idr
Internet-Draft
Intended status: Standards Track
Expires: 3 September 2026

Z. Zhang
K. Kompella
HPE
A. Mahale
Meta
R. Bhargava
Crusoe
A. Zhang
Westford Academy
2 March 2026

BGP Signaling for Multipath Traffic Engineering Junction States
draft-zzhang-idr-mpte-signaling-00

Abstract

Multi-Path Traffic Engineering (MPTE) combines Traffic Engineering with Multi-Path forwarding, offering a much desired TE solution for both traditional WAN and new AIML DC/DCI. MPTE tunnels are based on MPTE Directed Acyclic Graph (DAG) and can be signaled with extensions to RSVP-TE, PCEP, BGP. This document specifies the BGP protocol extensions and procedures for signaling MPTE DAGs.

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 3 September 2026.

Copyright Notice

Table of Contents

1. Introduction

[I-D.kompella-teas-mpte] describes the architecture and framework for Multipath Traffic Engineering (MPTE). A signaling approach was described, which could be implemented via extensions to RSVP, PCEP, or BGP. This document specifies how to signal MPTE Directed Acyclic Graphs (DAGs), in particular, how to provision junctions that make up an MPTE DAG, using a new AFI/SAFI, the MPTE AFI/SAFI, in BGP.

[I-D.ietf-bess-bgp-multicast-controller] specifies the BGP extensions to signal multicast replication states to multicast tree nodes. Much of the concepts and extensions can be used to signal MPTE Junction states. This section describes how that is achieved, and the difference between multicast signaling and MPTE signaling.

1.1. Mode of Operation

While the BGP signaling for MPTE is not limited to Data Centers (DCs), a DC using EBGP signaling is used as an example.

Assume the EBGP sessions between switches in the DC support the MPTE SAFI for the signaling of junction states. A future revision of this document will describe the use of Route Reflectors to a) isolate MPTE from the other functions of the EBGP mesh (basic routing), and b) scale sessions.

For each DAG, its Signaling Source (SS), which could be a controller or an ingress switch, originates a set of BGP routes of the MPTE SAFI, one for each junction node. The route is referred to as a Junction State route, and MUST carry a Route Target (RT) to target the route at the corresponding junction node. Once the route is propagated to the targeted node, the matched RT causes the route to be imported by the node and stopped from being propagated further. Before the matching, the route is propagated by the BGP infrastructure.

Before a junction node has at least one path set up to an egress, its upstream node should not start sending traffic to it. This ordered control is preferably done in a hop-by-hop fashion, like in the RSVP-TE case. When a junction has its local state set up for a DAG (starting with an egress node), it originates a RESV route for each of its PHOPs for the DAG, including encapsulation (e.g., an MPLS label) and BW information (e.g., the maximum traffic it expects from the PHOP). The upstream node repeats the process, and eventually the ingress node can start sending traffic. Note that a junction node may originate RESV routes before it receives from all its NHOPs. When more or updated RESV routes are received from its downstream, or when some of its downstream nodes are removed or no longer reachable, it will send updated RESV routes to its PHOPs.

As an option, the ordered control could be done by the signaling source (SS). In this case, the encapsulation information (e.g., MPLS labels) can be assigned by the SS and included in the junction routes (the label assignment options are detailed in [I-D.ietf-bess-bgp-multicast-controller]). After a junction node installs the forwarding state, it sends an acknowledgment route to the SS, which will tally the result and notify the ingress when and how much traffic can be put onto the DAG. This is as if the junction nodes were programmed with static routes, which shifts the burden/ complexity to the SS.

1.2. Collecting Topology/TE Information

Typically, Traffic Engineering uses the IGP (via TE extensions) to distribute topology and TE information. That is not an option for a DC that uses BGP signaling.

BGP-LS [RFC9552] is a mechanism using BGP extensions to collect link state and TE information that has been signaled by IGP. Typically, the information is distributed to a few collectors (e.g., controllers) from a few distributors (e.g., IGP border routers).

This document suggests using BGP LS [RFC9815] to distribute TE information for MPTE. Every switch is a distributor of its local information. If distributed calculation is used, each switch is also a collector of other switches' local information. More details will be provided.

1.3. Considerations for BGP Signaling

[I-D.kompella-teas-mpte] outlined the information carried in the JUNCTION message. When implemented in BGP, the MC ID, MPTED ID, MPTED Version and Tunnel Type are encoded in the NLRI of a new SAFI.

