SPRING
Internet-Draft
Intended status: Standards Track
Expires: December 4, 2020

S. Hegde
S. Sangli
M. Srivastava
Juniper Networks Inc.
X. Xu
Alibaba Inc.

June 2, 2020

BGP-LS Extensions for Inter-AS TE using EPE based mechanisms draft-hegde-idr-bgp-ls-epe-inter-as-03

Abstract

In certain network deployments, a single operator has multiple Autonomous Systems(AS) to facilitate ease of management. A multiple AS network design could also be a result of network mergers and acquisitions. In such scenarios, a centralized Inter-domain TE approach could provide most optimal allocation of resources and the most controlled path placement. BGP-LS-EPE [I-D.ietf-idr-bgpls-segment-routing-epe] describes an extension to BGP Link State (BGP-LS) for the advertisement of BGP Peering Segments along with their BGP peering node and inter-AS link information, so that efficient BGP Egress Peer Engineering (EPE) policies and strategies can be computed based on Segment Routing. This document describes extensions to the BGP-LS EPE to enable it to be used for inter-AS Traffic-Engineering (TE) purposes.

Requirements Language

The key words "MUST", "MUST NOT", "REQUIRED", "SHALL", "SHALL NOT", "SHOULD", "SHOULD NOT", "RECOMMENDED", "MAY", and "OPTIONAL" in this document are to be interpreted as described in RFC 2119 [RFC2119].

Status of This Memo

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

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

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

This Internet-Draft will expire on December 4, 2020.

Table of Contents

1. Introduction

Segment Routing (SR) leverages source routing. A node steers a packet through a controlled set of instructions, called segments, by prepending the packet with an SR header with segment identifiers (SID). A SID can represent any instruction, topological or service-based. SR segments allows to enforce a flow through any topological path or service function while maintaining per-flow state only at the ingress node of the SR domain.

As there is no per-path state in the network, the bandwidth management for the paths is expected to be handled by a centralized entity which has a complete view of:

1. Up-to-date topology of the network

2. Resources, States and Attributes of links and nodes of the network

3. Run-time utilization/availability of resources

The BGP Link-State extensions provide mechanisms whereby link-state and TE information can be propagated n a network and a consumer of such BGP LS updates may build topology, provide bandwidth calendering and other traffic engineering services. The centralized entity can be such a consumer (also referred to as controller). In the case of multi-AS networks, the controller needs to learn the per-AS network information and the inter-AS link information thus arriving at a consolidated Traffic Engineering Database which can be used to compute end-to-end Traffic Engineering Path. The controller can learn the topology, link-state and TE information from each of the AS networks either by participating in their IGPs or by listening to BGP LS updates [RFC7752]. Similar information about the inter-AS links can be learnt via BGP-LS EPE [I-D.ietf-idr-bgpls-segment-routing-epe] along with extensions defined in this documet.

2. Reference Topology

The controller learns TE attributes of all the links, including the inter-AS links and uses the attributes to compute constrained paths. The controller should be able to correlate the inter-AS links for bidirectional connectivity from both ASes.

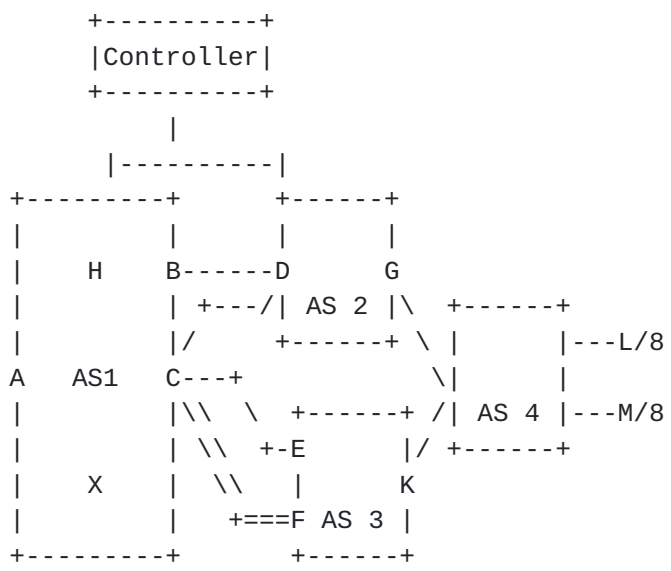


Figure 1: Reference Diagram

The reference diagram from [I-D.ietf-spring-segment-routing-central-epe] represents multiple Autonomous Systems connected to each other. When the Multiple ASes belong to same operator and are organised into separate domains for operational purposes, it is advantageous to support Traffic-Engineering across the ASes including the inter-AS links. The controller has visibility of all of the ASes by means of IGP topology exported via BGP-LS [RFC7752], or other means. In addition, the inter-AS links and the labels associated with the inter-AS links are exported via [I-D.ietf-idr-bgpls-segment-routing-epe]. The controller needs to correlate the information acquired from all of the ASes, including the inter-AS links in order to get a view of the unified topology so that it can build end-to-end Traffic-Engineered paths.

