         Label Switched Path (LSP) and Pseudowire (PW) Ping/Trace over
                  MPLS Network using Entropy Labels (EL)
                    draft-ietf-mpls-entropy-lsp-ping-01

Abstract

   The Multiprotocol Label Switching (MPLS) Label Switched Path (LSP)
   Ping and Traceroute are used to exercise specific paths of Equal-Cost
   Multipath (ECMP). When LSP is signaled to use Entropy Label (EL)
   described in RFC6790, the ability for LSP Ping and Traceroute
   operation to discover and exercise ECMP paths has been lost in
   scenarios which LSRs apply deviating load balance techniques. One
   such scenario is when some LSRs apply EL based load balancing while
   other LSRs apply non-EL based load balancing (ex: IP). Another
   scenario is when EL based LSP is stitched with another LSP which can
   be EL based or non-EL based.

   This document extends the MPLS LSP Ping and Traceroute mechanisms to
   restore the ability of exercising specific paths of ECMP over LSP
   which make use of Entropy Label. This document updates RFC4379,
   RFC6424 and RFC6790.

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 December 13, 2015.

Copyright Notice

Table of Contents

1. Introduction

1.1. Terminology

   The following acronyms/terminologies are used in this document:

   o MPLS - Multiprotocol Label Switching.

   o LSP - Label Switched Path.

   o LSR - Label Switching Router.

   o FEC - Forwarding Equivalent Class.

   o ECMP - Equal-Cost Multipath.

   o EL - Entropy Label.

   o ELI - Entropy Label Indicator.

   o GAL - Generic Associated Channel Label.

   o MS-PW - Multi-Segment Pseudowire.

   o Initiating LSR - LSR which sends MPLS echo request.

   o Responder LSR - LSR which receives MPLS echo request and sends
      MPLS echo reply.

   o IP Based Load Balancer - LSR which load balances on fields from IP
      header (and possibly fields from upper layers), and does not
      consider entropy label from label stack (i.e. Flow Label or
      Entropy Label) for load balancing purpose.

   o Label Based Load Balancer - LSR which load balances on entropy
      label from label stack (i.e. Flow Label or Entropy Label), and
      does not consider fields from IP header (and possibly fields from
      upper layers) for load balancing purpose.

   o Label and IP Based Load Balancer - LSR which load balances on both
      labels from label stack (including Flow Label or Entropy Label if
      present) and fields from IP header (and possibly fields from upper
      layers).

1.2. Prerequisite

   MPLS implementations employ wide variety of load balancing techniques
   in terms of fields used for hash "keys". [RFC4379] and [RFC6424] are
   designed to provide multipath support for subset of techniques.
   Intent of this document is to restore multipath support for those
   supported techniques which have been compromised by the introduction
   of [RFC6790] (i.e. Entropy Labels). Section 10 describes supported
   and unsupported cases, and it may be useful for one to visit this
   section first.

1.3. Background

   Section 3.3.1 of [RFC4379] specifies multipath information encoding
   in Downstream Mapping TLV (Section 3.3 of [RFC4379]) and Downstream
   Detailed Mapping TLV (Section 3.3 of [RFC6424]) which can be used by
   LSP Ping initiator to trace and validate all ECMP paths between
   ingress and egress. These encodings are sufficient when all the LSRs
   along the path(s), between ingress and egress, consider same set of
   "keys" as input for load balancing algorithm: all IP based or all
   label based.

   With introduction of [RFC6790], it is quite normal to see set of LSRs
   performing load balancing based on EL/ELI while others still follow
   the traditional way (IP based). This results in LSP Ping initiator
   not be able to trace and validate all ECMP paths in following
   scenarios:

   o One or more transit LSRs along LSP with ELI/EL in label stack do
      not perform ECMP load balancing based on EL (hashes based on
      "keys" including IP destination address). This scenario is not
      only possible but quite common due transit LSRs not implementing
      [RFC6790] or transit LSRs implementing [RFC6790] but not
      implementing suggested transit LSR behavior in Section 4.3 of
      [RFC6790].

   o Two or more LSPs stitched together with at least one of these LSP
      pushing ELI/EL in label stack. Such scenarios are described in
      [I-D.ravisingh-mpls-el-for-seamless-mpls].

   These scenarios will be quite common because every deployment of
   [RFC6790] will invariably end up with nodes that support ELI/EL and

nodes that do not. There will typically be areas that support ELI/EL
and areas that do not.

