Internet Engineering Task Force
Internet-Draft
Updates: 4379,6424 (if approved)
Intended status: Standards Track
Expires: December 15, 2014

N. Akiya
G. Swallow
Cisco Systems
S. Litkowski
B. Decraene
Orange
J. Drake
Juniper Networks

June 13, 2014

Label Switched Path (LSP) Ping/Trace Multipath Support for Link Aggregation Group (LAG) Interfaces
draft-akiya-mpls-lsp-ping-lag-multipath-00

Abstract

This document defines an extension to the Multiprotocol Label Switching (MPLS) Label Switched Path (LSP) Ping and Traceroute to describe Multipath Information for Link Aggregation (LAG) member links separately, thus allowing MPLS LSP Ping and Traceroute to discover and exercise specific paths of layer 2 Equal-Cost Multipath (ECMP) over 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 December 15, 2014.

Table of Contents

1. Introduction

1.1. Terminology

The following acronyms/terminologies are used in this document:

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

1.2. Background

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

- MPLS switching of LSP X over one or more member links of LAG Y is succeeding.
- MPLS switching of LSP X over one or more member links of LAG Y is failing.
- MPLS echo request for LSP X over LAG Y is load balanced over a member link which is MPLS switching successfully.

With above scenario, MPLS LSP Ping and Traceroute will not be able to detect the MPLS switching failure of problematic member link(s) of the LAG. In other words, lack of layer 2 ECMP discovery and exercise capability can produce an outcome where MPLS LSP Ping and Traceroute can be blind to MPLS switching failures over LAG interface that are impacting MPLS traffic. It is, thus, desirable to extend the MPLS

LSP Ping and Traceroute to have deterministic diagnostic coverage of LAG interfaces.

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 layer 2 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 layer 2 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). For example:

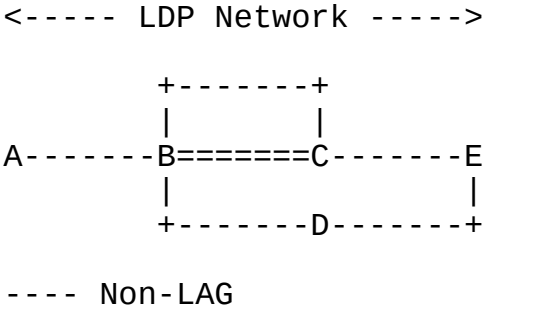


Figure 1: Example LDP Network

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 to discover L2 ECMP multipath information;
- In Section 4, a mechanism to validate L2 ECMP traversal in some LAG provisioning models;
- In Section 5, the LAG Interface Info TLV;
- In Section 6, the LAG Description Indicator flag;
- In Section 7, the Interface Index Sub-TLV;
- In Section 8, the Detailed Interface and Label Stack TLV.

3. Mechanism to Discover L2 ECMP Multipath

The MPLS echo request carries a DDMAP and the LAG Interface Info TLV (described in Section 5) to indicate that separate load balancing information for each layer 2 nexthop over LAG is desired in MPLS echo reply. Responder LSR:

- MUST add the LAG Interface Info TLV in the MPLS echo reply to provide acknowledgement back to the initiator. Downstream LAG Info Accommodation flag MUST be set in LAG Interface Info Flags.
- For each downstream that is a LAG interface:
  - MUST add DDMAP in the MPLS echo reply.
  - MUST set LAG Description Indicator flag in DS Flags (described in Section 6) of DDMAP.
  - All fields and Sub-TLVs, except for Multipath Data Sub-TLV and Interface Index Sub-TLV, are set/added to DDMAP to describe this LAG interface, as per [RFC6424].
  - For each LAG member link of this LAG interface:
    - MUST add Interface Index Sub-TLV (described in Section 7) with LAG Member Link Indicator flag set in Interface Index Flags, describing this LAG member link.
    - MUST add Multipath Data Sub-TLV for this LAG member link, if received DDMAP requested multipath information.

Each LAG member link is described with Interface Index Sub-TLV and conditionally with Multipath Data Sub-TLV (if multipath information is requested). If both Sub-TLVs are placed in the DDMAP to describe

a LAG member link, Interface Index Sub-TLV MUST be added first with Multipath Data Sub-TLV immediately following.

