IDR Working Group
Internet-Draft
Updates: 9085, 9086 (if approved)
Intended status: Standards Track
Expires: 31 January 2026

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

30 July 2025

Members draft-ietf-idr-bgp-ls-sr-epe-over-l2bundle-00

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 members. This document updates RFC9085 to allow the L2 Bundle Member Attributes TLV to be added to the BGP-LS Attribute associated with the Link NLRI of BGP peering link. 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.

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

* 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 members.

This document updates [RFC9085] to allow the L2 Bundle Member Attributes TLV to be added to the BGP-LS Attribute associated with the Link NLRI of BGP peering link. 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.

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 member link 1 has the largest available bandwidth. The operator of AS1 wishes to apply a BGP-EPE policy to steer certain flows from AS1 to AS2 via member link 1 of the Layer 2 bundle to ensure there is no over- subscription.

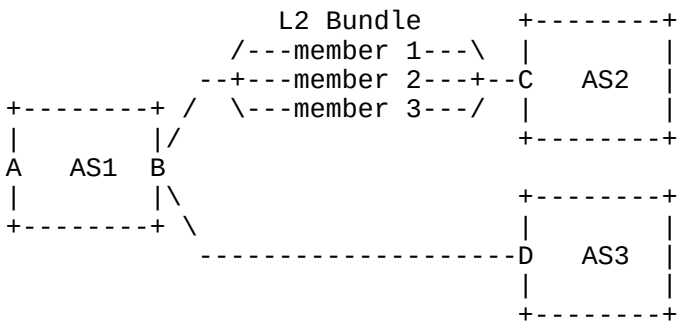

Figure 1: BGP-EPE over L2 Bundle

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.

In order to support BGP Egress Peer Engineering over Layer 2 bundle members, a BGP router needs to have the ability to assign Peer Adjacency Segments for member links. And, the Peer Adjacency Segments of bundle members need to be advertised in BGP-LS, which will be specified in this document.

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 BGP-EPE policies.

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.

In order to advertise the EPE Peer Adjacency SIDs for L2 bundle members in BGP-LS, the L2 Bundle Member Attributes TLVs [RFC9085] MUST also be included in the Link Attributes for the BGP-LS Link NLRI corresponding to the BGP peering session.

Section 2.2 of [RFC9085] restricted that the L2 Bundle Member Attributes TLV "should only be added to the BGP-LS Attribute associated with the Link NLRI that describes the link of the IGP node". This document updates [RFC9085] to allow the L2 Bundle Member Attributes TLV to be added to the BGP-LS Attribute associated with the Link NLRI of BGP peering link.

Each L2 Bundle Member Attributes TLV identifies an L2 bundle member, and includes the EPE Peer Adjacency SID for the associated L2 bundle member.

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.

3.1. SR-MPLS

For SR-MPLS, Section 5 of [RFC9086] defined the PeerAdj SID TLV and its usage for the BGP-LS advertisement of the BGP-EPE PeerAdj SID for L3 link. When advertising the SR-MPLS BGP-EPE Peer Adjacency SIDs for L2 bundle members, the PeerAdj SID TLV [RFC9086] MUST be carried in the L2 Bundle Member Attributes TLV to advertise the SR- MPLS Peer Adjacency 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.

When advertising SR-MPLS BGP-EPE Peer Adjacency SIDs for L2 bundle members, since L2 bundle information is considered a Layer 3 link attribute, it must be advertised in the BGP-LS Link NLRI. The details for LINK NLRI are the same as those for the PeerAdj SID, as described in Section 5.2 of [RFC9086]. This information mustnot be included in the BGP-LS Link NLRI that corresponds to the PeerNode SID, as defined in Section 5.1 of [RFC9086].

Note that for directly connected EBGP neighbors, if a BGP neighbor is established over an L2 Bundle, an additional BGP-LS Link NLRI(as described in Section 5.2 of [RFC9086]) must be generated to advertise Peer Link information when generating the BGP-LS Link NLRI (as described in Section 5.1 of [RFC9086]) corresponding to the PeerNode SID. The L2 Bundle Member Attributes TLV should be included under the BGP-LS Link Attribute TLVs.

The SR-MPLS BGP-EPE Peer Adjacency SIDs for L2 bundle members are advertised with a BGP-LS Link NLRI, where:

* BGP-LS Link NLRI: as described in Section 5.2 of [RFC9086].

* Link Attribute TLVs:

  - include the PeerAdj SID TLV [RFC9086] for Peer Link(Optional)

  - include the L2 Bundle Member Attributes TLV.

    o include the PeerAdj SID TLV [RFC9086] for each L2 Bundle Member.

3.2. SRv6

For SRv6, according to Section 4.1 of [RFC9514], the SRv6 End.X SID TLV is used for the advertisement of L3 link BGP EPE Peer Adjacency SID. When advertising the SRv6 BGP-EPE Peer Adjacency SIDs for L2 bundle members, 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.

Note Appendix A of [RFC9514], SRv6 BGP PeerNode is no longer advertised as BGP LINK NLRI. When advertising SRv6 BGP-EPE Peer Adjacency SIDs for L2 bundle members, since L2 bundle information is considered a Layer 3 link attribute, it must be advertised in the BGP-LS Link NLRI. The details for LINK NLRI are the same as those for the Peer Adjacency SID, as described in Section 5.2 of [RFC9086].

