Workgroup: Network Working Group
Internet-Draft:
draft-ietf-idr-sr-p2mp-policy-00
Published: 27 May 2022
Intended Status: Standards Track
Expires: 28 November 2022
Authors: H. Bidgoli, Ed. D. Voyer A. Stone
Nokia Bell Canada Nokia
R. Parekh S. Krier
Cisco System, Inc. Cisco System, Inc.
S. Agrawal
Cisco System, Inc.

Advertising p2mp policies in BGP

Abstract

SR P2MP policies are set of policies that enable architecture for P2MP service delivery.

A P2MP policy consists of candidate paths that connects the Root of the Tree to a set of Leaves. The P2MP policy is composed of replication segments. A replication segment is a forwarding instruction for a candidate path which is downloaded to the Root, transit nodes and the leaves.

This document specifies a new BGP SAFI with a new NLRI in order to advertise P2MP policy from a controller to a set of nodes.

This document introduces three new route types within this NLRI, one for P2MP policy and its candidate paths that need to be programmed on the Root node, one for the replication segment incoming SID which uniquely will identify the cross connect and another for each outgoing interface that the packets get replicated to. The last two route types are forwarding instructions that needs to be programmed on the Root, and optionally on Transit and Leaf nodes.

It should be noted that this document does not specify how the Root and the Leaves are discovered on the controller, it only describes how the P2MP Policy and Replication Segments are programmed from the controller to the nodes.

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 28 November 2022.

Copyright Notice

Table of Contents

1. Introduction

The draft [draft-ietf-pim-sr-p2mp-policy] defines a variant of the SR Policy [draft-ietf-spring-segment-routing-policy] for constructing a P2MP segment to support multicast service delivery.

A Point-to-Multipoint (P2MP) Policy contains a set of candidate paths and identifies a Root node and a set of Leaf nodes in a Segment Routing Domain. The draft also defines a Replication segment, which corresponds to the state of a P2MP segment on a particular node. The Replication segment is the forwarding instruction for a P2MP LSP at the Root, Transit and Leaf nodes.

For a P2MP segment, a controller may be used to compute a tree from a Root node to a set of Leaf nodes, optionally via a set of replication nodes. A packet is replicated at the root node and optionally on Replication nodes towards each Leaf node.

We define two types of a P2MP segment: Ingress Replication (aka Spray) and Downstream Replication (aka TreeSID).

A Point-to-Multipoint service delivery could be via Ingress Replication (aka Spray in some SR context), i.e., the root unicasts individual copies of traffic to each leaf. The corresponding P2MP segment consists of replication segments only for the root and the leaves.

A Point-to-Multipoint service delivery could also be via Downstream Replication (aka TreeSID in some SR context), i.e., the root and some downstream replication nodes replicate the traffic along the way as it traverses closer to the leaves.

It should be noted that two replication nodes can be connected directly, or they can be connected via unicast SR segment or a segment list.

The leaves and the root of a p2mp policy can be discovered via the multicast protocols or procedures like NG-MVPN [RFC6513] or manually configured on the PCC (CLI) or the PCE.

Based on the discovered root and leaves, the controller builds a P2MP policy and advertise it to the head-end router (i.e. the root of the P2MP Tree). The advertisement uses BGP extensions defined in this document. The controller also calculates the tree path and builds the replication segments on each segment of the tree, Root, Transit and Leaf nodes and downloads the forwarding instructions to the nodes via BGP extensions defined in this document.

SR p2mp policy is a variant of the SR policy and as such it reuses the concept of a candidate path. This draft reuses some of the concepts and TLVs mentioned in [draft-ietf-idr-segment-routing-te-policy]

A candidate path with in the P2MP policy can contain multiple path-instances. A path-instance can be viewed as a P2MP LSP. For candidate path global optimization purposes, two or more path-instances can be used to execute make before break procedures.

Each path-instance is a P2MP LSP as such each path-instance needs a set of replication segments to construct its forwarding instructions.