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

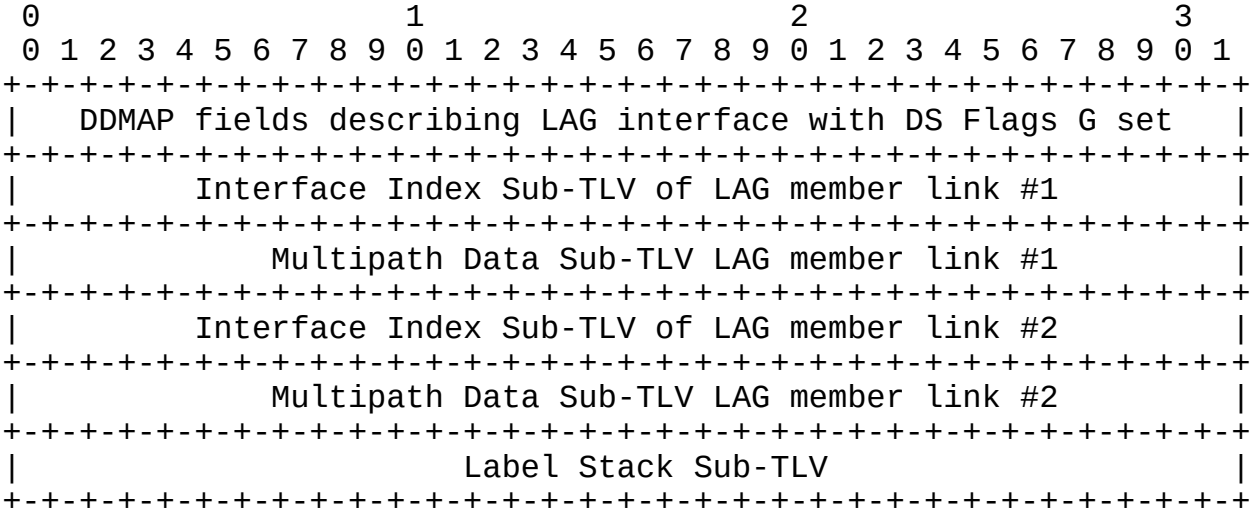


Figure 2: LAG Interface DDMAP Example

These procedures allow initiating LSR to:

- Identify whether responder LSR understands this mechanism.
- Identify whether each 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 interface index.

4. Mechanism to Validate L2 ECMP Traversal

The MPLS echo request is sent with a DDMAP with DS Flags I set and the optional LAG Interface Info TLV 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. Responder LSR MUST:

- Add LAG Interface Info TLV in the MPLS echo reply to provide acknowledgement back to the initiator. Upstream LAG Info Accommodation flag MUST be set in LAG Interface Info Flags.
- Add the Detailed Interface and Label Stack TLV (described in Section 8) in the MPLS echo reply.

- o Add the Incoming Interface Index Sub-TLV (described in Section 8.1.2) for LAG interfaces. The LAG Member Link Indicator flag MUST be set in Interface Index Flags, and the incoming Interface Index set to LAG member link which received the MPLS echo request.

Described procedures allow initiating LSR to know:

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

- o The actual incoming interface at LSR with TTL=n+1, including the interface index of LAG member link if incoming interface is a LAG interface.

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. LAG Interface Info TLV

The LAG Interface Info object is a new TLV that MAY be included in the MPLS echo request message. An MPLS echo request MUST NOT include more than one LAG Interface Info object. Presence of LAG Interface Info object is a request that responder LSR describes upstream and downstream LAG interfaces according to procedures defined in this document. If the responder LSR is able to accommodate this request, then the LAG Interface Info object MUST be included in the MPLS echo reply message.

LAG Interface Info TLV Type is TBD1. Length is 4. The Value field of LAG Interface TLV has following format:

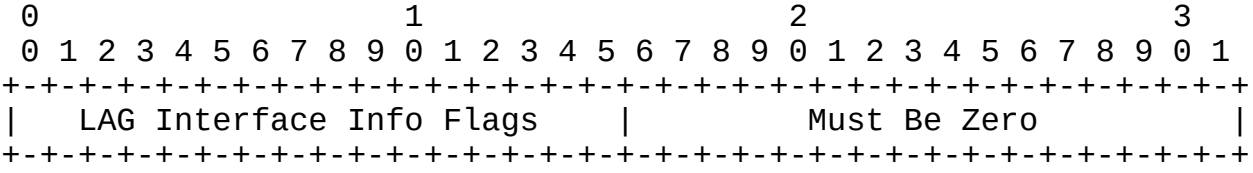
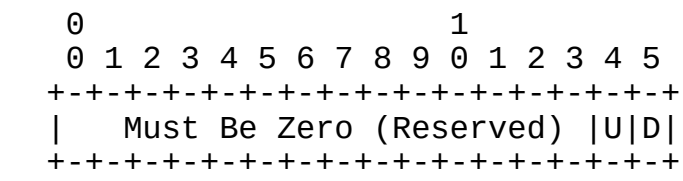


