MPLS Working Group
Internet-Draft
Intended status: Standards Track
Expires: 18 September 2024

R. Gandhi, Ed.
F. Brockners
Cisco Systems, Inc.
B. Wen
Comcast
B. Decraene
Orange
H. Song
Futurewei Technologies
17 March 2024

MPLS Network Action for Transporting In Situ OAM Data Fields
draft-gandhi-mpls-ioam-12

Abstract

In Situ Operations, Administration, and Maintenance (IOAM) is used for recording and collecting operational and telemetry information while the packet traverses a path between two points in the network. This document defines MPLS Network Action for transporting IOAM data fields.

Status of This Memo

Copyright Notice

Table of Contents

1. Introduction

In Situ Operations, Administration, and Maintenance (IOAM) is used for recording and collecting operational and telemetry information while the packet traverses a path between two points in the network. The term "in-situ" refers to the fact that the IOAM data fields are added to the data packets rather than being sent within the probe packets specifically dedicated to OAM. The IOAM data fields are defined in [RFC9197]. The IOAM data fields are further updated in [RFC9326] for direct export use-cases.

[I-D.ietf-mpls-mna-hdr] defines mechanisms for carrying MNA Sub-Stack (NAS) above the Bottom of the label stack (BOS) and [I-D.jags-mpls-ps-mna-hdr] defines mechanisms for carrying Post-Stack Network Actions in MPLS packets after the Bottom of the label stack. This document defines MPLS Network Action using these two mechanisms for transporting IOAM data fields.

2. Conventions

2.1. Requirement 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 BCP 14 [RFC2119] [RFC8174] when, and only when, they appear in all capitals, as shown here.

2.2. Abbreviations

The MPLS Network Action (MNA) terminology defined in [I-D.ietf-mpls-mna-hdr] and [I-D.jags-mpls-ps-mna-hdr] are used in this document.

Abbreviations used in this document:

ECMP Equal Cost Multi-Path

E2E Edge-To-Edge

HBH Hop-By-Hop

IHS Ingress-To-Egress (I2E), Hop-By-Hop (HBH) or Select Scope

IOAM In Situ Operations, Administration, and Maintenance

MPLS Multiprotocol Label Switching

MNA MPLS Network Action

NAI Network Action Indicator

NASL Network Action Sub-Stack Length

OAM Operations, Administration, and Maintenance

POT Proof-of-Transit

3. MPLS Extensions for IOAM Data Fields

3.1. Post-Stack Network Action for IOAM

The Post-Stack Network Action for IOAM is defined using the specification from [I-D.jags-mpls-ps-mna-hdr].

The Post-Stack Network Action encoding contains IOAM Post-Stack Network Action Opcode (value TBA5 allocated by IANA), length in number of 4-Octet units, and IOAM Option-Type with IOAM-Data-Fields in ancillary data as shown in Figure 1. The IOAM-Data-Fields MUST follow the definitions corresponding to their IOAM-Option-Types (e.g., see Section 4.4 of [RFC9197] and Section 3 of [RFC9326]).

```
 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
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+<-+
|N N N N|Version| PS-MNA-LEN    | TYPE = POST-STACK-MNA = 1      | TH
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-|<-+
|PS-NA-OP=IOAM|R|R|IOAM-HDR-LEN |R| BLOCK-NUMBER  |IOAM-OPT-TYPE|  |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+  I
|                                                               |  O
|                                                               |  A
~   IOAM Option and Data Space [RFC9197] [RFC9326]              ~  M
|                                                               |  |
|                                                               |  |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+<-+
.                                                               .
.                                                               .
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+<-+
|PS-NA-OP=IOAM|R|R|IOAM-HDR-LEN |R| BLOCK-NUMBER  |IOAM-OPT-TYPE|  |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+  I
|                                                               |  O
|                                                               |  A
~   IOAM Option and Data Space [RFC9197] [RFC9326]              ~  M
|                                                               |  |
|                                                               |  |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+<-+
|                                                               |
|                                                               |
~                 Optional Payload + Padding                    ~
|                                                               |
|                                                               |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
```

Figure 1: Post-Stack Network Action with IOAM Data Fields

The 4-Octet Top Header is added with the following fields in the the Post-Stack Network Action as defined in [I-D.jags-mpls-ps-mna-hdr]

NNNN: The first 4-bit (value TBA4) as defined in [I-D.jags-mpls-ps-mna-hdr].

PS-MNA-LEN: Length of the Post-Stack Network Action Stack in the packet in 4-Octet units including the Top Header (TH).

TYPE: TYPE is defined for POST-STACK-MNA (value 1) in [I-D.jags-mpls-ps-mna-hdr].

An IOAM-Data-Field is added in the Post-Stack Network Action containing the following fields:

IOAM-OPT-TYPE: 7-bit field defining the IOAM Option-Type, as defined in the "IOAM Option-Type Registry" specified in [RFC9197] and [RFC9326]).

IOAM-HDR-LEN: 7-bit unsigned integer. Length of the IOAM Data-Fields in 4-Octet units.

IOAM Option and Data Space: IOAM-Data-Fields as specified by the IOAM-OPT-Type field. IOAM-Data-Fields are defined corresponding to the IOAM-Option-Type (e.g., see Section 4.4 of [RFC9197] and Section 3 of [RFC9326].

BLOCK-NUMBER: The Block Number can be used to aggregate the IOAM data collected in data plane, e.g., to compute measurement metrics for each block of a data flow. It is also used to correlate the IOAM data on different nodes.

3.2. MNA Sub-Stack for IOAM Post-Stack Network Action

The P flag defined in [I-D.jags-mpls-ps-mna-hdr] is used to indicate the presence of Post-Stack Network Action and Ancillary Data.

A Post-Stack Network Action Indicator (P flag) MUST be set to "1" to indicate the presence of the Post-Stack Network Action with IOAM-Data-Fields in the Post-Stack Network Action (after the Bottom of the Stack).

The IHS scope field defined in [I-D.ietf-mpls-mna-hdr] is used to indicate that E2E or HBH or Select processing is required for the Post-Stack Network Action and Ancillary Data.

If both edge and intermediate nodes need to process the IOAM data fields then IHS scope MUST be set to "HBH, value 0x1". If only edge nodes need to process the IOAM data fields then IHS scope MUST be set to "I2E, value 0x0". The HBH scope allows to skip the IOAM data processing on the intermediate nodes i.e., avoids the need to parse all IOAM-Data-Fields to detect the HBH option type.

The MNA Label used in this document is a Base Special Purpose Label (bSPL value TBA1), which is also called Network Action Sub-Stack Indicator as allocated in [I-D.ietf-mpls-mna-hdr].

Opcode 1 is used to carry the Flag-Based Network Action Indicators (NAIs) in the Ancillary Data field [I-D.ietf-mpls-mna-hdr]. The Flag-Based Network Action Indicators is set to 0x0 to indicate that no Network Action other than the Post-Stack Network Action(s) is requested for the packet.

The U Flag for Unknown Action Handling is specified in [I-D.ietf-mpls-mna-hdr].