As pointed out in [RFC6790] the procedures of [RFC4379] with respect
to multipath information type {9} are incomplete. However [RFC6790]
does not actually update [RFC4379]. Further the specific EL location
is not clearly defined, particularly in the case of Flow Aware
Pseudowires [RFC6391]. This document defines a new FEC Stack sub-TLV
for the Entropy Label. Section 3 of this document updates the
procedures for multipath information type {9} described in [RFC4379].
Rest of this document describes extensions required to restore ECMP
discovery and tracing capabilities for scenarios described.

2. Overview

[RFC4379] describes LSP traceroute as an operation where the
initiating LSR send a series of MPLS echo requests towards the same
destination. The first packet in the series have the TTL set to 1.
When the echo reply is received from the LSR one hop away the second
echo request in the series is sent with the TTL set to 2, for each
echo request the TLL is incremented by one until a response is
received from the intended destination. Initiating LSR discovers and
exercises ECMP by obtaining multipath information from each transit
LSR and using specific destination IP address or specific entropy
label.

Notion of {x, y, z} from here on refers to Multipath information
types x, y or z.

LSP Ping initiating LSR sends MPLS echo request with multipath
information. This multipath information is described in DSMAP/DDMAP
TLV of echo request, and may contain set of IP addresses or set of
labels. Multipath information types {2, 4, 8} carry set of IP
addresses and multipath information type {9} carries set of labels.
Responder LSR (receiver of MPLS echo request) will determine the
subset of initiator specified multipath information which load
balances to each downstream (outgoing interface). Responder LSR
sends MPLS echo reply with resulting multipath information per
downstream (outgoing interface) back to the initiating LSR.
Initiating LSR is then able to use specific IP destination address or
specific label to exercise specific ECMP path on the responder LSR.

Current behavior is problematic in following scenarios:

o Initiating LSR sends IP multipath information, but responder LSR
   load balances on labels.

o Initiating LSR sends label multipath information, but responder
   LSR load balances on IP addresses.

o Initiating LSR sends existing multipath information to LSR which
   pushes ELI/EL in label stack, but the initiating LSR can only
   continue to discover and exercise specific path of ECMP, if the
   LSR which pushes ELI/EL responds with both IP addresses and
   associated EL corresponding to each IP address. This is because:

   * ELI/EL pushing LSR that is a stitching point will load balance
      based on IP address.

   * Downstream LSR(s) of ELI/EL pushing LSR may load balance based
      on ELs.

o Initiating LSR sends one of existing multipath information to ELI/
   EL pushing LSR, but initiating LSR can only continue to discover
   and exercise specific path of ECMP if ELI/EL pushing LSR responds
   with both labels and associated EL corresponding to label. This
   is because:

   * ELI/EL pushing LSR that is a stitching point will load balance
      based on EL from previous LSP and pushes new EL.

   * Downstream LSR(s) of ELI/EL pushing LSR may load balance based
      on new ELs.

The above scenarios point to how the existing multipath information
is insufficient when LSP traceroute is operated on an LSP with
Entropy Labels described by [RFC6790]. Therefore, this document
defines a multipath information type to be used in the DSMAP/DDMAP of
MPLS echo request/reply packets in Section 9.

In addition, responder LSR can reply with empty multipath information
if no IP address set or label set from received multipath information
matched load balancing to a downstream. Empty return is also
possible if initiating LSR sends multipath information of one type,
IP address or label, but responder LSR load balances on the other
type. To disambiguate between the two results, this document
introduces new flags in the DSMAP/DDMAP TLV to allow responder LSR to
describe the load balance technique being used.

It is required that all LSRs along the LSP understand new flags as
well as new multipath information type. It is also required that
initiating LSR can select both IP destination address and label to
use on transmitting MPLS echo request packets. Two additional DS
Flags are defined for the DSMAP and DDMAP TLVs in Section 8. These

two flags are used by the responder LSR to describe its load balance
behavior on received MPLS echo request.

Note that the terms "IP Based Load Balancer", "Label Based Load
Balancer" and "Label Based Load Balancer" are in context of how
received MPLS echo request is handled by the responder LSR.

3. Multipath Type 9

This section defines to which labels multipath type {9} applies.

[RFC4379] defined multipath type {9} for tracing of LSPs where label
based load-balancing is used. However, as pointed out in [RFC6790],
the procedures for using this type are incomplete as the specific
location of the label was not defined. It was assumed that the
presence of multipath type {9} implied the value of the bottom-of-
stack label should be varied by the values indicated by multipath to
determine their respective out-going interfaces.

