Workgroup: Routing area
Internet-Draft: draft-ietf-mpls-sr-epe-oam-08
Published: 9 February 2023
Intended Status: Standards Track
Expires: 13 August 2023
Authors: S. Hegde
Juniper Networks Inc.
M. Srivastava
Juniper Networks Inc.
K. Arora
Individual Contributor
S. Ninan
Individual Contributor
X. Xu
China Mobile

Label Switched Path (LSP) Ping/Traceroute for Segment Routing (SR) Egress Peer Engineering Segment Identifiers (SIDs) with MPLS Data Planes

Abstract

Egress Peer Engineering (EPE) is an application of Segment Routing to solve the problem of egress peer selection. The Segment Routing based BGP-EPE solution allows a centralized controller, e.g. a Software Defined Network (SDN) controller to program any egress peer. The EPE solution requires a node to program the PeerNode Segment Identifier(SID) describing a session between two nodes, the PeerAdj SID describing the link (one or more) that is used by sessions between peer nodes, and the PeerSet SID describing an arbitrary set of sessions or links between a local node and its peers. This document provides new sub-TLVs for EPE Segment Identifiers (SID) that would be used in the MPLS Target stack TLV (Type 1), in MPLS Ping and Traceroute procedures.

Table of Contents

1. Introduction
2. Theory of Operation
3. Requirements Language
4. FEC Definitions
   4.1. PeerAdj SID Sub-TLV
   4.2. PeerNode SID Sub-TLV
   4.3. PeerSet SID Sub-TLV
5. EPE-SID FEC validation
   5.1. EPE-SID FEC validiation
6. IANA Considerations
7. Security Considerations
8. Acknowledgments
9. References
   9.1. Normative References
   9.2. Informative References

Authors' Addresses

1. Introduction

Egress Peer Engineering (EPE) as defined in [RFC9087] is an effective mechanism to select the egress peer link based on different criteria. The EPE-SIDs provide means to represent egress peer links. 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 acquisitons. The inter-AS links connecting any two Autonomous Systems could be traffic engineered using EPE-SIDs in this case as well. It is important to be able to validate the control plane to forwarding plane synchronization for these SIDs so that any anomaly can be detected easily by the operator.

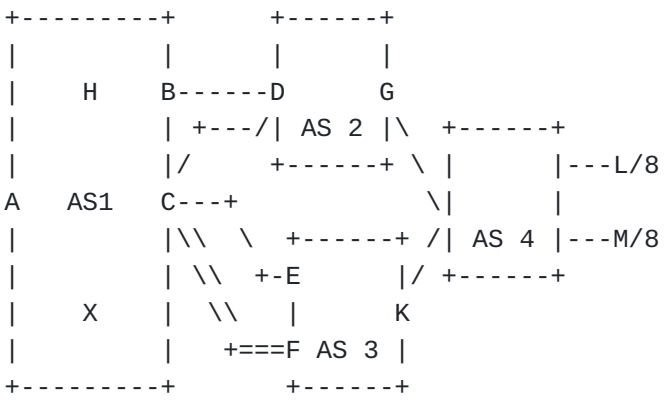

Figure 1: Reference Diagram

In this reference diagram, EPE-SIDs are advertised from AS1 to AS2 and AS3. In certain cases the EPE-SIDs advertised by the control plane may not be in synchronization with label programmed in data-plane. For example, on C a PeerAdj SID could be advertised to indicate it is for the link C->D. Due to some software anomaly the actual data forwarding on this PeerAdj SID could be happening over C->E link. If E had relevant data paths for further forwarding the packet, this kind of anomalies will go unnoticed by the operator. A FEC definition for the EPE-SIDs will define the details of the control plane association of the SID. The data plane validation of the SID will be done during the MPLS trace route procedure. When there is a multi-hop EBGP session between the ASBRs, PeerNode SID is advertised and traffic would be load-balanced between the interfaces connecting the two nodes. In the reference diagram C and F could have a PeerNode-SID advertised. When the OAM packet is received on F, it needs to validate that the packet came on one of the two interfaces connected to C.

