Index Sub-TLV (described in Section 10.1.2) in the Detailed Interface and Label Stack TLV. The "LAG Member Link Indicator flag" MUST be set in the Interface Index Flags field, and the Interface Index field set to the LAG member link which received the MPLS echo request.

These procedures allow initiator LSR to:

- Identify whether or not the responder LSR can describe the incoming LAG member link, by looking at the LSR Capability TLV.

- Utilize the Incoming Interface Index Sub-TLV in the Detailed Interface and Label Stack TLV to identify, if the incoming interface was a LAG, the identity of the incoming LAG member.

5.1.3. Additional Initiator LSR Procedures

Along with procedures described in Section 4, the procedures described in this section will allow an initiator LSR to know:

- The expected load balance information of every LAG member link, at LSR with TTL=n.

- With specific entropy, the expected interface index of the outgoing LAG member link at TTL=n.

- With specific entropy, the interface index of the incoming LAG member link at TTL=n+1.

Expectation is that there’s a relationship between the interface index of the outgoing LAG member link at TTL=n and the interface index of the incoming LAG member link at TTL=n+1 for all discovered entropies. In other words, set of entropies that load balances to outgoing LAG member link X at TTL=n should all reach the nexthop on same incoming LAG member link Y at TTL=n+1.

With additional logics, the initiator LSR can perform following checks in a scenario where the initiator knows that there is a LAG, with two LAG members, between TTL=n and TTL=n+1, and has the multipath information to traverse the two LAG members.

The initiator LSR sends two MPLS echo request messages to traverse the two LAG members at TTL=1:

- Success case:
  - One MPLS echo request message reaches TTL=n+1 on an LAG member 1.
  - The other MPLS echo request message reaches TTL=n+1 on an LAG member 2.

  The two MPLS echo request messages sent by the initiator LSR reach two different LAG members at the immediate downstream LSR.

- Error case:
  - One MPLS echo request message reaches TTL=n+1 on an LAG member 1.
  - The other MPLS echo request message also reaches TTL=n+1 on an LAG member 1.

  One or two MPLS echo request messages sent by the initiator LSR does not reach the immediate downstream LSR, or the two MPLS echo request messages reach a same LAG member at the immediate downstream LSR.

Note that defined procedures will provide a deterministic result for LAG interfaces that are back-to-back connected between routers (i.e. no L2 switch in between). If there is a L2 switch between LSR at TTL=n and LSR at TTL=n+1, there is no guarantee that traversal of every LAG member link at TTL=n will result in reaching different interface index at TTL=n+1. Issues resulting from LAG with L2 switch in between are further described in Appendix A. LAG provisioning models in operated network should be considered when analyzing the output of LSP Traceroute exercising L2 ECMPs.

5.2. Individual End-to-End Path Verification

When the Remote Interface Index Sub-TLVs are available from an LSR with TTL=n, then the validation of LAG member link traversal can be performed by the downstream LSR of TTL=n+1. The initiator LSR follows the procedures described in Section 4.3.

The DDMAP validation procedures by the downstream responder LSR are then updated to include the comparison of the incoming LAG member link (which MPLS echo request was received on) to the interface index described in the Remote Interface Index Sub-TLV in the DDMAP.

Failure of this comparison results in the return code being set to "Downstream Mapping Mismatch (5)".

A responder LSR that is not able to perform the above additional DDMAP validation procedures is considered to lack the upstream LAG capability. Thus, if the received MPLS echo request contained the LSR Capability TLV, then the responder LSR MUST include the LSR Capability TLV in the MPLS echo reply and the LSR Capability TLV MUST have the "Upstream LAG Info Accomodation flag" cleared.

6. LSR Capability TLV

The LSR Capability object is a new TLV that MAY be included in the MPLS echo request message and the MPLS echo reply message. An MPLS echo request message and an MPLS echo reply message MUST NOT include more than one LSR Capability object. Presence of an LSR Capability object in an MPLS echo request message is a request that a responder LSR includes an LSR Capability object in the MPLS echo reply message, with the LSR Capability object describing features and extensions supported. When the received MPLS echo request message contains an LSR Capability object, an responder LSR MUST include the LSR Capability object in the MPLS echo reply message.

