Internet Engineering Task Force (IETF)
Request for Comments: 9716
Category: Standards Track
ISSN: 2070-1721

S. Hegde
Juniper Networks, Inc.
K. Arora
Individual Contributor
M. Srivastava
Juniper Networks, Inc.
S. Ninan
Ciena
N. Kumar
Oracle

February 2025

Mechanisms for MPLS Ping and Traceroute Procedures in Inter-Domain Segment Routing Networks

Abstract

The Segment Routing (SR) architecture leverages source routing and can be directly applied to the use of an MPLS data plane. A Segment Routing over MPLS (SR-MPLS) network may consist of multiple IGP domains or multiple Autonomous Systems (ASes) under the control of the same organization. It is useful to have the Label Switched Path (LSP) ping and traceroute procedures when an SR end-to-end path traverses multiple ASes or IGP domains. This document outlines mechanisms to enable efficient LSP ping and traceroute procedures in inter-AS and inter-domain SR-MPLS networks. This is achieved through a straightforward extension to the Operations, Administration, and Maintenance (OAM) protocol, relying solely on data plane forwarding for handling echo replies on transit nodes.

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/rfc9716.

Table of Contents

1. Introduction

Many network deployments have built their networks consisting of multiple ASes either for the ease of operations or as a result of network mergers and acquisitions. SR can be deployed in such scenarios to provide end-to-end paths, traversing multiple Autonomous Systems (ASes).

[RFC8660] specifies SR with an MPLS data plane. [RFC8402] describes BGP peering segments, and [RFC9087] describes centralized BGP Egress Peer Engineering, which will help in steering packets from one AS to another. By utilizing these SR capabilities, it is possible to create paths that span multiple ASes.

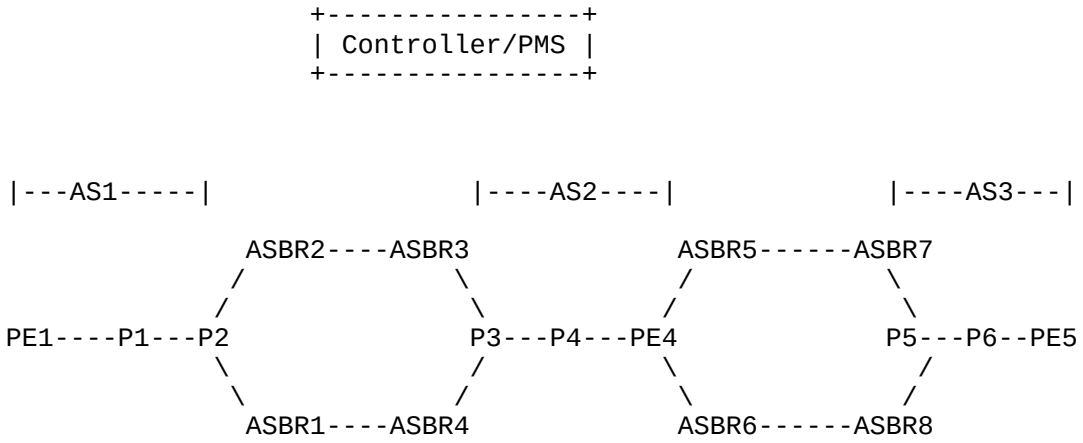

Figure 1: Inter-AS Segment Routing Topology

Autonomous System: AS1, AS2, AS3

Provider Edge: PE1, PE4, PE5

Provider: P1, P2, P3, P4, P5, P6

Autonomous System Boundary Router: ASBR1, ASBR2, ASBR3, ASBR4, ASBR5, ASBR6, ASBR7, ASBR8

For example, Figure 1 describes an inter-AS network scenario consisting of ASes AS1, AS2, and AS3. AS1, AS2, and AS3 are SR enabled, and the egress links have the following Segment Identifiers (SIDs) configured and advertised via [RFC9086]: PeerNode SID, PeerAdj SID, and PeerSet SID. The PeerNode SID, PeerAdj SID, and PeerSet SID are referred to as Egress Peer Engineering SIDs (EPE-SIDs) in this document. The controller or the head-end can build an end-to-end traffic-engineered path consisting of Node-SIDs, Adjacency-SIDs, and EPE-SIDs. It is useful for operators to be able to perform LSP ping and traceroute procedures on these inter-AS SR-MPLS paths, to detect and diagnose failed deliveries, and to determine the actual path that traffic takes through the network. LSP ping and traceroute procedures use IP connectivity for echo replies 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.

It is not always 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 mechanisms (see [RFC8287] and [RFC8029]). That is because there might not always be IP connectivity from a responding node back to the source address of the ping packet when the responding node is in a different AS from the source of the ping.

