Workgroup: Routing area
Internet-Draft:
draft-ietf-mpls-spring-inter-domain-oam-04
Published: 14 December 2022
Intended Status: Standards Track
Expires: 17 June 2023
Authors: S. Hegde, Juniper Networks Inc.
K. Arora, Juniper Networks Inc.
M. Srivastava, Juniper Networks Inc.
S. Ninan, Ciena
N. Kumar, Cisco Systems, Inc.

PMS/Head-end based MPLS Ping and Traceroute in Inter-domain SR Networks

Abstract

Segment Routing (SR) architecture leverages source routing and tunneling paradigms and can be directly applied to the use of a Multiprotocol Label Switching (MPLS) data plane. A network may consist of multiple IGP domains or multiple ASes under the control of same organization. It is useful to have the Label switched Path (LSP) Ping and traceroute procedures when an SR end-to-end path spans across multiple ASes or domains. This document describes mechanisms to facilitae LSP ping and traceroute in inter-AS/inter-domain SR-MPLS networks in an efficient manner with simple Operations, Administration, and Maintenance (OAM) protocol extension which uses dataplane forwarding alone for sending echo reply.

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 will expire on 17 June 2023.

Copyright Notice

Table of Contents

1. Introduction

```
                    +----------------+
                    | Controller/PMS |
                    +----------------+


 |---AS1-----|                |------AS2------|            |----AS3---|

                ASBR2----ASBR3                ASBR5------ASBR7
                /             \               /            \
               /               \             /              \
 PE1----P1---P2                 P3---P4---PE4                P5---P6--PE5
               \               /            \               /
                \             /              \             /
                 ASBR1----ASBR4              ASBR6------ASBR8
```

Figure 1: Inter-AS Segment Routing topology

Many network deployments have built their networks consisting of multiple Autonomous Systems either for ease of operations or as a result of network mergers and acquisitions. Segment Routing can be deployed in such scenarios to provide end to end paths, traversing multiple Autonomous systems(AS). These paths consist of Segment Identifiers(SID) of different type as per [RFC8402].

[RFC8660] specifies the forwarding plane behaviour to allow Segment Routing to operate on top of MPLS data plane. [RFC9087] describes BGP peering SIDs, which will help in steering packet from one Autonomous system to another. Using above SR capabilities, paths which span across multiple Autonomous systems can be created.

For example Figure 1 describes an inter-AS network scenario consisting of ASes AS1 and AS2. Both AS1 and AS2 are Segment Routing enabled and the EPE links have EPE labels configured and advertised via [RFC9086]. Controller or head-end can build end-to-end Traffic-Engineered path consisting of Node-SIDs, Adjacency-SIDs and EPE-SIDs. It is advantageous for operations to be able to perform LSP ping and traceroute procedures on these inter-AS SR-MPLS paths. LSP ping/traceroute procedures use IP connectivity for echo reply to reach the head-end. In inter-AS networks, IP connectivity may not be there from each router in the path.For example in Figure 1 P3 and P4 may not have IP connectivity for PE1.

[RFC8403] describes mechanisms to carry out the MPLS ping/traceroute from a Path Monitoring System (PMS). It is possible to build GRE tunnels or static routes to each router in the network to get IP connectivity for the reverse path. This mechanism is operationally very heavy and requires PMS to be capable of building huge number of GRE tunnels, which may not be feasible.

It is not possible to carry out LSP ping and Traceroute functionality on these paths to verify basic connectivity and fault isolation using existing LSP ping and Traceroute mechanism([RFC8287] and [RFC8029]). This is because, there exists no IP connectivity to source address of ping packet, which is in a different AS, from the destination of Ping/Traceroute.

[RFC7743] describes a Echo-relay based solution based on advertising a new Relay Node Address Stack TLV containing stack of Echo-relay IP addresses. These mechansims can be applied to segment routing networks as well. [RFC7743] mechanism requires the return ping packet to be processed in slow path or as a bump-in-the-wire on every relay node. The motivation of the current document is to provide an alternate mechanism for ping/traceroute in inter-domain segment routing networks.

This document describes a new mechanism which is efficient and simple and can be easily deployed in SR-MPLS networks. This mechanism uses MPLS path and no changes required in the forwarding path. Any MPLS capable node will be able to forward the echo-reply packet in fast path. The current draft describes a mechanism that uses Reply path TLV [RFC7110] to convey the reverse path. Three new sub-TLVs for Reply path TLV are defined, that faciliate encoding segment routing label stack. The TLV can either be derived by a smart application or controller which has a full topology view. This document also proposes mechanisms to derive the return path dynamically during traceroute procedures.

