BESS
Internet-Draft
Intended status: Standards Track
Expires: May 1, 2020

Z. Zhang
L. Giuliano
Juniper Networks
K. Patel
Arrcus
I. Wijnands
M. Mishra
Cisco Systems
A. Gulko
Refinitiv

October 29, 2019

BGP Based Multicast
draft-zzhang-bess-bgp-multicast-03

Abstract

This document specifies a BGP address family and related procedures that allow BGP to be used for setting up multicast distribution trees. This document also specifies procedures that enable BGP to be used for multicast source discovery, and for showing interest in receiving particular multicast flows. Taken together, these procedures allow BGP to be used as a replacement for other multicast routing protocols, such as PIM or mLDP. The BGP procedures specified here are based on the BGP multicast procedures that were originally designed for use by providers of Multicast Virtual Private Network service.

Requirements Language

The key words "MUST", "MUST NOT", "REQUIRED", "SHALL", "SHALL NOT", "SHOULD", "SHOULD NOT", "RECOMMENDED", "MAY", and "OPTIONAL" in this document are to be interpreted as described in RFC2119.

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 May 1, 2020.

Copyright Notice

Table of Contents

1. Introduction 3
1.1. Motivation 3
1.1.1. Native/unlabeled Multicast 3
1.1.2. Labeled Multicast 4
1.2. Overview 4
1.2.1. (x,g) Multicast 5
1.2.1.1. Source Discovery for ASM 5
1.2.1.2. ASM Shared-tree-only Mode 6
1.2.1.3. Integration with BGP-MVPN 7
1.2.2. BGP Inband Signaling for mLDP Tunnel 7
1.2.3. BGP Sessions 7
1.2.4. LAN and Parallel Links 8
1.2.5. Transition 9
2. Specification 10
2.1. BGP NLRIs and Attributes 10
2.1.1. S-PMSI A-D Route 11
2.1.2. Leaf A-D Route 11
2.1.3. Source Active A-D Route 12
2.1.4. S-PMSI A-D Route for C-multicast mLDP 13
2.1.5. Session Address Extended Community 13
2.2. Procedures 14
2.2.1. Source Discovery for ASM 14
2.2.2. Originating Tree Join Routes 14
2.2.2.1. (x,g) Multicast Tree 14
2.2.2.2. BGP Inband Signaling for mLDP Tunnel 15
2.2.3. Receiving Tree Join Routes 15

2.2.4. Withdrawl of Tree Join Routes 16
2.2.5. LAN procedures for (x,g) Unidirectional Tree 16
2.2.5.1. Originating S-PMSI A-D Routes 16
2.2.5.2. Receiving S-PMSI A-D Routes 17
2.2.6. Distributing Label for Upstream Traffic for Bidirectional Tree/Tunnel 17
3. Security Considerations 18
4. Acknowledgements 18
5. References 18
5.1. Normative References 18
5.2. Informative References 19
Authors' Addresses 20

1. Introduction

1.1. Motivation

This section provides some motivation for BGP signaling for native and labeled multicast. One target deployment would be a Data Center that requires multicast but uses BGP as its only routing protocol [RFC7938]. In such a deployment, it would be desirable to support multicast by extending the deployed routing protocol, without requiring the deployment of tree building protocols such as PIM, mLDP, RSVP-TE P2MP, and without requiring an IGP.

Additionally, compared to PIM, BGP based signaling has several advantage as described in the following section, and may be desired in non-DC deployment scenarios as well.

1.1.1. Native/unlabeled Multicast

Protocol Independent Multicast (PIM) has been the prevailing multicast protocol for many years. Despite its success, it has two drawbacks:

- o The ASM model, which is prevalent, introduces complexity in the following areas: source discovery procedures, need for Rendezvous Points (RPs) and group-to-RP mappings, need to switch between RP-rooted trees and source-rooted trees, etc.

- o Periodical protocol state refreshes due to soft state nature.

PIM-SSM removes much of the complexity of PIM-ASM by moving source discovery to the application layer. However, for various reasons, many legacy applications and devices still rely upon network-based source discovery. PIM-Port (PIM over Reliable Transport) solves the soft state issue, though its deployment has also been limited for two reasons:

- o It does not remove the ASM complexities.

- o In many of the scenarios where reliable transport is deemed important, BGP-based multicast (e.g. BGP-MVPN) has been used instead of PORT.

Partly because of the above mentioned problems, some Data Center operators have been avoiding deploying multicast in their networks.

