BESS
Internet-Draft
Intended status: Standards Track
Expires: December 11, 2020

Z. Zhang
Juniper Networks
R. Raszuk
Bloomberg LP
D. Pacella
Verizon
A. Gulko
Thomson Reuters

June 9, 2020

Controller Based BGP Multicast Signaling
draft-ietf-bess-bgp-multicast-controller-01

Abstract

This document specifies a way that one or more centralized controllers can use BGP to set up a multicast distribution tree in a network. In the case of labeled tree, the labels are assigned by the controllers either from the controllers' local label spaces, or from a common Segment Routing Global Block (SRGB), or from each routers Segment Routing Local Block (SRLB) that the controllers learn. In case of labeled unidirectional tree and label allocation from the common SRGB or from the controllers' local spaces, a single common label can be used for all routers on the tree to send and receive traffic with. Since the controllers calculate the trees, they can use sophisticated algorithms and constraints to achieve traffic engineering.

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.

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/.

Copyright Notice

Table of Contents

1. Overview 3
1.1. Introduction 3
1.2. Resilience 4
1.3. Signaling 5
1.4. Label Allocation 5
1.4.1. Using a Common per-tree Label for All Routers . . . 6
1.4.2. Upstream-assignment from Controller's Local Label Space 7
1.5. Determining Root/Leaves 8
1.5.1. PIM-SSM/Bidir or mLDP P2MP 8
1.5.2. PIM ASM 9
2. Specification 9
2.1. Enhancements to TEA 9
2.1.1. Any-Encapsulation Tunnel 9
2.1.2. Load-balancing Tunnel 9
2.1.3. RPF Sub-TLV 10
2.2. Context Label Wide Community 10
2.3. Procedures 10
3. Security Considerations 10
4. IANA Considerations 11
5. Acknowledgements 11
6. References 11
6.1. Normative References 11
6.2. Informative References 12

Authors' Addresses 12

1. Overview

1.1. Introduction

[I-D.zzhang-bess-bgp-multicast] describes a way to use BGP as a replacement signaling for PIM [RFC7761] or mLDP [RFC6388]. The BGP-based multicast signaling described there provides a mechanism for setting up both (s,g)/(*,g) multicast trees (as PIM does, but optionally with labels) and labeled (MPLS) multicast tunnels (as mLDP does). Each router on a tree performs essentially the same procedures as it would perform if using PIM or mLDP, but all the inter-router signaling is done using BGP.

These procedures allow the routers to set up a separate tree for each individual multicast (x,g) flow where the 'x' could be either 's' or '*', but they also allow the routers to set up trees that are used for more than one flow. In the latter case, the trees are often referred to as "multicast tunnels" or "multipoint tunnels", and specifically in this document they are mLDP tunnels (except that they are set up with BGP signaling). While it actually does not have to be restricted to mLDP tunnels, mLDP FEC is conveniently borrowed to identify the tunnel. In the rest of the document, the term tree and tunnel are used interchangeably.

The trees/tunnels are set up using the "receiver-initiated join" technique of PIM/mLDP, hop by hop from downstream routers towards the root. The BGP messages are either sent hop by hop between downstream routers and their upstream neighbors, or can be reflected by Route Reflectors (RRs).

As an alternative to each hop independently determining its upstream router and signaling upstream towards the root (following PIM/mLDP model), the entire tree can be calculated by a centralized controller, and the signaling can be entirely done from the controller, using the same BGP messages as defined in [I-D.zzhang-bess-bgp-multicast]. For that, some additional procedures and optimizations are specified in this document.

While it is outside the scope of this document, signaling from the controllers could be done via other means as well, like Netconf or any other SDN methods.

1.2. Resilience

Each router could establish direct BGP sessions with one or more controllers, or it could establish BGP sessions with RRs who in turn peer with controllers. For the same tree/tunnel, each controller may independently calculate the tree/tunnel and signal the routers on the tree/tunnel using MCAST-TREE S-PMSI/Leaf A-D routes [I-D.zzhang-bess-bgp-multicast]. How the tree/tunnel roots/leaves are discovered and how the calculation is done are outside the scope of this document.