[RFC8403] describes mechanisms to carry out MPLS ping and 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 the PMS to be capable of building a huge number of GRE tunnels or installing the necessary static routes, which may not be feasible.

[RFC7743] describes an Echo-relay-based solution that is predicated on advertising a new Relay Node Address Stack TLV containing a stack of Echo-relay IP addresses. These mechanisms can be applied to SR networks as well. The mechanism from [RFC7743] requires the return ping packet to be processed on the 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 and traceroute in inter-domain SR networks. The definition of the term "domain" as applicable to this document is defined in Section 1.1.

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

This document focuses on the inter-domain use case. The protocol

extensions described may also indicate the return path for other use cases, which are outside the scope of this document and are not further detailed here. The SRv6 data plane is also not covered in this document.

1.1. Definition of Domain

In this document, the term "domain" refers to an IGP domain where every node is visible to every other node for the purpose of shortest path computation, implying an IGP area or level. An Autonomous System (AS) comprises one or more IGP domains. The procedures described herein are applicable to paths constructed across multiple domains, including both inter-area and inter-AS paths. These procedures and deployment scenarios are relevant for inter-AS paths where the participating ASes are under closely coordinating administrations or single ownership. This document pertains to SR-MPLS networks where all nodes within each domain are SR capable. It also applies to SR-MPLS networks where SR functions as an overlay with SR-incapable underlay nodes. In such networks, the traceroute procedure is executed only on the overlay SR nodes.

1.2. 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.

2. Inter-Domain Networks with Multiple IGPs

When the network consists of a large number of nodes, the nodes are segregated into multiple IGP domains as shown in Figure 2. The connectivity to the remote PEs can be achieved by BGP advertisements with an MPLS label bound to the prefix as described in [RFC8277] or by building paths using a list of segments as described in [RFC8604].

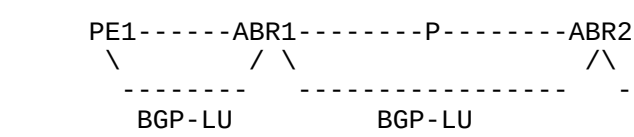

Figure 2: Inter-Domain Networks with Multiple IGPs

It is useful to support MPLS ping and traceroute mechanisms for these networks. The procedures described in this document for constructing the Reply Path TLV and its use in echo replies are equally applicable to networks consisting of multiple IGP domains that use BGP-Labeled Unicast (BGP-LU) or label stacking.

3. Reply Path TLV

The Reply Path (RP) TLV is defined in [RFC7110]. SR networks statically assign the labels to nodes, and a PMS/head-end may know the entire Link State Database (LSDB) along with assigned SIDs. The reverse path can be built from the PMS/head-end by stacking segments for the reverse path. The Reply Path TLV as defined in [RFC7110] is used to carry the return path. Reply Mode 5 (Reply via Specified Path) is defined in Section 4.1 of [RFC7110]. While using the procedures described in this document, the Reply Mode is set to 5 (Reply via Specified Path), and the 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 [RFC7110]. This document

defines three new sub-TLVs to encode the SR 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 Command Line Interface (CLI) input parameters in the form of labels, IPv4 addresses, IPv6 addresses, or a combination of these, which get encoded in the Reply Path TLV. Implementations may also provide mechanisms to acquire the LSDB of remote domains and compute the return path based on the acquired LSDB. 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 the head-end is to acquire the entire LSDB (of all domains) and build a return path for every node along the SR-MPLS path based on the knowledge of the LSDB. Another mechanism is to use a dynamically computed return path as described in Section 5.5.

Some networks may consist of IPv4-only domains and IPv6-only domains. Handling end-to-end MPLS OAM for such networks is out of the scope of this document. It is recommended to use dual-stack in such cases and use end-to-end IPv6 addresses for MPLS ping and traceroute procedures.

4. Segment Sub-TLV

Section 4 of [RFC9256] defines various Segment Types. The types of segments applicable to this document have been defined in this section for the use of MPLS OAM. The intention was to keep the definitions as close to those in [RFC9256] as possible, with modifications only when needed. One or more Segment sub-TLVs can be included in the Reply Path TLV. The Segment sub-TLVs included in a Reply Path TLV MAY be of different types.

The below types of Segment sub-TLVs apply to the Reply Path TLV. The code points for the sub-TLVs are taken from the IANA registry common to TLVs 1, 16, and 21. This document defines the usage and processing of the Type-A, Type-C, and Type-D Segment sub-TLVs when they appear in TLV 21 (Reply Path TLV). If these sub-TLVs appear in TLVs 1 or 16, appropriate error codes MUST be returned as defined in [RFC8029].