Section 7 defines a new FEC-Stack sub-TLV to indicate an entropy
label. These labels may appear anywhere in a label stack.

Multipath type {9} applies to the first label in the label-stack that
corresponds to an EL-FEC. If no such label is found, it applies to
the label at the bottom of the label stack.

4. Pseudowire Tracing

This section defines procedures for tracing pseudowires. These
procedures pertain to the use of multipath information type {9} as
well as type {TBD4}. In all cases below, when a control word is in
use the N-flag in the DDMAP or DSMAP MUST be set. Note that when a
control word is not in use the returned DDMAPs or DSMAPs may not be
accurate.

In order to trace a non Flow-Aware Pseudowire the initiator includes
an EL-FEC instead of the appropriate PW-FEC at the bottom of the FEC-
Stack. Tracing in this way will cause compliant routers to return
the proper outgoing interface. Note that this procedure only traces
to the end of the MPLS LSP that is under test and will not verify the
PW FEC. To actually verify the PW-FEC or in the case of a MS-PW, to
determine the next pseudowire label value, the initiator MUST repeat
that step of the trace, (i.e., repeating the TTL value used) but with
the FEC-Stack modified to contain the appropriate PW-FEC. Note that
these procedures are applicable to scenarios which an initiator is
able to vary the bottom label (i.e. pseudowire label). Possible
scenarios are tracing multiple non Flow-Aware Pseudowires on the same

endpoints or tracing a non Flow-Aware Pseudowire provisioned with multiple pseudowire labels.

In order to trace a Flow Aware Pseudowire, the initiator includes an EL-FEC at the bottom of the FEC-Stack and pushes the appropriate PW-FEC onto the FEC-Stack.

In order to trace through non-compliant routers the initiator forms an MPLS echo request message and includes a DDMAP or DSMAP with multipath type {9}. For a non Flow-Aware Pseudowire it includes the appropriate PW-FEC in the FEC-Stack. For a Flow Aware Pseudowire, the initiator includes a NIL-FEC at the bottom of the FEC-Stack and pushes the appropriate PW-FEC onto the FEC-Stack.

5. Initiating LSR Procedures

In order to facilitate the flow of the following text we speak in terms of a boolean called EL_LSP maintained by the initiating LSR. This value controls the multipath information type to be used in transmitted echo request packets. When the initiating LSR is transmitting an echo request packet with DSMAP/DDMAP with a non-zero multipath information type, then EL_LSP boolean MUST be consulted to determine the multipath information type to use.

In addition to procedures described in [RFC4379] as updated by Section 3 and [RFC6424], initiating LSR MUST operate with following procedures.

o When the initiating LSR pushes ELI/EL, initialize EL_LSP=True. Else set EL_LSP=False.

o When the initiating LSR is transmitting non-zero multipath information type:

   * If (EL_LSP), the initiating LSR MUST use multipath information type {TBD4} unless same responder LSR cannot handle type {TBD4}. When the initiating LSR is transmitting multipath information type {TBD4} in this case, both "IP Multipath Information" and "Label Multipath Information" MUST be included, and "IP Associated Label Multipath Information" MUST be omitted (NULL).

   * Else the initiating LSR MAY use multipath information type {2, 4, 8, 9, TBD4}. When the initiating LSR is transmitting multipath information type {TBD4} in this case, "IP Multipath Information" MUST be included, and "Label Multipath Information" and "IP Associated Label Multipath Information" MUST be omitted (NULL).

   o When the initiating LSR receives echo reply with {L=0, E=1} in DS
      flags with valid contents, set EL_LSP=True.

   In following conditions, the initiating LSR may have lost the ability
   to exercise specific ECMP paths. The initiating LSR MAY continue
   with "best effort".

   o Received echo reply contains empty multipath information.

   o Received echo reply contains {L=0, E=<any>} DS flags, but does not
      contain IP multipath information.

   o Received echo reply contains {L=1, E=<any>} DS flags, but does not
      contain label multipath information.

   o Received echo reply contains {L=<any>, E=1} DS flags, but does not
      contain associated label multipath information.

   o IP multipath information types {2, 4, 8} sent, and received echo
      reply with {L=1, E=0} in DS flags.

   o Multipath information type {TBD4} sent, and received echo reply
      with multipath information type other than {TBD4}.

6. Responder LSR Procedures

   Common Procedures:

   o The responder LSR receiving an MPLS echo request packet MUST first
      determine whether or not the initiating LSR supports this LSP Ping
      and Traceroute extension for Entropy Labels. If either of the
      following conditions are met, the responder LSR SHOULD determine
      that the initiating LSR supports this LSP Ping and Traceroute
      extension for Entropy Labels.