This document provides Target Forwarding Equivalence Class (FEC) stack TLV definitions for EPE-SIDs. Other procedures for MPLS Ping and Traceroute as defined in [RFC8287] section 7 and clarified by [RFC8690] are applicable for EPE-SIDs as well.

2. Theory of Operation

[RFC9086] provides mechanisms to advertise the EPE-SIDs in BGP-LS. These EPE-SIDs may be used to build Segment Routing paths as described in [I-D.ietf-idr-segment-routing-te-policy] or using Path Computation Element Protocol (PCEP) extensions as defined in [RFC8664]. Data plane monitoring for such paths which consist of EPE-SIDs will use extensions defined in this document to build the Target FEC stack TLV. The MPLS Ping and Traceroute procedures MAY be initiated by the head-end of the Segment Routing path or a

centralized topology-aware data plane monitoring system as described in [RFC8403]. The extensions in [I-D.ietf-idr-segment-routing-te-policy] and [RFC8664] do not define how to carry the details of the SID that can be used to construct the FEC. Such extensions are out of scope for this document. The node initiating the data plane monitoring may acquire the details of EPE-SIDs through BGP-LS advertisements as described in [RFC9086]. There may be other possible mechanisms to learn the definition of the SID from controller. Details of such mechanisms are out of scope for this document.

The EPE-SIDs are advertised for inter-AS links which run EBGP sessions. The procedures to operate EBGP sessions in a scenario with unnumbered interfaces is not very well defined in [RFC9086] and hence out of scope for this document. During AS migration scenario procedures described in [RFC7705] may be in force. In these scenarios, if the local and remote AS fields in the FEC as described in Section 4 carries the globally configured ASN and not the "local AS" as defined in [RFC7705], the FEC validation procedures may fail.

3. 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.

4. FEC Definitions

Three new sub-TLVs are defined for the Target FEC Stack TLV (Type 1), the Reverse-Path Target FEC Stack TLV (Type 16), and the Reply Path TLV (Type 21).

Sub-Type	Sub-TLV Name
TBD1	PeerAdj SID Sub-TLV
TBD2	PeerNode SID Sub-TLV
TBD3	PeerSet SID Sub-TLV

Figure 2: New sub-TLV types

4.1. PeerAdj SID Sub-TLV

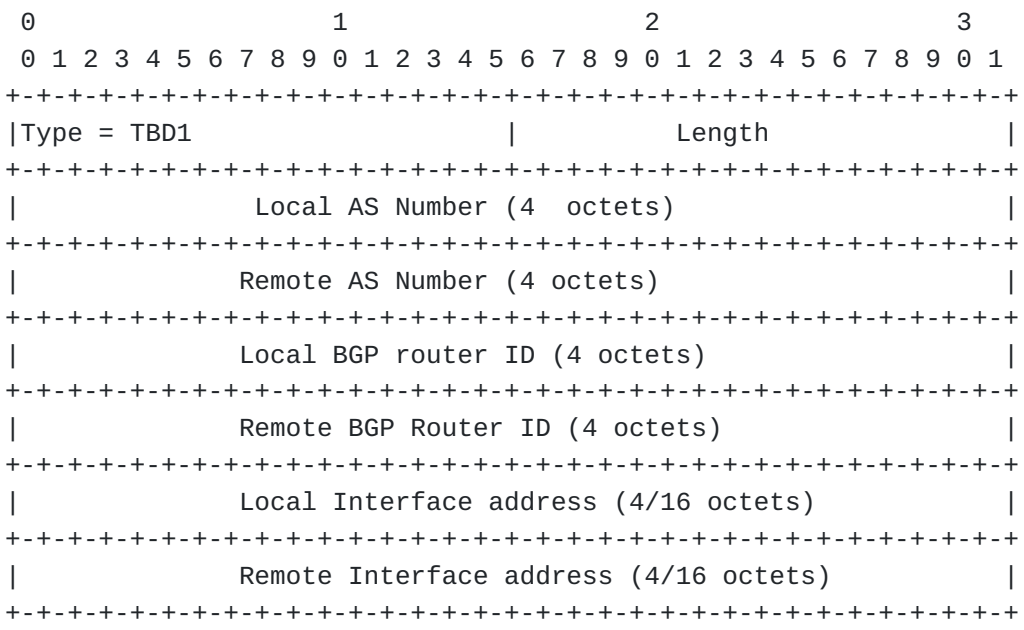

Figure 3: PeerAdj SID Sub-TLV

Type : TBD1

Length : variable based on IPV4/IPV6 interface address. Length excludes the length of Type and Length field.For IPV4 interface addresses length will be 24. In case of IPV6 address length will be 48.

