Internet Engineering Task Force (IETF)
Request for Comments: 9655
Category: Standards Track
ISSN: 2070-1721

D. Rathi, Ed.
Nokia
S. Hegde, Ed.
Juniper Networks Inc.
K. Arora
Individual Contributor
Z. Ali
N. Nainar
Cisco Systems, Inc.

November 2024

Egress Validation in Label Switched Path Ping and Traceroute Mechanisms

Abstract

The MPLS ping and traceroute mechanisms described in RFC 8029 and the related extensions for Segment Routing (SR) defined in RFC 8287 are highly valuable for validating control plane and data plane synchronization. In certain environments, only some intermediate or transit nodes may have been upgraded to support these validation procedures. A straightforward MPLS ping and traceroute mechanism allows traversal of any path without validation of the control plane state. RFC 8029 supports this mechanism with the Nil Forwarding Equivalence Class (FEC). The procedures outlined in RFC 8029 are primarily applicable when the Nil FEC is used as an intermediate FEC in the FEC stack. However, challenges arise when all labels in the label stack are represented using the Nil FEC.

This document introduces a new Type-Length-Value (TLV) as an extension to the existing Nil FEC. It describes MPLS ping and traceroute procedures using the Nil FEC with this extension to address and overcome these challenges.

Status of This Memo

This is an Internet Standards Track document.

This document is a product of the Internet Engineering Task Force (IETF). It represents the consensus of the IETF community. It has received public review and has been approved for publication by the Internet Engineering Steering Group (IESG). Further information on Internet Standards is available in Section 2 of RFC 7841.

Information about the current status of this document, any errata, and how to provide feedback on it may be obtained at https://www.rfc-editor.org/info/rfc9655.

Copyright Notice

Table of Contents

1. Introduction

Segment routing supports the creation of explicit paths by using one or more Link-State IGP Segments or BGP Segments defined in [RFC8402]. In certain use cases, the TE paths are built using mechanisms described in [RFC9256] by stacking the labels that represent the nodes and links in the explicit path. Controllers are often deployed to construct paths across multi-domain networks. In such deployments, the headend routers may have the link-state database of their domain and may not be aware of the FEC associated with labels that are used by the controller to build paths across multiple domains. A very useful Operations, Administration, and Maintenance (OAM) requirement is to be able to ping and trace these paths.

[RFC8029] describes a simple and efficient mechanism to detect data plane failures in MPLS Label Switched Paths (LSPs). It defines a probe message called an "MPLS echo request" and a response message called an "MPLS echo reply" for returning the result of the probe. SR-related extensions for these are specified in [RFC8287]. [RFC8029] provides mechanisms primarily to validate the data plane and secondarily to verify the consistency of the data plane with the control plane. It also provides the ability to traverse Equal-Cost Multipaths (ECMPs) and validate each of the ECMP paths. The Target FEC Stack TLV [RFC8029] contains sub-TLVs that carry information about the label. This information gets validated on each node for traceroute and on the egress for ping. The use of the Target FEC Stack TLV requires all nodes in the network to have implemented the validation procedures, but all intermediate nodes may not have been upgraded to support validation procedures. In such cases, it is useful to have the ability to traverse the paths in ping/traceroute mode without having to obtain the FEC for each label.

A simple MPLS echo request/reply mechanism allows for traversing the SR Policy path without validating the control plane state. [RFC8029] supports this mechanism with FECs like the Nil FEC and the Generic FECs (i.e., Generic IPv4 prefix and Generic IPv6 prefix). However, there are challenges in reusing the Nil FEC and Generic FECs for validation of SR Policies [RFC9256]. The Generic IPv4 prefix and Generic IPv6 prefix FECs are used when the protocol that is advertising the label is unknown. The information that is carried in the Generic FECs is the IPv4 or IPv6 prefix and prefix length. Thus, the Generic FEC types perform an additional control plane validation.

However, the Generic FECs and relevant validation procedures are not thoroughly detailed in [RFC8029]. The use case mostly specifies inter-AS (Autonomous System) VPNs as the motivation. Certain aspects of SR, such as anycast Segment Identifiers (SIDs), require clear guidelines on how the validation procedure should work. Also, the Generic FECs may not be widely supported, and if transit routers are not upgraded to support validation of Generic FECs, traceroute may fail. On the other hand, the Nil FEC consists of the label, and there is no other associated FEC information. The Nil FEC is used to traverse the path without validation for cases where the FEC is not defined or routers are not upgraded to support the FECs. Thus, it can be used to check any combination of segments on any data path. The procedures described in [RFC8029] are mostly applicable when the Nil FEC is used as an intermediate FEC in the FEC stack. Challenges arise when all labels in the label stack are represented using the Nil FEC.