      1. Received MPLS echo request contains the multipath information
          type {TBD4}.

      2. Received MPLS echo request contains a Target FEC Stack TLV
          that includes the Entropy Label FEC.

      If the initiating LSR is determined to not support this LSP Ping
      and Traceroute extension for Entropy Labels, then the responder
      LSR MUST NOT follow further procedures described in this section.
      Specifically, MPLS echo reply packets:

      * MUST have following DS Flags cleared (i.e., not set): "ELI/EL
         push indicator" and "Label based load balance indicator".

        *   MUST NOT use multipath information type {TBD4}.

   o The responder LSR receiving an MPLS echo request packet with
      multipath information type {TBD4} MUST validate following
      contents. Any deviation MUST result in the responder LSR to
      consider the packet as malformed and return code 1 (Malformed echo
      request received) in the MPLS echo reply packet.

        *   IP multipath information MUST be included.

        *   Label multipath information MAY be included.

        *   IP associated label multipath information MUST be omitted
            (NULL).

   Following subsections describe expected responder LSR procedures when
   echo reply is to include DSMAP/DDMAP TLVs, based on local load
   balance technique being employed. In case the responder LSR performs
   deviating load balance techniques per downstream basis, appropriate
   procedures matching to each downstream load balance technique MUST be
   operated.

6.1. IP Based Load Balancer & Not Pushing ELI/EL

   o The responder MUST set {L=0, E=0} in DS flags.

   o If multipath information type {2, 4, 8} is received, the responder
      MUST comply with [RFC4379] and [RFC6424].

   o If multipath information type {9} is received, the responder MUST
      reply with multipath type {0}.

   o If multipath information type {TBD4} is received, following
      procedures are to be used:

        *   The responder MUST reply with multipath information type
            {TBD4}.

        *   "Label Multipath Information" and "Associated Label Multipath
            Information" sections MUST be omitted (NULL).

        *   If no matching IP address is found, then "IPMultipathType"
            field MUST be set to multipath information type {0} and "IP
            Multipath Information" section MUST also be omitted (NULL).

        *   If at least one matching IP address is found, then
            "IPMultipathType" field MUST be set to appropriate multipath

        information type {2, 4, 8} and "IP Multipath Information"
        section MUST be included.

6.2. IP Based Load Balancer & Pushes ELI/EL

   o The responder MUST set {L=0, E=1} in DS flags.

   o If multipath information type {9} is received, the responder MUST
      reply with multipath type {0}.

   o If multipath type {2, 4, 8, TBD4} is received, following
      procedures are to be used:

      * The responder MUST respond with multipath type {TBD4}. See
         Section 9 for details of multipath type {TBD4}.

      * "Label Multipath Information" section MUST be omitted (i.e. is
         it not there).

      * IP address set specified in received IP multipath information
         MUST be used to determine the returning IP/Label pairs.

      * If received multipath information type was {TBD4}, received
         "Label Multipath Information" sections MUST NOT be used to
         determine the associated label portion of returning IP/Label
         pairs.

      * If no matching IP address is found, then "IPMultipathType"
         field MUST be set to multipath information type {0} and "IP
         Multipath Information" section MUST be omitted. In addition,
         "Assoc Label Multipath Length" MUST be set to 0, and
         "Associated Label Multipath Information" section MUST also be
         omitted.

      * If at least one matching IP address is found, then
         "IPMultipathType" field MUST be set to appropriate multipath
         information type {2, 4, 8} and "IP Multipath Information"
         section MUST be included. In addition, "Associated Label
         Multipath Information" section MUST be populated with list of
         labels corresponding to each IP address specified in "IP
         Multipath Information" section. "Assoc Label Multipath Length"
         MUST be set to a value representing length in octets of
         "Associated Label Multipath Information" field.

6.3. Label Based Load Balancer & Not Pushing ELI/EL

   o The responder MUST set {L=1, E=0} in DS flags.

   o If multipath information type {2, 4, 8} is received, the responder
      MUST reply with multipath type {0}.

   o If multipath information type {9} is received, the responder MUST
      comply with [RFC4379] and [RFC6424] as updated by Section 3.

   o If multipath information type {TBD4} is received, following
      procedures are to be used:

      * The responder MUST reply with multipath information type
         {TBD4}.

      * "IP Multipath Information" and "Associated Label Multipath
         Information" sections MUST be omitted (NULL).