On each router, BGP route selection rules will lead to one controller's route for the tree/tunnel being selected as the active route and used for setting up forwarding state. As long as all the routers on a tree/tunnel consistently pick the same controller's routes for the tree/tunnel, the setup should be consistent. If the tree/tunnel is labeled, different labels will be used from different controllers so there is no traffic loop issue even if the routers do not consistently select the same controlle's routes. In the unlabeled case, to ensure the consistency the selection SHOULD be solely based on the identifier of the controller, which could be carried in an Address Specific Extended Community (EC).

Another consistency issue is when a bidirectional tree/tunnel needs to be re-routed. Because this is no longer triggered hop-by-hop from downstream to upstream, it is possible that the upstream change happens before the downstream, causing traffic loop. In the unlabeled case, there is no good solution (other than that the controller issues upstream change only after it gets acknowledgement from downstream). In the labeled case, as long as a new label is used there should be no problem.

Besides the traffic loop issue, there could be transient traffic loss before both the upstream and downstream's forwarding state are updated. This could be mitigated if the upstream keep sending traffic on the old path (in addition to the new path) and the downstream keep accepting traffic on the old path (but not on the new path) for some time. It is a local matter when for the downstream to switch to the new path - it could be data driven (e.g., after traffic arrives on the new path) or timer driven.

For each tree, multiple disjoint instances could be calculated and signaled for live-live protection. Different labels are used for different instances, so that the leaves can differentiate incoming traffic on different instances. As far as transit routers are concerned, the instances are just independent. Note that the two instances are not expected to share common transit routers (it is otherwise outside the scope of this document/revision).

1.3. Signaling

Each router only receives S-PMSI/Leaf A-D routes from the controllers but does not originate or re-advertise those routes. The re-advertisement of a received route can be blocked based on the fact that a configured import RT matches the RT of the route, which indicates that this router is the target and consumer of the route hence it should not be re-advertised further. The routes includes the outgoing forwarding information in the form of Tunnel Encapsulation Attributes (TEA) [I-D.ietf-idr-tunnel-encaps], with enhancements specified in this document.

Suppose that for a particular tree, there are two downstream routers D1 and D2 for a particular upstream router U. A controller C may send two Leaf A-D routes to U, as if the two routes were originated by D1 and D2 but reflected by the controller. Alternatively, C could just send one route to U, with the Upstream Router's IP Address field set to U's IP address and the TEA specifying both the two downstreams and its upstream (see Section 2.1.3). In this case, the Originating Router's Address field of the Leaf A-D route is set to the controller's address. Note that for a TEA attached to a unicast NLRI, only one of the tunnels in a TEA is used for forwarding a particular packet, while all the tunnels in a TEA are used to reach multiple endpoints when it is attached to a multicast NLRI.

Note that, in case of labeled trees, the (x,g) or mLDP FEC signaling is actually not needed to transit routers but only needed on tunnel root/leaves. However, for consistency, the same signaling is used to all routers.

1.4. Label Allocation

In the case of labeled multicast signaled hop by hop towards the root, whether it's (x,g) multicast or "mLDP" tunnel, labels are assigned by a downstream router and advertised to its upstream router (from traffic direction point of view). In the case of controller based signaling, routers do not originate tree join (S-PMSI/Leaf A-D) routes anymore, so the controllers have to assign labels on behalf of routers, and there are three options for label assignment:

- From each router's SRLB that the controller learns
- From the common SRGB that the controller learns
- From the controller's local label space

Assignment from each router's SRLB is no different from each router assigning labels from its own local label space in the hop-by-hop

signaling case. The assignments for a router is independent of assignments for another router, even for the same tree.

Assignment from the controller's local label space is upstream-assigned [RFC5331]. It is used if the controller does not learn the common SRGB or each router's SRLB. Assignment from the SRGB [RFC8402] is only meaningful if all SRGBs are the same and a single common label is used for all the routers on a tree in case of unidirectional tree/tunnel (Section 1.4.1). Otherwise, assignment from SRLB is preferred.

