Workgroup: BESS
Internet-Draft:
draft-ietf-bess-bgp-multicast-07
Published: 2 December 2023
Intended Status: Standards Track
Expires: 4 June 2024
Authors: Z. Zhang
Juniper Networks
L. Giuliano
Juniper Networks
K. Patel
Arrcus
I. Wijnands
Arrcus
M. Mishra
Cisco Systems
A. Gulko
EdwardJones

BGP Based Multicast

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.

This document also describes how various signaling mechanisms can be used to set up end-to-end inter-region multicast trees.

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/.

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 4 June 2024.

Copyright Notice

Table of Contents

1. Introduction

1.1. Terminology

This document assumes the readers are familiar with basic multicast concepts. Some terminologies are included here for convenience.

PIM: Protocol Independent Multicast [RFC7761].

ASM: All/Any Source Multicast. A multicast mode where a receiver is interested in receiving traffic for a multicast group from anywhere.

SSM: Source-Specific Multicast [RFC7761].

PIM-ASM: PIM procedures for ASM.

PIM-SSM: PIM procedures for SSM.

PIM-Port: PIM Over Reliable Transportation [RFC6559].

PIM-Bidir: PIM procedures for bidirectional shared trees connecting multicast sources and receivers [RFC5015].

RP: Rendezvous Point, the root of the non-source-specific distribution tree for a multicast group [RFC7761].

RPA: Rendezvous Point Address for the root of a bidirectional distribution tree for a range of multicast groups [RFC5015].

RPL: Rendezvous Point Link, the link to which the RPA belongs.

RPF: Reverse Path Forwarding, by which multicast traffic is forwarded from the root of a multicast tree in the reverse path on which the tree leaves would reach the root.

FHR: First Hop Router (connecting to a multicast source) [RFC7761].

LHR: Last Hop Router (connecting to a multicast receiver) [RFC7761].

(x,G): An IP multicast flow/tree for a group G, where x is either an 'S' for a specific source or a '*' for all sources.

IGMP: Internet Group Management Protocol [RFC3376].

MLD: Multicast Listener Discovery [RFC3810].

P2MP: Point-to-MultiPoint.

MP2MP: MultiPoint-to-MultiPoint.

mLDP: Label Distribution Protocol Extensions for P2MP and MP2MP Label Switched Paths [RFC6388].

S-PMSI: Selective Provider Multicast Service Interface. In [RFC6514], the term refers to a pseudo-interface used to send customer multicast to a subset of Provider Edge routers, and S-PMSI Auto-Discovery (A-D) routes are used to advertise the binding of customer multicast flows to a tunnel that instantiates a pseudo-interface. In this document, S-PMSI A-D routes are used for various purposes as described in Section 2.2.5 and Section 2.2.6.

PTA: PMSI Tunnel Attribute [RFC6514]. An attribute carried in an S-PMSI A-D route that specifies the tunnel instantiating the PMSI.

EC: BGP Extended Community.

RT: BGP Route Target.

RTC: Route Target Constrain (RTC) [RFC4684].

1.2. Motivation

This section provides some motivation for BGP signaling for native and labeled multicast. One target deployment would be a Data Center

(DC) 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, 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.

1.2.1. Native/unlabeled Multicast

Protocol Independent Multicast (PIM) [RFC7761] has been the prevailing multicast protocol for many years. Despite its success, it has two drawbacks:

- 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.

- 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 Over Reliable Transport (PORT) [RFC6559] solves the soft state issue, though its deployment has also been limited for two reasons:

- It does not remove the ASM complexities.

- 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 Service Provider (SP) environment and has been widely deployed. While RFC 6514 enables an SP to provide MVPN service without running PIM on its backbone, it assumes that PIM (or mLDP) runs on the PE-CE links. [I-D.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 remain 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 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 set up with MPLS labels, with just minor enhancements in the signaling.

1.2.2. Labeled Multicast

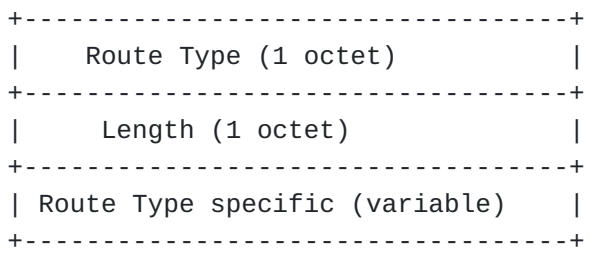
There could be two forms of labeled multicast signaled by BGP. The first one is labeled (x,G) multicast where 'x' stands for either 'S' or '*'. Basically, it is for a BGP-signaled multicast tree as described in the previous section but with labels. The second one is for mLDP tunnels with BGP signaling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

1.3. Overview

This overview section describes the mode of operation and some considerations.

