Label Switched Path (LSP) Ping/Trace Multipath Support for Link Aggregation Group (LAG) Interfaces
draft-ietf-mpls-lsp-ping-lag-multipath-05

Abstract

This document defines extensions to the MPLS Label Switched Path
(LSP) Ping and Traceroute mechanisms as specified in RFC 8029. The
extensions allow the MPLS LSP Ping and Traceroute mechanisms to
discover and exercise specific paths of Layer 2 (L2) Equal-Cost
Multipath (ECMP) over Link Aggregation Group (LAG) interfaces.
Additionally, a mechanism is defined to enable determination of the
capabilities of an LSR supported.

This document updates RFC8029.

Requirements Language

The key words "MUST", "MUST NOT", "REQUIRED", "SHALL", "SHALL NOT",
"SHOULD", "SHOULD NOT", "RECOMMENDED", "NOT RECOMMENDED", "MAY", and
"OPTIONAL" in this document are to be interpreted as described in
BCP14 [RFC2119][RFC8174] when, and only when, they appear in all
capitals, as shown here.

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 https://datatracker.ietf.org/drafts/current/.

Table of Contents

1. Introduction

1.1. Terminology

   The following acronyms/terms are used in this document:

   o MPLS - Multiprotocol Label Switching.

   o LSP - Label Switched Path.

   o LSR - Label Switching Router.

   o ECMP - Equal-Cost Multipath.

   o LAG - Link Aggregation Group.

   o Initiator LSR - LSR which sends MPLS echo request.

   o 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 mechanisms
   [RFC8029] are powerful tools designed to diagnose all available Layer
   3 (L3) paths of LSPs, including diagnostic coverage of L3 Equal-Cost

Multipath (ECMP). In many MPLS networks, Link Aggregation Group
(LAG) as defined in [IEEE802.1AX], which provides Layer 2 (L2) ECMP,
is often used for various reasons. MPLS LSP Ping and Traceroute
tools were not designed to discover and exercise specific paths of L2
ECMP. This raises a limitation for the following scenario when an
LSP traverses over a LAG:

o Label switching over some member links of the LAG is successful,
   but will be failed over other member links of the LAG.

o MPLS echo request for the LSP over the LAG is load balanced on one
   of the member links which is label switching successfully.

With the above scenarios, MPLS LSP Ping and Traceroute will not be
able to detect the label switching failure of the 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 a problematic LAG
interface. It is, thus, desirable to extend the MPLS LSP Ping and
Traceroute to have deterministic diagnostic coverage of LAG
interfaces.

The need for a solution of this problem was motivated by issues
encountered in live networks.

2. Overview of Solution

This document defines an optional TLV to discover the capabilities of
a responder LSR and extensions for use with the MPLS LSP Ping and
Traceroute mechanisms to describe Multipath Information for
individual LAG member links, thus allowing MPLS LSP Ping and
Traceroute to discover and exercise specific paths of L2 ECMP over
LAG interfaces. The reader is expected to be familiar with mechanics
of Downstream Mapping described in Section 3.3 of [RFC8029] and
Downstream Detailed Mapping TLV (DDMAP) described in Section 3.4 of
[RFC8029].

The solution consists of the MPLS echo request containing a DDMAP TLV
and the optional LSR capability TLV to indicate that separate load
balancing information for each L2 nexthop over LAG is desired in the
MPLS echo reply. The Responder LSR places the same optional LSR
capability TLV in the MPLS echo reply to provide acknowledgement back
to the initiator LSR. It also adds, for each downstream LAG member,
load balance information (i.e., multipath information and interface
index). This mechanism is applicable to all types of LSPs which can
traverse over LAG interfaces. Many LAGs are built from p2p links,
with router X and router X+1 having direct connectivity and the same
number of LAG members. It is possible to build LAGs asymmetrically

by using Ethernet switches between two routers. Appendix A lists
some use cases for which the mechanisms defined in this document may
not be applicable. Note that the mechanisms described in this
document do 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).