2. Conventions used in this document

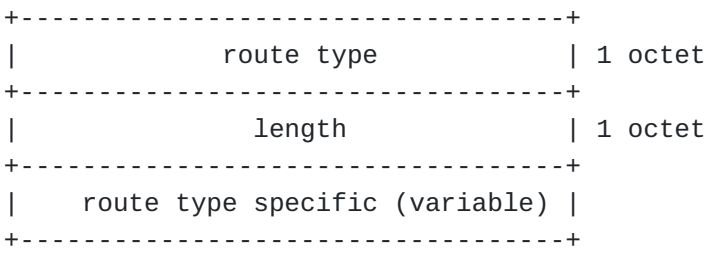
The key words "MUST", "MUST NOT", "REQUIRED", "SHALL", "SHALL NOT", "SHOULD", "SHOULD NOT", "RECOMMENDED", "MAY", and "OPTIONAL" in this document are to be interpreted as described in [RFC2119]

3. P2MP Policy and Replication Segment Encoding

3.1. P2MP Policy SAFI and NLRI

This document defines a new BGP NLRI, called the P2MP-POLICY NLRI.

A new SAFI is defined: the SR P2MP Policy SAFI, (Codepoint tbd assigned by IANA). The following is the format of the P2MP-POLICY NLRI:



*The Route type field defines the encoding of the rest of the P2MP- POLICY NLRI.

*The length field indicates the length in octets of the route type specific data, excluding route type and length

*This document defines the following route types:

-P2MP Policy route: TBD1, this is the actually P2MP policy on the root which contains the candidate paths, its prefrence and path instances.

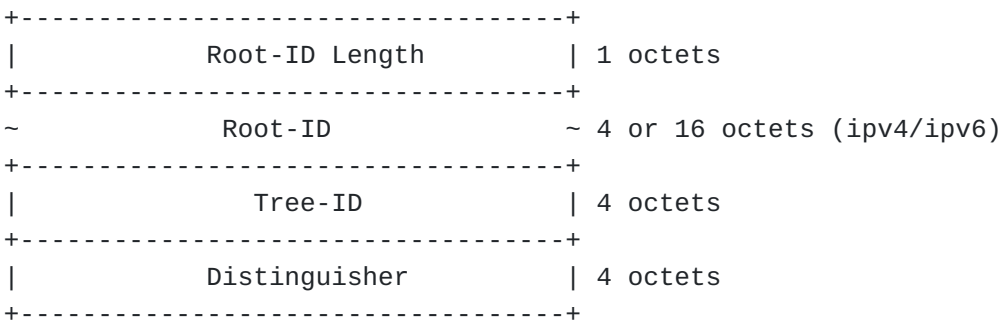
-Replication Segment Binding SID: TBD2, this is part of the replication segment and it is used for programming the incoming SID used to identify a P2MP cross connect.

-Replication Segment OIF: TBD3, this is a single Outgoing Interface for the P2MP cross connect. It also contains the outgoing SID.

The NLRI containing the SR P2MP Policy is carried in a BGP UPDATE message [RFC4271] using BGP multiprotocol extensions [RFC4760] with an AFI of 1 or 2 (IPv4 or IPv6) and with a SAFI of "TBD" (assigned by IANA from the "Subsequent Address Family Identifiers (SAFI) Parameters" registry).

All other recommendations of [draft-ietf-idr-segment-routing-te-policy] section SR Policy SAFI and NLRI, should be taken into account for P2MP policy.

3.1.1. P2MP Policy Route - Route Type TBD1



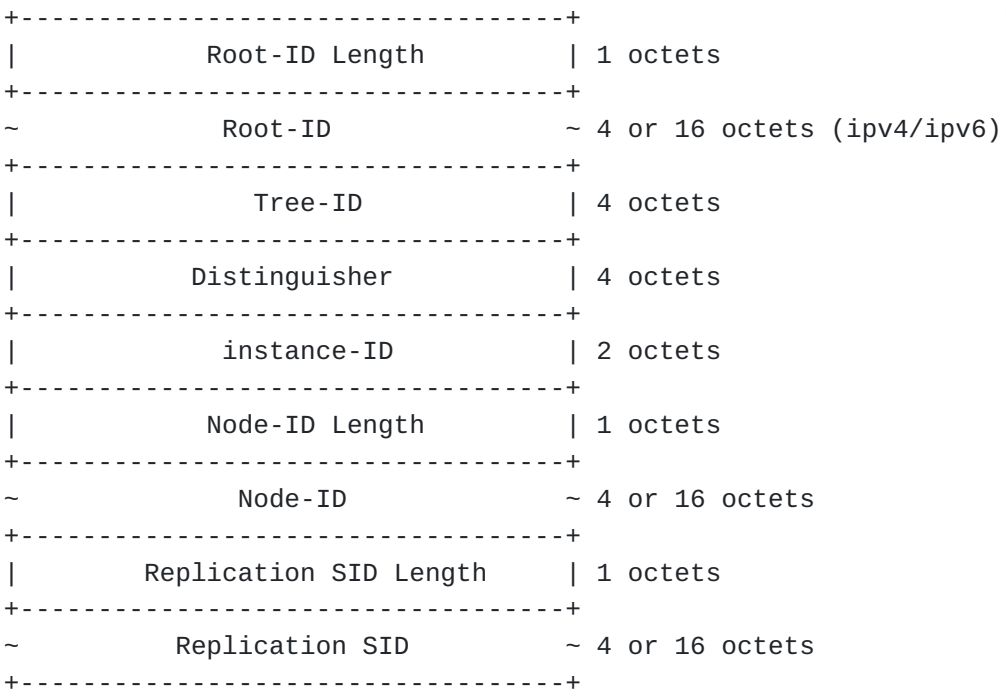
*Root-ID: IPv4/IPv6 address of the head-end (Root) of the p2mp tree, based on AFI.

