Workgroup: Network Working Group
Internet-Draft:
draft-ietf-idr-multinexthop-attribute-00
Published: 17 March 2024
Intended Status: Standards Track
Expires: 18 September 2024
Authors: K. Vairavakkalai, Ed. M. Jeyananth
Juniper Networks, Inc. Juniper Networks, Inc.
M. Nanduri Lingala
Microsoft AT&T

BGP MultiNexthop Attribute

Abstract

Today, a BGP speaker can advertise one nexthop for a set of NLRIs in an Update. This nexthop can be encoded in either the top-level BGP-Nexthop attribute (code 3), or inside the MP_REACH_NLRI attribute (code 14).

This document defines a new optional non-transitive BGP attribute called "MultiNexthop (MNH)" with IANA BGP attribute type code TBD, that can be used to carry an ordered set of one or more Nexthops in the same route, with forwaring information scoped on a per nexthop basis.

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 18 September 2024.

Copyright Notice

Copyright (c) 2024 IETF Trust and the persons identified as the document authors. All rights reserved.

This document is subject to BCP 78 and the IETF Trust's Legal Provisions Relating to IETF Documents (https://trustee.ietf.org/license-info) in effect on the date of publication of this document. Please review these documents carefully, as they describe your rights and restrictions with respect to this document. Code Components extracted from this document must include Revised BSD License text as described in Section 4.e of the Trust Legal Provisions and are provided without warranty as described in the Revised BSD License.

Table of Contents

1. Introduction

Today, a BGP speaker can advertise one nexthop for a set of NLRIs in an Update. This nexthop can be encoded in either the top-level BGP-Nexthop attribute (code 3), or inside the MP_REACH_NLRI attribute (code 14).

This document defines a new optional non-transitive BGP attribute called "MultiNexthop (MNH)" with IANA BGP attribute type code TBD, that can be used to carry an ordered set of one or more Nexthops in the same route, with forwaring information scoped on a per nexthop basis.

2. Terminology

SN: Service Node

iSN: Ingress Service Node

eSN: Egress Service Node

NLRI: Network Layer Reachability Information

AFI: Address Family Identifier

SAFI: Subsequent Address Family Identifier

PE : Provider Edge

RT : Route-Target extended community

RD : Route-Distinguisher

MPLS: Multi Protocol Label Switching

ECMP: Equal Cost Multi Path

WECMP: Weighted Equal Cost Multi Path

FRR: Fast Re Route

PNH : Protocol Next hop address carried in a BGP Update message

MNH: BGP MultiNextHop attribute

NFI: Nexthop Forwarding Information

FI: Forwarding Instruction

FA: Forwarding Argument

2.1. Definitions

MULTI_NEXT_HOP (aka MNH): BGP MultiNexthop attribute. The new attribute defined by this document.

MNH TLV: Top level TLV contained in a MULTI_NEXT_HOP.

NFI TLV: Nexthop Forwarding Information TLV, contained in a MNH TLV.

FI TLV: Forwarding Instruction TLV, contained in a NFI TLV.

FA TLV: Forwarding Argument TLV, contained as an argument to a FI in the FI TLV.

Service Family : BGP address family used for advertising routes for "data traffic" as opposed to tunnels (e.g. AFI/SAFIs 1/1 or 1/128).

Transport Family : BGP address family used for advertising tunnels, which are in turn used by service routes for resolution (e.g. AFI/SAFIs 1/4 or 1/76).

3. Motivation

For cases where multiple nexthops need to be advertised, BGP Addpath [RFC7911] is used with some address families. On some other address families like Flowspec, nexthop addresses are carried in one or more extended communities of specific type.

Though Addpath allows basic ability to advertise multiple-nexthops, it does not allow the sender to express the desired relationship between the multiple nexthops being advertised e.g., relative ordering, type of load balancing, fast reroute. These are local decisions at the receiving node based on local configuration and path selection between the various additional paths, which may tie-break on some arbitrary step like Router-Id or BGP nexthop address. Some scenarios with a BGP free core may benefit from having a mechanism, where egress node can signal multiple nexthops along with their relationship to ingress nodes.

It would be desirable to have a common way to carry one or more nexthops on a BGP route of any family.

This document defines a new optional non-transitive BGP attribute "MultiNexthop (MNH)" that can be used for this purpose.

The MNH attribute can be used in any BGP family that wants to carry one or more nexthops, with forwaring information scoped on a per nexthop basis. E.g. The MNH can be used to advertise MPLS label along with nexthop for labeled and unlabeled families (e.g. Inet Unicast, Inet6 Unicast, Flowspec) alike. Such that, mechanisms at the transport layer can work uniformly on labeled and unlabled BGP families to realize various usecases.

The MNH plays different role in "downstream allocation" scenario than "upstream allocation" scenario. E.g. for [RFC8277] families that advertise downstream allocated labels, the MNH can play the "Label Descriptor" role, describing the forwarding semantics of the label being advertised. This can be useful in network visualization and controller based traffic engineering (e.g. EPE).

4. Protocol Operations

4.1. Scope of Use, and Propagation

The MNH attribute is intended to be used in a BGP free core, between egress and ingress BGP speakers that understand this attribute. These BGP speakers may have an intra-AS or inter-AS BGP session between them.

To avoid un-intentionally leaking the MNH to another AS, via a BGP speaker that does not understand MNH attribute, it is defined as

"optional non-transitive". But this also means that a RR needs to be upgraded to support this attribute before any PEs in the network can make use of it.

Use of MNH on a BGP session is disabled by default. An implementation MUST provide configuration control on a per BGP neighbor address family basis, to enable MNH support. A sending BGP speaker MUST NOT advertise MNH on a BGP address family route update if MNH support is not enabled for the address family on the BGP session.

If the MNH attribute is received on a BGP session where MNH support is not enabled, the attribute MUST be treated as Unrecognized non-transitive. This rule provides additional protection against unintended propagation of this attribute, when both BGP speakers understand MNH but receiver has not enabled the support. A RFC3392 Capability is not used for this purpose, because it would cause session reset whenever 'MNH support' config is changed.

When a BGP speaker receives the MNH attribute on a BGP route in an address family that is MNH enabled, it propagates the attribute unchanged when readvertising the route with nexthop unchanged on a BGP session address family that is also MNH enabled. The BGP speaker excludes the MNH attribute when readvertising the route with nexthop unchanged on a BGP session that is not MNH enabled.

When a BGP speaker receives the MNH attribute on a BGP route in an address family that is MNH enabled, it processes the attribute but does not propagate it further when readvertising the route with nexthop altered. The BGP speaker MAY attach a new instance of MNH attribute when readvertising the route (with nexthop unchanged or altered) on a BGP session that is MNH enabled.

A BGP speaker re-advertising a BGP route with nexthop unchanged MAY add the MNH attribute on the reflected BGP route, on behalf of the originating BGP speaker. The "Advertising PNH field" in the MNH is set to the Nexthop field in BGP route being re-advertised.

4.2. Interaction of MNH with Nexthop (in attr codes 3, 14)

When adding a MultiNexthop attribute to an advertised BGP route, the speaker MUST put the same next-hop address in the Advertising PNH field as it put in the Nexthop field inside MP_REACH_NLRI attribute if one exists, or the NEXT_HOP attribute.

A speaker that recognizes the MNH attribute and does not change the PNH while readvertising the route, e.g. a Route Reflector, MUST propagate unchanged the MultiNexthop attribute in the readvertisement, satisfying the propagation scope constraints described in previous section.

A speaker that recognizes MNH attribute and changes the PNH while readvertising the route MUST remove the MNH attribute in the readvertisement. The speaker MAY however add a new MNH attribute to the re-dvertisement. While doing so the speaker MUST record in the "Advertising PNH" field the same next-hop address as used in MP_REACH_NLRI attribute if one exists, or the NEXT_HOP attribute.

A speaker receiving a MNH attribute SHOULD ignore it if the next-hop address contained in 'Advertising PNH' field is not the same as the nexthop address contained in MP_REACH_NLRI attribute if one exists, or the NEXT_HOP attribute.

In case of [RFC2545], the global (non link-local) IPv6 address should be used for this purpose.

As specified in [RFC7606] BGP update message can contain no more than one instance of MP_REACH attribute or NEXT_HOP attribute. Similarly, a BGP update MUST contain only one instance of MNH attribute. If the MNH attribute (whether recognized or unrecognized) appears more than once in an UPDATE message, then all the occurrences of the attribute other than the first one SHALL be discarded and the UPDATE message will continue to be processed.

4.3. Interaction with Addpath

[ADDPATH-GUIDELINES] suggests the following:

"Diverse path: A BGP path associated with a different BGP next-hop and BGP router than some other set of paths. The BGP router associated with a path is inferred from the ORIGINATOR_ID attribute or, if there is none, the BGP Identifier of the peer that advertised the path."

When selecting "diverse paths" for ADD_PATH as specified above, the MNH attribute should also be compared if it exists, to determine if two routes have "different BGP next-hop".

4.4. Path Selection Considerations

4.4.1. Determining IGP Cost

While tie breaking in the path-selection as described in [RFC4271], 9.1.2.2. step (e) viz. the "IGP cost to nexthop", consider the highest cost among the nexthop-legs present in this attribute.

The IGP cost thus calculated is also used when constructing AIGP TLV ([RFC7311])

4.5. Denoting Upstream or Downstream Semantics

MultiNexthop attribute may describe to a receiving speaker what the forwarding semantics of an Upstream-allocated label should be. This can be used with either labeled or unlabled BGP families.

A MultiNexthop attribute may also play "Downstream signaled Label Descriptor" role. A BGP speaker advertising a route carrying downstream allocated MPLS label MAY add this attribute to the BGP route, to "describe" to the receiving speaker what the label's forwarding semantics is at the Egress node.