LSR Capability TLV Type is TBD1. Length is 4. The value field of the LSR Capability TLV has following format:

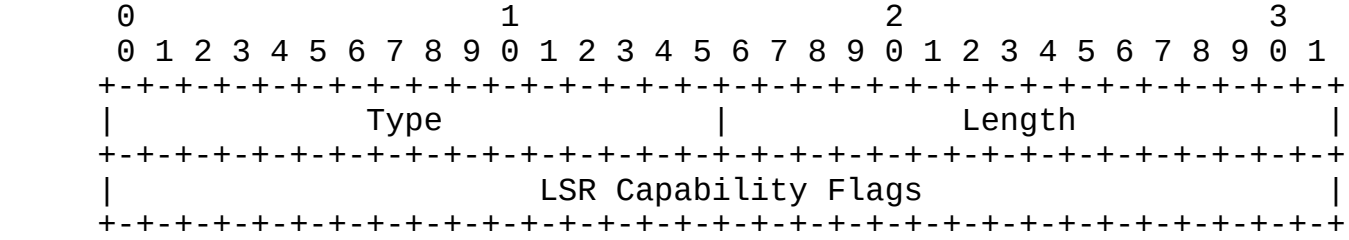
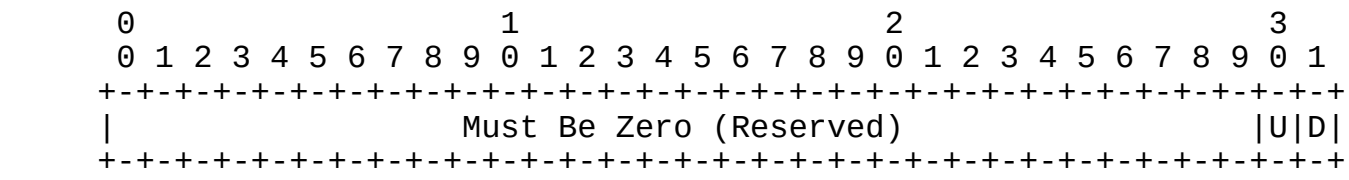


Figure 4: LSR Capability TLV

LSR Capability Flags

The LSR Capability Flags field is a bit vector with following format:



Two flags are defined: U and D. The remaining flags MUST be set to zero when sending and ignored on receipt. Both U and D flags MUST be cleared in MPLS echo request message when sending, and ignored on receipt. Neither, either or both U and D flags MAY be set in MPLS echo reply message.

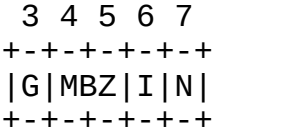
Flag	Name and Meaning
U	Upstream LAG Info Accommodation
	An LSR sets this flag when the node is capable of describing a LAG member link in the Incoming Interface Index Sub-TLV in the in the Detailed Interface and Label Stack TLV.
D	Downstream LAG Info Accommodation
	An LSR sets this flag when the node is capable of describing LAG member links in the Local Interface Index Sub-TLV and the Multipath Data Sub-TLV in the Downstream Detailed Mapping TLV.

7. LAG Description Indicator Flag: G

One flag, G, is added in DS Flags field of the DDMAP TLV. The G flag of the DS Flags field in the MPLS echo request message indicates the request for detailed LAG information from the responder LSR. In the MPLS echo reply message, the G flag MUST be set if the DDMAP TLV describes a LAG interface. It MUST be cleared otherwise.

DS Flags

DS Flags G is added, in Bit Number TBD5, in DS Flags bit vector.



RFC-Editor-Note: Please update above figure to place the flag G in the bit number TBD5.

Flag	Name and Meaning
G	LAG Description Indicator

When this flag is set in the MPLS echo request, responder is requested to respond with detailed LAG information. When this flag is set in the MPLS echo reply, the corresponding DDMAP describes a LAG interface.

8. Local Interface Index Sub-TLV

The Local Interface Index object is a Sub-TLV that MAY be included in a DDMAP TLV. Zero or more Local Interface Index object MAY appear in a DDMAP TLV. The Local Interface Index Sub-TLV describes the index assigned by the local LSR to the egress interface.