Local AS Number :

4 octet unsigned integer representing the Member-AS Number inside the Confederation [RFC5065]. The AS number corresponds to the AS to which PeerAdj SID advertising node belongs to.

Remote AS Number :

4 octet unsigned integer representing the Member-AS Number inside the Confederation [RFC5065]. The AS number corresponds to the AS of the remote node for which the PeerAdj SID is advertised.

Local BGP Router ID :

4 octet unsigned integer of the advertising node representing the BGP Identifier as defined in [RFC4271] and [RFC6286].

Remote BGP Router ID :

4 octet unsigned integer of the receiving node representing the BGP Identifier as defined in [RFC4271] and [RFC6286].

Local Interface Address :

In case of PeerAdj SID, Local interface address corresponding to the PeerAdj SID should be specified in this field. For IPV4,this field is 4 octets; for IPV6, this field is 16 octets. Link Local IPV6 addresses are for further study.

Remote Interface Address :

In case of PeerAdj SID Remote interface address corresponding to the PeerAdj SID should be apecified in this field. For IPV4, this field is 4 octets; for IPV6, this field is 16 octets. Link Local IPv6 addresses are for further study.

[RFC9086] mandates sending local interface ID and remote interface ID in the Link Descriptors and allows a value of 0 in the remote descriptors. It is useful to validate the incoming interface for a OAM packet and if the remote descriptor is 0 this validation is not possible. [RFC9086] allows optional link descriptors of local and remote interface addresses as described in section 4.2. This document RECOMMENDs sending these optional descriptors and use them to validate incoming interface. When these local and remote interface addresses are not available, an ingress node can send 0 in the local and/or remote interface address field. The receiver SHOULD skip the validation for the incoming interface if the address field contains 0.

4.2. PeerNode SID Sub-TLV

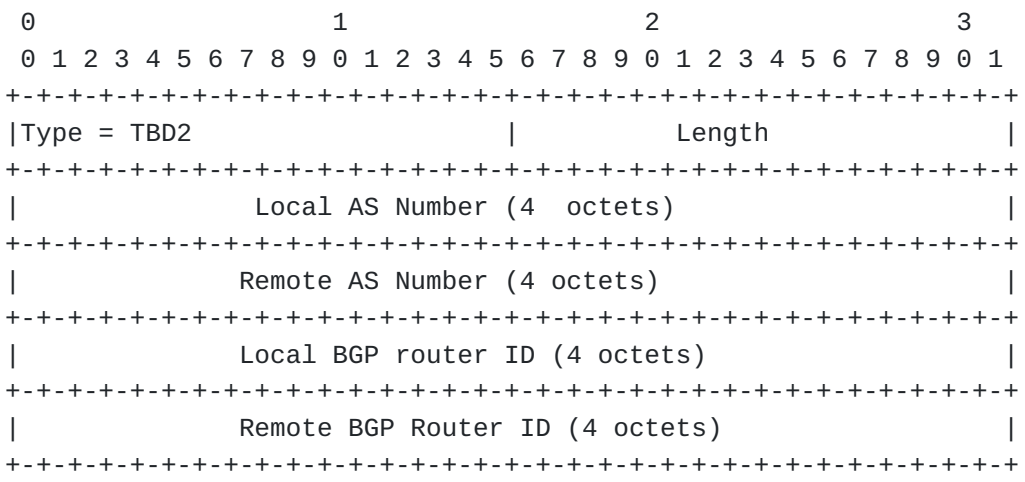

Figure 4: PeerNode SID Sub-TLV

Type : TBD2

Length : 16

Local AS Number :

4 octet unsigned integer representing the Member-AS Number inside the Confederation [RFC5065]. The AS number corresponds to the AS to which PeerNode SID advertising node belongs to.