The current document is focused on the inter-domain use case. However, the protocol extensions described in this document may be applied to indicate the return path for other use cases as well.

1.1. Definition of Domain

The term domain used in this document implies an IGP domain where every node is visible to every other node for the purposes of shortest path computation. The domain implies an IGP area or level. An Autonomous System (AS) consists of one or more IGP domains.The procedures described in this document are applicable to paths built across multiple domains which includes inter-area as well as inter-AS paths. This document is applicable to SR-MPLS networks where all nodes in each of the domains are SR capable. It is also applicable to SR-MPLS networks where SR acts an an overlay having SR incapable underlay nodes. In such networks, the traceroute procedure is executed only on the overlay SR nodes.

2. Inter domain networks with multiple IGPs

```
 |-Domain 1|-------Domain 2-----|--Domain 3-|


 PE1------ABR1--------P--------ABR2------PE4
  \        / \                 /\        /
   --------   -----------------   -------
    BGP-LU         BGP-LU          BGP-LU
```

Figure 2: Inter-domain networks with multiple IGPs

When the network consists of large number of nodes, the nodes are seggregated into multiple IGP domains. The connectivity to the remote PEs can be achieved using BGP-Labeled Unicast (BGP-LU) [RFC8277] or by stacking the labels for each domain as described in [RFC8604]. It is useful to support MPLS ping and traceroute mechanisms for these networks. The procedures described in this document for constructing Reply path TLV and its use in echo reply is equally applicable to networks consisting of multiple IGP domains that use BGP-LU or label stacking.

3. Reply Path TLV

Segment Routing networks statically assign the labels to nodes and PMS/Head-end may know the entire database. The reverse path can be built from PMS/Head-end by stacking segments for the reverse path. Reply path TLV as defined in [RFC7110] is used to carry the return path. While using the procedures described in this document, the reply mode is set to 5 (Reply via Specified Path), and Reply Path TLV is included in the echo request message as described in [RFC7110]. The Reply Path TLV is constructed as per Section 4.2 of RFC 7110.This document define three new sub-TLVs to encode the Segment Routing path.

The type of segment that the head-end chooses to send in the Reply Path TLV is governed by local policy. Implementations may provide CLI input parameters in Labels, IPv4 addresses or IPv6 addresses or a combination of these which gets encoded in the Reply path TLV. Implementations may also provide mechansims to acquire the database of remote domains and compute the return path based on the acquired database. For traceroute purposes, the return path will have to consider the reply being sent from every node along the path. The return path changes when the traceroute progresses and crosses each domain. One of the ways this can be implemented on headend is to acquire the entire database (of all domains) and build return path for every node along the SR-MPLS path based on the knowledge of the database. Another mechansim is to use dynamically computed return path as described in Section 8

Some networks may consist of pure IPV4 domains and pure IPv6 domains. Handling end-to-end MPLS OAM for such networks is out of scope for this document. It is recommended to use dual stack in such cases and use end-to-end IPv6 addresses for MPLS ping and trace route procedures.

4. Segment sub-TLV

[RFC9256] defines various types of segments. The types of segments applicable to this document have been defined in this section for the use of MPLS OAM. The motivation has been to keep the definitions same as in [RFC9256] with minimal modifications if it is absolutely needed. One or more segment sub-TLV can be included in the Reply Path TLV. The segment sub-TLVs included in a Reply Path TLV MAY be of different types.

Below types of segment sub-TLVs are applicable for the Reverse Path Segment List TLV.

Type A: SID only, in the form of MPLS Label

Type C: IPv4 Node Address with optional SID

Type D: IPv6 Node Address with optional SID for SR MPLS

4.1. Type A: SID only, in the form of MPLS Label

The Type-A Segment Sub-TLV encodes a single SID in the form of an MPLS label. The format is as follows:

```
 0                   1                   2                   3
 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|     Type                      |   Length                      |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|   Flags       |   RESERVED                                    |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|          Label                        | TC  |S|       TTL     |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
```

Figure 3: Type 1 Segment sub-TLV

where:

Type: TBD1(to be assigned by IANA from the registry "Sub-TLVs for TLV Types 1, 16, and 21").

Length is 8.

Flags: 1 octet of flags as defined in Section 4.4.

RESERVED: 3 octets of reserved bits. SHOULD be unset on transmission and MUST be ignored on receipt.

Label: 20 bits of label value.