The Local Interface Index Sub-TLV Type is TBD2. Length is 8, and the Value field has following format:

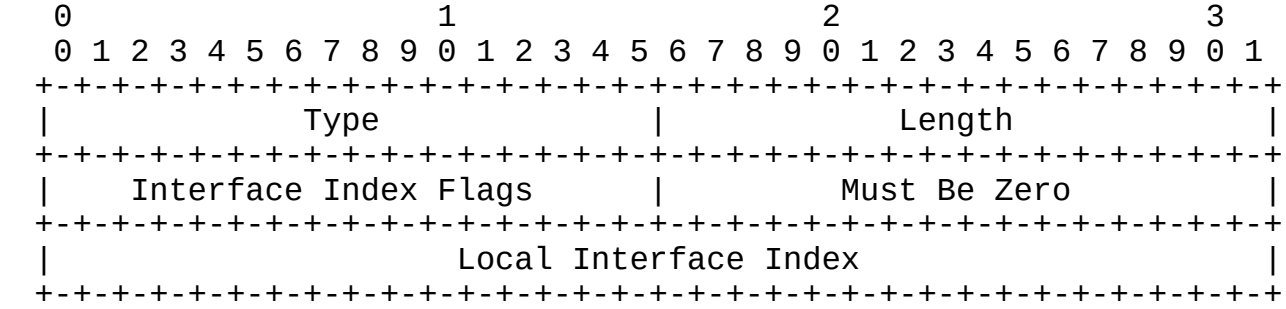
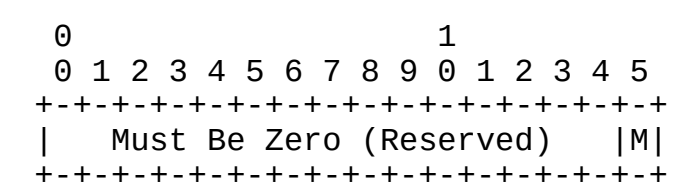


Figure 5: Local Interface Index Sub-TLV

Interface Index Flags

Interface Index Flags field is a bit vector with following format.



One flag is defined: M. The remaining flags MUST be set to zero when sending and ignored on receipt.

Flag	Name and Meaning
M	LAG Member Link Indicator

When this flag is set, interface index described in this sub-TLV is a member of a LAG.

Local Interface Index

An Index assigned by the LSR to this interface.

9. Remote Interface Index Sub-TLV

The Remote Interface Index object is a Sub-TLV that MAY be included in a DDMAP TLV. Zero or more Remote Interface Index object MAY appear in a DDMAP TLV. The Remote Interface Index Sub-TLV describes the index assigned by the downstream LSR to the ingress interface.

The Remote Interface Index Sub-TLV Type is TBD3. Length is 8, and the Value field has following format:

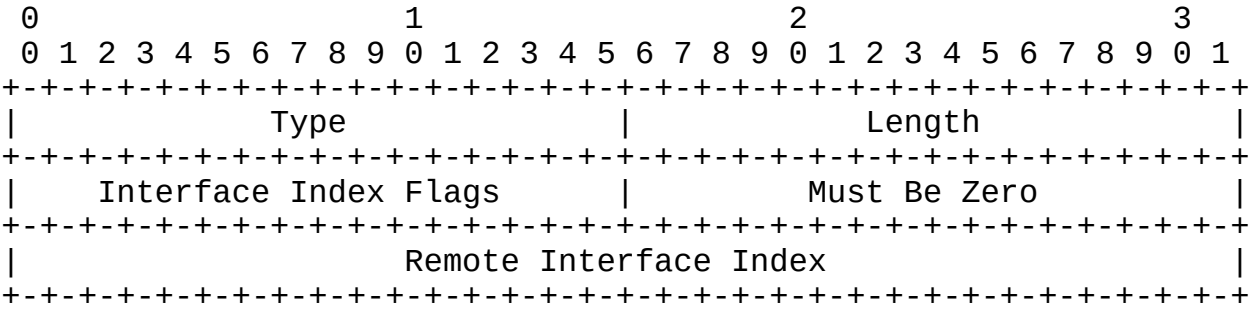
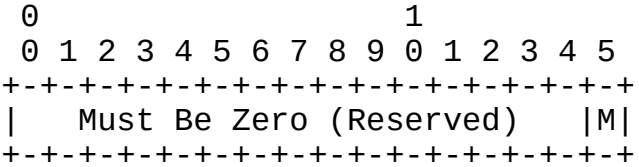