Section 2 discusses the problems associated with using the Nil FEC in an MPLS ping/traceroute procedure, and Sections 3 and 4 discuss simple extensions needed to solve the problem.

The problems and the solutions described in this document apply to the MPLS data plane. Segment Routing over IPv6 (SRv6) is out of scope for this document.

1.1. 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. Problem with Nil FEC

The purpose of the Nil FEC, as described in [RFC8029], is to ensure that transit tunnel information is hidden and, in some cases, to avoid false negatives when the FEC information is unknown.

This document uses a Nil FEC to represent the complete label stack in an MPLS echo request message in ping and traceroute mode. A single Nil FEC is used in the MPLS echo request message irrespective of the number of segments in the label stack. Section 4.4.1 of [RFC8029] notes:

> If the outermost FEC of the Target FEC stack is the Nil FEC, then the node MUST skip the Target FEC validation completely.

When a router in the label stack path receives an MPLS echo request message, there is no definite way to decide whether it is the intended egress router since the Nil FEC does not carry any information and no validation is performed by the router. Thus, there is a high possibility that the packet may be misforwarded to an incorrect destination but the MPLS echo reply might still return success.

To mitigate this issue, it is necessary to include additional information, along with the Nil FEC, in the MPLS echo request message in both ping and traceroute modes and to perform minimal validation on the egress/destination router. This will enable the router to send appropriate success and failure information to the headend router of the SR Policy. This supplementary information should assist in reporting transit router details to the headend router, which can be utilized by an offline application to validate the traceroute path.

Consequently, the inclusion of egress information in the MPLS echo request messages in ping and traceroute modes will facilitate the validation of the Nil FEC on the egress router, ensuring the correct destination. Egress information can be employed to verify any combination of segments on any path without requiring upgrades to transit nodes. The Egress TLV can be silently dropped if not recognized; alternately, it may be stepped over, or an error message may be sent (per [RFC8029] and the clarifications in [RFC9041] regarding code points in the range 32768-65535).

If a transit node does not recognize the Egress TLV and chooses to silently drop or step over the Egress TLV, the headend will continue to send the Egress TLV in the next echo request message, and if egress recognizes the Egress TLV, egress validation will be executed at the egress. If a transit node does not recognize the Egress TLV and chooses to send an error message, the headend will log the message for informational purposes and continue to send echo requests with the Egress TLV, with the TTL incremented. If the egress node does not recognize the Egress TLV and chooses to silently drop or step over the Egress TLV, egress validation will not be done, and the ping/traceroute procedure will proceed as if the Egress TLV were not received.

3. Egress TLV

The Egress TLV MAY be included in an MPLS echo request message. It is an optional TLV and, if present, MUST appear before the Target FEC Stack TLV in the MPLS echo request packet. This TLV can only be used in LSP ping/traceroute requests that are generated by the headend node of an LSP or SR Policy for which verification is performed. In cases where multiple Nil FECs are present in the Target FEC Stack TLV, the Egress TLV must be added corresponding to the ultimate egress of the label stack. Explicit paths can be created using Node-SID, Adj-SID, Binding SID, etc. The Address field of the Egress TLV must be derived from the path egress/destination. The format is as specified in Figure 1.

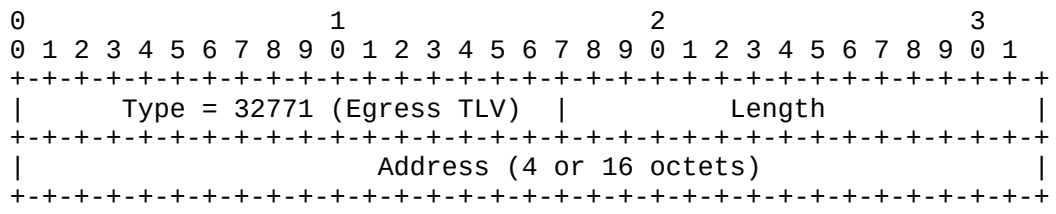

Figure 1: Egress TLV

Type: 32771 (Section 6.1)

Length: Variable (4 octets for IPv4 addresses and 16 octets for IPv6 addresses). Length excludes the length of the Type and Length fields.

Address: This field carries a valid 4-octet IPv4 address or a valid 16-octet IPv6 address. The address can be obtained from the egress of the path and corresponds to the last label in the label stack or the SR Policy Endpoint field [SR-POLICY-BGP].

4. Procedure

This section describes aspects of LSP ping and traceroute operations that require further considerations beyond those detailed in [RFC8029].