Remote AS Number :

4 octet unsigned integer representing the Member-AS Number inside the Confederation [RFC5065]. The AS number corresponds to the AS of the remote node for which the PeerNode SID is advertised.

Local BGP Router ID :

4 octet unsigned integer of the advertising node representing the BGP Identifier as defined in [RFC4271] and [RFC6286].

Remote BGP Router ID :

4 octet unsigned integer of the receiving node representing the BGP Identifier as defined in [RFC4271] and [RFC6286].

When there is a multi-hop EBGP session between two ASBRs, PeerNode SID is advertised for this session and traffic can be load balanced across these interfaces. An EPE controller that does bandiwdth management for these links should be aware of the links on which the traffic will be load-balanced. As per [RFC8029], the node advertising the EPE SIDs will send Downstream Detailed Mapping TLV (DDMT) specifying the details of nexthop interfaces, the OAM packet will be sent out. Based on this information controller MAY choose to verify the actual forwarding state with the topology information controller has. On the router, the validation procedures will include received DDMT validation as specified in [RFC8029] to verify the control and forwarding state synchronization on the two routers. Any descrepancies between controller's state and forwarding state will not be detected by the procedures described in the document.

4.3. PeerSet SID Sub-TLV

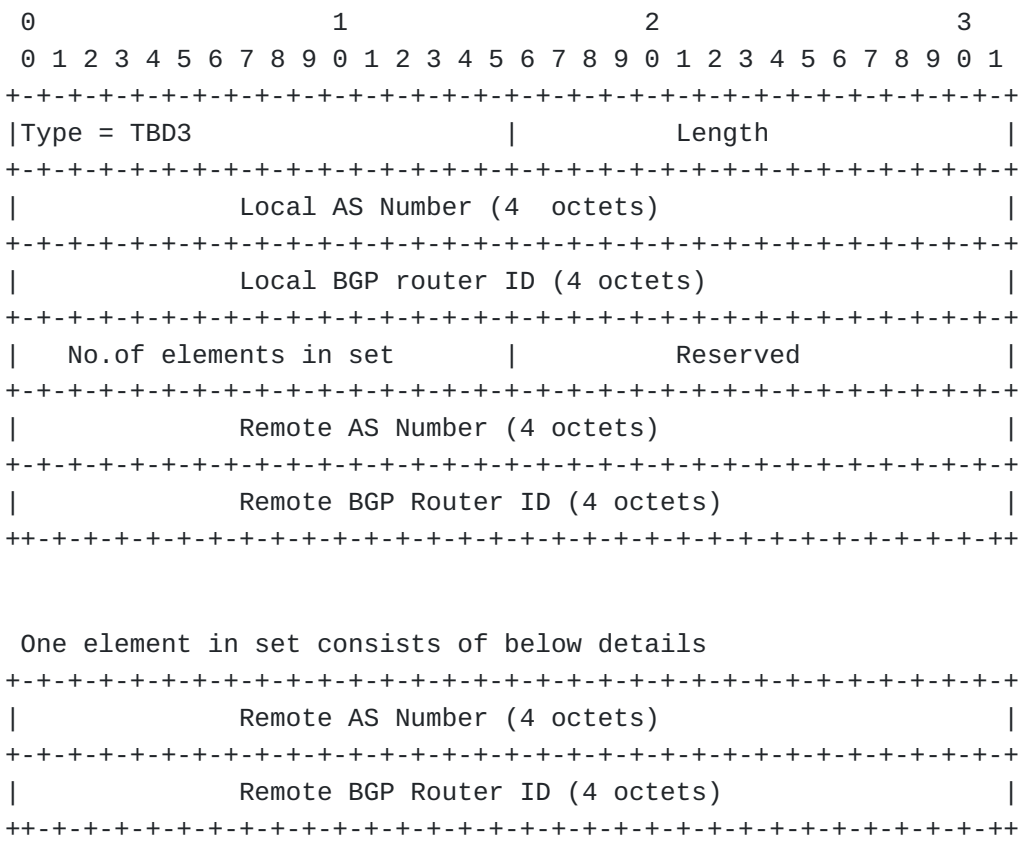

Figure 5: PeerSet SID Sub-TLV

Type : TBD3

Length : variable based on the number of elements in the set. The length field does not include the length of Type and Length fields.

