BESS
Internet-Draft
Intended status: Standards Track
Expires: August 23, 2021

Z. Zhang
Juniper Networks
R. Raszuk
NTT Network Innovations
D. Pacella
Verizon
A. Gulko
Edward Jones Wealth Management

February 19, 2021

Controller Based BGP Multicast Signaling
draft-ietf-bess-bgp-multicast-controller-06

Abstract

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

Requirements Language

The key words "MUST", "MUST NOT", "REQUIRED", "SHALL", "SHALL NOT", "SHOULD", "SHOULD NOT", "RECOMMENDED", "NOT RECOMMENDED", "MAY", and "OPTIONAL" in this document are to be interpreted as described in BCP 14 [RFC2119] [RFC8174] when, and only when, they appear in all capitals, as shown here.

Status of This Memo

This Internet-Draft is submitted in full conformance with the provisions of BCP 78 and BCP 79.

Internet-Drafts are working documents of the Internet Engineering Task Force (IETF). Note that other groups may also distribute working documents as Internet-Drafts. The list of current Internet-Drafts is at https://datatracker.ietf.org/drafts/current/.

Copyright Notice

Table of Contents

1. Overview

1.1. Introduction

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

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

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

As an alternative to each hop independently determining its upstream router and signaling upstream towards the root (following PIM/mLDP model), the entire tree can be calculated by a centralized controller, and the signaling can be entirely done from the controller, using the same BGP messages as defined in

[I-D.ietf-bess-bgp-multicast]. For that, some additional procedures and optimizations are specified in this document.

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

1.2. Resilience

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

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

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

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

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

1.3. Signaling

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

Suppose that for a particular tree, there are two downstream routers D1 and D2 for a particular upstream router U. A controller C may send two Leaf A-D routes to U, as if the two routes were originated by D1 and D2 but reflected by the controller. Alternatively, C could just send one route to U, with the Upstream Router's IP Address field set to U's IP address and the TEA specifying both the two downstreams and its upstream (see Section 3.1.4). In this case, the Originating Router's Address field of the Leaf A-D route is set to the controller's address. Note that for a TEA attached to a unicast NLRI, only one of the tunnels in a TEA is used for forwarding a particular packet, while all the tunnels in a TEA are used to reach multiple endpoints when it is attached to a multicast NLRI.

Notice that, in case of labeled trees, the (x,g), mLDP FEC, or SR-P2MP tree identification Section 1.7 signaling is actually not needed to transit routers but only needed to tunnel root/leaves. However, for consistency among the root/leaf/transit nodes, and for consistency with the hop-by-hop signaling, the same signaling (with tree identification encoded in the NLRI) is used to all routers.

Nonetheless, a new NLRI route type is defined to encode label/SID instead of tree identification in the NLRI, for scenarios where there is really no need to signal tree identification, e.g. as described in Section 2. On a tunnel root, the tree's binding SID can be encoded in the NLRI.

For a tree node to acknowledge to the controller that it has received the signaling and installed corresponding forwarding state, it

advertises a corresponding Leaf AD route, with the Originating Router's IP Address set to itself and with a Route Target to match the controller. For comparison, the tree signaling Leaf AD route from the controller has the Originating Router's IP Address set to the controller and the Route Target matching the tree node. The two Leaf AD routes (for controller to signal to a tree node and for a tree node to acknowledge back) differ only in those two aspects.

Notice that a leaf node may also send a Leaf A-D route to the controller to signal that it is a leaf of a tree (Section 1.5.1). That leaf-announcing route is different from the above mentioned acknowledgement route at least in the "Upstream Router's IP Address field" - the former has the controller's address while the latter has this node's address in the field. The RDs are likely different as well.

With the acknowledgement Leaf AD routes, the controller knows if tree setup is complete. The information can be used for many purposes, e.g. the controller may instruct the ingress to start forwarding traffic onto a tree only after it knows that the tree setup has completed.

1.4. Label Allocation

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

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

Assignment from each router's SRLB is no different from each router assigning labels from its own local label space in the hop-by-hop signaling case. The assignments for a router is independent of assignments for another router, even for the same tree.

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

unidirectional tree/tunnel (Section 1.4.1). Otherwise, assignment from SRLB is preferred.

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

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

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

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

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

Traditionally, p2mp mpls forwarding does not require explicit RPF check as a downstream router advertises a label only to its upstream router and all traffic with that incoming label is presumed to be from the upstream router and accepted. When a downstream router switches to a different upstream router a different label will be advertised, so it can determine if traffic is from its expected upstream neighbor purely based on the label. Now with a single common label used for all routers on a tree to send and receive traffic with, a router can no longer determine if the traffic is from its expected neighbor just based on that common tree label. Therefore, explicit RPF check is needed. Instead of interface based

RPF checking as in PIM case, neighbor based RPF checking is used - a label identifying the upstream neighbor precedes the tree label and the receiving router checks if that preceding neighbor label matches its expected upstream neighbor. Notice that this is similar to what's described in Section "9.1.1 Discarding Packets from Wrong PE" of RFC 6513 (an egress PE discards traffic sent from a wrong ingress PE). The only difference is one is used for label based forwarding and the other is used for (s,g) based forwarding. [note: for bidirectional trees, we may be able to use two labels per tree - one for upstream traffic and one for downstream traffic. This needs further verification].

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

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

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

This specification requires that, in case of upstream-assignment from a controller's local label space, each router D to assign, corresponding to each controller C, a context label that identifies the upstream-assigned label space used by that controller. This label, call it Lc-D, is communicated by D to C via BGP-LS [RFC 7752].

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

To achieve that, when C sends a Leaf A-D route, for each tunnel in the TEA, it includes a label stack Sub-TLV [I-D.ietf-idr-tunnel-encaps], with the outer label being the context label Lc-D (received by the controller from the corresponding downstream), the next label being the upstream neighbor label Lu, and the inner label being the label Lt assigned by the controller for the

tree. The router receiving the route will use the label stacks to send traffic to its downstreams.

For C to signal the expected label stack for D to receive traffic with, we overload a tunnel TLV in the TEA of the Leaf A-D route sent to D - if the tunnel TLV has a RPF sub-TLV (Section 3.1.4), then it indicates that this is actually for receiving traffic from the upstream.

1.5. Determining Root/Leaves

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

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

1.5.1. PIM-SSM/Bidir or mLDP

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

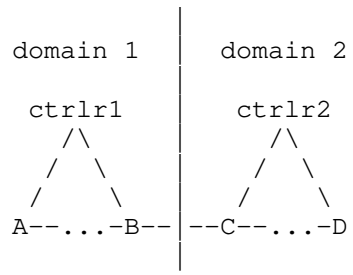
1.5.2. PIM ASM

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

1.6. Multiple Domains

An end to end multicast tree may span multiple routing domains, and the setup of the tree in each domain may be done differently as specified in [I-D.ietf-bess-bgp-multicast]. This section discusses a few aspects specific to controller signaling.

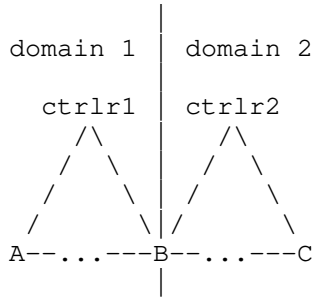
Consider two adjacent domains each with its own controller in the following configuration where router B is an upstream node of C for a multicast tree:



In the case of native (un-labeled) IP multicast, nothing special is needed. Controller 1 signals B to send traffic out of B-C link while Controller 2 signals C to accept traffic on the B-C link.

In the case of labeled IP multicast or mLDP tunnel, the controllers may be able to coordinate their actions such that Controller 1 signals B to send traffic out of B-C link with label X while Controller 2 signals C to accept traffic with the same label X on the B-C link. If the coordination is not possible, then C needs to use hop-by-hop BGP signaling to signal towards B, as specified in [I-D.ietf-bess-bgp-multicast].

The configuration could also be as following, where router B borders both domain 1 and domain 2 and is controlled by both controllers:



As discussed in Section 1.2, when B receives signaling from both Controller 1 and Controller 2, only one of the routes would be selected as the best route and used for programming the forwarding state of the corresponding segment. For B to stitch the two segments together, it is expected for B to know by provisioning that it is a border router so that B will look for the other segment (represented by the signaling from the other controller) and stitch the two together.

1.7. SR-P2MP

[I-D.voyer-pim-sr-p2mp-policy] describes an architecture to construct a Point-to-Multipoint (P2MP) tree to deliver Multi-point services in a Segment Routing domain. An SR P2MP tree is constructed by stitching together a set of Replication Segments that are specified in [I-D.voyer-spring-sr-replication-segment]. An SR Point-to-Multipoint (SR P2MP) Policy is used to define and instantiate a P2MP tree which is computed by a controller.

An SR P2MP tree is no different from an mLDP tunnel in MPLS forwarding plane. The difference is in control plane - instead of hop-by-hop mLDP signaling from leaves towards the root, to set up SR P2MP trees controllers program forwarding state (referred to as Replication Segments) to the root, leaves, and intermediate replication points using Netconf, PCEP, BGP or any other reasonable signaling/programming methods.

Procedures in this document can be used for controllers to set up SR P2MP trees with just an additional S-PMSI route type.

If/once the SR Replication Segment is extended to bi-redirectional, and SR MP2MP is introduced, the same procedures in this document would apply to SR MP2MP as well.

2. Alternative to BGP-MVPN

Multicast with BGP signaling from controllers can be an alternative to BGP-MVPN [RFC6514]. It is an attractive option especially when the controller can easily determine the source and leaf information.

With BGP-MVPN, distributed signaling is used for the following:

- Egress PEs advertise C-multicast (Type-6/7) Auto-Discovery (AD) routes to join C-multicast trees at the overlay (PE-PE)

- In case of ASM, ingress PEs advertise Source Active (Type-5) AD routes to signal sources so that egress PEs can establish Shortest Path Trees (SPT).

- PEs advertise I/S-PMSI (Type-1/2/3) AD routes to signal the binding of overlay/customer traffic to underlay/provider tunnels. For some types of tunnels, Leaf AD routes are advertised by egress PEs in response to I/S-PMSI AD routes to join the tunnels.

Based on the above signaled information, an ingress PE builds forwarding state to forward traffic arriving on the PE-CE interface to the provider tunnel (and local interfaces if there are local

downstream receivers), and an egress PE builds forwarding state to forward traffic arriving on a provider tunnel to local interfaces with downstream receivers.

Notice that multicast with BGP signaling from controllers essentially programs "static" forwarding state onto multicast tree nodes. As long as a controller can determine how a C-multicast flow should be forwarded on ingress/egress PEs, it can signal to the ingress/egress PEs using the procedures in this document to set up forwarding state, removing the need of the above-mentioned distributed signaling and processing.

For the controller to learn the egress PEs for a C-multicast tree (so that it can set up or find a corresponding provider tunnel), the egress PEs can advertise RTC routes that encodes ASM groups or advertise MCAST-TREE Leaf AD routes towards the controller to signal its desire to joins C-multicast trees, each carrying an extended community mapped from the Route Target for the VPN so that the controller knows which VPN it is for. The controller then advertises corresponding MCAST-TREE Leaf AD routes to set up C-multicast forwarding state on ingress and egress PEs. To encode the provider tunnel information in the MCAST-TREE Leaf AD route for an ingress PE, the TEA can explicitly list all replication branches of the tunnel, or just the corresponding SR-P2MP policy name, or just the binding SID.

If dynamic switching between inclusive and selective tunnels based on data rate is needed, the ingress PE can advertise/withdraw S-PMSI routes targeted only at the controllers, without Provider Tunnel Attribute attached. The controller then updates relevant MCAST-TREE Leaf AD routes to update C-multicast forwarding states on PEs to switch to a new tunnel.

3. Specification

3.1. Enhancements to TEA

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

3.1.1. Any-Encapsulation Tunnel

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

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

- An interface's local address - when a packet needs to sent out of the corresponding interface natively. On a LAN multicast MAC address MUST be used.

- A directly connected neighbor's interface address - when a packet needs to unicast to the address natively.

- An address that is not directly connected - when a packet needs to be tunneled to the address (any tunnel type/instance can be used).

3.1.2. Load-balancing Tunnel

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

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

3.1.3. Receiving MPLS Label Stack

While [I-D.ietf-bess-bgp-multicast] uses S-PMSI A-D routes to signal forwarding information for MP2MP upstream traffic, when controller signaling is used, a single Leaf A-D route is used for both upstream and downstream traffic. Since different upstream and downstream labels need to be used, a new "Receiving MPLS Label Stack" of type TBD is added as a tunnel sub-TLV in addition to the existing MPLS Label Stack sub-TLV. Other than type difference, the two are the encoded the same way.

The Receiving MPLS Label Stack sub-TLV is added to each downstream tunnel in the TEA of Leaf A-D route for an MP2MP tunnel to specify the forwarding information for upstream traffic from the corresponding downstream node. A label stack instead of a single

label is used because of the need for neighbor based RPF check, as further explained in the following section.

The Receiving MPLS Label Stack sub-TLV is also used for downstream traffic from the upstream for both P2MP and MP2MP, as specified below.

3.1.4. RPF Sub-TLV

The RPF sub-TLV has a type to be allocated by IANA and a one-octet length. The length is 0 currently, but if necessary in the future, sub-sub-TLVs could be placed in its value part. If the RPF sub-TLV appears in a tunnel, it indicates that the "tunnel" is for the upstream node instead of a downstream node. The tunnel contains an Receiving MPLS Label Stack sub-TLV for downstream traffic from the upstream node, and in case of MP2MP it also contains a regular MPLS Label Stack sub-TLV for upstream traffic to the upstream node.

The inner most label in the Receiving MPLS Label Stack is the incoming label identifying the tree (for comparison the inner most label for a regular MPLS Label Stack is the outgoing label). If the Receiving MPLS Label Stack sub-TLVe has more than one labels, the second inner most label in the stack identifies the expected upstream neighbor and explicit RPF checking needs to be set up for the tree label accordingly.

3.1.5. Tree Label Stack sub-TLV

The MPLS Label Stack sub-TLV can be used to specify the complete label stack used to send traffic, with the stack including both a transport label (stack) and label(s) that identify the (tree, neighbor) to the downstream node. There are cases where the controller only wants to specify the tree-identifying labels but leave the transport details to the router itself. For example, the router could locally determine a transport label (stack) and combine with the tree-identifying labels signaled from the controller to get the complete outgoing label stack.

For that purpose, a new Tree Label Stack sub-TLV is defined, with a one-octet length field. The value field contains a label stack with the same encoding as value part of the MPLS Label Stack sub-TLV, but the sub-TLV has a different type. A stack is specified because it may take up to three labels (see Section 1.4):

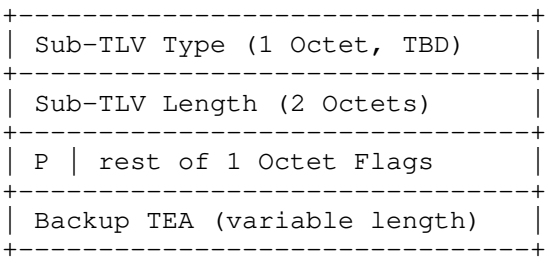
- If different nodes use different labels (allocated from the common SRGB or the node's SRLB) for a (tree, neighbor) tuple, only a single label is in the stack. This is similar to current mLDP hop by hop signaling case.

- o If different nodes use the same tree label, then an additional neighbor-identifying label is needed in front of the tree label.

- o For the previous bullet, if the neighbor-identifying label is allocated from the controller's local label space, then an additional context label is needed in front of the neighbor label.

3.1.6. Backup Tunnel sub-TLV

The Backup Tunnel sub-TLV is used to specify the backup paths for the tunnel. The length is two-octet. The value part encodes a one-octet flags field and a variable length Tunnel Encapsulation Attribute. If the tunnel goes down, traffic that is normally sent out of the tunnel is fast rerouted to the tunnels listed in the encoded TEA.

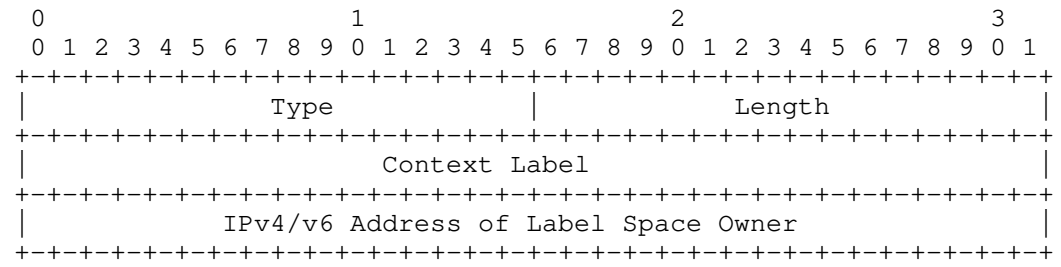


The backup tunnels can be going to the same or different nodes reached by the original tunnel.

If the tunnel carries a RPF sub-TLV and a Backup Tunnel sub-TLV, then both traffic arriving on the original tunnel and on the tunnels encoded in the Backup Tunnel sub-TLV's TEA can be accepted, if the Parallel (P-)bit in the flags field is set. If the P-bit is not set, then traffic arriving on the backup tunnel is accepted only if router has switched to receiving on the backup tunnel (this is the equivalent of PIM/mLDP MoFRR).

3.2. Context Label TLV in BGP-LS Node Attribute

For a router to signal the context label that it assigns for a controller (or any label allocator that assigns labels - from its local label space -- that will be received by this router), a new BGP-LS Node Attribute TLV is defined:



The Length field implies the type of the address. Multiple Context Label TLVs may be included in a Node Attribute, one for each label space owner.

An as example, a controller with address 11.11.11.11 allocates label 200 from its own label space, and router A assigns label 100 to identify this controller's label space. The router includes the Context Label TLV (100, 11.11.11.11) in its BGP-LS Node Attribute and the controller instructs router B to send traffic to router A with a label stack (100, 200), and router A uses label 100 to determine the Label FIB in which to look up label 200.

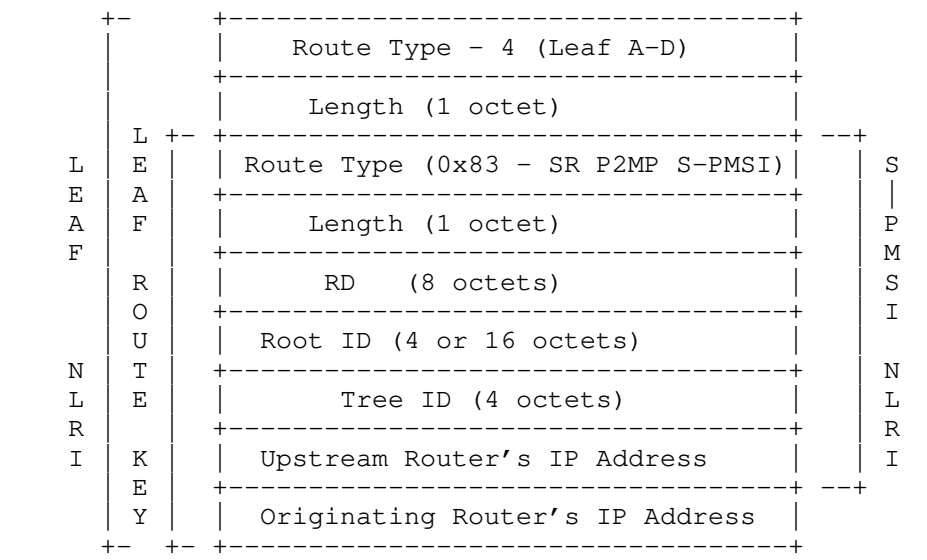
3.3. SR P2MP Signaling

An SR P2MP policy for an SR P2MP tree is identified by a (Root, Tree-id) tuple. It has a set of leaves and set of Candidate Paths (CPs). The policy is instantiated on the root of the tree, with corresponding Replication Segments - identified by (Root, Tree-id, Tree-Node-id) - instantiated on the tree nodes (root, leaves, and intermediate replication points). The Candidate Path is implicitly identified by the Route Distinguisher.

3.3.1. S-PMSI A-D Route for SR P2MP

With BGP signaled IP multicast trees and mLDP tunnels, the tree/ tunnel identification is encoded in the NLRI of S-PMSI A-D routes and corresponding Leaf A-D routes. The signaling sets up forwarding state on each node of the tree, so the NLRI also contains the identification of the node in the "Upstream Router's IP Address" field.

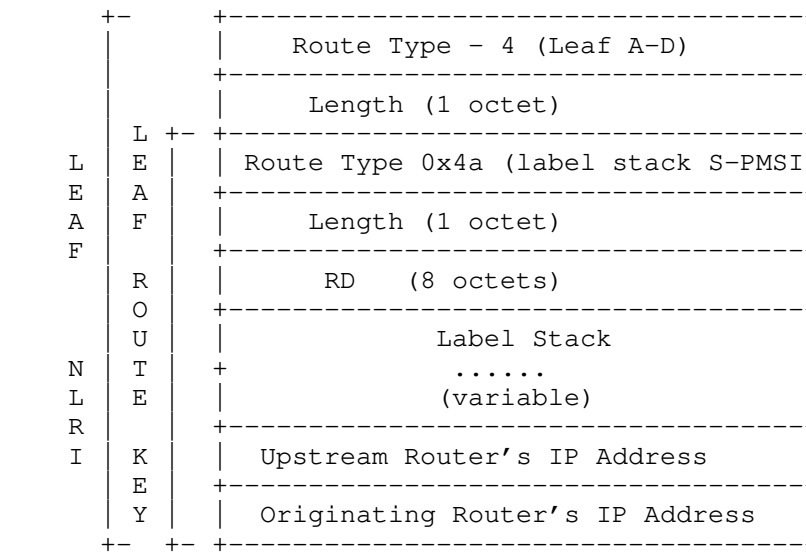
For SR P2MP, forwarding state are represented as Replication Segments and are signaled from controllers to tree nodes. A Replication Segment is identified in a new type of S-PMSI A-D route and corresponding Leaf A-D route (note that the "Leaf" term here does not refer to tree leaves):



Leaf A-D route for SR Replication Segment

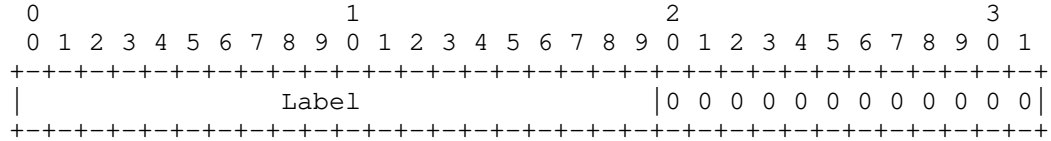
3.3.2. S-PMSI A-D Route for Encoding Label/SID

As described in Section 1.3, tree label/SID instead of tree identification could be encoded in the NLRI. For that a new Type-0x4a is defined for label stack S-PMSI. A Leaf AD route that embeds the label stack S-PMSI route has following format:



Leaf A-D route for tree identification by label stack

As discussed in Section 1.4.2, a label stack may have to be used to identify a tree in the data plane so a label stack is encoded here. The number of labels is derived from the Length field of the S-PMSI route. Each label stack entry is encoded as following:

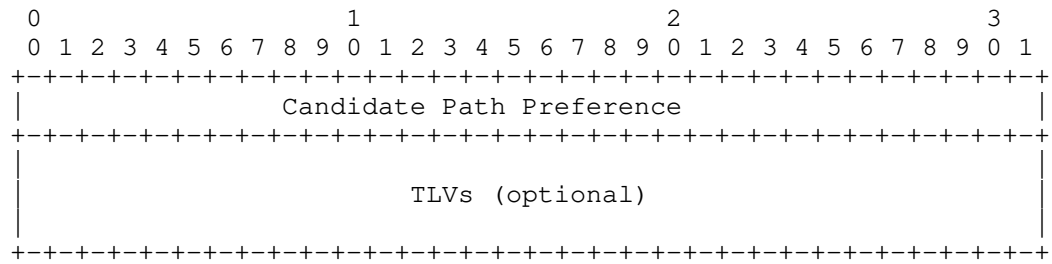


SRv6 case will be specified in future revisions.

3.3.3. BGP Community Container for SR P2MP Policy

The Leaf A-D route for Replication Segments signaled to the root is also used to signal (parts of) the SR P2MP Policy - the policy name, the set of leaves (optional, for informational purpose), preference of the CP and other information are all encoded in a newly defined BGP Community Container (BCC) [I-D.ietf-idr-wide-bgp-communities] called SR P2MP Policy BCC.

The SR P2MP Policy BCC has a BGP Community Container type to be assigned by IANA. It is composed of a fixed 4-octet Candidate Path Preference value, optionally followed by TLVs.