TC: 3 bits of traffic class

S: 1 bit Reserved

TTL: 1 octet of TTL.

The following applies to the Type-1 Segment sub-TLV:

The S bit SHOULD be zero upon transmission, and MUST be ignored upon reception.

If the originator wants the receiver to choose the TC value, it sets the TC field to zero.

If the originator wants the receiver to choose the TTL value, it sets the TTL field to 255.

If the originator wants to recommend a value for these fields, it puts those values in the TC and/or TTL fields.

The receiver MAY override the originator's values for these fields. This would be determined by local policy at the receiver. One possible policy would be to override the fields only if the fields have the default values specified above.

4.2. Type C: IPv4 Node Address with optional SID for SR-MPLS

The Type-C Segment Sub-TLV encodes an IPv4 node address, SR Algorithm and an optional SID in the form of an MPLS label. The format is as follows:

```
 0                   1                   2                   3
 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|     Type                      |   Length                      |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|   Flags       |  RESERVED (MBZ)             | SR Algorithm    |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|                 IPv4 Node Address (4 octets)                  |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|                SID (optional, 4 octets)                       |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
```

Figure 4: Type 3 Segment sub-TLV

where:

Type: TBD3(to be assigned by IANA from the registry "Sub-TLVs for TLV Types 1, 16, and 21").

Length is 8 or 12.

Flags: 1 octet of flags as defined in Section 4.4.

SR Algorithm: 1 octet specifying SR Algorithm as described in section 3.1.1 in [RFC8402], when A-Flag as defined in Section 4.4is present. SR Algorithm is used by the receiver to derive the Label. When A-flag is unset, this field has no meaning and thus MUST be set to zero on transmission and ignored on receipt.

RESERVED: 2 octets of reserved bits.MUST be set to zero when sending; MUST be ignored on receipt.

IPv4 Node Address: a 4 octet IPv4 address representing a node.

SID: optional :4 octet field containing label, TC, S and TTL as defined in Section 4.1

The following applies to the Type-3 Segment sub-TLV:

The IPv4 Node Address MUST be present.

The SID is optional and specifies a 4 octet MPLS SID containing label, TC, S and TTL as defined in Section 4.1.

If length is 8, then only the IPv4 Node Address is present.

If length is 12, then the IPv4 Node Address and the MPLS SID are present.When the MPLS SID field is present, it MUST be used for constructing the Reply Path TLV.

4.3. Type D: IPv6 Node Address with optional SID for SR MPLS

The Type-D Segment Sub-TLV encodes an IPv6 node address, SR Algorithm and an optional SID in the form of an MPLS label. The format is as follows:

```
 0                   1                   2                   3
 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|     Type                      |   Length                      |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|   Flags       |       RESERVED(MBZ)           | SR Algorithm  |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
//                IPv6 Node Address (16 octets)                //
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|                SID (optional, 4 octets)                       |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
```

Figure 5: Type 4 Segment sub-TLV

where:

Type: TBD4(to be assigned by IANA from the registry "Sub-TLVs for TLV Types 1, 16, and 21").

Length is 20 or 24.

Flags: 1 octet of flags as defined in Section 4.4.

SR Algorithm: 1 octet specifying SR Algorithm as described in section 3.1.1 in [RFC8402], when A-Flag as defined in Section 4.4 is present. SR Algorithm is used by the receiver to derive the label.W hen A-flag is unset, this field has no meaning and thus MUST be set to zero on transmission and ignored on receipt.

RESERVED: 2 octets of reserved bits.MUST be set to zero when sending; MUST be ignored on receipt..

IPv6 Node Address: a 16 octet IPv6 address representing a node.

SID: optional :4 octet field containing label, TC, S and TTL as defined in Section 4.1

The following applies to the Type-D Segment sub-TLV:

The IPv6 Node Address MUST be present.

The SID is optional and specifies a 4 octet MPLS SID containing label, TC, S and TTL as defined in Section 4.1 .

If length is 20, then only the IPv6 Node Address is present.

If length is 24, then the IPv6 Node Address and the MPLS SID are present. When the MPLS SID field is present, it MUST be used for constructing the Reply Path TLV.

4.4. Segment Flags

The Segment Types described above contain following flags in the "Flags" field (codes to be assigned by IANA from the registry "Segment sub-TLV Flags")

```
 0 1 2 3 4 5 6 7
+-+-+-+-+-+-+-+-+
| |A|           |
+-+-+-+-+-+-+-+-+
```

Figure 6: Flags

where:

A-Flag: This flag indicates the presence of SR Algorithm id in the "SR Algorithm" field applicable to various Segment Types.