Local AS Number :

4 octet unsigned integer representing the Member-AS Number inside the Confederation.[RFC5065]. The AS number corresponds to the AS to which PeerSet SID advertising node belongs to.

Remote AS Number :

4 octet unsigned integer representing the Member-AS Number inside the Confederation [RFC5065]. The AS number corresponds to the AS of the remote node for which the PeerSet SID is advertised.

Advertising BGP Router ID :

4 octet unsigned integer of the advertising node representing the BGP Identifier as defined in [RFC4271] and [RFC6286].

Receiving BGP Router ID :

4 octet unsigned integer of the receiving node representing the BGP Identifier as defined in [RFC4271] and [RFC6286].

No.of elements in set:

Number of remote ASes, the set SID load-balances on.

PeerSet SID may be associated with a number of PeerNode SIDs and PeerAdj SIDs. The remote AS number and the Router ID of each of these PeerNode SIDs PeerAdj SIDs MUST be included in the FEC.

5. EPE-SID FEC validation

When a remote ASBR of the EPE-SID advertisement receives the MPLS OAM packet with top FEC being the EPE-SID, it SHOULD perform validity checks on the content of the EPE-SID FEC sub-TLV. The basic length check should be performed on the received FEC.

```
PeerAdj SID
-----------
Length = 24 or 48

Peer Node SID
-------------
Length = 20 + No.of IPv4 interface pairs * 8 +
         No.of IPv6 interface pairs * 32

PeerSet SID
-----------
Length = 9 + no.of elements in the set *
         (8 + No.of IPv4 interface pairs * 8 +
          No.of IPv6 interface pairs * 32)
```

Figure 6: Length Validation

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

5.1. EPE-SID FEC validiation

4a. Segment Routing IGP-Prefix, IGP-Adjacency SID and EPE-SID Validation :

If the Label-stack-depth is 0 and the Target FEC Stack sub-TLV at FEC-stack-depth is TBD1 (PeerAdj SID sub-TLV)

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

- o Validate that the Receiving Node BGP Local AS matches with the remote AS field in the received PeerAdj SID FEC sub-TLV.

- o Validate that the Receiving Node BGP Router-ID matches with the Remote Router ID field in the received PeerAdj SID FEC.

- o Validate that there is a EBGP session with a peer having local AS number and BGP Router-ID as specified in the Local AS number and Local Router-ID field in the received PeerAdj SID FEC sub-TLV.

If the Remote interface address is not zero, validate the incoming interface. Set the Best-return-code to 35 "Mapping for this FEC is not associated with the incoming interface" [RFC8287] if any below conditions fail:

- o Validate the incoming interface on which the OAM packet was receieved, matches with the remote interface specified in the PeerAdj SID FEC sub-TLV

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

Else, if the Target FEC sub-TLV at FEC-stack-depth is TBD2 (PeerNode SID sub-TLV),

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

- o Validate that the Receiving Node BGP Local AS matches with the remote AS field in the received PeerNode SID FEC sub-TLV.

- o Validate that the Receiving Node BGP Router-ID matches with the Remote Router ID field in the received PeerNode SID FEC.

- o Validate that there is a EBGP session with a peer having local AS number and BGP Router-ID as specified in the Local AS number and Local Router-ID field in the received PeerNode SID FEC sub-TLV.

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

Else, if the Target FEC sub-TLV at FEC-stack-depth is TBD3 (PeerSet SID sub-TLV),

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

- o Validate that the Receiving Node BGP Local AS matches with one of the remote AS field in the received PeerSet SID FEC sub-TLV.

- o Validate that the Receiving Node BGP Router-ID matches with one of the Remote Router ID field in the received PeerSet SID FEC sub-TLV.

- o Validate that there is a EBGP session with a peer having local AS number and BGP Router-ID as specified in the Local AS number and Local Router-ID field in the received PeerSet SID FEC sub-TLV.

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

6. IANA Considerations

IANA is requested to allocated three new Target FEC stack sub-TLVs from the "Sub-TLVs for TLV types 1,16 and 21" subregistry in the "TLVs" registry of the "Multi-Protocol Label switching (MPLS) Label Switched Paths (LSPs) Ping parameters" namespace.