At a very high level, PIM Join messages or mLDP Label mapping messages are replaced by BGP updates of MCAST-TREE SAFI with the following NLRI format (Section 2.1):



Different route types are described in this section as they are encountered.

1.3.1. (x,G) Multicast

PIM/mLDP-like functionality is provided, using BGP-based join signaling and BGP-based source discovery in the case of ASM. The BGP-based join signaling supports both labeled multicast and IP multicast.

The same RPF procedures as in PIM/mLDP are used for each router to determine the RPF neighbor for a particular source or RPA (in the

case of Bidirectional Tree) or root. Except in the Bidirectional Tree case and a special case described in Section 1.3.1.2, 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 are replaced by control-driven mechanisms (Section 1.3.4).

One of the route types is Leaf A-D route - the equivalent of a PIM Join message or mLDP Label Mapping message. The Leaf A-D routes are targeted at the upstream neighbor by use of Route Targets. In some cases, S-PMSI A-D routes are also used, as described in some sections below.

If the BGP updates carry labels (via Tunnel Encapsulation Attribute [RFC9012]), 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 Local Area Networks (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) is 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 in a Leaf A-D route is the same one that it advertises to its downstreams using an S-PMSI A-D route.

1.3.1.1. 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.

An FHR originates Source Active (SA) A-D routes upon discovering sources for particular flows and advertises them to its peers. Route targets are used so that the SA routes only reach LHRs that are interested in receiving the traffic (Section 2.2.1).

Typically, a set of RRs are used and they maintain all Source Active routes but only distribute to interested LHRs on demand. The rest of the document assumes that RRs are used, even though that is not required.

That the set of RRs maintain all SA routes is comparable to that the RPs in PIM-ASM maintain all (S,G) states in the network. In fact, in PIM-ASM case the states are maintained in both the control plane and data plane, while in the case of BGP SA-based discovery, the states

are only maintained in the control plane, and the RRs can be placed outside the traffic path.

Note that the data-driven source discovery and subsequent control-driven tree setup means receivers will miss the initial packets of a multicast flow when it just starts or resumes. If it is important to avoid this, source discovery should be provided by the application layer instead of the network.

1.3.1.2. 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 the RP. In the latter case, either the paths 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 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 originated hop-by-hop towards the RP, and corresponding (*,G) forwarding states are established along the way, just like how (S,G) Leaf A-D routes are originated hop-by-hop towards the source and (S,G) forwarding sates are established along the way.

1.3.1.3. 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 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 on PE-CE interfaces.

1.3.2. BGP Inband Signaling for mLDP Tunnels

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 in the BGP NLRI, with MCAST-TREE SAFI and S-PMSI/Leaf A-D Routes for 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 bidirectional (*,G) case.

1.3.3. 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 (78).

How the BGP peer sessions are provisioned, whether EBGP or IBGP, whether statically, automatically, or programmably via an external controller, is outside the scope of this document.

In the 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 Extended Community (EC) corresponding to the router's local address 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 MCAST-TREE NLRIs.

FHRs and LHRs may also establish BGP sessions to some Route Reflectors for source discovery purposes (Section 1.3.1.1).

With the traditional PIM, the FHRs and LHRs refer to the PIM Desginated Routers (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 Desginated Forwarders (DFs) 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.

1.3.4. 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 the 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 purposes. 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.

There may be multiple LAN interfaces between two neighbors, and the upstream neighbor may send traffic on both LAN interfaces because of other downstream neighbors on both LANs. In this case, a downstream neighbor will choose one of the LANs to receive traffic - the RTs in the S-PMSI route enables the downstream neighbor to determine that its upstream neighbor is sending on both interfaces and it will only choose one on which to receive traffic.

1.3.5. 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 purposes, and routers connected to the same LAN MUST be transitioned during the same maintenance window.

In the case of PIM-SSM, any router can be transitioned at any time (except on a LAN). 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 the 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.

1.3.6. Inter-region Multicast

An end-to-end multicast tree or P2MP tunnel may span multiple regions, where a region could be an IGP area (or even a sub-area) or an Autonomous System (AS), and different multicast signaling could be used in different regions. There are several situations to consider.

1.3.6.1. Inband Signaling across a Region

With inband signaling, the multicast tree/tunnel is signaled through a region and internal routers in the region maintain corresponding per-tree/tunnel state. A downstream region and an upstream region may use the same or different signaling. For example, a (*/s, G) IP multicast tree with BGP signaling in a downstream region can be signaled with mLDP Inband Signaling [RFC6826] or with PIM across the

upstream region, and a p2mp tunnel with BGP signaling in the downstream region can be signaled with mLDP across the upstream region, or vice versa. A Regional Border Router (RBR) will stitch the upstream portion (e.g. PIM/mLDP-signaled) to the downstream portion (e.g. BGP-signaled).