BGP-MVPN [RFC6514] uses BGP to signal VPN customer multicast state over provider networks. It removes the above mentioned problems from the SP environment, and the deployment experiences have been encouraging. While RFC 6514 makes it possible for an SP to provide MVPN service without running PIM on its backbone, that RFC still assumes that PIM (or mLDP) runs on the PE-CE links. [draft-ietf-bess-mvpn-pe-ce] adapts the concept of BGP-MVPN to PE-CE links so that the use of PIM on the PE-CE links can be eliminated (though the PIM-ASM complexities still remains in the customer network), and this document extends it further to general topologies, so that they can be run on any router, as a replacement for PIM or mLDP.

With that, PIM can be completely eliminated from the network. PIM soft state is replaced by BGP hard state. For ASM, source specific trees are set up directly after simpler source discovery (data driven on FHRs and control driven elsewhere), all based on BGP. All the complexities related to source discovery and shared/source tree switch are also eliminated. Additionally, the trees can be setup with MPLS labels, with just minor enhancements in the signaling.

1.1.2. Labeled Multicast

There could be two forms of labeled multicast signaled by BGP. The first one is labeled (x,g) multicast where 'x' stands for either 's' or '*'. Basically, it is for BGP-signaled multicast tree as described in previous section but with labels. The second one is for mLDP tunnels with BGP signaling in part or whole through a BGP domain.

For both cases, BGP is used because other label distribution mechanisms like mLDP may not be desired by some operators. For example, a DC operator may prefer to have a BGP-only deployment.

1.2. Overview

1.2.1. (x,g) Multicast

PIM-like functionality is provided, using BGP-based join/prune signaling and BGP-based source discovery for ASM. The BGP-based join signaling supports both labeled multicast and IP multicast.

The same RPF procedures as in PIM are used for each router to determine the RPF neighbor for a particular source or RPA (in case of Bidirectional Tree). Except in the Bidirectional Tree case and a special case described in Section 1.2.1.2, no (*,G) join is used - LHR routers discover the sources for ASM and then join towards the sources directly. Data driven mechanisms like PIM Assert is replaced by control driven mechanisms (Section 1.2.4).

The joins are carried in BGP Updates with MCAST-TREE SAFI and S-PMSI/ Leaf A-D routes defined in this document. The updates are targeted at the upstream neighbor by use of Route Targets. [Note - earlier version of this draft uses C-multicast route to send joins. We're now switching to S-PMSI/Leaf routes for three reasons. a) when the routes go through RRs, we have to distinguish different routes based on upstream router and downstream router. This leads to Leaf routes. b) for labeled bidirectional trees, we need to signal "upstream fec". S-PMSI suits this very well. c) we may want to allow the option of setting up trees from the roots instead of from the leaves. S-PMSI suits that very well.]

If the BGP updates carry labels (via Tunnel Encapsulation Attribute [I-D.ietf-idr-tunnel-encaps]), then (s,g) multicast traffic can use the labels. This is very similar to mLDP Inband Signaling [RFC6826], except that there are no corresponding "mLDP tunnels" for the PIM trees. Similar to mLDP, labeled traffic on transit LANs are point to point. Of course, traffic sent to receivers on a LAN by a LHR is native multicast.

For labeled bidirectional (*,g) trees, downstream traffic (away from the RPA) can be forwarded as in the (s,g) case. For upstream traffic (towards RPA), the upstream neighbor needs to advertise a label for its downstream neighbors. The same label that the upstream neighbor advertises to its upstream is the same one that it advertises to its downstreams, using an S-PMSI A-D route.

1.2.1.1. Source Discovery for ASM

This document does not support ASM via shared trees (aka RP Tree, or RPT) with one exception discussed in the next section. Instead, FHRs, LHRs, and optionally RRs work together to propagate/discover source information via control plane and LHRs join source specific Shortest Path Trees (SPT) directly.

A FHR originates Source Active A-D routes upon discovering sources for particular flows and advertise them to its peers. It is desired that the SA routes only reach LHRs that are interested in receiving the traffic. To achieve that, the SA routes carry an IPv4 or IPv6 address specific Route Target. The Global Administrator field is set the group address of the flow, and the Local Administrator field is set to 0. An LHR advertises Route Target Membership routes, with the Route Target field in the NLRI set according to the groups it wants to receive traffic for, as how a FHR encode the Route Target in its Source Active routes. The propagation of the SA routes is subject to cooperative export filtering as specified in [RFC4684] and referred to as RTC mechanism in this document. That way, the LHR only receives Source Active routes for groups that it is interested in.

Typically, a set of RRs are used and they maintains all Source Active routes but only distribute to interested LHRs on demand (upon receiving corresponding Route Target Membership routes, which are triggered on LHRs when they receive IGMP/MLD membership routes). The rest of the document assumes that RRs are used, even though that is not required.