Today semantics of a downstream-allocated label is known only to the egress node advertising the label. The speaker receiving the label-binding doesn't know what the label's forwarding semantic at the advertiser is. In some environments, it may be useful to convey this information to the receiving speaker. This may help in better debugging and manageability, or enable the receiving speaker, which could also be some centralized controller, make better decisions about which label to use, based on the label's forwarding-semantic.

While doing upstream-label allocation, this attribute can be used to convey the forwarding-semantics at the receiving node should be. Details of the BGP protocol extensions required for signaling upstream-label allocation are out of scope of this document, and are described in [MPLS-NAMESPACES].

In rest of this document, the use of term "Label" will mean downstream allocated label, unless specified otherwise as upstream-allocated label.

When using the MultiNexthop attribute for IP-routes, the Upstream role is used. Since IP prefixes are by nature upstream allocated, global scope.

5. Encoding of BGP MultiNexthop (MNH) Attribute

"MultiNexthop (MNH)" is a new BGP optional non-transitive attribute (code TBD), that can be used to carry an ordered set of one or more Nexthops in the same route, with forwaring information scoped on a per nexthop basis. This attribute describes forwarding instructions using TLVs described in this document.

This section describes the organization and encoding of the MNH attribute.

```
      MNH Attribute: {
            PrimaryPath {
                [Forwarding Instruction 1],
                 ..
                [Forwarding Instruction n]
            }
            BackupPath {
                [Forwarding Instruction 1],
                 ..
                [Forwarding Instruction n]
            }
            LabelDescriptor {
                [Forwarding Instruction 1],
                 ..
                [Forwarding Instruction n]
            }
      }

      Forwarding Instruction: {
          {FwdAction, Forwarding Arguments}
      }
```

Figure 1: Overview of MNH Attribute Layout - Eye candy summary

A MNH attribute consists of one of more "MNH TLVs". A MNH TLV contains a Type and one unit of Nexthop Forwarding Information (NFI TLV).

A NFI TLV contains one or more Forwarding Instructions (FI TLV).

A Forwarding Instruction TLV contains a "Forwarding Action" and one more "Forwarding Arguments" (FA TLVs). The Forwarding Arguments describe the parameters required to complete a Forwarding Action.

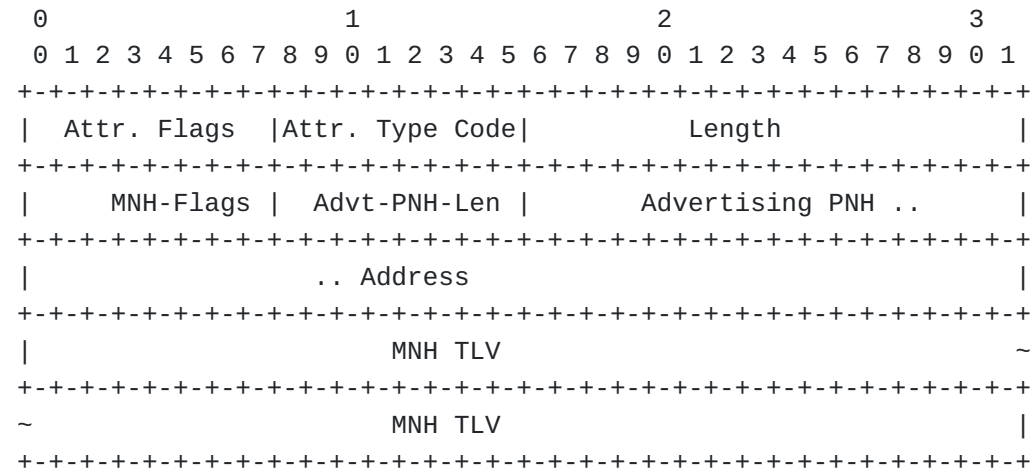


Figure 2: MultiNexthop - BGP Attribute

- Attr. Flags (1 octet)
  BGP Path-attribute flags. indicating an Optional Non-Transitive attribute. i.e. Optional bit set, Transitive bit reset.

- Attr. Type Code (1 octet)
  Type code allotted by IANA. TBD.

- Length (1 or 2 octets)
  One or Two bytes field stating length of attribute value in bytes

- MNH-Flags (1 octet)

```
 0 1 2 3 4 5 6 7
+-+-+-+-+-+-+-+-+
|R R R R R R R R|
+-+-+-+-+-+-+-+-+
```

  All bits are reserved.

  R: Reserved. MUST be set to zero, SHOULD be ignored by receiv

- Advt-PNH-Len (1 octet)
  Length in octets (4 for IPv4, 16 for IPv6, 12 for VPN-IPv4, 24 for VPN-IPv6) of Advertising PNH Address.

- Advertising PNH Address (Advt-PNH-Len octets)
  BGP Protocol Nexthop address advertised in NEXT_HOP or MP_REACH_N Used to sanity-check the MNH attribute. In case of RFC-2545, this the global (non link-local) IPv6 address.

- MNH TLVs: One or more MNH TLVs are carried in a MNH attr.
  MNH TLV is described in subsequent sections.

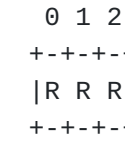
5.1. MNH TLV

The type of MNH TLV describes how the forwarding information carried in the MNH TLV is used.

```
 0                   1                   2                   3
 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|  MNH-TLV Flags| MNH. Type Code|          Length               |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|                             Value                             |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
```

Figure 3: MNH TLV

- MNH-TLV Flags (1 octet)



All bits are reserved.

R: Reserved. MUST be set to zero, SHOULD be ignored by receiv

MNH Type Code	Meaning
0	Reserved
1	Primary forwarding path
2	Backup forwarding path
3	Label Descriptor

- Length
  Length of Value portion in octects.

- Value
  Value portion contains the NFI TLV.

Type codes 1 and 2 are applicable for upstream allocated prefixes, example IP, MPLS, Flowspec routes.

Type code 4 describes the forwarding behavior given to downstream allocated MPLS label, adveritsed in BGP route.

Usage of Type code 1 in a BGP route containing IP prefix gives similar result as advertising the route with nexthop contained in BGP path-attributes: Nexthop (code 3) or MP_REACH_NLRI (code 14).

Upstream allocation for MPLS routes is achieved by using mechanisms explained in [MPLS-NAMESPACES].

If an invalid Type Code (like 0) is received, the TLV is ignored gracefully handing the error.

If an unknown Type Code is received, it SHOULD be ignored but propagated further when the MNH attribute is propagated, because nexthop is not changed.

If the received Type Code is incompatible for the prefix in BGP NLRI, the TLV should be ignored.

5.1.1. Primary Forwarding Path

Type Code = 1 means the TLV describes forwarding state to be programmed at receiving speaker as primary path nexthop leg. This TLV is used with Upstream allocated or global scope prefixes carried in BGP NLRI. Value part of this TLV contains Nexthop Forwarding Information TLV.

A BGP speaker uses the nexthop forwarding information received in this TLV as a primary path nexthop leg when programming the route for the NLRI prefix in its Forwarding table.

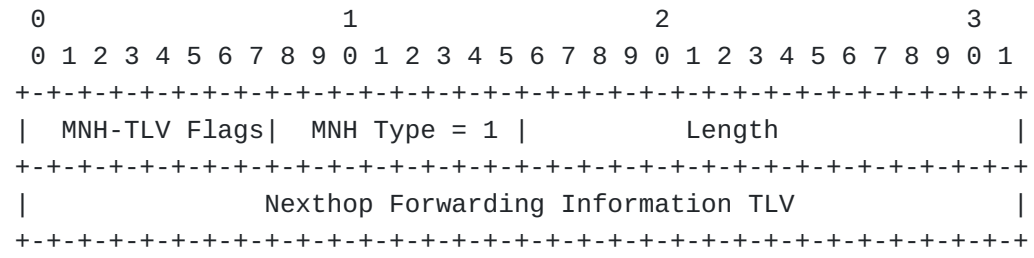


Figure 4: Primary forwarding path TLV

5.1.2. Backup Forwarding Path

Type Code = 2 means the TLV describes forwarding state to be programmed at receiving speaker as backup-path nexthop leg. This TLV is used with Upstream allocated prefixes or global scoped prefixes. Value part contains Nexthop Forwarding Information TLV.

Signaling a different nexthop for use as backup path is desired in some labeled forwarding scenarios, where two multihomed edge devices use each other as backup path to protect traffic when primary path fails.

This is required to avoid label advertisement oscillation between the multihomed PEs when they implement per-nexthop label allocation mode.

The label advertised by a PE1 for primary path advertisement is allocated/forwarded using external paths as primary leg and backup-path label from other multihomed PE2 as backup-path label. Such that primary-path label allocation at PE1 is not a function of the primary-path label advertised by PE2. Thus the primary path label remains stable at a PE and does not change when a new primary path label is received from the other multihomed PE. This prevents the label oscillation problem.