*Tree-ID: a unique 4 octets identifier of the p2mp tree on the head- end (root)router.

*Distinguisher: 4-octets value uniquely identifying the policy in the context of <Tree-ID, Originating Router's IP> tuple. The distinguisher has no semantic value and is solely used by the SR P2MP Policy originator to make unique (from an NLRI perspective) multiple occurrences of the same SR P2MP Policy.

3.1.2. Replication segment Route Binding SID- Route type TBD 2

There can be two type of replication segment, shared and non-shared. A shared replication segment can carry multiple MVPN services or it can be used for Facility Fast reroute protecting multiple P2MP trees. A non-shared tree is used when the label field of the PMSI Tunnel Attribute (PTA) is set to 0 as per [draft-ietf-bess-mvpn-evpn-sr-p2mp]. The Binding SID route type Programs the incoming replication SID on the replication node. Since a replication cross connect has a single incoming replication SID with a set of Outgoing Interfaces, this route type can be used to download the replication SID once for the cross connect.



*Root-ID: IPv4/IPv6 address of the head-end (Root) of the p2mp tree based on AFI.

*Tree-ID: a unique 4 octets identifier of the p2mp tree on the head- end router (Root)

*instance-id, identifies the path-instance with in the p2mp-policy. Each candidate path can have one, two or more path-instance. Path-instance is used for global optimization of the candidate path via make before break procedures. Instance-ID can be used

*Distinguisher: 4-octets value uniquely identifying the policy in the context of <Root-ID, Tree-ID> tuple. The distinguisher has no semantic value and is solely used by the SR P2MP Policy originator to make unique (from an NLRI perspective) multiple occurrences of the same SR P2MP Policy.