The Network Action Sub-Stack Length (NASL) is set to 0 if no additional Label Stack Entry (LSE) is added after the LSE carrying the P Flag.

4. Edge-To-Edge IOAM Network Action

4.1. MNA Sub-Stack for Edge-To-Edge IOAM

The IHS scope is set to "I2E, value 0x0" to indicate the scope of E2E IOAM-Data-Fields in the MPLS Sub-Stack as shown in Figure 2. Note that there may be additional Network Action (NA) LSEs added in the MNA Sub-Stack and NASL is updated with the number of LSEs after the LSE with P flag.

```
 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                                | TC  |S|  TTL          |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
.                                                               .
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|  MNA Label (bSPL value TBA1)          | TC  |S|  TTL          |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|  Opcode=1   | Flag-Based NAIs         |1|E2E|S| Res |U|NASL=0 |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
.                                                               .
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|                Packet as shown in Figure 1                    |
.                                                               .
+---------------------------------------------------------------+
```

Figure 2: Example MNA Sub-Stack for E2E IOAM

The E2E IOAM-Data-Fields carry the Option-Type(s) that require processing on the encapsulating and decapsulating nodes only. The IOAM Option-Type carried can be IOAM Edge-To-Edge Option-Type (value 3) defined in [RFC9197] as well as Direct Export (DEX) Option-Type (value 4) defined in [RFC9326]. The E2E IOAM-Data-Fields SHOULD NOT carry any IOAM Option-Type that require IOAM processing on the intermediate nodes as it will not be processed by them since IHS scope is set to "I2E, value 0x0".

4.2. Procedure for Edge-To-Edge IOAM Network Action

The E2E IOAM Network Action procedure is summarized as following:

* The encapsulating node inserts an MNA Sub-Stack with the MNA Label (bSPL value TBA1) with the Post-Stack Network Action Indicator (P flag) set to "1" and IHS scope set to "I2E, value 0x0" and one or more IOAM-Data-Fields in the Post-Stack Network Action in the MPLS packet.

* The intermediate nodes do not process IOAM-Data-Fields.

* The penultimate node MUST NOT remove the MNA Sub-Stack from the MPLS header so that the MNA Sub-Stack is received at the decapsulating node.

* The decapsulating node MAY punt the IOAM data from the packet with the receive timestamp to the slow path for IOAM-Data-Fields processing. The receive timestamp is required by the various E2E OAM use-cases, including streaming telemetry. Note that the packet is not necessarily punted to the control-plane.

* The decapsulating node processes the IOAM-Data-Fields using the procedures defined in [RFC9197] and [RFC9326]. An example of IOAM processing is to export the IOAM-Data-Fields for streaming telemetry.

* The decapsulating node MUST remove the Post-Stack Network Action with IOAM-Data-Fields from the received packet. The decapsulated packet is forwarded downstream or terminated locally similar to the regular data packets.

5. Hop-By-Hop IOAM Network Action

5.1. MNA Sub-Stack for Hop-By-Hop IOAM

The IHS scope is set to "HBH, value 0x1" to indicate the scope of HBH IOAM-Data-Fields in the MNA Sub-Stack as shown in Figure 3. Note that there may be additional Network Action (NA) LSEs added in the MNA Sub-Stack and NASL is updated based on the number of LSEs after the LSE with P flag.

```
 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                                | TC  |S|  TTL          |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
.                                                               .
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|  MNA Label (bSPL value TBA1)          | TC  |S|  TTL          |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|  Opcode=1   | Flag-Based NAIs         |1|HBH|S| Res |U|NASL=0 |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
.                                                               .
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|                Packet as shown in Figure 1                    |
.                                                               .
+---------------------------------------------------------------+
```

Figure 3: Example MNA Sub-Stack for HBH IOAM

The HBH IOAM-Data-Fields carry the Option-Type(s) that require processing at the intermediate and/or encapsulating and decapsulating nodes. The IOAM Option-Type carried can be IOAM Pre-allocated Trace Option-Type (value 0), IOAM Incremental Trace Option-Type (value 1) and IOAM Proof of Transit (POT) Option-Type (value 2), and Edge-To-Edge Option-Type (value 3) defined in [RFC9197] as well as Direct Export (DEX) Option-Type (value 4) defined in [RFC9326].

5.2. Procedure for Hop-By-Hop IOAM Network Action

The Hop-By-Hop IOAM Network Action procedure is summarized as following:

* The encapsulating node inserts an MNA Sub-Stack containing MNA Label (bSPL value TBA1) with the Post-Stack Network Action Indicator (P flag) set to "1" and IHS scope set to "HBH, value 0x1" and one or more IOAM-Data-Fields in the Post-Stack Network Action in the MPLS packet.

* The intermediate node enabled with HBH IOAM function processes the data packet including the IOAM-Data-Fields as defined in [RFC9197] and [RFC9326] when the node recognizes the HBH scope in the MNA Sub-Stack.

* The intermediate node MAY punt the IOAM data from the packet with the receive timestamp to the slow path for IOAM-Data-Fields processing when the node recognizes the HBH scope. The receive timestamp is required by the various HBH OAM use-cases, including streaming telemetry. Note that the packet is not necessarily punted to the control-plane.

* The intermediate node forwards the data packet downstream.

* The processing on the penultimate node is same as E2E case.

* The processing on the decapsulating node is same as E2E case.

5.3. Hop-By-Hop and Edge-To-Edge IOAM Network Actions

Both HBH and E2E IOAM NAs may be added in an MNA Sub-Stack. In this case, the HBH IOAM NAs MUST be added after the BOS and before the E2E IOAM NAs. When an intermediate node that processes the HBH IOAM NAs, encounters the first E2E IOAM NA, it will stop processing of the E2E NA.

6. Select IOAM Network Action

The procedure for carrying the IOAM NAs for Select Scope is the same as HBH Scope IOAM except the IHS scope is set to "Select, value 0x2". In this case, only the select nodes will process the IOAM-Data-Fields.

7. Considerations for IOAM

7.1. Considerations for ECMP

The encapsulating node needs to make sure the IOAM-Data-Fields do not start with a well-known IP Version Number (e.g. 0x4 for IPv4 and 0x6 for IPv6) as that can alter the hashing function for ECMP that uses the IP header. This is achieved by using the first nibble NNNN in the Top Header with a different value (value TBA4) after the MPLS label stack as defined in [I-D.jags-mpls-ps-mna-hdr].

7.2. Node Capability

The decapsulating node that has to remove the IOAM-Data-Fields and perform the IOAM function may not be capable of supporting it. The encapsulating node needs to know if the decapsulating node can support the IOAM function. The signaling extension for this capability exchange is outside the scope of this document.

The intermediate node that is not capable of supporting the IOAM functions defined in this document, can simply skip the IOAM processing.

The node that does not recognize the MNA Label received at the top of the label stack will drop the packet.