3. Fast Reroute Label

[I-D.ietf-spring-segment-routing-central-epe] section 3.6 describes mechanisms to provide Fast Reroute (FRR) protection for the EPE Labels. The BGP-LS EPE [I-D.ietf-idr-bgpls-segment-routing-epe] describes "B" bit to indicate that a PeerNodeSID or PeerAdjSID is eligible for backup. However, it does not specify what is the behaviour when the failure kicks-in. The controller needs to know which links are used for protection so that admission control and failure simulation can be done effectively and appropriate inter-AS links used for path construction.

This document defines a new flag "F" in the Peering SIDs TLV to indicate a SID as an FRR SID. With the "F" flag set, the protection for any peering SID can be specified using another PeerAdjSID, PeerNodeSID or PeerSetSID to the controller. If the protection is achieved by fallback to local IP lookup, FRR SID SHOULD not be advertised. The link(s) represented by the FRR SID will carry the traffic when there is a failure. These SIDs are included as an FRR SIDs in the peerAdjSID, PeerNodeSID and PeerSetSID advertisements.

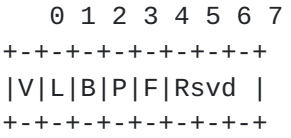


Figure 2: Peering SID TLV Flags Format

* F-Flag: FRR Label Flag: If set, the peer SID where the FRR Label appears is using backup links represented by FRR Label.

4. TE Link attributes of PeerNode SID

In any eBGP deployment, the peering session can be either single-hop of multi-hop. For single-hop eBGP sessions, the peering address is that of the directly attached interface to which the session is pinned down. For multi-hop eBGP session, the peering adress is reachable over more than one interface and that the peering session is not pinned down to any of the directly attached interfaces.

A Peer Node Segment is a segment describing a peer, including the SID (PeerNodeSID) allocated to it. The link descriptors for the PeerNodeSID include the addresses used by the BGP session encoded using TLVs as defined in [RFC7752]. Since the eBGP session can be either single-hop or multi-hop, the IP address used by BGP session as local/neigbour is not sufficient to identify the underlaying interface(s). Also, the controller needs to know the links associated with the PeerNodeSID, to be able derive TE link attributes. This can be achieved by including the interface local and remote addresses in the Link attributes in PeerNodeSID NLRI. This document defines a new link attribute TLV name Interface Address TLV. PeerNodeSID NLRI MAY optionally include Interface Address TLV.

5. TE Link attributes of PeerAdj SID

PeerAdjSID MUST be advertised for each inter-AS link for the purposes of inter-AS TE. The PeerAdjSID should contain link TE attributes such as bandwidth, admin-group etc. The PeerAdjSID should also

contain the local and remote interface IPv4/IPv6 addresses which is used for correlating the links. PeerNodeSID SHOULD contain the additional attribute of link local address which is used by the controller to find corresponding PeerAdjSID and hence the corresponding link TE attributes.

A peerAdj segment carries mandatory link descriptors as local and remote link id. Remote link id of the neighboring ASBR is not readily available. [I-D.ietf-idr-bgpls-segment-routing-epe] suggests to carry the value '0' for the remote link id. The Controller needs to associate the links in both directions to effectively handle failure notifications and for this purpose a unque remote link is necessary. The remote link ID cannot be manually configured on the router as the link-ids generally change over router reboot etc and hence manual configuration is operationally very difficult to manage. This document mandates advertisement of local and remote iterface addresses for the inetr-AS TE purposes.

The Unnumbered interface is not in the scope of this document.