Type-A: SID only, in the form of an MPLS label

Type-C: IPv4 Node Address with an optional SID

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

4.1. Type-A: SID Only, in the Form of an MPLS Label

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

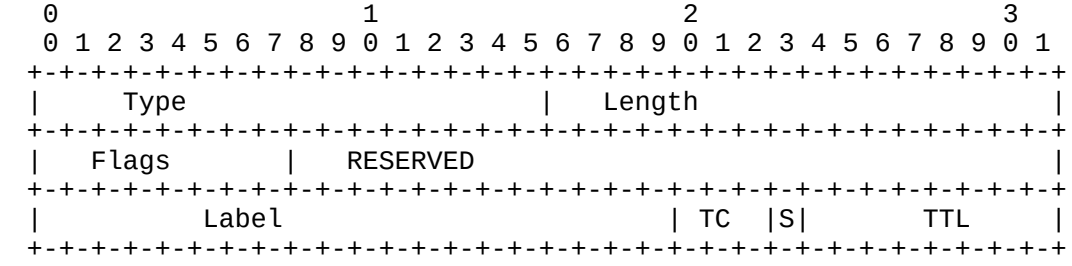

Figure 3: Type-A Segment Sub-TLV

Where:

Type: 2 octets. Carries value 46 (assigned by IANA from the "Sub-TLVs for TLV Types 1, 16, and 21" registry).

Length: 2 octets. Carries value 8. The length value excludes the length of the Type and Length fields.

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

RESERVED: 3 octets of reserved bits. MUST be set to zero when sending; MUST be ignored on receipt.

Label: 20 bits of label value.

TC: 3 bits of Traffic Class (TC). If the originator wants the receiver to choose the TC value, it MUST set the TC field to zero.

S: 1 bit Reserved. The S bit MUST be zero upon transmission and MUST be ignored upon reception.

TTL: 1 octet of TTL. If the originator wants the receiver to choose the TTL value, it MUST set the TTL field to 255.

The labels, TC, S, and TTL are collectively referred to as a SID.

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

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 an 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:

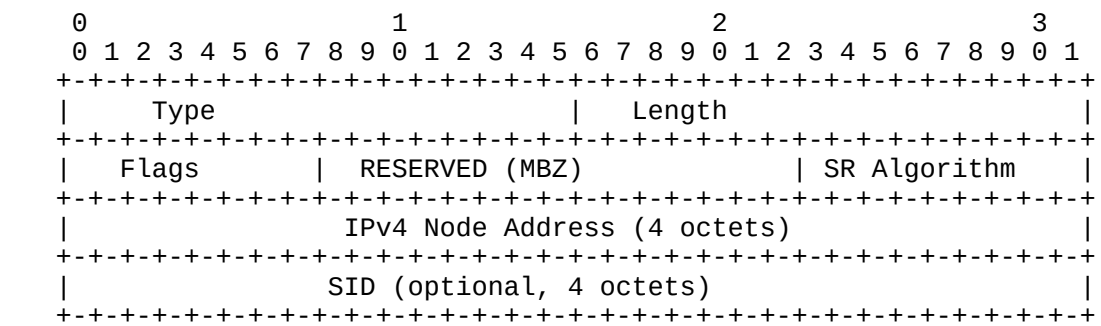

Figure 4: Type-C Segment Sub-TLV

Where:

Type: 47 (assigned by IANA from the "Sub-TLVs for TLV Types 1, 16, and 21" registry).

Length: 2 octets. Carries value 8 when no optional SID is included or value 12 when the optional SID is included.

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

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

SR Algorithm: 1 octet. When the A-Flag (as defined in Section 4.4) is present, this specifies the SR Algorithm as described in

Section 3.1.1 of [RFC8402] or the Flexible Algorithm as defined in [RFC9350]. The SR Algorithm is used by the receiver to derive the label. When the A-Flag is unset, this field has no meaning and thus MUST be set to zero (MBZ) on transmission and ignored on receipt.

IPv4 Node Address: 4-octet IPv4 address representing a node. The IPv4 Node Address MUST be present. It should be a stable address belonging to the node (e.g., loopback address).

SID: Optional 4-octet field containing the labels TC, S, and TTL as defined in Section 4.1. When the SID field is present, it MUST be used for constructing the Reply Path.