BGP Community Container for SR P2MP Policy

One optional TLV is to enclose the following optional Atoms TLVs that are already defined in [I-D.ietf-idr-wide-bgp-communities]:

- An IPv4 or IPv6 Prefix list - for the set of leaves
- A UTF-8 string - for the policy name

If more information for the policy are needed, more Atoms TLVs or SR P2MP Policy BCC specific TLVs can be defined.

The root receives one Leaf A-D route for each Candidate Path of the policy. Only one of the routes need to, though more than one MAY include the above listed optional Atom TLVs in the SR P2MP Policy BCC.

3.3.4. SR Policy Tunnel Type

The Tunnel Encapsulation Attribute (TEA) attached to Leaf A-D routes encodes all replication branch information. For example, if an SR explicit path is to be used to reach a particular downstream node, the TEA will include a tunnel that lists the entire label stack for that SR path, plus the label that identifies the SR P2MP tree to the downstream node.

That SR path may have been installed on the node as a unicast SR policy with a corresponding Binding SID. In stead of listing the entire label stack in an MPLS tunnel in the TEA, a different tunnel, SR Policy Tunnel [I-D.ietf-idr-segment-routing-te-policy], can be used as an alternative. The tunnel includes a Binding SID sub-TLV, an optional endpoint sub-TLV that identifies the downstream node, and an optional one-segment segment list that identifies to the downstream node the SR P2MP tree. When a node receives the Leaf A-D route with the TEA that contains an SR Policy Tunnel without a RPF sub-TLV, the Binding SID is used to locate corresponding outgoing segment lists used to reach the downstream node; the tree-identifying

segment from the optional one-segment segment list is added to to outgoing segment lists mapped from the binding SID to form the entire segment list used to send traffic to downstream node.

Note that, the SR Policy Tunnel is initially defined to instantiate an SR policy. For that use case it provides information associated with the policy, e.g., Binding SID, preference, and segment lists. The receiving node installs that policy and establishes the mapping from the Binding SID to the outgoing segments. The use of SR Policy Tunnel in this document is to refer to a pre-installed SR policy so the preference and segment lists are not used.

If a tunnel in the TEA carries a RPF sub-TLV, it is for the upstream node. The tunnel may be an MPLS tunnel in case of SR MPLS, and the Receiving MPLS Label Stack sub-TLV specifies the incoming label stack that identifies the tree and optionally the upstream neighbor. Alternatively, for both SR-MPLS and SRv6 an SR Policy Tunnel with the RPF sub-TLV can be used, in which the Binding SID sub-TLV is the SID for the tree.

If the node is the root and a Binding SID is allocated by the controller, the Binding SID is signaled to the root in a TEA tunnel with a RPF sub-TLV as above but without a destination sub-TLV.

4. Procedures

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

5. Security Considerations

This document does not introduce new security risks.

6. IANA Considerations

This document makes the following IANA requests:

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

- Assign "Member Tunnels", "Receiving MPLS Label Stack", "Tree Label Stack" and "RPF" sub-TLV types from the "BGP Tunnel Encapsulation Attribute Sub-TLVs" registry. The "Member Tunnels" sub-TLV has a two-octet value length (so the type should be in the 128-255 range), while the "Receiving MPLS Label Stack", "Tree Label" and "RPF" sub-TLV has a one-octet value length.

- Assign "Context Label TLV" type from the "BGP-LS Node Descriptor, Link Descriptor, Prefix Descriptor, and Attribute TLVs" registry.

- Assign "S-PMSI A-D Route for SR P2MP" route type from the "BGP MCAST-TREE Route Types" registry, with a suggested value of 0x83.

- Assign a new BGP Community Container type "SR P2MP Policy", and to create an "SR P2MP Policy Community Container TLV Registry", with an initial entry for "TLV for Atoms".

7. Acknowledgements

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

8. References

8.1. Normative References

[I-D.ietf-bess-bgp-multicast]
Zhang, Z., Giuliano, L., Patel, K., Wijnands, I., Mishra, M., and A. Gulko, "BGP Based Multicast", draft-ietf-bess-bgp-multicast-03 (work in progress), January 2021.

[I-D.ietf-idr-segment-routing-te-policy]
Previdi, S., Filsfils, C., Talaulikar, K., Mattes, P., Rosen, E., Jain, D., and S. Lin, "Advertising Segment Routing Policies in BGP", draft-ietf-idr-segment-routing-te-policy-11 (work in progress), November 2020.

[I-D.ietf-idr-tunnel-encaps]
Patel, K., Velde, G., Sangli, S., and J. Scudder, "The BGP Tunnel Encapsulation Attribute", draft-ietf-idr-tunnel-encaps-21 (work in progress), January 2021.

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

[I-D.voyer-pim-sr-p2mp-policy]
Voyer, D., Filsfils, C., Parekh, R., Bidgoli, H., and Z. Zhang, "Segment Routing Point-to-Multipoint Policy", draft-voyer-pim-sr-p2mp-policy-02 (work in progress), July 2020.

[I-D.voyer-spring-sr-replication-segment]
Voyer, D., Filsfils, C., Parekh, R., Bidgoli, H., and Z. Zhang, "SR Replication Segment for Multi-point Service Delivery", draft-voyer-spring-sr-replication-segment-04 (work in progress), July 2020.

[RFC2119] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, DOI 10.17487/RFC2119, March 1997, <https://www.rfc-editor.org/info/rfc2119>.

[RFC8174] Leiba, B., "Ambiguity of Uppercase vs Lowercase in RFC 2119 Key Words", BCP 14, RFC 8174, DOI 10.17487/RFC8174, May 2017, <https://www.rfc-editor.org/info/rfc8174>.

8.2. Informative References

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

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

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

[RFC8402] Filsfils, C., Ed., Previdi, S., Ed., Ginsberg, L., Decraene, B., Litkowski, S., and R. Shakir, "Segment Routing Architecture", RFC 8402, DOI 10.17487/RFC8402, July 2018, <https://www.rfc-editor.org/info/rfc8402>.

Authors' Addresses

Zhaohui Zhang
Juniper Networks

EMail: zzhang@juniper.net

Robert Raszuk
NTT Network Innovations

EMail: robert@raszuk.net

Dante Pacella
Verizon

EMail: dante.j.pacella@verizon.com

Arkadiy Gulko
Edward Jones Wealth Management

EMail: arkadiy.gulko@edwardjones.com

Network Working Group
Internet-Draft
Intended status: Informational
Expires: August 26, 2021

W. J
Y. Liu
China Mobile
S. Chen
Huawei

February 22, 2021

Traffic Steering using BGP Flowspec with SRv6 Policy
draft-jiang-idr-ts-flowspec-srv6-policy-03

Abstract

BGP Flow Specification (FlowSpec) [I-D.ietf-idr-rfc5575bis] has been proposed to distribute BGP FlowSpec NLRI to FlowSpec clients to mitigate (distributed) denial-of-service attacks, and to provide traffic filtering in the context of a BGP/MPLS VPN service. Recently, traffic steering applications in the context of SRv6 using FlowSpec aslo attract attention. This document introduces the usage of BGP FlowSpec to steer packets into an SRv6 Policy.

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 RFC 2119 [RFC2119].

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 August 26, 2021.

Table of Contents

1. Introduction

Segment Routing IPv6 (SRv6) is a protocol designed to forward IPv6 data packets on a network using the source routing model. SRv6 enables the ingress to add a segment routing header (SRH) [RFC8754] to an IPv6 packet and push an explicit IPv6 address stack into the SRH. After receiving the packet, each transit node updates the IPv6 destination IP address in the packet and segment list to implement hop-by-hop forwarding.

SRv6 Policy [I-D.ietf-spring-segment-routing-policy] is a tunneling technology developed based on SRv6. An SRv6 Policy is a set of candidate paths consisting of one or more segment lists, that is, segment ID (SID) lists. Each SID list identifies an end-to-end path from the source to the destination, instructing a device to forward traffic through the path rather than the shortest path computed using an IGP. The header of a packet steered into an SRv6 Policy is augmented with an ordered list of segments associated with that SRv6

Policy, so that other devices on the network can execute the instructions encapsulated into the list.

The headend of an SRv6 Policy may learn multiple candidate paths for an SRv6 Policy. Candidate paths may be learned via a number of different mechanisms, e.g., CLI, NetConf, PCEP, or BGP.

[I-D.ietf-idr-rfc5575bis] defines the flow specification (FlowSpec) that allows to convey flow specifications and traffic Action/Rules associated (rate- limiting, redirect, remark ...). BGP Flow specifications are encoded within the MP_REACH_NLRI and MP_UNREACH_NLRI attributes. Rules (Actions associated) are encoded in Extended Community attribute. The BGP Flow Specification function allows BGP Flow Specification routes that carry traffic policies to be transmitted to BGP Flow Specification peers to steer traffic.

This document proposes BGP flow specification usage that are used to steer data flow into an SRv6 Policy as well as to indicate Tailend function.

2. Definitions and Acronyms

- FlowSpec: Flow Specification
- SR: Segment Routing
- SRv6: IPv6 Segment Routing
- SID: Segment Identifier
- SRH: Segment Routing Header
- TE: Traffic Engineering

3. Operations

An SRv6 Policy [I-D.ietf-spring-segment-routing-policy] is identified through the tuple <headend, color, endpoint>. In the context of a specific headend, one may identify an SRv6 policy by the <color, endpoint> tuple. The headend is the node where the SRv6 policy is instantiated/implemented. The headend is specified as an IPv4 or IPv6 address and is expected to be unique in the domain. The endpoint indicates the destination of the SRv6 policy. The endpoint is specified as an IPv6 address and is expected to be unique in the domain. The color is a 32-bit numerical value that associates the SRv6 Policy, and it defines an application-level network Service Level Agreement (SLA) policy.

Assume one or multiple SRv6 Policies are already setup in the SRv6 HeadEnd device. In order to steer traffic into a specific SRv6 policy at the Headend, one can use the SRv6 color extended community and endpoint to map to a satisfying SRv6 policy, and steer traffic into this specific policy.

[I-D.ietf-idr-flowspec-redirect-ip] defines the redirect to IPv4 and IPv6 Next-hop action. The IPv6 next-hop address in the FlowSpec NLRI can be used to specify the endpoint of the SRv6 Policy. When the packets reach to the TailEnd device, some specific function imformation identifiers can be used decide how to further process the flows. Several endpoint functions are already defined, e.g., End.DT6: Endpoint with decapsulation and IPv6 table lookup, and End.DX6: Endpoint with decapsulation and IPv6 cross-connect. The BGP Prefix-SID defined in [RFC8669] is utilized to enable SRv6 VPN services [I-D.ietf-bess-srv6-services]. SRv6 Services TLVs within the BGP Prefix-SID Attribute can be used to indicate the endpoint functions.

This document proposes to carry the Color Extended Community and BGP Prefix-SID Attribute in the context of a Flowspec NLRI [I-D.ietf-idr-rfc5575bis] to an SRv6 Headend to steer traffic into one SRv6 policy, as well as to indicate specific Tailend functions.

In this document, the usage of at most one Color Extended Community in combination at most one BGP Prefix SID Attribute is discussed. For the case that a flowspec route carries multiple Color Extend Communities and/or a BGP Prefix SID Attribute, a protocol extension to Flowspec is required, and is thus out of the scope of this document.

However, the method proposed in this document still supports load balancing to the tailend device. To achieve that, the headend device CAN set up multiple paths in one SRv6 policy, and use a Flowspec route to indicate the specific SRv6 policy.

4. Application Example

In following scenario, BGP FlowSpec Controller signals the function imformation (SRv6 SID: Service_id_x) to the HeadEnd device.

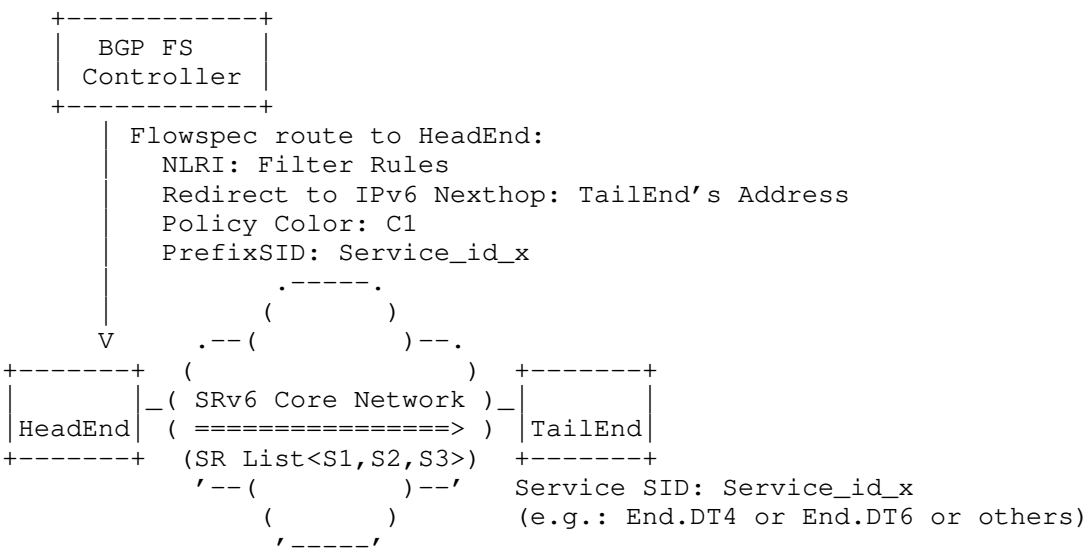


Figure 1: Steering the Flow into SRv6 Policy

When the headend device (as a Flowspec client) receives such instructions, it will steer the flows matching the criteria in the Flowspec route into the SRv6 Policy matching the tuple (Endpoint: TailEnd's Address, Color: C1). And the packets of such flows will be encapsulated with SRH using the SR List<S1, S2, S3, Service_id_x>. When the packets reach to the TailEnd device, they will be further procetssed per the function denoted by the Service_id_x.

For the cases of intra-AS and inter-AS traffic steering using this method, the usages of Flowspec Color Extended Community with BGP prefix SID are the same for both scenarios. The difference lie between the local SRv6 policy configurations. For the inter-domain case, the operator can configure an inter-domain SRv6 policy/path at the Headend device. For the intra-domain case, the operator can configure an intra-domain SRv6 policy/path at the Headend device.

5. IANA Considerations

No IANA actions are required for this document.

6. Security Considerations

This document does not change the security properties of SRv6 and BGP.

7. Contributors

The following people made significant contributions to this document:

Yunan Gu
Huawei
Email: guyunan@huawei.com

Shunwan Zhaung
Huawei
Email: zhuangshunwan@huawei.com

Haibo Wang
Huawei
Email: rainsword.wang@huawei.com

Jie Dong
Huawei
Email: jie.dong@huawei.com

8. Acknowledgements

TBD.

9. References

9.1. Normative References

[I-D.ietf-bess-srv6-services]
Dawra, G., Filsfils, C., Talaulikar, K., Raszuk, R., Decraene, B., Zhuang, S., and J. Rabadan, "SRv6 BGP based Overlay services", draft-ietf-bess-srv6-services-05 (work in progress), November 2020.

[I-D.ietf-idr-flowspec-redirect-ip]
Uttaro, J., Haas, J., Texier, M., Andy, A., Ray, S., Simpson, A., and W. Henderickx, "BGP Flow-Spec Redirect to IP Action", draft-ietf-idr-flowspec-redirect-ip-02 (work in progress), February 2015.

[I-D.ietf-idr-rfc5575bis]
Loibl, C., Hares, S., Raszuk, R., McPherson, D., and M. Bacher, "Dissemination of Flow Specification Rules", draft-ietf-idr-rfc5575bis-27 (work in progress), October 2020.

[I-D.ietf-idr-segment-routing-te-policy]
Previdi, S., Filsfils, C., Talaulikar, K., Mattes, P., Rosen, E., Jain, D., and S. Lin, "Advertising Segment Routing Policies in BGP", draft-ietf-idr-segment-routing-te-policy-11 (work in progress), November 2020.

[I-D.ietf-idr-tunnel-encaps]
Patel, K., Velde, G., Sangli, S., and J. Scudder, "The BGP Tunnel Encapsulation Attribute", draft-ietf-idr-tunnel-encaps-21 (work in progress), January 2021.

[I-D.ietf-spring-segment-routing-policy]
Filsfils, C., Talaulikar, K., Voyer, D., Bogdanov, A., and P. Mattes, "Segment Routing Policy Architecture", draft-ietf-spring-segment-routing-policy-09 (work in progress), November 2020.

[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>.

[RFC8669] Previdi, S., Filsfils, C., Lindem, A., Ed., Sreekantiah, A., and H. Gredler, "Segment Routing Prefix Segment Identifier Extensions for BGP", RFC 8669, DOI 10.17487/RFC8669, December 2019, <https://www.rfc-editor.org/info/rfc8669>.

9.2. Informative References

[RFC4456] Bates, T., Chen, E., and R. Chandra, "BGP Route Reflection: An Alternative to Full Mesh Internal BGP (IBGP)", RFC 4456, DOI 10.17487/RFC4456, April 2006, <https://www.rfc-editor.org/info/rfc4456>.

[RFC8754] Filsfils, C., Ed., Dukes, D., Ed., Previdi, S., Leddy, J., Matsushima, S., and D. Voyer, "IPv6 Segment Routing Header (SRH)", RFC 8754, DOI 10.17487/RFC8754, March 2020, <https://www.rfc-editor.org/info/rfc8754>.

Authors' Addresses

Wenying Jiang
China Mobile
Beijing
China

Email: jiangwenying@chinamobile.com

Yisong Liu
China Mobile
Beijing
China

Email: liuyisong@chinamobile.com

Shuanglong Chen
Huawei
Beijing
China

Email: chenshuanglong@huawei.com

Network Working Group
Internet-Draft
Intended status: Standards Track
Expires: August 19, 2021

K. Vairavakkalai
N. Venkataraman
B. Rajagopalan
Juniper Networks, Inc.
G. Mishra
Verizon Communications Inc.
M. Khaddam
Cox Communications Inc.
X. Xu
Alibaba Inc.
R. Szarecki
Google.

February 15, 2021

BGP Classful Transport Planes
draft-kaliraj-idr-bgp-classful-transport-planes-07

Abstract

This document specifies a mechanism, referred to as "service mapping", to express association of overlay routes with underlay routes satisfying a certain SLA, using BGP. The document describes a framework for classifying underlay routes into transport classes, and mapping service routes to specific transport class.

The "Transport class" construct maps to a desired SLA, and can be used to realize the "Topology Slice" in 5G Network slicing architecture.

This document specifies BGP protocol procedures that enable dissimination of such service mapping information that may span multiple co-operating administrative domains. These domains may be administetered by the same provider or closely co-ordinating provider networks.

It makes it possible to advertise multiple tunnels to the same destination address, thus avoiding need of multiple loopbacks on the egress node.

A new BGP transport layer address family (SAFI 76) is defined for this purpose that uses RFC-4364 technology and follows RFC-8277 NLRI encoding. This new address family is called "BGP Classful Transport", aka BGP CT.

It carries transport prefixes across tunnel domain boundaries (e.g. in Inter-AS Option-C networks), parallel to BGP LU (SAFI 4) . It dissiminates "Transport class" information for the transport prefixes

across the participating domains, which is not possible with BGP LU. This makes the end-to-end network a "Transport Class" aware tunneled network.

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 RFC 2119 [RFC2119].

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 August 19, 2021.

Copyright Notice

Table of Contents

1. Introduction

To facilitate service mapping, the tunnels in a network can be grouped by the purpose they serve into a "Transport Class". The tunnels could be created using any signaling protocol, such as LDP, RSVP, BGP LU or SPRING. The tunnels could also use native IP or IPv6, as long as they can carry MPLS payload. Tunnels may exist between different pair of end points. Multiple tunnels may exist between the same pair of end points.

Thus, a Transport Class consists of tunnels created by various protocols that satisfy the properties of the class. For example, a "Gold" transport class may consist of tunnels that traverse the shortest path with fast re-route protection, a "Silver" transport class may hold tunnels that traverse shortest paths without protection, a "To NbrAS Foo" transport class may hold tunnels that exit to neighboring AS Foo, and so on.

The extensions specified in this document can be used to create a BGP transport tunnel that potentially spans domains, while preserving its Transport Class. Examples of domain are Autonomous System (AS), or IGP area. Within each domain, there is a second level underlay tunnel used by BGP to cross the domain. The second level underlay tunnels could be hetrogeneous: Each domain may use a different type of tunnel (e.g. MPLS, IP, GRE), or use a differnet signaling protocol. A domain boundary is demarcated by a rewrite of BGP nexthop to 'self' while re-advertising tunnel routes in BGP. Examples of domain boundary are inter-AS links and inter-region ABRs. The path uses MPLS label-switching when crossing domain boundary and uses the native intra-AS tunnel of the desired transport class when traversing within a domain.

Overlay routes carry sufficient indication of the Transport Classes they should be encapsulated over, in form of BGP community called the "Mapping community". Based on the mapping community, "route resolution" procedure on the ingress node selects from the corresponding Transport Class an appropriate tunnel whose destination matches (LPM) the nexthop of the overlay route. If the overlay route is carried in BGP, the protocol nexthop (or, PNH) is generally carried as an attribute of the route.

The PNH of the overlay route is also referred to as "service endpoint" (SEP). The service endpoint may exist in the same domain as the service ingress node or lie in a different domain, adjacent or non-adjacent. In the former case, reachability to the SEP is provided by an intra-domain tunneling protocol, and in the latter case, reachability to the SEP is via BGP transport families.

In this architecture, the intra-domain transport protocols (e.g. RSVP, SRTE) are also "Transport Class aware", and they publish ingress routes in Transport RIB associated with the Transport Class, at the tunnel ingress node. These routes are then redistributed into BGP CT to be advertised to adjacent domains. It is outside the scope of this document how exactly the transport protocols are made transport class aware, though configuration on the tunnel ingress node is a simple mechanism to achieve it.

This document describes mechanisms to:

Model a "Transport Class" as "Transport RIB" on a router, consisting of tunnel ingress routes of a certain class.

Enable service routes to resolve over an intended Transport Class by virtue of carrying the appropriate "Mapping community". Which results in using the corresponding Transport RIB for finding nexthop reachability.

Advertise tunnel ingress routes in a Transport RIB via BGP without any path hiding, using BGP VPN technology and Add-path. Such that overlay routes in the receiving domains can also resolve over tunnels of associated Transport Class.

Provide a way for co-operating domains to reconcile any differences in extended community namespaces, and interoperate between different transport signaling protocols in each domain.

In this document we focus mainly on MPLS as the intra-domain transport tunnel forwarding, but the mechanisms described here would work in similar manner for non-MPLS (e.g. IP, GRE, UDP) transport tunnel forwarding technologies too.

This document assumes MPLS forwarding when crossing domain boundaries, as that is the defacto standard in deployed networks today. But mechanisms specified in this document can also support different forwarding technologies (e.g. SRv6). A future document may describe such adaptations, it is out of scope of this document.

The document Seamless Segment Routing [Seamless-SR] describes various use cases and applications of procedures described in this document.

2. Terminology

LSP: Label Switched Path.

TE : Traffic Engineering.

SN : Service Node.

BN : Border Node.

TN : Transport Node, P-router.

BGP-VPN : VPNs built using RFC4364 mechanisms.

RT : Route-Target extended community.

RD : Route-Distinguisher.

PNH : Protocol-Nexthop address carried in a BGP Update message.

SEP : Service End point, the PNH of a Service route.

LPM : Longest Prefix Match.

Service Family : BGP address family used for advertising routes for "data traffic", as opposed to tunnels.

Transport Family : BGP address family used for advertising tunnels, which are in turn used by service routes for resolution.

Transport Tunnel : A tunnel over which a service may place traffic. These tunnels can be GRE, UDP, LDP, RSVP, or SR-TE.

Tunnel Domain : A domain of the network containing SN and BN, under a single administrative control that has a tunnel between SN and BN. An end-to-end tunnel spanning several adjacent tunnel domains can be created by "stitching" them together using labels.

Transport Class : A group of transport tunnels offering the same type of service.

Transport Class RT : A Route-Target extended community used to identify a specific Transport Class.

Transport RIB : At the SN and BN, a Transport Class has an associted Transport RIB that holds its tunnel routes.

Transport Plane : An end to end plane comprising of transport tunnels belonging to same transport class. Tunnels of same transport class are stitched together by BGP route readvertisements with nexthop-self, to span across domain boundaries using Label-Swap forwarding mechanism similar to Inter-AS option-b.

Mapping Community : BGP Community/Extended-community on a service route, that maps it to resolve over a Transport Class.

3. Transport Class

A Transport Class is defined as a set of transport tunnels that share certain characteristics useful for underlay selection.

On the wire, a transport class is represented as the Transport Class RT, which is a new Route-Target extended community.

A Transport Class is configured at SN and BN, along with attributes like RD and Route-Target. Creation of a Transport Class instantiates

the associated Transport RIB and a Transport routing instance to contain them all.

