

MPLS Working Group
Internet-Draft
Intended status: Standards Track
Expires: March 17, 2021

R. Gandhi, Ed.
Z. Ali
C. Filsfils
F. Brockners
Cisco Systems, Inc.
B. Wen
V. Kozak
Comcast
September 13, 2020

MPLS Data Plane Encapsulation for In-situ OAM Data
draft-gandhi-mpls-ioam-sr-03

Abstract

In-situ Operations, Administration, and Maintenance (IOAM) records operational and telemetry information in the data packet while the packet traverses a path between two nodes in the network. This document defines how IOAM data fields are transported using the MPLS data plane encapsulation, including Segment Routing (SR) with MPLS data plane (SR-MPLS).

Status of This Memo

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

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

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 March 17, 2021.

Copyright Notice

Copyright (c) 2020 IETF Trust and the persons identified as the document authors. All rights reserved.

This document is subject to BCP 78 and the IETF Trust's Legal Provisions Relating to IETF Documents (<https://trustee.ietf.org/license-info>) in effect on the date of

publication of this document. Please review these documents carefully, as they describe your rights and restrictions with respect to this document. Code Components extracted from this document must include Simplified BSD License text as described in Section 4.e of the Trust Legal Provisions and are provided without warranty as described in the Simplified BSD License.

Table of Contents

1. Introduction	2
2. Conventions	3
2.1. Requirement Language	3
2.2. Abbreviations	3
3. IOAM Data Field Encapsulation in MPLS Header	3
3.1. Indicator Labels	6
4. Procedure for Edge-to-Edge IOAM	6
4.1. Edge-to-Edge IOAM Indicator Label Allocation	7
5. Procedure for Hop-by-Hop IOAM	7
5.1. Hop-by-Hop IOAM Indicator Label Allocation	8
6. Considerations for ECMP	8
7. Node Capability	9
8. Data Packets with SR-MPLS Header	9
9. Security Considerations	10
10. IANA Considerations	10
11. References	10
11.1. Normative References	10
11.2. Informative References	11
Acknowledgements	12
Contributors	12
Authors' Addresses	12

1. Introduction

In-situ Operations, Administration, and Maintenance (IOAM) records operational and telemetry information within the packet while the packet traverses a particular network domain. 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 or Performance Measurement (PM). The IOAM data fields are defined in [I-D.ietf-ippm-ioam-data], and can be used for various use-cases for OAM and PM. The IOAM data fields are further updated in [I-D.ietf-ippm-ioam-direct-export] for direct export use-cases and in [I-D.ietf-ippm-ioam-flags] for Loopback and Active flags.

This document defines how IOAM data fields are transported using the MPLS data plane encapsulations, including Segment Routing (SR) with MPLS data plane (SR-MPLS).

2. Conventions

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

2.2. Abbreviations

Abbreviations used in this document:

ECMP	Equal Cost Multi-Path
IOAM	In-situ Operations, Administration, and Maintenance
MPLS	Multiprotocol Label Switching
OAM	Operations, Administration, and Maintenance
PM	Performance Measurement
POT	Proof-of-Transit
PSID	Path Segment Identifier
SR	Segment Routing
SR-MPLS	Segment Routing with MPLS Data plane

3. IOAM Data Field Encapsulation in MPLS Header

The IOAM data fields defined in [I-D.ietf-ippm-ioam-data] are used. IOAM data fields are carried in the MPLS header as shown in Figure 1 and Figure 2. More than one trace options can be present in the IOAM data fields. The Indicator Label is added at the bottom of the MPLS label stack (S flag set to 1) to indicate the presence of the IOAM data field(s) in the MPLS header.

The data packets with IOAM data fields carry only one Indicator Label in the MPLS header. Any intermediate node that adds additional MPLS encapsulation in the MPLS header may further update the IOAM data fields in the header without inserting another Indicator Label.

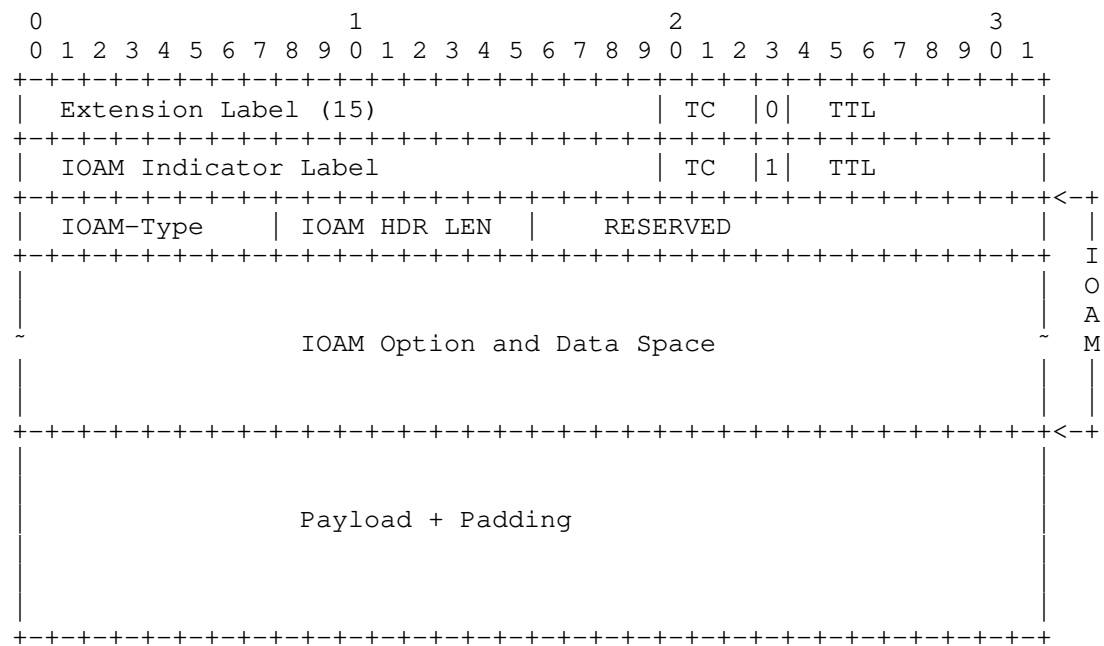


Figure 1: IOAM Encapsulation in MPLS Header

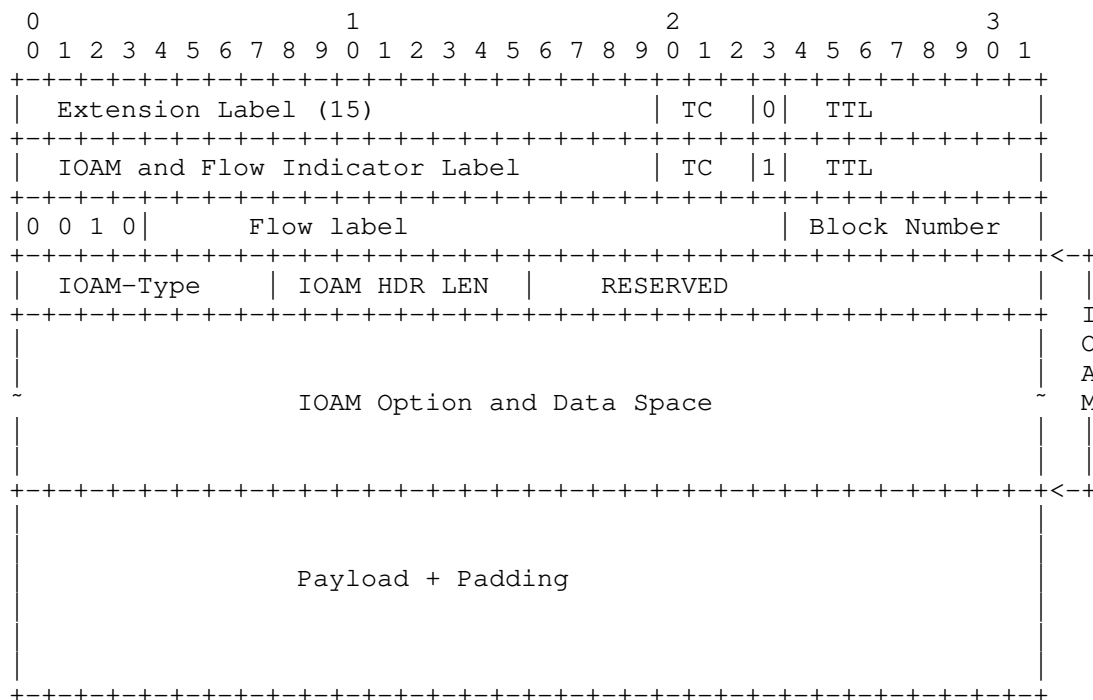


Figure 2: IOAM Encapsulation with Flow Label in MPLS Header

IOAM Indicator Label (IIL) and IOAM and Flow Indicator Label (IFIL) used are defined in this document.

The fields related to the encapsulation of IOAM data fields in the MPLS header are defined as follows:

IOAM-Type: 8-bit field defining the IOAM Option type, as defined in Section 7.2 of [I-D.ietf-ippm-ioam-data].

IOAM HDR LEN: 8-bit unsigned integer. Length of the IOAM HDR in 4-octet units.

RESERVED: 8-bit reserved field MUST be set to zero upon transmission and ignored upon receipt.

IOAM Option and Data Space: IOAM option header and data is present as defined by the IOAM-Type field, and is defined in Section 4 of [I-D.ietf-ippm-ioam-data].

3.1. Indicator Labels

IOAM Indicator Label (value TBA1 or TBA3) and IOAM and Flow Indicator Label (value TBA2 or TBA4) are used to indicate the presence of the IOAM data field in the MPLS header.

The IOAM and Flow Indicator Label (value TBA2 or TBA4) is used to carry a second label underneath with protocol value 0010b, 20-bit Flow Label and 8-bit Block Number.

- o The protocol value 0010b allows to avoid incorrect IP header-based hashing over ECMP paths that uses the value 0x4 (for IPv4) and value 0x6 (for IPv6) [RFC4928].
- o The Flow Label identifies the traffic flow that can be used for IOAM purpose, e.g. monitoring a specific traffic flow for latency.
- o The Block Number can be used to aggregate the IOAM data collected in data plane, e.g. compute measurement metrics for each block of a flow. It is also used to correlate the IOAM data on different nodes.

Different Indicator Labels are used for E2E and HbH IOAM to optimize processing on transit nodes and for checking if IOAM data fields need to be processed. If only edge nodes need to process IOAM data then E2E Indicator Label is used so that transit nodes can ignore it. If both edge and transit nodes need to process IOAM data then HbH Indicator Label is used.

The SR path computation needs to know the Maximum SID Depth (MSD) that can be imposed at each node/link of a given SR path [RFC8664]. This ensures that the SID stack depth of a computed path does not exceed the number of SIDs the node is capable of imposing. The MSD used for path computation MUST include the Indicator Labels.

4. Procedure for Edge-to-Edge IOAM

The Edge-to-Edge (E2E) IOAM includes IOAM Option-Type as Edge-to-Edge Option-Type [I-D.ietf-ippm-ioam-data]. This section summarizes the procedure for data encapsulation and decapsulation for Edge-to-Edge IOAM in MPLS header.

- o The encapsulating node inserts the IOAM Indicator Label or IOAM Flow Indicator Label with Flow Label and one or more IOAM data field(s) in the MPLS header. The procedure to generate the Flow Label is outside the scope of this document.

- o The decapsulating node "forwards and punts the timestamped copy" of the data packet including IOAM data fields when the node recognizes the IOAM Indicator Label and IOAM Flow Indicator Label. The copy of the data packet is punted to the slow path for OAM processing and is not necessarily punted to the control-plane. The receive timestamp is required by various E2E OAM use-cases.
- o The decapsulating node processes the IOAM data field(s) using the procedures defined in [I-D.ietf-ippm-ioam-data]. An example of IOAM processing may be to export the data fields, send data fields via Telemetry, etc.
- o The decapsulating node also pops the Indicator Label and the IOAM data fields from the MPLS header.

4.1. Edge-to-Edge IOAM Indicator Label Allocation

IOAM Indicator Label (value TBA1) and IOAM and Flow Indicator Label (value TBA2) are used to indicate the presence of the E2E IOAM data field in the MPLS header. The E2E IOAM Indicator Label and IOAM and Flow Indicator Label can be allocated using one of the following methods:

- o Labels assigned by IANA with value TBA1 and TBA2 from the Extended Special-Purpose MPLS Values [I-D.ietf-mpls-spl-terminology].
- o Labels allocated by a Controller from the global table of the decapsulating node. The Controller provisions the label on both encapsulating and decapsulating nodes.
- o Labels allocated by the decapsulating node. The signaling extension for this is outside the scope of this document.

5. Procedure for Hop-by-Hop IOAM

The Hop-by-Hop (HbH) IOAM includes IOAM Option-Types IOAM Pre-allocated Trace Option-Type, IOAM Incremental Trace Option-Type and IOAM Proof of Transit (POT) Option-Type [I-D.ietf-ippm-ioam-data]. This section summarizes the procedure for data encapsulation and decapsulation for Hop-by-hop IOAM in MPLS header.

- o The encapsulating node inserts the IOAM Indicator Label or IOAM Flow Indicator Label with Flow Label and one or more IOAM data field(s) in the MPLS header. The procedure to generate the Flow Label is outside the scope of this document.
- o The intermediate and decapsulating node enabled with IOAM functions "forwards and punts the timestamped copy" of the data

packet including IOAM data fields when the node recognizes the IOAM Indicator Label and IOAM Flow Indicator Label. The copy of the data packet is punted to the slow path for OAM processing and is not necessarily punted to the control-plane. The receive timestamp is required by various hop-by-hop OAM use-cases.

- o The intermediate and decapsulating node processes the IOAM data field(s) using the procedures defined in [I-D.ietf-ippm-ioam-data]. An example of IOAM processing may be to export the data fields, send data fields via Telemetry, etc.
- o The decapsulating node pops the Indicator Label and the IOAM data fields from the MPLS header.

5.1. Hop-by-Hop IOAM Indicator Label Allocation

IOAM Indicator Label (value TBA3) and IOAM and Flow Indicator Label (value TBA4) are used to indicate the presence of the HbH IOAM data field in the MPLS header. The HbH IOAM Indicator Label and IOAM and Flow Indicator Label can be allocated using one of the following methods:

- o Labels assigned by IANA with value TBA3 and TBA4 from the Extended Special-Purpose MPLS Values [I-D.ietf-mpls-spl-terminology].
- o Labels allocated by a Controller from the network-wide global table. The Controller provisions the labels on all nodes participating in IOAM functions along the data traffic path.

6. Considerations for ECMP

The encapsulating node needs to make sure the IOAM data field does not start with a well known IP protocol value (e.g. 0x4 for IPv4 and 0x6 for IPv6) as it can alter the hashing function for ECMP that uses the IP header. This can be achieved by using the IOAM and Flow Indicator Label (value TBA2 and TBA4) that follows by protocol value 0010b. This approach is consistent with utilizing 0000b or 0001b as the first nibble after the MPLS label stack, as described in [RFC4928] [RFC4385].

Note that the hashing function for ECMP that uses the labels from the MPLS header may now include the Indicator Label.

When entropy label [RFC6790] is used for hashing function for ECMP, the procedure defined in this document does not alter the hashing function.

7. Node Capability

The decapsulating node that has to pop the Indicator Label, 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 of the MPLS header.

8. Data Packets with SR-MPLS Header

Segment Routing (SR) technology leverages the source routing paradigm [RFC8660]. A node steers a packet through a controlled set of instructions, called segments, by pre-pending the packet with an SR header. In the SR with MPLS data plane (SR-MPLS), the SR header is instantiated through a label stack.

An example of data packet carrying the SR-MPLS header with Path Segment Identifier (PSID) [I-D.ietf-spring-mpls-path-segment] with IOAM encapsulation is shown in Figure 3. The Path Segment Identifier allows to identify the path associated with the data traffic being monitored for IOAM on the decapsulating node.

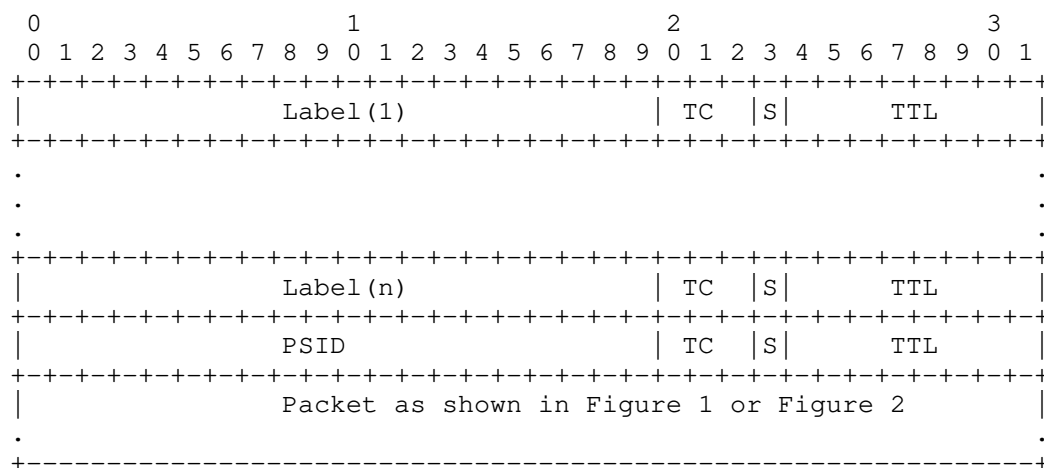


Figure 3: Data Packet with SR-MPLS Header

9. Security Considerations

The security considerations of SR-MPLS are discussed in [RFC8660], and the security considerations of IOAM in general are discussed in [I-D.ietf-ippm-ioam-data].

IOAM is considered a "per domain" feature, where one or several operators decide on leveraging and configuring IOAM according to their needs. Still, operators need to properly secure the IOAM domain to avoid malicious configuration and use, which could include injecting malicious IOAM packets into a domain.

10. IANA Considerations

IANA maintains the "Special-Purpose Multiprotocol Label Switching (MPLS) Label Values" registry (see <<https://www.iana.org/assignments/mpls-label-values/mpls-label-values.xml>>). IANA is requested to allocate IOAM Indicator Label value and IOAM and Flow Indicator value from the "Extended Special-Purpose MPLS Label Values" registry:

Value	Description	Reference
TBA1	E2E IOAM Indicator Label	This document
TBA2	E2E IOAM and Flow Indicator Label	This document
TBA3	HbH IOAM Indicator Label	This document
TBA4	HbH IOAM and Flow Indicator Label	This document

11. References

11.1. Normative References

[I-D.ietf-ippm-ioam-data]

Brockners, F., Bhandari, S., and T. Mizrahi, "Data Fields for In-situ OAM", draft-ietf-ippm-ioam-data-10 (work in progress), July 2020.