Figure 6: Remote Interface Index Sub-TLV

Interface Index Flags

Interface Index Flags field is a bit vector with following format.



One flag is defined: M. The remaining flags MUST be set to zero when sending and ignored on receipt.

Flag	Name and Meaning
M	LAG Member Link Indicator When this flag is set, interface index described in this sub-TLV is a member of a LAG.

Remote Interface Index

An Index assigned by the downstream LSR to the ingress interface.

10. Detailed Interface and Label Stack TLV

The "Detailed Interface and Label Stack" object is a TLV that MAY be included in a MPLS echo reply message to report the interface on which the MPLS echo request message was received and the label stack that was on the packet when it was received. A responder LSR MUST NOT insert more than one instance of this TLV. This TLV allows the initiator LSR to obtain the exact interface and label stack information as it appears at the responder LSR.

Detailed Interface and Label Stack TLV Type is TBD4. Length is K + Sub-TLV Length (sum of Sub-TLVs). K is the sum of all fields of this TLV prior to Sub-TLVs, but the length of K depends on the Address Type. Details of this information is described below. The Value field has following format:

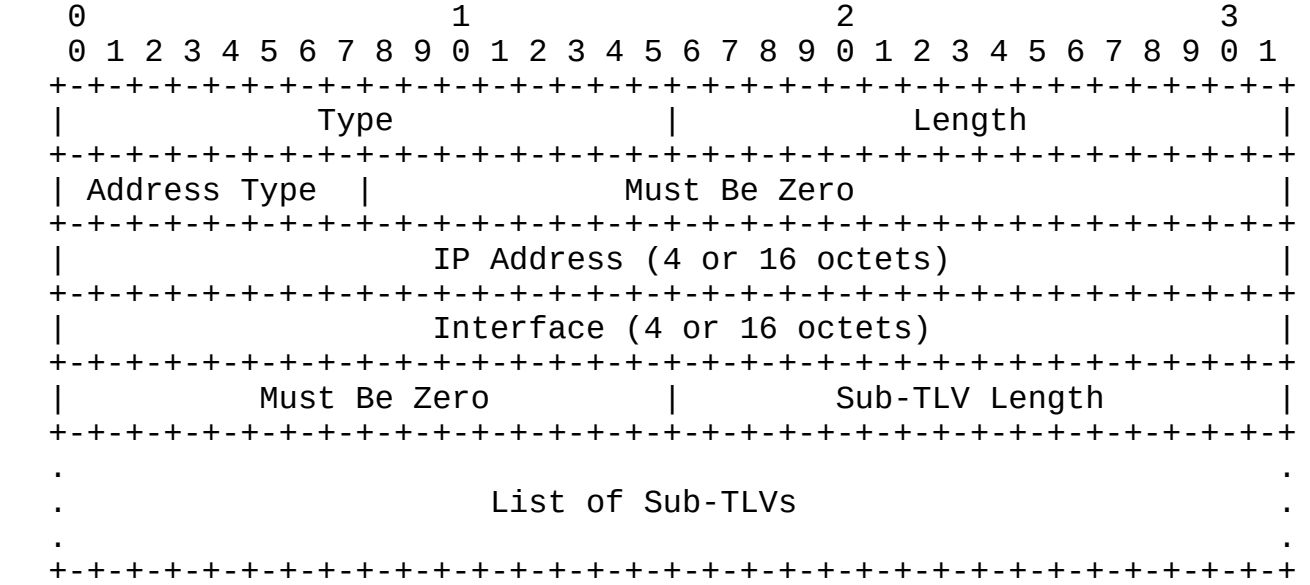


Figure 7: Detailed Interface and Label Stack TLV

The Detailed Interface and Label Stack TLV format is derived from the Interface and Label Stack TLV format (from [RFC4379]). Two changes are introduced. First is that label stack, which is of variable length, is converted into a sub-TLV. Second is that a new sub-TLV is added to describe an interface index. The fields of Detailed Interface and Label Stack TLV have the same use and meaning as in [RFC4379]. A summary of the fields taken from the Interface and Label Stack TLV is as below:

Address Type

The Address Type indicates if the interface is numbered or unnumbered. It also determines the length of the IP Address and Interface fields. The resulting total for the initial part of the TLV is listed in the table below as "K Octets". The Address Type is set to one of the following values:

Type #	Address Type	K Octets
1	IPv4 Numbered	16
2	IPv4 Unnumbered	16
3	IPv6 Numbered	40
4	IPv6 Unnumbered	28

IP Address and Interface

IPv4 addresses and interface indices are encoded in 4 octets; IPv6 addresses are encoded in 16 octets.

If the interface upon which the echo request message was received is numbered, then the Address Type MUST be set to IPv4 Numbered or IPv6 Numbered, the IP Address MUST be set to either the LSR’s Router ID or the interface address, and the Interface MUST be set to the interface address.

If the interface is unnumbered, the Address Type MUST be either IPv4 Unnumbered or IPv6 Unnumbered, the IP Address MUST be the LSR’s Router ID, and the Interface MUST be set to the index assigned to the interface.

Note: Usage of IPv6 Unnumbered has the same issue as [RFC4379], described in Section 3.4.2 of [I-D.ietf-mpls-ipv6-only-gap]. A solution should be considered an applied to both [RFC4379] and this document.

Sub-TLV Length

Total length in octets of the sub-TLVs associated with this TLV.

10.1. Sub-TLVs

This section defines the sub-TLVs that MAY be included as part of the Detailed Interface and Label Stack TLV.

Sub-Type	Value Field
1	Incoming Label stack
2	Incoming Interface Index

10.1.1. Incoming Label Stack Sub-TLV

The Incoming Label Stack sub-TLV contains the label stack as received by the LSR. If any TTL values have been changed by this LSR, they SHOULD be restored.

Incoming Label Stack Sub-TLV Type is 1. Length is variable, and the Value field has following format:

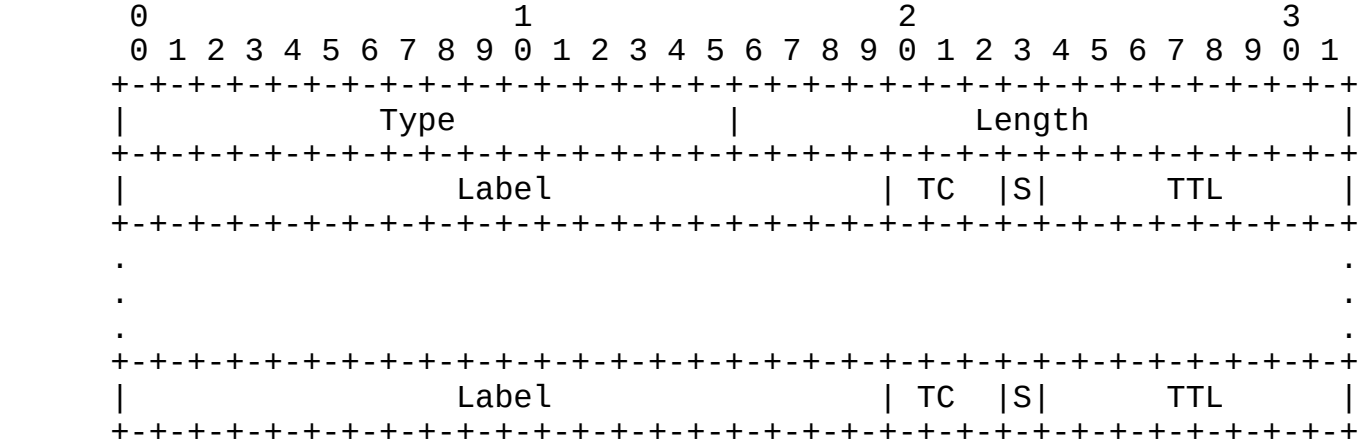


Figure 8: Incoming Label Stack Sub-TLV

10.1.2. Incoming Interface Index Sub-TLV

The Incoming Interface Index object is a Sub-TLV that MAY be included in a Detailed Interface and Label Stack TLV. The Incoming Interface Index Sub-TLV describes the index assigned by this LSR to the interface which received the MPLS echo request message.