The choice of which of the options to use depends on many factors. An operator may want to use a single common label per tree for ease of monitoring and debugging, but that requires explicit RPF checking and either SRGB or upstream assigned labels, which may not be supported due to either the software or hardware limitations (e.g. label imposition/disposition limits). In an SR network, assignment from the common SRGB if it's required to use a single common label per unidirectional tree, or otherwise assignment from SRLB is a good choice because it does not require support for context label spaces.

1.4.1. Using a Common per-tree Label for All Routers

MPLS labels only have local significance. For an LSP that goes through a series of routers, each router allocates a label independently and it swaps the incoming label (that it advertised to its upstream) to an outgoing label (that it received from its downstream) when it forwards a labeled packet. Even if the incoming and outgoing labels happen to be the same on a particular router, that is just incidental.

With Segment Routing, it is becoming a common practice that all routers use the same SRGB so that a SID maps to the same label on all routers. This makes it easier for operators to monitor and debug their network. The same concept applies to multicast trees as well - a common per-tree label is used for a router to receive traffic from its upstream neighbor and replicate traffic to all its downstream neighbor.

However, a common per-tree label can only be used for unidirectional trees. Additionally, it requires each router to do explicit RPF check, so that only packets from its expected upstream neighbor are accepted. Otherwise, traffic loop may form during topology changes, because the forwarding state update is no longer ordered.

Traditionally, p2mp mpls forwarding does not require explicit RPF check as a downstream router advertises a label only to its upstream router and all traffic with that incoming label is presumed to be

from the upstream router and accepted. When a downstream router switches to a different upstream router a different label will be advertised, so it can determine if traffic is from its expected upstream neighbor purely based on the label. Now with a single common label used for all routers on a tree to send and receive traffic with, a router can no longer determine if the traffic is from its expected neighbor just based on that common tree label. Therefore, explicit RPF check is needed. Instead of interface based RPF checking as in PIM case, neighbor based RPF checking is used - a label identifying the upstream neighbor precedes the tree label and the receiving router checks if that preceding neighbor label matches its expected upstream neighbor. Notice that this is similar to what's described in Section "9.1.1 Discarding Packets from Wrong PE" of RFC 6513 (an egress PE discards traffic sent from a wrong ingress PE). The only difference is one is used for label based forwarding and the other is used for (s,g) based forwarding. [note: for bidirectional trees, we may be able to use two labels per tree - one for upstream traffic and one for downstream traffic. This needs further verification].

Both the common per-tree label and the neighbor label are allocated either from the common SRGB or from the controller's local label space. In the latter case, an additional label identifying the controller's label space is needed, as described in the following section.

1.4.2. Upstream-assignment from Controller's Local Label Space

In this case in the multicast packet's label stack the tree label and upstream neighbor label (if used in case of single common-label per tree) are preceded by a downstream-assigned "context label". The context label identifies a context-specific label space (the controller's local label space), and the upstream-assigned label that follows it is looked up in that space.

This specification requires that, in case of upstream-assignment from a controller's local label space, each router D to assign, corresponding to each controller C, a context label that identifies the upstream-assigned label space used by that controller. This label, call it Lc-D, is communicated by D to C.

Suppose a controller is setting up unidirectional tree T. It assigns that tree the label Lt, and assigns label Lu to identify router U which is the upstream of router D on tree T. C needs to tell U: "to send a packet on the given tree/tunnel, one of the things you have to do is push Lt onto the packet's label stack, then push Lu, then push Lc-D onto the packet's label stack, then unicast the packet to D".

Controller C also needs to inform router D of the correspondence between <Lc-D, Lu, Lt> and tree T.

To achieve that, when C sends an S-PMSI/Leaf A-D route, for each tunnel in the TEA, it includes a label stack Sub-TLV [I-D.ietf-idr-tunnel-encaps], with the outer label being the context label Lc-D (received by the controller from the corresponding downstream), the next label being the upstream neighbor label Lu, and the inner label being the label Lt assigned by the controller for the tree. The router receiving the route will use the label stacks to send traffic to its downstreams.