The following figure and description provides an example using an LDP
network.

```
    <----- LDP Network ----->


            +-------+
            |       |
    A-------B=======C-------E
            |               |
            +-------D-------+


    ---- Non-LAG
    ==== LAG comprising of two member links
```
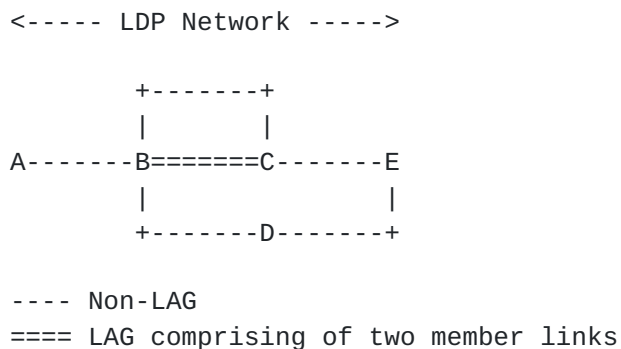
        Figure 1: Example LDP Network

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:

o In Section 3, a mechanism to discover capabilities of responder
   LSRs;

o In Section 4, a mechanism to discover L2 ECMP multipath
   information;

o In Section 5, a mechanism to validate L2 ECMP traversal;

o In Section 6, the LSR Capability TLV;

o In Section 7, the LAG Description Indicator flag;

o In Section 8, the Local Interface Index Sub-TLV;

o In Section 9, the Remote Interface Index Sub-TLV;

o In Section 10, the Detailed Interface and Label Stack TLV;

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 have 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 differences between:

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

   o 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.

   o 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 request message with the LSR Capability TLV
   included, if it knows the LSR Capability TLV, then it 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. Otherwise, an MPLS echo reply MUST be sent back to the
   initiator LSR with the return code set to "One or more of the TLVs
   was not understood". Then the initiator LSR can send another MPLS
   echo request without including the LSR Capability TLV.

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

3.1. Initiator LSR Procedures

If an initiator LSR does not know what capabilities a responder LSR
can support, it can send an MPLS each request message and carry the
LSR Capability TLV to the responder to discover the capabilities that
the responder LSR can support.

3.2. Responder LSR Procedures

When a responder LSR received an MPLS echo request message that
carries the LSR Capability TLV, the following procedures are used:

If the responder knows how to process the LSR Capability TLV, the
following procedures are used:

o The responder LSR MUST include the LSR Capability TLV in the MPLS
   echo reply message.

o If the responder LSR understands the "LAG Description Indicator
   flag":

   * Set the "Downstream LAG Info Accommodation flag" if the
      responder LSR is capable of describing outgoing LAG member
      links separately; otherwise, clear the "Downstream LAG Info
      Accommodation flag".

   * Set the "Upstream LAG Info Accommodation flag" if responder LSR
      is capable of describing incoming LAG member links separately;
      otherwise, clear the "Upstream LAG Info Accommodation flag".

o If the responder LSR does not understand the "LAG Description
   Indicator flag":

   * Clear both the "Downstream LAG Info Accommodation flag" and the
      "Upstream LAG Info Accommodation flag".

If the responder does not know the LSR Capability TLV, an MPLS echo
reply with the return code set to "One or more of the TLVs was not
understood" MUST be sent back to the initiator LSR.

4. Mechanism to Discover L2 ECMP Multipath

4.1. Initiator LSR Procedures

Through the "LSR Capability Discovery" as defined in Section 3, the
initiator LSR can understand whether the responder LSR can describe
incoming/outgoing LAG member links separately in the DDMAP TLV.

Once the initiator LSR knows the capabilities that a responder
supports, then it sends an MPLS echo request carrying a DDMAP with
the "LAG Description Indicator flag" (G) set to the responder LSR.
The "LAG Description Indicator flag" (G) indicates that separate load
balancing information for each L2 nexthop over a LAG is desired in
the MPLS echo reply. The new "LAG Description Indicator flag" is
described in Section 7.

