IDR Working Group
Internet Draft
Intended status: Standards Track
Expires: September 13, 2024

C. Lin
New H3C Technologies
Z. Li
China Mobile
R. Pang
China Unicom
K. Talaulikar
Cisco Systems
M. Chen
New H3C Technologies

March 17, 2024

Segment Routing BGP Egress Peer Engineering over Layer 2 Bundle
draft-lin-idr-sr-epe-over-l2bundle-05

Abstract

There are deployments where the Layer 3 interface on which a BGP peer session is established is a Layer 2 interface bundle. In order to allow BGP-EPE to control traffic flows on individual member links of the underlying Layer 2 bundle, BGP Peering SIDs need to be allocated to individual bundle member links, and advertisement of such BGP Peering SIDs in BGP-LS is required. This document describes how to support Segment Routing BGP Egress Peer Engineering over Layer 2 bundle.

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 September 13, 2024.

Table of Contents

1. Introduction

Segment Routing (SR) leverages the source routing paradigm. A node steers a packet through an ordered list of instructions called "segments". Segment Routing can be instantiated on both MPLS and IPv6 data planes, which are referred to as SR-MPLS and SRv6.

BGP Egress Peer Engineering (BGP-EPE) allows an ingress Provider Edge (PE) router within the domain to use a specific egress PE and a specific external interface/neighbor to reach a particular destination.

The SR architecture [RFC8402] defines three types of BGP Peering Segments that may be instantiated at a BGP node:

- Peer Node Segment (PeerNode SID): instruction to steer to a specific peer node

- Peer Adjacency Segment (PeerAdj SID): instruction to steer over a specific local interface towards a specific peer node

- o Peer Set Segment (PeerSet SID): instruction to load-balance to a set of specific peer nodes

[RFC9087] illustrates a centralized controller-based BGP-EPE solution involving SR path computation using the BGP Peering Segments. A centralized controller learns the BGP Peering SIDs via Border Gateway Protocol - Link State (BGP-LS) and then uses this information to program a BGP-EPE policy. [RFC9086] defines the extension to BGP-LS for advertisement of BGP Peering Segments along with their BGP peering node information.

There are deployments where the Layer 3 interface on which a BGP peer session is established is a Layer 2 interface bundle (L2 Bundle), for instance, a Link Aggregation Group (LAG) [IEEE802.1AX]. BGP-EPE may wish to control traffic flows on individual member links of the underlying Layer 2 bundle. In order to do so, BGP Peering SIDs need to be allocated to individual bundle member links, and advertisement of such BGP Peering SIDs in BGP-LS is required.

This document describes how to support Segment Routing BGP Egress Peer Engineering over Layer 2 bundle.

1.1. Requirements Language

The key words "MUST", "MUST NOT", "REQUIRED", "SHALL", "SHALL NOT", "SHOULD", "SHOULD NOT", "RECOMMENDED", "NOT RECOMMENDED", "MAY", and "OPTIONAL" in this document are to be interpreted as described in BCP 14 [RFC2119] [RFC8174] when, and only when, they appear in all capitals, as shown here.

2. Problem Statement

In the network depicted in Figure 1, B and C establish BGP peer session on a Layer 2 bundle. Assume that, the link delays of the members are different because they are over different transport paths, and member link 1 has the lowest delay.

The operator of AS1 wishes to apply a BGP-EPE policy to steer the time-sensitive traffic from AS1 to AS2 via member link 1 of the Layer 2 bundle.