The SRv6 BGP-EPE Peer Adjacency SIDs for L2 bundle members are advertised with a BGP-LS Link NLRI, where:

* BGP-LS Link NLRI: as described in Section 5.2 of [RFC9086].

* Link Attribute TLV:

  - include the SRv6 End.X SID TLV [RFC9514] for Peer

Link (Optional).

- include the L2 Bundle Member Attributes TLV.

  o include the SRv6 End.X SID TLV [RFC9514] for each 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. MC-LAG Bundles Considerations

In environments where MC-LAG (Multi-Chassis Link Aggregation Group) bundles are deployed across multiple devices, it is critical to implement mechanisms to prevent Broadcast, Unknown Unicast, and Multicast (BUM) traffic from looping and ensure a loop-free network. The following loop prevention mechanisms are included:

* Split Horizon Forwarding: Each MC-LAG device maintains a split horizon rule where it does not forward BUM traffic received from one MC-LAG member port to another MC-LAG member port. This prevents BUM frames from being forwarded back into the MC-LAG, creating loops.

* Designated Forwarder Election: In a typical MC-LAG configuration, one device is elected as the designated forwarder for BUM traffic. This ensures that only one device is responsible for forwarding BUM frames, preventing the possibility of multiple devices forwarding the same frame simultaneously and causing a loop.

* Consistent Hashing Algorithms: MC-LAG devices employ consistent hashing algorithms to ensure that traffic distribution across member links is stable and predictable. This minimizes the risk of reordering and helps in effective loop prevention.

By incorporating these mechanisms, MC-LAG deployments can effectively prevent BUM traffic from looping and ensure a stable, loop-free network.

6. Implementation Status

[Note to the RFC Editor - remove this section before publication, as well as remove the reference to [RFC7942].

This section records the status of known implementations of the protocol defined by this specification at the time of posting of this Internet-Draft, and is based on a proposal described in [RFC7942]. The description of implementations in this section is intended to assist the IETF in its decision processes in progressing drafts to RFCs. Please note that the listing of any individual implementation here does not imply endorsement by the IETF. Furthermore, no effort has been spent to verify the information presented here that was supplied by IETF contributors. This is not intended as, and must not be construed to be, a catalog of available implementations or their features. Readers are advised to note that other implementations may exist.

According to [RFC7942], "this will allow reviewers and working groups to assign due consideration to documents that have the benefit of running code, which may serve as evidence of valuable experimentation and feedback that have made the implemented protocols more mature. It is up to the individual working groups to use this information as they see fit".

6.1. New H3C Technologies

* Organization: New H3C Technologies.

* Implementation: H3C CR16000, CR19000 series routers implementation.

* Description: All sections including all the "MUST" and "SHOULD" clauses have been implemented in above-mentioned New H3C Products (running Version 7.1.110 and above).

* Maturity Level: Product

* Coverage: All sections.

* Version: Draft-00

* Licensing: N/A

* Implementation experience: Nothing specific.

* Contact: li_meng_limeng@h3c.com

* Last updated: July 19, 2025

6.2. ZTE Corp

* Organization: ZTE Corporation

* Implementation: ZTE’s M6000 Series Routers

* Description: This feature has been implemented in ZTE M6000 series routers and follows the definition and mechanism as defined in Section 3 including all the "MUST" and "SHOULD" clauses.

* Maturity Level: Beta

* Coverage: All

* Version: Draft-00

* Licensing: N/A

* Implementation experience: Nothing specific.

* Contact: zhu.xiaolong@zte.com.cn

* Last updated: July 19, 2025

7. Security Considerations

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

This document does not introduce any new security consideration.

8. IANA Considerations

This document has no IANA actions.

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>.

[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] 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>.

9.2. Informative References

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

[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 SR-MPLS, 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)

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

  - PeerAdj SID TLV (Label-1012)

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

  - PeerAdj SID TLV (Label-1013)

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)
* L2 Bundle Member Attribute TLV (Link Local Identifier describing the member link 2)
  - SRv6 End.X SID TLV (SID-A::A2)
* L2 Bundle Member Attribute TLV (Link Local Identifier describing the member link 3)
  - SRv6 End.X SID TLV (SID-A::A3)

Acknowledgements

Many thanks to Sasha Vainshtein, Acee Lindem, Chen Ran, Liyan Gong, Yongqing Zhu, Lan cheng, Wisdom Tan, Yisong Liu, Libin Liu, Liu Yao, Hongwei Li, Allan Michael, Huo Pengfei, Gyan Mishra, Dong Jie, Meng Liu, etc. for their valuable comments on this document.

Authors' Addresses

Changwang Lin
New H3C Technologies
8 Yongjia North Road
Beijing
Haidian District, 100094
China
Email: linchangwang.04414@h3c.com

Zhenqiang Li
China Mobile
32 Xuanwumen West Street
Beijing
Xicheng District, 100053
China
Email: lizhenqiang@chinamobile.com

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

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

Ran Chen
ZTE Corporation
China
Email: chen.ran@zte.com.cn