If all routers in the region have routes towards the source/root of the tree/tunnel then there is nothing different from the intra-region case. On the other hand, if internal routers do not have routes towards the source/root, e.g. as with Seamless MPLS [I-D.ietf-mpls-seamless-mpls], the internal routers need to do RPF towards an upstream RBR. To signal the RBR information to an internal upstream router, one of the following ways is used depending on the signaling method:

*With BGP signaling, the Leaf A-D Route carries a new BGP Extended Community referred to as Multicast RPF Address EC, similar to PIM RPF Vector [RFC5496] and mLDP Recursive FEC [RFC6512].

*With PIM signaling, PIM RPF Vector is used.

*With mLDP signaling, mLDP Recursive FEC is used.

1.3.6.2. Overlay Signaling Over a Region

With overlay signaling, a downstream RBR signals to its upstream RBR over the region and the internal routers do not maintain the state of the (overlay) tree/tunnel. This can be done with one of the following methods:

*mLDP P2MP tunnels can be signaled over the region via targeted LDP sessions [RFC7060].

*Both IP multicast tree and mLDP P2MP tunnels can be signaled over a region via BGP-MVPN procedures [RFC6514].

*Both IP multicast tree and mLDP P2MP tunnels can be signaled over a region via BGP as discussed in the rest of this section.

All three methods are actually very similar in concept. The upstream RBR tunnels packets to the downstream RBR, just as in the intra-region case when two routers on the tree/tunnel are not directly connected. The rest of this section only discusses BGP signaling.

When a downstream RBR determines that the route towards the source/root has a BGP Next Hop towards a BGP speaker capable of multicast signaling via BGP as specified in this document, it signals to that BGP speaker (via a RR or not).

Suppose an upstream RBR receives the signaling for the same tree/ tunnel from several downstream RBRs. It could use Ingress Replication to replicate packets directly to those downstream RBRs, or it could use underlay P2MP tunnels instead.

In the latter case, the upstream RBR advertises an S-PMSI A-D route with a PMSI Tunnel Attribute (PTA) specifying the underlay tunnel. This is very much like the "mLDP Over Targeted Sessions" [RFC7060] or BGP-MVPN [RFC6514] (though MCAST-VPN's C-Multicast routes are replaced with MCAST-TREE's Leaf A-D routes). If the mapping between overlay tree/tunnel and underlay tunnel is one-to-one, the MPLS Label field in the PTA is set to 0 or otherwise set to a Domain-wide Common Block (DCB) label [I-D.ietf-bess-mvpn-evpn-aggregation-label] or an upstream-assigned label corresponding to the overlay tree/tunnel.

1.3.6.3. Controller Based Signaling

[I-D.ietf-bess-bgp-multicast-controller] specifies the procedures for a controller to signal multicast forwarding state to each router on a multicast tree based on the controller's computation. Depending on deployment scenarios, in inter-region cases it is possible that the hop-by-hop signaling specified in this document and the controller based signaling may be used in different regions.

Consider a situation where an RBR is connected to three regions A, B, and C, where hop-by-hop signaling is used in A and B, while controller based signaling is used in C.

For a particular multicast tree, A is the upstream region, while B and C are two downstream regions. The RBR receives a Leaf A-D route from region B and a Leaf A-D route from C's controller, and sends a Leaf A-D route to its upstream router in A.

For a different tree, C is the upstream region while A and B are downstream. The RBR receives two Leaf A-D routes for the tree from regions A and B, and one Leaf A-D route from C's controller. Note that the RBR needs to signal to the controller that it is a leaf of the tree (because of the Leaf A-D routes received from regions A and B).

For both cases, the RBR stitches together different segments in different regions by creating forwarding state based on the Leaf A-D routes (optionally based on the S-PMSI A-D routes in region A and B in addition.)

1.3.7. BGP Classful Transport Planes

[I-D.ietf-idr-bgp-ct] specifies an experimental framework for classifying underlay routes into transport classes and mapping service routes to specific transport classes. An underlay route

signaled with BGP-CT SAFI carries a Transport Class Route Target (TC-RT) to both indicate the transport class that the route belongs to and to control the propagation and importation of the underlay route. The recipients of the underlay routes use the TC-RT to determine how the Protocol NH (PNH) is resolved. A service/overlay route may carry a mapping community that maps to a transport class that is used to resolve the service route's PNH.