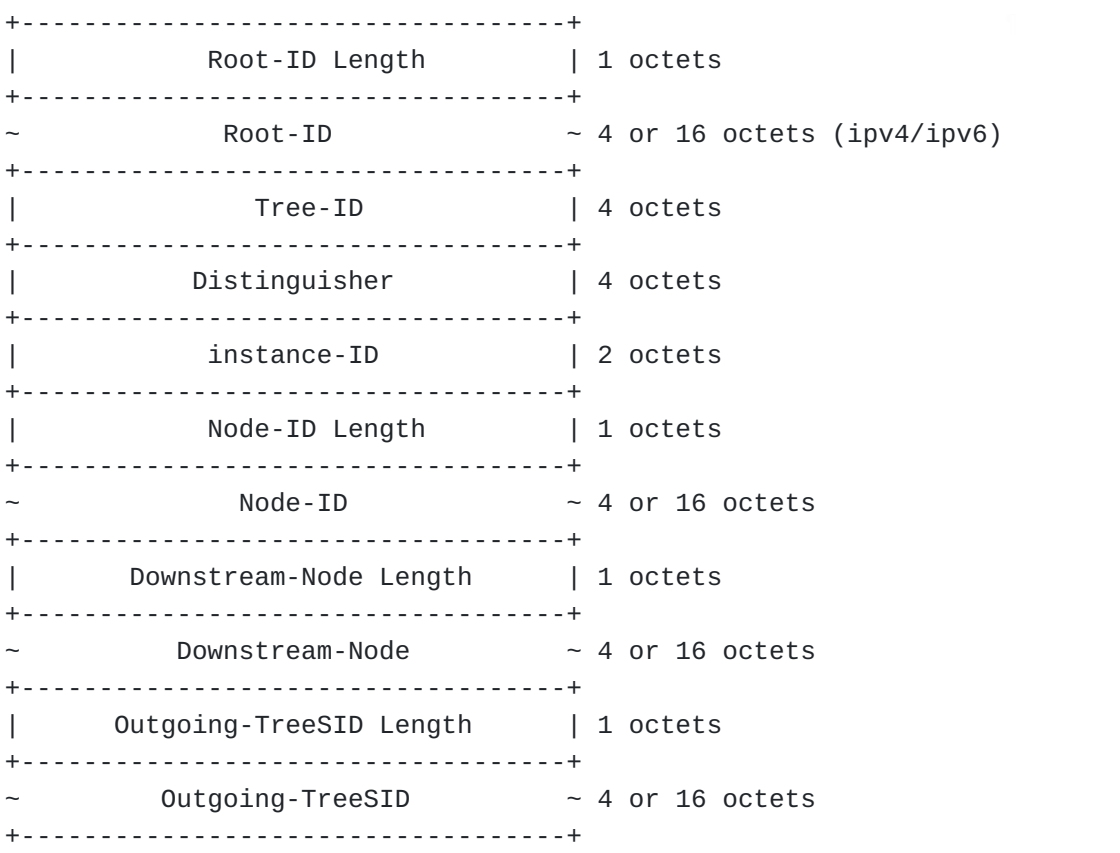
*Node-ID: This Node's IPv4/IPv6 address

*Replication SID: the incoming replication SID used to identify this replication point (MPLS or SRv6). Note the replication SID is not part of the NLRI key.

3.1.3. Replication segment Route OIF- Route type TBD 3

This route type is used to identify and program each out going interface individually for a replication cross connect. Downloading each OIF individually ensures easier modification and programming and will keep the programming of each OIF in par with [draft-ietf-

idr-segment-routing-te-policy] . Note: this route type can be used for shared and non-shared replication segment as it was explained in previous sections.



*Root-ID: IPv4/IPv6 address of the head-end (Root) of the p2mp tree based on AFI.

*Tree-ID: a unique 4 octets identifier of the p2mp tree on the head- end router (Root)

*instance-id, identifies the path-instance with in the p2mp-policy. Each candidate path can have one, two or more path-instance. Path-instance is used for global optimization of the candidate path via make before break procedures. Instance-ID can be used

*Distinguisher: 4-octets value uniquely identifying the policy in the context of <Root-ID, Tree-ID> tuple. The distinguisher has no semantic value and is solely used by the SR P2MP Policy originator to make unique (from an NLRI perspective) multiple occurrences of the same SR P2MP Policy.

*Node-ID: Node's IPv4/IPv6 address

*Downstream Node: Downstream Node Identifier

*Outgoing TreeSID: The outgoing SID for this branch (MPLS or SRv6). Note the outgoing-TreeSID is not part of the NLRI Key.

3.2. Tunnel Encapsulation Attribute

The content of this new NLRI is encoded in the tunnel Encapsulation Attribute originally defined in [RFC9012] using two new Tunnel-Type TLV (codepoint is TBD, assigned by IANA from the "BGP Tunnel Encapsulation Attribute Tunnel Types" registry) one for P2MP Policy and another for Replication segment.

3.2.1. SR P2MP policy encoding

```
SR P2MP Policy SAFI NLRI: <route-type p2mp-policy>
Attributes:
   Tunnel Encaps Attribute (23)
      Tunnel Type: (TBD, P2MP-Policy)
         Preference
         Policy Name
         Policy Candidate Path Name
         leaf-list (optional)
            remote-end point
            remote-end point
                     ...
         path-instance
            active-instance-id
            instance-id
            instance-id
                    ...
```

*Relevant only at the Root.

*SR P2MP-POLICY NLRI and P2MP Policy route type.

*Tunnel Encapsulation Attribute is defined in [RFC9012]

*Tunnel-Type is set to P2MP-Policy Tunnel-Type TBD (assigned by IANA from the "BGP Tunnel Encapsulation Attribute Tunnel Types" registry).

*Policy Name, Policy Candidate Path Name are defined in [draft-ietf-idr-segment-routing-te-policy]

*Preference, leaf-list, remote-end point and path- instance, instance-ids are defined in this document.

*Additional sub-TLVs may be defined in the future.