[I-D.ietf-ippm-ioam-direct-export]

Song, H., Gafni, B., Zhou, T., Li, Z., Brockners, F., Bhandari, S., Sivakolundu, R., and T. Mizrahi, "In-situ OAM Direct Exporting", draft-ietf-ippm-ioam-direct-export-01 (work in progress), August 2020.

- [I-D.ietf-ippm-ioam-flags]
Mizrahi, T., Brockners, F., Bhandari, S., Sivakolundu, R., Pignataro, C., Kfir, A., Gafni, B., Spiegel, M., and J. Lemon, "In-situ OAM Flags", draft-ietf-ippm-ioam-flags-02 (work in progress), July 2020.
- [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>>.
- [RFC8660] Bashandy, A., Ed., Filsfils, C., Ed., Previdi, S., Decraene, B., Litkowski, S., and R. Shakir, "Segment Routing with the MPLS Data Plane", RFC 8660, DOI 10.17487/RFC8660, December 2019, <<https://www.rfc-editor.org/info/rfc8660>>.
- [RFC8664] Sivabalan, S., Filsfils, C., Tantsura, J., Henderickx, W., and J. Hardwick, "Path Computation Element Communication Protocol (PCEP) Extensions for Segment Routing", RFC 8664, DOI 10.17487/RFC8664, December 2019, <<https://www.rfc-editor.org/info/rfc8664>>.

11.2. Informative References

- [I-D.ietf-mpls-spl-terminology]
Andersson, L., Kompella, K., and A. Farrel, "Special Purpose Label terminology", draft-ietf-mpls-spl-terminology-03 (work in progress), August 2020.
- [I-D.ietf-spring-mpls-path-segment]
Cheng, W., Li, H., Chen, M., Gandhi, R., and R. Zigler, "Path Segment in MPLS Based Segment Routing Network", draft-ietf-spring-mpls-path-segment-02 (work in progress), February 2020.
- [RFC4385] Bryant, S., Swallow, G., Martini, L., and D. McPherson, "Pseudowire Emulation Edge-to-Edge (PWE3) Control Word for Use over an MPLS PSN", RFC 4385, DOI 10.17487/RFC4385, February 2006, <<https://www.rfc-editor.org/info/rfc4385>>.

- [RFC4928] Swallow, G., Bryant, S., and L. Andersson, "Avoiding Equal Cost Multipath Treatment in MPLS Networks", BCP 128, RFC 4928, DOI 10.17487/RFC4928, June 2007, <<https://www.rfc-editor.org/info/rfc4928>>.
- [RFC6790] Kompella, K., Drake, J., Amante, S., Henderickx, W., and L. Yong, "The Use of Entropy Labels in MPLS Forwarding", RFC 6790, DOI 10.17487/RFC6790, November 2012, <<https://www.rfc-editor.org/info/rfc6790>>.

Acknowledgements

The authors would like to thank Patrick Khordoc, Shwetha Bhandari and Vengada Prasad Govindan for the discussions on IOAM. The authors would also like to thank Tarek Saad, Loa Andersson, Greg Mirsky, and Cheng Li for providing many useful comments.

Contributors

Sagar Soni
Cisco Systems, Inc.

Email: sagsoni@cisco.com

Authors' Addresses

Rakesh Gandhi (editor)
Cisco Systems, Inc.
Canada

Email: rgandhi@cisco.com

Zafar Ali
Cisco Systems, Inc.

Email: zali@cisco.com

Clarence Filsfils
Cisco Systems, Inc.
Belgium

Email: cf@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

Voitek Kozak
Comcast

Email: Voitek_Kozak@comcast.com

MPLS Working Group
Internet-Draft
Intended status: Standards Track
Expires: April 28, 2021

Y. Liu
G. Mirsky
ZTE Corporation
October 25, 2020

MPLS-based Service Function Path(SFP) Consistency Verification
draft-lm-mpls-sfc-path-verification-01

Abstract

This document proposes a method to verify the correlation between Service Function Chaining control and/or management plane view of the specified Service Function Path and the state of its data. It works for both SR service programming and MPLS-based NSH SFC.

This document defines the signaling of the Generic Associated Channel (G-ACh) over a Service Function Path (SFP) with an MPLS forwarding plane using the basic unit defined in RFC 8595. The document also describes the processing of the G-ACh by the elements of the SFP.

Status of This Memo

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

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

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 April 28, 2021.

Copyright Notice

Copyright (c) 2020 IETF Trust and the persons identified as the document authors. All rights reserved.

This document is subject to BCP 78 and the IETF Trust's Legal Provisions Relating to IETF Documents (<https://trustee.ietf.org/license-info>) in effect on the date of publication of this document. Please review these documents carefully, as they describe your rights and restrictions with respect

to this document. Code Components extracted from this document must include Simplified BSD License text as described in Section 4.e of the Trust Legal Provisions and are provided without warranty as described in the Simplified BSD License.

Table of Contents

1. Introduction	2
2. Conventions used in this document	3
2.1. Requirements Language	3
2.2. Acronyms	3
3. MPLS-based SFP Consistency Verification	4
3.1. Special-purpose Label in SFC-MPLS Environment	5
3.1.1. G-ACh over SFC-MPLS	6
3.2. MPLS-based SFP Consistency Verification	6
3.3. SFC Info Sub-TLV	7
3.4. SFC Basic Unit FEC Sub-TLV	8
3.5. Theory of Operation	9
3.5.1. MPLS-based service programming	10
3.5.2. Path Consistency in SFC-MPLS	10
3.6. Discussion	10
4. Security Considerations	11
5. IANA Considerations	11
5.1. SFC Validation TLV	11
5.1.1. Sub-TLVs for SFC Validation TLV	11
5.2. SFC Basic Unit sub-TLV	12
6. References	12
6.1. Normative References	12
6.2. Informative References	13
Authors' Addresses	13

1. Introduction

Service Function Chain (SFC) defined in [RFC7665] as an ordered set of service functions (SFs) to be applied to packets and/or frames, and/or flows selected as a result of classification.

SFC can be achieved through a variety of encapsulation methods, such as NSH [RFC8300], SR service programming [I-D.ietf-spring-sr-service-programming] and MPLS-based NSH SFC [RFC8595].

This document describes extensions to MPLS LSP ping [RFC8029] mechanisms to support verification between the control/management plane and the data plane state for both SR-MPLS service programming and MPLS-based NSH SFC.

An MPLS LSP ping is a component of the MPLS Operation, Administration, and Maintenance (OAM) toolset. OAM packets used to monitor a specific Service Function Path (SFP) can be transported over a Generic Associated Channel (G-ACh). This document defines the signaling of the G-ACh over an SFP with an MPLS forwarding plane using the basic unit defined in [RFC8595]. The document also describes the processing of the G-ACh by the elements of the SFP.

2. Conventions used in this document

2.1. Requirements Language

The key words "MUST", "MUST NOT", "REQUIRED", "SHALL", "SHALL NOT", "SHOULD", "SHOULD NOT", "RECOMMENDED", "NOT RECOMMENDED", "MAY", and "OPTIONAL" in this document are to be interpreted as described in BCP 14 [RFC2119] [RFC8174] when, and only when, they appear in all capitals, as shown here.

2.2. Acronyms

SFC: Service Function Chain

SFF: Service Function Forwarder

SF: Service Function

SFI: Instance of an SF

SFP: Service Function Path

RSP: Rendered Service Path

SFC-MPLS: SFC over an MPLS forwarding plane

SPL: Special-Purpose Label

bSPL: Base SPL

eSPL: Extended SPL

GAL: Generic Associated Channel Label

ELI: Entropy Label Indicator

OAM: Operation, Administration, and Maintenance

G-ACh: Generic Associated Channel

GAL: Generic Associated Channel Label

3. MPLS-based SFP Consistency Verification

MPLS echo request and reply messages [RFC8029] can be extended to support the verification of the consistency of an MPLS-based Service Function Path (SFP).

SR-MPLS/MPLS can be used to realize an SFP. Two methods have been defined:

- o [I-D.ietf-spring-sr-service-programming] describes how to achieve service function chaining in SR-enabled MPLS and IPv6 networks. In an SR-MPLS network, each SF is associated with an MPLS label. As a result, an SFP can be encoded as a stack of MPLS labels and pushed on top of the packet.
- o [RFC8595] provides another method to realize SFC in an MPLS network by means of using a logical representation of the Network Service Header (NSH) in an MPLS label stack. This method, throughout this document, is referred to as SFC over an MPLS data plane (SFC-MPLS). When an MPLS label stack is used to carry a logical NSH, a basic unit of representation is used, which can be present one or more times in the label stack. This unit comprises two MPLS labels, one carries a label to provide a context within the SFC scope (the SFC Context Label), and the other carries a label to show which SF is to be enacted (the SF Label). This two-label unit is shown in Figure 1.

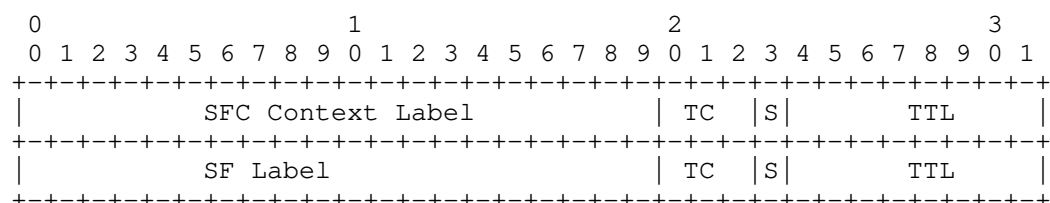


Figure 1: The Basic Unit of MPLS Label Stack for SFC

In MPLS Label Switched Paths (LSPs), MPLS LSP ping [RFC8029] is used to check the correctness of the data plane functioning and to verify the data plane against the control plane.

The proposed extension of MPLS LSP ping allows verification of the correlation between the control/management (if data model-based

central controller used) plane and the data plane state in SR-MPLS/MPLS-based SFC.

Generally, except for the designed specific functions, the packet processing functions supported by SFs are limited. SFs may not support control and/or management protocols operated over the G-ACh defined in [RFC5586], e.g., MPLS OAM protocols like LSP ping. To avoid such packets mishandled, SFFs are responsible for MPLS echo request processing. To support that processing, the basic unit can use the mechanism described in Section 3.1.

3.1. Special-purpose Label in SFC-MPLS Environment

When an SFC-MPLS is used, an SFF needs to identify an OAM packet with the SFP scope. To achieve that, this specification first defines the use of a base special-purpose label (bSPL) [RFC3032] or an extended special-purpose label (eSPL) [RFC7274] (referred in this document as SPL Unit) with the basic unit defined in [RFC8595]. And based on that, the use of Generic Associated Channel Label (GAL) [RFC5586] with the basic unit in the SFC-MPLS environment.

Special-purpose label (SPL), whether bSPL or eSPL, have special significance in the data and control planes. An ability to use an SPL in the basic unit allows for a closer functional match between the NSH-based SFC and SFC-MPLS. For example, Entropy Label Indicator (ELI) [RFC6790] with the basic unit can be used as the Flow ID TLV [I-D.ietf-sfc-nsh-tlv] to allow an SFF to balance SFC flows among SFs of the same type. An SPL MAY be used with the basic unit in SFC-MPLS, as displayed in Figure 2. Note that an SPL unit MAY be present in one or more basic units when MPLS label stacking is used to carry the SFC information.

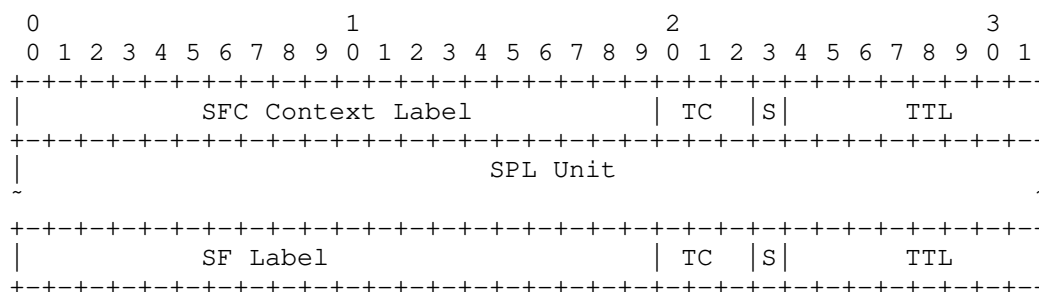


Figure 2: Special-purpose Label Unit with the Basic Unit of MPLS Label Stack for SFC

3.1.1. G-ACh over SFC-MPLS

SFC-MPLS environment could include instances of an SF (SFI) or SFC proxies that cannot properly process control and/or management protocol messages that are exchanged between nodes over the G-ACh associated with the particular SFP. To support OAM over G-ACh, it is beneficial to avoid handing over a test packet to the SFI or SFC proxy. Hence, this specification defines that if the Generic Associated Channel Label (GAL) immediately follows the SFC Context label [RFC8595], then the packet is recognized as an SFP OAM packet.

Below are the processing rules of an SFP OAM packet by an SFF:

- o An SFF MUST NOT pass the packet to a local SFI or SFC proxy.
- o The SFF MUST decrement SF Label entry's TTL value. If the resulting value equals zero, the SFF MUST pass the SFP OAM packet to the control plane for processing. An implementation that supports this specification MUST provide control to limit the rate of SFP OAM packets passed to the control plane for processing.
- o If the TTL value is not zero, the SFP OAM packet is processed as defined in Section 6, Section 7, and Section 8 [RFC8595], according to the type of MPLS forwarding used in the SFP.

3.2. MPLS-based SFP Consistency Verification

An MPLS SFC validation request/reply is an MPLS echo request/reply that includes an SFC validation TLV (shown in Figure 3).

Nodes examine and process the TLV only if configured to do so; other nodes MUST ignore the TLV and process the packet as a standard MPLS echo packet.

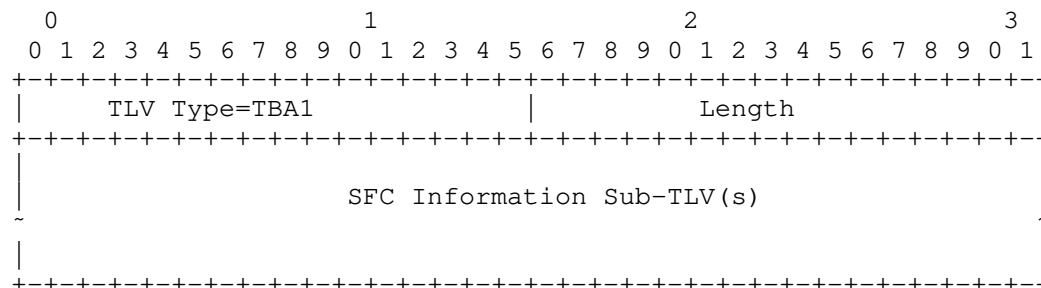


Figure 3: SFC Validation TLV

SFC Information Sub-TLV: The Sub-TLV, as presented in Figure 4, MUST NOT be included in an MPLS SFC validation request.

3.3. SFC Info Sub-TLV

Upon receiving the SFC validation request, the SFF MUST respond with an echo reply, which includes the SFC detailed information.

The SFC detailed information is recorded in SFC info sub-TLV.

Two types of sub-TLVs are defined in this section, and those are used in MPLS-based service programming [I-D.ietf-spring-sr-service-programming] and MPLS-based NSH [RFC8595] respectively.

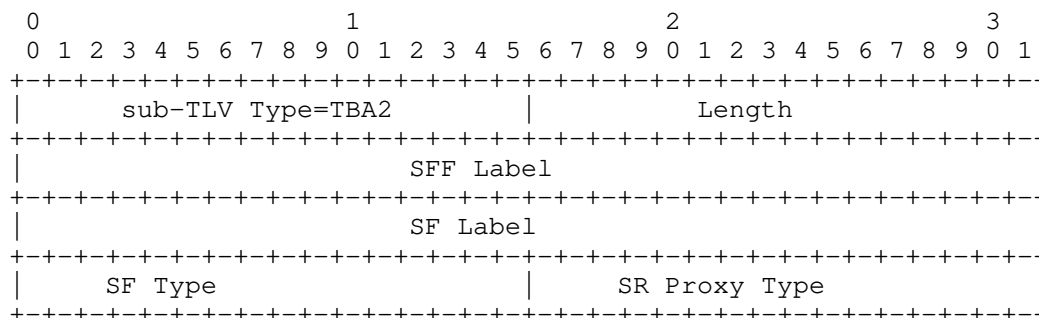


Figure 4: SFC Info Sub-TLV for SR-MPLS-based Service Programming

Type TBA2 sub-TLV: used in SR-MPLS-based service programming

SFF Label: represents the SID of the SFF

SF Label: represents the service SID of the SF or SR proxy

SF Type: indicates the type of SF, such as DPI, firewall, etc.

SR Proxy Type: It is defined in [I-D.ietf-spring-sr-service-programming] and indicates the type of SR proxy if it exists.

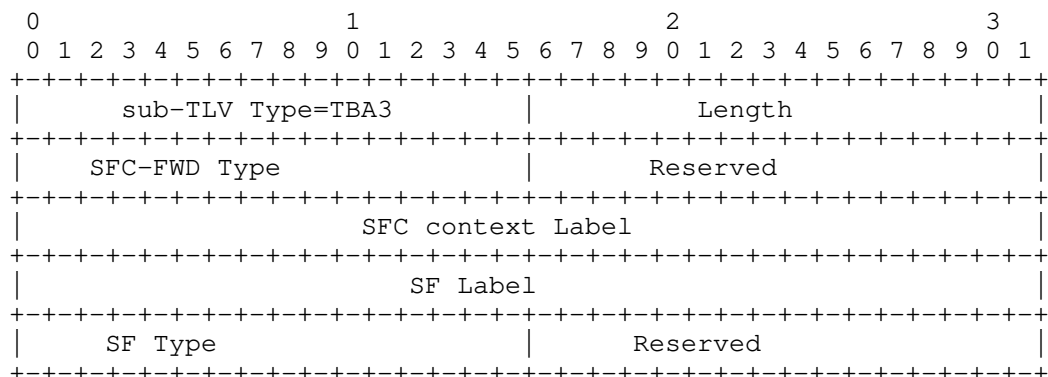


Figure 5: SFC Info Sub-TLV for MPLS-based NSH

Figure 5 presents the format of a sub-TLV for MPLS-based NSH. The fields are as follows:

Type TBA3 sub-TLV: used in MPLS-based NSH

SFC-FWD Type: indicates the forwarding type of the data plane, and has the following values:

0x10: MPLS-based NSH [RFC8595] label swapping

0x11: MPLS-based NSH [RFC8595] label stacking

SFC context Label: The meaning of the SFC context label depends on the SFC Type. If SFC-FWD Type is 0x10, the SFC context Label represents SPI. If SFC-FWD Type is 0x11, the SFC context Label represents the context label [RFC8595].