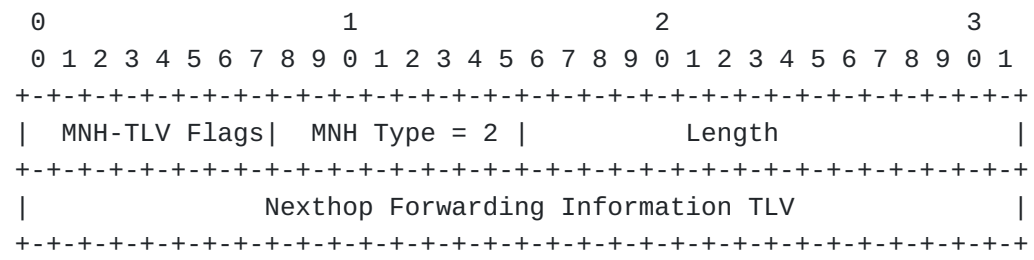


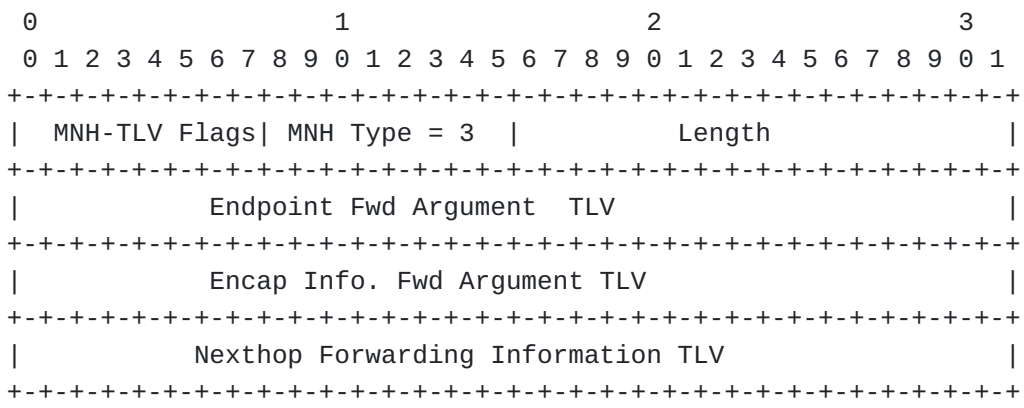
Figure 5: Backup forwarding path TLV

The backup path label allocated and advertised by a PE is a function of only the primary path. E.g. path to the CE device. So this label value does not change when a new label is received from the other multihomed PE

5.1.3. Label Descriptor

Type Code = 3 means the TLV describes forwarding state associated with downstream allocated MPLS label at the egress node identified in Endpoint FA TLV. Value part of this TLV contains Endpoint FA-TLV, Payload Info FA-TLV to identify the label being described, along with Nexthop Forwarding Information TLV that describes the forwarding state.

Signaling what a label advertised in BGP route signifies is helpful for debugging. The information provided by label descriptor can enable new usecases like network visualization and off box EPE decisions.



Endpoint Fwd Argument TLV:
Specifies the IP endpoint. Section 5.4.1.

Encap Info. Fwd Argument TLV:
Specifies the Label value being described. Section 5.4.3.

Nexthop Forwarding Information TLV:
Indicates the forwarding state. Described in Section 5.2.

Figure 6: Label Descriptor TLV

5.2. Nexthop Forwarding Information TLV

A Nexthop Forwarding Information TLV describes a MNH TLV. It contains one or more Forwarding Instruction TLVs. These Forwarding Instructions are the Forwarding Legs of the MNH.

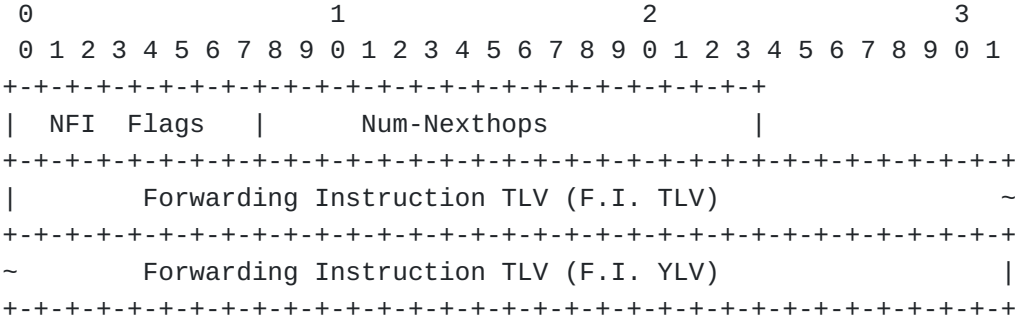
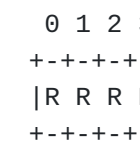


Figure 7: Nexthop Forwarding Information TLV

- NFI Flags (1 octet)



All bits are reserved.

R: Reserved. MUST be set to zero, SHOULD be ignored by receiv

- Num-Nexthops
  Number of F.I. TLVs.

- Forwarding Instruction TLV
  Each F.I. TLV describes a Nexthop Leg.
  Layout of Forwarding Instruction TLV is described in next sectio

5.3. Forwarding Instruction TLV

Each Forwarding Instruction TLV describes a Nexthop Leg. It expresses a "Forwarding Action" (FwdAction) along with arguments required to complete the action. The type of actions defined by this TLV are given below. The arguments are denoted by "Forwarding Argument TLVs". The Forwarding Argument TLVs takes appropriate values based on the FwdAction.

Each FwdAction should note the Arguments needed to complete the action. Any extranous arguments should be ignored. If the minimum set of arguments required to complete an action is not received, the Forwarding Instruction TLV should be ignored. Appropriate logging and diagnostic info MAY be provided by an implementation to help troubleshoot such scenarios.

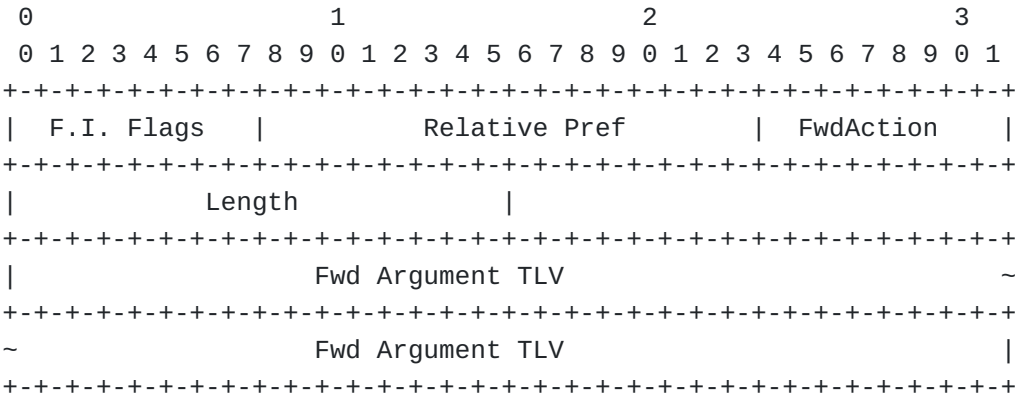
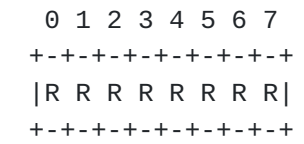


Figure 8: Forwarding Instruction TLV

- F.I. Flags (1 octet)



All bits are reserved.

R: Reserved. MUST be set to zero, SHOULD be ignored by receiv

- Relative Pref (2 octets)

Unsigned 2 octet integer specifying relative order or preference, a the many forwarding instructions, to use in FIB. All usable nexthop with lowest relative-pref are installed in FIB as primary-path. Thu multiple legs exist with that lowest relative-pref, ECMP is formed.

FwdAction	Meaning
0	Reserved
1	Forward
2	Pop-And-Forward
3	Swap
4	Push
5	Pop-And-Lookup
6	Replicate

Forwarding Instruction TLV with unknown FwdAction should be ignored, and rest of the attribute processed; gracefully handling the error. T may be appropriately logged for diagnosis.

- Length (2 octets)

Length in octets, of all Forwarding Argument TLVs.

Meaning of most of the above FwdAction semantics is well understood. FwdAction 1 is applicable for both IP and MPLS routes. FwdActions 2-5 are applicable for encapsulated payloads (like MPLS) only. FwdActions 1, 6 are applicable for Flowspec routes for Redirect and Mirror actions. FwdAction 6 can also be used to indicate multicast replication like functionality.

The "Forward" action means forward the IP/MPLS packet with the destination prefix (IP-dest-addr/MPLS-label) value unchanged. For IP routes, this is the forwarding-action given for next-hop addresses contained in BGP path-attributes: Nexthop (code 3) or MP_REACH_NLRI

(code 14). For MPLS routes, usage of this action is equivalent to SWAP with same label-value; one such usage is explained in [MPLS-NAMESPACES] when Upstream-label-allocation is in use.

The "Pop-And-Forward" action means Pop the payload header (e.g. MPLS-label) and forward the payload towards the Nexthop IP-address specified in the Endpoint Id TLV, using appropriate encapsulation to reach the Nexthop.

When applied to MPLS packet, the "Pop-And-Lookup" action may result in a MPLS-lookup or an upper-layer header (like IPv4, IPv6) lookup, depending on whether the label that was popped was the bottom of stack label.

If an incompatible FwdAction is received for a prefix-type, or an unsupported FwdAction is received, it is considered a semantic-error and MUST be dealt with as explained in "Error handling procedures" section.

5.4. Forwarding Argument TLV

The Forwarding Argument TLV describes various parameters required to execute a FwdAction.

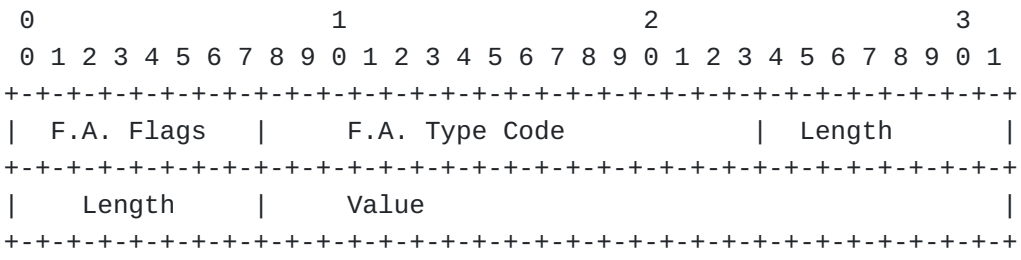


Figure 9: Forwarding Argument TLV

- F.A. Flags (1 octet)

```
          0 1 2 3 4 5 6 7
         +-+-+-+-+-+-+-+-+
         |R R R R R R R R|
         +-+-+-+-+-+-+-+-+
```

      All bits are reserved.