1.2.1.2. ASM Shared-tree-only Mode

It may be desired that only a shared tree is used to distribute all traffic for a particular ASM group from its RP to all LHRs, as described in Section 4.1 "PIM Shared Tree Forwarding" of [RFC7438]. This will significantly cut down the number of trees and works out very well in certain deployment scenarios. For example, all the sources could be connected to the RP, or clustered close to RP. In the latter case, either the path from FHRs to the RP do not intersect the shared tree so native forwarding can be used between the FHRs and the RP, or other means outside of this document could be used to forward traffic from FHRs to the RP.

For native forwarding from FHRs to the RP, SA routes may be used to announce the sources so that the RP can join source specific trees to pull traffic, but the group specific Route Target is not needed. The LHRs do not advertise the group specific Route Target Membership routes as they do not need the SA routes.

To establish the shared tree, (*,g) Leaf A-D routes are used as in the bidirectional tree case, though no forwarding state is established to forward traffic from downstream neighbors.

1.2.1.3. Integration with BGP-MVPN

For each VPN, the Source Active routes distribution in that VPN do not have to involve PEs at all unless there are sources/receivers directly connected to some PEs and they are independent of MVPN SA routes. For example, FHRs and LHRs establish BGP sessions with RRs of that particular VPN for the purpose of SA distribution.

After source discovery, BGP multicast signaling is done from LHRs towards the sources. When the signaling reaches an egress PE, BGP-MVPN signaling takes over, as if a PIM (s,g) join/prune was received on the PE-CE interface. When the BGP-MVPN signaling reaches the ingress PE, BGP multicast signaling as specified in this document takes over, similar to how BGP-MVPN triggers PIM (s,g) join/prune on PE-CE interfaces.

1.2.2. BGP Inband Signaling for mLDP Tunnel

Part of an (or the whole) mLDP tunnel can also be signaled via BGP and seamlessly integrated with the rest of mLDP tunnel signaled natively via mLDP. All the procedures are similar to mLDP except that the signaling is done via BGP. The mLDP FEC is encoded as the BGP NLRI, with MCAST-TREE SAFI and S-PMSI/Leaf A-D Routes for C-multicast mLDP defined in this document. The Leaf A-D routes correspond to mLDP Label Mapping messages, and the S-PMSI A-D routes are used to signal upstream FEC for MP2MP mLDP tunnels, similar to the bidirection (*,g) case.

1.2.3. BGP Sessions

In order for two BGP speakers to exchange MCAST-TREE NLRI, they must use BGP Capabilities Advertisement [RFC5492] to ensure that they both are capable of properly processing the MCAST-TREE NLRI. This is done as specified in [RFC4760], by using a capability code 1 (multiprotocol BGP) with an AFI of IPv4 (1) or IPv6 (2) and a SAFI of MCAST-TREE with a value to be assigned by IANA.

How the BGP peer sessions are provisioned, whether EBGP or IBGP, whether statically, automatically (e.g., based on IGP neighbor discovery), or programmably via an external controller, is outside the scope of this document.

In case of IBGP, it could be that every router peering with Route Reflectors, or hop by hop IBGP sessions could be used to exchange MCAST-TREE NLRIs for joins. In the latter case, unless desired otherwise for reasons outside of the scope of this document, the hop by hop IBGP sessions SHOULD only be used to exchange MCAST-TREE NLRIs.

When multihop BGP is used, a router advertises its local interface addresses, for the same purposes that the Address List TLV in LDP serves. This is achieved by advertising the interface address as host prefixes with IPv4/v6 Address Specific ECs corresponding to the router's local addresses used for its BGP sessions (Section 2.1.5).

Because the BGP Capability Advertisement is only between two peers, when the sessions are only via RRs, a router needs another way to determine if its neighbor is capable of signaling multicast via BGP. The interface address advertisement can be used for that purpose - the inclusion of a Session Address EC indicates that the BGP speaker identified in the EC supports the C-Multicast NLRI.

FHRs and LHRs may also establish BGP sessions to some Route Reflectors for source discovery purpose (Section 1.2.1.1).

With the traditional PIM, the FHRs and LHRs refer to the PIM DRs on the source or receiver networks. With BGP based multicast, PIM may not be running at all, and the FHRs and LHRs refer to the IGMP/MLD queriers, or the DF elected per [I-D.wijnands-bier-mld-lan-election]. Alternatively, if it is known that a network only has senders then no IGMP/MLD or DF election is needed - any router may generate SA routes. That will not cause any issue other than redundant SA routes being originated.