4.3. Type-D: IPv6 Node Address with an 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:

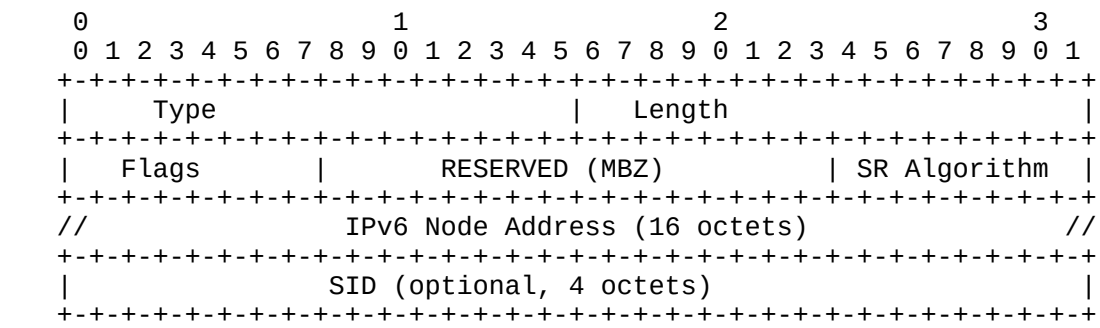

Figure 5: Type-D Segment Sub-TLV

Where:

Type: 48 (assigned by IANA from the "Sub-TLVs for TLV Types 1, 16, and 21" registry).

Length: 2 octets. Carries value 20 when no optional SID is included or value 24 when the optional SID is included.

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

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

SR Algorithm: 1 octet. When the A-Flag (as defined in Section 4.4) is present, this specifies the SR Algorithm as described in Section 3.1.1 of [RFC8402] or the Flexible Algorithm as defined in [RFC9350]. The SR Algorithm is used by the receiver to derive the label. When the A-Flag is unset, this field has no meaning and thus MUST be set to zero (MBZ) on transmission and ignored on receipt.

IPv6 Node Address: 16-octet IPv6 address of one interface of a node. The IPv6 Node Address MUST be present. It should be a stable address belonging to the node (e.g., loopback address).

SID: Optional 4-octet field containing the labels TC, S, and TTL as defined in Section 4.1. When the SID field is present, it MUST be used for constructing the Reply Path.

4.4. Segment Flags

The Segment Types described above contain the following flags in the

Flags field (codes assigned by IANA from the "Segment ID Sub-TLV Flags" registry):

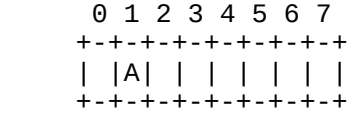

Figure 6: Flags

Where:

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

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

The following applies to the Segment Flags:

The A-Flag applies to Segment Type-C and Type-D. If the A-Flag appears with the Type-A Segment Type, it MUST be ignored.

5. Detailed Procedures

This section uses the term "initiator" for the node that initiates the MPLS ping or the MPLS traceroute procedure. The term "responder" is used for the node that receives the echo request and sends the echo reply. The term "egress node" is used to identify the last node where the MPLS ping or traceroute is destined to. In an MPLS network, any node can be an initiator, responder, or egress.

5.1. Sending an Echo Request

In the inter-AS scenario, the procedures outlined in this document are employed to specify the return path when IP connectivity to the initiator is unavailable. These procedures may also be utilized regardless of the availability of IP connectivity. The LSP ping initiator MUST set the Reply Mode of the echo request to 5 (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 SR Path in the reverse direction encoded as an ordered list of segments. The first segment MUST correspond to the top segment in the MPLS header that the responder MUST use while sending the echo reply.

5.2. Receiving an Echo Request

As described in [RFC7110], when the Reply Mode is set to 5 (Reply via Specified Path), the echo request must contain the Reply Path TLV. The absence of the Reply Path TLV is treated as a malformed echo request. When an echo request is received, if the responder does not support 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 must be sent back to the initiator according to the rules of [RFC8029]. If the echo request message contains a malformed Segment sub-TLV, such as an incorrect length field, an echo reply must be sent back to the initiator with the Return Code set to "Malformed echo request received" and the Subcode set to zero.

When a Reply Path TLV is received, the responder that supports processing it MUST use the segments in Reply Path TLV to build the echo reply. The responder MUST follow the normal Forwarding Equivalence Class (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 MUST be used to construct the MPLS packet to send out.