The operator may configure a SN/BN to classify a tunnel into an appropriate Transport Class, which causes the tunnel's ingress routes to be installed in the corresponding Transport RIB. At a BN, these tunnel routes may then be advertised into BGP CT.

Alternatively, a router receiving the transport routes in BGP with appropriate signaling information can associate those ingress routes to the appropriate Transport Class. E.g. for Classful Transport family (SAFI 76) routes, the Transport Class RT indicates the Transport Class. For BGP LU family(SAFI 4) routes, import processing based on Communities or inter-AS source-peer may be used to place the route in the desired Transport Class.

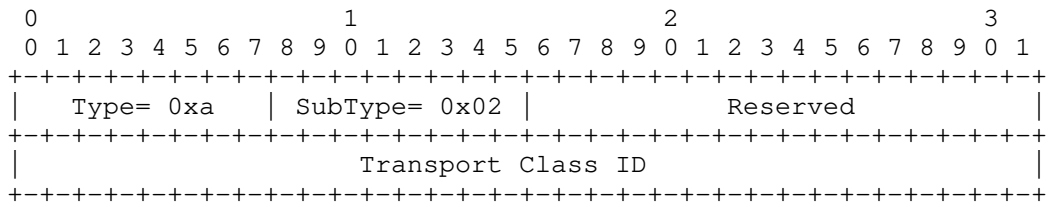
When the ingress route is received via SRTE [SRTE], which encodes the Transport Class as an integer 'Color' in the NLRI as "Color:Endpoint", the 'Color' is mapped to a Transport Class during import processing. SRTE ingress route for 'Endpoint' is installed in that transport class. The SRTE route when advertised out to BGP speakers will then be advertised in Classful Transport family with Transport Class RT and a new label. The MPLS swap route thus installed for the new label will pop the label and deliver decapsulated traffic into the path determined by SRTE route.

4. "Transport Class" Route Target Extended Community

This document defines a new type of Route Target, called "Transport Class" Route Target Extended Community.

"Transport Class" Route Target extended community is a transitive extended community EXT-COMM [RFC4360] of extended-type, with a new Format (Type high = 0xa) and SubType as 0x2 (Route Target).

This new Route Target Format has the following encoding:



"Transport Class" Route Target Extended Community

Type: 2 octets

Type field contains value 0xa.

SubType: 2 octets

Subtype field contain 0x2. This indicates 'Route Target'.

Transport Class ID: 4 octets

The least significant 32-bits of the value field contain the "Transport Class" identifier, which is a 32-bit integer.

The remaining 2 octets after SubType field are Reserved, they MUST be set to zero by originator, and ignored, left unaltered by receiver.

The "Transport class" Route Target Extended community follows the mechanisms for VPN route import, export as specified in BGP-VPN [RFC4364], and Route Target Contrain mechanisms as specified in VPN-RTC [RFC4684]

A BGP speaker that implements RT Constraint VPN-RTC [RFC4684] MUST apply the RT Constraint procedures to the "Transport class" Route Target Extended community as-well.

The Transport Class Route Target Extended community is carried on Classful Transport family routes, and allows associating them with appropriate Transport RIBs at receiving BGP speakers.

Use of the Transport Class Route Target Extended community with a new Type code avoids conflicts with any VPN Route Target assignments already in use for service families.

5. Transport RIB

A Transport RIB is a routing-only RIB that is not installed in forwarding path. However, the routes in this RIB are used to resolve reachability of overlay routes' PNH. Transport RIB is created when the Transport Class it represents is configured.

Overlay routes that want to use a specific Transport Class confine the scope of nexthop resolution to the set of routes contained in the corresponding Transport RIB. This Transport RIB is the "Routing Table" referred in Section 9.1.2.1 RFC4271 [1]

Routes in a Transport RIB are exported out in 'Classful Transport' address family.

6. Transport Routing Instance

A BGP VPN routing instance that is a container for the Transport RIB. It imports, and exports routes in this RIB with Transport Class RT. Tunnel destination addresses in this routing instance's context come from the "provider namespace". This is different from user VRFs for e.g., which contain prefixes in "customer namespace"

The Transport Routing instance uses the RD and RT configured for the Transport Class.

7. Nexthop Resolution Scheme

An implementation may provide an option for the service route to resolve over less preferred Transport Classes, should the resolution over preferred, or "primary" Transport Class fail.

To accomplish this, the set of service routes may be associated with a user-configured "resolution scheme", which consists of the primary Transport Class, and optionally, an ordered list of fallback Transport Classes.

A community called as "Mapping Community" is configured for a "resolution scheme". A Mapping community maps to exactly one resolution scheme. A resolution scheme comprises of one primary transport class and optionally one or more fallback transport classes.

A BGP route is associated with a resolution scheme during import processing. The first community on the route that matches a mapping community of a locally configured resolution scheme is considered the effective mapping community for the route. The resolution scheme thus found is used when resolving the route's PNH. If a route

contains more than one mapping community, it indicates that the route considers these multiple mapping communities as equivalent. So the first community that maps to a resolution scheme is chosen.

A transport route received in BGP Classful Transport family SHOULD use a resolution scheme that contains the primary Transport Class without any fallback to best effort tunnels. The primary Transport Class is identified by the Transport Class RT carried on the route. Thus Transport Class RT serves as the Mapping Community for Classful Transport routes.

A service route received in a BGP service family MAY map to a resolution scheme that contains the primary Transport Class identified by the mapping community on the route, and a fallback to best effort tunnels transport class. The primary Transport Class is identified by the Mapping community carried on the route. For e.g. the Extended Color community may serve as the Mapping Community for service routes. Color:0:<n> MAY map to a resolution scheme that has primary transport class <n>, and a fallback to best-effort transport class.

8. BGP Classful Transport Family NLRI

The Classful Transport family will use the existing AFI of IPv4 or IPv6, and a new SAFI 76 "Classful Transport" that will apply to both IPv4 and IPv6 AFIs.

The "Classful Transport" SAFI NLRI itself is encoded as specified in https://tools.ietf.org/html/rfc8277#section-2 [RFC8277].

When AFI is IPv4 the "Prefix" portion of Classful Transport family NLRI consists of an 8-byte RD followed by an IPv4 prefix. When AFI is IPv6 the "Prefix" consists of an 8-byte RD followed by an IPv6 prefix.

Attributes on a Classful Transport route include the Transport Class Route-Target extended community, which is used to leak the route into the right Transport RIBs on SNs and BNs in the network.

9. Comparison with other families using RFC-8277 encoding

SAFI 128 (Inet-VPN) is a RF8277 encoded family that carries service prefixes in the NLRI, where the prefixes come from the customer namespaces, and are contexualized into separate user virtual service RIBs called VRFs, using RFC4364 procedures.

SAFI 4 (BGP LU) is a RFC8277 encoded family that carries transport prefixes in the NLRI, where the prefixes come from the provider namespace.

SAFI 76 (Classful Transport) is a RFC8277 encoded family that carries transport prefixes in the NLRI, where the prefixes come from the provider namespace, but are contexualized into separate Transport RIBs, using RFC4364 procedures.

It is worth noting that SAFI 128 has been used to carry transport prefixes in "L3VPN Inter-AS Carrier's carrier" scenario, where BGP LU/LDP prefixes in CsC VRF are advertised in SAFI 128 towards the remote-end baby carrier.

In this document a new AFI/SAFI is used instead of reusing SAFI 128 to carry these transport routes, because it is operationally advantageous to segregate transport and service prefixes into separate address families, RIBs. E.g. It allows to safely enable "per-prefix" label allocation scheme for Classful Transport prefixes without affecting SAFI 128 service prefixes which may have huge scale. "per prefix" label allocation scheme keeps the routing churn local during topology changes.

A new family also facilitates having a different readvertisement path of the transport family routes in a network than the service route readvertisement path. viz. Service routes (Inet-VPN) are exchanged over an EBGP multihop session between Autonomous systems with nexthop unchanged; whereas Classful Transport routes are readvertised over EBGP single hop sessions with "nexthop-self" rewrite over inter-AS links.

The Classful Transport family is similar in vein to BGP LU, in that it carries transport prefixes. The only difference is, it also carries in Route Target an indication of which Transport Class the transport prefix belongs to, and uses RD to disambiguate multiple instances of the same transport prefix in a BGP Update.

10. Protocol Procedures

This section summarizes the procedures followed by various nodes speaking Classful Transport family

Preparing the network for deploying Classful Transport planes

- Operator decides on the Transport Classes that exist in the network, and allocates a Route-Target to identify each Transport Class.

Operator configures Transport Classes on the SNs and BNs in the network with unique Route-Distinguishers and Route-Targets.

Implementations may provide automatic generation and assignment of RD, RT values for a transport routing instance; they MAY also provide a way to manually override the automatic mechanism, in order to deal with any conflicts that may arise with existing RD, RT values in the different network domains participating in a deployment.

Origination of Classful Transport route:

At the ingress node of the tunnel's home domain, the tunneling protocols install routes in the Transport RIB associated with the Transport Class the tunnel belongs to.

The ingress node then advertises this tunnel destination into BGP as a Classful Transport family route with NLRI RD:TunnelEndpoint, attaching a 'Transport Class' Route Target that identifies the Transport Class. This BGP CT route is advertised to EBGP peers and IBGP peers which are RR-clients. This route MUST NOT be advertised to the IBGP peers who are not RR-clients.

Alternatively, the egress node of the tunnel i.e. the tunnel endpoint can originate the same BGP Classful Transport route, with NLRI RD:TunnelEndpoint and PNH TunnelEndpoint, which will resolve over the tunnel route at the ingress node. When the tunnel is up, the Classful Transport BGP route will become usable and get re-advertised.

Unique RD SHOULD be used by the originator of a Classful Transport route to disambiguate the multiple BGP advertisements for a transport end point.

Ingress node receiving Classful Transport route

On receiving a BGP Classful Transport route with a PNH that is not directly connected, e.g. an IBGP-route, a mapping community on the route (the Transport Class RT) indicates which Transport Class this route maps to. The routes in the associated Transport RIB are used to resolve the received PNH. If there does not exist a route in the Transport RIB matching the PNH, the Classful Transport route is considered unusable, and MUST NOT be re-advertised further.

Border node readvertising Classful Transport route with nexthop self:

The BN allocates an MPLS label to advertise upstream in Classful Transport NLRI. The BN also installs an MPLS swap-route for that label that swaps the incoming label with a label received from the downstream BGP speaker, or pops the incoming label. And then pushes received traffic to the transport tunnel or direct interface that the Classful Transport route's PNH resolved over.

The label SHOULD be allocated with "per-prefix" label allocation semantics. The prefix used as key is formed by stripping RD from the BGP CT NLRI prefix. This helps in avoiding BGP CT route churn through out the CT network when a failure happens in a domain. The failure is not propagated further than the BN closest to the failure.

The value of advertised MPLS label is locally significant, and is dynamic by default. The BN may provide option to allocate a value from a statically carved out range. This can be achieved using locally configured export policy, or via mechanisms described in BGP Prefix-SID [RFC8669].

Border node receiving Classful Transport route on EBGP :

If the route is received with PNH that is known to be directly connected, e.g. EBGP single-hop peering address, the directly connected interface is checked for MPLS forwarding capability. No other nexthop resolution process is performed, as the inter-AS link can be used for any Transport Class.

If the inter-AS links should honor Transport Class, then the BN SHOULD follow procedures of an Ingress node described above, and perform nexthop resolution process. The interface routes SHOULD be installed in the Transport RIB belonging to the associated Transport Class.

Avoiding path-hiding through Route Reflectors

When multiple BNs exist that advertise a RDn:PEn prefix to RRs, the RRs may hide all but one of the BNs, unless ADDPATH [RFC7911] is used for the Classful Transport family. This is similar to L3VPN option-B scenarios. Hence ADDPATH SHOULD be used for Classful Transport family, to avoid path-hiding through RRs.

Avoiding loop between Route Reflectors in forwarding path

Pair of redundant ABRs acting as RR with nexthop-self may chose each other as best path instead of the upstream ASBR, causing a traffic forwarding loop.

Implementations SHOULD provide a way to alter the tie-breaking rule specified in BGP RR [RFC4456] to tie-break on CLUSTER_LIST step before ROUTER-ID step, when performing path selection for BGP CT routes. RFC4456 considers pure RR which is not in forwarding path. When RR is in forwarding path and reflects routes with nexthop-self, which is the case for ABR BNs in a BGP transport network, this rule may cause loops. This document suggests the following modification to the BGP Decision Process Tie Breaking rules (Sect. 9.1.2.2, [RFC4271]) when doing path selection for BGP CT family routes:

The following rule SHOULD be inserted between Steps e) and f): a BGP Speaker SHOULD prefer a route with the shorter CLUSTER_LIST length. The CLUSTER_LIST length is zero if a route does not carry the CLUSTER_LIST attribute.

Some deployment considerations can also help in avoiding this problem:

IGP metric should be assigned such that "ABR to redundant ABR" cost is inferior than "ABR to upstream ASBR" cost.

Tunnels belonging to special Transport classes SHOULD NOT be provisioned between ABR to ABRs. This will ensure that the route received from an ABR with nexthop-self will not be usable at a redundant ABR.

This avoids possibility of such loops altogether, irrespective of whether the path selection modification mentioned above is implemented.

Ingress node receiving service route with mapping community

Service routes received with mapping community resolve using Transport RIBs determined by the resolution scheme. If the resolution process does not find an usable Classful Transport route or tunnel route in any of the Transport RIBs, the service route MUST be considered unusable for forwarding purpose.

Coordinating between domains using different community namespaces.

Cooperating option-C domains may sometimes not agree on RT, RD, Mapping-community or Transport Route Target values because of differences in community namespaces; e.g. during network mergers or renumbering for expansion. Such deployments may deploy mechanisms to map and rewrite the Route-target values on domain boundaries, using per ASBR import policies. This is no different than any other BGP VPN family. Mechanisms employed in inter-AS

VPN deployments may be used with the Classful Transport family also.

The resolution schemes SHOULD allow association with multiple mapping communities. This helps with renumbering, network mergers, or transitions.

Though RD can also be rewritten on domain boundaries, deploying unique RDs is strongly RECOMMENDED, because it helps in trouble shooting by uniquely identifying originator of a route, and avoids path-hiding.

This document defines a new format of Route-Target extended-community to carry Transport Class, this avoids collision with regular Route Target namespace used by service routes.

11. Scaling considerations

11.1. Avoiding unintended spread of CT routes across domains.

RFC8212 [RFC8212] suggests BGP speakers require explicit configuration of both BGP Import and Export Policies for any EBGP sessions, in order to receive or send routes on EBGP sessions.

It is recommended to follow this for BGP CT routes. It will prohibit unintended advertisement of transport routes through out the BGP CT transport domain which may span multiple AS. This will conserve usage of MPLS label and nexthop resources in the network. An ASBR of a domain can be provisioned to allow routes with only the Transport targets that are required by SNs in the domain.

11.2. Constrained distribution of PNHs to SNs (On Demand Nexthop)

This section describes how the number of Protocol Nexthops advertised to a SN or BN can be constrained using BGP Classsful Transport and VPN RTC [RFC4684]

An egress SN MAY advertise BGP CT route for RD:eSN with two Route Targets: transport-target:0:<TC> and a RT carrying <eSN>:<TC>. Where TC is the Transport Class identifier, and eSN is the IP-address used by SN as BGP nexthop in it’s service route advertisements.

transport-target:0:<TC> is the new type of route target (Transport Class RT) defined in this document. It is carried in BGP extended community attribute (BGP attribute code 16).

The RT carrying <eSN>:<TC> MAY be an IP-address specific regular RT (BGP attribute code 16), IPv6-address specific RT (BGP attribute code 25), or a Wide-communities based RT (BGP attribute code 34) as described in RTC-Ext [RTC-Ext]

An ingress SN MAY import BGP CT routes with Route Target carrying: <eSN>:<TC>. The ingress SN MAY learn the eSN values either by configuration, or it MAY discover them from the BGP nexthop field in the BGP VPN service routes received from eSN. A BGP ingress SN receiving a BGP service route with nexthop of eSN SHOULD generate a RTC/Extended-RTC route for Route Target prefix <Origin ASN>:<eSN>/[80|176] in order to learn BGP CT transport routes to reach eSN. This allows constrained distribution of the transport routes to the PNHs actually required by iSN.

When path of route propogation of BGP CT routes is same as the RTC routes, a BN would learn the RTC routes advertised by ingress SNs and propagate further. This will allow constraining distribution of BGP CT routes for a PNH to only the necessary BNs in the network, closer to the egress SN.

This mechanism provides "On Demand Nexthop" of BGP CT routes, which help with scaling of MPLS forwarding state at SN and BN.

But the amount of state carried in RTC family may become proportional to number of PNHs in the network. To strike a balance, the RTC route advertisements for <Origin ASN>:<eSN>/[80|176] MAY be confined to the BNs in home region of ingress-SN, or the BNs of a super core.

Such a BN in the core of the network SHOULD import BGP CT routes with Transport Class Route Target: 0:<TC>, and generate a RTC route for <Origin ASN>:0:<TC>/96, while not propagating the more specific RTC requests for specific PNHs. This will let the BN learn transport routes to all eSN nodes. But confine their propagation to ingress-SNs.

11.3. Limiting scope of visibility of PE loopback as PNHs

It may be even more desirable to limit the number of PNHs that are globaly visible in the network. This is possible using mechanism described in MPLS Namespaces [MPLS-NAMESPACES]

Such that advertisement of PE loopback addresses as next-hop in BGP service routes is confined to the region they belong to. An anycast IP-address called "Context Protocol Nexthop Address" abstracts the PEs in a region from other regions in the network, swapping the PE scoped service label with a CPNH scoped private namespace label.

This provides much greater advantage in terms of scaling and convergence. Changes to implement this feature are required only on the region's BNs and RR.

12. OAM considerations

Standard MPLS OAM procedures specified in [RFC8029] also apply to BGP Classful Transport.

The 'Target FEC Stack' sub-TLV for IPv4 Classful Transport has a Sub-Type of [TBD], and a length of 13. The Value field consists of the RD advertised with the Classful Transport prefix, the IPv4 prefix (with trailing 0 bits to make 32 bits in all), and a prefix length, encoded as follows:

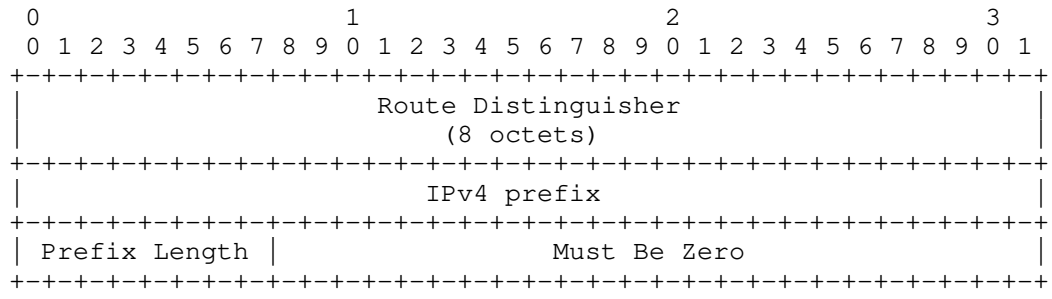


Figure 1: Classful Transport IPv4 FEC

The 'Target FEC Stack' sub-TLV for IPv6 Classful Transport has a Sub-Type of [TBD], and a length of 25. The Value field consists of the RD advertised with the Classful Transport prefix, the IPv6 prefix (with trailing 0 bits to make 128 bits in all), and a prefix length, encoded as follows:

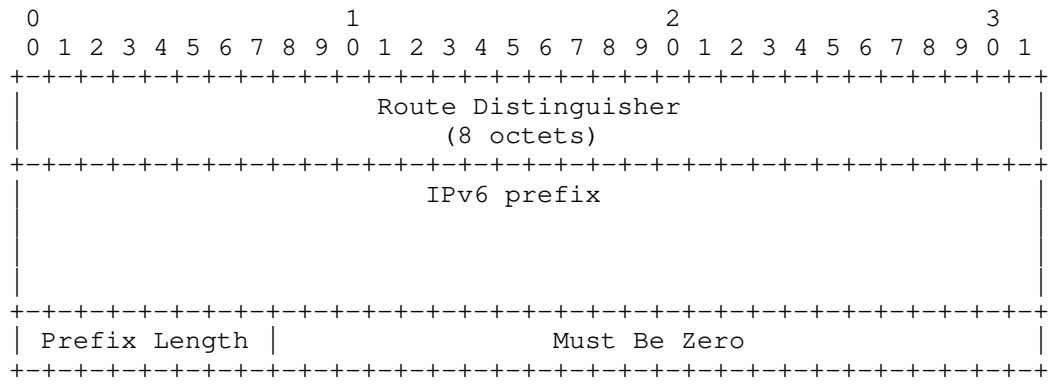


Figure 2: Classful Transport IPv6 FEC

13. Applicability to Network Slicing

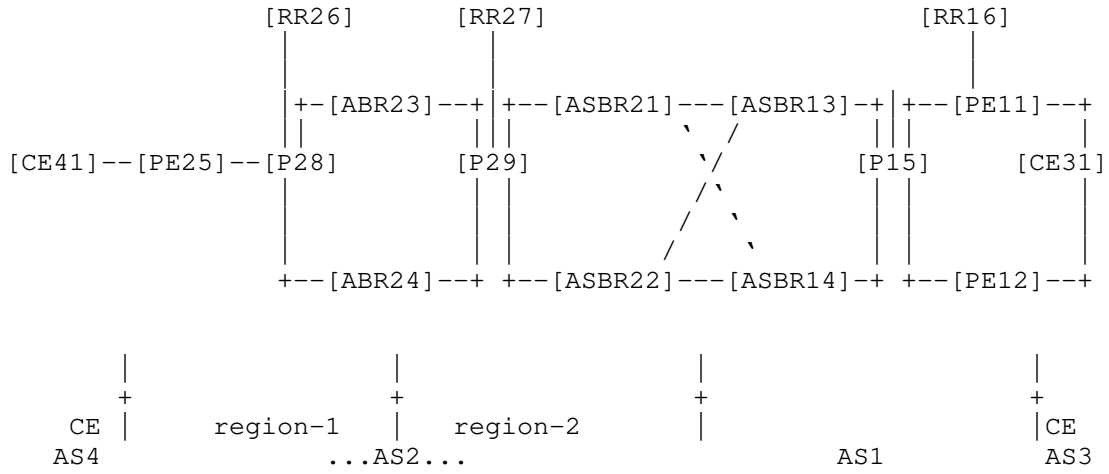
In Network Slicing, the Transport Slice Controller (TSC) sets up the Topology (e.g. RSVP, SR-TE tunnels with desired characteristics) and resources (e.g. polices/shapers) in a transport network to create a Transport slice. The Transport class construct described in this document represents the "Topology Slice" portion of this equation.

The TSC can use the Transport Class Identifier (Color value) to provision a transport tunnel in a specific Topology Slice.

Further, Network slice controller can use the Mapping community on the service route to map traffic to the desired Transport slice.

14. Illustration of procedures with example topology

14.1. Topology



This example shows a provider network that comprises of two Autonomous systems, AS1, AS2. They are serving customers AS3, AS4 respectively. Traffic direction being described is CE41 to CE31. CE31 may request a specific SLA, e.g. Gold for this traffic, when traversing these provider networks.

AS2 is further divided into two regions. So there are three tunnel domains in provider space. AS1 uses ISIS Flex-Algo intra-domain tunnels, whereas AS2 uses RSVP intra-domain tunnels.

The network has two Transport classes: Gold with transport class id 100, Bronze with transport class id 200. These transport classes are provisioned at the PEs and the Border nodes (ABRs, ASBRs) in the network.

Following tunnels exist for Gold transport class.

PE25_to_ABR23_gold - RSVP tunnel

PE25_to_ABR24_gold - RSVP tunnel

ABR23_to_ASBR22_gold - RSVP tunnel

ASBR13_to_PE11_gold - ISIS FlexAlgo tunnel

ASBR14_to_PE11_gold - ISIS FlexAlgo tunnel

Following tunnels exist for Bronze transport class.

PE25_to_ABR23_bronze - RSVP tunnel

ABR23_to_ASBR21_bronze - RSVP tunnel

ABR23_to_ASBR22_bronze - RSVP tunnel

ABR24_to_ASBR21_bronze - RSVP tunnel

ASBR13_to_PE12_bronze - ISIS FlexAlgo tunnel

ASBR14_to_PE11_bronze - ISIS FlexAlgo tunnel

These tunnels are either provisioned or auto-discovered to belong to transport class 100 or 200.

14.2. Service Layer route exchange

Service nodes PE11, PE12 negotiate service families (SAFI 1, 128) on the BGP session with RR16. Service helpers RR16, RR26 have multihop EBGP session to exchange service routes between the two AS. Similarly PE25 negotiates service families with RR26.