3.2.2. Replication segment Binding SID encoding

```
replication segment Binding SID SAFI NLRI:
              <route-type non-sahred/shared
               tree replication-segment-binding-sid>
```

This route type has no additional sub-TLVs, and it is only meant to download the incoming SID for the replication cross connect.

3.2.3. Replication segment OIF encoding

```
replication segment SAFI NLRI: <route-type replication-segment-binding-s
                                          or
                                          replication-segment-oif>
Attributes:
   Tunnel Encaps Attribute (23)(Optional)
      Tunnel Type: (TBD Replication-Segment-oif)
         segment-list
            weight (optional)
            protection (optional, must be present when protection flag i
            segment
            segment
              ...
         segment-list
            weight (optional)
            protection (optional, must be present when protection flag i
            segment
            segment
              ...
         segment-list (protection segment list)
            protection (protecting the first segment list, can't have we
            segment
            segment
            ...
       ...
    ...
```

*SR P2MP-POLICY NLRI and non-shared tree Replication segment route type or shared tree Replication segment route type.

*Tunnel Encapsulation Attribute is defined in [RFC9012].

*Tunnel-Type is set to Replication Segment OIF Tunnel Type, TBD (assigned by IANA from the "BGP Tunnel Encapsulation Attribute Tunnel Types" registry). Note: this tunnel-type is optional. The outgoing replication sid is part of the replication segment OIF Route type. As such if the P2MP Policy is creating a hop by hop tree then there is no need for this TEA. If the P2MP policy is programing a tree where 2 or more replication segments are connected via unicast SR Policy or SID List then this TEA can be used. The Segments are unicast SIDs or binding SID. IF binding

SID used then the segment-list can be bind to a existing unicast SR Policy.

*segment-list are defined in this document.

*Additional sub-TLVs may be defined in the future.

3.3. P2MP Policy Sub-TLVs

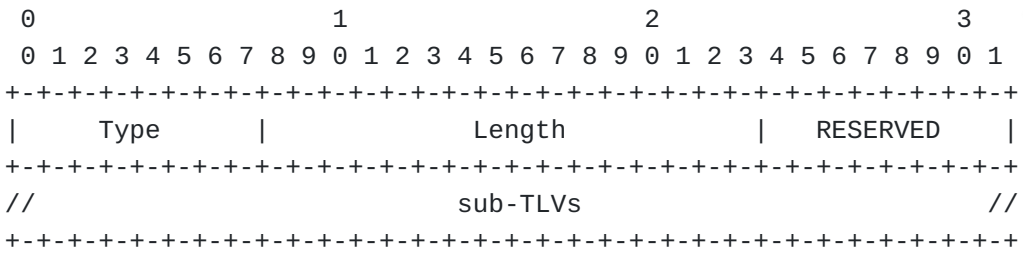
EACH P2MP policy NLRI represents a candidate path for a P2MP policy. A P2MP policy can have multiple candidate paths and would need multiple P2MP policy NRLI to download all the candidate paths.

3.3.1. preference Sub-TLV

As defined in preference Sub-TLV section in [draft-ietf-idr-segment-routing-te-policy] the candidate path with highest preference is the active candidate path.

3.3.2. leaf-list Sub-TLV

The leaf list sub-tlv identifies a set of leaves for the tree. Each leaf is a remote endpoint as defined in [RFC9012] The leaf-list sub-tlv is optional. The PCE can choose to download the leaf list every time it is configured or learns a new leaf. If the PCE chooses to download this optional sub-tlv it should download the entire set of the end-points every time the endpoint list has been modified. The leaf list has informational value only hence why it is optonal and it is not required for the root PE to operate. However, it must be noted that in some cases the end-points list can become very large with 100s of leaves.



*Type: TBD, 1 octet

*Length: 2 octets, the total length (not including the Type and Length fields) of the sub-TLVs encoded within the leaf-list sub-TLV.

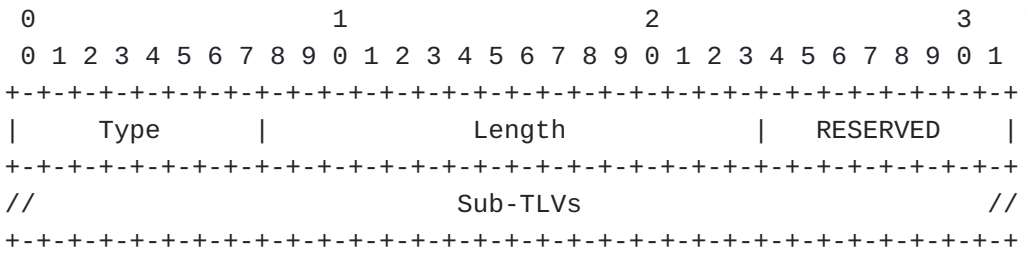
*RESERVED: 1 octet of reserved bits. SHOULD be unset on transmission and MUST be ignored on receipt.