          R: Reserved. MUST be set to zero, SHOULD be ignored by receiv

F.A. Type Code	Meaning
0	Reserved
1	Endpoint Identifier
2	Path Constraints
3	Payload encapsulation info signaling
4	Endpoint attributes advertisement

- Length (2 octets)

   Length in bytes of Value field.

5.4.1. Endpoint Identifier

   F.A. Type Code = 1. This Forwarding Argument TLV identifies an
   Endpoint of different types.

```
 0                   1                   2                   3
 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|  F.A. Flags   |     F.A. Type Code =1         |  Length       |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|    Length     | Endpoint Type |  Endpoint Len | Endpoint Value|
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|                  Endpoint Value                               |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
```

Figure 10: Endpoint Identifier

- F.A. Flags (1 octet)

```
 0 1 2 3 4 5 6 7
+-+-+-+-+-+-+-+-+
|R R R R R R R R|
+-+-+-+-+-+-+-+-+
```

R: Reserved. MUST be set to zero, SHOULD be ignored by receiv

- Length (2 octets)
  Length in bytes of Value field.

Endpoint Type	Value	Len (octets)
0	Reserved	
1	IPv4 Address	4
2	IPv6 Address	16
3	MPLS Label (Upstream allocated or Global scope)	4
4	Fwd Context RD	8
5	Fwd Context RT	8

- Endpoint Len (1 octet)

  Length in bytes of Endpoint Value field.

5.4.2. Path Constraints

F.A. Type Code = 2. This Forwarding Argument TLV defines constraints for path to the Endpoint.

```
 0                   1                   2                   3
 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|  F.A. Flags   |     F.A. Type Code = 2        |  Length       |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|    Length     | ConstrainType | Constrain Len | ConstrainValue|
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|                  ConstrainValue                               |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
```

Figure 11: Path Constraints

- F.A. Flags (1 octet)

```
 0 1 2 3 4 5 6 7
+-+-+-+-+-+-+-+-+
|R R R R R R R R|
+-+-+-+-+-+-+-+-+
```

R: Reserved. MUST be set to zero, SHOULD be ignored by receiv

- Length (2 octets)
  Length in bytes of Value field.

ConstrainType	Value	Len (octets)
0	Reserved	
1	Proximity check	2
2	Transport Class ID (Color)	4
3	Load balance factor	2

- Constrain Len (1 octet)

Length in bytes of Constrain Value field.

- Proximity check Flags (2 octets)
  Flags describing whether the nexthop endpoint is expected to be away, or multihop away. Format of flags is described in next sec

- Transport Class ID (Color):

This is a 32 bit identifier, associated with the Nexthop address. The Nexthop IP-address specified in "Endpoint Identifier" TLVs are resolved over tunnels of this color. Defined in [BGP-CT] [draft-kaliraj-idr-bgp-classful-transport-planes

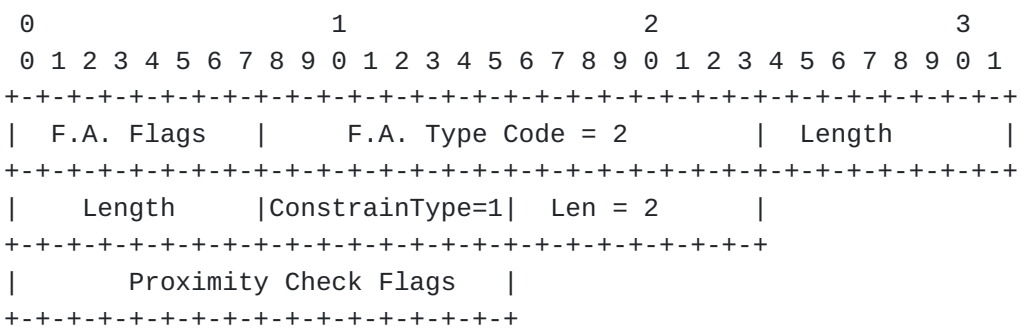
- Load balance factor (2 octets)
  Balance Percentage

5.4.2.1. Proximity Check

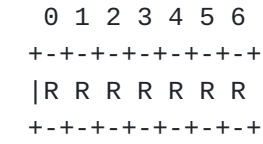
Usually EBGP singlehop received routes are expected to be one hop away, directly connected. And IBGP received routes are expected to be multihop away. Implementations today provide configuring exceptions to this rule.

The 'expected proximity' of the Nexthop can be signaled to the receiver using the Proximity check flags. Such that irrespective of whether the route is received from IBGP/EBGP peer, it can be treated as a single-hop away or multihop away nexthop.

The format of the Proximity check Sub-TLV is as follows:



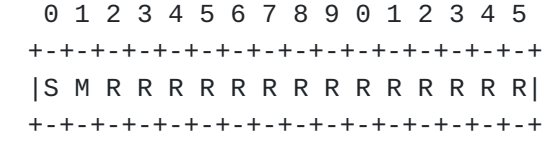
- F.A. Flags (1 octet)



R: Reserved. MUST be set to zero, SHOULD be ignored by receiv

- Length (2 octets)
  Length in bytes of Value field.

- Proximity check Flags (2 octets)



S: Restrict to Singlehop path.
M: Expect Multihop path.
R: Reserved. MUST be set to zero, SHOULD be ignored by receiv

Figure 12: Proximity check constrain

This TLV would be valid with Forwarding Instructions TLV with FwdAction of Forward, Pop-And-Forward, Swap or Push.

When S bit is set, receiver considers the nexthop valid only if it is directly connected to the receiver.

When M bit is set, receiver assumes that the nexthop can be multiple hops away, and resolves the path to the nexthop via another route.

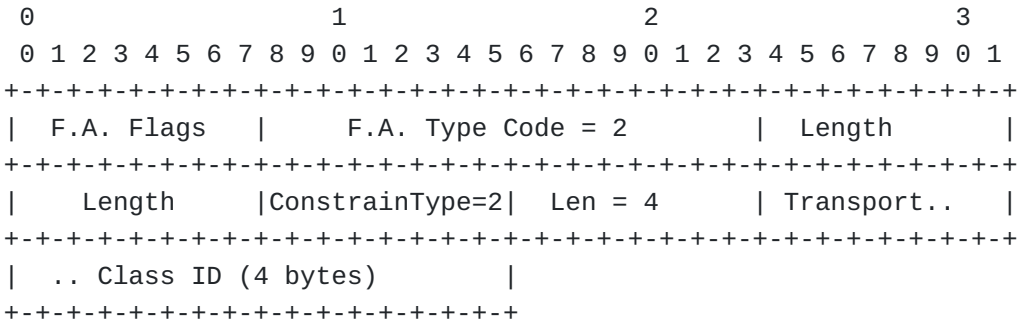
When both S and M bits are set, M bit behavior takes precedence. When both S and M bits are Clear, the current behavior of deriving proximity from peer type (EBGP is singlehop, IBGP is multihop) is followed.

5.4.2.2. Transport Class ID (Color)

The Nexthop can be associated with a Transport Class, so as to resolve a path that satisfies required Transport tunnel characteristics. Transport Class is defined in [BGP-CT]

Transport Class is a per-nexthop scoped attribute. Without MNH, the Transport class is applied to the nexthop IP-address encoded in the BGP-Nexthop attribute (code 3), or inside the MP_REACH_NLRI attribute (code 14). With MNH, the Transport Class can be specified per Nexthop-Leg (Forwarding Instruction TLV). It is applied to the IP-address encoded in the Endpoint Identifier TLV of type "IPv4 Address", "IPv6 Address" , "MPLS Label (Upstream allocated or Global scope)".

The format of the Transport Class ID Sub-TLV is as follows:



- F.A. Flags (1 octet)

R: Reserved. MUST be set to zero, SHOULD be ignored by receiv

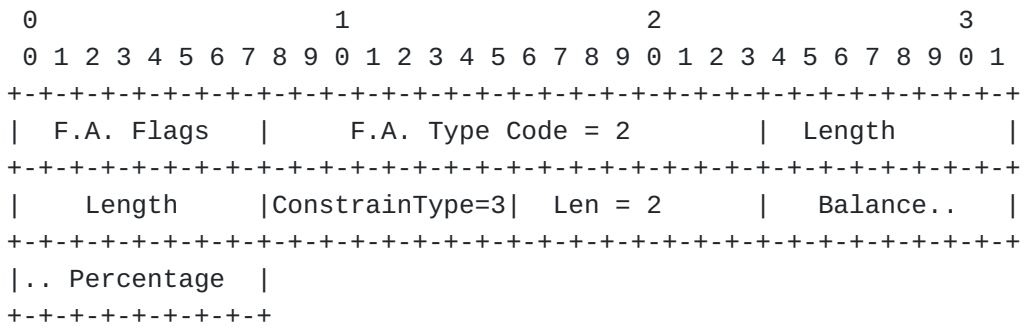
- Length (2 octets)
  Length in bytes of Value field.

- Transport Class ID (Color):
  This is a 32 bit identifier, associated with the Nexthop address.
  The Nexthop specified in Endpoint Identifier TLVs
  are resolved over tunnels of this color.
  Defined in [BGP-CT] [draft-kaliraj-idr-bgp-classful-transport-planes]

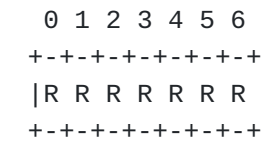
Figure 13: Transport Class ID (Color)

This TLV would be valid with Forwarding Instructions TLV with FwdAction of Forward, Swap or Push.