Forwarding happens using service routes at service nodes PE25, PE11, PE12 only. Routes received from CEs are not present in any other nodes' FIB in the network.

CE31 advertises a route for example prefix 31.31.31.31 with nexthop self to PE11, PE12. CE31 can attach a mapping community Color:0:100 on this route, to indicate its request for Gold SLA. Or, PE11 can attach the same using locally configured policies. Let us assume CE31 is getting VPN service from PE25.

The 31.31.31.31 route is readvertised in SAFI 128 by PE11 with nexthop self (1.1.1.1) and label V-L1, to RR16 with the mapping community Color:0:100 attached. This SAFI 128 route reaches PE25 via RR16, RR26 with the nexthop unchanged, as PE11 and label V-L1. Now PE25 can resolve the PNH 1.1.1.1 using transport routes received in BGP CT or BGP LU.

The IP FIB at PE25 will have a route for 31.31.31.31 with a nexthop thus found, that points to a Gold tunnel in ingress domain.

14.3. Transport Layer route propagation

ASBR13 negotiates BGP CT family with transport ASBRs ASBR21, ASBR22. They negotiate BGP CT family with RR27 in region 2. ABR23, ABR24 negotiate BGP CT family with RR27 in region 2 and RR26 in region 1. PE25 receives BGP CT routes from RR26. BGP LU family is also

negotiated on these sessions alongside BGP CT family. BGP LU carries "best effort" transport class routes, BGP CT carries gold, bronze transport class routes.

ASBR13 is provisioned with transport class 100, RD value 1.1.1.3:10 and a transport route target 0:100. And a Transport class 200 with RD value 1.1.1.3:20, and transport route target 0:200.

Similarly, these transoprt classes are also configured on ASBRs, ABRs and PEs, with same transport route target, but unique RDs.

Ingress route for ASBR13_to_PE11_gold is advertised by ASBR13 in BGP CT family to ASBRs ASBR21, ASBR22. This route is sent with a NLRI containing RD prefix 1.1.1.3:10:1.1.1.1, Label B-L1 and a route target extended community transport-target:0:100. MPLS swap route is installed at ASBR13 for B-L1 with a nexthop pointing to ASBR13_to_PE11_gold tunnel.

Ingress route for ASBR13_to_PE11_bronze is advertised by ASBR13 in BGP CT family to ASBRs ASBR21, ASBR22. This route is sent with a NLRI containing RD prefix 1.1.1.3:20:1.1.1.1, Label B-L2 and a route target extended community transport-target:0:200. MPLS swap route is installed at ASBR13 for label B-L2 with a nexthop pointing to ASBR13_to_PE11_bronze tunnel

ASBR21 receives BGP CT route 1.1.1.3:10:1.1.1.1 over the single hop EBGP sesion, and readvertises with nexthop self (loopback adderss 2.2.2.1) to RR27, advertising a new label B-L3. MPLS swap route is installed for label B-L3 at ASBR21 to swap to received label B-L1 and forwards to ASBR13. RR27 readvertises this BGP CT route to ABR23, ABR24.

ASBR22 receives BGP CT route 1.1.1.3:10:1.1.1.1 over the single hop EBGP sesion, and readvertises with nexthop self (loopback adderss 2.2.2.2) to RR27, advertising a new label B-L4. MPLS swap route is installed for label B-L4 at ASBR21 to swap to received label B-L2 and forwards to ASBR13. RR27 readvertises this BGP CT route to ABR23, ABR24.

Addpath is enabled for BGP CT family on the sessions between RR27 and ASBRs, ABRs. Such that routes for 1.1.1.3:10:1.1.1.1 with the nexthops ASBR21 and ASBR22 are reflected to ABR23, ABR24 without any path hiding. Thus giving ABR23 visibiity of both available nexthops for Gold SLA.

ABR23 receives the route with nexthop 2.2.2.1, label B-L3 from RR27. The route target "transport-target:0:100" on this route acts as mapping community, and instructs ABR23 to strictly resolve the

nexthop using transport class 100 routes only. ABR23 is unable to find a route for 2.2.2.1 with transport class 100. Thus it considers this route unusable and does not propagate it further. This prunes ASBR21 from Gold SLA tunneled path.

ABR23 also receives the route with nexthop 2.2.2.2, label B-L4 from RR27. The route target "transport-target:0:100" on this route acts as mapping community, and instructs ABR23 to strictly resolve the nexthop using transport class 100 routes only. ABR23 successfully resolves the nexthop to point to ABR23_to_ASBR22_gold tunnel. ABR23 readvertises this route with nexthop self (loopback address 2.2.2.3) and a new label B-L5 to RR26. Swap route for B-L5 is installed by ABR23 to swap to label B-L4, and forward into ABR23_to_ASBR22_gold tunnel.

RR26 reflects the route from ABR23 to PE25. PE25 receives the BGP CT route for prefix 1.1.1.3:10:1.1.1.1 with label B-L5, nexthop 2.2.2.3 and transport-target:0:100 from RR26. And it similarly resolves the nexthop 2.2.2.3 over transport class 100, pushing labels associated with PE25_to_ABR23_gold tunnel.

In this manner, the Gold transport LSP "ASBR13_to_PE11_gold" in egress-domain is extended by BGP CT until the ingress-node PE25 in ingress domain, to create an end-to-end Gold SLA path. MPLS swap routes are installed at ASBR13, ASBR22 and ABR23, when propagating the PE11 BGP CT Gold transport class route 1.1.1.3:10:1.1.1.1 with nexthop self towards PE25.

The BGP CT LSP thus formed, originates in PE25, and terminates in ASBR13, traversing over the Gold underlay LSPs in each domain. ASBR13 uses UHP to stitch the BGP CT LSP into the "ASBR13_to_PE11_gold" LSP to traverse the last domain, thus satisfying Gold SLA end-to-end.

When PE25 receives service route with nexthop 1.1.1.1 and mapping community Color:0:100, it resolves over this BGP CT route 1.1.1.3:10:1.1.1.1. Thus pushing label B-L5, and pushing as top label the labels associated with PE25_to_ABR23_gold tunnel.

14.4. Data plane view

14.4.1. Steady state

This section describes how the data plane looks like in steady state.

CE41 transmits an IP packet with destination as 31.31.31.31. On receiving this packet PE25 performs a lookup in the IP FIB associated with the CE41 interface. This lookup yeids the service route that

pushes the VPN service label V-L1, BGP CT label B-L5, and labels for PE25_to_ABR23_gold tunnel. Thus PE25 encapsulates the IP packet in MPLS packet with label V-L1(innermost), B-L5, and top label as PE25_to_ABR23_gold tunnel. This MPLS packet is thus transmitted to ABR23 using Gold SLA.

ABR23 decapsulates the packet received on PE25_to_ABR23_gold tunnel as required, and finds the MPLS packet with label B-L5. It performs lookup for label B-L5 in the global MPLS FIB. This yields the route that swaps label B-L5 with label B-L4, and pushes top label provided by ABR23_to_ASBR22_gold tunnel. Thus ABR23 transmits the MPLS packet with label B-L4 to ASBR22, on a tunnel that satisfies Gold SLA.

ASBR22 similarly performs a lookup for label B-L4 in global MPLS FIB, finds the route that swaps label B-L4 with label B-L2, and forwards to ASBR13 over the directly connected MPLS enabled interface. This interface is a common resource not dedicated to any specific transport class, in this example.

ASBR13 receives the MPLS packet with label B-L2, and performs a lookup in MPLS FIB, finds the route that pops label B-L2, and pushes labels associated with ASBR13_to_PE11_gold tunnel. This transmits the MPLS packet with VPN label V-L1 to PE11, using a tunnel that preserves Gold SLA in AS 1.

PE11 receives the MPLS packet with V-L1, and performs VPN forwarding. Thus transmitting the original IP payload from CE41 to CE31. The payload has traversed path satisfying Gold SLA end-to-end.

14.4.2. Absorbing failure of primary path

This section describes how the data plane reacts when gold path experiences a failure.

Let us assume tunnel ABR23_to_ASBR22_gold goes down, such that now end-to-end Gold path does not exist in the network. This makes the BGP CT route for RD prefix 1.1.1.1:10:1.1.1.1 unusable at ABR23. This makes ABR23 send a BGP withdrawal for 1.1.1.1:10:1.1.1.1 to RR26, which then withdraws the prefix from PE25.

Withdrawal for 1.1.1.1:10:1.1.1.1 allows PE25 to react to the loss of gold path to 1.1.1.1. Let us assume PE25 is provisioned to use best-effort transport class as the backup path. This withdrawal of BGP CT route allows PE25 to adjust the nexthop of the VPN Service-route to push the labels provided by the BGP LU route. That repairs the traffic to go via best effort path. PE25 can also be provisioned to use Bronze transport class as the backup path. The repair will happen in similar manner in that case as-well.

Traffic repair to absorb the failure happens at ingress node PE25, in a service prefix scale independent manner. This is called PIC (Prefix scale Independent Convergence). The repair time will be proportional to time taken for withdrawing the BGP CT route.

15. IANA Considerations

This document makes following requests of IANA.

15.1. New BGP SAFI

New BGP SAFI code for "Classful Transport". Value 76.

This will be used to create new AFI,SAFI pairs for IPv4, IPv6 Classful Transport families. viz:

- "Inet, Classful Transport". AFI/SAFI = "1/76" for carrying IPv4 Classful Transport prefixes.
- "Inet6, Classful Transport". AFI/SAFI = "2/76" for carrying IPv6 Classful Transport prefixes.

15.2. New Format for BGP Extended Community

Please assign a new Format (Type high = 0xa) of extended community EXT-COMM [RFC4360] called "Transport Class" from the following registries:

the "BGP Transitive Extended Community Types" registry, and

the "BGP Non-Transitive Extended Community Types" registry.

Please assign the same low-order six bits for both allocations.

This document uses this new Format with subtype 0x2 (route target), as a transitive extended community.

The Route Target thus formed is called "Transport Class" route target extended community.

Taking reference of RFC7153 [RFC7153] , following requests are made:

15.2.1. Existing registries to be modified

15.2.1.1. Registries for the "Type" Field

15.2.1.1.1. Transitive Types

This registry contains values of the high-order octet (the "Type" field) of a Transitive Extended Community.

Registry Name: BGP Transitive Extended Community Types

TYPE VALUE	NAME
+ 0x0a	Transitive Transport Class Extended Community (Sub-Types are defined in the "Transitive Transport Class Extended Community Sub-Types" registry)

15.2.1.1.2. Non-Transitive Types

This registry contains values of the high-order octet (the "Type" field) of a Non-transitive Extended Community.

Registry Name: BGP Non-Transitive Extended Community Types

TYPE VALUE	NAME
+ 0x4a	Non-Transitive Transport Class Extended Community (Sub-Types are defined in the "Non-Transitive Transport Class Extended Community Sub-Types" registry)

15.2.2. New registries to be created

15.2.2.1. Transitive "Transport Class" Extended Community Sub-Types Registry

This registry contains values of the second octet (the "Sub-Type" field) of an extended community when the value of the first octet (the "Type" field) is 0x07.

Registry Name: Transitive Transport Class Extended Community Sub-Types

RANGE	REGISTRATION PROCEDURE
0x00-0xBF	First Come First Served
0xC0-0xFF	IETF Review

SUB-TYPE VALUE	NAME
0x02	Route Target

15.2.2.2. Non-Transitive "Transport Class" Extended Community Sub-Types Registry

This registry contains values of the second octet (the "Sub-Type" field) of an extended community when the value of the first octet (the "Type" field) is 0x47.

Registry Name: Non-Transitive Transport Class Extended Community Sub-Types

RANGE	REGISTRATION PROCEDURE
0x00-0xBF	First Come First Served
0xC0-0xFF	IETF Review

SUB-TYPE VALUE	NAME
0x02	Route Target

15.3. MPLS OAM code points

The following two code points are sought for Target FEC Stack sub-TLVs:

- IPv4 BGP Classful Transport
- IPv6 BGP Classful Transport

16. Security Considerations

Mechanisms described in this document carry Transport routes in a new BGP address family. That minimizes possibility of these routes leaking outside the expected domain or mixing with service routes.

When redistributing between SAFI 4 and SAFI 76 Classful Transport routes, there is a possibility of SAFI 4 routes mixing with SAFI 1 service routes. To avoid such scenarios, it is RECOMMENDED that implementations support keeping SAFI 4 routes in a separate transport RIB, distinct from service RIB that contain SAFI 1 service routes.

17. Acknowledgements

The authors thank Jeff Haas, John Scudder, Navaneetha Krishnan, Ravi M R, Chandrasekar Ramachandran, Shradha Hegde, Richard Roberts, Krzysztof Szarkowicz, John E Drake, Srihari Sangli, Vijay Kestur, Santosh Kolenchery, Robert Raszuk, Ahmed Darwish for the valuable discussions and review comments.

The decision to not reuse SAFI 128 and create a new address-family to carry these transport-routes was based on suggestion made by Richard Roberts and Krzysztof Szarkowicz.

18. References

18.1. Normative References

[MPLS-NAMESPACES] Vairavakkalai, Ed., "Private MPLS-label namespaces", 08 2020, <https://tools.ietf.org/html/draft-kaliraj-bess-bgp-sig-private-mpls-labels-01#section-6.1>.

[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>.

[RFC4360] Sangli, S., Tappan, D., and Y. Rekhter, "BGP Extended Communities Attribute", RFC 4360, DOI 10.17487/RFC4360, February 2006, <https://www.rfc-editor.org/info/rfc4360>.

[RFC4364] Rosen, E. and Y. Rekhter, "BGP/MPLS IP Virtual Private Networks (VPNs)", RFC 4364, DOI 10.17487/RFC4364, February 2006, <https://www.rfc-editor.org/info/rfc4364>.

[RFC4456] Bates, T., Chen, E., and R. Chandra, "BGP Route Reflection: An Alternative to Full Mesh Internal BGP (IBGP)", RFC 4456, DOI 10.17487/RFC4456, April 2006, <https://www.rfc-editor.org/info/rfc4456>.

[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>.

[RFC7153] Rosen, E. and Y. Rekhter, "IANA Registries for BGP Extended Communities", RFC 7153, DOI 10.17487/RFC7153, March 2014, <https://www.rfc-editor.org/info/rfc7153>.

[RFC7911] Walton, D., Retana, A., Chen, E., and J. Scudder, "Advertisement of Multiple Paths in BGP", RFC 7911, DOI 10.17487/RFC7911, July 2016, <https://www.rfc-editor.org/info/rfc7911>.

[RFC8029] Kompella, K., Swallow, G., Pignataro, C., Ed., Kumar, N., Aldrin, S., and M. Chen, "Detecting Multiprotocol Label Switched (MPLS) Data-Plane Failures", RFC 8029, DOI 10.17487/RFC8029, March 2017, <https://www.rfc-editor.org/info/rfc8029>.

[RFC8212] Mauch, J., Snijders, J., and G. Hankins, "Default External BGP (EBGP) Route Propagation Behavior without Policies", RFC 8212, DOI 10.17487/RFC8212, July 2017, <https://www.rfc-editor.org/info/rfc8212>.

[RFC8277] Rosen, E., "Using BGP to Bind MPLS Labels to Address Prefixes", RFC 8277, DOI 10.17487/RFC8277, October 2017, <https://www.rfc-editor.org/info/rfc8277>.

[RFC8669] Previdi, S., Filsfils, C., Lindem, A., Ed., Sreekantiah, A., and H. Gredler, "Segment Routing Prefix Segment Identifier Extensions for BGP", RFC 8669, DOI 10.17487/RFC8669, December 2019, <https://www.rfc-editor.org/info/rfc8669>.

[RTC-Ext] Zhang, Z., Ed., "Route Target Constrain Extension", 07 2020, <https://tools.ietf.org/html/draft-zzhang-idr-bgp-rt-constrains-extension-00#section-2>.

[Seamless-SR] Hegde, Ed., "Seamless Segment Routing", 11 2020, <https://datatracker.ietf.org/doc/html/draft-hegde-spring-mpls-seamless-sr-03>.

[SRTE] Previdi, S., Ed., "Advertising Segment Routing Policies in BGP", 11 2019, <https://tools.ietf.org/html/draft-ietf-idr-segment-routing-te-policy-08>.

18.2. URIs

[1] https://www.rfc-editor.org/rfc/rfc4271#section-9.1.2.1

Authors' Addresses

Kaliraj Vairavakkalai
Juniper Networks, Inc.
1133 Innovation Way,
Sunnyvale, CA 94089
US

Email: kaliraj@juniper.net

Natrajan Venkataraman
Juniper Networks, Inc.
1133 Innovation Way,
Sunnyvale, CA 94089
US

Email: natv@juniper.net

Balaji Rajagopalan
Juniper Networks, Inc.
Electra, Exora Business Park˜Marathahalli - Sarjapur Outer Ring Road,
Bangalore, KA 560103
India

Email: balajir@juniper.net

Gyan Mishra
Verizon Communications Inc.
13101 Columbia Pike
Silver Spring, MD 20904
USA

Email: gyan.s.mishra@verizon.com

Mazen Khaddam
Cox Communications Inc.
Atlanta, GA
USA

Email: mazen.khaddam@cox.com

Xiaohu Xu
Alibaba Inc.
Beijing
China

Email: xiaohu.xxh@alibaba-inc.com

Rafal Jan Szarecki
Google.
1160 N Mathilda Ave, Bldg 5,
Sunnyvale,, CA 94089
USA

Email: szarecki@google.com

IDR
Internet-Draft
Intended status: Standards Track
Expires: April 5, 2021

F. Qin
China Mobile
H. Yuan
UnionPay
T. Zhou
G. Fioccola
Y. Wang
Huawei

October 2, 2020

BGP SR Policy Extensions to Enable IFIT
draft-qin-idr-sr-policy-ifit-04

Abstract

Segment Routing (SR) policy is a set of candidate SR paths consisting of one or more segment lists and necessary path attributes. It enables instantiation of an ordered list of segments with a specific intent for traffic steering. In-situ Flow Information Telemetry (IFIT) refers to network OAM data plane on-path telemetry techniques, in particular the most popular are In-situ OAM (IOAM) and Alternate Marking. This document defines extensions to BGP to distribute SR policies carrying IFIT information. So that IFIT methods can be enabled automatically when the SR policy is applied.

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 RFC 2119 [RFC2119].

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 April 5, 2021.

Table of Contents

1. Introduction

Segment Routing (SR) policy [I-D.ietf-spring-segment-routing-policy] is a set of candidate SR paths consisting of one or more segment lists and necessary path attributes. It enables instantiation of an ordered list of segments with a specific intent for traffic steering.

In-situ Flow Information Telemetry (IFIT) denotes a family of flow-oriented on-path telemetry techniques (e.g. IOAM, Alternate Marking), which can provide high-precision flow insight and real-time network issue notification (e.g., jitter, latency, packet loss).In

particular, IFIT refers to network OAM data plane on-path telemetry techniques, including In-situ OAM (IOAM) [I-D.ietf-ippm-ioam-data] and Alternate Marking [RFC8321]. It can provide flow information on the entire forwarding path on a per-packet basis in real time.

An automatic network requires the Service Level Agreement (SLA) monitoring on the deployed service. So that the system can quickly detect the SLA violation or the performance degradation, hence to change the service deployment. For this reason, the SR policy native IFIT can facilitate the closed loop control and enable the automation of SR service.

This document defines extensions to Border Gateway Protocol (BGP) to distribute SR policies carrying IFIT information. So that IFIT behavior can be enabled automatically when the SR policy is applied.

This BGP extension allows to signal the IFIT capabilities together with the SR-policy. In this way IFIT methods are automatically activated and running. The flexibility and dynamicity of the IFIT applications are given by the use of additional functions on the controller and on the network nodes, but this is out of scope here.

2. Motivation

IFIT Methods are being introduced in multiple protocols and below is a proper picture of the relevant documents for Segment Routing. Indeed the IFIT methods are becoming mature for Segment Routing over the MPLS data plane (SR-MPLS) and Segment Routing over IPv6 data plane (SRv6), that is the main focus of this draft:

IOAM: the reference documents for the data plane are [I-D.ietf-ippm-ioam-ipv6-options] for SRv6 and [I-D.gandhi-mpls-ioam-sr] for SR-MPLS.

Alternate Marking: the reference documents for the data plane are [I-D.ietf-6man-ipv6-alt-mark] for SRv6 and [I-D.ietf-mpls-rfc6374-sfl], [I-D.gandhi-mpls-rfc6374-sr] for SR-MPLS.

The definition of these data plane IFIT methods for SR-MPLS and SRv6 imply requirements for various routing protocols, such as BGP, and this document aims to define BGP extensions to distribute SR policies carrying IFIT information. This allows to signal the IFIT capabilities so IFIT methods are automatically configured and ready to run when the SR Policy candidate paths are distributed through BGP.

It is to be noted that, for PCEP, [I-D.chen-pce-pcep-ifit] proposes the extensions to PCEP to distribute paths carrying IFIT information and therefore to enable IFIT methods for SR policy too.

3. IFIT methods for SR Policy

In-situ Operations, Administration, and Maintenance (IOAM) [I-D.ietf-ippm-ioam-data] records operational and telemetry information in the packet while the packet traverses a path between two points in the network. In terms of the classification given in RFC 7799 [RFC7799] IOAM could be categorized as Hybrid Type 1. IOAM mechanisms can be leveraged where active OAM do not apply or do not offer the desired results. When SR policy enables the IOAM, the IOAM header will be inserted into every packet of the traffic that is steered into the SR paths.

The Alternate Marking [RFC8321]technique is an hybrid performance measurement method, per RFC 7799 [RFC7799] classification of measurement methods. Because this method is based on marking consecutive batches of packets. It can be used to measure packet loss, latency, and jitter on live traffic.

This document aims to define the control plane. While the relevant documents for the data plane application of IOAM and Alternate Marking are respectively [I-D.ietf-ippm-ioam-ipv6-options] and [I-D.ietf-6man-ipv6-alt-mark] for Segment Routing over IPv6 data plane (SRv6).

4. IFIT Attributes in SR Policy

As defined in [I-D.ietf-idr-segment-routing-te-policy], the SR Policy encoding structure is as follows:

```
SR Policy SAFI NLRI: <Distinguisher, Policy-Color, Endpoint>
Attributes:
   Tunnel Encaps Attribute (23)
      Tunnel Type: SR Policy
          Binding SID
          Preference
          Priority
          Policy Name
          Explicit NULL Label Policy (ENLP)
          Segment List
              Weight
              Segment
              Segment
              ...
          ...
```

A candidate path includes multiple SR paths, each of which is specified by a segment list. IFIT can be applied to the candidate path, so that all the SR paths can be monitored in the same way. The new SR Policy encoding structure is expressed as below:

```
SR Policy SAFI NLRI: <Distinguisher, Policy-Color, Endpoint>
Attributes:
   Tunnel Encaps Attribute (23)
      Tunnel Type: SR Policy
          Binding SID
          Preference
          Priority
          Policy Name
          Explicit NULL Label Policy (ENLP)
          IFIT Attributes
          Segment List
              Weight
              Segment
              Segment
              ...
          ...
```

IFIT attributes can be attached at the candidate path level as sub-TLVs. There may be different IFIT tools. The following sections will describe the requirement and usage of different IFIT tools, and define the corresponding sub-TLV encoding in BGP.

Note that the IFIT attributes here described can also be generalized and included as sub-TLVs for other SAFIs and NLRIs.

5. IFIT Attributes Sub-TLV

The format of the IFIT Attributes Sub-TLV is defined as follows:

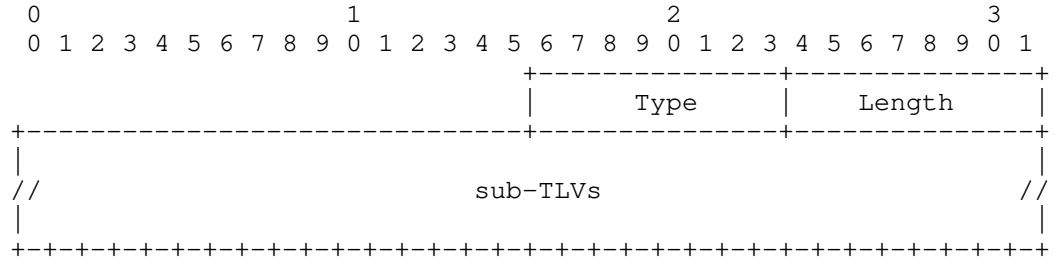


Fig. 1 IFIT Attributes Sub-TLV

Where:

Type: to be assigned by IANA.