For the tunnel information part, the ingress/egress nodes information and tunnel bandwidth are (for now) not encoded.

The junction bandwidth is in the NLRI as well, but not considered as part of the NLRI key.

All the NHOP and PHOP information is encoded into a Tunnel Encapsulation Attribute (TEA) [RFC9012], with extensions specified in [I-D.ietf-bess-bgp-multicast-controller] and this document. A TEA encodes a list of "tunnels", each of which could be a real tunnel or just an interface or neighbor.

As explained in [I-D.ietf-bess-bgp-multicast-controller], when a TEA is attached to an NLRI of MCAST-TREE SAFI, corresponding traffic is replicated across the downstream tunnels in the TEA. Otherwise (including in the MPTE case), traffic is load-balanced across the downstream (NHOP in the case of MPTE) tunnels. Other than that, most of the TEA extensions defined in [I-D.ietf-bess-bgp-multicast-controller] are applicable to MPTE, with the following notes:

* All tunnel types and sub-TLVs mentioned in [I-D.ietf-bess-bgp-multicast-controller] can be used.

* A tunnel with an RPF sub-TLV is for a PHOP.

* The NHOP load share is encoded in the Weight sub-TLV [RFC9830].

* In the case of labeled MPTE tunnels:

  - The Tree Label Stack sub-TLV is used to signal the outgoing label (stack) of an NHOP.

  - The Receiving MPLS Label Stack sub-TLV is used to signal the incoming label (stack) of a PHOP.

* For the MCAST-TREE case, only one tunnel has an RPF sub-TLV, and either there is only one tunnel with the Receiving MPLS Label Stack sub-TLV in the case of P2MP tunnel, or each tunnel has a Receiving MPLS Label Stack sub-TLV in the case of MP2MP tunnel.

* For the MPTE case, only and all the PHOP tunnels for labeled MPTE tunnels have a Receiving MPLS Label Stack sub-TLV unless ordered control is used.

* The indication of an egress point (on a pure egress or on a bud node) is an Any-Encapsulation tunnel without either the RPF sub-TLV or any sub-TLV that identifies a downstream interface/tunnel. In the bud node case, this tunnel has the Weight sub-TLV, indicating the load share as the traffic is load-balanced between local delivery and other NHOP tunnels.

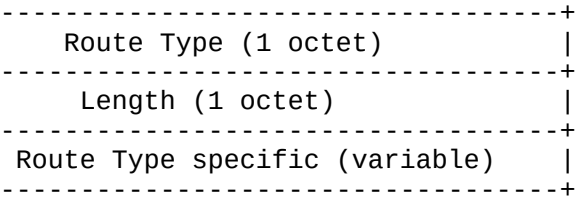
2. Specification

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.

2.1. AFI/SAFI and NLRI

This document defines a new SAFI type MPTE with value TBD1 for signaling MPTE junction states. When it is used with AFI 1, the IP addresses in the NLRI are IPv4. When it is used with AFI 2, the addresses in the NLRI are IPv6.

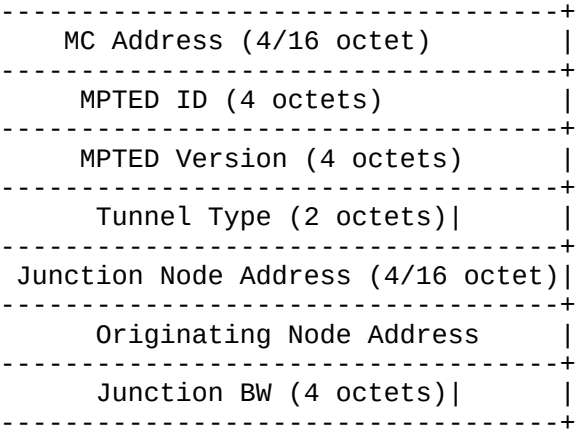
The NLRI is encoded as follows:



This document defines the following Route Types:

+ 1 - Junction State route
+ 2 - Junction RESV route

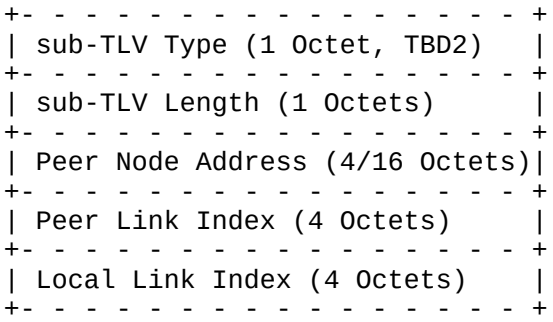
The Route Type specific part of the NLRI has the following format for both route types:



All the fields above, except the Junction BW, along with the route type and length are part of the NLRI key.