6. Link address TLV

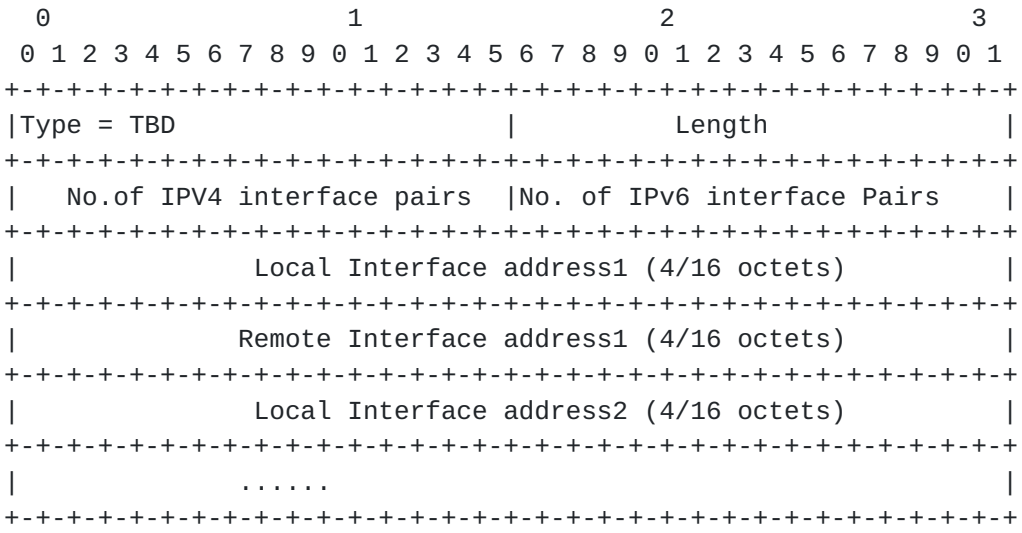


Figure 3: Link Address TLV carried as attribute

Type : TBD

Length : variable based on ipv4/ipv6 interface address

Number of IPv4 interface pairs:

Number of IPv6 interface pairs:

There may be a number of parallel interfaces and few or all of them may be used for the PeerNodeSID. These interfaces may have both IPV4 and IPV6 address or some interfaces may be IPv4 only and some IPv6 only. The total number of IPv4 and IPv6 interface address count is carried seperately in above fields.

Local Interface Address :

The interface local address ipv4/ipv6 which corresponds to the PeerNodeSID MUST be specified. For IPv4,this field is 4 octets; for IPv6, this field is 16 octets.

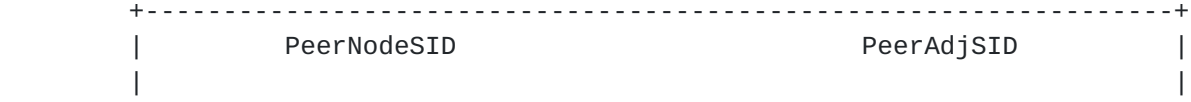
Remote Interface Address :