1.2.4. LAN and Parallel Links

There could be parallel links between two BGP peers. A single multi-hop session, whether IBGP or EBGP, between loopback addresses may be used. Except for LAN interfaces in case of unlabeled (x,g) unidirectional trees (note that transit LAN interface is not supported for BGP signaled (*,g) bidirectional tree and for mLDP tunnels, traffic on transit LAN is point to point between neighbors), any link between the two peers can be automatically used by a downstream peer to receive traffic from the upstream peer, and it is for the upstream peer to decide which link to use. If one of the links goes down, the upstream peer switches to a different link and there is no change needed on the downstream peer.

For unlabeled (x,g) unidirectional trees, the upstream peer MAY prefer LAN interfaces to send traffic, since multiple downstream peers may be reached simultaneously, or it may make a decision based on local policy, e.g., for load balancing purpose. Because different downstream peers might choose different upstream peers for RPF, when an upstream peer decides to use a LAN interface to send traffic, it originates an S-PMSI A-D route indicating that one or more LAN interface will be used. The route carries Route Targets specific to the LANs so that all the peers on the LANs import the route. If more

than one router originate the route specifying the same LAN for the same (s,g) or (*,g) flow, then assert procedure based on the S-PMSI A-D routes happens and assert losers will stop sending traffic to the LAN.

1.2.5. Transition

A network currently running PIM can be incrementally transitioned to BGP based multicast. At any time, a router supporting BGP based multicast can use PIM with some neighbors (upstream or downstream) and BGP with some other neighbors. PIM and BGP MUST not be used simultaneously between two neighbors for multicast purpose, and routers connected to the same LAN MUST be transitioned during the same maintenance window.

In case of PIM-SSM, any router can be transitioned at any time (except on a LAN all routers must be transitioned together). It may receive source tree joins from a mixed set of BGP and PIM downstream neighbors and send source tree joins to its upstream neighbor using either PIM or BGP signaling.

In case of PIM-ASM, the RPs are first upgraded to support BGP based multicast. They learn sources either via PIM procedures from PIM FHRs, or via Source Active A-D routes from BGP FHRs. In the former case, the RPs can originate proxy Source Active A-D routes. There may be a mixed set of RPs/RRs - some capable of both traditional PIM RP functionalities while some only redistribute SA routes.

Then any routers can be transitioned incrementally. A transitioned LHR router will pull Source Active A-D routes from the RPs/RRs when they receive IGMP/MLD (*,G) joins for ASM groups, and may send either PIM (s,g) joins or BGP Source Tree Join routes. A transitioned transit router may receive (*,g) PIM joins but only send source tree joins after pulling Source Active A-D routes from RPs/RRs.

Similarly, a network currently running mLDP can be incrementally transitioned to BGP signaling. Without the complication of ASM, any router can be transitioned at any time, even without the restriction of coordinated transition on a LAN. It may receive mixed mLDP label mapping or BGP updates from different downstream neighbors, and may exchange either mLDP label mapping or BGP updates with its upstream neighbors, depending on if the neighbor is using BGP based signaling or not.

2. Specification

2.1. BGP NLRIs and Attributes

The BGP Multiprotocol Extensions [RFC4760] allow BGP to carry routes from multiple different "AFI/SAFIs". This document defines a new SAFI known as a MCAST-TREE SAFI with a value to be assigned by the IANA. This SAFI is used along with the AFI of IPv4 (1) or IPv6 (2).

The MCAST-TREE NLRI defined below is carried in the BGP UPDATE messages [RFC4271] using the BGP multiprotocol extensions [RFC4760] with a AFI of IPv4 (1) or IPv6 (2) assigned by IANA and a MCAST-TREE SAFI with a value to be assigned by the IANA.

The Next hop field of MP_REACH_NLRI attribute SHALL be interpreted as an IPv4 adress whenever the length of the Next Hop address is 4 octets, and as an IPv6 address whenever the length of the Next Hop is address is 16 octets.

The NLRI field in the MP_REACH_NLRI and MP_UNREACH_NLRI is a prefix with a maximum length of 12 octers for IPv4 AFI and 36 octets for IPv6 AFI. The following is the format of the MCAST-TREE NLRI:

```
+-----------------------------------+
|    Route Type (1 octet)           |
+-----------------------------------+
|     Length (1 octet)              |
+-----------------------------------+
| Route Type specific (variable)    |
+-----------------------------------+
```

The Route Type field defines encoding of the rest of the MCAST-TREE NLRI. (Route Type specific MCAST-TREE NLRI).

The Length field indicates the length in octets of the Route Type specific field of MCAST-TREE NLRI.

The following new route types are defined:

```
       3 -  S-PMSI A-D Route for (x,g)
       4 -  Leaf A-D Route
       5 -  Source Active A-D Route
    0x43 -  S-PMSI A-D Route for C-multicast mLDP
```

Except for the Source Active A-D routes, the routes are to be consumed by targeted upstream/downstream neighbors, and are not propagated further. This can be achieved by outbound filtering based on the RTs that lead to the importation of the routes.