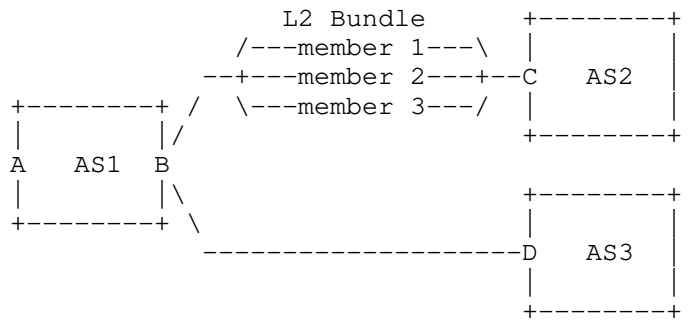

Figure 1: BGP-EPE over L2 Bundle

The existing Peer Adjacency SID can be allocated to the Layer 3 interface between B and C, which is a Layer 2 interface bundle. If steered by that Peer Adjacency SID, the traffic will be forwarded by load balancing among all the bundle member links. So, the existing mechanism cannot meet the requirement of steering traffic flows via individual member link.

3. Advertising Peer Adjacency Segment for L2 Bundle Member in BGP-LS

BGP peering segments are generally advertised in BGP-LS from a BGP node along with its peering topology information, in order to enable computation of efficient BGP-EPE policies and strategies.

When a BGP peer session is established over a Layer 2 interface bundle, an implementation MAY allocate one or more Peer Adjacency Segments for each member link. If so, it SHOULD advertise the Peer Adjacency Segments of bundle members in BGP-LS, using the method defined in this section.

3.1. MPLS-SR

For SR-MPLS, Section 5.2 of [RFC9086] described the BGP-LS advertisement of the PeerAdj SID for L3 link.

In order to advertise the PeerAdj SIDs for L2 bundle members in BGP-LS, the L2 Bundle Member Attributes TLVs [RFC9085] MUST also be included in the Link Attributes. Each L2 Bundle Member Attributes TLV identifies an L2 bundle member, and includes the PeerAdj SID TLV [RFC9086] to advertise the PeerAdj SID for the associated L2 bundle member.

This document updates [RFC9085] and [RFC9086] to allow the PeerAdj SID TLV to be included as a sub-TLV of the L2 Bundle Member Attributes TLV.

Note that the inclusion of a L2 Bundle Member Attributes TLV implies that the identified link is a member of the L2 bundle and that the member link is operationally up. If any member link fails, an implementation MUST withdraw the L2 Bundle Member Attributes TLV in BGP-LS, along with the Peer Adjacency Segments for the failed member link.

3.2. SRv6

For SRv6, according to Section 4.1 of [RFC9514], the advertisement of L3 link BGP EPE Peer Adjacency SID is the same as for SR-MPLS, except for using the SRv6 End.X SID TLV [RFC9514] instead of the PeerAdj SID TLV [RFC9086].

Similarly, when advertising the SRv6 BGP Peer Adjacency SIDs for L2 bundle members, the L2 Bundle Member Attributes TLVs [RFC9085] MUST also be included in the Link Attributes. The SRv6 End.X SID TLV [RFC9514] MUST be carried in the L2 Bundle Member Attributes TLV to advertise the SRv6 Peer Adjacency SID for the associated L2 bundle member.

4. Manageability Considerations

The manageability considerations described in [RFC9552] and [RFC9086] also apply to this document.

The operator MUST be provided with the options of configuring, enabling, and disabling the advertisement of Peer Adjacency Segment for L2 Bundle member links, as well as control of which information is advertised to which internal or external peer.

5. Security Considerations

The security considerations described in [RFC9552] and [RFC9086] also apply to this document.

This document does not introduce any new security consideration.

6. IANA Considerations

This document has no IANA actions.

7. References

7.1. 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>.

[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>.

[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>.

[RFC9085] Previdi, S., Talaulikar, K., Ed., Filsfils, C., Gredler, H., and M. Chen, "Border Gateway Protocol - Link State (BGP-LS) Extensions for Segment Routing", RFC 9085, DOI 10.17487/RFC9085, August 2021, <https://www.rfc-editor.org/info/rfc9085>.

[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>.

[RFC9514] Dawra, G., Filsfils, C., Talaulikar, K., Ed., Chen, M., Bernier, D., and B. Decraene, "Border Gateway Protocol - Link State (BGP-LS) Extensions for Segment Routing over IPv6 (SRv6)", RFC 9514, DOI 10.17487/RFC9514, December 2023, <https://www.rfc-editor.org/info/rfc9514>.