4.2. Responder LSR Procedures

When a responder LSR received an MPLS echo request message with the
"LAG Description Indicator flag" set, if the responder LSR
understands the "LAG Description Indicator flag" and is capable of
describing outgoing LAG member links separately, the following
procedures are used, regardless of whether or not outgoing interfaces
include LAG interfaces:

o For each downstream that is a LAG interface:

   * The responder LSR MUST include a DDMAP TLV when sending the
      MPLS echo reply.

   * The responder LSR MUST set the "LAG Description Indicator flag"
      in the DS Flags field of the DDMAP TLV.

   * In the DDMAP TLV, the Local Interface Index Sub-TLV, Remote
      Interface Index Sub-TLV and Multipath Data Sub-TLV are used to
      describe each LAG member link. All other fields of the DDMAP
      TLV are used to describe the LAG interface.

   * For each LAG member link of the 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 the received DDMAP TLV 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 TLV. The order of the Sub-TLVs in the DDMAP TLV
   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 TLV 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:

```
    0                   1                   2                   3
    0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1
   +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
   ~   DDMAP fields describing LAG interface with DS Flags G set   ~
   +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
   |[MANDATORY] Local Interface Index Sub-TLV of LAG member link #1|
   +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
   |[OPTIONAL] Remote Interface Index Sub-TLV of LAG member link #1|
   +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
   |[MANDATORY] Multipath Data Sub-TLV LAG member link #1          |
   +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
   |[MANDATORY] Local Interface Index Sub-TLV of LAG member link #2|
   +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
   |[OPTIONAL] Remote Interface Index Sub-TLV of LAG member link #2|
   +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
   |[MANDATORY] Multipath Data Sub-TLV LAG member link #2          |
   +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
   |                      Label Stack Sub-TLV                      |
   +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
```

          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 TLV, the value of Multipath Length field of
   the Multipath Data Sub-TLV is set to zero.

4.3. Additional Initiator LSR Procedures

The procedures above allow an initiator LSR to:

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

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

o 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:

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

o 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.

o The Multipath Data Sub-TLVs MUST be updated to include just one
   Multipath Data Sub-TLV. The initiator LSR MAY just keep 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.

o All other fields of the DDMAP are to comply with procedures
   described in [RFC8029].

Figure 3 is an example that shows how to use the DDMAP TLV to notify
which member link (link #1 in the example) will be chosen to send the
MPLS echo request message to the next downstream LSR:

```
 0                   1                   2                   3
 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
~   DDMAP fields describing LAG interface with DS Flags G set   ~
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|[OPTIONAL] Remote Interface Index Sub-TLV of LAG member link #1|
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|              Multipath Data Sub-TLV LAG member link #1        |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|                        Label Stack Sub-TLV                    |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
```

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. The Remote Interface Index Sub-TLVs that
describes the incoming LAG member links of the downstream LSR is
optional, because 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, 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

An MPLS echo request carrying a DDMAP TLV with the "Interface and
Label Stack Object Request flag" and "LAG Description Indicator flag"
set is sent 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"
   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:

   o When the "I" flag ("Interface and Label Stack Object Request
      flag") of the DDMAP TLV in the received MPLS echo request is set:

      * The responder LSR MUST add the Detailed Interface and Label
         Stack TLV (described in Section 10) in the MPLS echo reply.

      * If the incoming interface is 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:

   o Utilize the Incoming Interface Index Sub-TLV in the Detailed
      Interface and Label Stack TLV to derive, if the incoming interface
      is 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:

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

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

   o 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 the following
checks in a scenario where the initiator LSR 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 member links.

The initiator LSR sends two MPLS echo request messages to traverse
the two LAG member links at TTL=n+1:

o Success case:

   * One MPLS echo request message reaches TTL=n+1 on an LAG member
      link 1.

   * The other MPLS echo request message reaches TTL=n+1 on an LAG
      member link 2.