SF Label: The meaning of the SF label depends on the SFC-FWD Type. If SFC Type is 0x10, the SF Label represents SI. If SFC Type is 0x11, the SF Label represents the SFI index [RFC8595].

SF Type: It is defined in [I-D.ietf-bess-nsh-bgp-control-plane] and indicates the type of SF, such as DPI, firewall, etc.

3.4. SFC Basic Unit FEC Sub-TLV

Unlike standard MPLS forwarding, based on a single label, packet forwarding defined in [RFC8595] is based on the basic unit of MPLS label stack for SFC(SFC Context Label+SF Label). A new SFC Basic Unit FEC sub-TLV with Type value (TBA4) is defined in this document.

The SFC Basic Unit FEC sub-TLV MAY be used to carry the corresponding FEC of the basic unit.

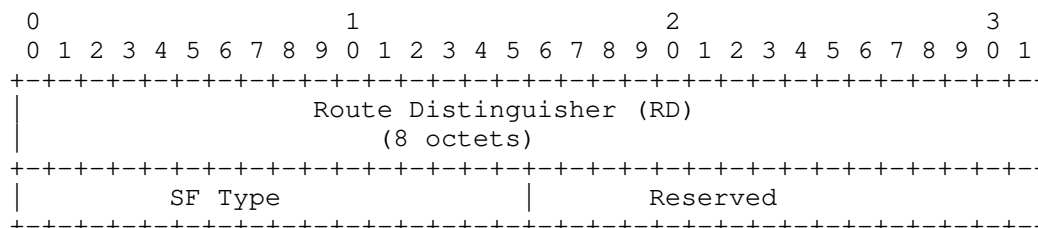


Figure 6: SFC Basic Unit sub-TLV

The format of the basic unit sub-TLV is shown in Figure 6 and includes the following fields:

Route Distinguisher (RD): 8 octets field in SFIR Route Type specific NLRI [I-D.ietf-bess-nsh-bgp-control-plane].

SF Type: 2 octets. It is defined in [I-D.ietf-bess-nsh-bgp-control-plane] and indicates the type of SF, such as DPI, firewall, etc.

SFC Basic Unit sub-TLV is defined for Target FEC Stack TLV(Type 1 TLV).

A node that receives an LSP ping with the Target FEC Stack TLV and the SFC Basic Unit FEC Sub-TLV included will check if it is its Route Distinguisher and whether it advertised that Service Function Type. If the validation is not passed, the SFF will generate an MPLS echo reply with an error code as defined in [RFC8029].

3.5. Theory of Operation

An MPLS SFC validation request is an MPLS echo request with an SFC validation TLV, and the echo request is sent with a label stack corresponding to the SFP being tested. To trace SFC-MPLS, the Generic Associated Channel Label (GAL) which immediately follows the SFC Context label is also included.

Sending an SFC echo request to the control plane is triggered by one of the following packet processing exceptions: IP TTL expiration, MPLS TTL expiration, or the receiver is the terminal SFF for an SFP.

After general packet sanity verifying [RFC8029], Section 3.5.1 and Section 3.5.2 in this document separately describe the following processing procedures in service programming and MPLS-based NSH.

After all SFFs on the SFP send back MPLS echo reply, the sender collects information about all traversed SFFs and SFs on the rendered service path (RSP).

3.5.1. MPLS-based service programming

[I-D.ietf-spring-sr-service-programming] describes how a service can be associated with a SID to achieve service function chaining. In an SR-MPLS network, the SFP is encoded as a stack of MPLS labels. That stack is pushed on top of the packet.

Before generating an echo reply, an SFF MUST parse through the label stack until the next label is not a local service SID to get all the SFs attached to the SFP on the SFP and record the corresponding Label-stack-depth.

The SFF then sends an MPLS echo reply with all the SF information recorded in SFC Information Sub-TLV, including the service SID and the SF type.

3.5.2. Path Consistency in SFC-MPLS

[RFC8595] describes how Service Function Chaining (SFC) can be achieved in an MPLS network using a logical representation of the Network Service Header (NSH) in an MPLS label stack.

SFC forwarding can be achieved by label swapping, label stacking, or the mix of both. When an SFF receives a packet containing an MPLS label stack, it examines the top basic unit of MPLS label stack for SFC, {SPI, SI} or {context label, SFI index}, to determine where to send the packet next.

As described in Section 3.1.1, the packet with GAL is recognized by the SFF as an SFP OAM packet. The SFF then decrements SF Label entry's TTL value. If the resulting value equals zero, the SFF passes the SFP OAM packet to the control plane for processing. The system that supports this specification generates a reply message, including in it the SFC info sub-TLV as described in Section 3.3.

3.6. Discussion

In [RFC8595], it says, "when an SFF receives a packet from any component of the SFC system (classifier, SFI, or another SFF), it MUST discard any packets with TTL set to zero". To trace SFC, it

should be changed to allow punting the packet to the control plane though under throttling control.

4. Security Considerations

This specification defines the processing of an SFP OAM packet. Such packets could be used as an attack vector. A system that supports this specification **MUST** provide control to limit the rate of SFP OAM packets sent to the control plane for processing.

5. IANA Considerations

This document requests assigning type values for TLVs and sub-TLVs from the IANA "Multiprotocol Label Switching (MPLS) Label Switched Paths (LSPs) Ping Parameters" registry as described in the sub-sections below.

5.1. SFC Validation TLV

IANA is requested to assign a type from the Standards Action range of the "TLVs" registry according to Table 1:

Value	Description	Reference
TBA1	SFC Validation	This document

Table 1: Type Value for SFC Validation TLV

5.1.1. Sub-TLVs for SFC Validation TLV

Two sub-TLVs of SFC Validation TLV is defined in this document according to Table 2.

Value	Description	Reference
TBA2	Info for SR-MPLS-based Service Programming	This document
TBA3	Info for MPLS-based NSH	This document

Table 2: Sub-TLV Values for SFC Validation TLV

5.2. SFC Basic Unit sub-TLV

IANA is requested to assign a type from the Standards Action range of the "Sub-TLVs for TLV Types 1, 16, and 21" registry according to Table 3:

Value	Description	Reference
TBA4	SFC Basic Unit	This document

Table 3: Type Value for SFC Basic Unit sub-TLV

6. References

6.1. Normative References

- [I-D.ietf-bess-nsh-bgp-control-plane]
Farrel, A., Drake, J., Rosen, E., Uttaro, J., and L. Jalil, "BGP Control Plane for the Network Service Header in Service Function Chaining", draft-ietf-bess-nsh-bgp-control-plane-18 (work in progress), August 2020.
- [I-D.ietf-spring-sr-service-programming]
Clad, F., Xu, X., Filsfils, C., daniel.bernier@bell.ca, d., Li, C., Decraene, B., Ma, S., Yadlapalli, C., Henderickx, W., and S. Salsano, "Service Programming with Segment Routing", draft-ietf-spring-sr-service-programming-03 (work in progress), September 2020.
- [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>>.
- [RFC5586] Bocci, M., Ed., Vigoureux, M., Ed., and S. Bryant, Ed., "MPLS Generic Associated Channel", RFC 5586, DOI 10.17487/RFC5586, June 2009, <<https://www.rfc-editor.org/info/rfc5586>>.

- [RFC7274] Kompella, K., Andersson, L., and A. Farrel, "Allocating and Retiring Special-Purpose MPLS Labels", RFC 7274, DOI 10.17487/RFC7274, June 2014, <<https://www.rfc-editor.org/info/rfc7274>>.
- [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>>.
- [RFC8300] Quinn, P., Ed., Elzur, U., Ed., and C. Pignataro, Ed., "Network Service Header (NSH)", RFC 8300, DOI 10.17487/RFC8300, January 2018, <<https://www.rfc-editor.org/info/rfc8300>>.
- [RFC8595] Farrel, A., Bryant, S., and J. Drake, "An MPLS-Based Forwarding Plane for Service Function Chaining", RFC 8595, DOI 10.17487/RFC8595, June 2019, <<https://www.rfc-editor.org/info/rfc8595>>.

6.2. Informative References

- [I-D.ietf-sfc-nsh-tlv]
Wei, Y., Elzur, U., and S. Majee, "Network Service Header Metadata Type 2 Variable-Length Context Headers", draft-ietf-sfc-nsh-tlv-03 (work in progress), May 2020.
- [RFC6790] Kompella, K., Drake, J., Amante, S., Henderickx, W., and L. Yong, "The Use of Entropy Labels in MPLS Forwarding", RFC 6790, DOI 10.17487/RFC6790, November 2012, <<https://www.rfc-editor.org/info/rfc6790>>.
- [RFC7665] Halpern, J., Ed. and C. Pignataro, Ed., "Service Function Chaining (SFC) Architecture", RFC 7665, DOI 10.17487/RFC7665, October 2015, <<https://www.rfc-editor.org/info/rfc7665>>.

Authors' Addresses

Liu Yao
ZTE Corporation
No. 50 Software Ave, Yuhuatai District
Nanjing
China

Email: liu.yao71@zte.com.cn

Greg Mirsky
ZTE Corporation

Email: gregimirsky@gmail.com

MPLS Working Group
Internet-Draft
Intended status: Standards Track
Expires: May 6, 2021

G. Mirsky
ZTE Corp.
G. Mishra
Verizon Inc.
D. Eastlake
Futureway Technologies
November 2, 2020

BFD for Multipoint Networks over Point-to-Multi-Point MPLS LSP
draft-mirsky-mpls-p2mp-bfd-12

Abstract

This document describes procedures for using Bidirectional Forwarding Detection (BFD) for multipoint networks to detect data plane failures in Multiprotocol Label Switching (MPLS) point-to-multipoint (p2mp) Label Switched Paths (LSPs) using active tails with unsolicited notifications mode. It also describes the applicability of LSP Ping, as in-band, and the control plane, as out-band, solutions to bootstrap a BFD session in this environment.

Status of This Memo

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

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

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 May 6, 2021.

Copyright Notice

Copyright (c) 2020 IETF Trust and the persons identified as the document authors. All rights reserved.

This document is subject to BCP 78 and the IETF Trust's Legal Provisions Relating to IETF Documents (<https://trustee.ietf.org/license-info>) in effect on the date of publication of this document. Please review these documents

carefully, as they describe your rights and restrictions with respect to this document. Code Components extracted from this document must include Simplified BSD License text as described in Section 4.e of the Trust Legal Provisions and are provided without warranty as described in the Simplified BSD License.

Table of Contents

1. Introduction	2
2. Conventions used in this document	2
2.1. Terminology	3
2.2. Requirements Language	3
3. Multipoint BFD Encapsulation	3
3.1. IP Encapsulation of Multipoint BFD	4
3.2. Non-IP Encapsulation of Multipoint BFD	4
4. Bootstrapping Multipoint BFD	5
4.1. LSP Ping	5
4.2. Operation of Multipoint BFD with Active Tail over P2MP MPLS LSP	6
4.3. Control Plane	7
5. Security Considerations	7
6. IANA Considerations	7
7. Acknowledgements	8
8. References	8
8.1. Normative References	8
8.2. Informative References	9
Authors' Addresses	10

1. Introduction

[RFC8562] defines a method of using Bidirectional Detection (BFD) [RFC5880] to monitor and detect unicast failures between the sender (head) and one or more receivers (tails) in multipoint or multicast networks. [RFC8562] added two BFD session types - MultipointHead and MultipointTail. Throughout this document, MultipointHead and MultipointTail refer to the value of the bfd.SessionType is set on a BFD endpoint. This document describes procedures for using such modes of BFD protocol to detect data plane failures in Multiprotocol Label Switching (MPLS) point-to-multipoint (p2mp) Label Switched Paths (LSPs). The document also describes the applicability of out-band solutions to bootstrap a BFD session in this environment.

2. Conventions used in this document

2.1. Terminology

MPLS: Multiprotocol Label Switching

LSP: Label Switched Path

BFD: Bidirectional Forwarding Detection

p2mp: Point-to-Multipoint

FEC: Forwarding Equivalence Class

G-ACh: Generic Associated Channel

ACH: Associated Channel Header

GAL: G-ACh Label

2.2. Requirements Language

The key words "MUST", "MUST NOT", "REQUIRED", "SHALL", "SHALL NOT", "SHOULD", "SHOULD NOT", "RECOMMENDED", "NOT RECOMMENDED", "MAY", and "OPTIONAL" in this document are to be interpreted as described in BCP 14 [RFC2119] [RFC8174] when, and only when, they appear in all capitals, as shown here.

3. Multipoint BFD Encapsulation

[RFC8562] uses BFD in the Demand mode from the very start of a point-to-multipoint (p2mp) BFD session. Because the head doesn't receive any BFD Control packet from a tail, the head of the p2mp BFD session transmits all BFD Control packets with the value of Your Discriminator field set to zero. As a result, a tail cannot demultiplex BFD sessions using Your Discriminator, as defined in [RFC5880]. [RFC8562] requires that to demultiplex BFD sessions, the tail uses the source IP address, My Discriminator, and the identity of the multipoint tree from which the BFD Control packet was received. The p2mp MPLS LSP label MAY provide the identification of the multipoint tree in case of an inclusive p-tree or upstream-assigned label in case of aggregate p-tree. If the BFD Control packet is encapsulated in IP/UDP, then the source IP address MUST be used to demultiplex the received BFD Control packet as described in Section 3.1. The non-IP encapsulation case is described in Section 3.2.

3.1. IP Encapsulation of Multipoint BFD

[RFC8562] defines IP/UDP encapsulation for multipoint BFD over p2mp MPLS LSP:

UDP destination port MUST be set to 3784;

destination IP address MUST be set to the loopback address 127.0.0.1/32 for IPv4, or the loopback address ::1/128 for IPv6 [RFC4291].

This specification further clarifies that:

if multiple alternative paths for the given p2mp LSP Forwarding Equivalence Class (FEC) exist, the MultipointHead SHOULD use Entropy Label [RFC6790] used for LSP Ping [RFC8029] to exercise that particular alternative path;

or the MultipointHead MAY use the UDP port number as discovered by LSP Ping traceroute [RFC8029] as the source UDP port number to possibly exercise that particular alternate path.

3.2. Non-IP Encapsulation of Multipoint BFD

In some environments, the overhead of extra IP/UDP encapsulations may be considered burdensome, making the use of more compact G-ACh encapsulation attractive. Also, the validation of the IP/UDP encapsulation of a BFD Control packet in a p2mp BFD session may fail because of a problem not related to neither MPLS label stack nor to BFD. Avoiding unnecessary encapsulation of p2mp BFD over an MPLS LSP improves the accuracy of the correlation of the detected failure and defect in MPLS LSP. Non-IP encapsulation for multipoint BFD over p2mp MPLS LSP MUST use Generic Associated Channel (G-ACh) Label (GAL) (see [RFC5586]) at the bottom of the label stack followed by an Associated Channel Header (ACH). If BFD Control packet in PW-ACH encapsulation (without IP/UDP Headers) is to be used in ACH, an implementation would not be able to verify the identity of the MultipointHead and, as a result, will not properly demultiplex BFD packets. Hence, a new channel type value is needed. The Channel Type field in ACH MUST be set to TBA1 value Section 6. To provide the identity of the MultipointHead for the particular multipoint BFD session, a Source Address TLV [RFC7212] MUST immediately follow a BFD Control message.

4. Bootstrapping Multipoint BFD

4.1. LSP Ping

LSP Ping is the part of the on-demand OAM toolset used to detect and localize defects in the data plane and verify the control plane against the data plane by ensuring that the LSP is mapped to the same FEC at both egress and ingress endpoints.

LSP Ping, as defined in [RFC6425], MAY be used to bootstrap MultipointTail. If LSP Ping used, it MUST include the Target FEC TLV and the BFD Discriminator TLV defined in [RFC5884]. The Target FEC TLV MUST use sub-TLVs defined in Section 3.1 [RFC6425]. It is RECOMMENDED to set the value of Reply Mode field to "Do not reply" [RFC8029] for the LSP Ping to bootstrap MultipointTail of the p2mp BFD session. Indeed, because BFD over a multipoint network is using BFD Demand mode, the LSP echo reply from a tail has no useful information to convey to the head, unlike in the case of the BFD over a p2p MPLS LSP [RFC5884]. A MultipointTail that receives the LSP Ping that includes the BFD Discriminator TLV:

- o MUST validate the LSP Ping;
- o MUST associate the received BFD Discriminator value with the p2mp LSP;
- o MUST create a p2mp BFD session and set `bfd.SessionType = MultipointTail` as described in [RFC8562];
- o MUST use the source IP address of LSP Ping, the value of BFD Discriminator from the BFD Discriminator TLV, and the identity of the p2mp LSP to properly demultiplex BFD sessions.

Besides bootstrapping a BFD session over a p2mp LSP, LSP Ping SHOULD be used to verify the control plane against the data plane periodically by checking that the p2mp LSP is mapped to the same FEC at the MultipointHead and all active MultipointTails. The rate of generation of these LSP Ping Echo request messages SHOULD be significantly less than the rate of generation of the BFD Control packets because LSP Ping requires more processing to validate the consistency between the data plane and the control plane. An implementation MAY provide configuration options to control the rate of generation of the periodic LSP Ping Echo request messages.

4.2. Operation of Multipoint BFD with Active Tail over P2MP MPLS LSP

[RFC8562] defined how the BFD Demand mode can be used in multipoint networks. When applied in MPLS, procedures specified in [RFC8562] allow an egress LSR to detect a failure of the part of the MPLS p2mp LSP from the ingress LSR. The ingress LSR is not aware of the state of the p2mp LSP. [RFC8563], using mechanisms defined in [RFC8562], defined an "active tail" behavior. An active tail might notify the head of the detected failure and responds to a poll sequence initiated by the head. The first method, referred to as Head Notification without Polling, is mentioned in Section 5.2.1 [RFC8563], is the simplest of all described in [RFC8563]. The use of this method in BFD over MPLS p2mp LSP is discussed in this document. Analysis of other methods of a head learning of the state of an MPLS p2mp LSP is outside the scope of this document.