5.3. Sending an Echo Reply

The echo reply message is sent as an MPLS packet with an MPLS label stack. The echo reply message MUST be constructed as described in [RFC8029]. An MPLS packet is constructed with an echo reply in the payload. The top label MUST be constructed from the first segment of the Reply Path TLV. The remaining labels MUST be constructed by following the order of the segments from the Reply Path TLV. The MPLS header of the echo reply MUST be constructed from the segments in the Reply Path TLV and MUST NOT add any other label. The S bit is set for the bottom label as per the MPLS specifications [RFC3032]. The responder MAY check the reachability of the top label in its own Label Forwarding Information Base (LFIB) before sending the echo reply. If the top label is unreachable, the responder SHOULD send the appropriate Return Code and follow the procedures as per Section 5.2 of [RFC7110]. The exception case is when the responder does not have IP reachability to the originator, in which case, it may not be possible to send an echo reply at all. Even if sent (by following a default route present on the responder, for example), the echo reply might not reach the originator. The node MAY provide necessary log information in case of unreachability. In certain scenarios, the head-end MAY choose to send Type-C/Type-D segments consisting of IPv4 addresses or IPv6 addresses when it is unable to derive the SID from available topology information. Optionally, the SID may also be associated with the Type-C/Type-D segment, if such information is available from the controller or via operator input. In such cases, the node sending the echo reply MUST derive the MPLS labels based on the Node-SIDs associated with the IPv4/IPv6 addresses. If an optional MPLS SID is present in the Type-C/Type-D segments, the SID MUST be used to encode the echo reply with MPLS labels. If the MPLS SID does not match with the IPv4 or IPv6 address field in the Type-C or Type-D SID, log information should be generated.

The Reply Path Return Code is set as described in Section 7.4 of [RFC7110]. According to Section 5.3 of [RFC7110], the Reply Path TLV is included in an echo reply indicating the specified return path that the echo reply message is required to follow.

When the node is configured to dynamically create a return path for the next echo request, the procedures described in Section 5.5 MUST be used. The Reply Path Return Code MUST be set to 0x0006, and the same Reply Path TLV or a new Reply Path TLV MUST be included in the echo reply.

5.4. Receiving an Echo Reply

The rules and processes 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 from this echo reply for building the next echo request" (as defined in this document), the Reply Path TLV from the echo reply MUST be sent in the next echo request with the TTL incremented by 1. If the initiator node does not support the Return Code "Use Reply Path TLV from this echo reply for building the next echo request", log information should be generated indicating the Return Code, and the operator may choose to specify the return path explicitly or use other mechanisms to verify the SR Policy. If the Return Code is 0x0007 "Local policy does not allow dynamic return path building", it indicates that the intermediate node does not support building the dynamic return path. Log information should be generated on the initiator receiving this Return Code, and the operator may choose to specify the return path explicitly or use other mechanisms to verify the SR Policy. If the TTL is already 255, the traceroute procedure MUST be ended with an appropriate log

message.

5.5. Building a 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 routers in the path to build the reverse path for MPLS traceroute procedures. For this purpose, the Reply Path TLV in the echo reply corresponds to the return path to be used in building the next echo request. A new Return Code "Use Reply Path TLV from this echo reply for building the next echo request" is defined in this document.

Value	Meaning
0x0006	Use Reply Path TLV from this echo reply for building the next echo request

Table 1

5.5.1. Procedures to Build the Return Path

To dynamically build the return path for the traceroute procedures, the domain border nodes along the path being traced should support the procedures described in this section. Local policy on the domain border nodes should determine whether the domain border node participates in building the return path dynamically during traceroute.

The head-end/PMS node may include its node label while initiating the traceroute procedure. When an Area Border Router (ABR) receives the echo request, if the local policy implies building a dynamic return path, the ABR should include its node label in the Reply Path TLV and send it in the echo reply. 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 the Segment Routing Global Block (SRGB) is not uniform across the network, which can be inferred from the LSDB, it is RECOMMENDED to add a Type-C or a Type-D segment. However, implementations MAY safely use other approaches if they see benefits in doing so. If the existing segment in the Reply Path TLV is a Type-C/Type-D segment, that segment should be converted to a Type-A segment based on the ABR’s own SRGB. This is because downstream nodes in the path 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 the next label after its label.

When an ASBR receives an echo request from another AS, and the ASBR is configured to build the return path dynamically, the ASBR should build a Reply Path TLV and include it in the echo reply. The Reply Path TLV should consist of its node label and an EPE-SID to the AS from where the traceroute message was received. A Reply Path Return Code of 0x0006 MUST be set in the echo reply to indicate that the next echo request MUST use the return path from the Reply Path TLV in the echo reply. ASBR should locally decide the outgoing interface for the echo reply packet. Generally, remote ASBR will choose the interface on which the incoming OAM packet was received to send the echo reply out. In case the ASBR identifies multiple paths to reach the initiator, it MUST choose to send one such path in the Reply Path TLV. The Reply Path TLV is built by adding two Segment sub-TLVs. The top Segment sub-TLV consists of the ASBR’s Node-SID, and the 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 the Reply Path TLV is a local policy. It is recommended to use the Type-C/Type-D segment for the top segment when the SRGB is not guaranteed to be uniform in the domain.