      * If no matching label is found, then "LbMultipathType" field
         MUST be set to multipath information type {0} and "Label
         Multipath Information" section MUST also be omitted (NULL).

      * If at least one matching label is found, then "LbMultipathType"
         field MUST be set to appropriate multipath information type {9}
         and "Label Multipath Information" section MUST be included.

6.4. Label Based Load Balancer & Pushes ELI/EL

   o The responder MUST set {L=1, E=1} in DS flags.

   o If multipath information type {2, 4, 8} is received, the responder
      MUST reply with multipath type {0}.

   o If multipath type {9, TBD4} is received, following procedures are
      to be used:

      * The responder MUST respond with multipath type {TBD4}.

      * "IP Multipath Information" section MUST be omitted.

      * Label set specified in received label multipath information
         MUST be used to determine the returning Label/Label pairs.

      * If received multipath information type was {TBD4}, received
         "Label Multipath Information" sections MUST NOT be used to
         determine the associated label portion of returning Label/Label
         pairs.

       * If no matching label is found, then "LbMultipathType" field
          MUST be set to multipath information type {0} and "Label
          Multipath Information" section MUST be omitted. In addition,
          "Assoc Label Multipath Length" MUST be set to 0, and
          "Associated Label Multipath Information" section MUST also be
          omitted.

       * If at least one matching label is found, then "LbMultipathType"
          field MUST be set to appropriate multipath information type {9}
          and "Label Multipath Information" section MUST be included. In
          addition, "Associated Label Multipath Information" section MUST
          be populated with list of labels corresponding to each label
          specified in "Label Multipath Information" section. "Assoc
          Label Multipath Length" MUST be set to a value representing
          length in octets of "Associated Label Multipath Information"
          field.

6.5. Flow Aware MS-PW Stitching LSR

   Stitching LSR that cross-connects Flow Aware Pseudowires behave in
   one of two ways:

   o Load balances on previous Flow Label, and carries over same Flow
      Label. For this case, stitching LSR is to behave as procedures
      described in Section 6.3.

   o Load balances on previous Flow Label, and replaces Flow Label with
      newly computed. For this case, stitching LSR is to behave as
      procedures described in Section 6.4.

7. Entropy Label FEC

   Entropy Label Indicator (ELI) is a reserved label that has no
   explicit FEC associated, and has label value 7 assigned from the
   reserved range. Use Nil FEC as Target FEC Stack sub-TLV to account
   for ELI in a Target FEC Stack TLV.

   Entropy Label (EL) is a special purpose label with label value being
   discretionary (i.e. label value may not be from the reserved range).
   For LSP verification mechanics to perform its purpose, it is
   necessary for a Target FEC Stack sub-TLV to clearly describe EL,
   particularly in the scenario where label stack does not carry ELI
   (ex: Flow Aware Pseudowire [RFC6391]). Therefore, this document
   defines a EL FEC to allow a Target FEC Stack sub-TLV to be added to
   the Target FEC Stack to account for EL.

   The Length is 4. Labels are 20-bit values treated as numbers.

```
    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
   +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
   |                  Label                |             MBZ         |
   +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
```
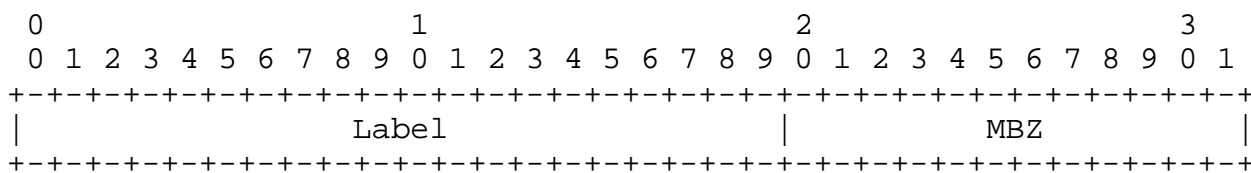
            Figure 1: Entropy Label FEC

   Label is the actual label value inserted in the label stack; the MBZ
   fields MUST be zero when sent and ignored on receipt.

8. DS Flags: L and E

   Two flags, L and E, are added in DS Flags field of the DSMAP/DDMAP
   TLVs. Both flags MUST NOT be set in echo request packets when
   sending, and ignored when received. Zero, one or both new flags MUST
   be set in echo reply packets.