As specified in [RFC8563] for the active tail mode, BFD variables MUST be as follows:

On an ingress LSR:

- o bfd.SessionType is MultipointHead;
- o bfd.RequiredMinRxInterval is set to nonzero, allowing egress LSRs to send BFD Control packets.

On an egress LSR:

- o bfd.SessionType is MultipointTail;
- o bfd.SilentTail is set to zero.

In Section 5.2.1 [RFC8563] is noted that "the tail sends unsolicited BFD packets in response to the detection of a multipoint path failure" but without the specifics on the information in the packet and frequency of transmissions. This document defines below the procedure of an active tail with unsolicited notifications for p2mp MPLS LSP.

Upon detecting the failure of the p2mp MPLS LSP, an egress LSR sends BFD Control packet with the following settings:

- o the Poll (P) bit is set;
- o the Status (Sta) field set to Down value;
- o the Diagnostic (Diag) field set to Control Detection Time Expired value;

- o the value of the Your Discriminator field is set to the value the egress LSR has been using to demultiplex that BFD multipoint session;
- o BFD Control packet is encapsulated in IP/UDP with the destination IP address of the ingress LSR and the UDP destination port number set to 4784 per [RFC5883]
- o these BFD Control packets are transmitted at the rate of one per second until either it receives the valid for this BFD session control packet with the Final (F) bit set from the ingress LSR or the defect condition clears.

To improve the likelihood of notifying the ingress LSR of the failure of the p2mp MPLS LSP, the egress LSR SHOULD initially transmit three BFD Control packets defined above in short succession.

An ingress LSR that has received the BFD Control packet, as described above, sends the unicast IP/UDP encapsulated BFD Control packet with the Final (F) bit set to the egress LSR.

4.3. Control Plane

The BGP-BFD Attribute [I-D.ietf-bess-mvpn-fast-failover] MAY be used to bootstrap multipoint BFD session on a tail.

5. Security Considerations

This document does not introduce new security aspects but inherits all security considerations from [RFC5880], [RFC5884], [RFC7726], [RFC8562], [RFC8029], and [RFC6425].

Also, BFD for p2mp MPLS LSP MUST follow the requirements listed in section 4.1 [RFC4687] to avoid congestion in the control plane or the data plane caused by the rate of generating BFD Control packets. An operator SHOULD consider the amount of extra traffic generated by p2mp BFD when selecting the interval at which the MultipointHead will transmit BFD Control packets. Also, the operator MAY consider the size of the packet the MultipointHead transmits periodically as using IP/UDP encapsulation, which adds up to 28 octets, more than 50% of the BFD Control packet length, comparing to G-ACh encapsulation.

6. IANA Considerations

IANA is requested to allocate value (TBA1) from its MPLS Generalized Associated Channel (G-ACh) Types registry.

Value	Description	Reference
TBA1	Multipoint BFD Session	This document

Table 1: Multipoint BFD Session G-ACh Type

7. Acknowledgements

The author sincerely appreciates the comments received from Andrew Malis and thought stimulating questions from Carlos Pignataro.

8. References

8.1. 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>>.
- [RFC5586] Bocci, M., Ed., Vigoureux, M., Ed., and S. Bryant, Ed., "MPLS Generic Associated Channel", RFC 5586, DOI 10.17487/RFC5586, June 2009, <<https://www.rfc-editor.org/info/rfc5586>>.
- [RFC5880] Katz, D. and D. Ward, "Bidirectional Forwarding Detection (BFD)", RFC 5880, DOI 10.17487/RFC5880, June 2010, <<https://www.rfc-editor.org/info/rfc5880>>.
- [RFC5883] Katz, D. and D. Ward, "Bidirectional Forwarding Detection (BFD) for Multihop Paths", RFC 5883, DOI 10.17487/RFC5883, June 2010, <<https://www.rfc-editor.org/info/rfc5883>>.
- [RFC5884] Aggarwal, R., Kompella, K., Nadeau, T., and G. Swallow, "Bidirectional Forwarding Detection (BFD) for MPLS Label Switched Paths (LSPs)", RFC 5884, DOI 10.17487/RFC5884, June 2010, <<https://www.rfc-editor.org/info/rfc5884>>.
- [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>>.

- [RFC6790] Kompella, K., Drake, J., Amante, S., Henderickx, W., and L. Yong, "The Use of Entropy Labels in MPLS Forwarding", RFC 6790, DOI 10.17487/RFC6790, November 2012, <<https://www.rfc-editor.org/info/rfc6790>>.
- [RFC7212] Frost, D., Bryant, S., and M. Bocci, "MPLS Generic Associated Channel (G-ACh) Advertisement Protocol", RFC 7212, DOI 10.17487/RFC7212, June 2014, <<https://www.rfc-editor.org/info/rfc7212>>.
- [RFC7726] Govindan, V., Rajaraman, K., Mirsky, G., Akiya, N., and S. Aldrin, "Clarifying Procedures for Establishing BFD Sessions for MPLS Label Switched Paths (LSPs)", RFC 7726, DOI 10.17487/RFC7726, January 2016, <<https://www.rfc-editor.org/info/rfc7726>>.
- [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>>.
- [RFC8562] Katz, D., Ward, D., Pallagatti, S., Ed., and G. Mirsky, Ed., "Bidirectional Forwarding Detection (BFD) for Multipoint Networks", RFC 8562, DOI 10.17487/RFC8562, April 2019, <<https://www.rfc-editor.org/info/rfc8562>>.
- [RFC8563] Katz, D., Ward, D., Pallagatti, S., Ed., and G. Mirsky, Ed., "Bidirectional Forwarding Detection (BFD) Multipoint Active Tails", RFC 8563, DOI 10.17487/RFC8563, April 2019, <<https://www.rfc-editor.org/info/rfc8563>>.

8.2. Informative References

- [I-D.ietf-bess-mvpn-fast-failover]
Morin, T., Kebler, R., and G. Mirsky, "Multicast VPN Fast Upstream Failover", draft-ietf-bess-mvpn-fast-failover-12 (work in progress), October 2020.
- [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>>.

[RFC4687] Yasukawa, S., Farrel, A., King, D., and T. Nadeau,
"Operations and Management (OAM) Requirements for Point-
to-Multipoint MPLS Networks", RFC 4687,
DOI 10.17487/RFC4687, September 2006,
<<https://www.rfc-editor.org/info/rfc4687>>.

Authors' Addresses

Greg Mirsky
ZTE Corp.

Email: gregimirsky@gmail.com

Gyan Mishra
Verizon Inc.

Email: gyan.s.mishra@verizon.com

Donald Eastlake, 3rd
Futureway Technologies
2386 Panoramic Circle
Apopka FL 32703
USA

Email: d3e3e3@gmail.com

Network Work group
Internet-Draft
Intended status: Standards Track
Expires: April 30, 2021

N. Nainar, Ed.
C. Pignataro, Ed.
Z. Ali
C. Filsfils
Cisco
T. Saad
Juniper
October 27, 2020

Segment Routing Generic TLV for MPLS Label Switched Path (LSP) Ping/
Traceroute
draft-nainar-mpls-spring-lsp-ping-sr-generic-sid-04

Abstract

RFC8402 introduces Segment Routing architecture that leverages source routing and tunneling paradigms and can be directly applied to the Multi Protocol Label Switching (MPLS) data plane. A node steers a packet through a controlled set of instructions called segments, by prepending the packet with Segment Routing header. SR architecture defines different types of segments with different forwarding semantics associated. SR can be applied to the MPLS directly and to IPv6 dataplane using a new routing header.

RFC8287 defines the extensions to MPLS LSP Ping and Traceroute for Segment Routing IGP-Prefix and IGP-Adjacency Segment Identifier (SIDs) with an MPLS data plane. Various SIDs are proposed as part of SR architecture with different associated instructions that raises a need to come up with new Target FEC Stack Sub-TLV for each such SIDs.

This document defines a new Target FEC Stack Sub-TLV that is used to validate the instruction associated with any SID.

Status of This Memo

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

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

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 April 30, 2021.

Copyright Notice

Copyright (c) 2020 IETF Trust and the persons identified as the document authors. All rights reserved.

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

Table of Contents

1. Introduction	2
1.1. Challenges with Existing Mechanism	3
2. Requirements notation	3
3. Terminology	4
4. Target FEC Stack sub-TLV for Segment Routing SID	4
4.1. Segment Routing Generic Label	4
4.2. FEC for Path validation	4
5. Procedures	5
5.1. SID to Interface Mapping	5
5.2. Initiator behavior	6
5.2.1. SRGL in Target FEC Stack TLV	6
5.3. Responder behavior	7
5.4. PHP flag behavior	7
6. IANA Considerations	8
6.1. New Target FEC Stack Sub-TLVs	8
6.2. Security Considerations	8
7. Acknowledgement	8
8. Contributors	8
9. References	8
9.1. Normative References	8
9.2. Informative References	9
Authors' Addresses	10

1. Introduction

[RFC8402] introduces and describes a Segment Routing architecture that leverages the source routing and tunneling paradigms. A node steers a packet through a controlled set of instructions called segments, by prepending the packet with Segment Routing header. A

detailed definition of the Segment Routing architecture is available in [RFC8402]

As described in [RFC8402] and [I-D.ietf-spring-segment-routing-mpls], the Segment Routing architecture can be directly applied to an MPLS data plane, the Segment identifier (Segment ID) will be of 20-bits size and the Segment Routing header is the label stack.

1.1. Challenges with Existing Mechanism

[RFC8287] defines the mechanism to perform LSP Ping and Traceroute for Segment Routing with MPLS data plane. [RFC8287] defines the Target FEC Stack Sub-TLVs for IGP-Prefix Segment ID and IGP-Adjacency Segment ID.

There are various other Segment IDs proposed by different documents that are applicable for SR architecture. [I-D.ietf-idr-bgp-prefix-sid] defines BGP Prefix Segment ID, [I-D.ietf-idr-bgppls-segment-routing-epe] defines BGP Peering Segment ID such as Peer Node SID, Peer Adj SID and Peer Set SID. [I-D.sivabalan-pce-binding-label-sid] defines Path Binding Segment ID. As SR evolves for different usecases, we may see more types of SIDs defined in the future. This raises a need to propose new Target FEC Stack Sub-TLV for each such Segment ID that may need specific or network wide software upgrade to support such new Target FEC Stack Sub-TLVs.

So instead of proposing different Target FEC Stack Sub-TLV for each SID, this document attempt to propose a SR Generic Label Sub-TLV for Target FEC Stack TLV with the procedure to validate the associated instruction.

This document describes the new Target FEC Stack Sub-TLV that carries the SID and the procedure to use LSP Ping and Traceroute using the new sub-tlv to support path validation and fault isolation for any SR Segment IDs. This document neither deprecates any existing Target FEC Stack Sub-TLVs nor precludes defining new Sub-TLVs in the future.

2. Requirements notation

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 RFC 2119 [RFC2119] RFC 8174 [RFC8174] when and only when, they appear in all capitals, as shown here.

3. Terminology

This document uses the terminologies defined in [RFC8402], [RFC8029], readers are expected to be familiar with it.

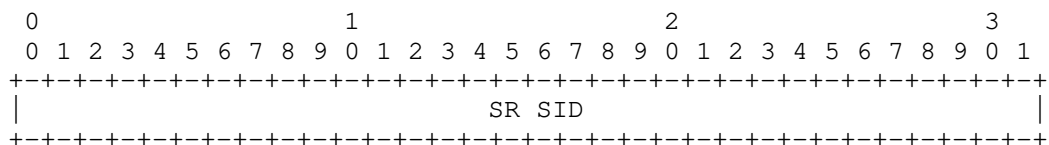
4. Target FEC Stack sub-TLV for Segment Routing SID

Following the procedure defined in [RFC8029], below defined Target FEC Stack Sub-TLV will be included for each labels in the stack. The below Sub-TLV is defined for Target FEC Stack TLV (Type 1), the Reverse-Path Target FEC Stack TLV (Type 16), and the Reply Path TLV (Type 21).

sub-Type	Value Field
TBD1	Segment Routing Generic Label (SRGL)

4.1. Segment Routing Generic Label

The format of the Sub-TLV is as specified below:



SR SID

Carries 20 bits of Segment ID that is used for validating the instruction.

4.2. FEC for Path validation

In SR architecture, any SID is associated with topology or service instruction. While the topology instruction steers the packet over best path or specific path, the service instruction instructs the type of service to be applied on the packet.

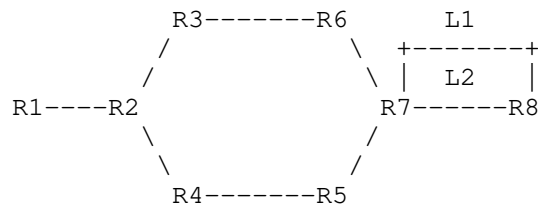


Figure 1: Segment Routing network

The Node Segment IDs for Rx for Algo 0 is 16000x. (Ex: For R1, it is 160001)
 The Node Segment IDs for Rx for Algo 128 is 16128x. (Ex: For R1, it is 161281)

9178 --> Adjacency Segment ID from R7 to R8 over link L1.
 9278 --> Adjacency Segment ID from R7 to R8 over link L2.
 9378 --> Parallel Adjacency Segment ID from R7 to R8 over Link L1 or L2.
 9187 --> Adjacency Segment ID from R8 to R7 over link L1.
 9287 --> Adjacency Segment ID from R8 to R7 over link L2.
 9387 --> Parallel Adjacency Segment ID from R8 to R7 over Link L1 or L2.

The instruction associated with any SID can be validated by verifying if the segment is terminated on the correct node and optionally received over the correct incoming interface. In Figure 1, in order to validate the SID 9178, R1 can use {(SID=9178)} as FEC in Target FEC Stack Sub-TLV.

5. Procedures

This section describes the procedure to validate SR Generic Label Sub-TLV.

5.1. SID to Interface Mapping

Any End point MAY maintain a SID to Interface mapping table that maintains the below:

- o All the local Prefix/Node SID with any SR enabled interface as incoming interface.
- o All the Adj-SIDs assigned by directly connected neighbor nodes with the relevant interface incoming interface.

In Figure 1, R8 maintains 160008 and 161288 with Incoming interface as any SR enabled interface. Similarly, R8 maintains 9178 with Link L1 as incoming interface, 9278 with Link L2 as incoming interface and 9378 with Link L1 or L2 as incoming interface.

How this mapping is populated and maintained is a local implementation matter. It can be populated based on the IGP database or can be based on a query to Path Computation Element (PCE) controller. The mapping can be persistent or on-demand triggered by receiving LSP Ping Request.

5.2. Initiator behavior

This section defines the Target FEC Stack TLV construction mechanism by an initiator when using SR Generic Label Sub-TLV.

Ping

Initiator MUST include FEC(s) corresponding to the destination segment.

Initiator MAY include FECs corresponding to some or all of segments imposed in the label stack by the initiator to communicate the segments traversed.

Traceroute

Initiator MUST initially include FECs corresponding to all of segments imposed in the label stack.

When a received echo reply contains FEC Stack Change TLV with one or more of original segment(s) being popped, initiator MAY remove corresponding FEC(s) from Target FEC Stack TLV in the next (TTL+1) traceroute request as defined in section 4.6 of [RFC8029].

When a received echo reply does not contain FEC Stack Change TLV, initiator MUST NOT attempt to remove FEC(s) from Target FEC Stack TLV in the next (TTL+1) traceroute request.

5.2.1. SRGL in Target FEC Stack TLV

When the last segment ID in the label stack is IGP Prefix SID, Adj-SID, Binding SID, BGP Prefix SID or BGP Peering SID, set the SR SID field to the Segment ID value advertised by the LSP End Point. When the SID is advertised as index, the Segment ID value MUST be derived based on the index and the SRGB advertised by the LSP End Point.

How the above values are derived is a local implementation matter. It can be manually defined using CLI knob while triggering the LSP Ping Request or can use other mechanisms like querying the local database.

5.3. Responder behavior

Step 4a defined in Section 7.4 of [RFC8287] is updated as below:

If the Label-stack-depth is 0 and Target FEC Stack Sub-TLV at FEC-stack-depth is TBD1 (SRGL) {

- * Set the Best-return-code to 10 when the responding node is not the LSP End Point for SR SID.
 - * Set the Best-return-code to 35, if Interface-I does not match the SID to Interface mapping for the received SR SID.
 - * set FEC-Status to 1, and return.
- }

If the Label-stack-depth is greater than 0 and Target FEC Stack Sub-TLV at FEC-stack-depth is TBD1 (SRGL), {

- * If the Label at Label-stack-depth is Imp-null {
 - + Set the Best-return-code to 10 when the responding node is not the LSP End Point for the SR SID.
 - + Set the Best-return-code to 35, if Interface-I does not match the SID to Interface mapping for the received SR SID.
 - + set FEC-Status to 1, and return.
- }
- * Else:
 - + Set the Best-return-code to 10 when the index derived from the label at Label-stack-depth is not advertised by LSP End Point.
 - + set FEC-Status to 1, and return.
- }

5.4. PHP flag behavior

Section 7.2 of [RFC8287] explains the behavior for FEC stack change for Adjacency Segment ID. The same procedure is applicable for BGP Peering SID as well.

6. IANA Considerations

6.1. New Target FEC Stack Sub-TLVs

IANA is requested to assign three new Sub-TLVs from "Sub-TLVs for TLV Types 1, 16 and 21" sub-registry from the "Multi-Protocol Label Switching (MPLS) Label Switched Paths (LSPs) Ping Parameters" [IANA-MPLS-LSP-PING] registry.

Sub-Type	Sub-TLV Name	Reference
TBD1	Segment Routing Generic Label	Section 4.1 of this document

6.2. Security Considerations

This document defines additional MPLS LSP Ping Sub-TLVs and follows the mechanisms defined in [RFC8029]. All the security considerations defined in [RFC8029] will be applicable for this document, and in addition, they do not impose any additional security challenges to be considered.

7. Acknowledgement

TBD

8. Contributors

Danial Johari, Cisco Systems

9. References

9.1. Normative References

[I-D.ietf-idr-bgp-prefix-sid]
Previdi, S., Filsfils, C., Lindem, A., Sreekantiah, A.,
and H. Gredler, "Segment Routing Prefix SID extensions for
BGP", draft-ietf-idr-bgp-prefix-sid-27 (work in progress),
June 2018.