5.4.2.3. Load Balance Factor



- F.A. Flags (1 octet)



R: Reserved. MUST be set to zero, SHOULD be ignored by receiv

- Length (2 octets)
  Length in bytes of Value field.

- Len (1 octet)
  Length of the Constrain Value field.

- Balance Percentage:
  This is the explicit "balance percentage" requested by the sender, for unequal load-balancing over these Nexthop-Descriptor-TLV legs. This balance percentage would override the implicit balance-percentage calculated using "Bandwidth" attribute sub-TLV.

Figure 14: Load Balance Factor

This sub-TLV would be valid with Forwarding Instructions TLV with FwdAction of Forward, Swap or Push.

This is the explicit "balance percentage" requested by the sender, for unequal load-balancing over these Nexthop-Descriptor-TLV legs. This balance percentage would override the implicit balance-percentage calculated using "Bandwidth" attribute sub-TLV

When the sum of "balance percentage" on the nexthop legs does not equal 100, it is scaled up or down to match 100. The individual balance percentages in each nexthop leg are also scaled up or down proportionally to determine the effective balance percentage per nexthop leg.

5.4.3. Payload Encapsulation Info

F.A. Type Code = 3. This Forwarding Argument TLV defines payload encapsulation information.

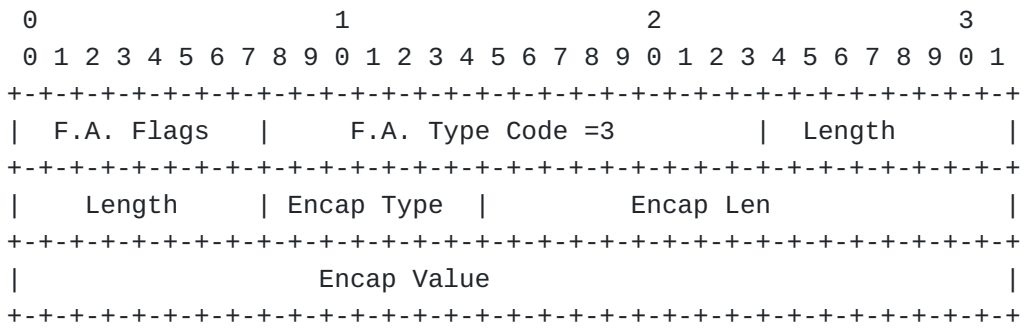
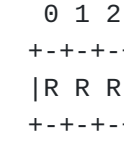


Figure 15: Payload Encapsulation Info

- F.A. Flags (1 octet)



R: Reserved. MUST be set to zero, SHOULD be ignored by receiv

- Length (2 octets)

Length in bytes of Value field.

Endcap Type	Value
0	Reserved
1	MPLS Label Info
2	SR MPLS label Index Info
3	SRv6 SID info
4	DSCP code point

- Encap Len (2 octets)

Length in octets of Encap Value field.

5.4.3.1. MPLS Label Info

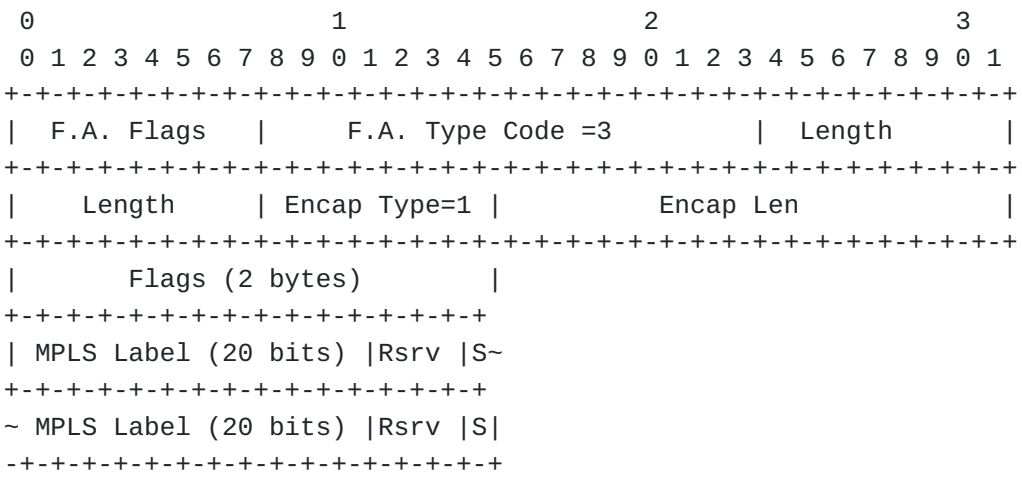
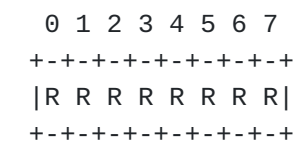


Figure 16: MPLS Label Info

- F.A. Flags (1 octet)



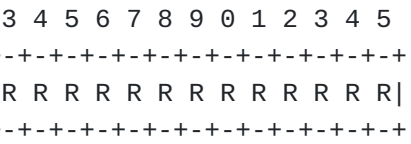
R: Reserved. MUST be set to zero, SHOULD be ignored by receiv

- Length (2 octets)
  Length in bytes of Value field.

- Encap Type
  = 1, to signify MPLS Label Info.

- Encap Len (2 octets)
  Length in bytes of following Encap Value field.

- Flags (2 octets):



E: ELC bit. Indicates if this egress NH is Entropy Label Capable.
   1 means the Entropy Label capable.
   0 means not capable to handle Entropy Label.

R: Reserved. MUST be set to zero, SHOULD be ignored by receiver.

- MPLS Label, Rsrv, S bit.
  20 bit MPLS Label stack encoded as in RFC 8277.
  S bit set on last label in label stack.

5.4.3.2. SR MPLS Label Index Info

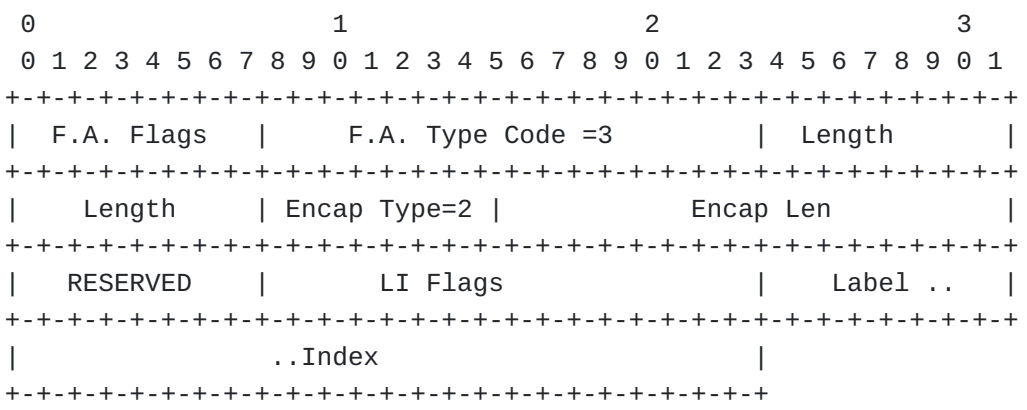
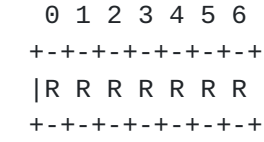


Figure 17: SR MPLS Label Index Info

- F.A. Flags (1 octet)



R: Reserved. MUST be set to zero, SHOULD be ignored by receiv

- Length (2 octets)
  Length in bytes of Value field.

- Encap Type
  = 2, to signify SR MPLS SID Info.

- Encap Len (2 octets)
  Length in bytes of following Encap Value field.

Rest of the value portion is encoded as specified in RFC-8669 sec 3.1.

- RESERVED: 8-bit field. MUST be set to zero, SHOULD be ignored by re

- LI Flags: 16 bits of flags. None defined. MUST be set to zero, SHOU

- Label Index:
  32-bit value representing the index value in the SRGB space.

5.4.3.3. SRv6 SID Info

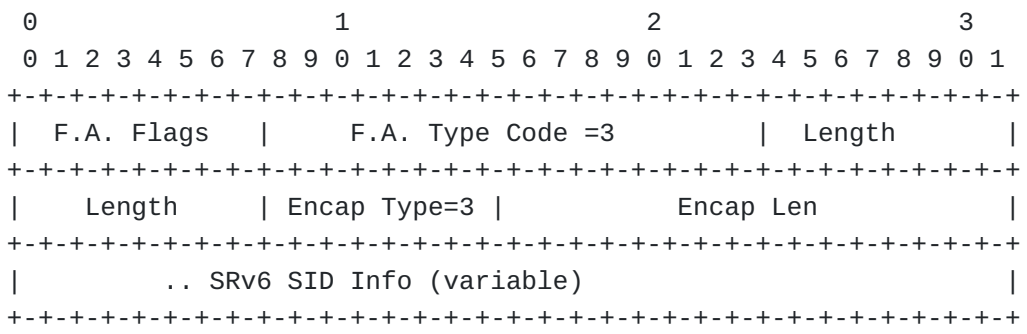
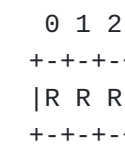


Figure 18: SRv6 SID Info

- F.A. Flags (1 octet)



R: Reserved. MUST be set to zero, SHOULD be ignored by receiv

- Length (2 octets)
  Length in bytes of Value field.

- Encap Type
  = 3, to signify SRv6 SID Info.