Figure 3: LAG Interface Info TLV

LAG Interface Info Flags

LAG Interface Info 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. Either or both U and D flags MAY be set in MPLS echo reply message.

Flag	Name and Meaning

U Upstream LAG Info Accommodation

When this flag is set, LSR is capable of placing Incoming Interface Index Sub-TLV, describing LAG member link, in the Detailed Interface and Label Stack TLV.

D Downstream LAG Info Accommodation

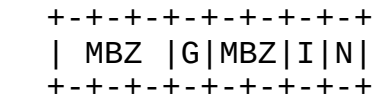
When this flag is set, LSR is capable of placing Interface Index Sub-TLV and Multipath Data Sub-TLV, describing LAG member link, in the Downstream Detailed Mapping TLV.

6. DDMAP TLV DS Flags: G

One flag, G, is added in DS Flags field of the DDMAP TLV. In the MPLS echo request message, G flag MUST be cleared when sending, and ignored on receipt. In the MPLS echo reply message, 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 3, in DS Flags bit vector.



Flag	Name and Meaning
G	LAG Description Indicator

When this flag is set, DDMAP describes a LAG interface.

7. Interface Index Sub-TLV

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

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

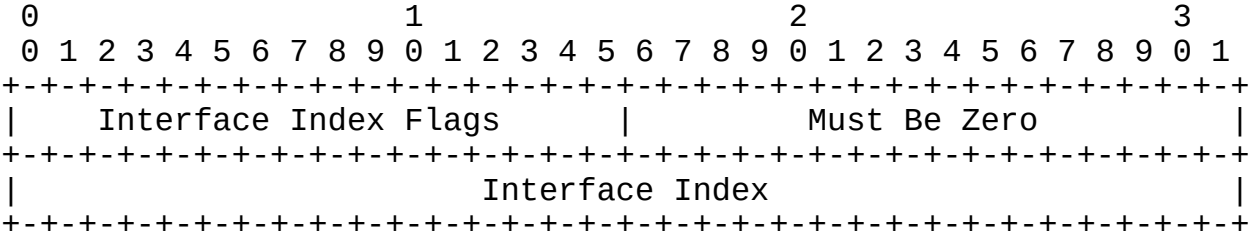
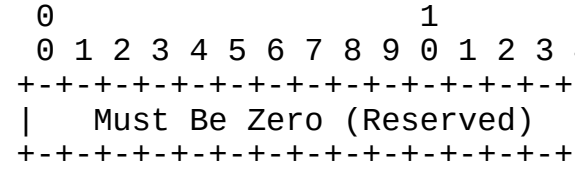


Figure 4: Interface Index Sub-TLV

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 member of a LAG.

Interface Index

Index assigned by the LSR to this interface.

8. 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 initiating 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 TBD3. Length is K + Sub-TLV Length, and the Value field has following format:

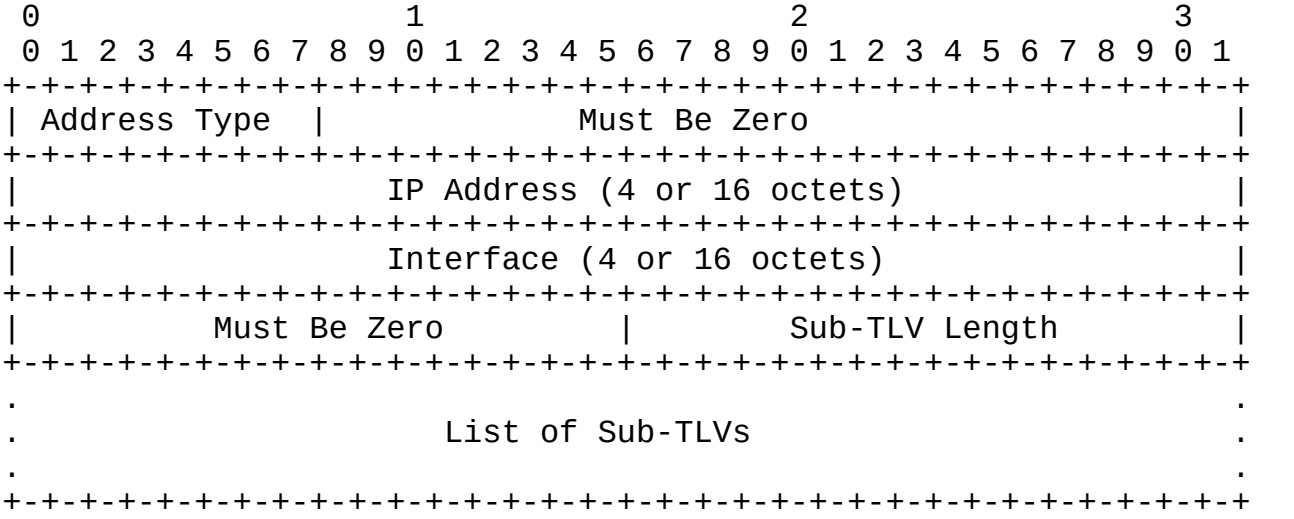


Figure 5: Detailed Interface and Label Stack TLV

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.

8.1. 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

8.1.1. 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:

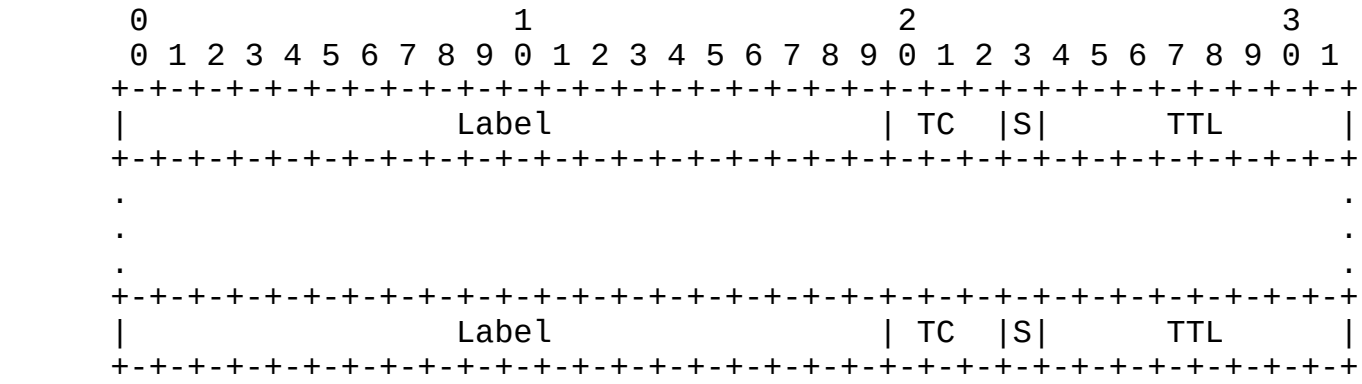


Figure 6: Incoming Label Stack Sub-TLV

8.1.2. 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 following format:

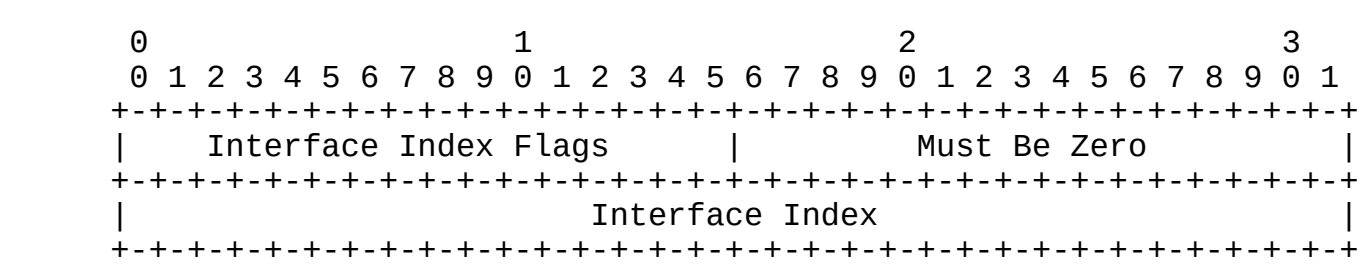
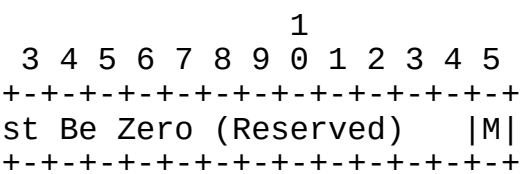


Figure 7: Incoming Interface Index Sub-TLV

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 sent and ignored on receipt.