Length: the total length of the value field not including Type and Length fields.

sub-TLVs currently defined:

* IOAM Pre-allocated Trace Option Sub-TLV,

* IOAM Incremental Trace Option Sub-TLV,

* IOAM Directly Export Option Sub-TLV,

* IOAM Edge-to-Edge Option Sub-TLV,

* Enhanced Alternate Marking (EAM) sub-TLV.

The presence of the IFIT Attributes Sub-TLV implies support of IFIT methods (IOAM and/or Alternate Marking). It is worth mentioning that IOAM and Alternate Marking can be activated one at a time or can coexist; so it is possible to have only IOAM or only Alternate Marking enabled as Sub-TLVs. The sub-TLVs currently defined for IOAM and Alternate Marking are detailed in the next sections.

5.1. IOAM Pre-allocated Trace Option Sub-TLV

The IOAM tracing data is expected to be collected at every node that a packet traverses to ensure visibility into the entire path a packet takes within an IOAM domain. The preallocated tracing option will create pre-allocated space for each node to populate its information.

The format of IOAM pre-allocated trace option sub-TLV is defined as follows:

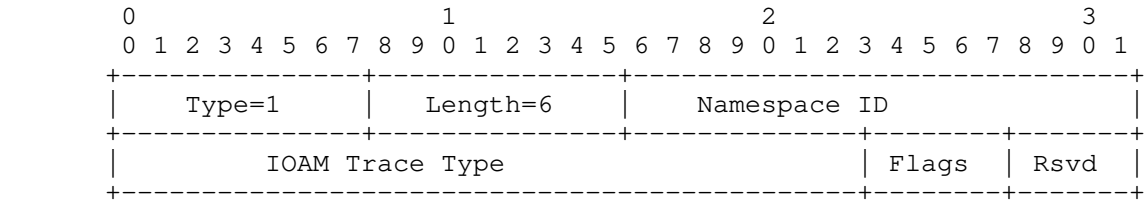


Fig. 2 IOAM Pre-allocated Trace Option Sub-TLV

Where:

Type: 1 (to be assigned by IANA).

Length: 6, it is the total length of the value field (not including Type and Length fields).

Namespace ID: A 16-bit identifier of an IOAM-Namespace. The definition is the same as described in section 4.4 of [I-D.ietf-ippm-ioam-data].

IOAM Trace Type: A 24-bit identifier which specifies which data types are used in the node data list. The definition is the same as described in section 4.4 of [I-D.ietf-ippm-ioam-data].

Flags: A 4-bit field. The definition is the same as described in [I-D.ietf-ippm-ioam-flags] and section 4.4 of [I-D.ietf-ippm-ioam-data].

Rsvd: A 4-bit field reserved for further usage. It MUST be zero.

5.2. IOAM Incremental Trace Option Sub-TLV

The incremental tracing option contains a variable node data fields where each node allocates and pushes its node data immediately following the option header.

The format of IOAM incremental trace option sub-TLV is defined as follows:

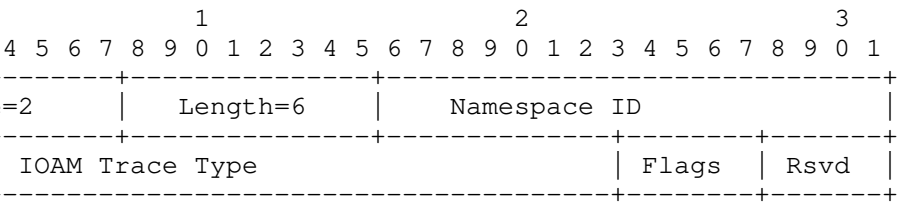


Fig. 3 IOAM Incremental Trace Option Sub-TLV

Where:

Type: 2 (to be assigned by IANA).

Length: 6, it is the total length of the value field (not including Type and Length fields).

All the other fields definistion is the same as the pre-allocated trace option sub-TLV in section 4.1.

5.3. IOAM Directly Export Option Sub-TLV

IOAM directly export option is used as a trigger for IOAM data to be directly exported to a collector without being pushed into in-flight data packets.

The format of IOAM directly export option sub-TLV is defined as follows:

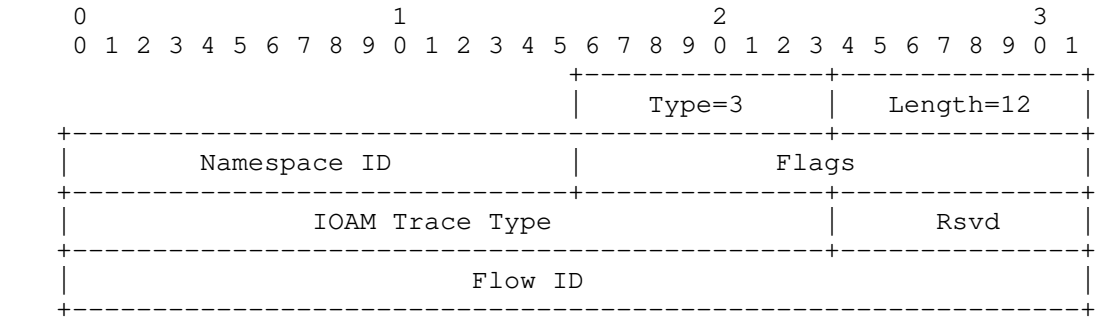


Fig. 4 IOAM Directly Export Option Sub-TLV

Where:

Type: 3 (to be assigned by IANA).

Length: 12, it is the total length of the value field (not including Type and Length fields).

Namespace ID: A 16-bit identifier of an IOAM-Namespace. The definition is the same as described in section 4.4 of [I-D.ietf-ippm-ioam-data].

IOAM Trace Type: A 24-bit identifier which specifies which data types are used in the node data list. The definition is the same as described in section 4.4 of [I-D.ietf-ippm-ioam-data].

Flags: A 16-bit field. The definition is the same as described in section 3.2 of [I-D.ietf-ippm-ioam-direct-export].

Flow ID: A 32-bit flow identifier. The definition is the same as described in section 3.2 of [I-D.ietf-ippm-ioam-direct-export].

Rsvd: A 4-bit field reserved for further usage. It MUST be zero.

5.4. IOAM Edge-to-Edge Option Sub-TLV

The IOAM edge to edge option is to carry data that is added by the IOAM encapsulating node and interpreted by IOAM decapsulating node.

The format of IOAM edge-to-edge option sub-TLV is defined as follows:

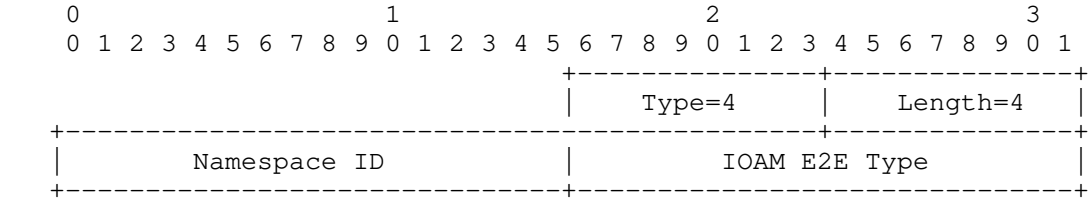


Fig. 5 IOAM Edge-to-Edge Option Sub-TLV

Where:

Type: 4 (to be assigned by IANA).

Length: 4, it is the total length of the value field (not including Type and Length fields).

Namespace ID: A 16-bit identifier of an IOAM-Namespace. The definition is the same as described in section 4.6 of [I-D.ietf-ippm-ioam-data].

IOAM E2E Type: A 16-bit identifier which specifies which data types are used in the E2E option data. The definition is the same as described in section 4.6 of [I-D.ietf-ippm-ioam-data].

5.5. Enhanced Alternate Marking (EAM) sub-TLV

The format of Enhanced Alternate Marking (EAM) sub-TLV is defined as follows:

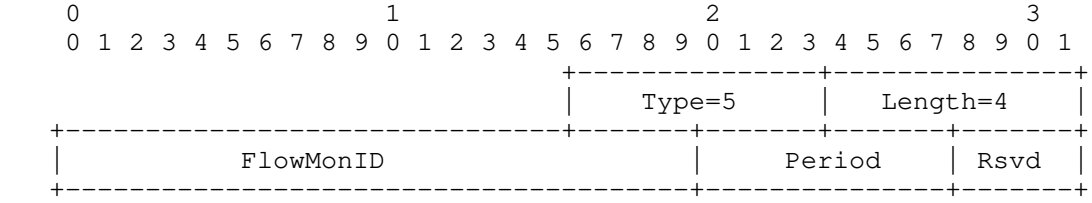


Fig. 6 Enhanced Alternate Marking Sub-TLV

Where:

Type: 5 (to be assigned by IANA).

Length: 4, it is the total length of the value field (not including Type and Length fields).

FlowMonID: A 20-bit identifier to uniquely identify a monitored flow within the measurement domain. The definition is the same as described in section 5.3 of [I-D.ietf-6man-ipv6-alt-mark].

Period: Time interval between two alternate marking period. The unit is second.

Rsvd: A 4-bit field reserved for further usage. It MUST be zero.

6. SR Policy Operations with IFIT Attributes

The details of SR Policy installation and use are specified in [I-D.ietf-spring-segment-routing-policy]. This document complements SR Policy Operations described in [I-D.ietf-idr-segment-routing-te-policy] by adding the IFIT Attributes.

The operations described in [I-D.ietf-idr-segment-routing-te-policy] are always valid. The only difference is the addition of IFIT Attributes Sub-TLVs for the SR Policy NLRI, that can affect its acceptance by a BGP speaker, but the implementation MAY provide an option for ignoring the unrecognized or unsupported IFIT sub-TLVs. SR Policy NLRIs that have been determined acceptable, usable and valid can be evaluated for propagation, including the IFIT information.

The error handling actions are also described in [I-D.ietf-idr-segment-routing-te-policy].

The validation of the IFIT Attributes sub-TLVs introduced in this document MUST be performed to determine if they are malformed or invalid. The validation of the individual fields of the IFIT Attributes sub-TLVs are handled by the SRPM (SR Policy Module).

7. IANA Considerations

This document defines a new sub-TLV in the registry "BGP Tunnel Encapsulation Attribute sub-TLVs" to be assigned by IANA:

Codepoint	Description	Reference
TBD1	IFIT Attributes Sub-TLV	This document

This document requests creation of a new registry called "IFIT Attributes Sub-TLVs". The allocation policy of this registry is "Specification Required" according to RFC 8126 [RFC8126].

Following initial Sub-TLV codepoints are assigned by this document:

Value	Description	Reference
1	IOAM Pre-allocated Trace Option Sub-TLV	This document
2	IOAM Incremental Trace Option Sub-TLV	This document
3	IOAM Directly Export Option Sub-TLV	This document
4	IOAM Edge-to-Edge Option Sub-TLV	This document
5	Enhanced Alternate Marking Sub-TLV	This document

8. Security Considerations

The security mechanisms of the base BGP security model apply to the extensions described in this document as well. See the Security Considerations section of [I-D.ietf-idr-segment-routing-te-policy].

SR operates within a trusted SR domain RFC 8402 [RFC8402] and its security considerations also apply to BGP sessions when carrying SR Policy information. The isolation of BGP SR Policy SAFI peering sessions may be used to ensure that the SR Policy information is not advertised outside the SR domain. Additionally, only trusted nodes (that include both routers and controller applications) within the SR domain must be configured to receive such information.

Implementation of IFIT methods (IOAM and Alternate Marking) are mindful of security and privacy concerns, as explained in [I-D.ietf-ippm-ioam-data] and RFC 8321 [RFC8321]. Anyway incorrect IFIT parameters in the BGP extension SHOULD not have an adverse effect on the SR Policy as well as on the network, since it affects only the operation of the telemetry methodology.

9. Acknowledgements

The authors of this document would like to thank Ketan Talaulikar, Joel Halpern, Jie Dong for their comments and review of this document.

10. References

10.1. Normative References

[I-D.ietf-6man-ipv6-alt-mark]
Fioccola, G., Zhou, T., Cociglio, M., Qin, F., and R. Pang, "IPv6 Application of the Alternate Marking Method", draft-ietf-6man-ipv6-alt-mark-01 (work in progress), June 2020.

[I-D.ietf-idr-segment-routing-te-policy]
Previdi, S., Filsfils, C., Talaulikar, K., Mattes, P., Rosen, E., Jain, D., and S. Lin, "Advertising Segment Routing Policies in BGP", draft-ietf-idr-segment-routing-te-policy-09 (work in progress), May 2020.

[I-D.ietf-ippm-ioam-data]
Brockners, F., Bhandari, S., and T. Mizrahi, "Data Fields for In-situ OAM", draft-ietf-ippm-ioam-data-10 (work in progress), July 2020.

[I-D.ietf-ippm-ioam-direct-export]
Song, H., Gafni, B., Zhou, T., Li, Z., Brockners, F., Bhandari, S., Sivakolundu, R., and T. Mizrahi, "In-situ OAM Direct Exporting", draft-ietf-ippm-ioam-direct-export-01 (work in progress), August 2020.

[I-D.ietf-ippm-ioam-flags]
Mizrahi, T., Brockners, F., Bhandari, S., Sivakolundu, R., Pignataro, C., Kfir, A., Gafni, B., Spiegel, M., and J. Lemon, "In-situ OAM Flags", draft-ietf-ippm-ioam-flags-02 (work in progress), July 2020.

[I-D.ietf-ippm-ioam-ipv6-options]
Bhandari, S., Brockners, F., Pignataro, C., Gredler, H., Leddy, J., Youell, S., Mizrahi, T., Kfir, A., Gafni, B., Lapukhov, P., Spiegel, M., Krishnan, S., Asati, R., and M. Smith, "In-situ OAM IPv6 Options", draft-ietf-ippm-ioam-ipv6-options-03 (work in progress), September 2020.

[I-D.ietf-spring-segment-routing-policy]
Filsfils, C., Talaulikar, K., Voyer, D., Bogdanov, A., and P. Mattes, "Segment Routing Policy Architecture", draft-ietf-spring-segment-routing-policy-08 (work in progress), July 2020.

[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>.

[RFC7799] Morton, A., "Active and Passive Metrics and Methods (with Hybrid Types In-Between)", RFC 7799, DOI 10.17487/RFC7799, May 2016, <https://www.rfc-editor.org/info/rfc7799>.

[RFC8126] Cotton, M., Leiba, B., and T. Narten, "Guidelines for Writing an IANA Considerations Section in RFCs", BCP 26, RFC 8126, DOI 10.17487/RFC8126, June 2017, <https://www.rfc-editor.org/info/rfc8126>.

[RFC8321] Fioccola, G., Ed., Capello, A., Cociglio, M., Castaldelli, L., Chen, M., Zheng, L., Mirsky, G., and T. Mizrahi, "Alternate-Marking Method for Passive and Hybrid Performance Monitoring", RFC 8321, DOI 10.17487/RFC8321, January 2018, <https://www.rfc-editor.org/info/rfc8321>.

[RFC8402] Filsfils, C., Ed., Previdi, S., Ed., Ginsberg, L., Decraene, B., Litkowski, S., and R. Shakir, "Segment Routing Architecture", RFC 8402, DOI 10.17487/RFC8402, July 2018, <https://www.rfc-editor.org/info/rfc8402>.

10.2. Informative References

[I-D.chen-pce-pcep-ifit]
Chen, H., Yuan, H., Zhou, T., Li, W., Fioccola, G., and Y. Wang, "Path Computation Element Communication Protocol (PCEP) Extensions to Enable IFIT", draft-chen-pce-pcep-ifit-01 (work in progress), September 2020.

[I-D.gandhi-mpls-ioam-sr]
Gandhi, R., Ali, Z., Filsfils, C., Brockners, F., Wen, B., and V. Kozak, "MPLS Data Plane Encapsulation for In-situ OAM Data", draft-gandhi-mpls-ioam-sr-03 (work in progress), September 2020.

[I-D.gandhi-mpls-rfc6374-sr]
Gandhi, R., Filsfils, C., Voyer, D., Salsano, S., and M. Chen, "Performance Measurement Using RFC 6374 for Segment Routing Networks with MPLS Data Plane", draft-gandhi-mpls-rfc6374-sr-05 (work in progress), June 2020.

[I-D.ietf-mpls-rfc6374-sfl]
Bryant, S., Swallow, G., Chen, M., Fioccola, G., and G. Mirsky, "RFC6374 Synonymous Flow Labels", draft-ietf-mpls-rfc6374-sfl-07 (work in progress), June 2020.

Appendix A.

Authors' Addresses

Fengwei Qin
China Mobile
No. 32 Xuanwumenxi Ave., Xicheng District
Beijing
China

Email: qinfengwei@chinamobile.com

Hang Yuan
UnionPay
1899 Gu-Tang Rd., Pudong
Shanghai
China

Email: yuanhang@unionpay.com

Tianran Zhou
Huawei
156 Beiqing Rd., Haidian District
Beijing
China

Email: zhoutianran@huawei.com

Giuseppe Fioccola
Huawei
Riesstrasse, 25
Munich
Germany

Email: giuseppe.fioccola@huawei.com

Yali Wang
Huawei
156 Beiqing Rd., Haidian District
Beijing
China

Email: wangyali11@huawei.com

IDR Working Group
Internet-Draft
Intended status: Standards Track
Expires: May 27, 2021

W. Wang
A. Wang
China Telecom
H. Wang
Huawei Technologies
G. Mishra
Verizon Inc.
S. Zhuang
J. Dong
Huawei Technologies

November 23, 2020

Route Distinguisher Outbound Route Filter (RD-ORF) for BGP-4
draft-wang-idr-rd-orf-05

Abstract

This draft defines a new Outbound Route Filter (ORF) type, called the Route Distinguisher ORF (RD-ORF). RD-ORF is applicable when the routers do not exchange VPN routing information directly (e.g. routers in single-domain connect via Route Reflector, or routers in Option B/Option AB/Option C cross-domain scenario).

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 May 27, 2021.

Table of Contents

1. Introduction

With the rapid growth of network scale, Route Reflector is introduced in order to reduce the network complexity. Routers in the same Autonomous System only need to establish iBGP session with RR to transmit routes.

In VPN scenario shown in Figure 1, PE1 - PE4 establish IBGP sessions with RR to ensure the routes can be transmitted within AS100, where PE1 and PE3 maintain VRFs of VPN1 and VPN2, PE2 maintains VPN1's VRF and PE4 maintains VPN2's VRF. RR don not maintain any VRFs.

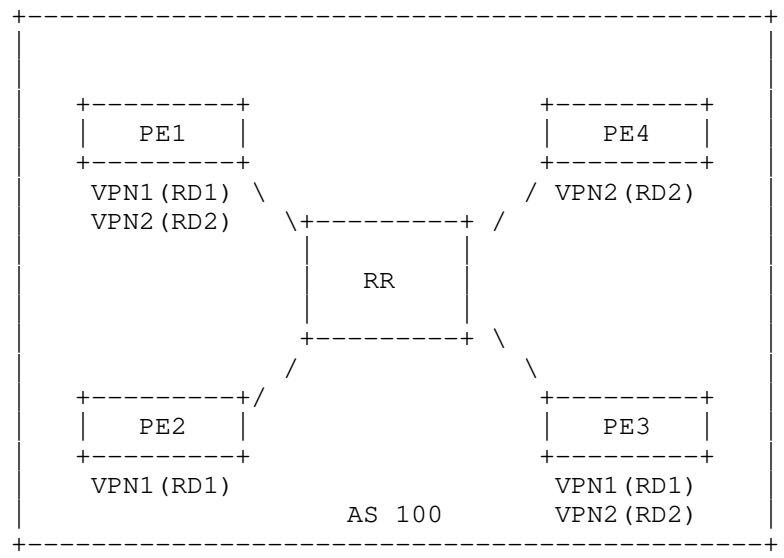


Figure 1: Single-domain scenario

When the VRF of VPN1 in PE1 overflows, due to PE1 and other PEs are not iBGP neighbors, BGP Maximum Prefix Features cannot work, so the problem on PE2 cannot be known.

Now, there are several solutions can be used to alleviate this problem:

- Route Target Constraint (RTC) as defined in [RFC4684]
- Address Prefix ORF as defined in [RFC5292]
- PE-CE edge peer Maximum Prefix
- Configure the Maximum Prefix for each VRF on edge nodes

However, there are limitations to existing solutions:

1) Route Target Constraint

RTC can only filter the VPN routes from the uninterested VRFs, if the "trashing routes" come from the interested VRF, filter on RTs will erase all prefixes from this VRF.

2) Address Prefix ORF

Using Address Prefix ORF to filter VPN routes need to pre-configuration, but it is impossible to know which prefix may cause overflow in advance.

3) PE-CE edge peer Maximum Prefix

This mechanism can only protect the edge between PE-CE, it can't be deployed within PE that peered via RR. Depending solely on the edge protection is dangerous, because if only one of the edge points being comprised/error-configured/attacked, then all of PEs within domain are under risk.

4) Configure the Maximum Prefix for each VRF on edge nodes

When a VRF overflows, it stops the import of routes and log the extra VPN routes into its RIB. However, PEs should parse the BGP updates. These processes will cost CPU cycles and further burden the overflowing PE.

This draft defines a new ORF-type, called the Route Distinguisher ORF (RD-ORF). Using RD-ORF mechanism, VPN routes of a VPN can be controlled based on source RD. This mechanism is event-driven and does not need to be pre-configured. When a VRF of a router overflows, the router will find out the main source RD of VPN routes in this VRF, and send a RD-ORF to its BGP peer that carrys the RD. If a BGP speaker receives a RD-ORF from its BGP peer, it will filter the VPN routes it tends to send according to the RD-ORF entry.

RD-ORF is applicable when the routers do not exchange VPN routing information directly (e.g. routers in single-domain connect via Route Reflector, or routers in Option B/Option AB/Option C cross-domain scenario).

2. Conventions used in this document

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] .

3. Terminology

The following terms are defined in this draft:

- RD: Route Distinguisher, defined in [RFC4364]
- ORF: Outbound Route Filter, defined in [RFC5291]
- AFI: Address Family Identifier, defined in [RFC4760]

- SAFI: Subsequent Address Family Identifier, defined in [RFC4760]
- EVPN: BGP/MPLS Ethernet VPN, defined in [RFC7432]
- RR: Router Reflector, provides a simple solution to the problem of IBGP full mesh connection in large-scale IBGP implementation.
- VRF: Virtual Routing Forwarding, a virtual routing table based on VPN instance.

4. RD-ORF Encoding

In this draft, we defined a new ORF type called Route Distinguisher Outbound Route Filter (RD-ORF). The ORF entries are carried in the BGP ROUTE-REFRESH message as defined in [RFC5291]. A BGP ROUTE-REFRESH message can carry one or more ORF entries, and MUST be regenerated when it is tended to be sent to other BGP peers. The ROUTE-REFRESH message which carries ORF entries contains the following fields:

- AFI (2 octets)
- SAFI (1 octet)
- When-to-refresh (1 octet): the value is IMMEDIATE or DEFER
- ORF Type (1 octet)
- Length of ORF entries (2 octets)

A RD-ORF entry contains a common part and type-specific part. The common part is encoded as follows:

- Action (2 bits): the value is ADD, REMOVE or REMOVE-ALL
- Match (1 bit): the value is PERMIT or DENY
- Reserved (5 bits)

RD-ORF also contains type-specific part. The encoding of the type-specific part is shown in Figure 2.

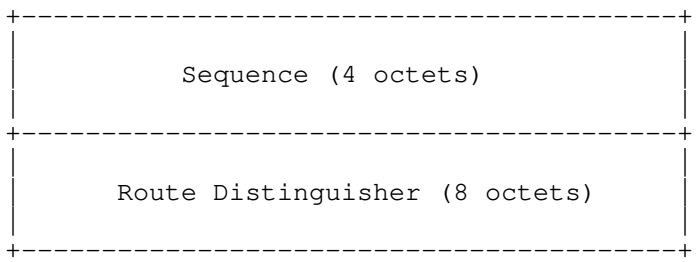


Figure 2: RD-ORF type-specific encoding

- Sequence: identifying the order in which RD-ORF is generated
- Route Distinguisher: distinguish the different user routes. The RD-ORF filters the VPN routes it tends to send based on Route Distinguisher.

Note that if the Action component of an ORF entry specifies REMOVE-ALL, the ORF entry does not include the type-specific part.

When the BGP ROUTE-REFRESH message carries RD-ORF entries, it must be set as follows:

- The ORF-Type MUST be set to RD-ORF.
- The AFI MUST be set to IPv4, IPv6, or Layer 2 VPN (L2VPN).
- If the AFI is set to IPv4 or IPv6, the SAFI MUST be set to MPLS-labeled VPN address.
- If the AFI is set to L2VPN, the SAFI MUST be set to BGP EVPN.
- The Match field MUST be equal to DENY.

5. Application in single-domain scenario

5.1. Addition of RD-ORF entries

The operation of RD-ORF mechanism on each device is independent, each of them makes a local judgement to determine whether it needs to send RD-ORF to its peers.