- Encap Len (2 octets)
  Length in bytes of following Encap Value field.

- SRv6 SID Info:
  SRv6 SID Information, as specified in RFC-9252 sec 3.1.

5.4.3.4. DSCP

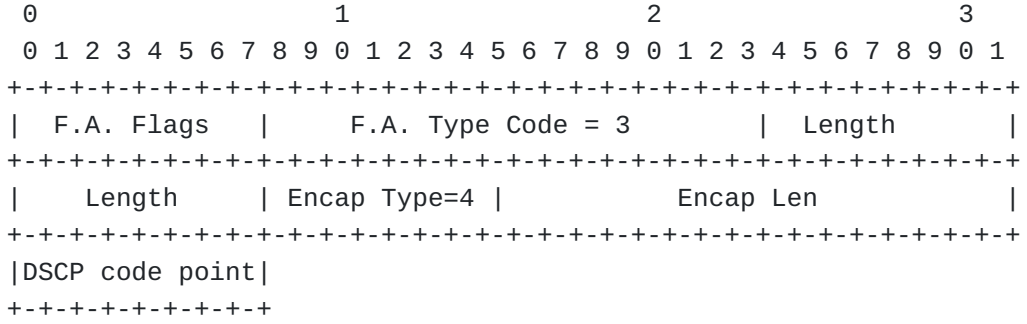


Figure 19: DSCP

- F.A. Flags (1 octet)

```
 0 1 2 3 4 5 6 7
+-+-+-+-+-+-+-+-+
|R R R R R R R R|
+-+-+-+-+-+-+-+-+
```

R: Reserved. MUST be set to zero, SHOULD be ignored by receiv

- Length (2 octets)
  Length in bytes of Value field.

- Encap Type
  = 4, to signify DSCP code point.

- Encap Len (2 octets)
  = 1, Length in bytes of following Encap Value field.

- DSCP code point:
  DS Field, as specified in RFC-2474 sec 3.

5.4.4. Endpoint Attributes

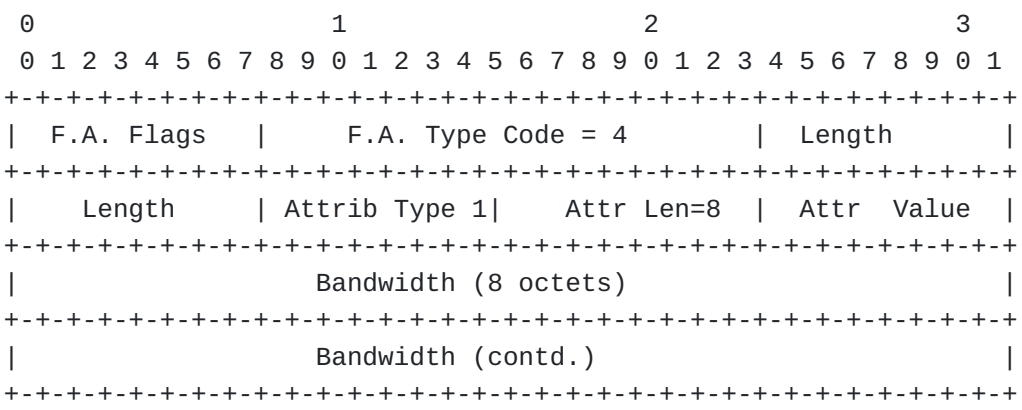
F.A. Type Code = 4. This Forwarding Argument TLV defines attributes of an endpoint.

```
 0                   1                   2                   3
 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|  F.A. Flags   |     F.A. Type Code = 4        |  Length       |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|    Length     | Attrib Type  |    Attr Len    |  Attr  Value  |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|                    Attr Value                                 |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
```

Figure 20: Endpoint attributes

EP Attrib Type	Attrib Value	Attrib Len (octets)
0	None	
1	Endpoint Bandwidth	8

5.4.4.1. Endpoint Bandwidth



- Len (2 octets)
  Length in bytes of remaining portion of SubTLV.

- Bandwidth
  The bandwidth to the endpoint expressed as 8 octets,
  units being bits per second.

Figure 21: Endpoint Bandwidth

This sub-TLV would be valid with Forwarding Instruction TLV with FwdAction of Forward, Swap or Push.

6. Error Handling

With MNH TLV Type = 4 (Label Descriptor), this attribute is used to describe the label advertised by the BGP-peer. If the value in the attribute is syntactically parse-able, but not semantically valid, the receiving speaker should deal with the error gracefully and MUST NOT tear down the BGP session. In such cases the rest of the BGP-update can be consumed if possibe.

With other MNH TLV Types, this attribute is used to specify the forwarding action at the receiving BGP-peer. If the value in the attribute is syntactically parse-able, but not semantically valid, the receiving speaker SHOULD deal with the error gracefully by ignoring the MNH attribute, and continue processing the route. It MUST NOT tear down the BGP session.

If a MNH TLV Type = 4 is received for an IP-route (SAFI Unicast), the MNH attribute SHOULD be ignored. Because IP route prefixes are upstream allocated by nature.

If a MNH TLV Type = 4 is received for an [MPLS-NAMESPACES] route, the MNH attribute SHOULD be ignored. Because the label prefix in MPLS-NAMESPACE family routes is upstream allocated.

The receiving BGP speaker MAY consider the "Num-Nexthops" value in a Nexthop Forwarding Information TLV not acceptable, based on it's

forwarding capabilities. In such cases, the MNH attribute SHOULD be considered Unusable, and not be used, ignored on receipt. The condition SHOULD be dealt gracefully and MUST NOT tear down the BGP session.

A TLV or sub-TLV of a certain Type in a MNH attribute can occur only once, unless specified otherwise by that type value. If multiple instances of such TLV or sub-TLV is received, the instances other than the first occurance are ignored.

If a TLV or sub-TLV of an unknown Type value is received, it is ignored and skipped. Remaining part of the MNH attribute if parseable is used

In case of length errors inside a TLV, such that the MNH attribute cannot be used, but the length value in MNH attribute itself is proper, the MNH attribute should be considered invalid and not used. But rest of the route update if parseable should be used. This follows the 'Attribute discard' approach described in [RFC7606] Section 2.

7. Scaling Considerations

The MNH attribute allows receiving multiple nexthops on the same BGP session. This flexibility also opens up the possibility that a peer can send large number of multipath (ECMP/UCMP/FRR) nexthops that may overwhelm the local system's forwarding plane. Prefix-limit based checks will not avoid this situation.

To keep the scaling limits under check, a BGP speaker MAY keep account of number of unique multipath nexthops that are received from a BGP peer, and impose a configurable max-limit on that. This is especially useful for EBGP peers.

A good scaling property of conveying multipath nexthops using the MNH attribute with N nexthop legs on one BGP session, as against BGP routes on N BGP sessions is that, it limits the amount of transitionary multipath combinatorial state in the latter model. Because the final multipath state is conveyed by one route update in deterministic manner, there is no transitionary multipath combinatorial explosion created during establishment of N sessions.

8. IANA Considerations

This document makes request to IANA to allocate the following codes in BGP attributes registry.

8.1. BGP Path Attributes

A new BGP attribute code TBD for "BGP MultiNexthop Attribute (MULTI_NEXT_HOP)", in "BGP Path Attributes" registry.

8.2. Capability Codes

This document makes request to IANA to allocate a BGP capability code TBD for "BGP MultiNexthop Attribute (MULTI_NEXT_HOP), in "Capability Codes" registry.

8.3. BGP MultiNextHop Attribute

This document requests IANA to create a new registry group for MultiNextHop attribute, and the following registries in it.

8.3.1. MultiNextHop (MNH) TLV Types

This is a Registry for Type codes in Section 5.1 "MULTI_NEXT_HOP TLV"

Under "Border Gateway Protocol (BGP) Parameters",

Registry Group: BGP MultiNextHop Attribute

Registry Name: MultiNexthop (MNH) TLV Types

MNH Type Code	Meaning
0	Reserved
1	Primary forwarding path
2	Backup forwarding path
3	Label Descriptor
4-254	Unassigned
255	Reserved

Reference: This document.

Registration Procedure(s)
Future assignments are to be made using either the Standards Ac process defined in [RFC2434], or the Early IANA Allocation proc defined in [RFC4020].

8.3.2. Forwarding Action Types

This is a Registry for Type codes in Section 5.3 "Forwarding Instruction TLV"

Under "Border Gateway Protocol (BGP) Parameters",

Registry Group: BGP MultiNextHop Attribute

Registry Name: Forwarding Action Types

FwdAction	Meaning
0	Reserved
1	Forward
2	Pop-And-Forward
3	Swap
4	Push
5	Pop-And-Lookup
6	Replicate
7-254	Unassigned
255	Reserved

Reference: This document.

Registration Procedure(s)
Future assignments are to be made using either the Standards
process defined in [RFC2434], or the Early IANA Allocation pr
defined in [RFC4020].

8.3.3. Forwarding Argument Types

This is a Registry for Type codes in Section 5.4 "Forwarding Arguments TLV"

Under "Border Gateway Protocol (BGP) Parameters",

Registry Group: BGP MultiNextHop Attribute

Registry Name: Forwarding Argument Types

F.A. Type Code	Meaning
0	Reserved
1	Endpoint Identifier
2	Path Constraints
3	Payload encapsulation info signaling
4	Endpoint attributes advertisement
5-65534	Unassigned
65535	Reserved

Reference: This document.

Registration Procedure(s)
Future assignments are to be made using either the Standards A process defined in [RFC2434], or the Early IANA Allocation pro defined in [RFC4020].

8.3.4. Endpoint Types

This is a Registry for Type codes in Section 5.4.1 "Endpoint Identifier" Forwarding Argument.