Incoming Interface Index Sub-TLV Type is 2. Length is 8, and the Value field has the same format as the Local Interface Index Sub-TLV described in Section 8, and has following format:

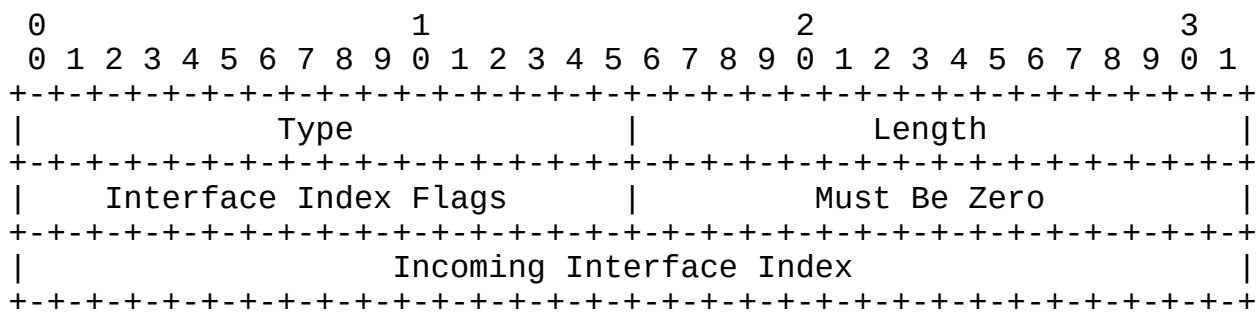
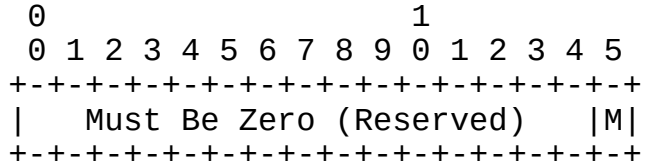


Figure 9: Incoming Interface Index Sub-TLV

Interface Index Flags

Interface Index Flags field is a bit vector with following format.



One flag is defined: M. The remaining flags MUST be set to zero when sending and ignored on receipt.

Flag	Name and Meaning
M	LAG Member Link Indicator

When this flag is set, interface index described in this sub-TLV is a member of a LAG.

Incoming Interface Index

An Index assigned by the LSR to this interface.

11. Security Considerations

This document extends LSP Traceroute mechanism to discover and exercise L2 ECMP paths. As a result of supporting the code points and procedures described in this document, additional processing are required by initiator LSRs and responder LSRs, especially to compute and handle increasing number of multipath information. Due to additional processing, it is critical that proper security measures described in [RFC4379] and [RFC6424] are followed.

The LSP Traceroute allows an initiator LSR to discover the paths of tested LSPs, providing deep knowledge of the MPLS network. Exposing such information to a malicious user is considered dangerous. To prevent leakage of vital information to untrusted users, a responder LSR MUST only accept MPLS echo request messages from trusted sources via filtering source IP address field of received MPLS echo request messages.

12. IANA Considerations

12.1. LSR Capability TLV

The IANA is requested to assign new value TBD1 for LSR Capability TLV from the "Multiprotocol Label Switching Architecture (MPLS) Label Switched Paths (LSPs) Ping Parameters - TLVs" registry.

Value	Meaning	Reference
TBD1	LSR Capability TLV	this document

12.1.1. LSR Capability Flags

The IANA is requested to create and maintain a registry entitled "LSR Capability Flags" with following registration procedures:

Registry Name: LAG Interface Info Flags

Bit number	Name	Reference
31	D: Downstream LAG Info Accommodation	this document
30	U: Upstream LAG Info Accommodation	this document
0-29	Unassigned	

Assignments of LSR Capability Flags are via Standards Action [RFC5226].