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

o Error case:

   * One MPLS echo request message reaches TTL=n+1 on an LAG member
      link 1.

   * The other MPLS echo request message also reaches TTL=n+1 on an
      LAG member link 1.

   One or two MPLS echo request messages sent by the initiator LSR
   cannot reach the immediate downstream LSR, or the two MPLS echo
   request messages reach at the immediate downstream LSR from the
   same LAG member link.

Note that the above defined procedures will provide a deterministic
result for LAG interfaces that are back-to-back connected between
LSRs (i.e. no L2 switch in between). If there is a L2 switch between
the LSR at TTL=n and the LSR at TTL=n+1, there is no guarantee that
traversal of every LAG member link at TTL=n will result in reaching
from different interface 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 for the downstream responder LSR are
then updated to include the comparison of the incoming LAG member
link to the interface index described in the Remote Interface Index
Sub-TLV in the DDMAP TLV. Failure of this comparison results in the
return code being set to "Downstream Mapping Mismatch (5)".

6. LSR Capability TLV

This document defines a new optional TLV which is referred to as the
"LSR Capability TLV. It 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 TLV. The presence of an LSR Capability TLV in an MPLS
echo request message is a request that a responder LSR includes an
LSR Capability TLV in the MPLS echo reply message, with the LSR
Capability TLV describing features and extensions that the responder
LSR supports.

When a responder LSR received an MPLS echo request message that
carries an LSR Capability TLV, if the responder LSR knows how to
process the LSR Capability TLV, an LSR Capability TLV MUST be
included in the MPLS echo reply message. Otherwise, if the responder
does not know the LSR Capability TLV, an MPLS echo reply with the
return code set to "One or more of the TLVs was not understood" MUST
be sent back to the initiator LSR.

The format of the LSR Capability TLV is as below:

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

```
 0                   1                   2                   3
 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|             Type              |             Length            |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|                     LSR Capability Flags                      |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
```

Figure 4: LSR Capability TLV

Where:

The Type is 2 octets in length and the value is TBD1.

The Length filed is 2 octets in length, and the value is 4.

The LSR Capability Flags is 4 octets in length, this document
defines following flags:

```
 0                   1                   2                   3
 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|                  Must Be Zero (Reserved)                  |U|D|
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
```

This document defines two flags. The remaining flags MUST be set
to zero when sending and ignored on receipt. Both the U and the D
flag MUST be cleared in the MPLS echo request message when
sending, and ignored on receipt. Neither, either or both the U
and the D flag MAY be set in the MPLS echo reply message.

Flag   Name and Meaning
----   ----------------

   U   Upstream LAG Info Accommodation

       An LSR sets this flag when the LSR is capable of
       describing a LAG member link in the Incoming Interface
       Index Sub-TLV in the Detailed Interface and
       Label Stack TLV.

   D   Downstream LAG Info Accommodation

       An LSR sets this flag when the LSR 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

This document defines a flag, the "G" flag (LAG Description
Indicator), in the DS Flags field of the DDMAP TLV.

The "G" flag 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.

The "G" flag is defined as below:

The Bit Number is TBD5.

```
    0 1 2 3 4 5 6 7
   +-+-+-+-+-+-+-+-+
   | MBZ |G|MBZ|I|N|
   +-+-+-+-+-+-+-+-+
```

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

 Flag   Name and Meaning
 ----   ----------------

   G   LAG Description Indicator

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

8. Local Interface Index Sub-TLV

   The Local Interface Index Sub-TLV is an optional TLV, it describes
   the interface index assigned by the local LSR to an egress interface.
   One or more Local Interface Index sub-TLVs MAY appear in a DDMAP TLV.

   The format of the Local Interface Index Sub-TLV is below:

```
    0                   1                   2                   3
    0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1
   +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
   |             Type              |             Length            |
   +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
   |                    Local Interface Index                     |
   +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
```

          Figure 5: Local Interface Index Sub-TLV

   Where:

   o The "Type" field is 2 octets in length, the value is TBD2.

   o The "Length" filed 2 octets in length, and the value is 4.

   o The "Local Interface Index" field is 4 octets in length, it is an
      interface index assigned by a local LSR to an egress interface.