Flag	Name and Meaning
M	LAG Member Link Indicator

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

Interface Index

Index assigned by the LSR to this interface.

9. Security Considerations

This document extends LSP Traceroute mechanism to discover and exercise layer 2 ECMP paths. Additional processing are required for initiating LSR and responder LSR, 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.

10. IANA Considerations

10.1. LAG Interface Info TLV

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

Value	Meaning	Reference
TBD1	LAG Interface Info TLV	this document

10.2. Interface Index Sub-TLV

The IANA is requested to assign new value TBD2 for 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	Interface Index Sub-TLV	this document

10.3. Detailed Interface and Label Stack TLV

The IANA is requested to assign new value TBD3 for Detailed Interface and Label Stack TLV from the "Multiprotocol Label Switching Architecture (MPLS) Label Switched Paths (LSPs) Ping Parameters - TLVs" registry.

Value	Meaning	Reference
TBD3	Detailed Interface and Label Stack TLV	this document

10.4. New Sub-Registry

10.4.1. DS Flags

[RFC4379] defines the Downstream Mapping TLV, which has the Type 2 assigned from the "Multi-Protocol Label Switching (MPLS) Label Switched Paths (LSPs) Ping Parameters - TLVs" registry. [RFC6424] defines the Downstream Detailed Mapping TLV, which has the Type 20 assigned from the "Multi-Protocol Label Switching (MPLS) Label Switched Paths (LSPs) Ping Parameters - TLVs" registry. DSMAP has been deprecated by DDMAP, but both TLVs shares a field: "DS Flags". This document requires allocation of a new value in the "DS Flags" field, which is not maintained by IANA today. Therefore, this document requests IANA to create new registries within [IANA-MPLS-LSP-PING] protocol to maintain "DS Flags" field. Initial values for this registry, "DS Flags", are described below.

Bit number	Name	Reference
7	N: Treat as a Non-IP Packet	RFC4379
6	I: Interface and Label Stack Object Request	RFC4379
5-4	Unassigned	
3	G: LAG Description Indicator	this document
2-0	Unassigned	

Assignments of DS Flags are via Standards Action [RFC5226] or IESG Approval [RFC5226].

Note that "DS Flags" is a field included in two TLVs defined in "Multi-Protocol Label Switching (MPLS) Label Switched Paths (LSPs) Ping Parameters - TLVs" registry: Downstream Mapping TLV (value 2) and Downstream Detailed Mapping TLV (value 20). Modification to "DS Flags" registry will affect both TLVs.

Also note that [I-D.akiya-mpls-entropy-lsp-ping] makes request to create a new retry for "DS Flags", with new values being added for Bit Number 4 and 5. If [I-D.akiya-mpls-entropy-lsp-ping] becomes RFC and "DS Flags" IANA registry is created as result, then this document simply requests Bit Number 3 (G: LAG Description Indicator) to be added to the registry.

10.4.2. Sub-TLVs for TLV Type TBD3

The IANA is requested to make a new "Sub-TLVs for TLV Type TBD3" sub-registry 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 TBD3", are described below.

Sub-Type	Name	Reference
1	Incoming Label Stack	this document
2	Incoming Interface Index	this document
4-65535	Unassigned	

Assignments of Sub-Types are via Standards Action [RFC5226] or IESG Approval [RFC5226].

11. Acknowledgements

TBD

12. References

12.1. Normative References

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

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

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

12.2. Informative References

[I-D.akiya-mpls-entropy-lsp-ping]
Akiya, N., Swallow, G., and C. Pignataro, "Label Switched Path (LSP) Ping/Trace over MPLS Network using Entropy Labels (EL)", draft-akiya-mpls-entropy-lsp-ping-01 (work in progress), December 2013.

[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-00 (work in progress), April 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, May 2008.

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

R1 ==== S1 ==== R2

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

```
                 ____
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
Cisco Systems

Email: nobo@cisco.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