Unused bits in the Flag octet SHOULD be set to zero upon transmission and MUST be ignored upon receipt.

The following applies to the Segment Flags:

A-Flag is applicable to Segment Types C, D. If A-Flag appears with any other Segment Type, it MUST be ignored.

5. SRv6 Dataplane

SRv6 dataplane is not in the scope of this document and will be addressed in a separate document.

6. Detailed Procedures

6.1. Sending an echo request

In the inter-AS scenario when there is no reverse path connectivity, the procedures described in this document should be used. LSP ping initiator MUST set the Reply Mode of the echo request to "Reply via Specified Path", and a Reply Path TLV MUST be carried in the echo request message correspondingly. The Reply Path TLV must contain the Segment Routing Path in the reverse direction encoded as an ordered list of segments. The first Segment MUST correspond to the top

Segment in MPLS header that the responder MUST use while sending the echo reply.

6.2. Receiving an echo request

As described in [RFC7110], when Reply mode is set to 5 (Reply via Specified Path),The echo request MUST contain the Reply path TLV. Absence of Reply path TLV is treated as malformed echo request. when an echo request is received, if the egress LSR does not know the Reply Mode 5 defined in [RFC7110], an echo reply with the return code set to "Malformed echo request received" and the Subcode set to zero will be sent back to the ingress LSR according to the rules of [RFC8029]. When a Reply Path TLV is received, and the responder that supports processing it, it MUST use the segments in Reply Path TLV to build the echo reply.The responder MUST follow the normal FEC validation procedures as described in [RFC8029] and [RFC8287] and this document does not suggest any change to those procedures. When the echo reply has to be sent out the Reply Path TLV is used to construct the MPLS packet to send out.

6.3. Sending an echo reply

The echo reply message is sent as MPLS packet with a MPLS label stack. The echo reply message MUST be constructed as described in the [RFC8029]. An MPLS packet is constructed with echo reply in the payload. The top label MUST be constructed from the first Segment from the Reply Path TLV. The remaining labels MUST follow the order from the Reply Path TLV. The responder MAY check the reachability of the top label in its own Label Forwarding Information Base (LFIB) before sending the echo reply. In certain scenarios the head-end may choose to send Type 3/Type 4 segments consisting of IPV4 address or IPv6 address. Optionally a SID may also be assiciated with Type 3/ Type4 segment. In such cases the node sending the echo reply MUST derive the MPLS labels based on Node-SIDs associated with the IPv4 / IPv6 addresses or from the optional MPLS SIDs in the type 3/ type 4 segments and encode the echo reply with MPLS labels.

The reply path return code MUST be set as described in section 7.4 of [RFC7110]. The Reply Path TLV MUST be included in echo reply indicating the specified return path that the echo reply message is required to follow as described in section 5.3 of [RFC7110].

When the node is configured to dynamically create return path for next echo request, the procedures described in Section 8 MUST be used. The reply path return code MUST be set to TBA1 and same Reply Path TLV or a new Reply Path TLV MUST be included in the echo reply.

6.4. Receiving an echo reply

The rules and process defined in Section 4.6 of [RFC8029] and section 5.4 of [RFC7110] apply here. In addition, if the Reply path return code is "Use Reply Path TLV in echo reply for next echo request", the Reply Path TLV from the echo Reply MUST be sent in the next echo request with TTL incremented by 1.

7. Detailed Example

Example topologies given in Figure 1 and Figure 2 will be used in below sections to explain LSP Ping and Traceroute procedures. The PMS/Head-end has complete view of topology. PE1, P1, P2, ASBR1 and ASBR2 are in AS1. Similarly ASBR3, ASBR4, P3, P4 and PE4 are in AS2.

AS1 and AS2 have Segment Routing enabled. IGPs like OSPF/ISIS are used to flood SIDs in each Autonomous System. The ASBR1, ASBR2, ASBR3, ASBR4 advertise BGP EPE SIDs for the inter-AS links. Topology of AS1 and AS2 are advertised via BGP-Link State (BGP-LS) to the controller/PMS or Head-end node. The EPE-SIDs are also advertised via BGP-LS as described in [RFC9086]. The example uses EPE-SIDs for the inter-AS links but the same could be achieved using adjacency-SIDs advertised for a passive IGP link.

The description in the document uses below notations for Segment Identifiers(SIDs).

Node SIDs : N-PE1, N-P1, N-ASBR1 N-ABR1, N-ABR2etc.