In the case of multicast, the selection of the link/tunnel between an upstream and downstream tree node may be subject to the transport class that the tree is for (in the case of an underlay tree) or the class of transport that the tree should use (in the case of an overlay tree). In both the underlay and overlay case, the transport class is indicated by a mapping community attached to the BGP multicast routes, which could be a color community or any community intended for mapping to the transport.

The mapping community not only affects an upstream node's selection of link/tunnel to a downstream node, but may also affect a downstream node's selection of its upstream node (i.e. the RPF procedure).

[I-D.ietf-idr-bgp-car] is another experimental mechanism that provides class/color-aware routing. Multicast signaled by BGP may be integrated with that as well, but it is outside the scope of this document.

1.3.8. Flexible Algorithm and Multi-topology

Similar to classful transport, in the case of multi-topology [RFC4915] [RFC5120] or Flexible Algorithm [RFC9350], a multicast tree may be required to do RPF based on a particular topology or Flexible Algorithm (IPA). To signal that, the BGP-MCAST Leaf A-D route may carry an extended community to encode the topology and/or IPA. Note that this could also be an operator-defined mapping community that maps to a transport class (that is associated with a topology or a Flexible Algorithm).

In the grand scheme of inter-region scenario, if mLDP is to be used with Flexible Algorithm or Multi-topology for signaling in a particular region, [I-D.ietf-mpls-mldp-multi-topology] specifies how topology and/or IPA are encoded.

Similarly, in the case of PIM, [RFC6420] specifies how topology information is encoded in PIM signaling and a similar mechanism can be specified for Flexible Algorithm. However, that, and potentially encoding transport class in PIM/mLDP are outside the scope of this document.

2. Specification

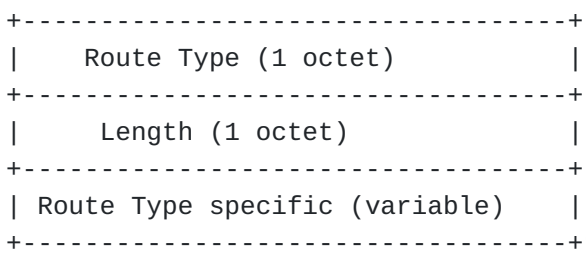
2.1. BGP NLRIs and Attributes

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 value 78 assigned by the IANA.

The MCAST-TREE NLRI defined below is carried in the BGP UPDATE messages [RFC4271] using the BGP multiprotocol extensions [RFC4760] with an AFI of IPv4 (1) or IPv6 (2) and a MCAST-TREE SAFI (78).

The Next hop field of MP_REACH_NLRI attribute SHALL be interpreted as an IPv4 address 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 octets for IPv4 AFI and 36 octets for IPv6 AFI. The following is the format of the MCAST-TREE NLRI:



The Route Type field defines the encoding of the rest of the 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 mLDP
```

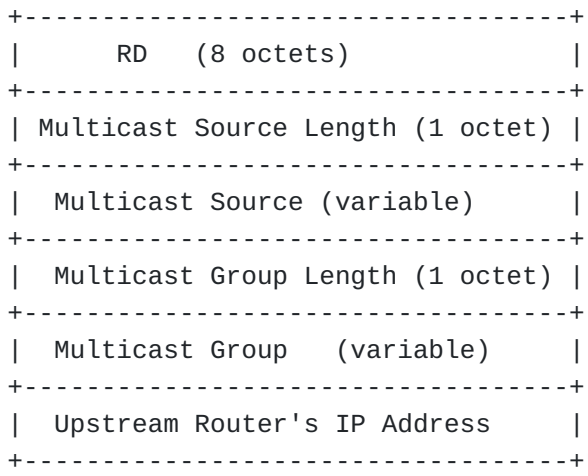
Except for the Source Active A-D routes, the routes are to be consumed by targeted upstream/downstream neighbors and are not propagated further. This can be achieved by outbound filtering based on the RTs that lead to the importation of the routes.

The Type-3/4 routes MAY carry a Tunnel Encapsulation Attribute (TEA) [RFC9012]. The Type-0x43 route MUST carry a TEA. When used for mLDP, the Type-4 route MUST carry a TEA. The TEA includes one tunnel entry

with an MPLS Label Stack Sub-TLV that includes one label. This is the label associated with the (x,G) labeled tree or mLDP tunnel.

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:



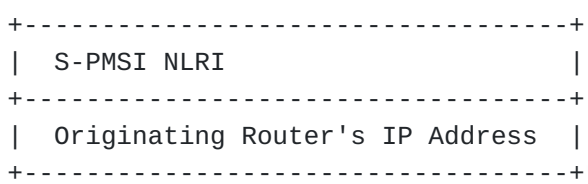
If the Multicast Source (or Group) field contains an IPv4 address, then the value of the Multicast Source (or Group) Length field is 32. If the Multicast Source (or Group) field contains an IPv6 address, then the value of the Multicast Source (or Group) Length field is 128.

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

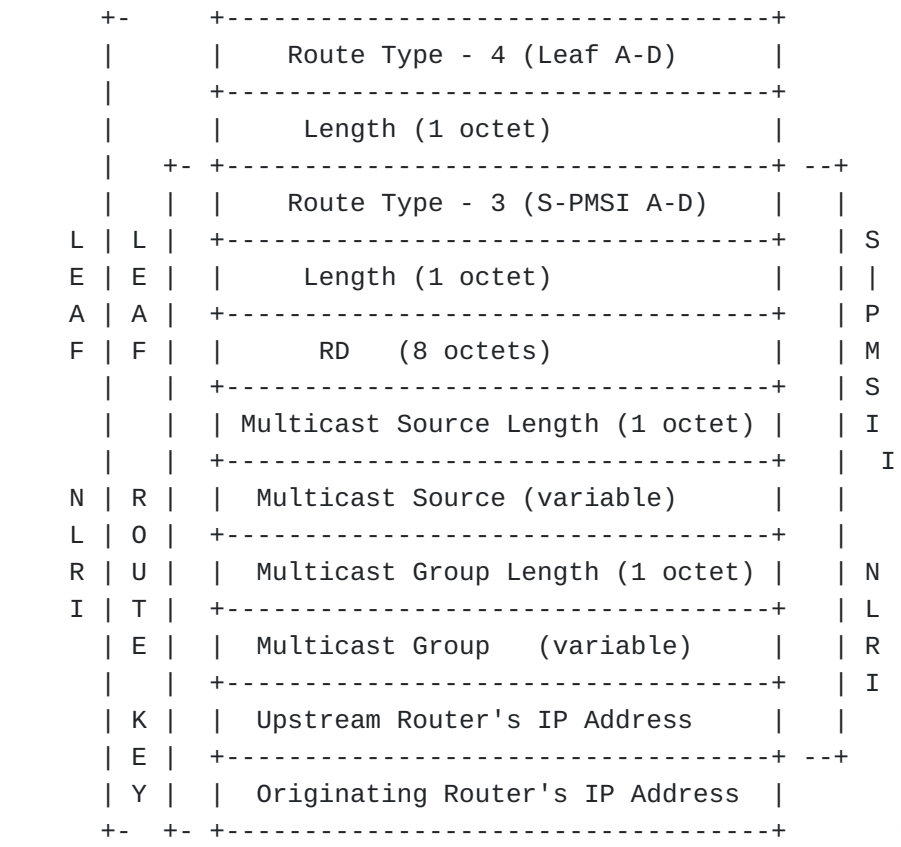
There are three usages for S-PMSI A-D route. They're described in Section 1.3.6.2, Section 2.2.5 and Section 2.2.6 respectively.

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 Address.

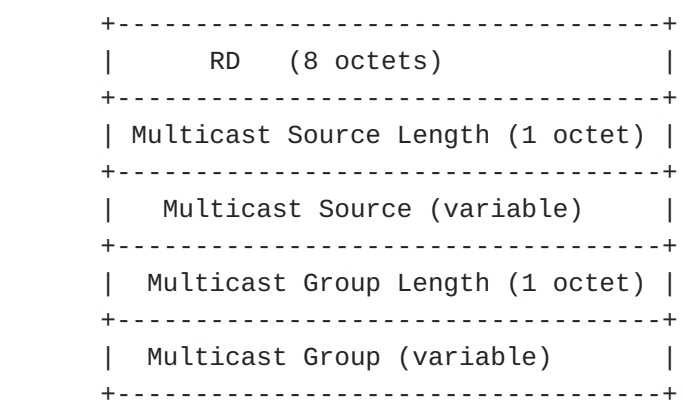


For example, the entire NLRI of a Leaf A-D route for (x,G) tree is as following:



2.1.3. Source Active A-D Route

Similar to what is defined in [RFC6514], a Source Active A-D Route Type specific MCAST NLRI consists of the following:



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 1.3.1.1.

2.1.4. S-PMSI A-D Route for mLDP

The route is used to signal upstream FEC for an MP2MP mLDP tunnel. The route key include a Route Distinguisher, the mLDP FEC and the Upstream Router's IP Address field.

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 a Session Address Extended Community. This is an IPv4/IPv6 Address Specific EC with the Global Administrator Field set to the local address used for its multihop sessions and the Local Administrator Field set to the prefix length corresponding to the interface's network mask.

As an IPv4 example, a router has two interfaces with address 192.0.2.1/28 and 198.51.100.1/24 respectively (notice the different prefix lengths), and a loopback address 203.0.113.1/32 that is used for BGP sessions. It advertises prefix 192.0.2.1/32 with a Session Address EC 203.0.113.1:28 and 198.51.100.1/32 with a Session Address EC 203.0.113.1:24. If it also uses another loopback address 203.0.113.101/32 for other BGP sessions, then the routes will additionally carry Session Address EC 203.0.113.101:28 and 203.0.113.101:24 respectively.

As an IPv6 example, a router has two interfaces with address 2001:DB8::1:1/112 and 2001:DB8::2:1/120 respectively (notice the different prefix lengths), and a loopback address 2001:DB8::3:1/128 that is used for BGP sessions. It advertises prefix 2001:DB8::1:1/128 with a Session Address EC "2001:DB8::3:1":112 (the quoted part is the IPv6 address for the Global Administrator field and the "112" is the Local Administrator field) and prefix 2001:DB8::2:1/128 with a Session Address EC "2001:DB8::3:1":120. If it also uses another loopback address 2001:DB8:3:101/128 for other BGP sessions, then the routes will additionally carry Session Address EC "2001:DB8::3:101": 112 and "2001:DB8::3:101":120 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 RPF 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, in LDP this is done via the

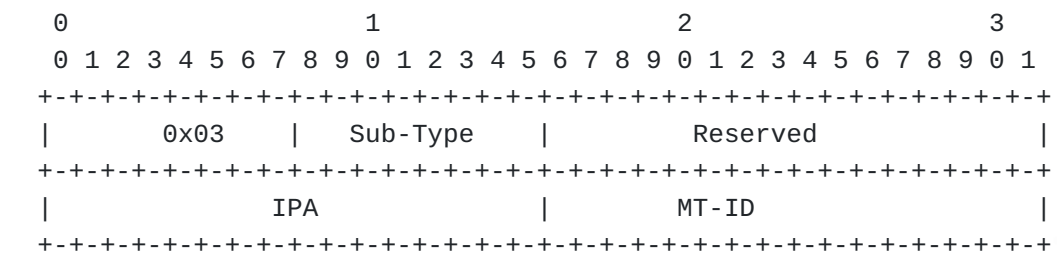
(interface address, session address) mapping that is built by the LDP Address Messages.