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

An ASBR that receives the echo request from a neighbor belonging to the same AS MUST look at the Reply Path TLV received in the echo request. If the Reply Path TLV consists of a Type-C/Type-D segment, it MUST convert the Type-C/Type-D segment to a Type-A segment by deriving a label from its own SRGB. The ASBR MUST set the Reply Path Return Code to 0x0006 and send the newly constructed Reply Path TLV in the echo reply.

Internal nodes or non-domain border nodes might not set the Reply Path TLV Return Code to 0x0006 in the echo reply message as there is no change in the return path. In these cases, the head-end node/PMS that initiates the traceroute procedure MUST continue to send the previously sent Reply Path TLV in the echo request message in every subsequent 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, an ASBR that supports this document MUST set the Return Code to 0x0007 to indicate that local policies do not allow the dynamic return path building.

Value	Meaning
0x0007	Local policy does not allow dynamic return path building

Table 2

6. Security Considerations

The procedures described in this document enable LSP ping and traceroute procedures to be executed across multiple IGP domains or multiple ASes that belong to the same administration or closely cooperating administrations. It is assumed that sharing domain internal information across such domains does not pose a security risk. However, the procedures described in this document may be used by an attacker to extract the domain's internal information. An operator MUST deploy appropriate filter policies as described in [RFC8029] to restrict the LSP ping and traceroute packets based on origin. It is also RECOMMENDED that an operator deploy security mechanisms such as Media Access Control Security (MACsec) [IEEE-802.1AE] on inter-domain links or security-vulnerable links to prevent spoofing attacks.

All the security considerations defined in [RFC8029] will be applicable for this document. Appropriate filter policies SHOULD be applied at the edges to prevent attackers from getting into the network. In the event of such a security breach, the network devices MUST have mechanisms to prevent denial-of-service attacks as described in [RFC8029].

7. IANA Considerations

7.1. Segment Sub-TLV

IANA has assigned three new sub-TLVs from the "Sub-TLVs for TLV Types 1, 16, and 21" registry of the "Multiprotocol Label Switching (MPLS) Label Switched Paths (LSPs) Ping Parameters" registry group.

Sub-Type	Sub-TLV Name	Reference
46	SID only, in the form of MPLS label	Section 4.1 of RFC 9716
47	IPv4 Node Address with an optional SID for SR-MPLS	Section 4.2 of RFC 9716
48	IPv6 Node Address with an optional SID for SR-MPLS	Section 4.3 of RFC 9716

Table 3

The code points for the Segment sub-TLVs have been registered in the Standards Action range (0-16383).

7.2. New Registry for Segment ID Sub-TLV Flags

IANA has created a new "Segment ID Sub-TLV Flags" registry (see Section 4.4) under the "Multiprotocol Label Switching (MPLS) Label Switched Paths (LSPs) Ping Parameters" registry group.

This registry tracks the assignment of 8 flags in the Segment ID sub-TLV flags field. The flags are numbered from 0 (the most significant bit and transmitted first) to 7.

New entries are assigned by Standards Action. Initial entries in the registry are as follows:

Bit Number	Name	Reference
1	A-Flag	Section 4.4 of RFC 9716

Table 4

7.3. Reply Path Return Codes Registry

IANA has assigned new Return Codes in the "Reply Path Return Codes" registry under the "Multiprotocol Label Switching (MPLS) Label Switched Paths (LSPs) Ping Parameters" registry group.

Value	Meaning	Reference
0x0006	Use Reply Path TLV from this echo reply for building the next echo request	RFC 9716
0x0007	Local policy does not allow dynamic return path building	RFC 9716

Table 5

The Return Codes have been registered in the Standards Action range

(0x0000-0xFFFB).

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>.

[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>.

8.2. Informative References

[IEEE-802.1AE] IEEE, "IEEE Standard for Local and metropolitan area networks-Media Access Control (MAC) Security", IEEE Std 8021.AE-2018, DOI 10.1109/IEEESTD.2018.8585421, December 2018, <https://ieeexplore.ieee.org/document/8585421>.

[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>.

[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>.

[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>.

[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>.

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

[RFC9552] Talaulikar, K., Ed., "Distribution of Link-State and Traffic Engineering Information Using BGP", RFC 9552, DOI 10.17487/RFC9552, December 2023, <https://www.rfc-editor.org/info/rfc9552>.

Appendix A. Examples

This section elaborates examples of the inter-domain ping and traceroute procedures described in this document.

A.1. Detailed Example

The example topology given in Figure 1 will be used in the below sections to explain LSP ping and traceroute procedures. The PMS/ head-end has a complete view of the topology. PE1, P1, P2, ASBR1, and ASBR2 are in AS1. Similarly, ASBR3, ASBR4, P3, P4, and PE4 are in AS2.