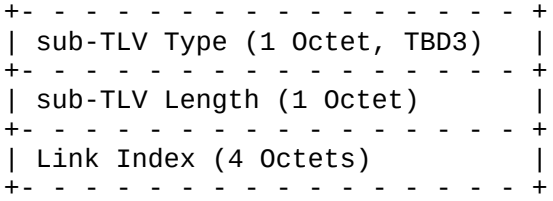
2.2. Full Link Identifier sub-TLV

The Full Link Identifier sub-TLV is used to identify an unnumbered interface by the Peer Node Address, Peer Link Index and Local Link Index. It is used for unnumbered PHOPs in the Junction State routes.



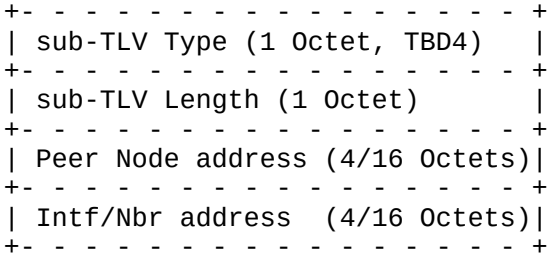
2.3. Link Index sub-TLV

The Link Index sub-TLV encodes the Link ID on a node receiving the route. It is used for unnumbered PHOPs in the Junction RESV routes .



2.4. Interface and Node Address sub-TLV

The Interface and Node Address sub-TLV encodes the local or neighbor address on an interface, and the address of the node that the interface connects to. The type of address (IPv4/IPv6) is inferred from the sub-TLV Length.



2.5. Procedures

2.5.1. Originating Junction State Routes

After the MC calculates an MPTED, the SS originates a Junction State route for each junction node. The route carries an IP Address Specific RT, with the Global Administrator field set to the junction node's address and the Local Administrator field set to 0.

The route carries a Tunnel Encapsulation Attribute (TEA). Each tunnel in the TEA corresponds to a PHOP or NHOP:

* Each tunnel is an Any-Encapsulation tunnel, with a Full Link Identifier sub-TLV or an Interface and Node Address sub-TLV to identify an incoming/outgoing link (in addition to other sub-TLVs).

* Each PHOP tunnel MUST also include the following sub-TLVs:

  - One RPF sub-TLV to indicate it is a PHOP

  - One Receiving MPLS Label Stack sub-TLV to encode the incoming label unless ordered control is used.

* Each NHOP tunnel MUST include one Tree Label Stack sub-TLV to encode the outgoing label unless ordered control is used. It MAY include a Weight sub-TLV to encode the NHOP share. If one NHOP tunnel includes a Weight sub-TLV, then all NHOP tunnels MUST include a Weight sub-TLV.

2.5.2. Receiving Junction State Routes

Each node X that receives an MPTE route with an RT whose Global Administrator field does not match its loopback address propagates the route to all its neighbors (except the one from which it received the route).

If the RT matches its own loopback address, X MUST import it, and MUST stop re-advertising the route upon match and importation.

Once the route is imported, X installs forwarding states as described in the following sections, in the case of MPLS when ordered control is not used (other tunnel types or ordered control will be specified in a future revision).

2.5.2.1. Building Forwarding Nexthop

When a data packet is received, an IP address or MPLS label lookup is done to produce the forwarding information about how the packet should be forwarded. The forwarding information is referred to as forwarding nexthop in this document, or simply nexthop when it is not ambiguous.

The forwarding nexthop for a junction is built by checking each NHOP tunnel. The Interface and Node Address sub-TLV or the Full Link Identifier sub-TLV identifies the outgoing interface/neighbor, and the Tree Label sub-TLV identifies the outgoing label. The Weight sub-TLV provides the load-balancing share for the link, and bandwidth reservation can be done based on the Junction Bandwidth in the NLRI and the Weight sub-TLV in the NHOP.

2.5.2.2. Installing Routes

For each PHOP tunnel in the TEA, a label route is installed with the label value in the Receiving Label Stack sub-TLV, pointing to the forwarding nexthop built as specified above.