[I-D.ietf-idr-bgppls-segment-routing-epe]
Previdi, S., Talaulikar, K., Filsfils, C., Patel, K., Ray,
S., and J. Dong, "BGP-LS extensions for Segment Routing
BGP Egress Peer Engineering", draft-ietf-idr-bgppls-
segment-routing-epe-19 (work in progress), May 2019.

- [I-D.sivabalan-pce-binding-label-sid]
Sivabalan, S., Filsfils, C., Tantsura, J., Hardwick, J.,
Previdi, S., and C. Li, "Carrying Binding Label/Segment-ID
in PCE-based Networks.", draft-sivabalan-pce-binding-
label-sid-07 (work in progress), July 2019.
- [RFC2119] Bradner, S., "Key words for use in RFCs to Indicate
Requirement Levels", BCP 14, RFC 2119,
DOI 10.17487/RFC2119, March 1997,
<<https://www.rfc-editor.org/info/rfc2119>>.
- [RFC8029] Kompella, K., Swallow, G., Pignataro, C., Ed., Kumar, N.,
Aldrin, S., and M. Chen, "Detecting Multiprotocol Label
Switched (MPLS) Data-Plane Failures", RFC 8029,
DOI 10.17487/RFC8029, March 2017,
<<https://www.rfc-editor.org/info/rfc8029>>.
- [RFC8174] Leiba, B., "Ambiguity of Uppercase vs Lowercase in RFC
2119 Key Words", BCP 14, RFC 8174, DOI 10.17487/RFC8174,
May 2017, <<https://www.rfc-editor.org/info/rfc8174>>.
- [RFC8287] Kumar, N., Ed., Pignataro, C., Ed., Swallow, G., Akiya,
N., Kini, S., and M. Chen, "Label Switched Path (LSP)
Ping/Traceroute for Segment Routing (SR) IGP-Prefix and
IGP-Adjacency Segment Identifiers (SIDs) with MPLS Data
Planes", RFC 8287, DOI 10.17487/RFC8287, December 2017,
<<https://www.rfc-editor.org/info/rfc8287>>.
- [RFC8402] Filsfils, C., Ed., Previdi, S., Ed., Ginsberg, L.,
Decraene, B., Litkowski, S., and R. Shakir, "Segment
Routing Architecture", RFC 8402, DOI 10.17487/RFC8402,
July 2018, <<https://www.rfc-editor.org/info/rfc8402>>.

9.2. Informative References

- [I-D.ietf-spring-segment-routing-mpls]
Bashandy, A., Filsfils, C., Previdi, S., Decraene, B.,
Litkowski, S., and R. Shakir, "Segment Routing with MPLS
data plane", draft-ietf-spring-segment-routing-mpls-22
(work in progress), May 2019.
- [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](http://www.iana.org/assignments/mpls-lsp-ping-parameters/mpls-lsp-ping-parameters.xhtml)>.

Authors' Addresses

Nagendra Kumar Nainar (editor)
Cisco Systems, Inc.
7200-12 Kit Creek Road
Research Triangle Park, NC 27709-4987
US

Email: naikumar@cisco.com

Carlos Pignataro (editor)
Cisco Systems, Inc.
7200-11 Kit Creek Road
Research Triangle Park, NC 27709-4987
US

Email: cpignata@cisco.com

Zafar Ali
Cisco Systems, Inc.

Email: zali@cisco.com

Clarence Filsfils
Cisco Systems, Inc.

Email: cfilsfil@cisco.com

Tarek Saad
Juniper Networks

Email: tsaad@juniper.net

Routing area
Internet-Draft
Intended status: Standards Track
Expires: May 6, 2021

D. Rath
K. Arora
S. Hegde
Juniper Networks Inc.
November 2, 2020

Egress TLV for Nil FEC in Label Switched Path Ping and Traceroute
Mechanisms
draft-rathi-mppls-egress-tlv-for-nil-fec-01

Abstract

Segment routing supports the creation of explicit paths using adjacency-sids, node-sids, and anycast-sids. The SR-TE paths are built by stacking the labels that represent the nodes and links in the explicit path. A very useful Operations And Maintenance (OAM) requirement is to be able to ping and trace these paths. A simple mppls ping/traceroute mechanism comprises of ability to traverse the SR-TE path without having to validate the control plane state. This document describes mppls ping and traceroute procedures using Nil FEC with additional extensions.

Requirements Language

The key words "MUST", "MUST NOT", "REQUIRED", "SHALL", "SHALL NOT", "SHOULD", "SHOULD NOT", "RECOMMENDED", "NOT RECOMMENDED", "MAY", and "OPTIONAL" in this document are to be interpreted as described in BCP 14 [RFC2119] [RFC8174] when, and only when, they appear in all capitals, as shown here.

Status of This Memo

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

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

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 May 6, 2021.

Copyright Notice

Copyright (c) 2020 IETF Trust and the persons identified as the document authors. All rights reserved.

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

Table of Contents

1. Introduction	2
2. Problem with Nil FEC	3
3. Egress TLV	3
4. Procedure	4
4.1. Sending Egress TLV in MPLS Echo Request	4
4.2. Receiving Egress TLV in MPLS Echo Request	5
5. Backward Compatibility	6
6. Security Considerations	6
7. IANA Considerations	6
7.1. New TLV	6
8. Acknowledgements	6
9. References	6
9.1. Normative References	6
9.2. Informative References	7
Authors' Addresses	7

1. Introduction

MPLS ping and traceroute mechanism as described in [RFC8029] and related extensions for SR as defined in [RFC8287] is very useful to precisely validate the control plane and data plane synchronization. It also provides ability to traverse multiple ECMP paths and validate each of the ECMP paths.

In certain usecases, the traffic engineered (TE) paths are built using mechanisms described in [I.D-ietf-spring-segment-routing-policy]. When the TE paths are built by the controller, the head-end routers may not have the complete database of the network and may not be aware of the FEC associated with labels that are used in the label stack. Use of Target FEC also requires all nodes in the network to have implemented

the validation procedures. All intermediate nodes may not have been upgraded to support validation procedures.

In such cases, it is useful to have ability to traverse the paths using ping and traceroute without having to obtain the Forwarding Equivalence Class (FEC) for each label. RFC 8029 supports this mechanism with Nil FEC. Nil FEC consists of the label and there is no other associated FEC information. The procedures described in RFC 8029 are mostly applicable when the Nil FEC is used where the Nil FEC is an intermediate FEC in the label stack. When all labels are represented using Nil FEC, it poses some challenges.

Section 2 discusses the problems associated with using all Nil FECs in a MPLS ping/traceroute procedure and Section 3 and Section 4 discusses simple extensions needed to solve the problem.

2. Problem with Nil FEC

The purpose of Nil FEC as described in [RFC8029] is to ensure hiding of transit tunnel information and in some cases to avoid false negatives when the FEC information is not known.

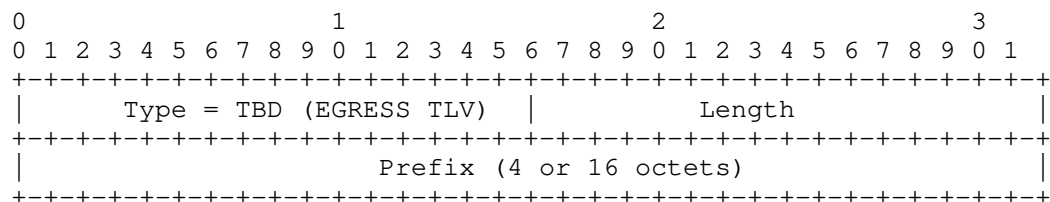
The MPLS ping/traceroute packet consists of only single Nil FEC corresponding to the complete label stack irrespective of number of segments in the label-stack. When router in the label-stack path receives MPLS ping/traceroute packets, there is no definite way to decide on whether its egress or transit since Nil FEC does not carry any information. So there is high possibility that the packet may be mis-forwarded to incorrect destination but the ping/traceroute might still show success.

To avoid this problem, there is a need to add additional information in the MPLS ping/traceroute packet along with Nil FEC that will help to do needed validation on each router of the label-stack path and sends proper information to ingress router on success and failure.

Thus it will be useful to add egress information in ping/traceroute packet that will help in validating Nil-FEC on each receiving router on label-stack path to ensure the correct destination.

3. Egress TLV

The Egress object is a TLV that MAY be included in an MPLS Echo Request message. Its an optional TLV and should appear before FEC-stack TLV in the MPLS Echo Request packet. In case multiple Nil FEC is present in Target FEC Stack TLV, Egress TLV should be added corresponding to the ultimate egress of the label-stack. The format is as specified below:



Type : TBD

Length : variable based on IPV4/IPV6 prefix. Length excludes the length of Type and length field. Length will be 4 octets for IPV4 and 16 octets for IPV6.

Prefix : This field carries the valid IPv4 prefix of length 4 octets or valid IPv6 Prefix of length 16 octets. It can be obtained from egress of Nil FEC corresponding to last label in the label-stack or SR-TE policy endpoint field [I.D-ietf-idr-segment-routing-te-policy].

4. Procedure

This section describes aspects of LSP Ping and Traceroute operations that require further considerations beyond [RFC8029].

4.1. Sending Egress TLV in MPLS Echo Request

As stated earlier, when the sender node builds a Echo Request with target FEC Stack TLV, Egress TLV SHOULD appear before Target FEC-stack TLV in MPLS Echo Request packet.

Ping

When the sender node builds a Echo Request with target FEC Stack TLV that contains a single Nil FEC corresponding to the last segment of the SR-TE path, sender node MUST add a Egress TLV with prefix obtained from SR-TE policy endpoint field [I.D-ietf-idr-segment-routing-te-policy] to indicate the egress for this Nil FEC in the Echo Request packet. In case endpoint is not specified or is equal to 0, sender MUST use the prefix corresponding to last segment of the SR-TE path as prefix for Egress TLV.

Traceroute

When the sender node builds a Echo Request with target FEC Stack TLV that contains a single Nil FEC corresponding to complete segment-list of the SR-TE path, sender node MUST add a Egress TLV with prefix obtained from SR-TE policy endpoint field

[I.D-ietf-idr-segment-routing-te-policy] to indicate the egress for this Nil FEC in the Echo Request packet. In case of multiple Nil FEC, Egress TLV SHOULD be added with prefix that indicate endpoint for last Nil-FEC corresponding to respective segment in label-stack. In case endpoint is not specified or is equal to 0, sender MUST use the prefix corresponding to the last segment endpoint of the SR-TE path i.e. ultimate egress as prefix for Egress TLV.

Consider the SR-TE policy configured with label-stack as 1001, 1002 , 1003 and end point as X on ingress router N1 to reach egress router N3. Segment 1003 belongs to N3 that has prefix X configured on it locally.

In Ping Echo Request, with target FEC Stack TLV that contains a single Nil FEC corresponding to 1003, should add Egress TLV for endpoint X with type as EGRESS-TLV, length depends on if X is IPv4 or IPv6 address and prefix as X.

In Traceroute Echo Request, with target FEC Stack TLV that contains a single Nil FEC corresponding to complete label-stack (1001, 1002, 1003) or multiple Nil-FEC corresponding to each label in label-stack, should add single Egress TLV for endpoint X with type as EGRESS-TLV, length depends on if X is IPv4 or IPv6 address and prefix as X or endpoint of segment 1003. In case X is not present or is set to 0, sender should use endpoint of segment 1003 as prefix for Egress TLV.

4.2. Receiving Egress TLV in MPLS Echo Request

No change in the processing for Nil FEC as defined in [RFC8029] in Target FEC stack TLV Node that receives an MPLS echo request.

Additional processing done for Egress TLV on receiver node as follows:

1. Get the prefix from the Egress TLV
2. Look up for an exact match of the prefix to any of locally configured interface as well as loopback address.
3. If the Label-stack-depth is 0 and the Target FEC Stack sub-TLV at FEC-stack-depth is Nil FEC and look up for EGRESS TLV prefix succeeds, set Best-return-code to 3 ("Replying router is an egress for the FEC at stack-depth") and Best-return-subcode to 1 to report egress ok in MPLS Echo Reply message.
4. If the Label-stack-depth greater than 0 and the Target FEC Stack sub-TLV at FEC-stack-depth is Nil FEC and look up for EGRESS TLV prefix fails, set Best-return-code to 8 ("Label switched at stack-

depth") and Best-return-subcode to Label-stack-depth to report transit switching in MPLS Echo Reply message.

5. If the Label-stack-depth is 0 and the Target FEC Stack sub-TLV at FEC-stack-depth is Nil FEC and look up for EGRESS TLV prefix fails, set the Best-return-code to 10, "Mapping for this FEC is not the given label at stack-depth" and Best-return-subcode to 1.

6. If the Label-stack-depth is greater than 0 and the Target FEC Stack sub-TLV at FEC-stack-depth is Nil FEC and look up for EGRESS TLV prefix succeeds, set the Best-return-code to 10, "Mapping for this FEC is not the given label at stack-depth" and Best-return-subcode to Label-stack-depth.

5. Backward Compatibility

The extension proposed in this document is backward compatible with procedures described in [RFC8029].

6. Security Considerations

TBD

7. IANA Considerations

7.1. New TLV

IANA need to assign new value for EGRESS TLV in the "Multi-Protocol Label Switching (MPLS) Label Switched Paths (LSPs) Ping Parameters" TLV registry [IANA].

EGRESS TLV : (TBD)

8. Acknowledgements

TBD.

9. References

9.1. Normative References

[I.D-ietf-idr-segment-routing-te-policy]
Filsfils, C., Ed., Previdi, S., Ed., Talaulikar, K.,
Mattes, P., Rosen, E., Jain, D., and S. Lin, "Advertising
Segment Routing Policies in BGP", draft-ietf-idr-segment-
routing-te-policy-09, work in progress, may 2020,
<[https://datatracker.ietf.org/doc/html/draft-ietf-idr-
segment-routing-te-policy-09](https://datatracker.ietf.org/doc/html/draft-ietf-idr-segment-routing-te-policy-09)>.

- [I.D-ietf-spring-segment-routing-policy]
Filsfils, C., Talaulikar, K., Bogdanov, A., Mattes, P.,
and D. Voyer, "Segment Routing Policy Architecture",
draft-ietf-spring-segment-routing-policy-08, work in
progress, July 2020,
<<https://datatracker.ietf.org/doc/html/draft-ietf-spring-segment-routing-policy-08>>.
- [RFC2119] Bradner, S., "Key words for use in RFCs to Indicate
Requirement Levels", BCP 14, RFC 2119,
DOI 10.17487/RFC2119, March 1997,
<<https://www.rfc-editor.org/info/rfc2119>>.
- [RFC8029] Kompella, K., Swallow, G., Pignataro, C., Ed., Kumar, N.,
Aldrin, S., and M. Chen, "Detecting Multiprotocol Label
Switched (MPLS) Data-Plane Failures", RFC 8029,
DOI 10.17487/RFC8029, March 2017,
<<https://www.rfc-editor.org/info/rfc8029>>.
- [RFC8174] Leiba, B., "Ambiguity of Uppercase vs Lowercase in RFC
2119 Key Words", BCP 14, RFC 8174, DOI 10.17487/RFC8174,
May 2017, <<https://www.rfc-editor.org/info/rfc8174>>.
- [RFC8287] Kumar, N., Ed., Pignataro, C., Ed., Swallow, G., Akiya,
N., Kini, S., and M. Chen, "Label Switched Path (LSP)
Ping/Traceroute for Segment Routing (SR) IGP-Prefix and
IGP-Adjacency Segment Identifiers (SIDs) with MPLS Data
Planes", RFC 8287, DOI 10.17487/RFC8287, December 2017,
<<https://www.rfc-editor.org/info/rfc8287>>.

9.2. Informative References

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

Authors' Addresses

Deepti N. Rath
Juniper Networks Inc.
Exora Business Park
Bangalore, KA 560103
India

Email: deeptir@juniper.net

Kapil Arora
Juniper Networks Inc.
Exora Business Park
Bangalore, KA 560103
India

Email: kapilaro@juniper.net

Shraddha Hegde
Juniper Networks Inc.
Exora Business Park
Bangalore, KA 560103
India

Email: shraddha@juniper.net

MPLS
Internet-Draft
Intended status: Informational
Expires: January 14, 2021

Q. Xiong
G. Mirsky
ZTE Corporation
W. Cheng
China Mobile
July 13, 2020

The Use of Path Segment in SR-MPLS and MPLS Interworking
draft-xiong-mpls-path-segment-sr-mpls-interworking-02

Abstract

This document illustrates the SR-MPLS and MPLS interworking scenarios to support end-to-end bidirectional tunnel across multiple domains with the use of Path Segments.

Status of This Memo

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

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

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

This Internet-Draft will expire on January 14, 2021.

Copyright Notice

Copyright (c) 2020 IETF Trust and the persons identified as the document authors. All rights reserved.

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

Table of Contents

1. Introduction	2
2. Conventions used in this document	3
2.1. Terminology	3
2.2. Requirements Language	4
3. SR-MPLS Interworking with MPLS	4
3.1. Stitching of Path Segments	5
3.2. Nesting of Path Segments	6
4. Security Considerations	7
5. Acknowledgements	7
6. IANA Considerations	7
7. Normative References	8
Authors' Addresses	8

1. Introduction

Segment Routing (SR) leverages the source routing paradigm. A node steers a packet through an SR Policy instantiated as an ordered list of instructions called "segments". SR supports a per-flow explicit routing while maintaining per-flow state only at the ingress nodes of the SR domain. Segment Routing can be instantiated on MPLS data plane which is referred to as SR-MPLS [RFC8660]. SR-MPLS leverages the MPLS label stack to construct the SR path.

IP/MPLS technology can be deployed in domains, which may serve as an access, aggregation, or core network. Further, using SR architecture, the IP/MPLS network may be upgraded to support the SR-MPLS technology. As such transformation is performed incrementally, by one domain at the time, operators are faced with a requirement to support the interworking between MPLS and SR-MPLS networks at the boundaries to provide the end-to-end bidirectional service. As defined in [RFC8402], the headend of an SR Policy binds a Binding Segment ID (B-SID) to its policy. The B-SID could be bound to a SID List or selected path and used to stitch the SR list and the SR Label Switched Paths (LSP) across multiple domains. The use of the B-SID is recommended to reduce the size of the label stack and stitch the SR LSPs.

In some scenarios, for example, a mobile backhaul transport network, it is required to provide end-to-end bidirectional path across SR and MPLS networks. The Path Segment as defined in [I-D.ietf-spring-mpls-path-segment] can be used to support bidirectional tunnel scenarios such as SR path Performance Measurement (PM), end-to-end 1+1 SR path protection and bidirectional SR paths correlation.