AS1 and AS2 have SR enabled. IGPs like OSPF/IS-IS are used to flood SIDs in each AS. ASBR1, ASBR2, ASBR3, and ASBR4 advertise BGP EPE-SIDs for the inter-AS links. The topologies 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 this document uses the notations below for SIDs.

Node-SIDs: N-PE1, N-P1, N-ASBR1, etc.

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

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

A.1.1. Procedures for Segment Routing LSP Ping

Consider an SR-MPLS path from PE1 to PE4 consisting of a label stack [N-P1, N-ASBR1, EPE-ASBR1-ASBR4, N-PE4] from Figure 1. In order to perform MPLS ping procedures on this path, the remote end (PE4) needs IP connectivity to head-end PE1 for the echo reply to travel back to PE1. In a deployment that uses a controller-computed inter-domain path, there may be no IP connectivity from PE4 to PE1 as they lie in different ASes.

PE1 sends an echo request message to the endpoint PE4 along the path that consists of label stacks [N-P1, N-ASBR1, EPE-ASBR1-ASBR4, N-PE4]. PE1 adds the return path from PE4 to PE1 in the echo request message in the Reply Path TLV. As an example, the Reply Path TLV for PE1 to PE4 for LSP ping is [N-ASBR4, EPE-ASBR4-ASBR1, N-PE1]. This example path provides the entire return path up to the head-end node PE1. The mechanism used to construct the return path is implementation dependent.

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 an MPLS echo request, PE4 first validates the FEC in the echo request. PE4 then builds a label stack to send the response from PE4 to PE1 by copying the labels from the Reply Path TLV. PE4 builds the echo reply packet with the MPLS label stack constructed, imposes MPLS headers on top of the echo reply packet, and sends out the packet to PE1. This segment list stack can successfully steer the reply back to the head-end node (PE1).

Let us consider a case when the P3 node does not have a route to reach N-PE4. On P3, a ping packet would be dropped, and the head-end node (PE1) will not receive an echo reply indicating failure.

A.1.2. Procedures for SR LSP Traceroute

A.1.2.1. Procedures for SR LSP Traceroute with the Same SRGB on All Nodes

The traceroute procedure involves visiting every node on the path and obtaining echo replies from every node. In this section, we describe the traceroute mechanisms when the head-end/PMS has complete visibility of the LSDB. The head-end/PMS computes the return path from each node in the entire SR-MPLS path that is being tracerouted. The return path computation is implementation dependent. As the head-end/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 the border node’s Node-SID, the EPE-SID in the reverse direction also needs to be added to the label stack.

The inter-domain/inter-AS traceroute procedure uses the TTL expiry mechanism as specified in [RFC8029] and [RFC8287]. Every echo request packet head-end/PMS will include the appropriate return path in the Reply Path TLV. The node that receives the echo request will follow procedures described in Sections 5.1 and 5.2 to send out an echo reply.

For example:

Let us consider the topology from Figure 1. Let us consider an 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 the first echo request with the TTL set to 1 and includes a Reply Path TLV consisting of a Type-A segment containing a label derived from its own SRGB. Note that the type of segment used in constructing the return path is determined by local policy. If the entire network has the same SRGB configured, Type-A segments can be used. The TTL expires on P1, and P1 sends an echo reply using the return path. Note that implementations may choose to exclude the Reply Path TLV until the traceroute reaches the first domain border as the return IP path to PE1 is expected to be available inside the first domain.

The TTL is set to 2, and the next echo request is sent out. Until the traceroute procedure reaches the domain border node ASBR1, the same return path TLV consisting of a single label (PE1’s node label) is used. When an echo request reaches the border node ASBR1, and an echo reply is received from ASBR1, the next echo request needs to include an additional label as ASBR1 is a border node. The head-end node has complete visibility of the network LSDB learned via BGP-LS (see [RFC9552] and [RFC9086]) and can derive the details of ASBR nodes. 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: [EPE-ASBR4-ASBR1, N-PE1]. The echo reply from ASBR4 will use this return path to send the reply.

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

Note that the above return path building procedure requires the LSDB of all the domains to be available at the head-end/PMS.

Let us consider a case when the P3 node does not have a route to reach N-PE4. When the TTL of the packet is 5, the packet reaches P3, its TTL becomes zero, and it is sent to the control plane. The FEC validation procedures are executed, and the echo reply is sent using the labels in the Reply Path TLV, which is [N-PE1, EPE-ASBR4-ASBR1, N-ASBR4]. The head-end PE1 increases the TTL to 6 and sends the next echo request. The packet is dropped at P3 as there is no route on P3 to forward to N-PE4. The traceroute identifies that the path [N-P1, N-ASBR1, EPE-ASBR1-ASBR4, N-PE4] is broken at P3.