Adjacency SIDs : Adj-PE1-P1, Adj-P1-P2 etc.

EPE SIDS : EPE-ASBR2-ASBR3, EPE-ASBR1-ASBR4, EPE-ASBR3-ASBR2 etc.

Let us consider a traffic engineered path built from PE1 to PE4 with Segment List stack as below. N-P1, N-ASBR1, EPE-ASBR1-ASBR4, N-PE4 for following procedures. This stack may be programmed by controller/PMS or Head-end router PE1 may have imported the whole topology information from BGP-LS and computed the inter-AS path.

7.1. Procedures for Segment Routing LSP ping

To perform LSP ping procedure on an SR-MPLS-Path from PE1 to PE4 consisting of label stacks [N-P1,N-ASBR1,EPE-ASBR1-ASBR4, N-PE4], The remote end(PE4) needs IP connectivity to head end(PE1) for the Segment Routing ping to succeed, because echo reply needs to travel back to PE1 from PE4. But in typical deployment scenario there will be no IP route from PE4 to PE1 as they belong to different ASes.

PE1 adds return Path from PE4 to PE1 in the MPLS echo request using multiple Segments in "Reply Path TLV" as defined above. An example

Reply path TLV for PE1 to PE4 for LSP ping is [N-ASBR4, EPE-ASBR4-ASBR1, N-PE1]. An implementation may also build a return Path consisting of labels to reach its own AS. Once the label stack is popped-off the echo reply message will be exposed. The further packet forwarding will be based on IP lookup. An example return Path for this case could be [N-ASBR4, EPE-ASBR4-ASBR1].

On receiving MPLS echo request PE4 first validates FEC in the echo request. PE4 then builds label stack to send the response from PE4 to PE1 by copying the labels from "Reply Path TLV". PE4 builds the echo reply packet with the MPLS label stack constructed and imposes MPLS headers on top of echo reply packet and sends out the packet towards PE1. This Segment List stack can successfully steer reply back to Head-end node(PE1).

7.2. Procedures for Segment Routing LSP Traceroute

Traceroute procedure involves visiting every node on the path and echo reply sent from every node. In this section, we describe the traceroute mechanims when the headend/PMS has complete visibility of the database. Headend/PMS computes the return path from each node in the entire SR-MPLS path that is being tracerouted. The return path computation is implementation dependant. As the headend/PMS completely controls the return path, it can use proprietary computations to build the return path.

One of the ways the return path can be built, is to use the principle of building label stacks by adding each domain border node's Node SID on the return path label stack as the traceroute progresses. For inter-AS networks, in addition to border node's Node-SID, EPE-SID in the reverse direction also need to be added to the label stack.

The Inter-domain/inter-as traceroute procedure uses the TTL expiry mechansim as specified in [RFC8029] and [RFC8287]. Every echo request packet Headend/PMS MUST include the appropriate return path in the Reply Path TLV. The node that receives the echo request MUST follow procedures described in section Section 6.1 and section Section 6.2 to send out echo reply.

For Example:

Let us consider a topology from Figure 1. Let us consider a SR-MPLS path [N-P1,N-ASBR1,EPE-ASBR1-ASBR4, N-PE4]. The traceroute is being executed for this inter-AS path for destination PE4. PE1 sends first echo request with TTL set to 1 and includes Reply path TLV consisting of Type 1 Segment containing label derived from its own SR Global Block (SRGB). Note that the type of segment used in constructing the return Path is local policy. If the entire network has same SRGB configured, Type 1 segments can be used.The TTL expires on P1 and the P1 sends echo reply using the return path. Note that implementations may choose to exclude Reply path TLV until traceroute reaches the first domain border as the return IP path to PE1 is expected to be available inside the first domain.

TTL is set to 2 and the next echo request is sent out. Until the traceroute procedure reaches the domain border node ASBR1, same return path TLV consisting of single Label (PE1's node Label)is used. When echo request reaches ASBR1, and echo reply is received, the next echo request needs to include additional label as ASBR1 is a border node. The Reply path TLV is built based on the forward path. As the forward path consists of EPE-ASBR1-ASBR4, an EPE-SID in the reverse direction is included in the Reply path TLV. The return path now consists of two labels [N-PE1, EPE-ASBR4-ASBR1]. The echo reply from ASBR4 will use this return path to send the reply.