This document illustrates the SR-MPLS and MPLS interworking scenarios to support end-to-end bidirectional tunnel across multiple domains with the use of Path Segments.

2. Conventions used in this document

2.1. Terminology

ABR: Area Border Routers. Routers used to connect two IGP areas (areas in OSPF or levels in IS-IS).

AS: Autonomous System. An Autonomous System is composed by one or more IGP areas.

ASBR: Autonomous System Border Router. A router used to connect together ASes of the same or different service providers via one or more inter-AS links.

Border Node: An ABR that interconnects two or more IGP areas.

Border Link: Two ASes are interconnected with ASBRs.

B-SID: Binding Segment ID.

Domains: Autonomous System (AS) or IGP Area. An Autonomous System is composed of one or more IGP areas.

e-PSID: end-to-end Path Segment.

IGP: Interior Gateway Protocol.

N-PSID: Nesting of Path Segments.

PM: Performance Measurement.

SID: Segment ID.

SR: Segment Routing.

SR-MPLS: Segment Routing with MPLS data plane.

S-PSID: Stitching of Path Segments.

VPN: Virtual Private Network.

2.2. Requirements Language

The key words "MUST", "MUST NOT", "REQUIRED", "SHALL", "SHALL NOT", "SHOULD", "SHOULD NOT", "RECOMMENDED", "NOT RECOMMENDED", "MAY", and "OPTIONAL" in this document are to be interpreted as described in BCP 14 [RFC2119] [RFC8174] when, and only when, they appear in all capitals, as shown here.

3. SR-MPLS Interworking with MPLS

It is required to establish the end-to-end Virtual Private Network (VPN) service across the access network, aggregation network, and core network. For example, SR-MPLS may be deployed in access and core network, and MPLS may be deployed in the aggregation network. The network interworking should be taken into account in deployment are the following:

- o Border Node or Border Link
- o Stitching of Path Segments or Nesting of Path Segments
- o End-to-end Path Monitoring

The domains of the networks may be IGP Areas or ASes. The SR-MPLS and MPLS networks can be interconnected with a border node between IGP areas or border links between ASes. MPLS domain can be deployed between two SR-MPLS domains, as Figure 1 shows. The packets being transmitted along the SR path in SR-MPLS domains by using the SID list at the ingress node. And the path in MPLS domains can be pre-configuration either via NMS or via the MPLS control plane signaling. This document takes border node scenarios across IGP Areas domains for example. The border link scenarios are in future discussion.

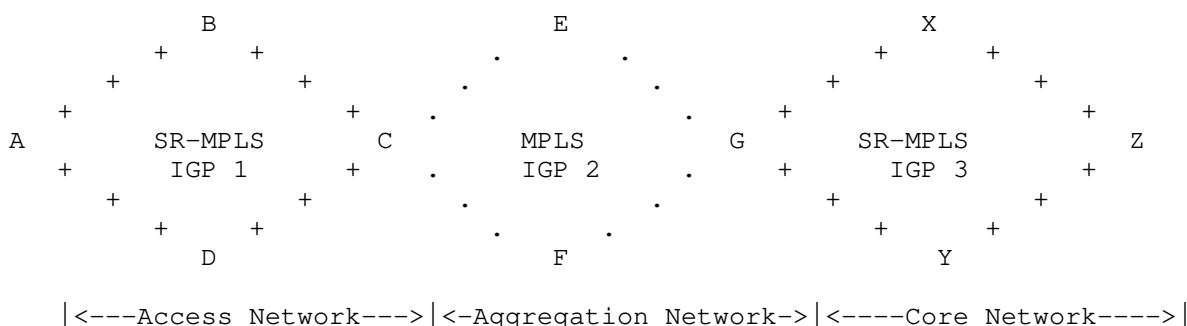


Figure 1: SR-MPLS and MPLS interworking Scenario

The VPN service across the SR-MPLS and MPLS domains is an end-to-end bidirectional path. In the SR-MPLS network, a Path Segment uniquely identifies an SR path and can be used for the end-to-end bidirectional path. This document illustrates the end-to-end Path Segment used in the interworking scenario including the stitching and nesting models. As described in [I-D.ietf-spring-mpls-path-segment], an end-to-end path segment or PSID (e-PSID), is also referred to as Nesting of Path SID (N-PSID) in nesting model or Stitching of Path SID (S-PSID) in stitching model.

3.1. Stitching of Path Segments

It is a common requirement that SR-MPLS needs to interwork with MPLS when SR is incrementally deployed in the MPLS domain. Figure 2 shows the stitching of Path Segments in SR-MPLS interworking with MPLS. The SR-LSPs and IP/MPLS LSPs are established independently in each domain which consist of SID list or MPLS label. The end-to-end bidirectional path acrossing the SR-MPLS and MPLS networks is split into multiple segments which can be identified by the S-PSID. The end-to-end path is terminated at the egress node in egress domain. The S-PSID will be popped out at the border node in each domain and correlated to the S-PSID of next domain.

The correlation of S-PSIDs can bind the segments of end-to-end path. The S-PSIDs are valid in the corresponding domain and the border nodes maintain the forwarding entries of that S-PSID segment that maps to the next S-PSID and the related path segments. In the headend node, the S-PSID can correlate the inter-domain path of reverse direction and bind the two unidirectional paths. The stitching of Path Segments can support the end-to-end path stitching and monitoring.

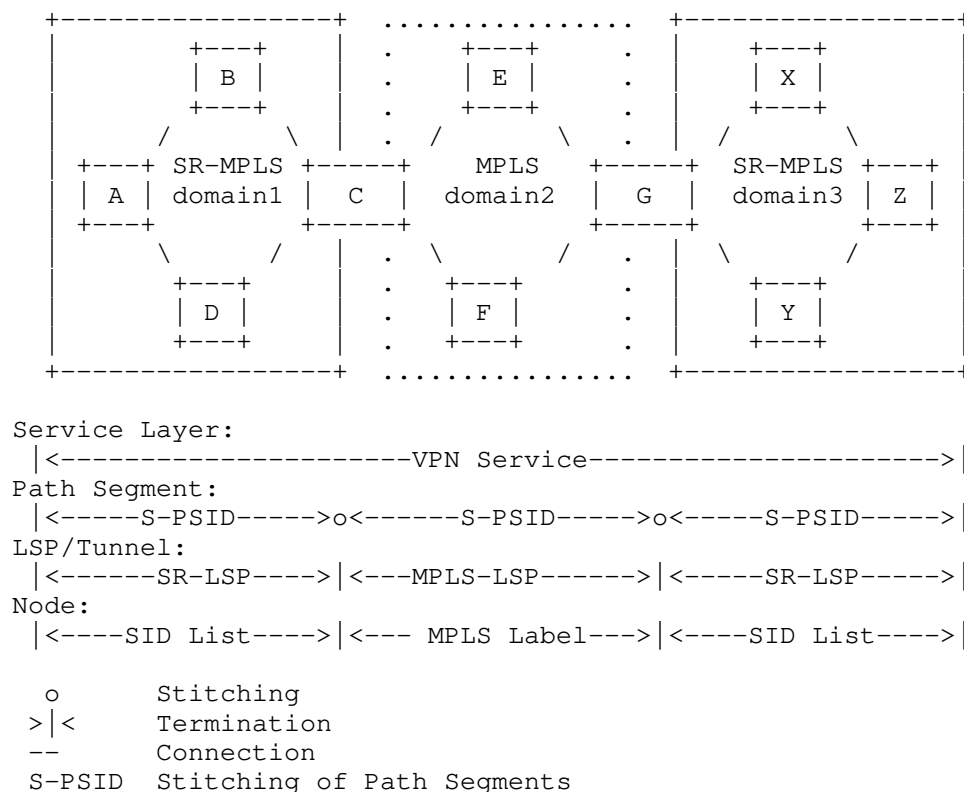


Figure 2: Stitching of Path Segments in SR-MPLS and MPLS interworking

3.2. Nesting of Path Segments

Figure 3 displays the nesting of Path Segments in SR-MPLS and MPLS interworking. The SR-LSPs and IP/MPLS LSPs are established in respective domain which consist of SID list or MPLS label. The SR-LSPs and IP/MPLS LSPs may be stitched across domains with B-SID. Comparing with S-PSID in the stitching model, the N-PSID presents end-to-end encapsulation in the packet from an SR-MPLS domain to an MPLS domain which is encapsulated at the ingress nodes and decapsulated at the egress nodes. The transit nodes, even the border nodes of domains, are not aware of the N-PSID.

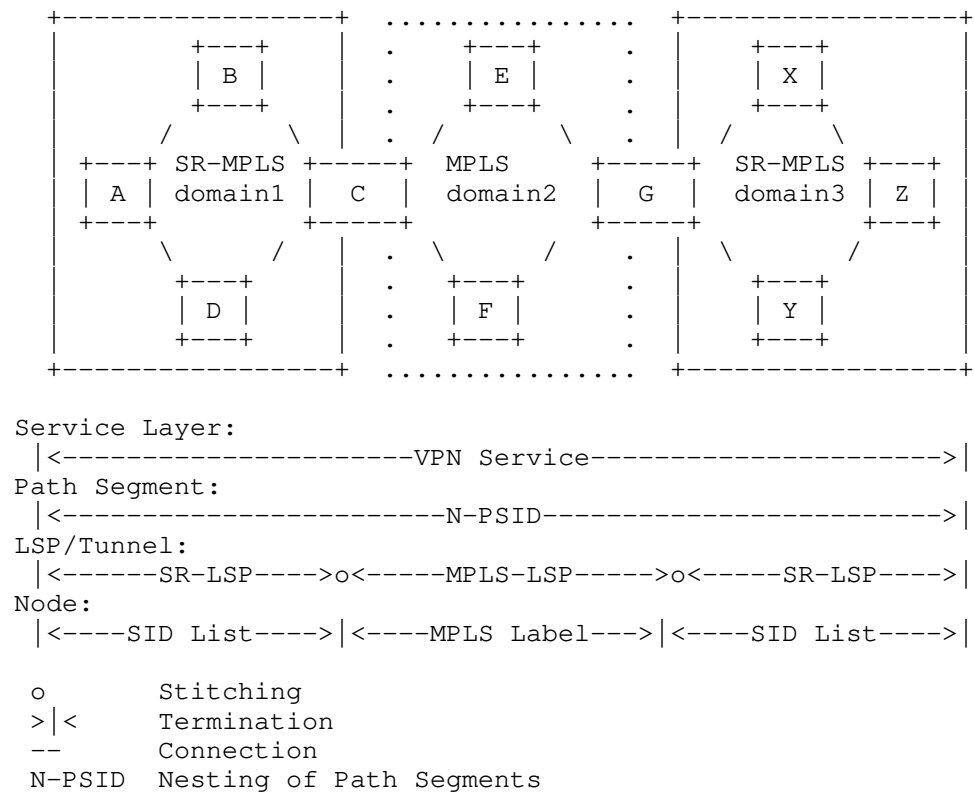


Figure 3: Nesting of Path Segments in SR-MPLS and MPLS interworking

4. Security Considerations

TBA

5. Acknowledgements

TBA

6. IANA Considerations

TBA

7. Normative References

- [I-D.ietf-spring-mpls-path-segment]
Cheng, W., Li, H., Chen, M., Gandhi, R., and R. Zigler,
"Path Segment in MPLS Based Segment Routing Network",
draft-ietf-spring-mpls-path-segment-02 (work in progress),
February 2020.
- [I-D.xiong-spring-path-segment-sr-inter-domain]
Xiong, Q., Mirsky, G., and W. Cheng, "The Use of Path
Segment in SR Inter-domain Scenarios", draft-xiong-spring-
path-segment-sr-inter-domain-01 (work in progress),
October 2019.
- [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>>.
- [RFC8402] Filsfils, C., Ed., Previdi, S., Ed., Ginsberg, L.,
Decraene, B., Litkowski, S., and R. Shakir, "Segment
Routing Architecture", RFC 8402, DOI 10.17487/RFC8402,
July 2018, <<https://www.rfc-editor.org/info/rfc8402>>.
- [RFC8660] Bashandy, A., Ed., Filsfils, C., Ed., Previdi, S.,
Decraene, B., Litkowski, S., and R. Shakir, "Segment
Routing with the MPLS Data Plane", RFC 8660,
DOI 10.17487/RFC8660, December 2019,
<<https://www.rfc-editor.org/info/rfc8660>>.

Authors' Addresses

Quan Xiong
ZTE Corporation
No.6 Huashi Park Rd
Wuhan, Hubei 430223
China

Phone: +86 27 83531060
Email: xiong.quan@zte.com.cn

Greg Mirsky
ZTE Corporation
USA

Email: gregimirsky@gmail.com

Weiqiang Cheng
China Mobile
Beijing
China

Email: chengweiqiang@chinamobile.com

MPLS Working Group
Internet-Draft
Intended status: Standards Track
Expires: May 3, 2021

X. Min
P. ShaoFu
ZTE Corp.
October 30, 2020

Label Switched Path (LSP) Ping/Traceroute for Segment Routing (SR) Path
Segment Identifiers (SIDs) with MPLS Data Planes
draft-xp-mpls-spring-lsp-ping-path-sid-00

Abstract

Path Segment is a type of SR segment, which is used to identify an SR path. This document provides Target Forwarding Equivalence Class (FEC) stack TLV definitions for Path Segment Identifiers.

Status of This Memo

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

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

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 May 3, 2021.

Copyright Notice

Copyright (c) 2020 IETF Trust and the persons identified as the document authors. All rights reserved.

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

Table of Contents

1. Introduction	2
2. Conventions	2
2.1. Requirements Language	2
2.2. Terminology	3
3. Path Segment ID Sub-TLV	3
3.1. SR Candidate Path's Path SID	3
3.2. SR Segment List's Path SID	6
4. Path-SID FEC Validation	8
5. Security Considerations	14
6. IANA Considerations	14
7. Acknowledgements	15
8. References	15
8.1. Normative References	15
8.2. Informative References	16
Authors' Addresses	17

1. Introduction

Path Segment is a type of SR segment, which is used to identify an SR path. Path Segment in MPLS based segment routing network is defined in [I-D.ietf-spring-mpls-path-segment].

When Path Segment is used, it's inserted by the ingress node of the SR path, and then processed by the egress node of the SR path. The position of Path Segment Label within the MPLS label stack is immediately following the segment list of the SR path. Note that the Path Segment would not be popped up until it reaches the egress node.

This document provides Target Forwarding Equivalence Class (FEC) stack TLV definitions for Path-SIDs. Procedures for LSP Ping and Traceroute as defined in [RFC8287] and [RFC8690] are applicable to Path-SIDs as well.

2. Conventions

2.1. Requirements Language

The key words "MUST", "MUST NOT", "REQUIRED", "SHALL", "SHALL NOT", "SHOULD", "SHOULD NOT", "RECOMMENDED", "NOT RECOMMENDED", "MAY", and "OPTIONAL" in this document are to be interpreted as described in BCP 14 [RFC2119] [RFC8174] when, and only when, they appear in all capitals, as shown here.

2.2. Terminology

This document uses the terminology defined in [RFC8402] and [RFC8029], readers are expected to be familiar with those terms.

3. Path Segment ID Sub-TLV

Analogous to what's defined in Section 5 of [RFC8287] and Section 4 of [I-D.ietf-mppls-sr-epe-oam], two new sub-TLVs are defined for the Target FEC Stack TLV (Type 1), the Reverse-Path Target FEC Stack TLV (Type 16), and the Reply Path TLV (Type 21).

Sub-Type	Sub-TLV Name
TBD1	SR Candidate Path's Path SID
TBD2	SR Segment List's Path SID

As specified in Section 3 of [I-D.ietf-idr-sr-policy-path-segment], the Path Segment can be used to identify an SR path (specified by SID list) or an SR candidate path, so two different Target FEC sub-TLVs need to be defined for Path Segment ID. When a Path Segment is used to identify an SR path, then the Target FEC sub-TLV of SR Segment List's Path SID would be used to validate the control plane to forwarding plane synchronization for this Path-SID; When a Path Segment is used to identify an SR candidate path, then the Target FEC sub-TLV of SR Candidate Path's Path SID would be used to validate the control plane to forwarding plane synchronization for this Path-SID.

3.1. SR Candidate Path's Path SID

The format of SR Candidate Path's Path SID sub-TLV is as specified below:

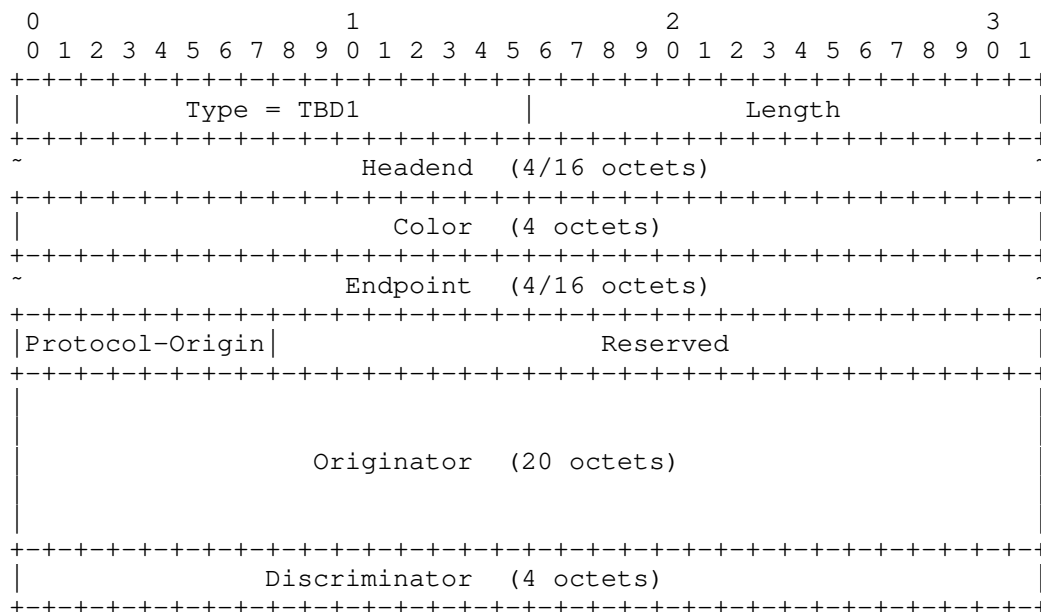


Figure 1: SR Candidate Path's Path SID sub-TLV

Type

This field is set to the value (TBD1) which indicates that it's a SR Candidate Path's Path SID sub-TLV.