9. Remote Interface Index Sub-TLV

   The Remote Interface Index Sub-TLV is an optional TLV, it describes
   the interface index assigned by a downstream LSR to an ingress
   interface. One or more Remote Interface Index sub-TLVs MAY appear in
   a DDMAP TLV.

   The format of the Remote Interface Index Sub-TLV is as below:

```
    0                   1                   2                   3
    0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1
   +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
   |             Type              |             Length            |
   +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
   |                     Remote Interface Index                    |
   +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
```
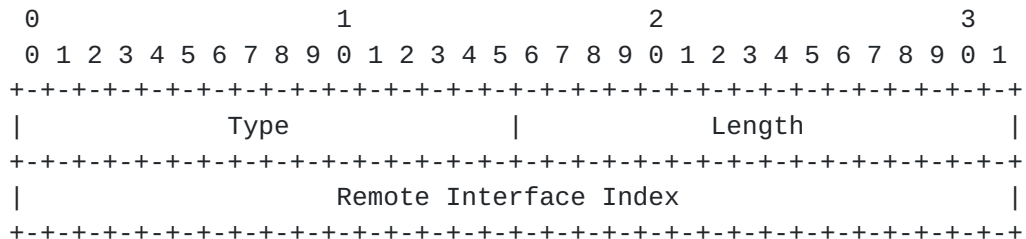
          Figure 6: Remote Interface Index Sub-TLV

   Where:

   o The "Type" field is 2 octets in length, and the value is TBD3.

   o The "Length" field is 2 octets in length, and the value is 4.

   o The "Remote Interface Index" is 4 octets in length, it is an
      interface index assigned by a downstream LSR to an ingress
      interface.

10. Detailed Interface and Label Stack TLV

   The "Detailed Interface and Label Stack" object is a TLV that MAY be
   included in an 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 into the MPLS echo
   reply message. 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:

```
  0                   1                   2                   3
   0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1
  +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
  |              Type             |             Length            |
  +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
  | Address Type  |            Must Be Zero                       |
  +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
  |                  IP Address (4 or 16 octets)                  |
  +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
  |                  Interface (4 or 16 octets)                   |
  +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
  .                                                               .
  .                     List of Sub-TLVs                          .
  .                                                               .
  +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
```
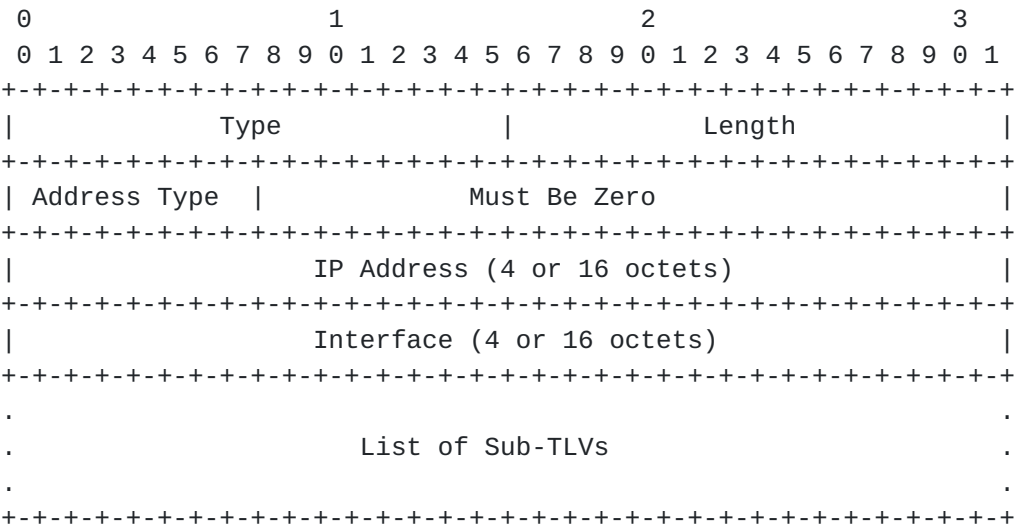
           Figure 7: Detailed Interface and Label Stack TLV