2.5.2.3. Sending Junction State Route Acknowledgment

Each junction sends an acknowledgment back to the SS. Unless ordered control is used, the SS makes sure that all junctions are properly programmed before the tunnel is put into use.

The acknowledgement is simply the same Junction State route modified as follows:

* The Originating Node’s Address is set to the junction node’s address.

* The Route Target is set to target the SS.

2.5.3. Ordered Control

When hop-by-hop Ordered Control is used, the Junction State route does not carry encapsulation information (e.g., labels) in the PHOPs/NHOPs, and the junction’s forwarding state is not installed until at least one Junction RESV route has been received from one of the NHOPs. Each junction originates a Junction RESV route targeted at each of its upstream junctions. The route type specific part of the NRLI is set according to the Junction State route, with the Junction Node Address set to that of the upstream junction, which is from either the Interface and Node Address sub-TLV or the Full Link Identifier sub-TLV. The Originating Node Address is set to that of

this junction. A Route Target is used to target the route at the upstream junction.

The Junction BW is set to the total BW to be reserved on the upstream junction for this junction. A TEA is attached, with only PHOP tunnels toward the upstream junctions . The PHOP tunnel includes one of the following:

* A Tunnel Egress Endpoint sub-TLV, in which the address is set to the interface/neighbor address in the Interface and Node Address sub-TLV in the corresponding PHOP in the corresponding Junction State route.

* A Link Index sub-TLV, in which the Link Index is the Peer Link ID in the Full Link identifier sub-TLV in the corresponding PHOP in the corresponding Junction State route.

2.5.4. Route Update and Withdrawal

When a junction is updated (e.g., with added/removed/updated PHOPs/ NHOPs), the corresponding Junction State route is updated accordingly. If a junction is deleted, the corresponding Junction State route is withdrawn. Corresponding acknowledgement and reservation routes are updated, originated, or withdrawn accordingly.

2.5.5. Routes For Other Messages

To be added.

3. Security Considerations

To be added.

4. IANA Considerations

To be added.

5. Acknowledgments

The authors thank Vishnu Pavan Beeram, Chandrasekar Ramachandran, Sudharsana Venkataraman, and Jai Hari M K for their comments and suggestions.

6. References

6.1. Normative References

[I-D.ietf-bess-bgp-multicast-controller]
Zhang, Z. J., Raszuk, R., Pacella, D., and A. Gulko, "Controller-based BGP Multicast Signaling", Work in Progress, Internet-Draft, draft-ietf-bess-bgp-multicast-controller-16, 28 February 2025, <https://datatracker.ietf.org/doc/html/draft-ietf-bess-bgp-multicast-controller-16>.

[I-D.kompella-teas-mpte]
Kompella, K., Jalil, L., Khaddam, M., and A. Smith, "Multipath Traffic Engineering", Work in Progress, Internet-Draft, draft-kompella-teas-mpte-01, 7 July 2025, <https://datatracker.ietf.org/doc/html/draft-kompella-teas-mpte-01>.

[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>.

[RFC9012] Patel, K., Van de Velde, G., Sangli, S., and J. Scudder, "The BGP Tunnel Encapsulation Attribute", RFC 9012, DOI 10.17487/RFC9012, April 2021, <https://www.rfc-editor.org/info/rfc9012>.

[RFC9815] Patel, K., Lindem, A., Zandi, S., and W. Henderickx, "BGP Link State (BGP-LS) Shortest Path First (SPF) Routing", RFC 9815, DOI 10.17487/RFC9815, July 2025, <https://www.rfc-editor.org/info/rfc9815>.

[RFC9830] Previdi, S., Filsfils, C., Talaulikar, K., Ed., Mattes, P., and D. Jain, "Advertising Segment Routing Policies in BGP", RFC 9830, DOI 10.17487/RFC9830, September 2025, <https://www.rfc-editor.org/info/rfc9830>.

6.2. Informative References

[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>.

Authors' Addresses

Zhaohui Zhang
HPE
Email: zhaohui.zhang@hpe.com

Kireeti Kompella
HPE
Email: kireeti.ietf@gmail.com

Aditya Mahale
Meta
Email: aditya.ietf@gmail.com

Raghav Bhargava
Crusoe
Email: raghavbhargava12@gmail.com

Aaron Zhang
Westford Academy
Email: aaronzhang194@gmail.com