In general, every VRF on PE is configured a Maximum Prefix, the trigger of RD-ORF mechanism can be set as the number of VPN routes in VRF reach 80% of the Maximum Prefix. For RR, it doesn't have VRF and the machanism can be triggered by other conditions, such as the RR's memory/CPU utilization reaches 80%.

When the RD-ORF mechanism is triggered, the device must send an alarm information to network operators.

5.1.1. Operation process of RD-ORF mechanism on source PE

In scenario shown in Figure 1, when the VRF of VPN1 in PE1 overflows, PE1 will do analysis and calculation locally to find out the main source of VPN routes in this VRF, assuming it is PE3. Then, PE1 will resolve the corresponding RD of VPN routes from BGP UPDATE message, and generate a BGP ROUTE-REFRESH message contains a RD-ORF entry, and send it to RR. The message contains the following fields:

- AFI is set to IPv4 , IPv6 or L2 VPN
- SAFI is set to "MPLS-labeled VPN address" or "BGP EVPN"
- When-to-refresh is set to IMMEDIATE
- ORF Type is set to RD-ORF
- Length of ORF entries depends on the type of Source Address sub-TLV (21, 23 or 33 octets)
- Action is set to ADD
- Match is set to DENY
- Sequence is set to 1
- Route Distinguisher is set to RD1

It noted that the Sequence can uniquely identifies an RD-ORF entry. All VRFs share the sequence field, and the corresponding sequence of RD-ORF sent by each VRF will be recorded on the device.

Sometimes, several VRFs in a PE may import VPN routes carries the same RT, as shown in Figure 3.

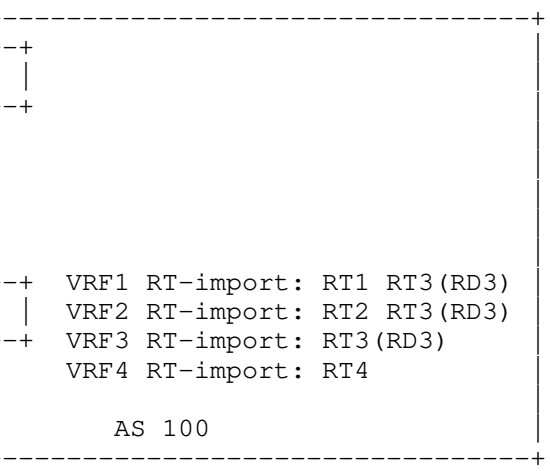


Figure 3: The scenario of several VRFs in a PE import VPN routes carries the same RT

In this scenario, VRF1, VRF2 and VRF3 import VPN routes carries RT3, which contains RD3. VRF1, VRF2 and VRF3 have different maximum prefix. When the VPN routes carrying RT3 cause the overflow of VRF3, PE will send a BGP ROUTE-REFRESH message containing a RD-ORF entry to RR, which Route Distinguisher field is equal to RD3. RR will stop sending associated VPN routes to PE. However, this will cause VRF1 to fail to receive VPN routes containing RD3.

The local determination of the PE can be used to inhibit the PE from sending RD-ORF entries. When the resources of the device are not exhausted, only prevent the overflowed VRF from importing related VPN routes without sending RD-ORF, unless all the VRFs that import the RD overflow.

5.1.2. Operation process of RD-ORF mechanism on RR

When RR receives the ROUTE-REFRESH message, it checks <AFI/SAFI, ORF-Type, Sequence, Route Distinguisher> to find whether it received the latest entry or not. If not, RR will discard the entry; otherwise, RR will add the RD-ORF entry into its Adj-RIB-out.

Before sending a VPN route toward PE1, RR will check its Adj-RIB-out and find there is a filter associated with RD1. Then, RR will stop sending that VPN route to PE1.

If the processing capacity of RR reaches the limit (e.g. RR's memory/CPU utilization reaches 80%), RR will find out the peer that sends the most routing entries to it, assuming it is PE3. Then, RR

will generate a BGP ROUTE-REFRESH message contains a RD-ORF entry based on the result of calculation, and send it to PE3.

5.1.3. Operation process of RD-ORF mechanism on target PE

After receiving the ROUTE-REFRESH message that carries a RD-ORF entry, PE3 will check if it receives the latest entry. If not, PE3 will discard it; otherwise, PE3 will add the RD-ORF entry into its Adj-RIB-out.

Before sending a VPN route toward RR, PE3 will check its Adj-RIB-out and find the RD-ORF entry prevent it from sending VPN route which carries RD1 to RR. Then, PE3 will stop sending that VPN route.

The BGP Maximum Prefix Features can be configured to protect PE-CE peering at the edge. Therefore, in general, CEs will not cause the overflow of PEs. If the boundary protection measures fail and cause the overflow, the PE can calculate and find the CEs in corresponding VRF, and break down the associated BGP sessions.

5.2. Withdraw of RD-ORF entries

When the RD-ORF mechanism is triggered, the alarm information will be generated and sent to the network operators. Operators should manually configure the network to resume normal operation. Due to devices can record the RD-ORF entries sent by each VRF, operators can find the entries needs to be withdrawn, and trigger the withdraw process as described in [RFC5291] manually. After returning to normal, the device sends withdraw ORF entries to its peers who have previously received ORF entries.

6. Applications in cross-domain scenarios

6.1. Application in Option B/Option AB cross-domain scenario

The Option B/Option AB cross-domain scenario is shown in Figure 4:

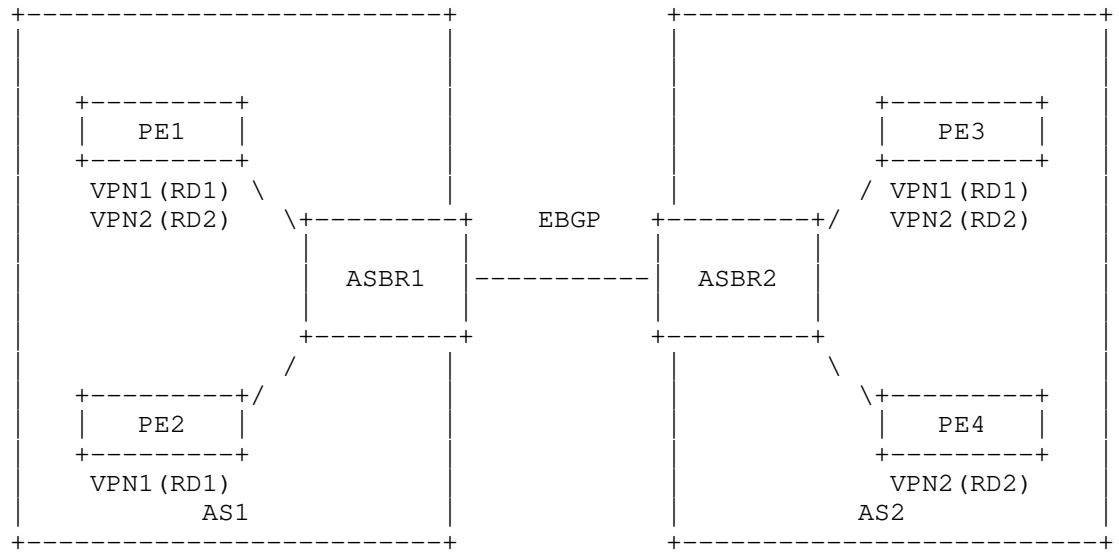


Figure 4: The Option B/Option AB cross-domain scenario

In Option B cross-domain scenario, PE1 - PE4 are responsible for maintaining VPN routing information in AS1 and AS2. There is a direct link between ASBR1 and ASBR2 via EBGP. In AS1, PE1 and PE2 establish IBGP sessions with ASBR1 to ensure the routes can be transmitted in AS1. In AS2, PE3 and PE4 establish IBGP session with ASBR2.

Due to the maintenance of VPN routes is only done by PEs. ASBRs cannot know whether the PEs' ability to handle VPN routes has reached the upper limit or not, so it needs the RD-ORF to control the number of routes.

Assume that PE1 - PE4 can transmit VPN routes through the network architecture shown in Figure 4. When the VRF of VPN1 in PE1 overflows, the RD-ORF mechanism will be implemented as follows:

1) PE1 will check and find out the main source of VPN routes in this VRF is PE3. Then, PE1 will resolve the corresponding RD from BGP UPDATE message, and generate a BGP ROUTE-REFRESH message contains an RD-ORF entry, and send it to ASBR1.

2) When ASBR1 receives the ROUTE-REFRESH message, it checks whether it receives the latest RD-ORF entry. If not, ASBR1 will discard the entry; Otherwise, ASBR1 will add the RD-ORF entry into its Adj-RIB-out.

Before sending a VPN route toward PE1, RR will check its Adj-RIB-out and find there is a filter associated with RD1. Then, ASBR1 will stop sending that VPN route.

Besides, ASBR1 will locally determine if it needs to send an RD-ORF entry to ASBR2. The judgment criteria refers to Section 5.1.2.

3) If ASBR2/PE3 receives the RD-ORF entry, it will repeat the above process.

When the RD-ORF mechanism is triggered, network operators need to manually configure the network to return to resume normal operation. The withdraw of RD-ORF entries refers to Section 5.2.

In Option AB cross-domain scenario, ASBRs maintain a VRF for a VPN. However, due to VPN routes in all VRFs use the same BGP session, ASBRs cannot prevent the overflow of a certain VRF by breaking down a BGP session. The operation process of RD-ORF is similar to that in Option B scenario.

6.2. Application in Option C cross-domain scenario

The Option C cross-domain scenario is shown in Figure 5:

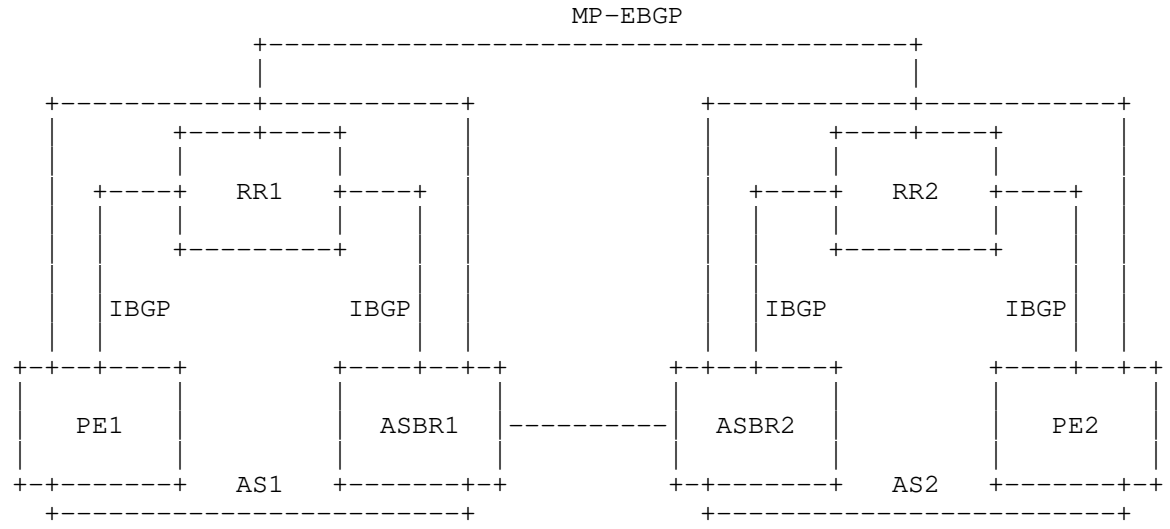


Figure 5: The Option C cross-domain scenario

In this scenario, PE1 and PE2 are responsible for maintaining VPN routing information in AS1 and AS2. In order to reduce the complexity that full-mesh brings to the network, RR1 and RR2

establish MP-EBGP session to transmit labeled routes. In AS1, PE1 and ASBR1 establish IBGP session with RR1 to ensure the routes can be transmitted in AS1. In AS2, PE2 and ASBR2 establish IBGP session with RR2.

Due to the maintenance of VPN routes is only done by PEs. RRs cannot know whether the PEs' ability to handle VPN routes has reached the upper limit or not, so it needs the RD-ORF to control the number of routes.

The operating mechanism of RD-ORF is similar to the description in Section 6.1.

7. Security Considerations

A BGP speaker will maintain the RD-ORF entries in Adj-RIB-out, this behavior consumes its memory and compute resources. To avoid the excessive consumption of resources, [RFC5291] specifies that a BGP speaker can only accept ORF entries transmitted by its interested peers.

8. IANA Considerations

This document defines a new Outbound Route Filter type - Route Distinguisher Outbound Route Filter (RD-ORF). The code point is from the "BGP Outbound Route Filtering (ORF) Types". It is recommended to set the code point of RD-ORF to 66.

9. Acknowledgement

Thanks Robert Raszuk, Jim Uttaro, Jakob Heitz, Jeff Tantsura, Rajiv Asati, John E Drake and Gert Doering for their valuable comments on this draft.

10. Normative References

[I-D.ietf-bess-evpn-inter-subnet-forwarding]
Sajassi, A., Salam, S., Thoria, S., Drake, J., and J. Rabadan, "Integrated Routing and Bridging in EVPN", draft-ietf-bess-evpn-inter-subnet-forwarding-11 (work in progress), October 2020.

[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>.

[RFC4360] Sangli, S., Tappan, D., and Y. Rekhter, "BGP Extended Communities Attribute", RFC 4360, DOI 10.17487/RFC4360, February 2006, <https://www.rfc-editor.org/info/rfc4360>.

[RFC4364] Rosen, E. and Y. Rekhter, "BGP/MPLS IP Virtual Private Networks (VPNs)", RFC 4364, DOI 10.17487/RFC4364, February 2006, <https://www.rfc-editor.org/info/rfc4364>.

[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>.

[RFC5291] Chen, E. and Y. Rekhter, "Outbound Route Filtering Capability for BGP-4", RFC 5291, DOI 10.17487/RFC5291, August 2008, <https://www.rfc-editor.org/info/rfc5291>.

[RFC5292] Chen, E. and S. Sangli, "Address-Prefix-Based Outbound Route Filter for BGP-4", RFC 5292, DOI 10.17487/RFC5292, August 2008, <https://www.rfc-editor.org/info/rfc5292>.

[RFC7432] Sajassi, A., Ed., Aggarwal, R., Bitar, N., Isaac, A., Uttaro, J., Drake, J., and W. Henderickx, "BGP MPLS-Based Ethernet VPN", RFC 7432, DOI 10.17487/RFC7432, February 2015, <https://www.rfc-editor.org/info/rfc7432>.

Authors' Addresses

Wei Wang
China Telecom
Beiqijia Town, Changping District
Beijing, Beijing 102209
China

Email: wangw36@chinatelecom.cn

Aijun Wang
China Telecom
Beiqijia Town, Changping District
Beijing, Beijing 102209
China

Email: wangaj3@chinatelecom.cn

Haibo Wang
Huawei Technologies
Huawei Building, No.156 Beiqing Rd.
Beijing, Beijing 100095
China

Email: rainsword.wang@huawei.com

Gyan S. Mishra
Verizon Inc.
13101 Columbia Pike
Silver Spring MD 20904
United States of America

Phone: 301 502-1347
Email: gyan.s.mishra@verizon.com

Shunwan Zhuang
Huawei Technologies
Huawei Building, No.156 Beiqing Rd.
Beijing, Beijing 100095
China

Email: zhuangshunwan@huawei.com

Jie Dong
Huawei Technologies
Huawei Building, No.156 Beiqing Rd.
Beijing, Beijing 100095
China

Email: jie.dong@huawei.com

IDR Working Group
Internet-Draft
Intended status: Standards Track
Expires: July 30, 2021

C. Xie
C. Li
China Telecom
J. Dong
Z. Li
Huawei Technologies

January 26, 2021

BGP-LS with Multi-topology for Segment Routing based Virtual Transport Networks
draft-xie-idr-bgpls-sr-vtn-mt-02

Abstract

Enhanced VPN (VPN+) aims to provide enhanced VPN service to support some applications' needs of enhanced isolation and stringent performance requirements. VPN+ requires integration between the overlay VPN and the underlay network. A Virtual Transport Network (VTN) is a virtual underlay network which consists of a customized network topology and a set of network resource allocated from the physical network. A VTN could be used as the underlay to support one or a group of VPN+ services.

When Segment Routing is used as the data plane of VTNs, each VTN can be allocated with a group of SIDs to identify the topology and resource attributes of network segments in the VTN. The association between the network topology, the network resource attributes and the SR SIDs may need to be distributed to a centralized network controller. For network scenarios where each VTN can be identified by a unique topology ID, this document describes a mechanism to distribute the information of SR based VTNs using BGP-LS with Multi-Topology.

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 RFC 2119 [RFC2119].

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 July 30, 2021.

Table of Contents

1. Introduction

Enhanced VPN (VPN+) is an enhancement to VPN services to support the needs of new applications, particularly including the applications that are associated with 5G services. These applications require enhanced isolation and have more stringent performance requirements than that can be provided with traditional overlay VPNs. Thus these

properties require integration between the overlay connectivity and the characteristics provided by the underlay networks. [I-D.ietf-teas-enhanced-vpn] specifies the framework of enhanced VPN and describes the candidate component technologies in different network planes and layers. An enhanced VPN can be used for 5G network slicing, and will also be of use in more generic scenarios.

To meet the requirement of enhanced VPN services, a number of Virtual Transport Networks (VTNs) need to be created, each consists of a subset of the underlay network topology and a set of network resources allocated from the underlay network to meet the requirement of one or a group of VPN+ services.

[I-D.ietf-spring-resource-aware-segments] introduces resource awareness to Segment Routing (SR) [RFC8402], by associating existing type of SIDs with network resource attributes (e.g. bandwidth, processing or storage resources). These resource-aware SIDs retain their original functionality, with the additional semantics of identifying the set of network resources available for the packet processing action.[I-D.dong-spring-sr-for-enhanced-vpn] describes the use of resource-aware segments to build SR based VTNs. To allow the network controller and network nodes to perform VTN-specific explicit path computation and/or shortest path computation, the group of resource-aware SIDs allocated by network nodes to each VTN and the associated topology and resource attributes need to be distributed in the control plane. When a centralized network controller is used for VTN-specific path computation, especially when a VTN spans multiple IGP areas or multiple Autonomous Systems (ASes), BGP-LS is needed to advertise the VTN information in each IGP area or AS to the network controller, so that the controller could use the collected information to build the view of inter-area or inter-AS SR VTNs.

In some network scenarios, each VTN can be identified by a unique topology ID [RFC5120], [I-D.xie-lsr-isis-sr-vtn-mt] describes an IGP mechanism to advertise the association between the topology, resource attributes and the SR SIDs for each VTN. This document describes a mechanism to distribute the information of SR based VTNs to the network controller using BGP-LS with Multi-Topology.

2. Advertisement of SR VTN Topology Attribute

[I-D.xie-lsr-isis-sr-vtn-mt] describes the IS-IS Multi-topology based mechanisms to distribute the topology attributes of SR based VTNs. This section describes the corresponding BGP-LS mechanism to distribute both the intra-domain and inter-domain topology attributes of SR based VTNs.

2.1. Intra-domain Topology Advertisement

In section 4.2.2.1 of [I-D.ietf-idr-rfc7752bis], Multi-Topology Identifier (MT-ID) TLV is defined, which can contain one or more IS-IS or OSPF Multi-Topology IDs. The MT-ID TLV MAY be present in a Link Descriptor, a Prefix Descriptor, or the BGP-LS Attribute of a Node NLRI.

[I-D.ietf-idr-bgp-ls-segment-routing-ext] defines the BGP-LS extensions to carry the segment routing information using TLVs of BGP-LS Attribute. When MTR is used with SR-MPLS data plane, topology-specific prefix-SIDs and topology-specific Adj-SIDs can be carried in the BGP-LS Attribute associated with the prefix NLRI and link NLRI respectively, the MT-ID TLV is carried in the prefix descriptor or link descriptor to identify the corresponding topology of the SIDs.

[I-D.ietf-idr-bgpls-srv6-ext] defines the BGP-LS extensions to advertise SRv6 segments along with their functions and attributes. When MTR is used with SRv6 data plane, the SRv6 Locator TLV is carried in the BGP-LS Attribute associated with the prefix-NLRI, the MT-ID TLV can be carried in the prefix descriptor to identify the corresponding topology of the SRv6 Locator. The SRv6 End.X SIDs are carried in the BGP-LS Attribute associated with the link NLRI, the MT-ID TLV can be carried in the link descriptor to identify the corresponding topology of the End.X SIDs. The SRv6 SID NLRI is defined to advertise other types of SRv6 SIDs, in which the SRv6 SID Descriptors can include the MT-ID TLV so as to advertise topology-specific SRv6 SIDs.

[I-D.ietf-idr-rfc7752bis] also defines the rules of the usage of MT-ID TLV:

"In a Link or Prefix Descriptor, only a single MT-ID TLV containing the MT-ID of the topology where the link or the prefix is reachable is allowed. In case one wants to advertise multiple topologies for a given Link Descriptor or Prefix Descriptor, multiple NLRIs MUST be generated where each NLRI contains a single unique MT-ID."

Editor’s note: the above rules indicates that only one MT-ID is allowed to be carried the Link or Prefix descriptors. When a link or prefix needs to be advertised in multiple topologies, multiple NLRIs needs to be generated to report all the topologies the link or prefix participates in, together with the topology-specific segment routing information and link attributes. This may increase the number of BGP Updates needed for advertising MT-specific topology attributes, and may introduce additional processing burden to both the sending BGP speaker and the receiving network controller. When the number of

topologies in a network is not a small number, some optimization may be needed for the reporting of multi-topology information and the associated segment routing information in BGP-LS. Based on the WG's opinion, this will be elaborated in a future version.

2.2. Inter-Domain Topology Advertisement

[I-D.ietf-idr-bgpls-segment-routing-epe] and [I-D.ietf-idr-bgpls-srv6-ext] defines the BGP-LS extensions for advertisement of BGP topology information between ASes and the BGP Peering Segment Identifiers. Such information could be used by a network controller for the computation and instantiation of inter-AS traffic engineering SR paths.

In some network scenarios, there are needs to create VTNs which span multiple ASes. The inter-domain VTNs could have different inter-domain connectivity, and may be associated with different set of network resources in each domain and also on the inter-domain links. In order to build the multi-domain SR based VTNs, it is necessary to advertise the topology and resource attribute of each VTN and the associated BGP Peering SIDs on the inter-domain links.

Depending on the requirement of inter-domain VTNs, different mechanism can be used on the inter-domain connection:

- One EBGP session between two ASes can be established over multiple underlying links. In this case, different underlying links can be used for different inter-domain VTNs which requires link isolation between each other. In another similar case, the EBGP session is established over a single link, while the network resource (e.g. bandwidth) on this link can be partitioned into several pieces, each of which can be considered as a virtual member link. A VTN is associated with one of the physical or virtual member links. In both cases, different BGP Peer-Adj-SIDs or SRv6 End.X SID SHOULD be allocated to each underlying physical or virtual member link, the association between the BGP Peer Adj-SID/End.X SID and the identifier of the VTN SHOULD be advertised by the ASBR.

- For inter-domain connection between two ASes, multiple EBGP sessions can be established between different set of peering ASBRs. It is possible that some of these BGP sessions are used for one multi-domain VTN, while some other BGP sessions are used for another multi-domain VTN. In this case, different BGP Peer Node SIDs are allocated to each BGP session and are advertised using the mechanism in [I-D.ietf-idr-bgpls-segment-routing-epe] and [I-D.ietf-idr-bgpls-srv6-ext], the association between the BGP Peer Node SIDs and the identifier of the VTN SHOULD be advertised by the ASBR.

- At the AS-level topology, different multi-domain VTNs may have different inter-domain connectivity. Different BGP Peer Set SIDs MAY be allocated to represent the groups of BGP peers which can be used for load-balancing in each multi-domain VTN.

When MT-ID is used consistently in multiple ASes covered by a VTN, the topology-specific BGP peering SIDs can be advertised with the MT-ID carried in the corresponding Link NLRI. This can be achieved with the existing mechanisms as defined in [RFC7752][I-D.ietf-idr-bgpls-segment-routing-epe] and [I-D.ietf-idr-bgpls-srv6-ext].

In network scenarios where consistent usage of MT-ID among multiple domains can not be expected, a global-significant VTN-ID needs to be introduced to define the inter-domain topologies. Within each domain, the MT based mechanism could be reused for intra-domain topology advertisement. The detailed mechanism is specified in [I-D.dong-idr-bgpls-sr-enhanced-vpn].

3. Advertisement of SR VTN Resource Attribute

[I-D.xie-lsr-isis-sr-vtn-mt] specifies the mechanism to advertise the resource information associated with each VTN. This section describes the corresponding BGP-LS mechanisms.