2.1.6. Multicast RPF Address Extended Community

This is an IP or IPv6 Address Specific EC with the Global Admin Field set to the address of the upstream RBR and the Local Admin Field set to 0.

2.1.7. Topology/IPA Extended Community

This is a Transitive Opaque Extended Community with the following format:



IPA is the Flexible Algorithm number and MT-ID is the Multi-Topology Identifier to be used for setting a multicast tree. The usage of this EC is described in Section 1.3.8.

2.2. Procedures

2.2.1. Source Discovery for ASM

When an FHR first receives a multicast packet addressed to an ASM group, it originates a Source Active route.

The FHRs withdraw 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.

The SA routes carry an IPv4 or IPv6 address specific Route Target. The Global Administrator field is set to the group address of the flow, and the Local Administrator field is set to 0 or a pre-assigned domain-wide unique value that identifies a VPN.

When an LHR needs to join an ASM group (e.g., as the result of receiving a (*,G) IGMP/MLD join), it advertises a Route Target Membership route, with the Route Target field in the NLRI set according to the group, as how an FHR encodes 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 Route Target Constrain (RTC) mechanism in this document. With

that, the LHR only receives Source Active routes for groups that it is interested in.

Upon the receiving of the Source Active A-D routes, the LHR originates Leaf A-D routes as described below, 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).

2.2.2. Originating Tree Join Routes

2.2.2.1. (x,G) Multicast Tree

When a router 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 or a pre-assigned domain-wide unique value that identifies a VPN. The route is advertised the route to the RPF neighbor (in the case of EBGP or hop-by-hop IBGP), or to one or more RRs.

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.

As the Leaf A-D route is originated, the router sets up the 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, a corresponding S-PMSI A-D route advertised by the upstream router is used to determine the tunnel used to receive traffic, as described in Section 1.3.6.2

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 Multicast Only Fast ReRroute (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 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 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 for a bidirectional group, 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 the case of bidirectional tree/tunnel. If the downstream neighbor is not directly connected, the tunnel announced in a corresponding S-PMSI route is used, as described in Section 1.3.6.2.

2.2.4. Withdrawal 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 that does not include a PTA. 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.

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 originats 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 switchover in case the assert winner fails).

If this router receives an S-PMSI A-D route from its upstream neighbor with multiple RTs for the LANs that this router is on, it MUST select only one of the LAN interfaces to receive traffic. Which LAN interface is selected is a local decision.

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 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 tunnel is used to reach downstream neighbors as described in Section 1.3.6.2.

3. IANA Considerations

IANA has assigned BGP SAFI value 78 for the MCAST-TREE SAFI.

This document requests IANA to create a new "BGP MCAST-TREE Route Types" registry, referencing this document. The following initial values are defined:

```
    0~2 -  Reserved
    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 mLDP
```

This document requests IANA to assign two Sub-type values from Transitive IPv4-Address-Specific Extended Community Sub-types Registry for Session Address EC and Multicast RPF Address EC respectively.

This document requests IANA to assign two Sub-Type values from Transitive IPv6-Address-Specific Extended Community Types Registry for Session Address EC and Multicast RPF Address EC respectively.

This document requests IANA to assign one Sub-Type value from Transitive Opaque Extended Community Types Registry for the Topology/ IPA EC.

4. Security Considerations

This document shares many of the mechanisms and concepts of MVPN and, accordingly, can reuse many of the security considerations described in RFC6513 and RFC6514, though the distinctions made on PE-CE links and relationships in those documents are not relevant.

This document describes interworking with several multicast control protocols, including PIM-SM, PIM-SSM, PIM-Bidir, mLDP and IGMP/MLD. Security considerations specified for those protocols are applicable to this document.

Implementations should include Multicast Damping procedures specified in RFC7899 to protect the control plane from excessive churn due to multicast dynamicity. Implementations should also include the ability to rate-limit join state creation on a per-peer and per-RIB basis, as well as rate-limit Source Active A-D route propagation on a per-source, per-peer and per-RIB basis to configurable thresholds.

5. 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 to find solutions to some issues. We also thank Luay Jalil, James Uttaro and Shraddha Hegde for their comments and support for the work. Special thanks go to Joe Halpern for his thorough review and comments that significantly improved the document quality.

6. References

6.1. 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>.

[RFC4271] Rekhter, Y., Ed., Li, T., Ed., and S. Hares, Ed., "A Border Gateway Protocol 4 (BGP-4)", RFC 4271, DOI 10.17487/RFC4271, January 2006, <https://www.rfc-editor.org/info/rfc4271>.

[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>.

[RFC4760] Bates, T., Chandra, R., Katz, D., and Y. Rekhter, "Multiprotocol Extensions for BGP-4", RFC 4760, DOI 10.17487/RFC4760, January 2007, <https://www.rfc-editor.org/info/rfc4760>.

[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>.

[RFC5492] Scudder, J. and R. Chandra, "Capabilities Advertisement with BGP-4", RFC 5492, DOI 10.17487/RFC5492, February 2009, <https://www.rfc-editor.org/info/rfc5492>.

[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>.

[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>.

[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>.

[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>.

6.2. Informative 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-11, 21 August 2023, <https://datatracker.ietf.org/doc/html/draft-ietf-bess-bgp-multicast-controller-11>.

[I-D.ietf-bess-mvpn-evpn-aggregation-label] Zhang, Z. J., Rosen, E. C., Lin, W., Li, Z., and I. Wijnands, "MVPN/EVPN Tunnel Aggregation with Common Labels", Work in Progress, Internet-Draft, draft-ietf-bess-mvpn-evpn-aggregation-label-14, 4 October 2023, <https://datatracker.ietf.org/doc/html/draft-ietf-bess-mvpn-evpn-aggregation-label-14>.

[I-D.ietf-bess-mvpn-pe-ce] Patel, K., Rosen, E. C., and Y. Rekhter, "BGP as an MVPN PE-CE Protocol", Work in Progress, Internet-Draft, draft-ietf-bess-mvpn-pe-ce-01, 5 October 2015, <https://datatracker.ietf.org/doc/html/draft-ietf-bess-mvpn-pe-ce-01>.

[I-D.ietf-idr-bgp-car] Rao, D., Agrawal, S., and Co-authors, "BGP Color-Aware Routing (CAR)", Work in Progress, Internet-Draft, draft-ietf-idr-bgp-car-04, 14 November 2023, <https://datatracker.ietf.org/doc/html/draft-ietf-idr-bgp-car-04>.