*sub-TLVs: One or more remote endpoint sub-TLVs. Note the remote endpoint object is defined in [RFC9012]

3.3.3. path-instance Sub-TLV

The path instance sub-tlv contains a set of instance-ids (P2MP LSPs). These LSPs can be used for MBB procedure under a candidate path. Each LSP Instance-id has a unique id (4 octets) with in the <root node, P2MP policy>, in other word it is unique per <root node,tree-id>. The PCE SHOULD always download all instance-ids to the node. The active instance is identified via the active instance-id sub-tlv.

The P2MP LSP and its replication segments should be configured from root to the leaves first before the PCE switches that active instance-id to this new instance.



*Type: TBD, 1 octet

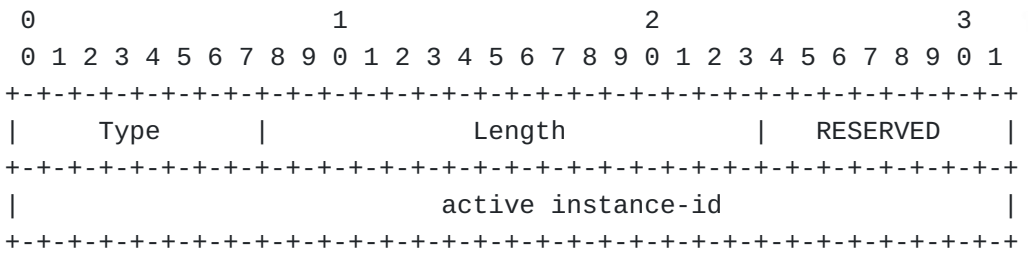
*Length: 2 octets, the total length (not including the Type and Length fields) of the sub-TLVs encoded within the Segment List sub-TLV.

*RESERVED: 1 octet of reserved bits. SHOULD be unset on transmission and MUST be ignored on receipt

*sub-TLVs: * active instance-id * one or more instance-id

3.3.3.1. active instance-id Sub-TLV

The Active instance-id is used to identify the P2MP LSP which should be active amongst the collection of instances.



*Type: TBD.

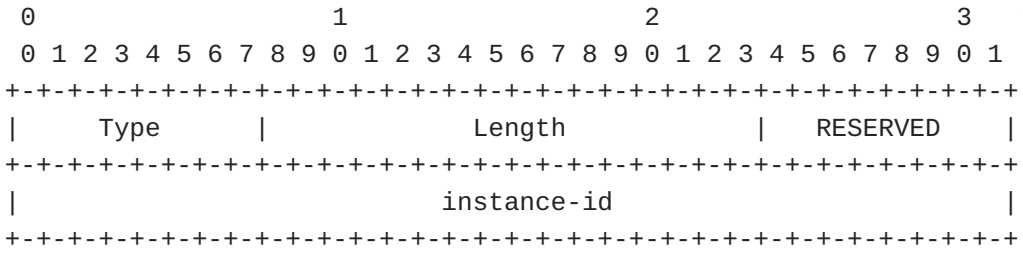
*Length: the total length (not including the Type and Length fields) of the sub-TLVs encoded within the Segment List sub-TLV.

*RESERVED: 1 octet of reserved bits. SHOULD be unset on transmission and MUST be ignored on receipt.

*active instant-id: The identifier of the active instance-id

3.3.3.2. instance-id Sub-TLV

Multiple Instance-ids can be programmed for a candidate path.



*Type: TBD

*Length: the total length (not including the Type and Length fields) of the sub-TLVs encoded within the Segment List sub-TLV.

*RESERVED: 1 octet of reserved bits. SHOULD be unset on transmission and MUST be ignored on receipt.

*instan-id: a 32 bit unique identifier. The instance-id is unique with in the context of the <root node, p2mp policy>

3.4. Replication segment Sub-TLVs

3.4.1. Segment list Sub-TLV

The segment list Sub-TLV is defined in [draft-ietf-spring-segment-routing-policy]. The segment-list Sub-TLV contains one or more segment Sub-TLVs. Two replication segments can be directly connected via a replication sid or can be connected via a unicast segment list and a replication sid. In the later case the replication sid needs to be at the bottom of the unicast segment list.