The information of the network resources associated with a VTN can be specified by carrying the TE Link attribute TLVs in BGP-LS Attribute [RFC7752], with the associated MT-ID carried in the corresponding Link NLRI.

When Maximum Link Bandwidth sub-TLV is carried in the BGP-LS attribute associated with the Link NLRI of a VTN, it indicates the amount of link bandwidth resource allocated to the corresponding VTN on the link. The bandwidth allocated to a VTN can be exclusive for traffic in the corresponding VTN. The advertisement of other TE attributes in BGP-LS for each VTN is for further study.

4. Scalability Considerations

The mechanism described in this document requires that each VTN mapped to an independent topology, and for the inter-domain VTNs, the MT-IDs used in each involved domain need to be consistent. Reusing MT-IDs as the identifier of VTN can avoid introducing new identifiers in the control plane, while it also has some limitations. For example, when multiple VTNs shares the same topology, each VTN still need to be identified using different MT-IDs in the control plane, thus independent path computation needs be executed for each VTN. The number of VTNs supported in a network may be dependent on the

number of topologies supported, which is related to the control plane overhead.

5. Security Considerations

This document introduces no additional security vulnerabilities to BGP-LS.

The mechanism proposed in this document is subject to the same vulnerabilities as any other protocol that relies on BGP-LS.

6. IANA Considerations

This document does not request any IANA actions.

7. Acknowledgments

The authors would like to thank Shunwan Zhuang for the review and discussion of this document.

8. References

8.1. Normative References

[I-D.dong-spring-sr-for-enhanced-vpn]
Dong, J., Bryant, S., Miyasaka, T., Zhu, Y., Qin, F., Li, Z., and F. Clad, "Segment Routing based Virtual Transport Network (VTN) for Enhanced VPN", draft-dong-spring-sr-for-enhanced-vpn-13 (work in progress), January 2021.

[I-D.ietf-idr-bgp-ls-segment-routing-ext]
Previdi, S., Talaulikar, K., Filsfils, C., Gredler, H., and M. Chen, "BGP Link-State extensions for Segment Routing", draft-ietf-idr-bgp-ls-segment-routing-ext-16 (work in progress), June 2019.

[I-D.ietf-idr-bgpls-segment-routing-epe]
Previdi, S., Talaulikar, K., Filsfils, C., Patel, K., Ray, S., and J. Dong, "BGP-LS extensions for Segment Routing BGP Egress Peer Engineering", draft-ietf-idr-bgpls-segment-routing-epe-19 (work in progress), May 2019.

[I-D.ietf-idr-bgpls-srv6-ext]
Dawra, G., Filsfils, C., Talaulikar, K., Chen, M., daniel.bernier@bell.ca, d., and B. Decraene, "BGP Link State Extensions for SRv6", draft-ietf-idr-bgpls-srv6-ext-05 (work in progress), November 2020.

[I-D.ietf-idr-rfc7752bis]
Talaulikar, K., "Distribution of Link-State and Traffic Engineering Information Using BGP", draft-ietf-idr-rfc7752bis-05 (work in progress), November 2020.

[I-D.ietf-spring-resource-aware-segments]
Dong, J., Bryant, S., Miyasaka, T., Zhu, Y., Qin, F., Li, Z., and F. Clad, "Introducing Resource Awareness to SR Segments", draft-ietf-spring-resource-aware-segments-01 (work in progress), January 2021.

[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>.

[RFC5029] Vasseur, JP. and S. Previdi, "Definition of an IS-IS Link Attribute Sub-TLV", RFC 5029, DOI 10.17487/RFC5029, September 2007, <https://www.rfc-editor.org/info/rfc5029>.

[RFC7752] Gredler, H., Ed., Medved, J., Previdi, S., Farrel, A., and S. Ray, "North-Bound Distribution of Link-State and Traffic Engineering (TE) Information Using BGP", RFC 7752, DOI 10.17487/RFC7752, March 2016, <https://www.rfc-editor.org/info/rfc7752>.

[RFC8402] Filsfils, C., Ed., Previdi, S., Ed., Ginsberg, L., Decraene, B., Litkowski, S., and R. Shakir, "Segment Routing Architecture", RFC 8402, DOI 10.17487/RFC8402, July 2018, <https://www.rfc-editor.org/info/rfc8402>.

8.2. Informative References

[I-D.dong-idr-bgpls-sr-enhanced-vpn]
Dong, J., Hu, Z., Li, Z., Tang, X., and R. Pang, "BGP-LS Extensions for Segment Routing based Enhanced VPN", draft-dong-idr-bgpls-sr-enhanced-vpn-02 (work in progress), June 2020.

[I-D.dong-lsr-sr-enhanced-vpn]
Dong, J., Hu, Z., Li, Z., Tang, X., Pang, R., JooHeon, L., and S. Bryant, "IGP Extensions for Segment Routing based Enhanced VPN", draft-dong-lsr-sr-enhanced-vpn-04 (work in progress), June 2020.

[I-D.ietf-lsr-isis-srv6-extensions]
Psenak, P., Filsfils, C., Bashandy, A., Decraene, B., and Z. Hu, "IS-IS Extension to Support Segment Routing over IPv6 Dataplane", draft-ietf-lsr-isis-srv6-extensions-11 (work in progress), October 2020.

[I-D.ietf-teas-enhanced-vpn]
Dong, J., Bryant, S., Li, Z., Miyasaka, T., and Y. Lee, "A Framework for Enhanced Virtual Private Networks (VPN+) Service", draft-ietf-teas-enhanced-vpn-06 (work in progress), July 2020.

[I-D.xie-lsr-isis-sr-vtn-mt]
Xie, C., Ma, C., Dong, J., and Z. Li, "Using IS-IS Multi-Topology (MT) for Segment Routing based Virtual Transport Network", draft-xie-lsr-isis-sr-vtn-mt-02 (work in progress), October 2020.

[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>.

[RFC8667] Previdi, S., Ed., Ginsberg, L., Ed., Filsfils, C., Bashandy, A., Gredler, H., and B. Decraene, "IS-IS Extensions for Segment Routing", RFC 8667, DOI 10.17487/RFC8667, December 2019, <https://www.rfc-editor.org/info/rfc8667>.

Authors' Addresses

Chongfeng Xie
China Telecom
China Telecom Beijing Information Science & Technology, Beiqijia
Beijing 102209
China

Email: xiechf@chinatelecom.cn

Cong Li
China Telecom
China Telecom Beijing Information Science & Technology, Beiqijia
Beijing 102209
China

Email: licong@chinatelecom.cn

Jie Dong
Huawei Technologies
Huawei Campus, No. 156 Beiqing Road
Beijing 100095
China

Email: jie.dong@huawei.com

Zhenbin Li
Huawei Technologies
Huawei Campus, No. 156 Beiqing Road
Beijing 100095
China

Email: lizhenbin@huawei.com

Network Working Group
Internet-Draft
Intended status: Standards Track
Expires: May 20, 2021

Y. Zhu
China Telecom
Z. Hu
S. Peng
Huawei Technologies
R. Mwehaire
MTN Uganda Ltd.

November 16, 2020

Signaling Maximum Transmission Unit (MTU) using BGP-LS
draft-zhu-idr-bgp-ls-path-mtu-05

Abstract

BGP Link State (BGP-LS) describes a mechanism by which link-state and TE information can be collected from networks and shared with external components using the BGP routing protocol. The centralized controller (PCE/SDN) completes the service path calculation based on the information transmitted by the BGP-LS and delivers the result to the Path Computation Client (PCC) through the PCEP or BGP protocol.

Segment Routing (SR) leverages the source routing paradigm, which can be directly applied to the MPLS architecture with no change on the forwarding plane and applied to the IPv6 architecture, with a new type of routing header, called SRH. The SR uses the IGP protocol as the control protocol. Compared to the MPLS tunneling technology, the SR does not require additional signaling. Therefore, the SR does not support the negotiation of the Path MTU. Since multiple labels or SRv6 SIDs are pushed in the packets, it is more likely that the packet size exceeds the path mtu of SR tunnel.

This document specifies the extensions to BGP Link State (BGP-LS) to carry maximum transmission unit (MTU) messages of link. The PCE/SDN calculates the Path MTU while completing the service path calculation based on the information transmitted by the BGP-LS.

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 RFC 2119 [RFC2119].

Status of This Memo

This Internet-Draft is submitted in full conformance with the provisions of BCP 78 and BCP 79.

This Internet-Draft will expire on May 20, 2021.

Table of Contents

1. Introduction

[RFC7752]describes the implementation mechanism of BGP-LS by which link-state and TE information can be collected from networks and shared with external components using the BGP routing protocol [RFC4271]. BGP-LS allows the necessary Link-State Database (LSDB)

and Traffic Engineering Database (TEDB) information to be collected from the IGP within the network, filtered according to configurable policy, and distributed to the PCE as necessary.

The appropriate MTU size guarantees efficient data transmission. If the MTU size is too small and the packet size is large, fragmentation may occur too much and packets are discarded by the QoS queue. If the MTU configuration is too large, packet transmission may be slow. Path MTU is the maximum length of a packet that can pass through a path without fragmentation. [RFC1191] describes a technique for dynamically discovering the maximum transmission unit (MTU) of an arbitrary internet path.

The traditional MPLS tunneling technology has signaling for establishing a path. [RFC3988] defines the mechanism for automatically discovering the Path MTU of LSPs. For a certain FEC, the LSR compares the MTU advertised by all downstream devices with the MTU of the FEC output interface in the local device, and calculates the minimum value for the upstream device.

[RFC3209] specify the mechanism of MTU signaling in RSVP-TE. The ingress node of the RSVP-TE tunnel sends a Path message to the downstream device. The Adspec object in the Path message carries the MTU. Each node along the tunnel receives a Path message, compares the MTU value in the Adspec object with the interface MTU value and MPLS MTU configured on the physical output interface of the local tunnel, obtains the minimum MTU value, and puts it into the newly constructed Path message and continues to send it to the downstream equipment. Thus, the MTU carried in the Path message received by the Egress node is the minimum value of the path MTU. The Egress node brings the negotiated Path MTU back to the Ingress node through the Resv message.

Segment Routing (SR) described in [RFC8402] leverages the source routing paradigm. Segment Routing can be directly applied to the MPLS architecture with no change on the forwarding plane [RFC8660] and applied to the IPv6 architecture with a new type of routing header called the SR header (SRH) [RFC8754]. [I-D.ietf-idr-bgp-ls-segment-routing-ext] defines SR extensions to BGP-LS and specifies the TLVs and sub-TLVs for advertising SR information. Based on the SR information reported by the BGP-LS, the SDN can calculate the end-to-end explicit SR-TE paths or SR Policies.

Nevertheless, Segment Routing is a tunneling technology based on the IGP protocol as the control protocol, and there is no additional signaling for establishing the path. so the Segment Routing tunnel cannot currently support the negotiation mechanism of the MTU. Multiple labels or SRv6 SIDs are pushed in the packets. This causes

the length of the packets encapsulated in the Segment Routing tunnel to increase during packet forwarding. This is more likely to cause packet size exceed the traditional MPLS packet size.

This document specify the extension to BGP Link State (BGP-LS) to carry link maximum transmission unit (MTU) messages.

2. Terminology

This draft refers to the terms defined in [RFC8201], [RFC4821] and [RFC3988].

MTU: Maximum Transmission Unit, the size in bytes of the largest IP packet, including the IP header and payload, that can be transmitted on a link or path. Note that this could more properly be called the IP MTU, to be consistent with how other standards organizations use the acronym MTU.

Link MTU: The Maximum Transmission Unit, i.e., maximum IP packet size in bytes, that can be conveyed in one piece over a link. Be aware that this definition is different from the definition used by other standards organizations.

For IETF documents, link MTU is uniformly defined as the IP MTU over the link. This includes the IP header, but excludes link layer headers and other framing that is not part of IP or the IP payload.

Be aware that other standards organizations generally define link MTU to include the link layer headers.

For the MPLS data plane, this size includes the IP header and data (or other payload) and the label stack but does not include any lower-layer headers. A link may be an interface (such as Ethernet or Packet-over-SONET), a tunnel (such as GRE or IPsec), or an LSP.

Path: The set of links traversed by a packet between a source node and a destination node.

Path MTU, or PMTU: The minimum link MTU of all the links in a path between a source node and a destination node.

3. Deploying scenarios

This document suggests a solution to extension to BGP Link State (BGP-LS) to carry maximum transmission unit (MTU) messages. The MTU information of the link is acquired through the process of collecting link state and TE information by BGP-LS. Concretely, a router maintains one or more databases for storing link-state information about nodes and links in any given area. The router's BGP process can retrieve topology from these IGP, BGP and other sources, and distribute it to a consumer, either directly or via a peer BGP speaker (typically a dedicated Route Reflector). [RFC7176] specifies a possible way of using the ISIS mechanism and extensions for link MTU Sub-TLV. In the case of inter-AS scenario (e.g., BGP EPE), the link MTU of the inter-AS link can be collected via BGP-LS directly.

As per [RFC7752], the collection of link-state and TE information and its distribution to consumers is shown in the following figure.

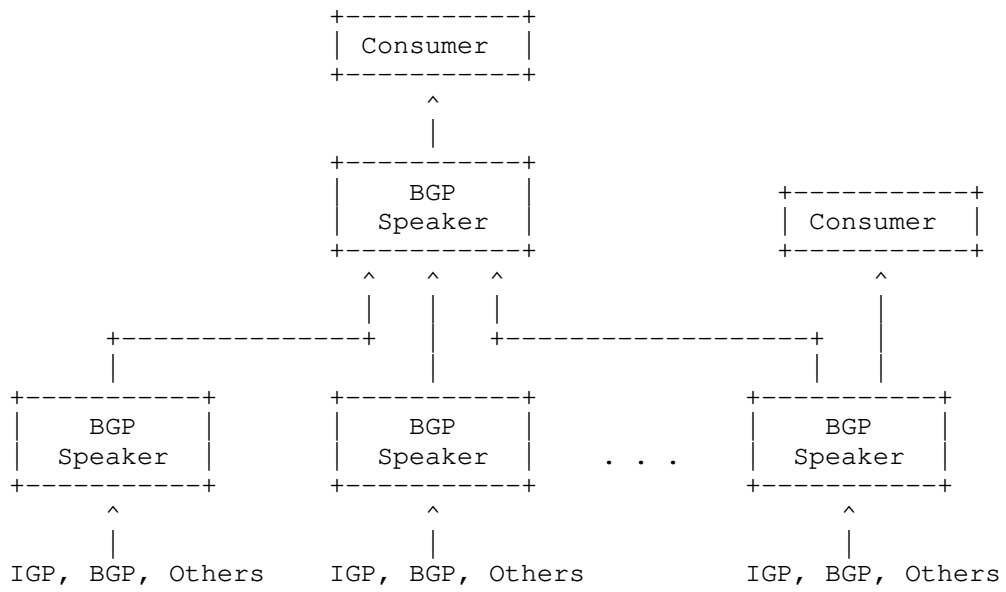


Figure 1: Collection of Link-State and TE Information

Please note that this signaled MTU may be different from the actual MTU, which is usually from configuration mismatches in a control plane and a data plane component.

4. BGP_LS Extensions for Link MTU

[RFC7752] defines the BGP-LS NLRI that can be a Node NLRI, a Link NLRI or a Prefix NLRI. The corresponding BGP-LS attribute is a Node Attribute, a Link Attribute or a Prefix Attribute. [RFC7752] defines the TLVs that map link-state information to BGP-LS NLRI and the BGP-LS attribute. Therefore, according to this document, a new sub-TLV is added to the Link Attribute TLV. It is an independent attribute TLV that can be used for the link NLRI advertised with all the Protocol IDs.

The format of the sub-TLV is as shown below.

- TYPE - TBD
- LENGTH - Total length of the value field, it should be 2
- VALUE - 2-byte MTU value of the link

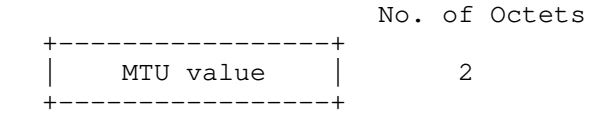


Figure 2. Sub-TLV Format for Link MTU

Whenever there is a change in MTU value represented by Link Attribute TLV, BGP-LS should re-originate the respective TLV with the new MTU value.

5. IANA Considerations

This document requests assigning a new code-point from the BGP-LS Link Descriptor and Attribute TLVs registry as specified in section 4.

Value	Description	Reference
TBD	Link MTU	This document

6. Security Considerations

This document does not introduce security issues beyond those discussed in RFC7752.

7. Acknowledgements

8. Contributors

Gang Yan
Huawei
China

Email:yangang@huawei.com

Junda Yao
Huawei
China

Email:yaojunda@huawei.com

9. References

9.1. Normative References

[RFC2119] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, DOI 10.17487/RFC2119, March 1997, <https://www.rfc-editor.org/info/rfc2119>.

9.2. Informative References

[I-D.ietf-idr-bgp-ls-segment-routing-ext] Previdi, S., Talaulikar, K., Filsfils, C., Gredler, H., and M. Chen, "BGP Link-State extensions for Segment Routing", draft-ietf-idr-bgp-ls-segment-routing-ext-16 (work in progress), June 2019.

[RFC1191] Mogul, J. and S. Deering, "Path MTU discovery", RFC 1191, DOI 10.17487/RFC1191, November 1990, <https://www.rfc-editor.org/info/rfc1191>.

[RFC3209] Awduche, D., Berger, L., Gan, D., Li, T., Srinivasan, V., and G. Swallow, "RSVP-TE: Extensions to RSVP for LSP Tunnels", RFC 3209, DOI 10.17487/RFC3209, December 2001, <https://www.rfc-editor.org/info/rfc3209>.

[RFC3988] Black, B. and K. Kompella, "Maximum Transmission Unit Signalling Extensions for the Label Distribution Protocol", RFC 3988, DOI 10.17487/RFC3988, January 2005, <https://www.rfc-editor.org/info/rfc3988>.

[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>.

[RFC4821] Mathis, M. and J. Heffner, "Packetization Layer Path MTU Discovery", RFC 4821, DOI 10.17487/RFC4821, March 2007, <https://www.rfc-editor.org/info/rfc4821>.

[RFC7176] Eastlake 3rd, D., Senevirathne, T., Ghanwani, A., Dutt, D., and A. Banerjee, "Transparent Interconnection of Lots of Links (TRILL) Use of IS-IS", RFC 7176, DOI 10.17487/RFC7176, May 2014, <https://www.rfc-editor.org/info/rfc7176>.

[RFC7752] Gredler, H., Ed., Medved, J., Previdi, S., Farrel, A., and S. Ray, "North-Bound Distribution of Link-State and Traffic Engineering (TE) Information Using BGP", RFC 7752, DOI 10.17487/RFC7752, March 2016, <https://www.rfc-editor.org/info/rfc7752>.

[RFC8201] McCann, J., Deering, S., Mogul, J., and R. Hinden, Ed., "Path MTU Discovery for IP version 6", STD 87, RFC 8201, DOI 10.17487/RFC8201, July 2017, <https://www.rfc-editor.org/info/rfc8201>.

[RFC8402] Filsfils, C., Ed., Previdi, S., Ed., Ginsberg, L., Decraene, B., Litkowski, S., and R. Shakir, "Segment Routing Architecture", RFC 8402, DOI 10.17487/RFC8402, July 2018, <https://www.rfc-editor.org/info/rfc8402>.

[RFC8660] Bashandy, A., Ed., Filsfils, C., Ed., Previdi, S., Decraene, B., Litkowski, S., and R. Shakir, "Segment Routing with the MPLS Data Plane", RFC 8660, DOI 10.17487/RFC8660, December 2019, <https://www.rfc-editor.org/info/rfc8660>.

[RFC8754] Filsfils, C., Ed., Dukes, D., Ed., Previdi, S., Leddy, J., Matsushima, S., and D. Voyer, "IPv6 Segment Routing Header (SRH)", RFC 8754, DOI 10.17487/RFC8754, March 2020, <https://www.rfc-editor.org/info/rfc8754>.

Authors' Addresses

Yongqing Zhu
China Telecom
109, West Zhongshan Road, Tianhe District.
Guangzhou 510000
China

Email: zhuyq8@chinatelecom.cn

Zhibo Hu
Huawei Technologies
Huawei Bld., No.156 Beiqing Rd.
Beijing 100095
China

Email: huzhibo@huawei.com

Shuping Peng
Huawei Technologies
Huawei Bld., No.156 Beiqing Rd.
Beijing 100095
China

Email: pengshuping@huawei.com

Robbins Mwehaire
MTN Uganda Ltd.
Uganda

Email: Robbins.Mwehair@mtn.com

idr
Internet-Draft
Updates: 4684 (if approved)
Intended status: Standards Track
Expires: July 15, 2021

Z. Zhang
J. Haas
Juniper Networks

January 11, 2021

Generic Route Constraint Distribution Mechanism for BGP
draft-zzhang-idr-bgp-rt-constrains-extension-01

Abstract

This document defines a mechanism based upon Constrained Route Distribution for BGP (RFC 4684) that works with various types of BGP Community-like Path Attributes. Similar to RFC 4684, this mechanism can be used to build a route distribution graph to limit the propagation of BGP Routes. Unlike RFC 4684, this mechanism is not restricted to BGP Extended Communities (RFC 4360).

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 July 15, 2021.

Table of Contents

1. Introduction

1.1. Constrained Route Distribution

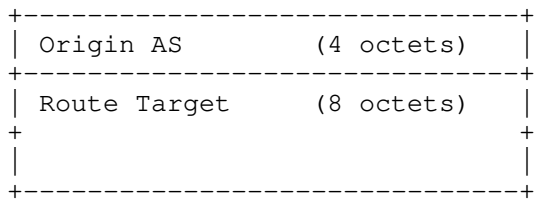
In BGP/MPLS Layer 3 VPNs [RFC4364], Route Target Extended Communities [RFC4360] are used to control VPN membership. Networks providing VPN services may be large. In such networks, VPN routes for a given VPN may be only needed at a small subset of Provider Edge (PE) routers.

The Constrained Route Distribution feature [RFC4684] assists in scaling such large VPN networks by building a distribution graph of VPN routes through the BGP routing infrastructure. Much of the benefit of this feature comes from BGP routers, such as Route Reflectors [RFC4456], avoiding the work of sending all VPN routes to a PE that may simply discard unneded routes. Instead, the PE may receive only the VPN routes for VPNs located on that PE.

1.2. Brief Summary of Constrained Route Distribution Procedure

BGP Speakers implementing [RFC4684] advertise their interest in receiving VPN routes that contain specific Route Target Extended Communities by advertising Route Target membership NLRI.

The format of the Route Target membership NLRI in [RFC4684] follows. It may be of length from 0 to 96 bits.



The Origin AS contains the Autonomous System number of the originator of this NLRI.

The Route Target contains a BGP Route Target Extended Community, or a prefix of a BGP Route Target Extended Community.

Route Target membership NLRI act as a filter mechanism on VPN routes. The BGP Speaker receiving these Route Target membership NLRI from another BGP Speaker will propagate VPN routes that match these membership NLRI. VPN routes that do not match these membership NLRI will not be propagated.

The propagation of Route Target membership NLRI from an originating PE router to other interested BGP Speakers builds a distribution graph for VPN routes matching the desired Route Targets.

1.3. Need for a Generic Route Constraint Distribution Mechanism

Since BGP/MPLS Layer 3 VPNs were introduced, many new BGP VPN features have been created that leverage the original concepts in [RFC4364]. While many of these new features similarly use Route Target Extended Communities for VPN membership, some use other Extended Communities. That is, they utilize a different Type/Sub-Type code than those defined in [RFC4360].

While [RFC4684] is explicit about being utilized for Route Targets, the definition of a Route Target has become more fluid as VPN features have been introduced; for example, ES-Import from [RFC7432]. It could be observed that that [RFC4684] is capable of being used on any type of [RFC4360] BGP Extended Community, for any VPN route type. However, other attributes are coming to be used for idenitifying VPN routes and a procedure that is only applicable to Extended Communities cannot be used.

[RFC5701] introduced the IPv6 Address Specific BGP Extended Community Attribute. This type of BGP Community permits the encoding of an IPv6 address as the Global Administrator of a route. Similar to the [RFC4360] Extended Communities, the IPv6 Address Specific type carries a Type and Sub-Type field. One of the Type/Sub-Type allocations is for an IPv6 address specific Route Target. This permits operators to leverage IPv6 addressing when building their VPNs.

IPv6 Extensions for Route Target Distribution [I-D.ietf-idr-bgp-ipv6-rt-constrain] proposes to permit matching for IPv6 address specific Extended Communities using [RFC4684] by overloading the NLRI length for Route Target membership NLRI for NLRI longer than 96 bits. (See [RFC4684], Section 4.) However, this doesn't account for Route Target membership NLRI length shorter than 96 bits. These shorter prefixes permit matching of many more specific Route Targets from a less specific Route Target membership BGP Route. Therefore, a different mechanism is needed for safely matching IPv6 address specific Route Targets.