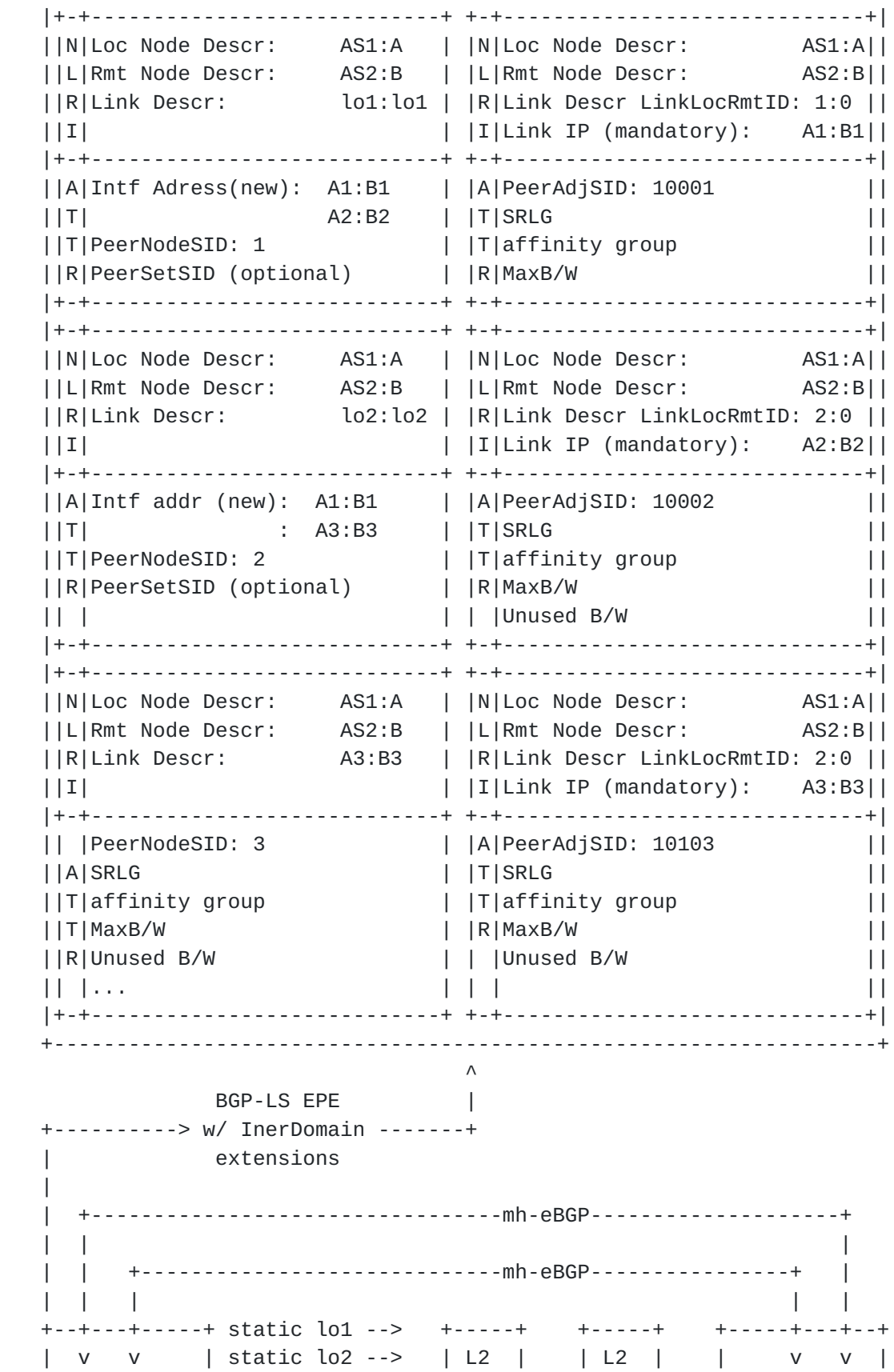
The interface remote address ipv4/ipv6 which corresponds to the PeerNodeSID MUST be specified. For IPv4,this field is 4 octets; for IPv6, this field is 16 octets.

There can be multiple Layer 3 interfaces on which a peerNodeSId loadbalances the traffic. All such interfaces local/remote address MUST be included when a Link address TLV is added.When a single Layer 3 interface consists of multiple addresses or when a link has both IPv4 and IPv6 addresses configured, It is sufficient to include one such pair (either IPV4 or IPv6)address for the PeerNodeSID advertisement. When a PeerNodeSID load-balances over few interfaces with IPv4 only address and few interfaces with IPv6 address then the Link address TLV should list all IPv4 address pairs together followed by IPv6 address pairs.

7. Example Advertisements

The below diagram represents two ASBR routers and inter-AS links between them. The inter-AS links could be connected via switches L1 and L2 as shown in the diagram or via Point-to-point links A2->B2, A3->B3 as shown in the diagram below. In the below example, lets assume peerNodeSID 1 is configured to use peerAdjSID 10002 then PeerNodeSID 1 will have the B bit set which means the PeerNodeSID 1 is eligible for backup. Label 10002 is added to the PeerNodeSID with a "F" bit set, which means 10002 is a backup for PeerNodeSID 1.




|N|Loc Node Descr: AS1:A
|L|Rmt Node Descr: AS2:B
|R|Link Descr: lo1:lo1
|I|
|A|Intf Adress(new): A1:B1
|T| A2:B2
|T|PeerNodeSID: 1
|R|PeerSetSID (optional)
|N|Loc Node Descr: AS1:A
|L|Rmt Node Descr: AS2:B
|R|Link Descr LinkLocRmtID: 1:0
|I|Link IP (mandatory): A1:B1
|A|PeerAdjSID: 10001
|T|SRLG
|T|affinity group
|R|MaxB/W
|N|Loc Node Descr: AS1:A
|L|Rmt Node Descr: AS2:B
|R|Link Descr: lo2:lo2
|I|
|A|Intf addr (new): A1:B1
|T| : A3:B3
|T|PeerNodeSID: 2
|R|PeerSetSID (optional)
|N|Loc Node Descr: AS1:A
|L|Rmt Node Descr: AS2:B
|R|Link Descr LinkLocRmtID: 2:0
|I|Link IP (mandatory): A2:B2
|A|PeerAdjSID: 10002
|T|SRLG
|T|affinity group
|R|MaxB/W
| |Unused B/W
|N|Loc Node Descr: AS1:A
|L|Rmt Node Descr: AS2:B
|R|Link Descr: A3:B3
|I|
|PeerNodeSID: 3
|A|SRLG
|T|affinity group
|T|MaxB/W
|R|Unused B/W
|...
|N|Loc Node Descr: AS1:A
|L|Rmt Node Descr: AS2:B
|R|Link Descr LinkLocRmtID: 2:0
|I|Link IP (mandatory): A3:B3
|A|PeerAdjSID: 10103
|T|SRLG
|T|affinity group
|R|MaxB/W
| |Unused B/W
BGP-LS EPE
w/ InerDomain
extensions
mh-eBGP
mh-eBGP
static lo1 -->
static lo2 -->
L2
L2