7.3. Nested MPLS Encapsulation

When a packet is received with MPLS Encapsulated IOAM Network Action, the nested MPLS encapsulating node that needs to add different IOAM Network Action, the node MUST add a new MNA Sub-Stack with the IOAM Network Action as part of the new MPLS encapsulation.

7.4. Readable Label Depth Consideration

The encapsulating node needs to make sure that the IOAM-Data-Fields in Post-Stack Network Action are added within the Readable Label Depth (RLD) of the downstream MNA capable nodes in order for them to be able to process the IOAM.

7.5. Post-Stack Network Action Opcodes

The Post-Stack Network Action Offset (Opcode TBA2), as well as Post-Stack and In-Stack Network Action Ordering (Opcode TBA3) both defined in [I-D.jags-mpls-ps-mna-hdr] can be used along with IOAM Post-Stack Network Action Opcode (TBA5) defined in this document as shown in Figure 4.

```
 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                                | TC  |S|  TTL          |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
.                                                               .
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|  MNA Label (bSPL value TBA1)          | TC  |S|  TTL          |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|  Opcode=TBA2| Offset in 4 octets      |R|IHS|S| Res |U|NASL=0 |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
.                                                               .
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|                Packet as shown in Figure 1                    |
.                                                               .
+---------------------------------------------------------------+
```

Figure 4: Example MNA Sub-Stack for IOAM Using Opcode TBA2

8. Security Considerations

The security considerations of IOAM in general are discussed in [RFC9197] and [RFC9326]) and apply to the procedure defined in this document.

The usage of MPLS extensions defined in this document for IOAM is intended for deployment in a single network administrative domain. As such, it assumes that the operator enabling the IOAM operation has previously verified the integrity of the path. Still, operators need to properly secure the IOAM in the domain to avoid malicious configuration and use, which could include injecting malicious IOAM packets into the domain.

9. IANA Considerations

Registry "Post-Stack Network Action Opcode" will be created by [I-D.jags-mpls-ps-mna-hdr]. The PS-NA-OPCODE for IOAM is allocated from this registry.

Value	Name	Description	Reference
TBA5	IOAM	In Situ OAM	This document

Table 1: Post-Stack Network Action Opcode

10. References

10.1. Normative References

[I-D.ietf-mpls-mna-hdr]
Rajamanickam, J., Ed., Gandhi, R., Ed., Zigler, R., Song, H., and K. Kompella, "MPLS Network Action Sub-Stack Solution", Work in Progress, Internet-Draft, draft-ietf-mpls-mna-hdr-04, October 2023, <https://www.ietf.org/archive/id/draft-ietf-mpls-mna-hdr-04.txt>.

[I-D.jags-mpls-ps-mna-hdr]
Rajamanickam, J., Ed., Gandhi, R., Ed., Zigler, R., Li, T., and J. Dong, "Post-Stack MPLS Network Action (MNA) Solution", Work in Progress, Internet-Draft, draft-jags-mpls-ps-mna-hdr-01, October 2023, <https://www.ietf.org/archive/id/draft-jags-mpls-ps-mna-hdr-01.txt>.

[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>.

[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>.

[RFC9197] Brockners, F., Bhandari, S., and T. Mizrahi, "Data Fields for In Situ Operations, Administration, and Maintenance", RFC 9197, May 2022, <https://www.rfc-editor.org/info/rfc9197>.

[RFC9326] Song, H., Gafni, B., Brockners, F., Bhandari, S., and T. Mizrahi, "In Situ Operations, Administration, and Maintenance OAM Direct Exporting", RFC 9326, November 2022, <https://www.rfc-editor.org/info/rfc9326>.

10.2. Informative References

Acknowledgements

The authors would like to thank Patrick Khordoc, Sagar Soni, Shwetha Bhandari, Clarence Filsfils, and Vengada Prasad Govindan for the discussions on IOAM. The authors would also like to thank Tarek Saad, Loa Andersson, Greg Mirsky, Stewart Bryant, Xiao Min, Jaganbabu Rajamanickam, and Cheng Li for providing many useful comments. The authors would also like to thank Mach Chen, Andrew Malis, Matthew Bocci, and Nick Delregno for the MPLS-RT reviews of the early version of this document.

Contributors

The following people have substantially contributed to this document:

Zafar Ali
Cisco Systems, Inc.
Email: zali@cisco.com

Voitek Kozak
Comcast
Email: Voitek_Kozak@comcast.com

Authors' Addresses

Rakesh Gandhi (editor)
Cisco Systems, Inc.
Canada
Email: rgandhi@cisco.com

Frank Brockners
Cisco Systems, Inc.
Hansaallee 249, 3rd Floor
DUESSELDORF, NORDRHEIN-WESTFALEN 40549
Germany
Email: fbrockne@cisco.com

Bin Wen
Comcast
Email: Bin_Wen@cable.comcast.com

Bruno Decraene
Orange
Email: bruno.decraene@orange.com

Haoyu Song
Futurewei Technologies
United States of America
Email: haoyu.song@futurewei.com

MPLS Working Group
Internet-Draft
Intended status: Standards Track
Expires: 20 October 2024

IJ. Wijnands
Individual
M. Mishra (Editor)
K. Raza
Cisco Systems, Inc.
Z. Zhang
Juniper Networks
A. Gulko
Edward Jones wealth management
18 April 2024

mLDP Extensions for Multi-Topology Routing
draft-ietf-mpls-mldp-multi-topology-05

Abstract

Multi-Topology Routing (MTR) is a technology to enable service differentiation within an IP network. Flexible Algorithm (FA) is another mechanism of creating a sub-topology within a topology using defined topology constraints and computation algorithm. In order to deploy mLDP (Multipoint label distribution protocol) in a network that supports MTR and/or FA, mLDP is required to become topology and FA aware. This document specifies extensions to mLDP to support MTR with FA such that when building a Multipoint LSPs(Label Switched Paths) it can follow a particular topology and algorithm.

Status of This Memo

Copyright Notice

Table of Contents

1. Glossary

MT - Multi-Topology

MT-ID - Multi-Topology Identifier

MTR - Multi-Topology Routing

IGP - Interior Gateway Protocol

MP - Multipoint (P2MP or MP2MP)

LDP - Label Distribution Protocol

mLDP - Multipoint LDP

P2MP - Point-to-Multipoint

MP2MP - Multipoint-to-Multipoint

FEC - Forwarding Equivalence Class

LSP - Label Switched Path

FA - Flexible Algorithm

IPA - IGP Algorithm

2. Introduction

Multi-Topology Routing (MTR) is a technology to enable service differentiation within an IP network. IGP protocols (OSPF and IS-IS) and LDP have already been extended to support MTR. To support MTR, an IGP maintains independent IP topologies, termed as "Multi-Topologies" (MT), and computes/installs routes per topology. OSPF extensions [RFC4915] and ISIS extensions [RFC5120] specify the MT extensions under respective IGPs. To support IGP MT, similar LDP extensions [RFC7307] have been specified to make LDP MT-aware and be able to setup unicast Label Switched Paths (LSPs) along IGP MT routing paths.