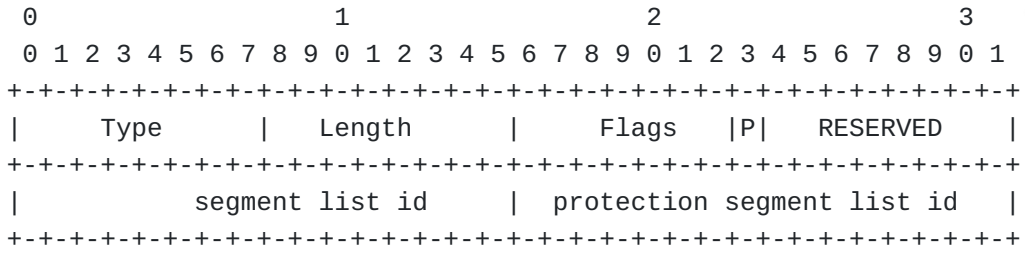
3.4.2. Weight sub-tlv

The Weight sub-TLV is optional and is as defined in [draft-ietf-idr-segment-routing-te-policy]. With in the downstream node sub-tlv, there can be one or more segment list used for ECMP. In this case the weight sub-tlv can provide weighted ECMP.

3.4.3. Protection sub-tlv

Protection sub-tlv is optional, if FRR is desired for the downstream node this sub-tlv can be used to identify the protection segment list. To identify protection segment list this sub-tlv provides a segment list identifier. If protection is desired under the endpoint all the segment lists should have this sub-tlv. A protection segment list can not have a weight sub-tlv and it can not participate in ECMP. That said a segment list that is being protected can have a weight sub-tlv and participate in ECMP.

In general protection segment list is used only if replication segments are directly connected and there is no unicast segment list connecting two replication segment. If there is a unicast replication segment connecting the two replication sid, then the unicast protection mechanism can be exercise and there is no need for this protection sub-tlv, hence why this sub-tlv is optional.



*Type : tbd, 1 octet.

*Length: 8

*Flag: 1 octet, the P bit is set when this segment list is protected by another segment list for the downstream node

*segment list id: the segment list id

*protection segment list id: the segment list id that is being used as protection.

3.4.4. Segment Sub-TLV

The segment sub-Tlv is identified in [draft-ietf-idr-segment-routing-te-policy]. As it was mentioned before two replication segments can be connected directly to each other or via a segment list. If they are connected directly to each other then the segment list can be constructed via:

*If the replication segment is steered via IPv4 or IPv6 nexthops or interface then the segment type E or G can be used with the new R flag set.

*If the replication segment is steered via a SR Unicast node or adjacency SID then segment type A can be used with the new R flag set. Unicast SR segment types can also be configured for steering.

If they are connected via SR domain then the segment list can contain multiple different types of SIDs, such as Node, Adjacency or Binding SIDe. In this case the replication sid is at the bottom of the stack and of type A with the R flag set. The SR node/adjacency or binding sids steer the packet through a SR domain until it reaches another replication segment. where the bottom of the stack replication sid identifies the forwarding information on that replication segment.

It should be noted that the segment sub-TLV is only used to program the unicast SR Segment or outgoing interface for the replication SID outgoing interface. The outgoing tree SID it self is programmed in the appropriate route type.

4. P2MP Policy Operation

Inline with [draft-ietf-idr-segment-routing-te-policy] the consumer of an P2MP Policy is not the BGP process. The BGP process is used for distributing the P2MP policy NLRI and its route-types but its installation and use is outside the scope of BGP. The detail for P2MP Policy can be found in [draft-ietf-pim-sr-p2mp-policy]

4.1. Configuration and advertisement of P2MP Policies

The controller usually is connected to the receivers via a route reflector. As such one or more route-target SHOULD be attached to the advertisement of P2MP Policy NLRI and its route-type. Each route target identifies one head-end (root nodes) for P2MP Policy route or one or more head-end, transit and leaf nodes for the Non- Shared/ Shared Tree Replication Segment route, for the advertised P2MP Policy.

4.2. Reception of an P2MP Policy NLRI

When a BGP speaker receives an P2MP Policy NLRI the following rules apply:

*The P2MP Policy update MUST have either the NO_ADVERTISE community or at least one route-target extended community in IPv4-address format. If a router supporting this document receives an P2MP Policy update with no route-target extended communities and no NO_ADVERTISE community, the update MUST NOT be processed. Furthermore, it SHOULD be considered to be malformed, and the "treat-as-withdraw" strategy of [RFC7606] is applied.