[RFC9552] K. Talaulikar, "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>.

7.2. Informative References

[IEEE802.1AX] IEEE, "IEEE Standard for Local and metropolitan area networks -- Link Aggregation", IEEE 802.1AX, <https://ieeexplore.ieee.org/document/7055197>.

[RFC8668] Ginsberg, L., Ed., Bashandy, A., Filsfils, C., Nanduri, M., and E. Aries, "Advertising Layer 2 Bundle Member Link Attributes in IS-IS", RFC 8668, DOI 10.17487/RFC8668, December 2019, <https://www.rfc-editor.org/info/rfc8668>.

[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>.

Appendix A. Example

This section shows an example of how Node B in Figure 1 allocates and advertises Peer Adjacency Segments for L2 bundle members.

B allocates a PeerAdj SID for the Layer 2 interface bundle to peer C, along with a PeerAdj SID for each member link. B programs its forwarding table accordingly:

PeerAdj SID		Outgoing Interface
IF on SR-MPLS Data Plane	IF on SRv6 Data Plane	
1010	A::A0	L2 Bundle to C
1011	A::A1	Member link 1 to C
1012	A::A2	Member link 2 to C
1013	A::A3	Member link 3 to C

B signals the related BGP-LS Link NLRI and Link Attributes including the PeerAdj SID for L3 parent link to the BGP-EPE controller, as specified in Section 5.2 of [RFC9086]. In addition, B also advertises L2 Bundle Member Attribute TLVs carrying the PeerAdj SIDs for L2 bundle members.

For MPLS-SR, the Link Attributes are as follows:

- PeerAdj SID TLV (Label-1010)
- L2 Bundle Member Attribute TLV (Link Local Identifier describing the member link 1)
  - PeerAdj SID TLV (Label-1011)

    * (Optional) Min/Max Unidirectional Link Delay TLV (Delay of member link 1)

- L2 Bundle Member Attribute TLV (Link Local Identifier describing the member link 2)

    * PeerAdj SID TLV (Label-1012)

    * (Optional) Min/Max Unidirectional Link Delay TLV (Delay of member link 2)

- L2 Bundle Member Attribute TLV (Link Local Identifier describing the member link 3)

    * PeerAdj SID TLV (Label-1013)

    * (Optional) Min/Max Unidirectional Link Delay TLV (Delay of member link 3)

For SRv6, the Link Attributes are as follows:

- SRv6 End.X SID TLV (SID-A::A0)

- L2 Bundle Member Attribute TLV (Link Local Identifier describing the member link 1)

    * SRv6 End.X SID TLV (SID-A::A1)

    * (Optional) Min/Max Unidirectional Link Delay TLV (Delay of member link 1)

- L2 Bundle Member Attribute TLV (Link Local Identifier describing the member link 2)

    * SRv6 End.X SID TLV (SID-A::A2)

    * (Optional) Min/Max Unidirectional Link Delay TLV (Delay of member link 2)

- L2 Bundle Member Attribute TLV (Link Local Identifier describing the member link 3)

    * SRv6 End.X SID TLV (SID-A::A3)

    * (Optional) Min/Max Unidirectional Link Delay TLV (Delay of member link 3)

Acknowledgements

The authors would like to thank Sasha Vainshtein for his review and comments of this document.

Authors' Addresses

Changwang Lin
New H3C Technologies
China
Email: linchangwang.04414@h3c.com

Zhenqiang Li
China Mobile
China
Email: lizhenqiang@chinamobile.com

Ran Pang
China Unicom
China
Email: pangran@chinaunicom.cn

Ketan Talaulikar
Cisco Systems
India
Email: ketant.ietf@gmail.com

Mengxiao Chen
New H3C Technologies
China
Email: chen.mengxiao@h3c.com