The simplest change would be to utilize a new AFI/SAFI for IPv6 Route Target Distribution that only matches IPv6 address specific Route Targets. It can be further observed that various forms of BGP "Community" types continue to evolve to suit a variety of BGP route filtering needs, including those not intended for VPN services. Examples of these include BGP Large Communities [RFC8092], BGP Wide Communities [I-D.ietf-idr-wide-bgp-communities], and Bitmask Route Targets [I-D.zzhang-idr-bitmask-route-target].

This document proposes a mechanism to match arbitrary BGP Community-like attributes, including those with Route Target-like semantics, for building Constrained Route Distribution graphs for BGP routes containing those attributes.

2. Community-like Attributes

2.1. Definition of Community-like Attributes

BGP Communities were originally introduced in [RFC1997]. That RFC contains the definition, "A community is a group of destinations which share some common property." Recall that in BGP-4 [RFC4271], a BGP Route is defined as a pairing of destinations (NLRI) with Path Attributes.

In practice, a Community is implemented as an element of a BGP Path Attribute that is used to mark a prefix in a way that protocol and BGP policy mechanisms may be used to interact with that BGP Route.

Since [RFC1997], this idea of marking BGP Routes has been extended to other mechanisms such as BGP Extended Communities [RFC4360], and BGP Large Communities [RFC8092]. Other similar mechanisms are regularly considered for standardization.

For purposes of this document, a Community-like Attribute (CLA) has the semantics of being an attribute of a BGP Path Attribute that is intended to interact with protocol mechanisms and may enable policy mechanisms to interact with that BGP Route. Thus, classic [RFC1997] BGP Communities, BGP Extended Communities, and Large BGP Communities are all CLAs.

2.2. Prefix Structure of BGP Community-like Attributes

[RFC4684] provides for matching less-specific BGP Extended Communities by utilizing a shorter NLRI length for the Route Target membership NLRI. To highlight situations where such summarization is useful, consider the various forms of Route Target extended community from [RFC4360]. In each of those types, the Sub-Type field is 0x02, with the Type selecting the format:

- 0x00 - 2-octet Global Administrator field, 4-octet Local Administrator field.
- 0x01 - 4-octet Global Administrator field, 2-octet Local Administrator field.
- 0x02 - 4-octet IPv4 Address Global Administrator field, 2-octet Local Administrator field.

The Global Administrator field for Route Targets is typically an Autonomous System number.

Summarization offers several useful options where the Sub-Type of the Route Target Extended Community is 0x02. Examples include:

- Type = 0x00 and NLRI length = 48: Match all 2-octet Global Administrator fields of a given value; for example Origin AS 64511:Route Target 64496:*.
- Type = 0x01 and NLRI length = 64: Match all 4-octet Global Administrator fields of a given value; for example Origin AS 64511:Route Target 65551:*.
- Type = 0x03 and NLRI length = 88: Match all IPv4 Address Global Administrator fields of a given value; for example Origin AS 64511:Route Target 192.0.2.*:*.

Similarly, for inter-domain purposes, matching all Route Target Membership NLRI for a given Origin AS may be useful:

- NLRI length = 32; for example Origin AS 64511:*. This matches all classes of Extended Community originated from AS 64511.
- NLRI length = 44; for example Origin AS 64511:0x0002:*. This matches all Extended Communities originated from AS 64511 that have the first two octets as 0x0002, which includes the class of Extended Communities that are 2-octet Global Administrator Route Target types.

It's even possible to utilize a Prefix Length that splits a well defined field. When the structure of that field is understood, clever operators may be able to generate summaries. It should be noted that understanding the intent of such summarization may be difficult to discern from the NLRI in question. Some examples:

- NLRI length = 31; for example Origin AS 6451[01]:*. This matches all classes of Extended Community originated from Origin ASes 64510 and 64511.
- NLRI length = 47; for example Origin AS 64511:0x0002:*. This matches all two-octet AS-Specific Extended Communities originated from AS 64511 that include Route Targets (0x02) and Route Origins (0x03).

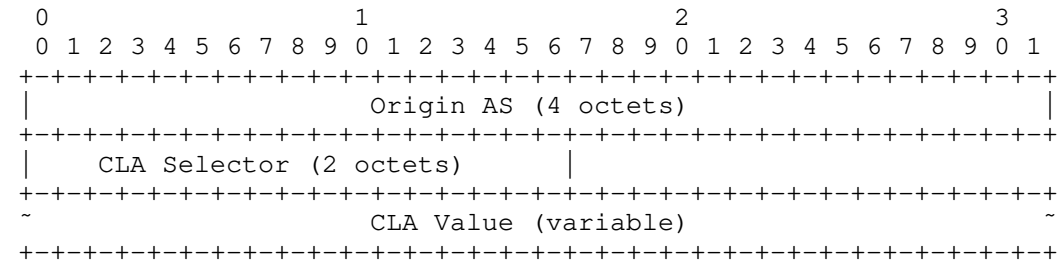
The purpose of highlighting that a variable NLRI length can be applied in these ways is to demonstrate the flexibility of summarization. This is most true when the structure of that

attribute is arranged most general to most specific; that is, Global to Local Admin as we have in Extended Communities.

3. Specification

3.1. NLRI Definition

To support applying Constrained Route Distribution procedures to BGP Community-like attributes, the following NLRI is defined. The "Generic Route Constraint Distribution Mechanism" NLRI uses a new SAFI (TBD) with the following format:



It can be observed that the format of this NLRI emulates the format of the Route Target membership NLRI from [RFC4684], with the addition of the CLA selector to permit the recipient to correctly interpret the CLA value.

3.2. NLRI Length Encoding

To support potentially large Community-like Values, the NLRI length field is encoded using 1 or 2 octets using the same mechanism as [RFC5575], Section 4. The text from that RFC is copied here:

> If the NLRI length value is smaller than 240 (0xf0 hex), the length field can be encoded as a single octet. Otherwise, it is encoded as an extended-length 2-octet value in which the most significant nibble of the first byte is all ones.
>
> In the figure above, values less-than 240 are encoded using two hex digits (0xnn). Values above 240 are encoded using 3 hex digits (0xfnnn). The highest value that can be represented with this encoding is 4095. The value 241 is encoded as 0xf0f1.

3.3. Operation

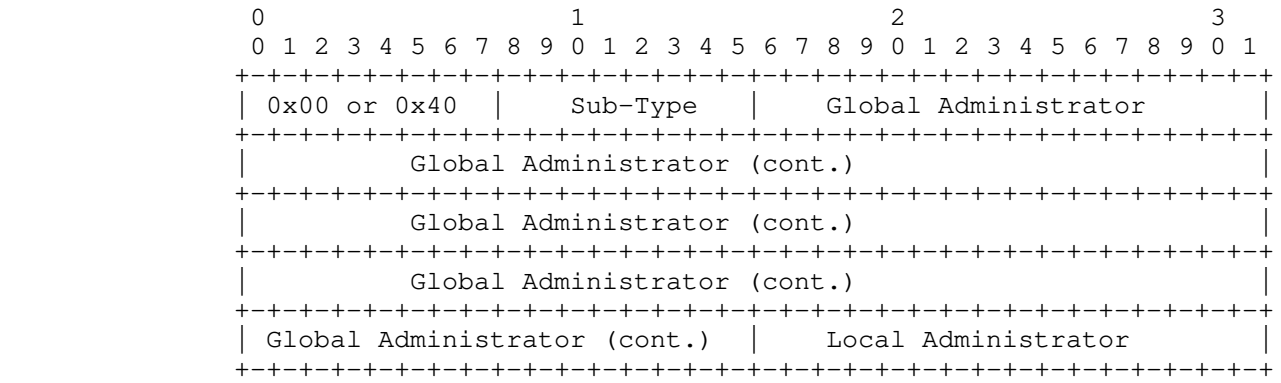
The two-octet CLA Selector identifies the type of Community-like attribute in a BGP route to apply the Constrained Route Distribution procedures to. The value of this field, registered with IANA, may identify Community-like attributes that exist in a given BGP Path Attribute, or internal fields of structured BGP Path Attributes. Examples of a stand-alone BGP Path Attribute may be [RFC1997] classic BGP Communities or [RFC8092] Large BGP Communities. Examples of internal community values may be Bitmask Route Targets [I-D.zzhang-idr-bitmask-route-target] defined inside a BGP Wide Community Container, or newly defined sub-TLVs in a BGP Tunnel Encapsulation Attribute [I-D.ietf-idr-tunnel-encaps].

The Community-like Attribute is encoded in the CLA Value field. Sufficient octets are encoded for the Prefix Length of this NLRI.

4. Examples

4.1. IPv6 Specific Extended Communities

[RFC5701] defines IPv6 Specific Extended Communities. Its structure, from the RFC is:



Where Global Administrator is 16 octets in length, and Local Administrator is 2 octets in length. The community is a fixed length of 20 octets.

The Community Selector for Large BGP Communities is assigned 1, per this document.

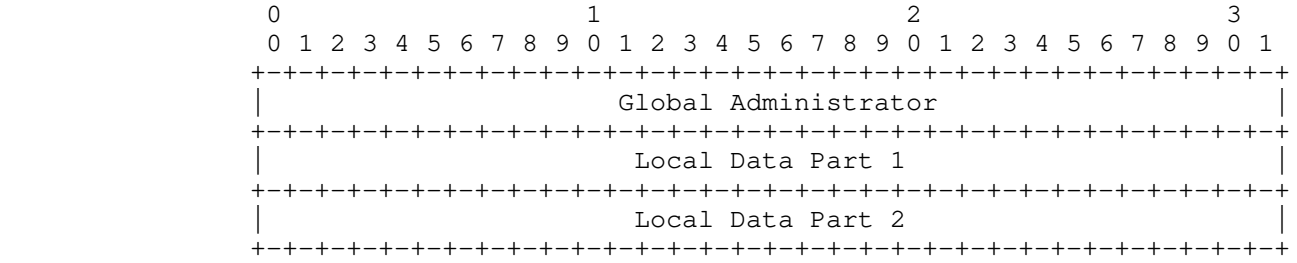
The encoding for a Generic Route Constraint Distribution Mechanism NLRI for an IPv6 Specific Extended Community for an Origin AS of

64511, for the IPv6 Specific Extended Community [2001:DB8::2]:100 would be:

```
NLRI length          = 0xd0 (208)
Origin AS            = 0x0000fbff (64511)
Community Selector   = 0x0001 (2)         # IPv6 Specific
                                          # Extended Community
Community-like Value = 0x0001000f (65551) # Global Administrator
                       0x2001 0DB8 0000 0000 0000 0000 0000 0000
                       0x0000 0000 0000 0000 0000 0000 0000 0002
                                          # Global Administrator
                       0x00000064 (100)   # Local Administrator
```

4.2. Large BGP Communities

[RFC8092] defines Large BGP Communities. Its structure, from the RFC is:



Where each of the fields Global Administrator, Local Data Part 1, and Local Data Part 2 are 4 octets in length. The community is a fixed length of 12 octets.

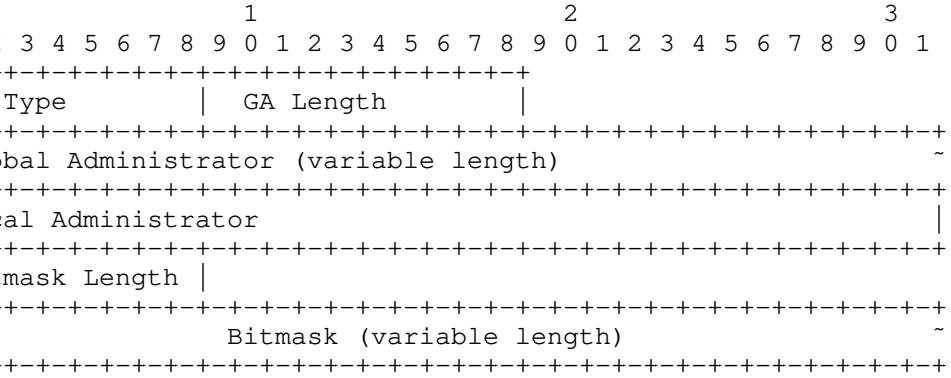
The Community Selector for Large BGP Communities is assigned 2, per this document.

The encoding for a Generic Route Cosntraint Mechanism NLRI for Large BGP Communities for an Origin AS of 64511, for Large BGP Community 65551:100:16777215 would be:

```
NLRI length          = 0x90 (144)
Origin AS            = 0x0000fbff (64511)
Community Selector   = 0x0001 (2)            # Large BGP Community
Community-like Value = 0x0001000f (65551)    # Global Administrator
                       0x00000064 (100)      # Local Data Part 1
                       0x00ffffff (16777215) # Local Data Part 2
```

4.3. Bitmask Route Target

[I-D.zzhang-idr-bitmask-route-target] defines Bitmask Route Targets. Bitmask Route Targets are encoded within the BGP Community Container Path Attribute, which is defined in [I-D.ietf-idr-wide-bgp-communities]. The structure of the Bitmask Route Target, from the Internet-Draft, is:



GA Type and GA Length are 1 octet in length.

Local Administrator is 4 octets in length.

The Bitmask is a number of octets that will fit the Bitmask Length.

The following GA Types and corresponding lengths are defined:

- 1: AS Number, 4 octets
- 2: IPv4 Address, 4 octets
- 3: IPv6 Address, 16 octets

The Community Selector for Bitmask Route Targets is assigned 3, per this document.

The Bitmask Route Target, a Community-like attribute, is carried as the payload (that is, the value portion) of another Path Attribute. The Generic Route Constraint Distribution Mechanism NLRI is not constructed to match any of the outer portions of the Community Container; rather it matches only the payload, that is, the Bitmask Route Target itself.

4.3.1. AS Number Bitmask Route Target

The encoding for a Generic Route Constraint Distribution Mechanism NLRI for Origin AS 64511 for an AS-Number based Bitmask Route Target for AS 65551 with Local Administrator value 100 and a bitmask of 0xc0ffee (3 octets) would be:

```
NLRI length          = 0xa0 (160)
Origin AS            = 0x0000fbff (64511)
Community Selector   = 0x0002 (3)          # Bitmask Route Target
Community-Like Value = 0x01 (1)            # GA Type AS Number
                       0x04 (4)            # GA Length
                       0x0001000f (65551)  # Global Administrator
                       0x00000064 (100)    # Local Administrator
                       0x03 (3)            # Bitmask Length
                       0xc0ffee            # Bitmask
```

4.3.2. IPv6 Address Bitmask Route Target

The encoding for a Generic Route Constraint Distribution Mechanism NLRI for Origin AS 64511 for an AS-Number based Bitmask Route Target for 2001:DB8::2 with Local Administrator value 100 and a bitmask of 0xc0ffee (3 octets) would be:

```
NLRI length          = 0xf108 (264)
Origin AS            = 0x0000fbff (64511)
Community Selector   = 0x0002 (2)          # Bitmask Route Target
Community-Like Value = 0x01 (1)            # GA Type IPv6 Address
                       0x04 (16)           # GA Length
                       0x2001 0DB8 0000 0000 0000 0000 0000 0000
                       0x0000 0000 0000 0000 0000 0000 0000 0002
                                           # Global Administrator
                       0x00000064 (100)    # Local Administrator
                       0x03 (3)            # Bitmask Length
                       0xc0ffee            # Bitmask
```

5. Security Considerations

This document does not change security aspects discussed in [RFC4684].

6. IANA Considerations

This document requests IANA to assign a new SAFI, the "Generic Route Constraint Distribution Mechanism" from the First Come First Served "Subsequent Address Family Identifiers (SAFI) Parameters" registry.

This documument requests IANA to create a new registry, the Generic Route Constraint CLA Selector Registry. It should have the following initial values and registration policies assigned:

Value	Description	Defining Specification for Community-like attribute (CLA)	Reference for this Value
0	RESERVED	-	This document
1	IPv6 Address Specific BGP Extended Communities	RFC 5701	This document
2	Large BGP Communities	RFC 8092	This document
3	Bitmask Route Targets	draft-zzhang-idr-bitmask-route-target	This document
4..64511	Available for first come, first served allocation.		
255	RESERVED	-	This document

7. Acknowledgements

The authors would like to thank John Scudder for his comments and suggestions.

8. References

8.1. 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>.

[RFC4360] Sangli, S., Tappan, D., and Y. Rekhter, "BGP Extended Communities Attribute", RFC 4360, DOI 10.17487/RFC4360, February 2006, <https://www.rfc-editor.org/info/rfc4360>.

[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>.

[RFC5575] Marques, P., Sheth, N., Raszuk, R., Greene, B., Mauch, J., and D. McPherson, "Dissemination of Flow Specification Rules", RFC 5575, DOI 10.17487/RFC5575, August 2009, <https://www.rfc-editor.org/info/rfc5575>.

[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>.

8.2. Informative References

[I-D.ietf-idr-bgp-ipv6-rt-constrain]
Patel, K., Raszuk, R., Djernaes, M., Dong, J., and M. Chen, "IPv6 Extensions for Route Target Distribution", draft-ietf-idr-bgp-ipv6-rt-constrain-12 (work in progress), April 2018.

[I-D.ietf-idr-tunnel-encaps]
Patel, K., Velde, G., Sangli, S., and J. Scudder, "The BGP Tunnel Encapsulation Attribute", draft-ietf-idr-tunnel-encaps-22 (work in progress), January 2021.

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

[I-D.zzhang-idr-bitmask-route-target]
Zhang, Z., Sangli, S., and J. Haas, "Bitmask Route Target", draft-zzhang-idr-bitmask-route-target-00 (work in progress), July 2020.

[RFC1997] Chandra, R., Traina, P., and T. Li, "BGP Communities Attribute", RFC 1997, DOI 10.17487/RFC1997, August 1996, <https://www.rfc-editor.org/info/rfc1997>.

[RFC4364] Rosen, E. and Y. Rekhter, "BGP/MPLS IP Virtual Private Networks (VPNs)", RFC 4364, DOI 10.17487/RFC4364, February 2006, <https://www.rfc-editor.org/info/rfc4364>.

[RFC4456] Bates, T., Chen, E., and R. Chandra, "BGP Route Reflection: An Alternative to Full Mesh Internal BGP (IBGP)", RFC 4456, DOI 10.17487/RFC4456, April 2006, <https://www.rfc-editor.org/info/rfc4456>.

[RFC5398] Huston, G., "Autonomous System (AS) Number Reservation for Documentation Use", RFC 5398, DOI 10.17487/RFC5398, December 2008, <https://www.rfc-editor.org/info/rfc5398>.

[RFC5701] Rekhter, Y., "IPv6 Address Specific BGP Extended Community Attribute", RFC 5701, DOI 10.17487/RFC5701, November 2009, <https://www.rfc-editor.org/info/rfc5701>.

[RFC7432] Sajassi, A., Ed., Aggarwal, R., Bitar, N., Isaac, A., Uttaro, J., Drake, J., and W. Henderickx, "BGP MPLS-Based Ethernet VPN", RFC 7432, DOI 10.17487/RFC7432, February 2015, <https://www.rfc-editor.org/info/rfc7432>.

[RFC8092] Heitz, J., Ed., Snijders, J., Ed., Patel, K., Bagdonas, I., and N. Hilliard, "BGP Large Communities Attribute", RFC 8092, DOI 10.17487/RFC8092, February 2017, <https://www.rfc-editor.org/info/rfc8092>.

[RFC8126] Cotton, M., Leiba, B., and T. Narten, "Guidelines for Writing an IANA Considerations Section in RFCs", BCP 26, RFC 8126, DOI 10.17487/RFC8126, June 2017, <https://www.rfc-editor.org/info/rfc8126>.

Authors' Addresses

Zhaohui Zhang
Juniper Networks

EMail: zzhang@juniper.net

Jeffrey Haas
Juniper Networks

EMail: jhaas@juniper.net

idr
Internet-Draft
Intended status: Standards Track
Expires: January 13, 2021

Z. Zhang
S. Sangli
J. Haas
Juniper Networks

July 12, 2020

Bitmask Route Target
draft-zzhang-idr-bitmask-route-target-00

Abstract

This document specifies a new type of Route Target called Bitmask Route Target as a BGP Community Container. The key element of a Bitmask Route Target is a Bitmask. Two Bitmask Route Targets are considered equivalent for the purpose of controlling route propagation (via Route Target Constraints) and importation if the result of logical "AND" operation of the Bitmask of the two is non-zero.

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 January 13, 2021.

Table of Contents

1. Introduction

The importation and propagation of BGP routes can be controled using Route Targets [RFC4364] and Route Target Constrains [RFC4684]. Both relies on comparing two Route Targets based on full match of the 8-octet encoding.

There are situations where it is desired to consider two Route Targets to be equivalent (hence the route could be imported or propagated) as long as certain bits have matching set values. This document defines a new type of Route Target for that purpose.

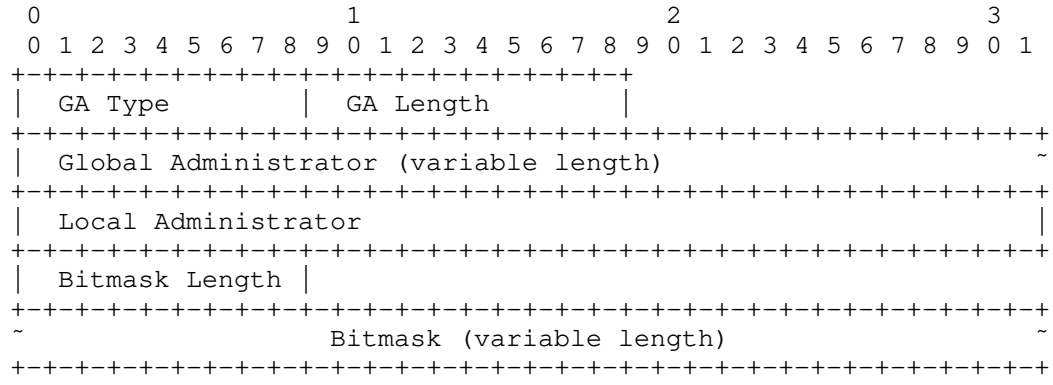
An example use case of this Bitmask Route Target is documented in [I-D.zzhang-teas-network-slicing-with-flex-te].

The use of Bitmask Route Target with Route Target Constrains is specified separately in [I-D.zzhang-idr-bgp-route-target-constrains-extension].

2. Specification

The Bitmask Route Target is a Transitive BGP Community Container of type TBD [I-D.ietf-idr-wide-bgp-communities].

The container includes a 1-octet Global Administrator (GA) Type, 1-octet GA Length, a variable length GA, a 4-octet Local Administrator (LA), a 1-octet Bitmask Length in number of octets, and the Bitmask.



The following GA Types and corresponding lengths are defined in this document:

- TBD1: AS Number, 4-octet
- TBD2: IPv4 Address, 4-octet
- TBD3: IPv6 Address, 16-octet

A Bitmask Route Targets A is considered to match Bitmask Route Target B for the purpose of controlling propagation and importation of a route with an attached Bitmask Route Target B if the following conditions are met:

- The GA Type, GA Length, GA, and LA fields in A and B match.
- The result of the logical "AND" operation of the Bitmask field in A and B is not 0. If A and B have different Bitmask Lengths, the smaller one is used to truncate the longer Bitmask.

3. Security Considerations

This document does not change security aspects as discussed in [RFC4364] and [I-D.ietf-idr-wide-bgp-communities].

4. IANA Considerations

This document requests IANA to assign a BGP Community Container Type for the Bitmask Route Target from the "BGP Community Container Types" registry.

This document requests IANA to setup a "Bitmask Route Target Global Administrator Type Registry" and assign three type values as listed in Section 2. Allocation from the first half of the number is based on standardization and allocation from the second half is First Come First Serve.

5. Acknowledgements

The authors thank John Scudder for his comments and suggestions.

6. References

6.1. Normative References

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

[RFC2119] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, DOI 10.17487/RFC2119, March 1997, <https://www.rfc-editor.org/info/rfc2119>.

[RFC8174] Leiba, B., "Ambiguity of Uppercase vs Lowercase in RFC 2119 Key Words", BCP 14, RFC 8174, DOI 10.17487/RFC8174, May 2017, <https://www.rfc-editor.org/info/rfc8174>.

6.2. Informative References

[RFC4364] Rosen, E. and Y. Rekhter, "BGP/MPLS IP Virtual Private Networks (VPNs)", RFC 4364, DOI 10.17487/RFC4364, February 2006, <https://www.rfc-editor.org/info/rfc4364>.

[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>.

Authors' Addresses

Zhaohui Zhang
Juniper Networks

EMail: zzhang@juniper.net

Srihari Sangli
Juniper Networks

EMail: ssangli@juniper.net

Jeffrey Haas
Juniper Networks

EMail: jhaas@juniper.net