12.2. Local Interface Index Sub-TLV

The IANA is requested to assign new value TBD2 (from the range 4-31743) for the Local Interface Index Sub-TLV from the "Multiprotocol Label Switching Architecture (MPLS) Label Switched Paths (LSPs) Ping Parameters - TLVs" registry, "Sub-TLVs for TLV Types 20" sub-registry.

Value	Meaning	Reference
TBD2	Local Interface Index Sub-TLV	this document

12.2.1. Interface Index Flags

The IANA is requested to create and maintain a registry entitled "Interface Index Flags" with following registration procedures:

Registry Name: Interface Index Flags

Bit number	Name	Reference
15	M: LAG Member Link Indicator	this document
0-14	Unassigned	

Assignments of Interface Index Flags are via Standards Action [RFC5226].

Note that this registry is used by the Interface Index Flags field of following Sub-TLVs:

- The Local Interface Index Sub-TLV which may be present in the "Downstream Detailed Mapping" TLV.
- The Remote Interface Index Sub-TLV which may be present in the "Downstream Detailed Mapping" TLV.
- The Incoming Interface Index Sub-TLV which may be present in the "Detailed Interface and Label Stack" TLV.

12.3. Remote Interface Index Sub-TLV

The IANA is requested to assign new value TBD3 (from the range 32768-49161) for the Remote Interface Index Sub-TLV from the "Multiprotocol Label Switching Architecture (MPLS) Label Switched Paths (LSPs) Ping Parameters - TLVs" registry, "Sub-TLVs for TLV Types 20" sub-registry.

Value	Meaning	Reference
TBD3	Remote Interface Index Sub-TLV	this document

12.4. Detailed Interface and Label Stack TLV

The IANA is requested to assign new value TBD4 for Detailed Interface and Label Stack TLV from the "Multiprotocol Label Switching Architecture (MPLS) Label Switched Paths (LSPs) Ping Parameters - TLVs" registry ([IANA-MPLS-LSP-PING]).

Value	Meaning	Reference
TBD4	Detailed Interface and Label Stack TLV	this document

12.4.1. Sub-TLVs for TLV Type TBD4

The IANA is requested to create and maintain a sub-registry entitled "Sub-TLVs for TLV Type TBD4" under "Multiprotocol Label Switching Architecture (MPLS) Label Switched Paths (LSPs) Ping Parameters - TLVs" registry.

Initial values for this sub-registry, "Sub-TLVs for TLV Types TBD4", are described below.

Sub-Type	Name	Reference
1	Incoming Label Stack	this document
2	Incoming Interface Index	this document
3-16383	Unassigned (mandatory TLVs)	
16384-31743	Experimental	
32768-49161	Unassigned (optional TLVs)	
49162-64511	Experimental	

Assignments of Sub-Types in the mandatory and optional spaces are are via Standards Action [RFC5226]. Assignments of Sub-Types in the experimental space is via Specification Required [RFC5226].

12.5. DS Flags

The IANA is requested to assign a new bit number from the "DS flags" sub-registry from the "Multi-Protocol Label Switching (MPLS) Label Switched Paths (LSPs) Ping Parameters - TLVs" registry ([IANA-MPLS-LSP-PING]).

Note: the "DS flags" sub-registry is created by [RFC7537].

Bit number	Name	Reference
TBD5	G: LAG Description Indicator	this document

13. Acknowledgements

The authors would like to thank Nagendra Kumar and Sam Aldrin for providing useful comments and suggestions. The authors would like to thank Loa Andersson for performing a detailed review and providing number of comments.

The authors also would like to extend sincere thanks to the MPLS RT review members who took time to review and provide comments. The members are Eric Osborne, Mach Chen and Yimin Shen. The suggestion by Mach Chen to generalize and create the LSR Capability TLV was tremendously helpful for this document and likely for future documents extending the MPLS LSP Ping and Traceroute mechanism. The suggestion by Yimin Shen to create two separate validation procedures had a big impact to the contents of this document.