A more light weight mechanism to define constraint-based topologies is the Flexible Algorithm (FA) [RFC9350]. FA can be seen as creating a sub-topology within a topology using defined topology constraints and computation algorithm. This can be done within a MTR topology or the default Topology. An instance of such a sub-topology is identified by a 1 octet value as documented in [RFC9350]). Flexible Algorithm is a mechanism to create a sub-topology, but in the future different algorithms might be defined on how to achieve that. For that reason, in the remainder of this document, we'll refer to this as the IGP Algorithm (IPA).

Multipoint LDP (mLDP) refers to extensions in LDP to setup multi-point LSPs (point-to-multipoint (P2MP) or multipoint-to-multipoint (MP2MP)), by means of set of extensions and procedures defined in [RFC6388]. In order to deploy mLDP in a network that supports MTR and FA, mLDP is required to become topology and algorithm aware. This document specifies extensions to mLDP to support MTR/IPA such

that when building a Multi-Point LSPs it can follow a particular topology and alogirthm. This means that the identifier for the particular topology to be used by mLDP have to become a 2-tuple (MTR Topology Id, IGP Algorithm).

3. Specification of Requirements

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 BCP 14 [RFC2119] [RFC8174] when, and only when, they appear in all capitals, as shown here.

4. MT Scoped mLDP FECs

As defined in [RFC7307], MPLS Multi-Topology Identifier (MT-ID) is an identifier that is used to associate an LSP with a certain MTR topology. In the context of MP LSPs, this identifier is part of the mLDP FEC encoding so that LDP peers are able to setup an MP LSP via their own defined MTR policy. In order to avoid conflicting MTR policies for the same mLDP FEC, the MT-ID needs to be a part of the FEC, so that different MT-ID values will result in unique MP-LSP FEC elements.

The same applies to the IPA. The IPA needs to be encoded as part of the mLDP FEC to create unique MP-LSPs and at the same time is used to signal to mLDP (hop-by-hop) which Algorithm needs to be used to create the MP-LSP.

Since the MT-ID and IPA are part of the FEC, they apply to all the LDP messages that potentially include an mLDP FEC element.

4.1. MP FEC Extensions for MT

Following subsections propose the extensions to bind an mLDP FEC to a topology. The mLDP MT extensions reuse some of the extensions specified in [RFC7307].

4.1.1. MP FEC Element

Base mLDP specification [RFC6388] defines MP FEC Element as follows:

```
 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
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
| MP FEC type   |       Address Family          |    AF Length  |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|                Root Node Address                              |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|    Opaque Length              |       Opaque Value            |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+                               +
~                                                               ~
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
```

Figure 1: MP FEC Element Format [RFC6388]

Where "Root Node Address" encoding is as defined for given "Address Family", and whose length (in octets) is specified by the "AF Length" field.

To extend MP FEC elements for MT, the {MT-ID, IPA} is a tuple that is relevant in the context of the root address of the MP LSP. The {MT-ID, IPA} tuple determines in which (sub)-topology the root address needs to be resolved. Since the {MT-ID, IPA} tuple should be considered part of the mLDP FEC, the most natural place to encode this tuple is as part of the root address. While encoding it, we also propose to use "MT IP" Address Families as described in following sub section.

4.1.2. MT IP Address Families

[RFC7307] has specified new address families, named "MT IP" and "MT IPv6", to allow specification of an IP prefix within a topology scope. In addition to using this address family for mLDP, we also use 8 bits of the 16 bits Reserved field to encode the IGP Algorithm (IPA) Registry. The resulting format of the data associated with these new Address Families is as follows:

```
 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
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|                      IPv4 Address                             |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|    Reserved   |      IPA      |        MT-ID                  |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+


 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
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|                      IPv6 Address                             |
|                                                               |
|                                                               |
|                                                               |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|    Reserved   |      IPA      |        MT-ID                  |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
```

Figure 2: Modified MT IP Address Families Data Format

Where:

IPv4/IPv6 Address: An IP address corresponding to "MT IP" and "MT IPv6" address families respectively.

IPA: The IGP Algorithm, values are from the IGP Algorithm registry.

Reserved: This 8-bit field MUST be zero on transmission and ignored on receipt.

4.1.3. MT MP FEC Element

By using extended MT IP Address Family, the resultant MT MP FEC element is to be encoded as follows:

```
 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
 +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
 | MP FEC type   |  AF (MT IP/ MT IPv6)          |    AF Length  |
 +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
 |                       Root Node Address                       |
 +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
 |    Reserved   |      IPA      |        MT-ID                  |
 +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
 |    Opaque Length              |       Opaque Value            |
 +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+                               +
 ~                                                               ~
 +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
```

Figure 3: IP MT-Scoped MP FEC Element Format

In the context of this document, the applicable LDP FECs for MT mLDP include:

* MP FEC Elements:

  - P2MP (type 0x6)

  - MP2MP-up (type 0x7)

  - MP2MP-down (type 0x8)

* Typed Wildcard FEC Element (type 0x5)

In case of "Typed Wildcard FEC Element", the sub FEC Element type MUST be one of the MP FECs listed above.

This specification allows the use of Topology-scoped mLDP FECs in LDP label and notification messages, as applicable.

[RFC6514] defines the PMSI tunnel attribute for MVPN. When the Tunnel Type is set to mLDP P2MP LSP, the Tunnel Identifier is a P2MP FEC Element. When the Tunnel Type is set to mLDP Multipoint-to-Multipoint (MP2MP) LSP, the Tunnel Identifier is an MP2MP FEC Element. For deploying mLDP in a network that supports MTR and FA, New MT MP FEC element SHOULD be used as the Tunnel Identifier.

4.2. Topology IDs

This document assumes the same definitions and procedures associated with MPLS MT-ID as defined in [RFC7307] specification.

5. MT Multipoint Capability

"MT Multipoint Capability" is a new LDP capability, defined in accordance with LDP Capability definition guidelines [RFC5561], that is to be advertised to its peers by an mLDP speaker to announce its capability to support MTR and the procedures specified in this document. This capability MAY be sent either in an Initialization message at the session establishment time, or in a Capability message dynamically during the lifetime of a session (only if "Dynamic Announcement" capability [RFC5561] has been successfully negotiated with the peer).

The format of this capability is as follows:

```
 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
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|U|F|  MT Multipoint Cap.(IANA) |            Length             |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|S| Reserved    |
+-+-+-+-+-+-+-+-+
```

Figure 4: MT Multipoint Capability TLV Format

Where:

U- and F-bits: MUST be 1 and 0, respectively, as per Section 3 of LDP Capabilities [RFC5561].

MT Multipoint Capaility: TLV type (IANA assigned).

Length: The length (in octets) of TLV. The value of this field MUST be 1 as there is no Capability-specific data [RFC5561] that follows in the TLV.

S-bit: Set to 1 to announce and 0 to withdraw the capability (as per [RFC5561].