The next echo request after visiting the border node ASBR4 will update the return path with Node-SID label of ASBR4. The return path beyond ASBR4 will be [N-PE1, EPE-ASBR4-ASBR1, N-ASBR4]. This same return path is used until the traceroute procedure reaches next set of border nodes. When there are multiple ASes the traceroute procedure will continue by adding a set of Node labels and EPE labels as the border nodes are visited.

Note that the above return path building procedure requires the database of all the domains to be available at the headend/PMS.

The above description assumed the same SRGB is configured on all nodes along the path. The SRGB may differ from one node to another node and the SR architecture [RFC8402] allows the nodes to use different SRGB. In such scenarios PE1 sends Type 3 (or Type 4 in case of IPv6 networks) segment with Node address of PE1 and with optional MPLS SID associated with the Node address. The receiving node derives the label for the return path based on its own SRGB. When the traceroute procedure crosses the border ASBR1, headend PE1 should send type 1 segment for N-PE1 based on the label derived from ASBR1's SRGB. This is required because in AS2, ASBR4, P3, P4 etc may not have the topology information to derive SRGB for PE1. After the traceroute procedure reaches ASBR4 the return path will be [N-PE1(type1 with label based on ASBR1's SRGB), EPE-ASBR4-ASBR1, N-ASBR4 (Type 3)].

In order to extend the example to multiple ASes consisting of 3 or more ASes, let us consider a traceroute from PE1 to PE5 in Figure 1. In this example, the PE1 to PE5 path has to cross 3 domains AS1, AS2 and AS3. Let us consider a path from PE1 to PE5 that goes through [PE1, ASBR1, ASBR4, ASBR6, ASBR8,PE5]. When the traceroute procedure is visiting the nodes in AS1, the Reply path TLV sent from headend consists of [N-PE1]. When the traceroute procedure reaches the ASBR4, the return Path consists of [N-PE1, EPE-ASBR4-ASBR1]. While visiting nodes in AS2, the traceroute procedure consists of Reply Path TLV [N-PE1, EPE-ASBR4-ASBR1, N-ASBR4]. similarly, while visiting the ASBR8 Reply Path TLV adds the EPE SID from ASBR8 to ASBR6. While visiting nodes in AS3 Node-SId of ASBR8 would also be added which makes the return Path [N-PE1, EPE-ASBR4-ASBR1, N-ASBR4, EPE-ASBR8-ASBR6, N-ASBR8]

Let us consider another example from topology Figure 2. This topology consists of multi-domain IGP with common border node between the domains. This could be achieved with multi-area or multi-level IGP or multiple instances of IGP deployed on same node. The return path computation for this topology is similar to the multi-AS computation except that the return path consists of single border node label. When traceroute procedure is visiting node P, the return path consists of [N-PE1, N-ABR1].

8. Building Reply Path TLV dynamically

In some cases, the head-end may not have complete visibility of Inter-AS/Inter-domain topology. In such cases, it can rely on downstream routers to build the reverse path for MPLS traceroute procedures. For this purpose, Reply Path TLV in the echo reply corresponds to the return path to be used in next echo request.

Value	Meaning
TBA1	Use Reply Path TLV in echo reply for next echo request.

Figure 7: Reply path return Code

8.1. The procedures to build the return path

In order to dynamically build the return Path for traceroute procedures, the domain border nodes along the path being tracerouted MUST support the procedures described in this section. Local policy on the domain border nodes SHOULD determine whether the domain border node participates in building return path dynamically during traceroute.

Headend/PMS node MAY include its own node label while initiating traceroute procedure. When an ABR receives the echo request, if the local policy implies building dynamic return path, ABR MUST include its own Node label. If there is a Reply Path TLV included in the received echo request message, the ABR's node label is added before the existing segments. The type of segment added is based on local policy. In cases when SRGB is not uniform across the network, it is

RECOMMENDED to add type 3 or type 4 segment. If the existing segment in the Reply Path TLV is a type 3/type 4 segment, that segment MUST be converted to Type 1 segment based on ABR's own SRGB.This is because downstream nodes will not know what SRGB to use to translate the IP address to a label. As the ABR added its own Node label, it is guaranteed that this ABR will be in the return path and will be forwarding the traffic based on next label after its own label.