    DS Flags
    ----

        0 1 2 3 4 5 6 7
       +-+-+-+-+-+-+-+-+
       | MBZ |L|E|I|N|
       +-+-+-+-+-+-+-+-+
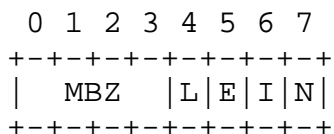
   RFC-Editor-Note: Please update above figure to place the flag E in
   the bit number TBD2 and the flag L in the bit number TBD3.

    Flag   Name and Meaning
    ----   ----------------
       L   Label based load balance indicator
           This flag MUST be set to zero in the echo request. LSR
           which performs load balancing on a label MUST set this
           flag in the echo reply. LSR which performs load
           balancing on IP MUST NOT set this flag in the echo
           reply.

       E   ELI/EL push indicator
           This flag MUST be set to zero in the echo request. LSR
           which pushes ELI/EL MUST set this flag in the echo
           reply. LSR which does not push ELI/EL MUST NOT set
           this flag in the echo reply.

   Two flags result in four load balancing techniques which echo reply
   generating LSR can indicate:

   o {L=0, E=0} LSR load balances based on IP and does not push ELI/EL.

o {L=0, E=1} LSR load balances based on IP and pushes ELI/EL.

o {L=1, E=0} LSR load balances based on label and does not push ELI/
   EL.

o {L=1, E=1} LSR load balances based on label and pushes ELI/EL.

9. New Multipath Information Type: TBD4

   One new multipath information type is added to be used in DSMAP/DDMAP
   TLVs. New multipath type has value of TBD4.

     Key   Type   Multipath Information
     ---   ----------------   ---------------------
     TBD4   IP and label set   IP addresses and label prefixes

   Multipath type TBD4 is comprised of three sections. One section to
   describe IP address set. One section to describe label set. One
   section to describe another label set which associates to either IP
   address set or label set specified in the other section.