An mLDP speaker that has successfully advertised and negotiated "MT Multipoint" capability MUST support the following:

1. Topology-scoped mLDP FECs in LDP messages (Section 4.1)

2. Topology-scoped mLDP forwarding setup (Section 7)

6. MT Applicability on FEC-based features

6.1. Typed Wildcard MP FEC Elements

[RFC5918] extends base LDP and defines Typed Wildcard FEC Element framework. Typed Wildcard FEC element can be used in any LDP message to specify a wildcard operation for a given type of FEC.

The MT extensions proposed in document do not require any extension in procedures for Typed Wildcard FEC Element support [RFC5918], and these procedures apply as-is to Multipoint MT FEC wildcarding. Like Typed Wildcard MT Prefix FEC Element, as defined in [RFC7307], the MT extensions allow use of "MT IP" or "MT IPv6" in the Address Family field of the Typed Wildcard MP FEC element in order to use wildcard operations for MP FECs in the context of a given (sub)-topology as identified by the MT-ID and IPA field.

This document proposes following format and encoding for a Typed Wildcard MP FEC element:

```
 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
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|Typed Wcard (5)| Type = MP FEC |   Len = 6     |  AF = MT IP ..|
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|... or MT IPv6 |    Reserved   |      IPA      |     MT-ID     |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|MT ID (contd.) |
+-+-+-+-+-+-+-+-+
```

Figure 5: Typed Wildcard MT MP FEC Element

Where:

Type: One of MP FEC Element type (P2MP, MP2MPup, MP2MP-down).

MT ID: MPLS MT ID

IPA: The IGP Algorithm, values are from the IGP Algorithm registry.

The proposed format allows an LSR to perform wildcard MP FEC operations under the scope of a (sub-)topology.

6.2. End-of-LIB

[RFC5919] specifies extensions and procedures that allows an LDP speaker to signal its End-of-LIB (i.e. convergence) for a given FEC type towards a peer. MT extensions for MP FEC do not require any change in these procedures and they apply as-is to MT MP FEC elements. This means that an MT mLDP speaker MAY signal its convergence per (sub-)topology using MT Typed Wildcard MP FEC element.

7. Topology-Scoped Signaling and Forwarding

Since the {MT-ID, IPA} tuple is part of an mLDP FEC, there is no need to support the concept of multiple (sub-)topology forwarding tables in mLDP. Each MP LSP will be unique due to the tuple being part of the FEC. There is also no need to have specific label forwarding tables per topology, and each MP LSP will have its own unique local label in the table. However, In order to implement MTR in an mLDP network, the selection procedures for upstream LSR and downstream forwarding interface need to be changed.

7.1. Upstream LSR selection

The procedures as described in RFC-6388 section-2.4.1.1 depend on the best path to reach the root. When the {MT-ID, IPA} tuple is signaled as part of the FEC, this tuple is used to select the (sub-)topology that must be used to find the best path to the root address. Using the next-hop from this best path, a LDP peer is selected following the procedures as defined in [RFC6388].

7.2. Downstream forwarding interface selection

The procedures as described in RFC-6388 section-2.4.1.2 describe how a downstream forwarding interface is selected. In these procedures, any interface leading to the downstream LDP neighbor can be considered as candidate forwarding interface. When the {MT-ID, IPA} tuple is part of the FEC, this is no longer true. An interface must only be selected if it is part of the same (sub-)topology that was signaled in the mLDP FEC element. Besides this restriction, the other procedures in [RFC6388] apply.

8. LSP Ping Extensions

[RFC6425] defines procedures to detect data plane failures in Multipoint MPLS LSPs. Section 3.1.2 of [RFC6425] defines new Sub-Types and Sub-TLVs for Multipoint LDP FECs to be sent in "Target FEC Stack" TLV of an MPLS echo request message [RFC8029].

To support LSP ping for MT Multipoint LSPs, this document uses existing sub-types "P2MP LDP FEC Stack" and "MP2MP LDP FEC Stack" defined in [RFC6425]. The proposed extension is to specify "MT IP" or "MT IPv6" in the "Address Family" field, set the "Address Length" field to 8 (for MT IP) or 20 (for MT IPv6), and encode the sub-TLV with additional {MT-ID, IPA} information as an extension to the "Root LSR Address" field. The resultant format of sub-tlv is as follows:

```
 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
 +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
 |Address Family (MT IP/MT IPv6) | Address Length|               |
 +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+               |
 ~                   Root LSR Address (Cont.)                    ~
 |                                                               |
 +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
 |    Reserved   |      IPA      |        MT-ID                  |
 +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
 |        Opaque Length          |         Opaque Value ...      |
 +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+                               +
 ~                                                               ~
 |                               +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
 |                               |
 +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
```

Figure 6: Multipoint LDP FEC Stack Sub-TLV Format for MT

The rules and procedures of using this new sub-TLV in an MPLS echo request message are same as defined for P2MP/MP2MP LDP FEC Stack Sub-TLV in [RFC6425] with only difference being that Root LSR address is now (sub-)topology scoped.

9. Implementation Status

[Note to the RFC Editor - remove this section before publication, as well as remove the reference to [RFC7942]

This section records the status of known implementations of the protocol defined by this specification at the time of posting of this Internet-Draft, and is based on a proposal described in [RFC7942] . The description of implementations in this section is intended to assist the IETF in its decision processes in progressing drafts to RFCs. Please note that the listing of any individual implementation here does not imply endorsement by the IETF. Furthermore, no effort has been spent to verify the information presented here that was supplied by IETF contributors. This is not intended as, and must not be construed to be, a catalog of available implementations or their features. Readers are advised to note that other implementations may exist.

According to [RFC7942] , "this will allow reviewers and working groups to assign due consideration to documents that have the benefit of running code, which may serve as evidence of valuable experimentation and feedback that have made the implemented protocols more mature. It is up to the individual working groups to use this information as they see fit".

9.1. Cisco Systems

The feature has been implemented on IOS-XR.

* Organization: Cisco Systems

* Implementation: Cisco systems IOS-XR has an implementation

* Description: The implementation has been done.

* Maturity Level: Product

* Contact: mankamis@cisco.com

10. Security Considerations

This extension to mLDP does not introduce any new security considerations beyond that already apply to the base LDP specification [RFC5036], base mLDP specification [RFC6388], and MPLS security framework [RFC5920].

11. IANA Considerations

This document defines a new LDP capability parameter TLV. IANA is requested to assign the lowest available value after 0x0500 from "TLV Type Name Space" in the "Label Distribution Protocol (LDP) Parameters" registry within "Label Distribution Protocol (LDP) Name Spaces" as the new code point for the LDP TLV code point.

Value	Description	Reference	Notes/Registration Date
TBA	MT Multipoint Capability	This document	

Figure 7: IANA Code Point

12. Contributor

Anuj Budhiraja Cisco systems

13. Acknowledgments

The authors would like to acknowledge Eric Rosen for his input on this specification.

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>.

[RFC4915] Psenak, P., Mirtorabi, S., Roy, A., Nguyen, L., and P. Pillay-Esnault, "Multi-Topology (MT) Routing in OSPF", RFC 4915, DOI 10.17487/RFC4915, June 2007, <https://www.rfc-editor.org/info/rfc4915>.