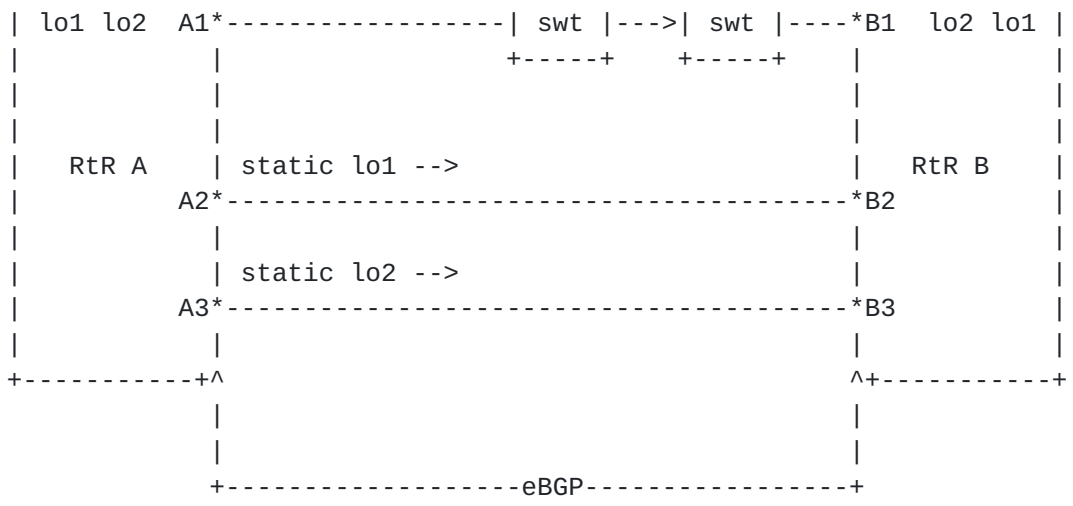


Figure 4: Example Advertisements

8. Backward Compatibility

The extension proposed in this document is backward compatible with procedures described in [I-D.ietf-idr-bgpls-segment-routing-epe] and [I-D.ietf-spring-segment-routing-central-epe]

9. Security Considerations

TBD

10. IANA Considerations

New attribute TLV in BGP-LS Node Descriptor, Link Descriptor, Prefix Descriptor, and Attribute TLVs registry

TLV Code Point	Description	IS-IS TLV /Sub-TLV	Reference (RFC/Section)
TBD	Link address TLV	NA	This draft

Figure 5: IANA code point

11. Acknowledgements

Thanks to Julian Lucek and Rafal Jan Szarecki for careful review and suggestions.

12. References

12.1. Normative References

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

[I-D.ietf-spring-segment-routing-central-epe]
Filsfils, C., Previdi, S., Dawra, G., Aries, E., and D. Afanasiev, "Segment Routing Centralized BGP Egress Peer Engineering", draft-ietf-spring-segment-routing-central-epe-10 (work in progress), December 2017.

[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>.

[RFC7752] Gredler, H., Ed., Medved, J., Previdi, S., Farrel, A., and S. Ray, "North-Bound Distribution of Link-State and Traffic Engineering (TE) Information Using BGP", RFC 7752, DOI 10.17487/RFC7752, March 2016, <https://www.rfc-editor.org/info/rfc7752>.

12.2. Informative References

[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>.

Authors' Addresses

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

Email: shraddha@juniper.net

Srihari Sangli
Juniper Networks Inc.
Exora Business Park
Bangalore, KA 560103
India

Email: ssangli@juniper.net

Mukul Srivastava
Juniper Networks Inc.

Email: msri@juniper.net

Xiaohu Xu
Alibaba Inc.
Beijing
China

Email: xiaohu.xxh@alibaba-inc.com