14. References

14.1. Normative References

[RFC2119] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, DOI 10.17487/RFC2119, March 1997, <http://www.rfc-editor.org/info/rfc2119>.

[RFC4379] Kompella, K. and G. Swallow, "Detecting Multi-Protocol Label Switched (MPLS) Data Plane Failures", RFC 4379, DOI 10.17487/RFC4379, February 2006, <http://www.rfc-editor.org/info/rfc4379>.

[RFC6424] Bahadur, N., Kompella, K., and G. Swallow, "Mechanism for Performing Label Switched Path Ping (LSP Ping) over MPLS Tunnels", RFC 6424, DOI 10.17487/RFC6424, November 2011, <http://www.rfc-editor.org/info/rfc6424>.

[RFC7537] Decraene, B., Akiya, N., Pignataro, C., Andersson, L., and S. Aldrin, "IANA Registries for LSP Ping Code Points", RFC 7537, DOI 10.17487/RFC7537, May 2015, <http://www.rfc-editor.org/info/rfc7537>.

14.2. Informative References

[I-D.ietf-mpls-ipv6-only-gap] George, W. and C. Pignataro, "Gap Analysis for Operating IPv6-only MPLS Networks", draft-ietf-mpls-ipv6-only-gap-04 (work in progress), November 2014.

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

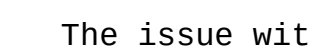
[IEEE802.1AX]
IEEE Std. 802.1AX, "IEEE Standard for Local and metropolitan area networks - Link Aggregation", November 2008.

[RFC5226] Narten, T. and H. Alvestrand, "Guidelines for Writing an IANA Considerations Section in RFCs", BCP 26, RFC 5226, DOI 10.17487/RFC5226, May 2008, <http://www.rfc-editor.org/info/rfc5226>.

Appendix A. LAG with L2 Switch Issues

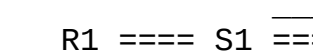
Several flavors of "LAG with L2 switch" provisioning models are described in this section, with MPLS data plane ECMP traversal validation issues with each.

A.1. Equal Numbers of LAG Members



The issue with this LAG provisioning model is that packets traversing a LAG member from R1 to S1 can get load balanced by S1 towards R2. Therefore, MPLS echo request messages traversing specific LAG member from R1 to S1 can actually reach R2 via any LAG members, and sender of MPLS echo request messages have no knowledge of this nor no way to control this traversal. In the worst case, MPLS echo request messages with specific entropies to exercise every LAG members from R1 to S1 can all reach R2 via same LAG member. Thus it is impossible for MPLS echo request sender to verify that packets intended to traverse specific LAG member from R1 to S1 did actually traverse that LAG member, and to deterministically exercise "receive" processing of every LAG member on R2.

A.2. Deviating Numbers of LAG Members



There are deviating number of LAG members on the two sides of the L2 switch. The issue with this LAG provisioning model is the same as previous model, sender of MPLS echo request messages have no knowledge of L2 load balance algorithm nor entropy values to control the traversal.

A.3. LAG Only on Right

```
R1 ---- S1 ==== R2
```

The issue with this LAG provisioning model is that there is no way for MPLS echo request sender to deterministically exercise both LAG members from S1 to R2. And without such, "receive" processing of R2 on each LAG member cannot be verified.

A.4. LAG Only on Left

```
R1 ==== S1 ---- R2
```

MPLS echo request sender has knowledge of how to traverse both LAG members from R1 to S1. However, both types of packets will terminate on the non-LAG interface at R2. It becomes impossible for MPLS echo request sender to know that MPLS echo request messages intended to traverse a specific LAG member from R1 to S1 did indeed traverse that LAG member.

Authors' Addresses

Nobo Akiya
Big Switch Networks

Email: nobo.akiya.dev@gmail.com

George Swallow
Cisco Systems

Email: swallow@cisco.com

Stephane Litkowski
Orange

Email: stephane.litkowski@orange.com

Bruno Decraene
Orange

Email: bruno.decraene@orange.com

John E. Drake
Juniper Networks

Email: jdrake@juniper.net