   Multipath information type TBD4 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
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|IPMultipathType|     IP Multipath Length       | Reserved(MBZ) |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
~                                                               ~
|                  (IP Multipath Information)                   |
~                                                               ~
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|LbMultipathType|    Label Multipath Length     | Reserved(MBZ) |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
~                                                               ~
|                (Label Multipath Information)                  |
~                                                               ~
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|  Assoc Label Multipath Length |           Reserved(MBZ)       |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
~                                                               ~
|            (Associated Label Multipath Information)           |
~                                                               ~
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
```

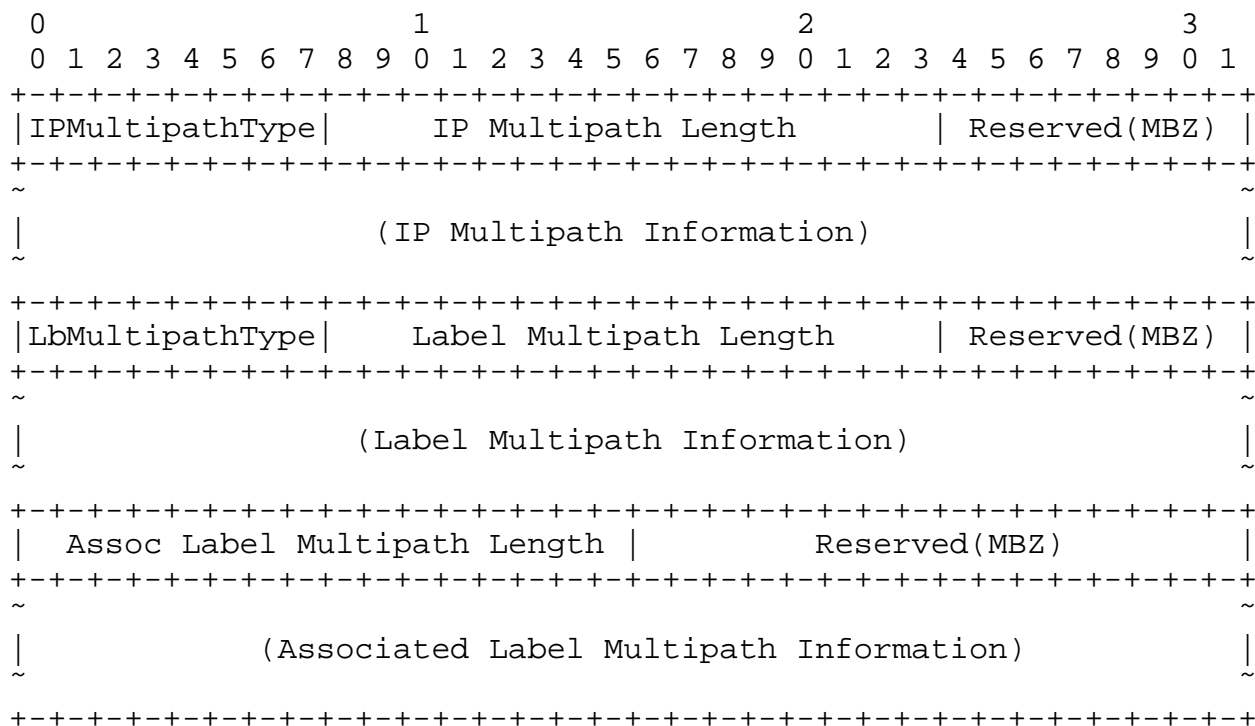   Figure 2: Multipath Information Type TBD4

   o IPMultipathType

   * 0 when "IP Multipath Information" is omitted. Otherwise one of
      IP multipath information values: {2, 4, 8}.

   o IP Multipath Information

   * This section is omitted when "IPMultipathType" is 0. Otherwise
      this section reuses IP multipath information from [RFC4379].
      Specifically, multipath information for values {2, 4, 8} can be
      used.

   o LbMultipathType

   * 0 when "Label Multipath Information" is omitted. Otherwise
      label multipath information value {9}.

   o Label Multipath Information

   * This section is omitted when "LbMultipathType" is 0. Otherwise
      this section reuses label multipath information from [RFC4379].
      Specifically, multipath information for value {9} can be used.

   o Associated Label Multipath Information

   * "Assoc Label Multipath Length" is a 16 bit field of multipath
      information which indicates length in octets of the associated
      label multipath information.

   * "Associated Label Multipath Information" is a list of labels
      with each label described in 24 bits. This section MUST be
      omitted in an MPLS echo request message. A midpoint which
      pushes ELI/EL labels SHOULD include "Assoc Label Multipath
      Information" in its MPLS echo reply message, along with either
      "IP Multipath Information" or "Label Multipath Information".
      Each specified associated label described in this section maps
      to specific IP address OR label described in the "IP Multipath
      Information" section or "Label Multipath Information" section.
      For example, if 3 IP addresses are specified in the "IP
      Multipath Information" section, then there MUST be 3 labels
      described in this section. First label maps to the lowest IP
      address specified, second label maps to the second lowest IP
      address specified and third label maps to the third lowest IP
      address specified.

10. Supported and Unsupported Cases

   MPLS architecture never defined strict rules on how implementations
   are to identify hash "keys" for load balancing purpose. As result,
   implementations may be of following load balancer types:

1. IP Based Load Balancer.
2. Label Based Load Balancer.
3. Label and IP Based Load Balancer.

For cases (2) and (3), implementation can include different sets of
labels from the label stack for load balancing purpose. Thus
following sub-cases are possible:

a. Entire label stack.
b. Top N labels from label stack where number of labels in label
    stack is >N.
c. Bottom N labels from label stack where number of labels in label
    stack is >N.

In a scenario where there is one Flow Label or Entropy Label present
in the label stack, following further cases are possible for (2b),
(2c), (3b) and (3c):

1. N labels from label stack include Flow Label or Entropy Label.
2. N labels from label stack does not include Flow Label or Entropy
    Label.

Also in a scenario where there are multiple Entropy Labels present in
the label stack, it is possible for implementations to employ
deviating techniques:

o Search for entropy stops at the first Entropy Label.
o Search for entropy includes any Entropy Label found plus continues
   to search for entropy in the label stack.

Furthermore, handling of reserved (i.e. special) labels varies among
implementations:

o Reserved labels are used in the hash as any other label would be
   (a bad practice).
o Reserved labels are skipped over and, for implementations limited
   to N labels, the reserved labels do not count towards the limit of
   N.
o Reserved labels are skipped over and, for implementations limited
   to N labels, the reserved labels count towards the limit of N.

It is important to point this out since presence of GAL will affect
those implementations which include reserved labels for load
balancing purpose.

As can be seen from above, there are many flavors of potential load
balancing implementations. Attempting for any OAM tools to support
ECMP discovery and traversal over all flavors of such will require

fairly complex procedures and implementations to support those
complex procedures. Complexities in OAM tools will produce minimal
benefits if majority of implementations are expected to employ small
subset of cases described above.

o Section 4.3 of [RFC6790] states that implementations, for load
   balancing purpose, parsing beyond the label stack after finding
   Entropy Label is "limited incremental value". Therefore, it is
   expected that most implementations will be of types "IP Based Load
   Balancer" or "Label Based Load Balancer".

o Section 2.4.5.1 of [RFC7325] recommends that search for entropies
   from the label stack should terminate upon finding the first
   Entropy Label. Therefore, it is expected that implementations
   will only include the first (top-most) Entropy Label when there
   are multiple Entropy Labels in the label stack.

o It is expected that, in most cases, number of labels in the label
   stack will not exceed number of labels (N) which implementations
   can include for load balancing purpose.

o It is expected that labels in the label stack, besides Flow Label
   and Entropy Label, are constant for the lifetime of a single LSP
   multipath traceroute operation. Therefore, deviating load
   balancing implementations with respect to reserved labels should
   not affect this tool.

Thus [RFC4379], [RFC6424] and this document will support cases (1)
and (2a1), where only the first (top-most) Entropy Label is included
when there are multiple Entropy Labels in the label stack.

11. Security Considerations

This document extends LSP Traceroute mechanism to discover and
exercise ECMP paths when LSP uses ELI/EL in label stack. Additional
processings are required for responder and initiator nodes.
Responder node that pushes ELI/EL will need to compute and return
multipath data including associated EL. Initiator node will need to
store and handle both IP multipath and label multipath information,
and include destination IP addresses and/or ELs in MPLS echo request
packet as well as in carried multipath information to downstream
nodes. Due to additional processing, it is critical that proper
security measures described in [RFC4379] and [RFC6424] are followed.

12. IANA Considerations

12.1. DS Flags

The IANA is requested to assign new bit numbers 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
 ---------- ---------------------------------------    ---------
       TBD2 E: ELI/EL push indicator                   this document
       TBD3 L: Label based load balance indicator      this document
```