For C to signal the expected label stack for D to receive traffic with, we overload a tunnel TLV in the TEA of the Leaf A-D route sent to D - if the tunnel TLV has a RPF sub-TLV (Section 2.1.3), then it indicates that this is actually for receiving traffic from the upstream.

Note that the use of TEA to specify downstream and upstream forwarding information also apply to label assignment from the common SRGB or each router's SRLB, with the differences that the context label is not needed in the SRGB/SRLB case, and that in SRLB case only a Label Stack Sub-TLV with a single SRLB label is used for upstream and downstream forwarding information (no RPF Label Stack Sub-TLV is needed) in the SRLB case.

1.5. Determining Root/Leaves

For the controller to calculate a tree, it needs to determine the root and leaves of the tree. This may be based on provisioning (static or dynamically programmed), or based on BGP signaling using the BGP multicast messages defined in [I-D.zzhang-bess-bgp-multicast], as described in the following two sections.

In both cases, the BGP updates are targeted at the controller, via an address specific Route Target with Global Administration Field set to the controller's address and the Local Administration Field set to 0, or a value pre-assigned to identify a VPN.

1.5.1. PIM-SSM/Bidir or mLDP P2MP

In this case, the PIM Last Hop Routers (LHRs) with interested receivers or mLDP P2MP tunnel leaves encode a Leaf A-D route with the Upstream Router's IP Address field set to the controller's address and the Originating Router's IP Address set to the address of the LHR or the P2MP tunnel leaf. The encoded PIM SSM source or mLDP FEC

provides root information and the Originating Router's IP Address provides leaves information.

1.5.2. PIM ASM

In this case, the First Hop Routers (FHRs) originate Source Active routes which provides root information, and the LHRs originate Leaf A-D routes, encoded as in the PIM-SSM case except that it is (*,G) instead of (S,G). The Leaf A-D routes provide leaf information.

2. Specification

2.1. Enhancements to TEA

This document specifies two new Tunnel Types and new sub-TLV. The type codes will be assigned by IANA from the "BGP Tunnel Encapsulation Attribute Tunnel Types".

2.1.1. Any-Encapsulation Tunnel

When a multicast packet needs to be sent from an upstream node to a downstream node, it may not matter how it is sent - natively when the two nodes are directly connected or tunneled otherwise. In case of tunneling, it may not matter what kind of tunnel is used - MPLS, GRE, IPinIP, or whatever.

To support this, an "Any-Encapsulation" tunnel type is defined. This tunnel MUST have a Tunnel Endpoint Sub-TLV and SHOULD NOT have any other Sub-TLVs. The Tunnel Endpoint Sub-TLV specifies an IP address, which could be any of the following:

- An interface's local address - when a packet needs to sent out of the corresponding interface natively.
- An interface's remote address - when a packet needs to sent to the address natively.
- An address that is not directly connected - when a packet needs to be tunneled to the address (any tunnel type/instance can be used).

2.1.2. Load-balancing Tunnel

Consider that a multicast packet needs to be sent to a downstream node, which could be reached via four paths P1~P4. If it does not matter which of path is taken, an "Any-Encapsulation" tunnel with the Tunnel Endpoint Sub-TLV specifying the downstream node's loopback address works well. If the controller wants to specify that only P1~P2 should be used, then a "Load-balancing" tunnel needs to be

used, listing P1 and P2 as member tunnels of the "Load-balancing" tunnel.

A load-balancing tunnel has one "Member Tunnels" Sub-TLV defined in this document. The Sub-TLV is a list of tunnels, each specifying a way to reach the downstream. A packet will be sent out of one of the tunnels listed in the Member Tunnels Sub-TLV of the load-balancing tunnel.