[RFC5120] Przygienda, T., Shen, N., and N. Sheth, "M-ISIS: Multi Topology (MT) Routing in Intermediate System to Intermediate Systems (IS-ISs)", RFC 5120, DOI 10.17487/RFC5120, February 2008, <https://www.rfc-editor.org/info/rfc5120>.

[RFC6388] Wijnands, IJ., Ed., Minei, I., Ed., Kompella, K., and B. Thomas, "Label Distribution Protocol Extensions for Point-to-Multipoint and Multipoint-to-Multipoint Label Switched Paths", RFC 6388, DOI 10.17487/RFC6388, November 2011, <https://www.rfc-editor.org/info/rfc6388>.

[RFC6425] Saxena, S., Ed., Swallow, G., Ali, Z., Farrel, A., Yasukawa, S., and T. Nadeau, "Detecting Data-Plane Failures in Point-to-Multipoint MPLS - Extensions to LSP Ping", RFC 6425, DOI 10.17487/RFC6425, November 2011, <https://www.rfc-editor.org/info/rfc6425>.

[RFC6514] Aggarwal, R., Rosen, E., Morin, T., and Y. Rekhter, "BGP Encodings and Procedures for Multicast in MPLS/BGP IP VPNs", RFC 6514, DOI 10.17487/RFC6514, February 2012, <https://www.rfc-editor.org/info/rfc6514>.

[RFC7307] Zhao, Q., Raza, K., Zhou, C., Fang, L., Li, L., and D. King, "LDP Extensions for Multi-Topology", RFC 7307, DOI 10.17487/RFC7307, July 2014, <https://www.rfc-editor.org/info/rfc7307>.

[RFC7942] Sheffer, Y. and A. Farrel, "Improving Awareness of Running Code: The Implementation Status Section", BCP 205, RFC 7942, DOI 10.17487/RFC7942, July 2016, <https://www.rfc-editor.org/info/rfc7942>.

[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>.

[RFC9350] Psenak, P., Ed., Hegde, S., Filsfils, C., Talaulikar, K., and A. Gulko, "IGP Flexible Algorithm", RFC 9350, DOI 10.17487/RFC9350, February 2023, <https://www.rfc-editor.org/info/rfc9350>.

14.2. Informative References

[RFC5036] Andersson, L., Ed., Minei, I., Ed., and B. Thomas, Ed., "LDP Specification", RFC 5036, DOI 10.17487/RFC5036, October 2007, <https://www.rfc-editor.org/info/rfc5036>.

[RFC5561] Thomas, B., Raza, K., Aggarwal, S., Aggarwal, R., and JL. Le Roux, "LDP Capabilities", RFC 5561, DOI 10.17487/RFC5561, July 2009, <https://www.rfc-editor.org/info/rfc5561>.

[RFC5918] Asati, R., Minei, I., and B. Thomas, "Label Distribution Protocol (LDP) 'Typed Wildcard' Forward Equivalence Class (FEC)", RFC 5918, DOI 10.17487/RFC5918, August 2010, <https://www.rfc-editor.org/info/rfc5918>.

[RFC5919] Asati, R., Mohapatra, P., Chen, E., and B. Thomas, "Signaling LDP Label Advertisement Completion", RFC 5919, DOI 10.17487/RFC5919, August 2010, <https://www.rfc-editor.org/info/rfc5919>.

[RFC5920] Fang, L., Ed., "Security Framework for MPLS and GMPLS Networks", RFC 5920, DOI 10.17487/RFC5920, July 2010, <https://www.rfc-editor.org/info/rfc5920>.

Authors' Addresses

IJsbrand Wijnands
Individual
Email: ice@braindump.be

Mankamana Mishra
Cisco Systems, Inc.
821 Alder Drive
Milpitas, CA 95035
United States of America
Email: mankamis@cisco.com

Kamran Raza
Cisco Systems, Inc.
2000 Innovation Drive
Kanata ON K2K-3E8
Canada
Email: skraza@cisco.com

Zhaohui Zhang
Juniper Networks
10 Technology Park Dr.
Westford, MA 01886
United States of America
Email: zzhang@juniper.net

Arkadiy Gulko
Edward Jones wealth management
United States of America

Email: Arkadiy.gulko@edwardjones.com

MPLS WG
Internet-Draft
Updates: 7506, 8029 (if approved)
Intended status: Standards Track
Expires: 12 June 2023

K. Kompella
R. Bonica
Juniper Networks
G. Mirsky
Ericsson

9 December 2022

Deprecating the Use of Router Alert in LSP Ping
draft-kompella-mpls-lspping-norao-02

Abstract

LSP ping messages (RFC 8029) are encapsulated in IP headers that include a Router Alert Option (RAO). The rationale for including an RAO is questionable. Furthermore, RFC6398 identifies security vulnerabilities associated with the RAO.

Therefore, this document removes the RAO from LSP ping message encapsulations. It updates RFCs 7506 and 8029.

This document also recommends the use of an IPv6 loopback address (::1/128) and discourages the use of an IPv4 loopback address mapped to IPv6.

Status of This Memo

Copyright Notice

Table of Contents

1. Introduction

LSP ping [RFC8029] detects data-plane failures in MPLS Label Switched Paths (LSPs). It can operate in "ping mode" or "traceroute mode". When operating in ping mode, it verifies end-to-end LSP continuity. When operating in traceroute mode, it can localize failures to a particular node along an LSP.

LSP ping defines a probe message, called the "MPLS echo request". It also defines a response message, called the "MPLS echo reply". Both messages are encapsulated in UDP and IP. The echo request message is further encapsulated in an MPLS label stack.

When operating in ping mode, LSP ping sends a single echo request message, with the MPLS TTL set to a high value (e.g., 255). This message is intended to reach the egress Label Switching Router (LSR). When operating in traceroute mode, MPLS ping sends multiple echo request messages. It manipulates the MPLS TTL so that the first message expires on the first LSR along the path and subsequent messages expire on subsequent LSRs.

The IP header that encapsulates an echo request message must include a Router Alert Option (RAO), while the IP header that encapsulates an echo reply message may include an RAO. In both cases, the rationale

for including an RAO is questionable. Furthermore, [RFC6398] identifies security vulnerabilities associated with the RAO and recommends against its use outside of controlled environments.

Therefore, this document removes the RAO from both LSP ping message encapsulations. It updates RFCs 7506 [RFC7506] and 8029.

1.1. Terminology

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 BCP 14 [RFC2119] [RFC8174] when, and only when, they appear in all capitals, as shown here.

LSP: Label Switched Path

LSR: Label Switching Router

RAO: Router Alert Option

2. Router Alert for LSP Ping (RFC 8029)

2.1. Echo Request

While the MPLS echo request message must traverse every node in the LSP under test, it must not traverse any other node. Specifically, the message must not be forwarded beyond the egress Label Switching Router (LSR).