12.2. Multpath Type

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

Note: the "Multipath Type" sub-registry is created by [RFC7537].

```
 Value      Meaning                                  Reference
 ---------- ---------------------------------------- ---------
   TBD4       IP and label set                       this document
```

12.3. Entropy Label FEC

The IANA is requested to assign a new sub-TLV from the "Sub-TLVs for
TLV Types 1 and 16" section from the "Multi-Protocol Label Switching
(MPLS) Label Switched Paths (LSPs) Ping Parameters - TLVs" registry
([IANA-MPLS-LSP-PING]).

```
 Sub-Type Sub-TLV Name            Reference
 -------- ------------            ---------
     TBD1 Entropy Label FEC       this document
```

13. Acknowledgements

Authors would like to thank Loa Andersson, Curtis Villamizar, Daniel
King, Sriganesh Kini and Victor Ji for performing thorough review and
providing valuable comments.

14. Contributing Authors

   Nagendra Kumar
   Cisco Systems
   Email: naikumar@cisco.com

15. References

15.1. Normative References

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

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

   [RFC6790] Kompella, K., Drake, J., Amante, S., Henderickx, W., and
              L. Yong, "The Use of Entropy Labels in MPLS Forwarding",
              RFC 6790, November 2012.

   [RFC7537] Decraene, B., Akiya, N., Pignataro, C., Andersson, L., and
              S. Aldrin, "IANA Registries for LSP Ping Code Points", RFC
              7537, May 2015.

15.2. Informative References

   [I-D.ravisingh-mpls-el-for-seamless-mpls]
              Singh, R., Shen, Y., and J. Drake, "Entropy label for
              seamless MPLS", draft-ravisingh-mpls-el-for-seamless-
              mpls-04 (work in progress), October 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>.

   [RFC6391] Bryant, S., Filsfils, C., Drafz, U., Kompella, V., Regan,
              J., and S. Amante, "Flow-Aware Transport of Pseudowires
              over an MPLS Packet Switched Network", RFC 6391, November
              2011.

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

   [RFC7325] Villamizar, C., Kompella, K., Amante, S., Malis, A., and
              C. Pignataro, "MPLS Forwarding Compliance and Performance
              Requirements", RFC 7325, August 2014.

Authors' Addresses

   Nobo Akiya
   Big Switch Networks

   Email: nobo.akiya.dev@gmail.com

   George Swallow
   Cisco Systems

   Email: swallow@cisco.com

   Carlos Pignataro
   Cisco Systems

   Email: cpignata@cisco.com

   Andrew G. Malis
   Huawei Technologies

   Email: agmalis@gmail.com

   Sam Aldrin
   Google

   Email: aldrin.ietf@gmail.com