2.1.3. RPF Sub-TLV

The RPF sub-TLV has a type to be allocated by IANA and a one-octet length. The length is 0 currently, but if necessary in the future, sub-sub-TLVs could be placed in its value part. If the RPF sub-TLV appears in a tunnel, it indicates that the "tunnel" is for the upstream instead of a downstream. If a Label Stack sub-TLV is also present, then the inner most label in the stack is the incoming label identifying the tree (for comparison the inner most label for a downstream tunnel, i.e., when the RPF sub-TLV is not present, is the outgoing label). If the Label Stack sub-TLVe has more than one labels, the second inner most label in the stack identifies the expected upstream neighbor and explicit RPF checking needs to be set up for the tree label accordingly.

2.2. Context Label Wide Community

For a router to signal the context label that it assigns for a controller (or any label allocator that assigns labels that will be seen by this router), it attaches a Context Label Wide Community [I-D.ietf-idr-wide-bgp-communities] to the host route for its own address used in its BGP session towards the controllers (directly or via RRs). This is a new wide community that specifies the (Label Allocator, Context Label) tuple, and the exact format will be specified in a future revision.

2.3. Procedures

Details to be added. The general idea is described in the introduction section.

3. Security Considerations

This document does not introduce new security risks.

4. IANA Considerations

This document makes the following IANA requests:

- "Any-Encapsulation" and "Load-balancing" tunnel types from the "BGP Tunnel Encapsulation Attribute Tunnel Types" registry

- "Member Tunnels" and "RPF" sub-TLV types from the "BGP Tunnel Encapsulation Attribute Sub-TLVs" registry. The "Member Tunnels" sub-TLV has a two-octet value length. while the "RPF" sub-TLV has a one-octet value length.

-

5. Acknowledgements

The authors Eric Rosen for his questions, suggestions, and help finding solutions to some issues like the neighbor based explicit RPF checking. The authors also thank Lenny Giuliano, Sanoj Vivekanandan and IJsbrand Wijnands for their review and comments.

6. References

6.1. Normative References

[I-D.ietf-idr-tunnel-encaps]
Patel, K., Velde, G., and S. Ramachandra, "The BGP Tunnel Encapsulation Attribute", draft-ietf-idr-tunnel-encaps-15 (work in progress), December 2019.

[I-D.ietf-idr-wide-bgp-communities]
Raszuk, R., Haas, J., Lange, A., Decraene, B., Amante, S., and P. Jakma, "BGP Community Container Attribute", draft-ietf-idr-wide-bgp-communities-05 (work in progress), July 2018.

[I-D.zzhang-bess-bgp-multicast]
Zhang, Z., Giuliano, L., Patel, K., Wijnands, I., mishra, m., and A. Gulko, "BGP Based Multicast", draft-zzhang-bess-bgp-multicast-03 (work in progress), October 2019.

[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>.

6.2. Informative References

[RFC6388] Wijnands, IJ., Ed., Minei, I., Ed., Kompella, K., and B. Thomas, "Label Distribution Protocol Extensions for Point-to-Multipoint and Multipoint-to-Multipoint Label Switched Paths", RFC 6388, DOI 10.17487/RFC6388, November 2011, <https://www.rfc-editor.org/info/rfc6388>.

[RFC6513] Rosen, E., Ed. and R. Aggarwal, Ed., "Multicast in MPLS/BGP IP VPNs", RFC 6513, DOI 10.17487/RFC6513, February 2012, <https://www.rfc-editor.org/info/rfc6513>.

[RFC7761] Fenner, B., Handley, M., Holbrook, H., Kouvelas, I., Parekh, R., Zhang, Z., and L. Zheng, "Protocol Independent Multicast - Sparse Mode (PIM-SM): Protocol Specification (Revised)", STD 83, RFC 7761, DOI 10.17487/RFC7761, March 2016, <https://www.rfc-editor.org/info/rfc7761>.

[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

Zhaohui Zhang
Juniper Networks

EMail: zzhang@juniper.net

Robert Raszuk
Bloomberg LP

EMail: robert@raszuk.net

Dante Pacella
Verizon

EMail: dante.j.pacella@verizon.com

Arkadiy Gulko
Thomson Reuters

EMail: arkadiy.gulko@thomsonreuters.com