Network Working Group
Internet Draft
Intended status: Standards Track
Expires: September 13, 2023

C. Lin
M. Chen
H. Li
New H3C Technologies

March 13, 2023

Segment Routing BGP Egress Peer Engineering over Layer 2 Bundle
draft-lin-idr-sr-epe-over-l2bundle-01

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 also 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, 2023.

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 MPLS-SR 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

- 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 also 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.

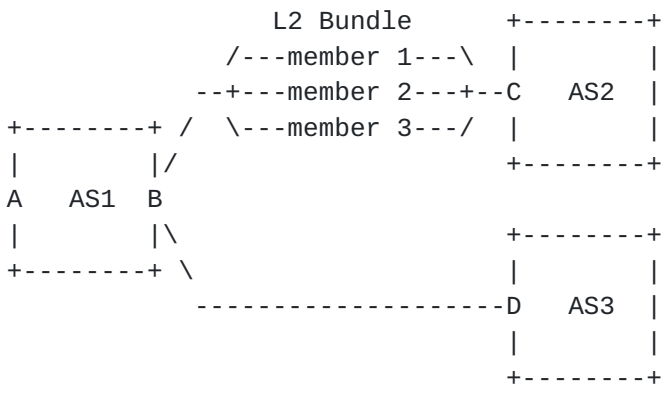

Figure 1: BGP-EPE over L2 Bundle

The existing PeerAdj SID can be allocated to the Layer 3 interface between B and C, which is a Layer 2 interface bundle. If steered by that PeerAdj 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. Peer Adjacency Segment for L2 Bundle Member Link

This document extends Peer Adjacency Segments to be allocated to the individual member links of the Layer 2 interface bundle.

The semantics of a Peer Adjacency Segment for L2 Bundle Member Link are:

o SR operation: NEXT.

o Next-Hop: forwarding across the bundle member link, which the segment is associated with, to the peer connected through the parent L3 interface.

On the MPLS-SR data plane, Peer Adjacency Segments for L2 Bundle Member Links are instantiated as MPLS labels. On the SRv6 data plane, Peer Adjacency Segments for L2 Bunlde Members are instantiated as SRv6 End.X SIDs.

3.1. Advertising 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 advertising Peer Adjacency Segments for L2 Bunlde Member Links in BGP-LS, they can be carried in the sub-TLVs of L2 Bundle Member

Attributes TLVs [RFC9085]. Derived from IS-IS advertisements of L2 Bundle [RFC8668], each L2 Bundle Member Attributes TLV identifies an L2 Bundle Member link, which in turn is associated with a parent L3 link. The L3 link is described by the Link NLRI, and the L2 Bundle Member Attributes TLV is associated with the Link NLRI. The L2 Bundle Member Attributes TLV MAY include sub-TLVs that describe attributes associated with the bundle member.

In order to advertise Peer Adjacency Segments for L2 Bundle Member Links in BGP-LS, a BGP-LS Link NLRI is advertised to describe the parent L3 link to the BGP peer, which is similar with advertising a PeerAdj SID for the parent L3 link as specified in Section 5.2 of [RFC9086]. Then, multiple L2 Bundle Member Attributes TLVs are included in the Link Attribute TLVs associated with the parent L3 link, each identifying an L2 Bundle member link. The L2 Bundle Member Attributes TLV MAY carry the following sub-TLVs:

- PeerAdj SID TLV (Type-1102) [RFC9086] contains a Peer Adjacency Segment for the associated L2 Bundle member link on MPLS-SR data plane.

- SRv6 End.X SID TLV (Type-1106) [I-D.ietf-idr-bgpls-srv6-ext] contains a Peer Adjacency Segment for the associated L2 Bundle member link on SRv6 data plane.

- Other BGP-LS Attribute TLVs used to describe the associated L2 Bundle member link [RFC9085], such as administrative group (color), link bandwidth, and link delay.

3.2. Example

Take the allocating and advertising of Peer Adjacency Segments on Node B in Figure 1 as an example.

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 MPLS-SR 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 NLRI to the BGP-EPE controller, which is described in the following.

Descriptors:

o Local Node Descriptors (B's router-ID, ASN-AS1)

o Remote Node Descriptors (C's router-ID, ASN-AS2)

o Link Descriptors (Link Local/Remote Identifiers describing the Layer 2 bundle, IPv4 or IPv6 Interface Address, IPv4 or IPv6 Neighbor Address)

Attributes (MPLS-SR data plane):

o PeerAdj SID TLV (Label-1010)

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

  * PeerAdj SID TLV (Label-1011)

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

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

  * PeerAdj SID TLV (Label-1012)

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

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

* PeerAdj SID TLV (Label-1013)

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

Attributes (SRv6 data plane):

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

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

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

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

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

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

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

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

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

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

4. Considerations of Defining New Types of BGP Peering Segments

The solution in Section 3 reuses the existing Peer Adjacency Segments. It may also be possible to define a new type of BGP Peering Segments for L2 Bundle member links. Then, a new type of BGP Peering SIDs TLV is required, having the same format as described in Figure 2 of [RFC9086].

One of the main advantages of reusing Peer Adjacency Segments is that, no new BGP-LS TLVs is introduced. This mechanism is similar with the BGP-LS advertisements of IGP Adj-SIDs for an L2 Bundle and its member links [RFC9085] [I-D.ietf-idr-bgpls-srv6-ext].

5. Security Considerations

TBD

6. IANA Considerations

TBD

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>.

[I-D.ietf-idr-bgpls-srv6-ext] Dawra, G., Filsfils, C., Talaulikar, K., Chen, M., Bernier, D., and B. Decraene, "BGP Link State Extensions for SRv6", Work in Progress, Internet-Draft, draft-ietf-idr-bgpls-srv6-ext-14, 17 February 2023, <https://www.ietf.org/archive/id/draft-ietf-idr-bgpls-srv6-ext-14.txt>.

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>.

Authors' Addresses

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

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

Hao Li
New H3C Technologies
China
Email: lihao@h3c.com