PeerAdj SID Sub-TLV : TBD1

PeerNode SID Sub-TLV: TBD2

PeerSet SID Sub-TLV : TBD3

The three lowest free values from the Standard Tracks range should be allocated if possible.

7. Security Considerations

The EPE-SIDs are advertised for egress links for Egress Peer Engineering purposes or for inter-AS links between co-operating ASes. When co-operating domains are involved, they can allow the

packets arriving on trusted interfaces to reach the control plane and get processed. When EPE-SIDs which are created for egress TE links where the neighbor AS is an independent entity, it may not allow packets arriving from external world to reach the control plane. In such deployments MPLS OAM packets will be dropped by the neighboring AS that receives the MPLS OAM packet. In MPLS traceroute applications, when the AS boundary is crossed with the EPE-SIDs, the FEC stack is changed. [RFC8287] does not mandate that the initiator upon receiving an MPLS Echo Reply message that includes the FEC Stack Change TLV with one or more of the original segments being popped remove a corresponding FEC(s) from the Target FEC Stack TLV in the next (TTL+1) traceroute request. If an initiator does not remove the FECs belonging to the previous AS that has traversed, it MAY expose the internal AS information to the following AS being traversed in traceroute.

8. Acknowledgments

Thanks to Loa Andersson, Dhruv Dhody, Ketan Talaulikar, Italo Busi and Alexander Vainshtein, Deepti Rathi for careful review and comments.

9. References

9.1. 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>.

[RFC8690] Nainar, N., Pignataro, C., Iqbal, F., and A. Vainshtein, "Clarification of Segment ID Sub-TLV Length for RFC

8287", RFC 8690, DOI 10.17487/RFC8690, December 2019, <https://www.rfc-editor.org/info/rfc8690>.

[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>.

9.2. Informative References

[I-D.ietf-idr-segment-routing-te-policy]
Previdi, S., Filsfils, C., Talaulikar, K., Mattes, P., Jain, D., and S. Lin, "Advertising Segment Routing Policies in BGP", Work in Progress, Internet-Draft, draft-ietf-idr-segment-routing-te-policy-20, 27 July 2022, <https://www.ietf.org/archive/id/draft-ietf-idr-segment-routing-te-policy-20.txt>.

[RFC4271] Rekhter, Y., Ed., Li, T., Ed., and S. Hares, Ed., "A Border Gateway Protocol 4 (BGP-4)", RFC 4271, DOI 10.17487/RFC4271, January 2006, <https://www.rfc-editor.org/info/rfc4271>.

[RFC5065] Traina, P., McPherson, D., and J. Scudder, "Autonomous System Confederations for BGP", RFC 5065, DOI 10.17487/RFC5065, August 2007, <https://www.rfc-editor.org/info/rfc5065>.

[RFC6286] Chen, E. and J. Yuan, "Autonomous-System-Wide Unique BGP Identifier for BGP-4", RFC 6286, DOI 10.17487/RFC6286, June 2011, <https://www.rfc-editor.org/info/rfc6286>.

[RFC7705] George, W. and S. Amante, "Autonomous System Migration Mechanisms and Their Effects on the BGP AS_PATH Attribute", RFC 7705, DOI 10.17487/RFC7705, November 2015, <https://www.rfc-editor.org/info/rfc7705>.

[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>.

[RFC8664] Sivabalan, S., Filsfils, C., Tantsura, J., Henderickx, W., and J. Hardwick, "Path Computation Element Communication Protocol (PCEP) Extensions for Segment

Routing", RFC 8664, DOI 10.17487/RFC8664, December 2019, <https://www.rfc-editor.org/info/rfc8664>.

[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>.

Authors' Addresses

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

Email: shraddha@juniper.net

Mukul Srivastava
Juniper Networks Inc.

Email: msri@juniper.net

Kapil Arora
Individual Contributor

Email: kapil.it@gmail.com

Samson Ninan
Individual Contributor

Email: samson.cse@gmail.com

Xiaohu Xu
China Mobile
Beijing
China

Email: xuxiaohu@cmss.chinamobile.com