Under "Border Gateway Protocol (BGP) Parameters",

Registry Group: BGP MultiNextHop Attribute

Registry Name: Endpoint Types

Endpoint Type	Value
0	Reserved
1	IPv4 Address
2	IPv6 Address
3	MPLS Label
4	Fwd Context RD
5	Fwd Context RT
6-254	Unassigned
255	Reserved

Reference: This document.

Registration Procedure(s)
Future assignments are to be made using either the Standards A process defined in [RFC2434], or the Early IANA Allocation pro defined in [RFC4020].

8.3.5. Path Constrain Types

This is a Registry for Type codes in Section 5.4.2 "Path Constrain" Forwarding Argument.

Under "Border Gateway Protocol (BGP) Parameters",

Registry Group: BGP MultiNextHop Attribute

Registry Name: Path Constrain Types

ConstrainType	Value
0	Reserved
1	Proximity check
2	Transport Class ID (Color)
3	Load balance factor
4-254	Unassigned
255	Reserved

Reference: This document.

Registration Procedure(s)
Future assignments are to be made using either the Standards
process defined in [RFC2434], or the Early IANA Allocation pr
defined in [RFC4020].

8.3.6. Encapsulation Types

This is a Registry for Type codes in Section 5.4.3 "Payload Encapsulation Info" Forwarding Argument.

Under "Border Gateway Protocol (BGP) Parameters",

Registry Group: BGP MultiNextHop Attribute

Registry Name: Encapsulation Types

Encap Type	Value
0	Reserved
1	MPLS Label Info
2	SR MPLS label Index Info
3	SRv6 SID info
4	DSCP code point
5-254	Unassigned
255	Reserved

Reference: This document.

Registration Procedure(s)
Future assignments are to be made using either the Standards process defined in [RFC2434], or the Early IANA Allocation p defined in [RFC4020].

8.3.7. Endpoint Attribute Types

This is a Registry for Type codes in Section 5.4.4 "Endpoint attributes" Forwarding Argument.

Under "Border Gateway Protocol (BGP) Parameters",

Registry Group: BGP MultiNextHop Attribute

Registry Name: Endpoint Attribute Types

EP Attrib Type	Attrib Value
0	Reserved
1	Bandwidth
2-254	Unassigned
255	Reserved

Reference: This document.

Registration Procedure(s)
Future assignments are to be made using either the Standards process defined in [RFC2434], or the Early IANA Allocation pr defined in [RFC4020].

Note to RFC Editor: this section may be removed on publication as an RFC.

9. Security Considerations

The MNH attribute is defined as optional non-transitive BGP attribute, such that it does not accidentally get propagated or leaked via BGP speakers that dont support this feature, especially does not unintentionally leak across EBGP boundaries.

MNH may be used to advertise nexthop with MPLS label in various BGP families. In scenarios where MPLS is enabled on link to a device in an untrusted domain, e.g. a PE-CE link or ASBR-ASBR inter-AS link, security can be provided against MPLS label spoofing by using MPLS context tables as described in MPLS enabled CE (Appendix A.10.2). Such that only MPLS traffic with labels advertised to the BGP speaker are allowed to forward. However, the PE may not be able to perform any checks based on inner payload in the MPLS packet since it performs label swap forwarding. Such 'inner payload' based checks may be offloaded to a downstream node that forwards and processes inner payload, e.g., an IP router having full FIB. These security aspects should be considered when using MPLS enabled CE devices.

Contributors

Reshma Das
Juniper Networks, Inc.
1133 Innovation Way,
Sunnyvale, CA 94089
United States of America

Email: dreshma@juniper.net

Natrajan Venkataraman
Juniper Networks, Inc.
1133 Innovation Way,
Sunnyvale, CA 94089
United States of America

Email: natv@juniper.net

Acknowledgements

Thanks to Jeff Haas, Robert Raszuk, Ron Bonica for the review, discussions and input to the draft.

Thanks to Blaine Williams and Satya Mohanty for the discussions on some usecases.

References

Normative References

[RFC2545] Marques, P. and F. Dupont, "Use of BGP-4 Multiprotocol Extensions for IPv6 Inter-Domain Routing", RFC 2545, DOI 10.17487/RFC2545, March 1999, <https://www.rfc-editor.org/info/rfc2545>.

[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>.

[RFC7311] Mohapatra, P., Fernando, R., Rosen, E., and J. Uttaro, "The Accumulated IGP Metric Attribute for BGP", RFC 7311, DOI 10.17487/RFC7311, August 2014, <https://www.rfc-editor.org/info/rfc7311>.

[RFC7606] Chen, E., Ed., Scudder, J., Ed., Mohapatra, P., and K. Patel, "Revised Error Handling for BGP UPDATE Messages", RFC 7606, DOI 10.17487/RFC7606, August 2015, <https://www.rfc-editor.org/info/rfc7606>.

[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>.

[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>.

Informative References

[ADDPATH-GUIDELINES] Uttaro, Ed., "BGP Flow-Spec Redirect to IP Action", 25 April 2016, <https://datatracker.ietf.org/doc/html/draft-ietf-idr-add-paths-guidelines-08#section-2>.

[BGP-CT] Vairavakkalai, Ed. and Venkataraman, Ed., "BGP Classful Transport Planes", 17 March 2023, <https://datatracker.ietf.org/doc/html/draft-ietf-idr-bgp-ct-28>.

[FLWSPC-REDIR-IP] Simpson, Ed., "BGP Flow-Spec Redirect to IP Action", 2 February 2015, <https://datatracker.ietf.org/doc/html/draft-ietf-idr-flowspec-redirect-ip#section-3>.

[MPLS-NAMESPACES] Vairavakkalai, Ed., "BGP Signaled MPLS Namespaces", 10 July 2023, <https://datatracker.ietf.org/

doc/html/draft-kaliraj-bess-bgp-sig-private-mpls-labels-06>.

[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>.

[RFC2474] Nichols, K., Blake, S., Baker, F., and D. Black, "Definition of the Differentiated Services Field (DS Field) in the IPv4 and IPv6 Headers", RFC 2474, DOI 10.17487/RFC2474, December 1998, <https://www.rfc-editor.org/info/rfc2474>.

[SRTE-COLOR-ONLY] Filsfils, Ed., "BGP Flow-Spec Redirect to IP Action", 21 February 2018, <https://tools.ietf.org/html/draft-filsfils-spring-segment-routing-policy-06#section-8.8.1>.

Appendix A. Example of Usecases

This section describes various example usecases of the MNH attribute.

A.1. Signaling WECMP to Ingress Node

This section describes how MNH can be used to provide weighted equal cost multipath in a network fabric, while not increasing RIB scale.

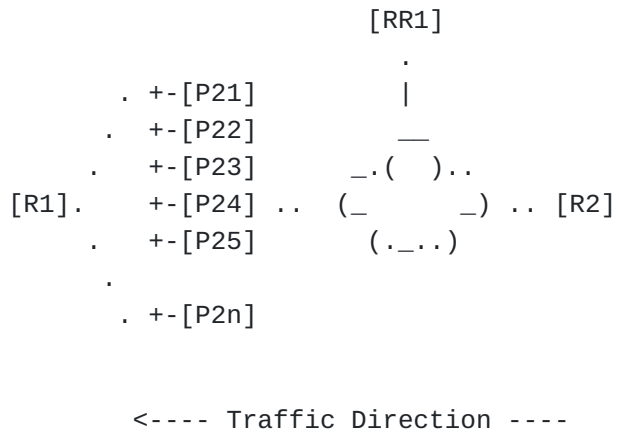


Figure 22: WECMP without increasing RIB scale

Figure 22 shows a network with BGP speaker R1 connected to a number of routers P21 .. P2n in its region. R1 is eSN and R2 is iSN for the IP traffic in consideration. BGP service families IPv4 Unicast (AFI/SAFI: 1/1) and IPv6 Unicast (AFI/SAFI: 2/1) are negotiated on the

BGP sessions between RR1 - R1 and RR1 - R2. RR1 reflects the BGP routes between R1 and R2 with next hop unchanged.

When MNH is not in use, R1 advertises "n" BGP Addpath routes for a service prefix Pfx1, each having a distinct next hop, P21 .. P2n, and desired Link Bandwidth Extended Community. These Addpath routes will be received by R2, which can do WECMP based on the Link Bandwidth Extended Communities attached on the routes. This model increases RIB scale by "n" times, so that WECMP can be achieved.

When MNH is used in this network, R1 advertises a single BGP route for prefix Pfx1, which contains a MNH attribute with "n" next hops, each carrying the desired link bandwidth using Section 5.4.2.3 or Section 5.4.4.1

This allows achieving WECMP in the network without increasing RIB scale.

A.2. Signaling Optimal Forwarding Exitpoints to Ingress Node

In a BGP free core, one can dynamically signal to the ingress-node, how traffic should be load-balanced towards a set of exit nodes, in one BGP-route containing this attribute.

Example, for prefix1, perform equal load balancing towards exit nodes A, B; where as for prefix2, perform weighted load balancing (40%, 30%, 30%) towards exit nodes A, B, C.

Example, for prefix1, use PE1 as primary-nexthop and use PE2 as a backup-nexthop.

A.3. Load balancing to multiple CEs in a VRF

This section describes how MNH can be used to provide load balancing and entropy in a provider network for traffic destined to multiple CEs in a VRF, without increasing RIB scale.