The Type-3/4 routes MAY carry a Tunnel Encapsulation Attribute (TEA) [I-D.ietf-idr-tunnel-encaps]. The Type-0x43 route MUST carry a TEA. When used for mLDP, the Type-4 route MUST carry a TEA. Only the MPLS tunnel type for the TEA is considered. Others are outside the scope of this document.

2.1.1. S-PMSI A-D Route

Similar to defined in RFC 6514, an S-PMSI A-D Route Type specific MCAST-TREE NLRI consists of the following, though it does not have an RD:

```
+-----------------------------------+
| Multicast Source Length (1 octet) |
+-----------------------------------+
|  Multicast Source (variable)      |
+-----------------------------------+
|  Multicast Group Length (1 octet) |
+-----------------------------------+
|  Multicast Group   (variable)     |
+-----------------------------------+
|  Upstream Router's IP Address     |
+-----------------------------------+
```

If the Multicast Source (or Group) field contains an IPv4 address, then the value of the Multicast Source (or Group) Length field is 32. If the Multicast Source (or Group) field contains an IPv6 address, then the value of the Multicast Source (or Group) Length field is 128.

Usage of other values of the Multicast Source Length and Multicast Group Length fields is outside the scope of this document.

There are two usages for S-PMSI A-D route. They're described in Section 2.2.5 and Section 2.2.6 respectively.

2.1.2. Leaf A-D Route

Similar to the Leaf A-D route in [RFC6514], a MCAST-TREE Leaf A-D route's route key includes the corresponding S-PMSI NLRI, plus the Originating Router's IP Addr. The difference is that there is no RD.

```
      +-----------------------------------+
      |  S-PMSI NLRI                      |
      +-----------------------------------+
      |  Originating Router's IP Addrress |
      +-----------------------------------+
```

For example, the entire NLRI of a Leaf A-D route for (x,g) tree is as following:

```
      +-     +-----------------------------------+
      |      |    Route Type - 4 (Leaf A-D)      |
      |      +-----------------------------------+
      |      |     Length (1 octet)              |
      |   +- +-----------------------------------+ --+
      |   |  |    Route Type - 3 (S-PMSI A-D)    |   |
    L | L |  +-----------------------------------+   | S
    E | E |  |     Length (1 octet)              |   | |
    A | A |  +-----------------------------------+   | P
    F | F |  | Multicast Source Length (1 octet) |   | M
      |   |  +-----------------------------------+   | S
    N | R |  |  Multicast Source (variable)      |   | I
    L | O |  +-----------------------------------+   |
    R | U |  |  Multicast Group Length (1 octet) |   | N
    I | T |  +-----------------------------------+   | L
      | E |  |  Multicast Group   (variable)     |   | R
      |   |  +-----------------------------------+   | I
      | K |  |  Upstream Router's IP Address     |   |
      | E |  +-----------------------------------+ --+
      | Y |  |  Originating Router's IP Addrress |
      +-  +- +-----------------------------------+
```

Even though the MCAST-TREE Leaf A-D route is unsolicited, unlike the Leaf A-D route for GTM in [RFC7524], it is encoded as if a corresponding S-PMSI A-D route had been received.

When used for signaling mLDP tunnels, even though the Leaf A-D route is unsolicited, unlike the "Route-type 0x44 Leaf A-D route for C-multicast mLDP" as in [RFC7441], it is Route-type 4 and encoded as if a corresponding S-PMSI A-D route had been received.

2.1.3. Source Active A-D Route

Similar to defined in RFC 6514, a Source Active A-D Route Type specific MCAST NLRI consists of the following:

```
+----------------------------------+
| Multicast Source Length (1 octet) |
+----------------------------------+
|   Multicast Source (variable)    |
+----------------------------------+
|  Multicast Group Length (1 octet) |
+----------------------------------+
|  Multicast Group (variable)      |
+----------------------------------+
```

The definition of the source/length and group/length fields are the same as in the S-PMSI A-D routes.

Usage of Source Active A-D routes is described in Section 1.2.1.1.

2.1.4. S-PMSI A-D Route for C-multicast mLDP

The route is used to signal upstream FEC for an MP2MP mLDP tunnel. The route key include the mLDP FEC and the Upstream Router's IP Address field. The encoding is similar to the same route in [RFC7441], though there is no RD.

2.1.5. Session Address Extended Community

For two BGP speakers to determine if they are directly connected, each will advertise their local interface addresses, with an Session Address Extended Community. This is an Address Specific EC, with the Global Admin Field set to the local address used for its multihop sessions and the Local Admin Field set to the prefix length corresponding to the interface's network mask.