To achieve this, RFC 8029 proposes the following:

1. When the echo request message is encapsulated in IPv4, the IPv4 destination address must be chosen from the subnet 127/8. When the echo request message is encapsulated in IPv6, the IPv6 destination address must be chosen from the subnet 0:0:0:0:0:FFFF:7F00:0/104.

2. When the echo request message is encapsulated in IPv4, the IPv4 TTL must be equal to 1. When the echo request message is encapsulated in IPv6, the IPv6 Hop Limit must be equal to 1.

3. When the echo request message is encapsulated in IPv4, the IPv4 header must include an RAO. When the echo request message is encapsulated in IPv6, the IPv6 header chain must include a Hop-by-hop extension header and the Hop-by-hop extension header must include an RAO.

Currently, ALL of these are required. However, any one is sufficient to prevent forwarding the packet beyond the egress LSR.

Therefore, this document changes RFC 8029 in that Requirement 3 is removed.

The authors are not aware of any implementation that relies on the RAO to prevent packets from being forwarded beyond the egress LSR.

2.2. Echo Reply

An LSP ping replies to the MPLS echo message with an MPLS echo reply message. It has four reply modes:

1. Do not reply

2. Reply via an IPv4/IPv6 UDP packet

3. Reply via an IPv4/IPv6 UDP packet with Router Alert

4. Reply via application-level control channel

The rationale for mode 3 is questionable, if not wholly misguided. According to RFC 8029, "If the normal IP return path is deemed unreliable, one may use 3 (Reply via an IPv4/IPv6 UDP packet with Router Alert)."

However, it is not clear that the use of the RAO increases the reliability of the return path. In fact, one can argue it decreases the reliability in many instances, due to the additional burden of processing the RAO. This document changes RGC 8020 in that mode 3 are removed.

The authors are not aware of any implementations of mode 3.

3. Update to RFC 7506

RFC 7506 defines the IPv6 Router Alert Option for MPLS Operations, Administration, and Management. This document reclassifies RFC 7506 as Historic.

4. Update to RFC 8029

[RFC8029] requires that the IPv6 Destination Address used in IP/UDP encapsulation of an echo request packet is selected from the IPv4 loopback address range mapped to IPv6. Such packets do not have the same behavior as prescribed in[RFC1122] for an IPv4 loopback addressed packet.

[RFC4291] defines ::1/128 as the single IPv6 loopback address. Considering that this specification updates section 2.1 of [RFC8029] regarding the selection of an IPv6 destination address for an echo request message:

* For IPv6, the IPv6 loopback address ::1/128 SHOULD be used.

* The sender of an echo request MAY select the IPv6 destination address from the 0:0:0:0:0:FFFF:7F00/104 range.

* To exercise all paths in an ECMP environment, the entropy other than the IP destination address SHOULD be used.

LSP Ping implementations SHOULD ignore RAO options when they arrive on incoming echo request and echo reply messages.

5. Backwards Compatibility

LSP Ping implementations SHOULD ignore RAO options when they arrive on incoming echo request and echo reply messages.

This document requests that the IPv6 RAO value for MPLS OAM (69) in [IANA-IPV6-RAO] is marked as "Deprecated". It also requests tha that Reply Mode 3 ("Reply via an IPv4/IPv6 UDP packet with Router Alert") in [IANA-LSP-PING] is marked as "Deprecated".

We interpret "DEPRECATED" in this context to mean that the deprecated values should not be used in new implementations, and that deployed implementations that use these values continue to work seamlessly.

6. IANA Considerations

If this document is approved, mark the IPv6 RAO value of MPLS OAM (69) in [IANA-IPV6-RAO] as "Deprecated". [RFC8126] offers a formal description of the word "Deprecated".

Also, mark Reply Mode 3 ("Reply via an IPv4/IPv6 UDP packet with Router Alert") in [IANA-LSP-PING] as "Deprecated".

7. Security Considerations

The recommendations this document makes do not compromise security.

8. Normative References

[IANA-IPV6-RAO]
IANA, "IPv6 Router Alert Option Values", n.d., <https://www.iana.org/assignments/ipv6-routeralert-values>.

[IANA-LSP-PING]
IANA, "Multiprotocol Label Switching (MPLS) Label Switched Paths (LSPs) Ping Parameters", n.d., <https://www.iana.org/assignments/mpls-lsp-ping-parameters/mpls-lsp-ping-parameters.xml>.

[RFC1122] Braden, R., Ed., "Requirements for Internet Hosts - Communication Layers", STD 3, RFC 1122, DOI 10.17487/RFC1122, October 1989, <https://www.rfc-editor.org/info/rfc1122>.

[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>.

[RFC4291] Hinden, R. and S. Deering, "IP Version 6 Addressing Architecture", RFC 4291, DOI 10.17487/RFC4291, February 2006, <https://www.rfc-editor.org/info/rfc4291>.

[RFC6398] Le Faucheur, F., Ed., "IP Router Alert Considerations and Usage", BCP 168, RFC 6398, DOI 10.17487/RFC6398, October 2011, <https://www.rfc-editor.org/info/rfc6398>.

[RFC7506] Raza, K., Akiya, N., and C. Pignataro, "IPv6 Router Alert Option for MPLS Operations, Administration, and Maintenance (OAM)", RFC 7506, DOI 10.17487/RFC7506, April 2015, <https://www.rfc-editor.org/info/rfc7506>.

[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>.

[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>.

[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>.

Authors' Addresses

Kireeti Kompella
Juniper Networks
1133 Innovation Way
Sunnyvale, CA 94089
United States
Email: kireeti.ietf@gmail.com

Ron Bonica
Juniper Networks
1133 Innovation Way
Sunnyvale, CA 94089
United States
Email: rbonica@juniper.net

Greg Mirsky
Ericsson
Email: gregimirsky@gmail.com

Network Working Group
Internet-Draft
Intended status: Standards Track
Expires: 12 January 2023

L. Li
J. Dong
S. Chen
Huawei

11 July 2022

Genralized IPv6 Tunnel for MPLS
draft-li-mpls-gip6-mpls-00

Abstract

With the development of new services, MPLS faces many problems and technical challenges. This document defines the method to implement MPLS through the GIP6 tunnel.

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 [RFC2119].

Status of This Memo

Copyright Notice

Table of Contents

1. Introduction

The GIP6 tunnel is defined in [draft-li-rtgwg-generic-ipv6-tunnel-00] which is to solve the challenges of the existing IP tunnels to support new features such as alternate marking, IOAM, network resource partition and APN.

With the development of new services, MPLS also faces many problems and technical challenges. For example, it is difficult to encapsulate new forwarding attributes because of lack of the metadata extensibility. This document defines how to implement MPLS through the GIP6 tunnel which is to solve the possible problems effectively.