A.1.2.2. Procedures for SR LSP Traceroute with Different SRGBs

Appendix A.1.2.1 assumes 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 SRGBs. In such scenarios, PE1 finds out the difference in the SRGB by looking into the LSDB. Then, it sends the Type-C segment (or the Type-D segment, in the case of IPv6 networks) with the node address of PE1 and with an 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, head-end PE1 should send a Type-A segment for N-PE1 based on the label derived from ASBR1’s SRGB. This is required because 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 (Type-A with the label based on ASBR1's SRGB), EPE-ASBR4-ASBR1, N-ASBR4 (Type-C)].

If the packet needs to follow a return path specific to an algorithm (as defined in [RFC9350]), a Type-C Segment sub-TLV with a corresponding algorithm field set should be used. The A-Flag should be set to indicate that the SID corresponding to the algorithm should be used.

To extend the example to three 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 three 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 the head-end 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 the Reply Path TLV [N-PE1, EPE-ASBR4-ASBR1, N-ASBR4]. Similarly, while visiting ASBR8, the EPE-SID from ASBR8 to ASBR6 is added to the Reply Path TLV. While visiting nodes in AS3, the 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 the topology in Figure 2. This topology consists of multi-domain IGP with a common border node between the domains. This could be achieved with multi-area or multi-level IGP or with multiple instances of IGP deployed on the same node. The return path computation for this topology is similar to multi-AS computation, except that the return path consists of a single border node label.

A.1.3. Procedures for Building Reply Path TLV Dynamically

Let us consider the topology from Figure 1. Let us consider an SR Policy path built from PE1 to PE4 with the following label stack: N-P1, N-ASBR1, EPE-ASBR1-ASBR4, N-PE4. PE1 begins traceroute procedures with the 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 an echo reply with the same Reply Path TLV with the 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 until 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 is included so that the head-end continues to use the same return path in the echo request that it used to send the previous echo request.

When ASBR1 receives the echo request, in the case it receives the Type-C/Type-D segment in the Reply Path TLV in the echo request, it converts that Type-C/Type-D segment to Type-A based on its own SRGB. When ASBR4 receives the echo request, it should form this Reply Path TLV using its Node-SID (N-ASBR4) and EPE-SID (EPE-ASRB4-ASBR1) labels and set the Reply Path Return Code to 0x0006. 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. This is because the first label (N-ASBR4) can direct the echo reply to ASBR4 and the second one (EPE-ASBR4-ASBR1) can direct the echo reply to AS1. Once the echo reply reaches AS1, normal IP forwarding or the N-PE1 helps it to reach PE1.

The example described in the above paragraphs can be extended to

multiple ASes. This is done by following the same procedure for each ASBR, i.e., adding Node-SIDs and EPE-SIDs on receiving echo requests from neighboring ASes.

Let us consider the topology from Figure 2. It consists of multiple IGP domains with multiple areas/levels or separate IGP instances. There is a single border node that separates the two domains. In this case, PE1 sends a traceroute packet with the TTL set to 1 and includes N-PE1 in the Reply Path TLV. ABR1 receives the echo request, adds its node label to the Reply Path TLV (while sending the echo reply), and sets the Reply Path Return Code to 0x0006. The Reply Path TLV in the echo reply from ABR1 consists of [N-ABR1, N-PE1]. The next echo request with a TTL of 2 reaches the P node. It is an internal node, so it does not change the return path. The echo request with a TTL of 3 reaches ABR2, and it adds its node label so the Reply Path TLV sent in the echo reply will be [N-ABR2, N-ABR1, N-PE1]. The echo request with a TTL of 4 reaches PE4, and it sends an echo reply Return Code as an egress. PE4 does not include any Reply Path TLVs in the echo reply. The above example assumes a uniform SRGB throughout the domain. In the case of different SRGBs, the top segment will be a Type-C/Type-D segment and all other segments will be Type-A. Each border node converts the Type-C/Type-D segment to Type-A before adding its segment to the Reply Path TLV.

Acknowledgments

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

Contributors

Carlos Pignataro
NC State University
Email: cpignata@gmail.com

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

Authors' Addresses

Shraddha Hegde
Juniper Networks, Inc.
Exora Business Park
Bangalore 560103
KA
India
Email: shraddha@juniper.net

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

Mukul Srivastava
Juniper Networks, Inc.
Email: msri@juniper.net

Samson Ninan
Ciena
Email: samson.cse@gmail.com

Nagendra Kumar
Oracle
Email: nagendrakumar.nainar@gmail.com