   The Detailed Interface and Label Stack TLV format is derived from the
   Interface and Label Stack TLV format (from [RFC8029]). Two changes
   are introduced. The first is that the label stack is converted into
   a sub-TLV. The second is that a new sub-TLV is added to describe an
   interface index. These fields of Detailed Interface and Label Stack
   TLV have the same use and meaning as in [RFC8029]. A summary of
   these fields 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 length of the
        initial part of the TLV is listed 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 [RFC8029],
described in Section 3.4.2 of [RFC7439]. A solution should be
considered an applied to both [RFC8029] and this document.

10.1. Sub-TLVs

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

```
        Sub-Type     Sub-TLV Name
        ---------    ------------
           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 an 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 its
format is as below:

```
   0                   1                   2                   3
   0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1
  +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
  |             Type              |             Length            |
  +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
  |               Label                   | TC  |S|     TTL       |
  +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
  .                                                               .
  .                                                               .
  .                                                               .
  +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
  |               Label                   | TC  |S|     TTL       |
  +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
```
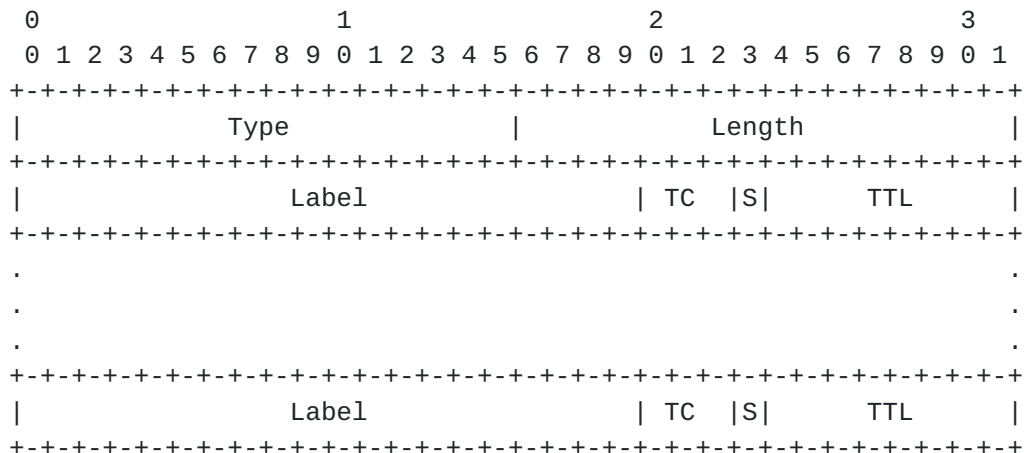
Figure 8: Incoming Label Stack Sub-TLV

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 a local LSR to the
   interface which received the MPLS echo request message.

   Incoming Interface Index Sub-TLV Type is 2. Length is 8, and its
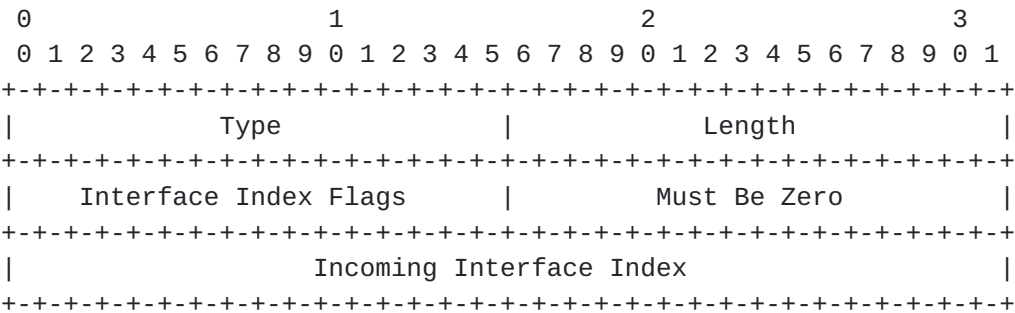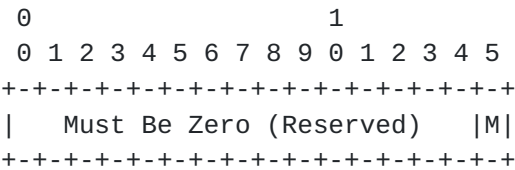   format is as below:

```
    0                   1                   2                   3
    0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1
   +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
   |             Type              |             Length            |
   +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
   |     Interface Index Flags     |          Must Be Zero         |
   +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
   |                    Incoming Interface Index                   |
   +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
```

         Figure 9: Incoming Interface Index Sub-TLV

   Interface Index Flags

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

```
    0                   1
    0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5
   +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
   |   Must Be Zero (Reserved)   |M|
   +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
```

      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 the additional multipath information. Due to additional
processing, it is critical that proper security measures described in
[RFC8029] are followed.

The LSP Traceroute allows an initiator LSR to discover the paths of
tested LSPs, providing detailed 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.[RFC8029] provides additional
recommendations to avoid attacks and recommendations to follow if an
operator desires to prevent tracing.

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
[RFC8126].

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
[RFC8126].

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

o The Local Interface Index Sub-TLV which may be present in the
   "Downstream Detailed Mapping" TLV.

o The Remote Interface Index Sub-TLV which may be present in the
   "Downstream Detailed Mapping" TLV.

o 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 via
Standards Action [RFC8126]. Assignments of Sub-Types in the
experimental space is via Specification Required [RFC8126].

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 [RFC8029].

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

13. Acknowledgements

The authors would like to thank Nagendra Kumar, 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,
           <https://www.rfc-editor.org/info/rfc2119>.

[RFC8029] Kompella, K., Swallow, G., Pignataro, C., Ed., Kumar, N.,
           Aldrin, S., and M. Chen, "Detecting Multiprotocol Label
           Switched (MPLS) Data-Plane Failures", RFC 8029,
           DOI 10.17487/RFC8029, March 2017,
           <https://www.rfc-editor.org/info/rfc8029>.

[RFC8174] Leiba, B., "Ambiguity of Uppercase vs Lowercase in RFC
           2119 Key Words", BCP 14, RFC 8174, DOI 10.17487/RFC8174,
           May 2017, <https://www.rfc-editor.org/info/rfc8174>.

14.2. Informative References

[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.

   [RFC7439] George, W., Ed. and C. Pignataro, Ed., "Gap Analysis for
             Operating IPv6-Only MPLS Networks", RFC 7439,
             DOI 10.17487/RFC7439, January 2015,
             <https://www.rfc-editor.org/info/rfc7439>.

   [RFC8126] Cotton, M., Leiba, B., and T. Narten, "Guidelines for
             Writing an IANA Considerations Section in RFCs", BCP 26,
             RFC 8126, DOI 10.17487/RFC8126, June 2017,
             <https://www.rfc-editor.org/info/rfc8126>.

Appendix A. LAG with intermediate L2 Switch Issues

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

A.1. Equal Numbers of LAG Members

   R1 ==== S1 ==== R2

   The issue with this LAG provisioning model is that packets traversing
   a LAG member from Router 1 (R1) to intermediate L2 switch (S1) can
   get load balanced by S1 towards Router 2 (R2). Therefore, MPLS echo
   request messages traversing a specific LAG member from R1 to S1 can
   actually reach R2 via any of the LAG members, and the sender of MPLS
   echo request messages has 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

                _____
   R1 ==== S1 ==== R2

   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

   Mach(Guoyi) Chen
   Huawei

      Email: mach.chen@huawei.com