For example, if a router has two interfaces with address 10.10.10.1/24 and 10.12.0.1/16 respectively (notice the different network mask), and a loopback address 11.11.11.1/32 that is used for BGP sessions, then it will advertise prefix 10.10.10.1/32 with a Session Address EC 11.11.11.1:24 and 10.12.0.1/32 with a Session Address EC 11.11.11.1:16. If it also uses another loopback address 11.11.11.11/32 for other BGP sessions, then the routes will additionally carry Session Address EC 11.11.11.11:24 and 11.11.11.11:16 respectively.

This achieves what the Address List TLV in LDP Address Messages achieves, and can also be used to indicate that a router supports the BGP multicast signaling procedures specified in this document.

Only those interface addresses that will be used as resolved nexthops in the RIB need to be advertised with the Session Address EC. For example, the RPF lookup may say that the resolved nexthop address is

A1, so the router needs to find out the corresponding BGP speaker with address A1 through the (interface address, session address) mapping built according to the interface address NLRI with the Session Address EC. For comparison with LDP, this is done via the (interface address, session address) mapping that is built by the LDP Address Messages.

2.2. Procedures

2.2.1. Source Discovery for ASM

When a FHR first receives a multicast packet addressed to an ASM group, it originates a Source Active route. It carries a IP/IPv6 Address Specific RT, with the Global Admin Field set to the group address and the Local Admin Field set to 0. The route is advertised to its peers, who will re-advertise further based on the RTC mechanisms. Note that typically the route is advertised only to the RRs.

The FHRs withdraws the Source Active route after a certain amount of time since it last received a packet of an (s,g) flow. The amount of time to wait is a local matter.

2.2.2. Originating Tree Join Routes

Note that in this document, tree join routes are S-PMSI/Leaf A-D routes.

2.2.2.1. (x,g) Multicast Tree

When a router learns from IGMP/MLD or a downstream PIM/BGP peer that it needs to join a particular (s,g) tree, it determines the RPF nexthop address wrt the source, following the same RPF procedures as defined for PIM. It further finds the BGP router that advertised the nexthop address as one of its local addresses.

If the RPF neighbor supports MCAST-TREE SAFI, this router originates a Leaf A-D route. Although it is unsolicited, it is constructed as if there was a corresponding S-PMSI A-D route. The Upstream Router's IP Address field is set to the RPF neighbor's session address (learnt via the EC attached to the host route for the RPF nexthop address). An Address Specific RT corresponding to the session address is attached to the route, with the Global Administrative Field set to the session address and the local administrative field set to 0.

Similarly, when a router learns that it needs to join a bi-directional tree for a particular group, it determines the RPF neighbor wrt the RPA. If the neighbor supports MCAST-TREE SAFI, it

originates a Leaf A-D Route and advertises the route to the RPF neighbor (in case of EBGP or hop-by-hop IBGP), or one or more RRs.

When a router first learns that it needs to receive traffic for an ASM group, either because of a local (*,g) IGMP/MLD report or a downstream PIM (*,g) join, it originates a RTC route with the NLRI's AS field set to its AS number and the Route Target field set to an address based RT, with the Global Administrator field set to group address and the Local Administrator field set to 0. The route is advertised to its peers (most practically some RRs), so that the router can receive matching Source Active A-D routes. Upon the receiving of the Source Active A-D routes, the router originates Leaf A-D routes as described above, as long as it still needs to receive traffic for the flows (i.e., the corresponding IGMP/MLD membership exists or join from downstream PIM/BGP neighbor exists).

When a Leaf A-D route is originated by this router, it sets up corresponding forwarding state such that the expected incoming interface list includes all non-LAN interfaces directly connecting to the upstream neighbor. LAN interfaces are added upon receiving corresponding S-PMSI A-D route (Section 2.2.5.2). If the upstream neighbor is not directly connected, tunnels may be used - details to be included in future revisions.

When the upstream neighbor changes, the previously advertised Leaf A-D route is withdrawn. If there is a new upstream neighbor, a new Leaf A-D route is originated, corresponding to the new neighbor. Because NLRIs are different for the old and new Leaf A-D routes, make-before-break as well as MoFRR [RFC7431] can be achieved.

2.2.2.2. BGP Inband Signaling for mLDP Tunnel

The same mLDP procedures as defined in [RFC6388] are followed, except that where a label mapping message is sent in [RFC6388], a Leaf A-D route is sent if the the upstream neighbor supports BGP based signaling.

2.2.3. Receiving Tree Join Routes

A router (auto-)configures Import RTs matching itself so that it can import tree join routes from their peers. Note that in this document, tree join routes are S-PMSI/Leaf A-D routes.