4.1. Sending Egress TLV in MPLS Echo Request

As previously mentioned, when the sender node constructs an echo request with a Target FEC Stack TLV, the Egress TLV, if present, MUST appear before the Target FEC Stack TLV in the MPLS echo request packet.

4.1.1. Ping Mode

When the sender node constructs an echo request with a Target FEC Stack TLV that contains a single Nil FEC corresponding to the last segment of the SR Policy path, the sender node MUST add an Egress TLV with the address obtained from the SR Policy Endpoint field [SR-POLICY-BGP]. The Label value in the Nil FEC MAY be set to zero when a single Nil FEC is added for multiple labels in the label stack. In case the endpoint is not specified or is equal to zero (Section 8.8.1 of [RFC9256]), the sender MUST use the address corresponding to the last segment of the SR Policy in the Address field of the Egress TLV. Some specific cases on how to derive the Address field in the Egress TLV are listed below:

* If the last SID in the SR Policy is an Adj-SID, the Address field in the Egress TLV is derived from the node at the remote end of the corresponding adjacency.

* If the last SID in the SR Policy is a Binding SID, the Address field in the Egress TLV is derived from the last node of the path represented by the Binding SID.

4.1.2. Traceroute Mode

When the sender node builds an echo request with a Target FEC Stack TLV that contains a Nil FEC corresponding to the last segment of the segment list of the SR Policy, the sender node MUST add an Egress TLV with the address obtained from the SR Policy Endpoint field [SR-POLICY-BGP].

Although there is no requirement to do so, an implementation MAY send multiple Nil FECs if that makes it easier for the implementation. If the SR Policy headend sends multiple Nil FECs, the last one MUST correspond to the Egress TLV. The Label value in the Nil FEC MAY be set to zero for the last Nil FEC. If the endpoint is not specified or is equal to zero (Section 8.8.1 of [RFC9256]), the sender MUST use the address corresponding to the last segment endpoint of the SR Policy path (i.e., the ultimate egress is used as the address in the Egress TLV).

4.1.3. Detailed Example

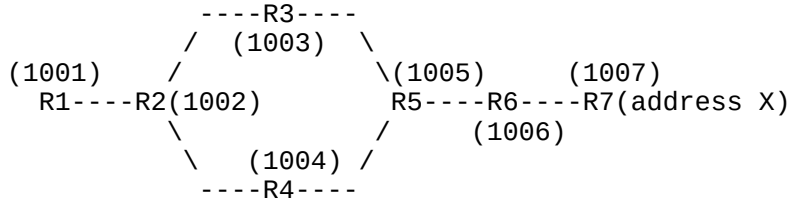

Figure 2: Egress TLV Processing in Sample Topology

Consider the SR Policy configured on router R1 to destination X, configured with label stack as 1002, 1004, 1007. Segment 1007 belongs to R7, which has the address X locally configured on it.

Let us look at an example of a ping echo request message. The echo request message contains a Target FEC Stack TLV with the Nil FEC sub-TLV. An Egress TLV is added before the Target FEC Stack TLV. The Address field contains X (corresponding to a locally configured address on R7). X could be an IPv4 or IPv6 address, and the Length

field in the Egress TLV will be either 4 or 16 octets, based on the address type of address X.

Let us look at an example of an echo request message in a traceroute packet. The echo request message contains a Target FEC Stack TLV with the Nil FEC sub-TLV corresponding to the complete label stack (1002, 1004, 1007). An Egress TLV is added before the Target FEC Stack TLV. The Address field contains X (corresponding to a locally configured address on destination R7). X could be an IPv4 or IPv6 address, and the Length field in the Egress TLV will be either 4 or 16 octets, based on the address type of address X. If the destination/endpoint is set to zero (as in the case of the color-only SR Policy), the sender should use the endpoint of segment 1007 (the last segment in the segment list) as the address for the Egress TLV.

4.2. Receiving Egress TLV in MPLS Echo Request

Any node that receives an MPLS echo request message and processes it is referred to as the "receiver". In the case of the ping procedure, the actual destination/egress is the receiver. In the case of traceroute, every node is a receiver. This document does not propose any change in the processing of the Nil FEC (as defined in [RFC8029]) in the node that receives an MPLS echo request with a Target FEC Stack TLV. The presence of the Egress TLV does not affect the validation of the Target FEC Stack sub-TLV at FEC-stack-depth if it is different than Nil FEC.

Additional processing MUST be done for the Egress TLV on the receiver node as follows. Note that <RSC> refers to the Return Subcode.

1. If the Label-stack-depth is greater than 0 and the Target FEC Stack sub-TLV at FEC-stack-depth is Nil FEC, set Best-return-code to 8 ("Label switched at stack-depth <RSC>") and Best-rtn-subcode to Label-stack-depth to report transit switching in the MPLS echo reply message.