When an ASBR receives an echo request from another AS, and ASBR is configured to build the return path dynamically, ASBR MUST build a Reply Path TLV and include it in the echo reply. The Reply Path TLV MUST consist of its own node label and an EPE-SID to the AS from where the traceroute message was received. A Reply path return code of TBA1 MUST be set in the echo reply to indicate that next echo request should use the return Path from the Reply Path TLV in the echo reply. ASBR MUST locally decide the outgoing interface for the echo reply packet. Generally, remote ASBR will choose interface on which the incoming OAM packet was receieved to send the echo reply out. Reply Path TLV is built by adding two segment sub TLVs. The top segment sub TLV consists of the ASBR's Node SID and second segment consists of the EPE SID in the reverse direction to reach the AS from which the OAM packet was received.The type of segment chosen to build Reply Path TLV is a local policy. It is RECOMMENDED to use type 3/type4 segment for the top segment when the SRGB is not gurateed to be uniform in the domain.

Irrespective of which type of segment is included in the Reply Path TLV, the responder of echo request MUST always translate the Reply Path TLV to a label stack and build MPLS header for the the echo reply packet. This procedure can be applied to an end-to-end path consisting of multiple ASes. Each ASBR that receives echo request from another AS adds its Node-SID and EPE-SID on top of existing segments in the Reply Path TLV.

An ASBR that receives the echo request from a neighbor belonging to same AS, MUST look at the Reply Path TLV received in the echo request. If the Reply Path TLV consists of a Type 3/Type 4 segment, it MUST convert the Type 3/4 segment to Type 1 segment by deriving label from its own SRGB. The ASBR MUST set the reply path return code to TBA1 and send the newly constructed Reply Path TLV in the echo reply.

Internal nodes or non domain border nodes MAY not set the Reply Path TLV return code to TBA1 in the echo reply message as there is no change in the return Path. In these cases, the headend node/PMS that initiates the traceroute procedure MUST continue to send previously sent Reply Path TLV in the echo request message in every next echo request.

Note that an ASBR's local policy may prohibit it from participating in the dynamic traceroute procedures. If such an ASBR is encountered in the forward path, dynamic return path building procedures will fail. In such cases, ASBR that supports this document MUST set the return code TBA2 to indicate local policies do not allow the dynamic return path building.

Value	Meaning
TBA2	Local policy does not allow dynamic return Path building

Figure 8: Local policy return Code

8.2. Details with example

Let us consider a topology from Figure 1. Let us consider a SR policy path built from PE1 to PE4 with a label stack as below. N-P1, N-ASBR1, EPE-ASBR1-ASBR4, N-PE4. PE1 begins traceroute with TTL set to 1 and includes [N-PE1] in the Reply Path TLV. The traceroute packet TTL expires on P1 and P1 processes the traceroute as per the procedures described in [RFC8029] and [RFC8287]. P1 sends echo reply with the same Reply Path TLV with reply path return code set to 6. The return code of the echo reply itself is set to the return code as per [RFC8029] and [RFC8287]. This traceroute doesn't need any changes to the Reply Path TLV till it leaves AS1. The same Reply Path TLV that is received may be included in the echo reply by P1 and P2 or no Reply Path TLV included so that headend continues to use same return path in echo request that it used to send previous echo request.

When ASBR1 receives the echo request, in case it recieved type3/type 4 segment in the Reply Path TLV in the echo request, it converts that type 3/4 segment to Type 1 based on its own SRGB. When ASBR4 receives the echo request, it should form this Reply Path TLV using its own Node SID(N-ASBR4) and EPE SID (EPE-ASRB4-ASBR1) labels and set the reply path return code to TBA1. Then PE1 should use this Reply Path TLV in subsequent echo requests. In this example, when the subsequent echo request reaches P3, it should use this Reply Path TLV for sending the echo reply. The same Reply Path TLV is sufficient for any router in AS2 to send the reply. Because the first label(N-ASBR4) can direct echo reply to ASBR4 and second one (EPE-ASBR4-ASBR1) to direct echo reply to AS1. Once echo reply reaches AS1, normal IP forwarding or the N-PE1 helps it to reach PE1.

The example described in above paragraphs can be extended to multiple ASes by following the same procedure of each ASBR adding Node-SID and EPE-SID on receieving echo request from neighboring AS.

Let us consider a topology from Figure 2. It consists of multiple IGP domains with multiple area/levels or separate IGP instances. There is a single border node that seperates the two domains. In this case, PE1 sends traceroute packet with TTL set to 1 and includes N-PE1 in the Reply path TLV. ABR1 receives the echo request and while sending echo reply adds its own node Label to the Reply Path TLV and sets the Reply path return code to TBA1. The Reply path TLV in the echo reply from ABR1 consists of [N-PE1, N-ABR1]. Next echo request with TTL 2 reaches P node. It is an internal node so it does not change the return Path. echo request with TTL 3 reaches ABR2 and it adds its own Node label so the Reply path TLV sent in echo reply will be [N-PE1, N-ABR1, N-ABR2]. echo request with TTL 4 reaches PE4 and it sends echo reply return code as Egress. PE4 does not include any Reply Path TLV in echo reply. The above example assumes uniform SRGB throughout the domain. In case of different SRGBs, the top segment will be a type 3/4 segment and all other segments will be type 1. Each border node converts the type 3/type 4 segment to type 1 before adding its own segment to the Reply Path TLV.