*If one or more route-targets are present, then at least one route- target MUST match one of the BGP Identifiers of the receiver in order for the update to be considered usable. The BGP Identifier is defined in [RFC4271] as a 4 octet IPv4 address. Therefore the route- target extended community MUST be of the same format.

*If one or more route-targets are present and no one matches any of the local BGP Identifiers, then, while the P2MP Policy NLRI is acceptable, it is not usable on the receiver node.

4.3. Global Optimization for P2MP LSPs

When a P2MP LSP needs to be optimized for any reason (i.e. it is taking on an FRR Path or new routers are added to the network) a global optimization is possible. Note that optimization works per candidate path. Each candidate path is capable of global optimization. To do so each candidate path contains two or more path- instances. Each path instance is a P2MP LSP, each P2MP LSP is identified via a path-instance-id (equivalent to an lsp-id [RFC3209]). After calculating an optimized P2MP LSP path the PCE will program the candidate path with a 2nd path instance and its set of replication segments for this path-instance on the root, transit and leaf nodes. After the optimized LSP replication segments are downloaded a MBB procedure is performed and the previous instance of the path instance is deleted and removed from head-end node and its corresponding replication segments from head-end, transit and leaves.

5. IANA Consideration

*A new SAFI is defined: the SR P2MP Policy SAFI, (Codepoint tbd assigned by IANA)

*3 new Route type field defines the encoding of the rest of the P2MP- POLICY SAFI

  -P2MP Policy Route

  -Replication Segment Binding Sid

  -Replication Segment OIF

*Two new Tunnel type to be assigned by IANA

  -P2MP-Policy Tunnel-Type

  -Replication Segment OIF Tunnel Type

6. Security Considerations

TBD

7. Acknowledgments

8. References

8.1. Normative References

[RFC2119] "S. Bradner "Key Words for use in RFCs to Indicate Requirement levels"", October 2019.

8.2. Informative References

[draft-ietf-bess-mvpn-evpn-sr-p2mp] "R. Parekh, C. Filsfils, A.V. Venkateswaran, H. Bidgoli, D. Voyer, Z. Zhang "Multicast and Ethernet VPN with Segment Routing P2MP"".

[draft-ietf-idr-segment-routing-te-policy] "s. Previdi, C. Filsfils, K. Talaulikar, P. Mattes, D. Jain, S. Lin "Advertise Segment Routing Policies in BGP"".

[draft-ietf-pim-sr-p2mp-policy] "D. Yoyer, C. Filsfils, R.Prekh, H.bidgoli, Z. Zhang, "draft-ietf-pim-sr-p2mp-policy"", October 2019.

[draft-ietf-spring-segment-routing-policy] "C. Filsfils, K. Talaulikar, D. Voyer, A. Bogdanov, P. Mattes "Segment Routing Policy Architecture"".

[RFC4271] "Y. Rekhter, T. Li, S. Hares "A Border Gateway Protocol 4 (BGP-4)"".

[RFC4760] "T. Bates, R. Chandra, D. Katz, Y. Rekhter "Multiprotocol Extensions for BGP-4"".

[RFC6513] "E. Rosen, R. Aggarwal "Multicast in MPLS/BGP IP VPNs"".

[RFC7606] "e. Chen, J. Scudder, P. Mohapatra, K. Patel "Revised Error handling for BGP UPDATE Messages"".

[RFC9012] "K. Patel, G. Van de Velde, S. Sangli, J. Scudder "The BGP Tunnel Encapsulation Attribute"".

Authors' Addresses

Hooman Bidgoli (editor)
Nokia
Ottawa

Canada

Email: hooman.bidgoli@nokia.com

Daniel Voyer
Bell Canada
Montreal
Canada

Email: daniel.yover@bell.ca

Andrew Stone
Nokia
Ottawa
Canada

Email: andrew.stone@nokia.com

Rishabh Parekh
Cisco System, Inc.
San Jose,
United States of America

Email: riparekh@cisco.com

Serge Krier
Cisco System, Inc.
Rixensart
Belgium

Email: sekrier@cisco.com

Swadesh Agrewal
Cisco System, Inc.
San Jose,
United States of America

Email: swaagraw@cisco.com