When a router receives a tree join route and imports it, it determines if it needs to originate its own corresponding route and advertise further upstream wrt the source/RPA or mLDP tunnel root. If this router is the FHR or is on the RPL or is the tunnel root,

then it does not need to. Otherwise the procedures in Section 2.2.2 are followed.

Additionally, the router sets up its corresponding forwarding state such that traffic will be sent to the downstream neighbor, and received from the downstream neighbor in case of birectional tree/ tunnel. If the downstream neighbor is not directly connected, tunnels may be used - details to be included in future revisions.

2.2.4. Withdrawl of Tree Join Routes

For a particular tree or tunnel, if a downstream neighbor withdraws its Leaf A-D route, the neighbor is removed from the corresponding forwarding state. If all downstream neighbors withdraw their tree join routes and this router no longer has local receivers, it withdraws the tree join routes that it previously originated.

As mentioned earlier, when the upstream neighbor changes, the previously advertised Leaf A-D route is also withdrawn. The corresponding incoming interfaces are also removed from the corresponding forwarding state.

2.2.5. LAN procedures for (x,g) Unidirectional Tree

For a unidirectional (x,g) multicast tree, if there is a LAN interface connecting to the downstream neighbor, it MAY be preferred over non-LAN interfaces, but an S-PMSI A-D route MUST be originated to facilitate the analog of the Assert process (Section 2.2.5.1).

2.2.5.1. Originating S-PMSI A-D Routes

If this router chooses to use a LAN interface to send traffic to its neighbors for a particular (s,g) or (*,g) flow, it MUST announce that by originating a corresponding S-PMSI A-D route. The Tunnel Type in the PMSI Tunnel Attribute (PTA) is set to 0 (no tunnel information Present). The LAN interface is identified by an IP address specific RT, with the Global Administrative Field set to the LAN interface's address prefix and the Local Administrative Field set to the prefix length. The RT also serves the purpose of restricting the importing of the route by all routers on the LAN. An operator MUST ensure that RTs encoded as above are not used for other purposes. Practically that should not be unreasonable.

If multiple LAN interfaces are to be used (to reach different sets of neighbors), then the route will include multiple RTs, one for each used LAN interface as described above.

The S-PMSI A-D routes may also be used to announce tunnels that could be used to send traffic to downstream neighbors that are not directly connected. Details may be added in future revisions.

2.2.5.2. Receiving S-PMSI A-D Routes

A router (auto-)configures an Import RT for each of its LAN interfaces over which BGP is used for multicast signaling. The construction of the RT is described in the previous section.

When a router R1 imports an S-PMSI A-D route for flow (x,g) from router R2, R1 checks to see if it also originating an S-PMSI A-D route with the same NLRI except the Upstream Router's IP Address field. When a router R1 originates an S-PMSI A-D route, it checks to see if it also has installed an S-PMSI A-D route, from some other router R2, with the same NLRI except the Upstream Router's IP Address field. In either case, R1 checks to see if the two routes have an RT in common and the RT is encoded as in Section 2.2.5.1. If so, then there is a LAN attached to both R1 and R2, and both routers are prepared to send (S,G) traffic onto that LAN. This kicks off the assert procedure to elect a winner - the one with the highest Upstream Router's IP Address in the NLRI wins. An assert loser will not include the corresponding LAN interface in its outgoing interface list, but it keeps the S-PMSI A-D route that it originates.

If this router does not have a matching S-PMSI route of its own with some common RTs, and the originator of the received S-PMSI route is a chosen upstream neighbor for the corresponding flow, then this router updates its forwarding state to include the LAN interface in the incoming interface list. When the last S-PMSI route with a RT matching the LAN is withdrawn later, the LAN interface is removed from the incoming interface list.

Note that a downstream router on the LAN does not participate in the assert procedure. It adds/keeps the LAN interface in the expected incoming interfaces as long as its chosen upstream peer originates the S-PMSI AD route. It does not switch to the assert winner as its upstream. An assert loser MAY keep sending joins upstream based on local policy even if it has no other downstream neighbors (this could be used for fast switch over in case the assert winner would fail).

2.2.6. Distributing Label for Upstream Traffic for Bidirectional Tree/ Tunnel

For MP2MP mLDP tunnels or labeled (*,g) bidirectional trees, an upstream router needs to advertise a label to all its downstream neighbors so that the downstream neighbors can send traffic to itself.

For MP2MP mLDP tunnels, the same procedures for mLDP are followed except that instead of MP2MP-U Label Mapping messages, S-PMSI A-D Routes for C-Multicast mLDP are used.