9. Security Considerations

The procedures described in this document enable LSP ping and traceroute to be executed across multiple domains or multiple ASes that belong to same adminstration or closely co-operating administration. It is assumed that sharing domain internal information across such domains does not pose security risk. However procedures described in this document may be used by an attacker to extract the domain internal information. An operator MUST deploy appropriate filter policies as described in [RFC8029] to restrict the LSP ping/traceroute packets based on origin. It is also suggested that an operator SHOULD deploy security mechanisms such as MACSEC on inter-domain links or security vulnerable links to prevent spoofing attacks.

10. IANA Considerations

Sub-TLVs for TLV Types 1, 16, and 21

SID only in the form of MPLS label : TBD (Range 32768-65535)

IPv4 Node Address with optional SID for SR-MPLS : TBD (Range 32768-65535)

IPv6 Node Address with optional SID for SR-MPLS : TBD (Range 32768-65535)

Reply Path Return Codes Registry

TBA1:Use Reply Path TLV in echo reply for next echo request.

TBA2:Local policy does not allow dynamic return Path building.

11. Contributors

1.Carlos Pignataro

Cisco Systems, Inc.

cpignata@cisco.com

2. Zafar Ali

Cisco Systems, Inc.

zali@cisco.com

12. Acknowledgments

Thanks to Bruno Decreane for suggesting use of generic Segment sub-TLV. Thanks to Adrian Farrel, Huub van Helvoort, Dhruv Dhody, Dongjie, for careful review and comments. Thanks to Mach Chen for suggesting to use Reply Path TLV. Thanks to Gregory Mirsky for detailed review which helped improve the readability of the document to a great extent.

13. References

13.1. 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>.

[RFC7110] Chen, M., Cao, W., Ning, S., Jounay, F., and S. Delord, "Return Path Specified Label Switched Path (LSP) Ping", RFC 7110, DOI 10.17487/RFC7110, January 2014, <https://www.rfc-editor.org/info/rfc7110>.

[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>.

[RFC9087] Filsfils, C., Ed., Previdi, S., Dawra, G., Ed., Aries, E., and D. Afanasiev, "Segment Routing Centralized BGP Egress Peer Engineering", RFC 9087, DOI 10.17487/RFC9087, August 2021, <https://www.rfc-editor.org/info/rfc9087>.

13.2. Informative References

[RFC7743] Luo, J., Ed., Jin, L., Ed., Nadeau, T., Ed., and G. Swallow, Ed., "Relayed Echo Reply Mechanism for Label Switched Path (LSP) Ping", RFC 7743, DOI 10.17487/RFC7743, January 2016, <https://www.rfc-editor.org/info/rfc7743>.

[RFC8277] Rosen, E., "Using BGP to Bind MPLS Labels to Address Prefixes", RFC 8277, DOI 10.17487/RFC8277, October 2017, <https://www.rfc-editor.org/info/rfc8277>.

[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>.

[RFC8604] Filsfils, C., Ed., Previdi, S., Dawra, G., Ed., Henderickx, W., and D. Cooper, "Interconnecting Millions of Endpoints with Segment Routing", RFC 8604, DOI 10.17487/RFC8604, June 2019, <https://www.rfc-editor.org/info/rfc8604>.

[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>.

[RFC9086] Previdi, S., Talaulikar, K., Ed., Filsfils, C., Patel, K., Ray, S., and J. Dong, "Border Gateway Protocol - Link State (BGP-LS) Extensions for Segment Routing BGP Egress Peer Engineering", RFC 9086, DOI 10.17487/RFC9086, August 2021, <https://www.rfc-editor.org/info/rfc9086>.

[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>.

Authors' Addresses

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

Email: shraddha@juniper.net

Kapil Arora
Juniper Networks Inc.

Email: kapilaro@juniper.net

Mukul Srivastava
Juniper Networks Inc.

Email: msri@juniper.net

Samson Ninan
Ciena

Email: samson.cse@gmail.com

Nagendra Kumar
Cisco Systems, Inc.

Email: naikumar@cisco.com