[I-D.ietf-idr-bgp-ct] Vairavakkalai, K. and N. Venkataraman, "BGP Classful Transport Planes", Work in Progress, Internet-Draft, draft-ietf-idr-bgp-ct-18, 5 November 2023, <https://datatracker.ietf.org/doc/html/draft-ietf-idr-bgp-ct-18>.

[I-D.ietf-mpls-mldp-multi-topology]
Wijnands, I., Mishra, M. P., Raza, S., Budhiraja, A., Zhang, Z. J., and A. Gulko, "mLDP Extensions for Multi-Topology Routing", Work in Progress, Internet-Draft, draft-ietf-mpls-mldp-multi-topology-03, 2 August 2023, <https://datatracker.ietf.org/doc/html/draft-ietf-mpls-mldp-multi-topology-03>.

[I-D.ietf-mpls-seamless-mpls]
Leymann, N., Decraene, B., Filsfils, C., Konstantynowicz, M., and D. Steinberg, "Seamless MPLS Architecture", Work in Progress, Internet-Draft, draft-ietf-mpls-seamless-mpls-07, 28 June 2014, <https://datatracker.ietf.org/doc/html/draft-ietf-mpls-seamless-mpls-07>.

[I-D.wijnands-bier-mld-lan-election] Wijnands, I., Pfister, P., and Z. J. Zhang, "Generic Multicast Router Election on LAN's",

Work in Progress, Internet-Draft, draft-wijnands-bier-mld-lan-election-02, 19 October 2023, <https://datatracker.ietf.org/doc/html/draft-wijnands-bier-mld-lan-election-02>.

[RFC3376] Cain, B., Deering, S., Kouvelas, I., Fenner, B., and A. Thyagarajan, "Internet Group Management Protocol, Version 3", RFC 3376, DOI 10.17487/RFC3376, October 2002, <https://www.rfc-editor.org/info/rfc3376>.

[RFC3810] Vida, R., Ed. and L. Costa, Ed., "Multicast Listener Discovery Version 2 (MLDv2) for IPv6", RFC 3810, DOI 10.17487/RFC3810, June 2004, <https://www.rfc-editor.org/info/rfc3810>.

[RFC4915] Psenak, P., Mirtorabi, S., Roy, A., Nguyen, L., and P. Pillay-Esnault, "Multi-Topology (MT) Routing in OSPF", RFC 4915, DOI 10.17487/RFC4915, June 2007, <https://www.rfc-editor.org/info/rfc4915>.

[RFC5120] Przygienda, T., Shen, N., and N. Sheth, "M-ISIS: Multi Topology (MT) Routing in Intermediate System to Intermediate Systems (IS-ISs)", RFC 5120, DOI 10.17487/RFC5120, February 2008, <https://www.rfc-editor.org/info/rfc5120>.

[RFC5496] Wijnands, IJ., Boers, A., and E. Rosen, "The Reverse Path Forwarding (RPF) Vector TLV", RFC 5496, DOI 10.17487/RFC5496, March 2009, <https://www.rfc-editor.org/info/rfc5496>.

[RFC6420] Cai, Y. and H. Ou, "PIM Multi-Topology ID (MT-ID) Join Attribute", RFC 6420, DOI 10.17487/RFC6420, November 2011, <https://www.rfc-editor.org/info/rfc6420>.

[RFC6512] Wijnands, IJ., Rosen, E., Napierala, M., and N. Leymann, "Using Multipoint LDP When the Backbone Has No Route to the Root", RFC 6512, DOI 10.17487/RFC6512, February 2012, <https://www.rfc-editor.org/info/rfc6512>.

[RFC6559] Farinacci, D., Wijnands, IJ., Venaas, S., and M. Napierala, "A Reliable Transport Mechanism for PIM", RFC 6559, DOI 10.17487/RFC6559, March 2012, <https://www.rfc-editor.org/info/rfc6559>.

[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>.

[RFC7060] Napierala, M., Rosen, E., and IJ. Wijnands, "Using LDP Multipoint Extensions on Targeted LDP Sessions", RFC 7060, DOI 10.17487/RFC7060, November 2013, <https://www.rfc-editor.org/info/rfc7060>.

[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>.

[RFC7438] Wijnands, IJ., Ed., Rosen, E., Gulko, A., Joorde, U., and J. Tantsura, "Multipoint LDP (mLDP) In-Band Signaling with Wildcards", RFC 7438, DOI 10.17487/RFC7438, January 2015, <https://www.rfc-editor.org/info/rfc7438>.

[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>.

[RFC9350] Psenak, P., Ed., Hegde, S., Filsfils, C., Talaulikar, K., and A. Gulko, "IGP Flexible Algorithm", RFC 9350, DOI 10.17487/RFC9350, February 2023, <https://www.rfc-editor.org/info/rfc9350>.

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
Arrcus

Email: ice@braindump.be

Mankamana Mishra
Cisco Systems

Email: mankamis@cisco.com

Arkadiy Gulko
EdwardJones

Email: arkadiy.gulko@edwardjones.com