2. Terminology

APN: Application-Aware Networking

GIP6: Generic Ipv6 Tunnel

GRE: Generic Routing Encapsulation

IFIT: In-situ Flow Information Telemetry

MP2P: Multi Point To Point

MPLS: Multiprotocol Label Switching

PM: Performance Monitor

SFL: Synonymos Flow Labels

SR-MPLS: Segment Routing Multiprotocol Label Switching

VXLAN: Virtual eXtensible Local Area Network

3. Problem Statement

With the development of new services, MPLS faces the following the technical problems and challenges of MPLS:

1. MPLS is lack of the source indication and MP2P connections may occur. This causes the difficulty and complex process for OAM over MPLS. Although SFL([RFC8957]) is defined, there is few implementation.

2. The payload type (for example, L2 or L3 packets) cannot be directly determined because there is no payload indication.

3. There is no metadata extensibility and it is difficult to encapsulate new forwarding attributes for the new features such as IETF network slicing, IFIT, and APN.

4. The process of the ECMP function is complex and affects forwarding performance. Entropy labels or flow labels are placed at the bottom of the label stack for processing and the internal IP header information may have to be parsed for the purpose of ECMP.

4. GIP6 Tunnel for MPLS

[I-D.li-rtgwg-generalized-ipv6-tunnel] defines the GIP6 tunnel to support both new features and the existing functions for the IP tunnels based on the extension of the IPv6 extension header. If the GIP6 tunnel is used for MPLS, there can be the following advantages:

1. The IPv6 source address is used to form a source identifier.

2. The IPv6 NH can indicate the payload type.

3. IPv6 flow labels are used to implement ECMP.

4. The encapsulations for the new features have been defined well in the IPv6 and can be reused easily.

It is simple and can benefit forwarding performance. Moreover, there have been many implementations and deployments.

In order to support MPLS based on the GIP6 tunnel, the method to carry MPLS label stack information is defined as follows:

1. A special IPv6 prefix MUST be used to indicate that it is followed by MPLS label encapsulation. The special IPv6 prefix can be specified or a well-known IPv6 prefix to be assigned.

2. The IPv6 special prefix can be followed by multiple MPLS label encapsulations to form a 128-bit IPv6 MPLS SID (Type 1). The format of the IPv6 MPLS SID (Type 1) is shown in the following figure.

```
 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
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|               Special Prefix(4 octets)                        |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|               MPLS Label Encap (4 octets)                     |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|               MPLS Label Encap (4 octets)                     |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|               MPLS Label Encap (4 octets)                     |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
```

Figure 1. IPv6 MPLS SID (TYPE 1)

Special prefix: TBD

MPLS Label Encap: For details, please refer to section 2.1 in [RFC3032].

3. IPv6 MPLS SID (Type 1) can be placed in the IPv6 destination address.

Processing of the first label following the special prefix is as follows:

(1) If the local action of the MPLS label is POP, the followed label encapsulations are shifted left by 32 bits after the label is popped. The following figure shows the process.

Before POP MPLS Label Encap:

```
     uSID-Block     Active-Label      Next-Label       Last-Label
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
| Special Prefix | Lable-1 Encap | Label-2 Encap| Label-3 Encap(EOL)|
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
```

After POP MPLS Label Encap:

```
     uSID-Block     Active-Label      Next-Label       Last-Label
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
| Special Prefix | Lable-2 Encap | Label-3 Encap(EOL) | 0000...0000 |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
```

Figure 2. Pop MPLS Label Encapsulation in IPv6 DA

(2) If the local action of the MPLS label is SWAP, the label encapsulation is changed to the new label after swap.

4. If all the MPLS label stack cannot be placed in the IPv6 destination address, IPv6 RH can be used to house the remaining MPLS label stack.

(1) IPv6 MPLS SID (Type 2) is defined to house multiple (<= 4) label encapsulations. The format of the IPv6 MPLS SID (Type 2) is shown in the following figure.

```
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
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|               MPLS Label Encap (4 octets)                   |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|               MPLS Label Encap (4 octets)                   |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|               MPLS Label Encap (4 octets)                   |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|               MPLS Label Encap (4 octets)                   |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
```

MPLS Label Encap: For details, see section 2.1 in [RFC3032].

(2) IPv6 MPLS SID (Type 2) is used as the segment in the RH. After all of the label encapsulations in the IPv6 destination address are popped, the first label encapsulation in the segment indicated by the SL of the RH will be processed as follows:

-- If the local action of the MPLS label is POP, the followed label encapsulations are shifted left by 32 bits after the label is popped.

-- If the local action of the MPLS label is SWAP, the label encapsulation is changed to the new label after swap.

After all the label encapsulations in the segment are popped, SL minus 1. Then the first label encapsulation in the segment indicted by the new SL will go on to be processed as the above procedures.

A new type of RH can be defined to contain IPv6 MPLS SID (Type 2) or SRv6 SRH can be reused for the purpose.

5. When find the S flag of the label encapsulation in the IPv6 destination address or the RH to be processed is set, this means the bottom of the label stack is reached and the process of the label stack in the GIP6 will end after the label is popped.

If an intermediate node requires to push a label or a label stack, there can be two modes: Encap mode and Inserting mode.

1) Encap mode: with this mode, a new IPv6 MPLS packet header is encapsulated outside the original MPLS packet, and the MPLS label (stack) is encapsulated in the new IPv6 MPLS packet header as the above procedures.

2) Inserting mode: All the label encapsulations in the IPv6 destination address and the IPv6 RH (if exist) need to be shifted right and the new label (stack) can be placed immediately following the special prefix in the IPv6 destination address. The process is complex and not recommended.

5. Control Plane Considerations

GIP6 only provides a way to carry MPLS label encapsulations in the data plane. The existing MPLS control plane does not need to be changed. That is, MPLS labels on the control plane can still be distributed for IPv4, IPv6, L2, etc.

6. Security Considerations

TBD.

7. IANA Considerations

TBD.

8. References

[I-D.li-rtgwg-generalized-ipv6-tunnel]
Zhenbin, Shuanglong, and Haibo, "Generalized IPv6 Tunnel (GIP6)", Work in Progress, Internet-Draft, draft-li-rtgwg-generalized-ipv6-tunnel-00, 10 July 2022, <https://www.ietf.org/archive/id/draft-li-rtgwg-generalized-ipv6-tunnel-00.txt>.

[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>.

[RFC3032] Rosen, E., Tappan, D., Fedorkow, G., Rekhter, Y., Farinacci, D., Li, T., and A. Conta, "MPLS Label Stack Encoding", RFC 3032, DOI 10.17487/RFC3032, January 2001, <https://www.rfc-editor.org/info/rfc3032>.

[RFC8957] Bryant, S., Chen, M., Swallow, G., Sivabalan, S., and G. Mirsky, "Synonymous Flow Label Framework", RFC 8957, DOI 10.17487/RFC8957, January 2021, <https://www.rfc-editor.org/info/rfc8957>.

Authors' Addresses

Zhenbin Li
Huawei
156 Beiqing Road
Beijing, 100095
P.R. China
Email: lizhenbin@huawei.com

Jie Dong
Huawei
156 Beiqing Road
Beijing,100095
P.R. China
Email: jie.dong@huawei.com

Shuanglong Chen
Huawei
156 Beiqing Road
Beijing,100095
P.R. China
Email: chenshuanglong@huawei.com