Length

This field is set to the length of the sub-TLV's Value field in octets. If Headend and Endpoint fields are in IPv4 address format which is 4 octets long, it MUST be set to 40; If Headend and Endpoint fields are in IPv6 address format which is 16 octets long, it MUST be set to 64.

Headend

This field identifies the headend of an SR Policy, the same as defined in Section 2.1 of [I-D.ietf-spring-segment-routing-policy]. The headend is a 4-octet IPv4 address or a 16-octet IPv6 address.

Color

This field associates the SR Policy with an intent, the same as defined in Section 2.1 of [I-D.ietf-spring-segment-routing-policy]. The color is a 4-octet numerical value.

Endpoint

This field identifies the endpoint of an SR Policy, the same as defined in Section 2.1 of [I-D.ietf-spring-segment-routing-policy]. The endpoint is a 4-octet IPv4 address or a 16-octet IPv6 address.

Protocol-Origin

This field identifies the component or protocol that originates or signals the candidate path for an SR Policy, the same as defined in Section 2.3 of [I-D.ietf-spring-segment-routing-policy]. The protocol-origin is a 1-octet value that follows the recommendation from Table 1 of Section 2.3 of [I-D.ietf-spring-segment-routing-policy], which specifies value 10 for "PCEP", value 20 for "BGP SR Policy" and value 30 for "Via Configuration".

Originator

This field identifies the node which provisioned or signaled the candidate path for an SR Policy, the same as defined in Section 2.4 of [I-D.ietf-spring-segment-routing-policy]. The originator is a 20-octet numerical value formed by the concatenation of the fields of the tuple <ASN, node-address>, among which ASN is a 4-octet number and node address is a 16-octet value (an IPv6 address or an IPv4 address encoded in the lowest 4 octets). When Protocol-Origin is respectively "Via Configuration", or "PCEP", or "BGP SR Policy", the values of ASN and node address follow the specification in Section 2.4 of [I-D.ietf-spring-segment-routing-policy].

Discriminator

This field uniquely identifies a candidate path within the context of an SR policy, the same as defined in Section 2.5 of [I-D.ietf-spring-segment-routing-policy]. The discriminator is a 4-octet value. When Protocol-Origin is respectively "Via Configuration", or "PCEP", or "BGP SR Policy", the value of discriminator follows the specification in Section 2.5 of [I-D.ietf-spring-segment-routing-policy].

3.2. SR Segment List's Path SID

The format of SR Segment List's Path SID sub-TLV is as specified below:

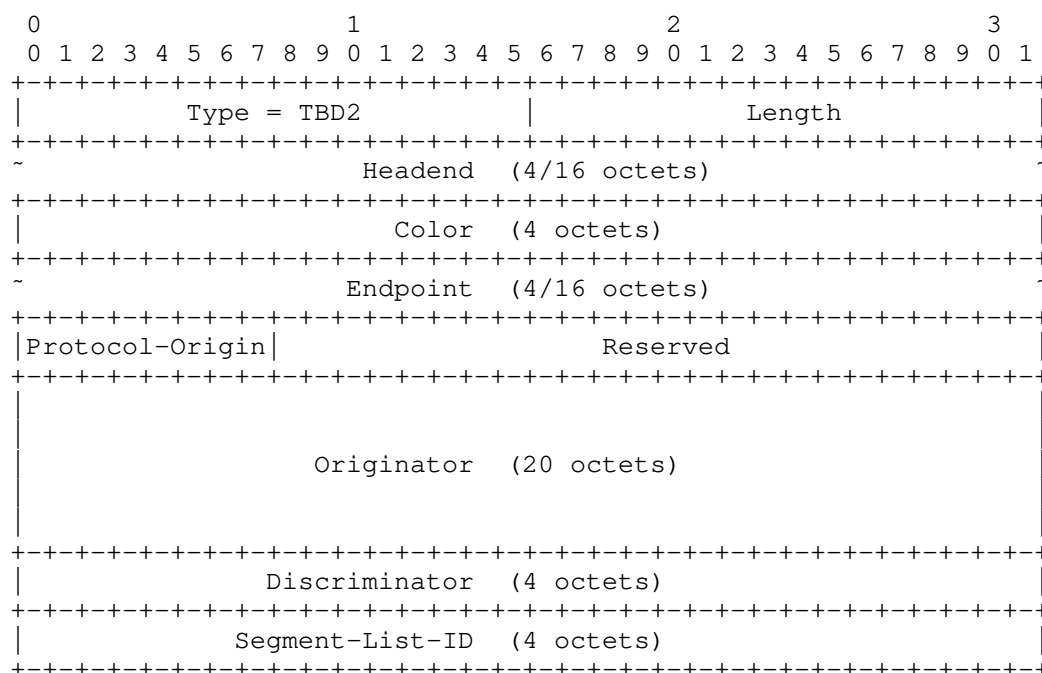


Figure 2: SR Segment List's Path SID sub-TLV

Type

This field is set to the value (TBD2) which indicates that it's a SR Segment List's Path SID sub-TLV.

Length

This field is set to the length of the sub-TLV's Value field in octets. If Headend and Endpoint fields are in IPv4 address format which is 4 octets long, it MUST be set to 44; If Headend and Endpoint fields are in IPv6 address format which is 16 octets long, it MUST be set to 68.

Headend

This field identifies the headend of an SR Policy, the same as defined in Section 2.1 of [I-D.ietf-spring-segment-routing-policy]. The headend is a 4-octet IPv4 address or a 16-octet IPv6 address.

Color

This field associates the SR Policy with an intent, the same as defined in Section 2.1 of [I-D.ietf-spring-segment-routing-policy]. The color is a 4-octet numerical value.

Endpoint

This field identifies the endpoint of an SR Policy, the same as defined in Section 2.1 of [I-D.ietf-spring-segment-routing-policy]. The endpoint is a 4-octet IPv4 address or a 16-octet IPv6 address.

Protocol-Origin

This field identifies the component or protocol that originates or signals the candidate path for an SR Policy, the same as defined in Section 2.3 of [I-D.ietf-spring-segment-routing-policy]. The protocol-origin is a 1-octet value that follows the recommendation from Table 1 of Section 2.3 of [I-D.ietf-spring-segment-routing-policy], which specifies value 10 for "PCEP", value 20 for "BGP SR Policy" and value 30 for "Via Configuration".

Originator

This field identifies the node which provisioned or signaled the candidate path for an SR Policy, the same as defined in Section 2.4 of [I-D.ietf-spring-segment-routing-policy]. The originator is a 20-octet numerical value formed by the concatenation of the fields of the tuple <ASN, node-address>, among which ASN is a 4-octet number and node address is a 16-octet value (an IPv6 address or an IPv4 address encoded in the lowest 4 octets). When Protocol-Origin is respectively "Via Configuration", or "PCEP", or "BGP SR Policy", the values of ASN and node address follow the specification in Section 2.4 of [I-D.ietf-spring-segment-routing-policy].

Discriminator

This field uniquely identifies a candidate path within the context of an SR policy, the same as defined in Section 2.5 of [I-D.ietf-spring-segment-routing-policy]. The discriminator is a 4-octet value. When Protocol-Origin is respectively "Via Configuration", or "PCEP", or "BGP SR Policy", the value of discriminator follows the specification in Section 2.5 of [I-D.ietf-spring-segment-routing-policy].

Segment-List-ID

This field identifies an SR path within the context of a candidate path of an SR Policy, the same as "List Identifier" defined in Section 2.2 of [I-D.lsp-idr-sr-path-protection]. The segment-list-ID is a 4-octet identifier of the corresponding segment list.

4. Path-SID FEC Validation

The MPLS LSP Ping/Traceroute procedures MAY be initiated by the headend of the Segment Routing path or a centralized topology-aware data plane monitoring system as described in [RFC8403]. For the Path-SID, the responder nodes that receive echo request and send echo reply MUST be the endpoint of the Segment Routing path.

When an endpoint receives the LSP echo request packet with top FEC being the Path-SID, it SHOULD perform validity checks on the content of the Path-SID FEC sub-TLV. The basic length check should be performed on the received FEC.

SR Candidate Path's Path SID

Length = 40 or 64

SR Segment List's Path SID

Length = 44 or 68

If a malformed FEC sub-TLV is received, then a return code of 1, "Malformed echo request received" as defined in [RFC8029] SHOULD be sent. The below section augments the section 7.4 of [RFC8287].

4a. Segment Routing Path-SID Validation:

If the Label-stack-depth is 0 and the Target FEC Stack sub-TLV at FEC-stack-depth is TBD1 (SR Candidate Path's Path SID sub-TLV), {

Set the Best-return-code to 10, "Mapping for this FEC is not the given label at stack-depth <RSC>" if any below conditions fail:

- + Validate that the Path Segment ID is signaled or provisioned for the SR Candidate Path {
 - When the Protocol-Origin field in the received SR Candidate Path's Path SID sub-TLV is 10, "PCEP" is used as the signaling protocol. And then validate that the Path Segment ID matches with the tuples identifying the SR Candidate Path within PCEP {
 - o Validate that the signaled headend defined in [I-D.ietf-pce-segment-routing-policy-cp] and [I-D.ietf-pce-sr-path-segment], for the Path SID, matches with the headend field in the received SR Candidate Path's Path SID sub-TLV.
 - o Validate that the signaled color defined in [I-D.ietf-pce-segment-routing-policy-cp] and [I-D.ietf-pce-sr-path-segment], for the Path SID, matches with the color field in the received SR Candidate Path's Path SID sub-TLV.
 - o Validate that the signaled end-point defined in [I-D.ietf-pce-segment-routing-policy-cp] and [I-D.ietf-pce-sr-path-segment], for the Path SID, matches with the endpoint field in the received SR Candidate Path's Path SID sub-TLV.
 - o Validate that the signaled both originator ASN and originator address defined in [I-D.ietf-pce-segment-routing-policy-cp] and [I-D.ietf-pce-sr-path-segment], for the Path SID, matches with the originator field in the received SR Candidate Path's Path SID sub-TLV.
 - o Validate that the signaled discriminator defined in [I-D.ietf-pce-segment-routing-policy-cp] and [I-D.ietf-pce-sr-path-segment], for the Path SID, matches with the discriminator field in the received SR Candidate Path's Path SID sub-TLV.
 - }
 - When the Protocol-Origin field in the received SR Candidate Path's Path SID sub-TLV is 20, "BGP SR Policy"

is used as the signaling protocol. And then validate that the Path Segment ID matches with the tuples identifying the SR Candidate Path within BGP SR Policy {

- o Validate that the signaled headend defined in [I-D.ietf-idr-segment-routing-te-policy] and [I-D.ietf-idr-sr-policy-path-segment], for the Path SID, matches with the headend field in the received SR Candidate Path's Path SID sub-TLV.
- o Validate that the signaled policy color defined in [I-D.ietf-idr-segment-routing-te-policy] and [I-D.ietf-idr-sr-policy-path-segment], for the Path SID, matches with the color field in the received SR Candidate Path's Path SID sub-TLV.
- o Validate that the signaled endpoint defined in [I-D.ietf-idr-segment-routing-te-policy] and [I-D.ietf-idr-sr-policy-path-segment], for the Path SID, matches with the endpoint field in the received SR Candidate Path's Path SID sub-TLV.
- o Validate that the signaled both ASN and BGP Router-ID defined in [I-D.ietf-idr-segment-routing-te-policy] and [I-D.ietf-idr-sr-policy-path-segment], for the Path SID, matches with the originator field in the received SR Candidate Path's Path SID sub-TLV.
- o Validate that the signaled distinguisher defined in [I-D.ietf-idr-segment-routing-te-policy] and [I-D.ietf-idr-sr-policy-path-segment], for the Path SID, matches with the discriminator field in the received SR Candidate Path's Path SID sub-TLV.

}

- When the Protocol-Origin field in the received SR Candidate Path's Path SID sub-TLV is 30, "Via Configuration" is used. And then validate that the Path Segment ID matches with the tuples identifying the SR Candidate Path within Configuration {
 - o Validate that the provisioned headend defined in [I-D.ietf-spring-sr-policy-yang], for the Path SID, matches with the headend field in the received SR Candidate Path's Path SID sub-TLV.

- o Validate that the provisioned color defined in [I-D.ietf-spring-sr-policy-yang], for the Path SID, matches with the color field in the received SR Candidate Path's Path SID sub-TLV.
 - o Validate that the provisioned endpoint defined in [I-D.ietf-spring-sr-policy-yang], for the Path SID, matches with the endpoint field in the received SR Candidate Path's Path SID sub-TLV.
 - o Validate that the provisioned originator defined in [I-D.ietf-spring-sr-policy-yang], for the Path SID, matches with the originator field in the received SR Candidate Path's Path SID sub-TLV.
 - o Validate that the provisioned discriminator defined in [I-D.ietf-spring-sr-policy-yang], for the Path SID, matches with the discriminator field in the received SR Candidate Path's Path SID sub-TLV.
- }
- }

If all the above validations have passed, set the return code to 3 "Replying router is an egress for the FEC at stack-depth <RSC>".

Set FEC-Status to 1 and return.

}

Else, if the Label-stack-depth is 0 and the Target FEC Stack sub-TLV at FEC-stack-depth is TBD2 (SR Segment List's Path SID sub-TLV), {

Set the Best-return-code to 10, "Mapping for this FEC is not the given label at stack-depth <RSC>" if any below conditions fail:

- + Validate that the Path Segment ID is signaled or provisioned for the SR Segment List {
 - When the Protocol-Origin field in the received SR Segment List's Path SID sub-TLV is 10, "PCEP" is used as the signaling protocol. And then validate that the Path Segment ID matches with the tuples identifying the SR Segment List within PCEP {

- o Validate that the signaled headend defined in [I-D.ietf-pce-segment-routing-policy-cp] and [I-D.ietf-pce-sr-path-segment], for the Path SID, matches with the headend field in the received SR Segment List's Path SID sub-TLV.
 - o Validate that the signaled color defined in [I-D.ietf-pce-segment-routing-policy-cp] and [I-D.ietf-pce-sr-path-segment], for the Path SID, matches with the color field in the received SR Segment List's Path SID sub-TLV.
 - o Validate that the signaled end-point defined in [I-D.ietf-pce-segment-routing-policy-cp] and [I-D.ietf-pce-sr-path-segment], for the Path SID, matches with the endpoint field in the received SR Segment List's Path SID sub-TLV.
 - o Validate that the signaled both originator ASN and originator address defined in [I-D.ietf-pce-segment-routing-policy-cp] and [I-D.ietf-pce-sr-path-segment], for the Path SID, matches with the originator field in the received SR Segment List's Path SID sub-TLV.
 - o Validate that the signaled discriminator defined in [I-D.ietf-pce-segment-routing-policy-cp] and [I-D.ietf-pce-sr-path-segment], for the Path SID, matches with the discriminator field in the received SR Segment List's Path SID sub-TLV.
 - o Validate that the signaled Segment-List-ID by PCEP, for the Path SID, matches with the Segment-List-ID field in the received SR Segment List's Path SID sub-TLV.
- }
- When the Protocol-Origin field in the received SR Segment List's Path SID sub-TLV is 20, "BGP SR Policy" is used as the signaling protocol. And then validate that the Path Segment ID matches with the tuples identifying the SR Segment List within BGP SR Policy {
 - o Validate that the signaled headend defined in [I-D.ietf-idr-segment-routing-te-policy] and [I-D.ietf-idr-sr-policy-path-segment], for the Path

SID, matches with the headend field in the received SR Segment List's Path SID sub-TLV.

- o Validate that the signaled policy color defined in [I-D.ietf-idr-segment-routing-te-policy] and [I-D.ietf-idr-sr-policy-path-segment], for the Path SID, matches with the color field in the received SR Segment List's Path SID sub-TLV.
 - o Validate that the signaled endpoint defined in [I-D.ietf-idr-segment-routing-te-policy] and [I-D.ietf-idr-sr-policy-path-segment], for the Path SID, matches with the endpoint field in the received SR Segment List's Path SID sub-TLV.
 - o Validate that the signaled both ASN and BGP Router-ID defined in [I-D.ietf-idr-segment-routing-te-policy] and [I-D.ietf-idr-sr-policy-path-segment], for the Path SID, matches with the originator field in the received SR Segment List's Path SID sub-TLV.
 - o Validate that the signaled distinguisher defined in [I-D.ietf-idr-segment-routing-te-policy] and [I-D.ietf-idr-sr-policy-path-segment], for the Path SID, matches with the discriminator field in the received SR Segment List's Path SID sub-TLV.
 - o Validate that the signaled List Identifier defined in [I-D.lsp-idr-sr-path-protection], for the Path SID, matches with the Segment-List-ID field in the received SR Segment List's Path SID sub-TLV.
- }
- When the Protocol-Origin field in the received SR Segment List's Path SID sub-TLV is 30, "Via Configuration" is used. And then validate that the Path Segment ID matches with the tuples identifying the SR Segment List within Configuration {
 - o Validate that the provisioned headend defined in [I-D.ietf-spring-sr-policy-yang], for the Path SID, matches with the headend field in the received SR Segment List's Path SID sub-TLV.
 - o Validate that the provisioned color defined in [I-D.ietf-spring-sr-policy-yang], for the Path SID,

matches with the color field in the received SR Segment List's Path SID sub-TLV.

- o Validate that the provisioned endpoint defined in [I-D.ietf-spring-sr-policy-yang], for the Path SID, matches with the endpoint field in the received SR Segment List's Path SID sub-TLV.
- o Validate that the provisioned originator defined in [I-D.ietf-spring-sr-policy-yang], for the Path SID, matches with the originator field in the received SR Segment List's Path SID sub-TLV.
- o Validate that the provisioned discriminator defined in [I-D.ietf-spring-sr-policy-yang], for the Path SID, matches with the discriminator field in the received SR Segment List's Path SID sub-TLV.
- o Validate that the provisioned Segment-List-ID through Yang Model, for the Path SID, matches with the Segment-List-ID field in the received SR Segment List's Path SID sub-TLV.

}

}

If all the above validations have passed, set the return code to 3 "Replying router is an egress for the FEC at stack-depth <RSC>".

Set FEC-Status to 1 and return.

}

5. Security Considerations

This document does not raise any additional security issues beyond those of the specifications referred to in the list of normative references.

6. IANA Considerations

IANA is requested to assign two new sub-TLVs from the "sub-TLVs for TLV Types 1, 16, and 21" subregistry of the "Multi-Protocol Label Switching (MPLS) Label Switched Paths (LSPs) Ping Parameters" registry [IANA].