2. If the Label-stack-depth is 0 and the Target FEC Stack sub-TLV at FEC-stack-depth is Nil FEC, then do a lookup for an exact match of the Address field of the Egress TLV to any of the locally configured interfaces or loopback addresses.

    a. If the Egress TLV address lookup succeeds, set Best-return-code to 36 ("Replying router is an egress for the address in the Egress TLV for the FEC at stack depth <RSC>") (Section 6.2) in the MPLS echo reply message.

    b. If the Egress TLV address lookup fails, set the Best-return-code to 10 ("Mapping for this FEC is not the given label at stack-depth <RSC>").

3. In some cases, multiple Nil FECs (one corresponding to each label in the label stack), along with the Egress TLV, are sent from the SR Policy headend. When the packet reaches the egress, the number of labels in the received packet (size of stack-R) becomes zero, or a label with the Bottom-of-Stack bit set to 1 is processed. All Nil FEC sub-TLVs MUST be removed, and the Egress TLV MUST be validated.

5. Backward Compatibility

The extensions defined in this document are backward compatible with the procedures described in [RFC8029]. A router that does not support the Egress TLV will ignore it and use the Nil FEC procedures described in [RFC8029].

When the egress node in the path does not support the extensions

defined in this document, egress validation will not be done, and Best-return-code will be set to 3 ("Replying router is an egress for the FEC at stack-depth <RSC>") and Best-rtn-subcode to stack-depth in the MPLS echo reply message.

When the transit node in the path does not support the extensions defined in this document, Best-return-code will be set to 8 ("Label switched at stack-depth <RSC>") and Best-rtn-subcode to Label-stack-depth to report transit switching in the MPLS echo reply message.

6. IANA Considerations

6.1. New TLV

IANA has added the following entry to the "TLVs" registry within the "Multiprotocol Label Switching (MPLS) Label Switched Paths (LSPs) Ping Parameters" registry group [IANA-MPLS-LSP]:

Type	TLV Name	Reference
32771	Egress TLV	RFC 9655

Table 1: TLVs Registry

6.2. New Return Code

IANA has added the following entry to the "Return Codes" registry within the "Multiprotocol Label Switching (MPLS) Label Switched Paths (LSPs) Ping Parameters" registry group [IANA-MPLS-LSP]:

Value	Meaning	Reference
36	Replying router is an egress for the address in the Egress TLV for the FEC at stack depth <RSC>	RFC 9655

Table 2: Return Codes Registry

7. Security Considerations

This document defines an additional TLV for MPLS LSP ping and conforms to the mechanisms defined in [RFC8029]. All the security considerations defined in [RFC8287] apply to this document. This document does not introduce any additional security challenges to be considered.

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>.

[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>.

[RFC9041] Andersson, L., Chen, M., Pignataro, C., and T. Saad, "Updating the MPLS Label Switched Paths (LSPs) Ping Parameters IANA Registry", RFC 9041, DOI 10.17487/RFC9041, July 2021, <https://www.rfc-editor.org/info/rfc9041>.

[RFC9256] Filsfils, C., Talaulikar, K., Ed., Voyer, D., Bogdanov, A., and P. Mattes, "Segment Routing Policy Architecture", RFC 9256, DOI 10.17487/RFC9256, July 2022, <https://www.rfc-editor.org/info/rfc9256>.

8.2. Informative References

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

[SR-POLICY-BGP]
Previdi, S., Filsfils, C., Talaulikar, K., Ed., Mattes, P., and D. Jain, "Advertising Segment Routing Policies in BGP", Work in Progress, Internet-Draft, draft-ietf-idr-sr-policy-safi-10, 7 November 2024, <https://datatracker.ietf.org/doc/html/draft-ietf-idr-sr-policy-safi-10>.

Acknowledgements

The authors would like to thank Stewart Bryant, Greg Mirsky, Alexander Vainshtein, Sanga Mitra Rajgopal, and Adrian Farrel for their careful review and comments.

Authors' Addresses

Deepti N. Rathi (editor)
Nokia
Manyata Embassy Business Park
Bangalore 560045
Karnataka
India
Email: deepti.nirmalkumarji_rathi@nokia.com

Shraddha Hegde (editor)
Juniper Networks Inc.
Exora Business Park
Bangalore 560103
Karnataka
India
Email: shraddha@juniper.net

Kapil Arora
Individual Contributor
Email: kapil.it@gmail.com

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

Nagendra Kumar Nainar
Cisco Systems, Inc.
Email: naikumar@cisco.com