For labeled (*,g) bidirectional trees, for a Leaf A-D route received from a downstream neighbor, a corresponding S-PMSI A-D route is sent back to the downstream router.

In both cases, a single S-PMSI A-D route is originated for each tree from this router, but with multiple RTs (one for each downstream neighbor on the tree). A TEA specifies a label allocated by the upstream router for its downstream neighbors to send traffic with. Note that this is still a "downstream allocated" label (the upstream router is "downstream" from traffic direction point of view).

The S-PMSI routes do not carry a PTA, unless a P2MP tunnel is used to reach downstream neighbors. Such use case is out of scope of this document for now and may be specified in the future.

3. Security Considerations

This document does not introduce new security risks.

4. Acknowledgements

The authors thank Marco Rodrigues for his initial idea/ask of using BGP for multicast signaling beyond MVPN. We thank Eric Rosen for his questions, suggestions, and help finding solutions to some issues. We also thank Luay Jalil and James Uttaro for their comments and support for the work.

5. References

5.1. 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-14 (work in progress), September 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>.

[RFC4601] Fenner, B., Handley, M., Holbrook, H., and I. Kouvelas, "Protocol Independent Multicast - Sparse Mode (PIM-SM): Protocol Specification (Revised)", RFC 4601, DOI 10.17487/RFC4601, August 2006, <https://www.rfc-editor.org/info/rfc4601>.

[RFC4684] Marques, P., Bonica, R., Fang, L., Martini, L., Raszuk, R., Patel, K., and J. Guichard, "Constrained Route Distribution for Border Gateway Protocol/MultiProtocol Label Switching (BGP/MPLS) Internet Protocol (IP) Virtual Private Networks (VPNs)", RFC 4684, DOI 10.17487/RFC4684, November 2006, <https://www.rfc-editor.org/info/rfc4684>.

[RFC5015] Handley, M., Kouvelas, I., Speakman, T., and L. Vicisano, "Bidirectional Protocol Independent Multicast (BIDIR-PIM)", RFC 5015, DOI 10.17487/RFC5015, October 2007, <https://www.rfc-editor.org/info/rfc5015>.

[RFC6514] Aggarwal, R., Rosen, E., Morin, T., and Y. Rekhter, "BGP Encodings and Procedures for Multicast in MPLS/BGP IP VPNs", RFC 6514, DOI 10.17487/RFC6514, February 2012, <https://www.rfc-editor.org/info/rfc6514>.

[RFC7441] Wijnands, IJ., Rosen, E., and U. Joorde, "Encoding Multipoint LDP (mLDP) Forwarding Equivalence Classes (FECs) in the NLRI of BGP MCAST-VPN Routes", RFC 7441, DOI 10.17487/RFC7441, January 2015, <https://www.rfc-editor.org/info/rfc7441>.

5.2. Informative References

[I-D.ietf-bess-mvpn-pe-ce]
Patel, K., Rosen, E., and Y. Rekhter, "BGP as an MVPN PE-CE Protocol", draft-ietf-bess-mvpn-pe-ce-01 (work in progress), October 2015.

[I-D.wijnands-bier-mld-lan-election]
Wijnands, I., Pfister, P., and Z. Zhang, "Generic Multicast Router Election on LAN's", draft-wijnands-bier-mld-lan-election-01 (work in progress), July 2016.

[RFC6826] Wijnands, IJ., Ed., Eckert, T., Leymann, N., and M. Napierala, "Multipoint LDP In-Band Signaling for Point-to-Multipoint and Multipoint-to-Multipoint Label Switched Paths", RFC 6826, DOI 10.17487/RFC6826, January 2013, <https://www.rfc-editor.org/info/rfc6826>.

[RFC7431] Karan, A., Filsfils, C., Wijnands, IJ., Ed., and B. Decraene, "Multicast-Only Fast Reroute", RFC 7431, DOI 10.17487/RFC7431, August 2015, <https://www.rfc-editor.org/info/rfc7431>.

[RFC7938] Lapukhov, P., Premji, A., and J. Mitchell, Ed., "Use of BGP for Routing in Large-Scale Data Centers", RFC 7938, DOI 10.17487/RFC7938, August 2016, <https://www.rfc-editor.org/info/rfc7938>.

Authors' Addresses

Zhaohui Zhang
Juniper Networks

EMail: zzhang@juniper.net

Lenny Giuliano
Juniper Networks

EMail: lenny@juniper.net

Keyur Patel
Arrcus

EMail: keyur@arrcus.com

IJsbrand Wijnands
Cisco Systems

EMail: ice@cisco.com

Mankamana Mishra
Cisco Systems

EMail: mankamis@cisco.com

Arkadiy Gulko
Refinitiv

EMail: arkadiy.gulko@refinitiv.com