Sub-Type	Sub-TLV Name	Reference
TBD1	SR Candidate Path's Path SID	Section 3.1
TBD2	SR Segment List's Path SID	Section 3.2

7. Acknowledgements

The authors would like to acknowledge Zhao Detao for his thorough review and very helpful comments.

8. References

8.1. Normative References

- [I-D.ietf-spring-mpls-path-segment]
Cheng, W., Li, H., Chen, M., Gandhi, R., and R. Zigler,
"Path Segment in MPLS Based Segment Routing Network",
draft-ietf-spring-mpls-path-segment-03 (work in progress),
September 2020.
- [RFC2119] Bradner, S., "Key words for use in RFCs to Indicate
Requirement Levels", BCP 14, RFC 2119,
DOI 10.17487/RFC2119, March 1997,
<<https://www.rfc-editor.org/info/rfc2119>>.
- [RFC8029] Kompella, K., Swallow, G., Pignataro, C., Ed., Kumar, N.,
Aldrin, S., and M. Chen, "Detecting Multiprotocol Label
Switched (MPLS) Data-Plane Failures", RFC 8029,
DOI 10.17487/RFC8029, March 2017,
<<https://www.rfc-editor.org/info/rfc8029>>.
- [RFC8174] Leiba, B., "Ambiguity of Uppercase vs Lowercase in RFC
2119 Key Words", BCP 14, RFC 8174, DOI 10.17487/RFC8174,
May 2017, <<https://www.rfc-editor.org/info/rfc8174>>.
- [RFC8287] Kumar, N., Ed., Pignataro, C., Ed., Swallow, G., Akiya,
N., Kini, S., and M. Chen, "Label Switched Path (LSP)
Ping/Traceroute for Segment Routing (SR) IGP-Prefix and
IGP-Adjacency Segment Identifiers (SIDs) with MPLS Data
Planes", RFC 8287, DOI 10.17487/RFC8287, December 2017,
<<https://www.rfc-editor.org/info/rfc8287>>.
- [RFC8690] Nainar, N., Pignataro, C., Iqbal, F., and A. Vainshtein,
"Clarification of Segment ID Sub-TLV Length for RFC 8287",
RFC 8690, DOI 10.17487/RFC8690, December 2019,
<<https://www.rfc-editor.org/info/rfc8690>>.

8.2. Informative References

- [I-D.ietf-idr-segment-routing-te-policy]
Previdi, S., Filsfils, C., Talaulikar, K., Mattes, P., Rosen, E., Jain, D., and S. Lin, "Advertising Segment Routing Policies in BGP", draft-ietf-idr-segment-routing-te-policy-09 (work in progress), May 2020.
- [I-D.ietf-idr-sr-policy-path-segment]
Li, C., Li, Z., Telecom, C., Cheng, W., and K. Talaulikar, "SR Policy Extensions for Path Segment and Bidirectional Path", draft-ietf-idr-sr-policy-path-segment-01 (work in progress), August 2020.
- [I-D.ietf-mpls-sr-epe-oam]
Hegde, S., Arora, K., Srivastava, M., Ninan, S., and X. Xu, "Label Switched Path (LSP) Ping/Traceroute for Segment Routing (SR) Egress Peer Engineering Segment Identifiers (SIDs) with MPLS Data Planes", draft-ietf-mpls-sr-epe-oam-00 (work in progress), June 2020.
- [I-D.ietf-pce-segment-routing-policy-cpl]
Koldychev, M., Sivabalan, S., Barth, C., Peng, S., and H. Bidgoli, "PCEP extension to support Segment Routing Policy Candidate Paths", draft-ietf-pce-segment-routing-policy-cp-01 (work in progress), October 2020.
- [I-D.ietf-pce-sr-path-segment]
Li, C., Chen, M., Cheng, W., Gandhi, R., and Q. Xiong, "Path Computation Element Communication Protocol (PCEP) Extension for Path Segment in Segment Routing (SR)", draft-ietf-pce-sr-path-segment-01 (work in progress), May 2020.
- [I-D.ietf-spring-segment-routing-policy]
Filsfils, C., Talaulikar, K., Voyer, D., Bogdanov, A., and P. Mattes, "Segment Routing Policy Architecture", draft-ietf-spring-segment-routing-policy-08 (work in progress), July 2020.
- [I-D.ietf-spring-sr-policy-yang]
Raza, K., Sawaya, R., Shunwan, Z., Voyer, D., Durrani, M., Matsushima, S., and V. Beeram, "YANG Data Model for Segment Routing Policy", draft-ietf-spring-sr-policy-yang-00 (work in progress), September 2020.

[I-D.lp-idr-sr-path-protection]

Yao, L. and S. Peng, "BGP extensions of SR policy for path protection", draft-lp-idr-sr-path-protection-00 (work in progress), October 2020.

[IANA]

Internet Assigned Numbers Authority (IANA), "Multi-Protocol Label Switching (MPLS) Label Switched Paths (LSPs) Ping Parameters", <<http://www.iana.org/assignments/mppls-lsp-ping-parameters>>.

[RFC8402]

Filsfils, C., Ed., Previdi, S., Ed., Ginsberg, L., Decraene, B., Litkowski, S., and R. Shakir, "Segment Routing Architecture", RFC 8402, DOI 10.17487/RFC8402, July 2018, <<https://www.rfc-editor.org/info/rfc8402>>.

[RFC8403]

Geib, R., Ed., Filsfils, C., Pignataro, C., Ed., and N. Kumar, "A Scalable and Topology-Aware MPLS Data-Plane Monitoring System", RFC 8403, DOI 10.17487/RFC8403, July 2018, <<https://www.rfc-editor.org/info/rfc8403>>.

Authors' Addresses

Xiao Min
ZTE Corp.
Nanjing
China

Email: xiao.min2@zte.com.cn

Peng Shaofu
ZTE Corp.
Nanjing
China

Email: peng.shaofu@zte.com.cn

tswwg
Internet-Draft
Intended status: Standards Track
Expires: May 5, 2021

Z. Zhang
R. Bonica
K. Kompella
Juniper Networks
November 01, 2020

Generic Transport Functions
draft-zzhang-tswwg-generic-transport-functions-00

Abstract

Some functionalities (e.g. fragmentation/reassembly and Encapsulating Security Payload) provided by IPv6 can be viewed as independent of IPv6 or even IP entirely. This document proposes to provide those functionalities at different layers (e.g., MPLS, BIER or even Ethernet) independent of IP.

Status of This Memo

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

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

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 May 5, 2021.

Copyright Notice

Copyright (c) 2020 IETF Trust and the persons identified as the document authors. All rights reserved.

This document is subject to BCP 78 and the IETF Trust's Legal Provisions Relating to IETF Documents (<https://trustee.ietf.org/license-info>) in effect on the date of publication of this document. Please review these documents carefully, as they describe your rights and restrictions with respect to this document. Code Components extracted from this document must include Simplified BSD License text as described in Section 4.e of

the Trust Legal Provisions and are provided without warranty as described in the Simplified BSD License.

Table of Contents

1. Introduction	2
2. Specifications	4
2.1. Generic Fragmentation Header	4
2.2. MPLS Signaling	5
2.2.1. BGP Signaling	5
2.2.2. IGP Signaling	6
2.3. Generic ESP/Authentication Header	6
3. Security Considerations	6
4. IANA Considerations	6
5. Acknowledgements	7
6. References	7
6.1. Normative References	7
6.2. Informative References	8
Authors' Addresses	8

1. Introduction

Consider an operator providing Ethernet services such as pseudowires, VPLS or EVPN. The Ethernet frames that a Provider Edge (PE) device receives from a Customer Edge (CE) device may have a larger size than the PE-PE path MTU (pMTU) in the provider network. This could be because

1. the provider network is built upon virtual connections (e.g. pseudowires) provided by another infrastructure provider, or
2. the customer network uses jumbo frames while the provider network does not, or
3. the provider-side overhead for transporting customers packets across the network pushes past the pMTU.

In any case, the provider simply cannot require its customers to change their MTU.

To get those large frames across the provider network, currently the only workaround is to encapsulate the frames in IP (with or without GRE) and then fragment the IP packets. Even if MPLS is used for service delimiting, IP is used for transportation (MPLS over IP/GRE). This may not be desirable in certain deployment scenarios, where MPLS is the preferred transport or IP encapsulation overhead is deemed excessive.

IPv6 fragmentation and reassembly are based on the IPv6 Fragmentation header below [RFC8200]:

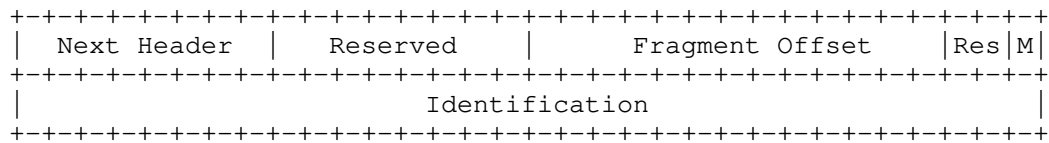


Figure 1: IPv6 Fragmentation Header

This document proposes reusing this header in non-IP contexts, since the fragmentation/reassembly function is actually independent of IPv6 except the following aspects:

- o The fragment header is identified as such by the "previous" header.
- o The "Next Header" value is from the "Internet Protocol Numbers" registry.
- o The "Identification" value is unique in the (source, destination) context provided by the IPv6 header

The "Identification" field, in conjunction with the IPv6 source and destination identifies fragments of the original packet, for the purpose of reassembly.

Therefore, the fragmentation/reassembly function can be applied at other layers as long as a) the fragment header is identified as such; and b) the context for packet identification is provided. Examples of such layers include MPLS, BIER, and Ethernet (if IEEE determines it is so desired).

For the layers where the IETF is concerned, the "Next Header" value will still be from the "Internet Protocol Numbers" registry when the function is applied at non-IP layers.

For the same consideration, the IP Encapsulating Security Payload (ESP) [RFC4303] could also be applied at other layers if ESP is desired there. For example, if for whatever reason the Ethernet service provider wants to provide ESP between its PEs, it could do so without requiring IP encapsulation if ESP is applied at non-IP layers.

The possibility of applying some other IP functions (e.g. Authentication Header [RFC4302]) is for further study.

2. Specifications

2.1. Generic Fragmentation Header

For generic fragmentation/reassembly functionality independent of IP, the following Generic Fragmentation Header (GFH) is defined:

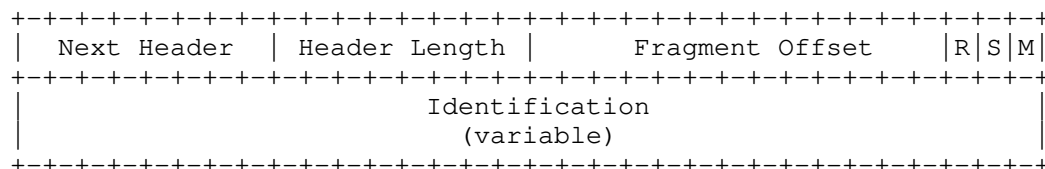


Figure 2: Generic Fragmentation Header

The "Next Header", "Fragment Offset" and "M" flag bit fields are as in the IPv6 Fragmentation Header.

Header Length: the number of octets of the entire header.

R: The "R" flag bit is reserved. It MUST be 0 on transmitting and ignored on receiving.

Identification: at least 4-octet long.

S: If the "S" flag bit is clear, the context for the Identification field is provided by the outer header, and only the source-identifying information in the outer header is used. If the "S" flag bit is set, the variable Identification field encodes both source-identifying information (e.g. the IP address of the node adding the GFH) and an identification number unique within that source.

The outer header MUST identify that a Generic Fragmentation Header follows and MAY carry source-identifying information.

If the outer header is BIER, a TBD value for the "proto" field in the BIER header identifies that a GFH follows. If the "S" flag bit is clear, the "BFIR-id" field in the BIER header provides the context for the "Identification" field.

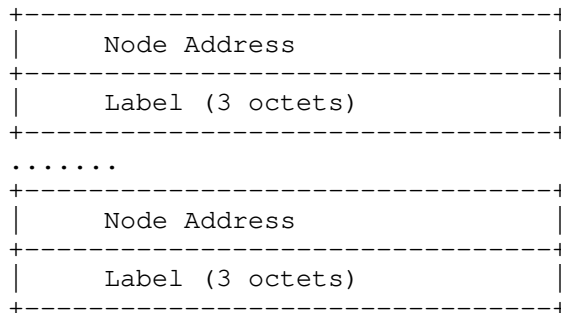
If the outer header is MPLS, the "S" flag bit MAY be clear if the the label preceeding the GFH identifies the sending BFR in addition to indicating that a GFH follows (see Section 2.2).

2.2. MPLS Signaling

When GFH is used with MPLS, the preceeding label needs to indicate that a GFH follows, and optionally identify the node that does the fragmentation. The label can be signaled via BGP or IGP as sepcified below.

2.2.1. BGP Signaling

This document defines a new transitive BGP "GFH Labels" attribute, very similar to the "PE Distinguisher Labels" attribute defined in [RFC6514] (and the text below is adapted from Section 8 of [RFC6514]):



The Label field contains an MPLS label encoded as 3 octets, where the high-order 20 bits contain the label value. The Node Address MAY be 0, meaning that the following label only indicates a GFH follows when the label is used in the label stack of a data packet.

The Node Address MAY also be a unicast address, indicating that the following label when used in the label stack of a data packet will both indicate that a GFH follows and identify the sending node.

If a node supports GFH with MPLS, it attaches the attribute in the BGP routes for its local addresses. A border router SHOULD remove the attribute if no node beyond the border will use GFH with MPLS to send traffic to the corresponding addresses.

A router that supports the attribute considers this attribute to be malformed if the Node Address field does not contain a unicast address or 0. The attribute is also considered to be malformed if: (a) the Node Address field is expected to be an IPv4 address, and the length of the attribute is not a multiple of 7 or (b) the Node Address field is expected to be an IPv6 address, and the length of the attribute is not a multiple of 19. The Address Family Indicator (AFI) of the BGP route that the attribute is attached to provides the

information on whether the Node Address field contains an IPv4 or IPv6 address. Each of the Node Addresses in the attribute MUST be of the same address family as the route that is carrying the attribute.

2.2.2. IGP Signaling

This document defines an OSPFv2 "GFH Labels" sub-TLV of OSPFv2 Extended Prefix TLV [RFC7684], with the value part being the same as BGP "GFH Labels" attribute above. If an OSPFv2 router supports GFH with MPLS, it includes the GFH Labels sub-TLV in the Extended Prefix TLV that is attached to its local addresses advertised in its OSPFv2 Extended Prefix Opaque LSA.

Similarly, This document defines an OSPFv3 "GFH Labels" sub-TLV of OSPFv3 Intra/Inter-Area-Prefix TLVs [RFC8362], with the value part being the same as BGP "GFH Labels" attribute above. If an OSPFv3 router supports GFH with MPLS, it includes the GFH Labels sub-TLV in the Intra-Area-Prefix TLV for its local addresses.

This document also defines an ISIS "GFH Labels" sub-TLV of ISIS prefix-reachability TLV [RFC5120] [RFC5305] [RFC5308], with the value part being the same as BGP "GFH Labels" attribute above. If an ISIS router supports GFH with MPLS, it includes the sub-TLV to the prefix-reachability TLV for its local addresses.

For both OSPF and ISIS, when advertising a prefix from one area/level to another, if there is a "GFH Labels TLV" attached in the source area/level, the TLV SHOULD be attached in the target area/level and the prefix SHOULD NOT be summarized.

2.3. Generic ESP/Authentication Header

To be specified in future revisions.

3. Security Considerations

To be provided.

4. IANA Considerations

This document makes the following IANA requests:

- o A new BGP Attribute type for "GFH Labels" from the BGP Path Attributes registry
- o A new OSPFv2 sub-TLV type for "GFH Labels" from the OSPFv2 Extended Prefix TLV Sub-TLVs registry

- o A new OSPFv3 sub-TLV type for "GFH Labels" from the OSPFv3 Extended-LSA sub-TLV registry
- o A new BIER Next Protocol Identifier value for GFH from BIER Next Protocol Identifiers registry

5. Acknowledgements

6. References

6.1. Normative References

- [RFC4303] Kent, S., "IP Encapsulating Security Payload (ESP)", RFC 4303, DOI 10.17487/RFC4303, December 2005, <<https://www.rfc-editor.org/info/rfc4303>>.
- [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>>.
- [RFC5305] Li, T. and H. Smit, "IS-IS Extensions for Traffic Engineering", RFC 5305, DOI 10.17487/RFC5305, October 2008, <<https://www.rfc-editor.org/info/rfc5305>>.
- [RFC5308] Hopps, C., "Routing IPv6 with IS-IS", RFC 5308, DOI 10.17487/RFC5308, October 2008, <<https://www.rfc-editor.org/info/rfc5308>>.
- [RFC7684] Psenak, P., Gredler, H., Shakir, R., Henderickx, W., Tantsura, J., and A. Lindem, "OSPFv2 Prefix/Link Attribute Advertisement", RFC 7684, DOI 10.17487/RFC7684, November 2015, <<https://www.rfc-editor.org/info/rfc7684>>.
- [RFC8200] Deering, S. and R. Hinden, "Internet Protocol, Version 6 (IPv6) Specification", STD 86, RFC 8200, DOI 10.17487/RFC8200, July 2017, <<https://www.rfc-editor.org/info/rfc8200>>.
- [RFC8362] Lindem, A., Roy, A., Goethals, D., Reddy Vallem, V., and F. Baker, "OSPFv3 Link State Advertisement (LSA) Extensibility", RFC 8362, DOI 10.17487/RFC8362, April 2018, <<https://www.rfc-editor.org/info/rfc8362>>.

6.2. Informative References

- [RFC4302] Kent, S., "IP Authentication Header", RFC 4302, DOI 10.17487/RFC4302, December 2005, <<https://www.rfc-editor.org/info/rfc4302>>.
- [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>>.

Authors' Addresses

Zhaohui Zhang
Juniper Networks
1133 Innovation Way
Sunnyvale 94089
USA

Phone: +1 408 745 2000
Email: zzhang@juniper.net

Ron Bonica
Juniper Networks
1133 Innovation Way
Sunnyvale 94089
USA

Phone: +1 408 745 2000
Email: rbonica@juniper.net

Kireeti Kompella
Juniper Networks
1133 Innovation Way
Sunnyvale 94089
USA

Phone: +1 408 745 2000
Email: kireeti@juniper.net