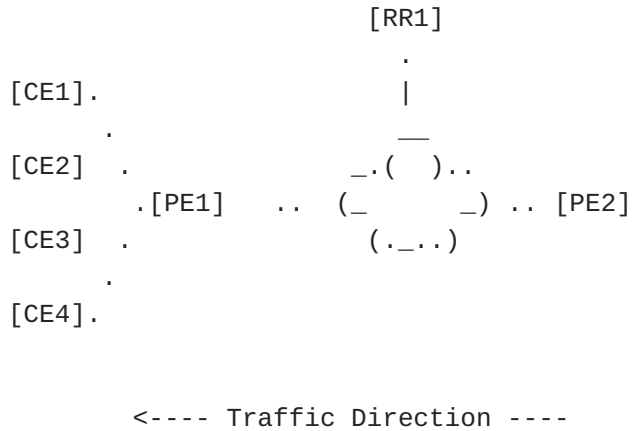


Figure 23: Load balancing to multiple CEs in a VRF

Figure 23 shows a L3VPN network with multiple CE devices connected to the same VRF at PE1. The VRF is configured with a RD: RD1, and uses "per next hop" label allocation mode to advertise the CE routes to L3VPN core. PE1 is eSN and PE2 is iSN for the IP traffic in consideration. CE1..CE4 advertise route for same prefix Pfx1 in BGP service families IPv4 Unicast (AFI/SAFI: 1/1) negotiated on the BGP sessions between the CEs and PE1. BGP L3VPN address family (AFI/SAFI: 1/128) is negotiated between PE1 - RR1, and RR1 - PE2. RR1 reflects the BGP routes between PE1 and PE2 with next hop unchanged.

PE1 would typically advertise to RR1 only the best path for prefix Pfx1 out of routes received from CE1..CE4. Using per CE RD or Addpath for L3VPN family may allow PE1 to advertise all CE routes to the RR, with an increase in RIB scale. This model increases RIB scale by "n" times, where 'n' is the number of CEs.

When MNH is used in this network, PE1 advertises a single BGP L3VPN route for prefix Pfx1, which contains a MNH attribute with "n" next hops, each carrying the label pointing towards a particular CE, using Section 5.4.3 along with the Section 5.4.1

This allows the network to direct traffic to a specific CE, and better loadbalance traffic in the provider network, with entropy provided by the per CE VPN labels, without increasing RIB scale.

A.4. Choosing a Received Label Based on it's Forwarding Semantic at Advertising Node

In Downstream label allocation case, the MNH plays role of "Label descriptor" and describes the forwarding treatment given to the label at the advertising speaker. The receiving speaker can benefit from this information as in the following examples:

- For a Prefix, a label with FRR enabled nexthop-set can be preferred to another label with a nexthop-set that doesn't provide FRR.

- For a Prefix, a label pointing to 10g nexthop can be preferred to another label pointing to a 1g nexthop

- Set of labels advertised can be aggregated, if they have same forwarding semantics (e.g. VPN per-prefix-label case)

A.5. Signaling Desired Forwarding Behavior for MPLS Upstream labels at Receiving Node

In Upstream label allocation case, the receiving speaker's forwarding-state can be controlled by the advertising speaker, thus enabling a standardized API to program desired MPLS forwarding-state at the receiving node. This is described in the [MPLS-NAMESPACES]

A.6. Load Balancing over EBGP Parallel Links

Consider N parallel links between two EBGP speakers. There are different models possible to do load balancing over these links:

N single-hop EBGP sessions over the N links. Interface addresses are used as next-hops. N copies of the RIB are exchanged to form N-way ECMP paths. The routes advertised on the N sessions can be attached with Link bandwidth comunity to perform weighted ECMP.

1 multi-hop EBGP session between loopback addresses, reachable via static route over the N links. Loopback addresses are used as next-hops. 1 copy of the RIB is exchanged with loopback address as nexthop. And a static route can be configured to the loopback address to perform desired N-way ECMP path. M loopbacks are configured in this model, to achieve M different load balancing schemes: ECMP, weighted ECMP, Fast-reroute enabled paths etc.

1 multi-hop EBGP session between loopback addresses, reachable via static route over the N links. Interface addresses are used as next-hops, without using additional loopbacks. 1 copy of the RIB is exchanged with MNH attribute to form N-way ECMP paths, weighted ECMP, Fast-reroute backup paths etc. BFD may be used to these directly connected BGP nexthops to detect liveness.

A.7. Flowspec Routes with Multiple "Redirect IP" next hops

There are existing protocol machinery which can benefit from the ability of MNH to clearly specify fallback behavior when multiple nexthops are involved. One example is the scenario described in [FLWSPC-REDIR-IP] where multiple Redirect-to-IP nexthop addresses exist for a Flowspec prefix. In such a scenario, the receiving

speakers may redirect the traffic to different nexthops, based on variables like IGP-cost. If instead, the MNH was used to specify the redirect-to-IP nexthop, then the order of preference between the different nexthops can be clearly specified using one flowspec route carrying a MNH containing those different nexthop-addresses specifying the desired preference-order. Such that, irrespective of IGP-cost, the receiving speakers will redirect the flow towards the same traffic collector device.

A.8. Color-Only Resolution next hop

Another existing protocol machinery that manufactures nexthop addresses from overloaded extended color community is specified in [SRTE-COLOR-ONLY]. In a way, the color field is overloaded to carry one anycast BGP next-hop with pre-specified fallback options. This approach gives us only two next-hops to play with. The 'BGP nexthop address' and the 'Color-only nexthop'

Instead, the MNH could be used to achieve the same result with more flexibility. Multiple BGP nexthops can be carried, each resolving over a desired Transport class (Color), and with customizable fallback order. And the solution will work for non-SRTE networks as-well.

A.9. Avoid Label Advertisement Oscillation Between Multihomed PEs.

In a MPLS network, a router may be multihomed to two PEs. The PEs may re-advertise routes received from the router to the IBGP core with self as nexthop and a "per nexthop" label. The PEs may also protect failure of primary path to the router by using the IBGP path via the other multihomed PE as a backup path.

In this scenario, label allocation oscillation may occur when one PE advertises a new label to the other PE. Reception of a new label results in change of nexthop, as the label is used as back nexthop leg, and per-nexthop label allocation is in use. Thus a new label is allocated and advertised. And when this new label is received by the first PE, it allocates a new label in turn. This process repeats.

This oscillation can be stopped only if the primary path label allocated by a PE does not depend on the primary path label advertised by other PE. A PE needs to be able to advertise multiple labels, one for use as primary path and another to be used as bacakup path by the receiver.

MNH attribute allows to advertise a Backup forwarding path label using Section 5.1.2 in addition to Primary forwarding path label using Section 5.1.1

A.10. Signaling Intent over PE-CE Attachment Circuit

BGP CT specifies procedures for Intent Driven Service Mapping in a service provider network, and defines 'Transport Class' construct to represent an Intent.

It may be desirable to allow a CE device to indicate in the data packet it sends what treatment it desires (the Intent) when the packet is forwarded within the provider network.

This section describes the mechanisms that enable such signaling. These procedures use existing AFIs 1 or 2, and service families (SAFI 1) on the PE-CE attachment circuit, with a new BGP attribute.

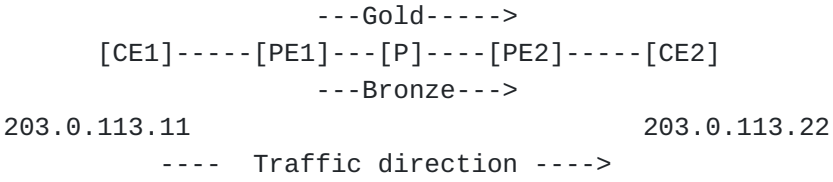


Figure 24: Example Topology with PE-CE Links

A.10.1. Using DSCP in MultiNexthop Attribute

Such an indication can be in form of DSCP code point ([RFC2474]) in the IP header.

In RFC2474, a Forwarding Class Selector maps to a PHB (Per-hop Behavior). The Transport Class construct is a PHB at transport layer.

Let PE1 be configured to map DSCP1 to Gold Transport class, and DSCP2 to Bronze Transport class. Based on the DSCP code point received on the IP traffic from CE1, PE1 forwards the IP packet over a Gold or Bronze tunnel. Thus, the forwarding is not based on just the destination IP address, but also the DSCP code point. This is known as Class Based Forwarding (CBF). Today CBF is configured at the PE1 device roles and CE1 doesn't receive any indication in BGP signaling regarding what DSCP code points are being offered by the provider network.

With a BGP MultiNexthop Attribute attached to a AFI/SAFI 1/1 service route, it is possible to extend the PE-CE BGP signaling (if used) to communicate such information to the CE1. In the preceding example, the MNH contains two Next hop Legs, described by two Forwarding Instruction TLVs. Each Next hop Leg contains PE1's peering self address in Endpoint Identifier TLV (Section 5.4.1), the color Gold or Bronze encoded in the Transport class ID TLV (Section 5.4.2.2, Figure 13), and associated DSCP code point indicating Gold or Bronze

transport class encoded in the Payload Encapsulation Info TLV (Section 5.4.3.4, Section 5.4.3). This allows the CE to discover what transport classes exist in the provider network, and which DSCP codepoint to encode so that traffic is forwarded using the desired transport class in the provided network.

A.10.2. MPLS-enabled CE

If the PE-CE link is MPLS enabled, a distinct MPLS label can also be used to express Intent in data packets from CE. Enabling MPLS forwarding on PE-CE links comes with some security implications. This section gives details on these aspects.

Consider the ingress PE1 receiving a VPN prefix RD:Pfx1 received with VPN label VL1, next hop as PE2 and a mapping community containing TC1 as 'Transport class ID'. PE1 can allocate a MPLS Label PVL1 for the tuple "VPN Label, PNH Address, Transport class ID" and advertise to CE1.

Label PVL1 may identifies a service function at any node in the network, e.g. a Firewall device or egress node PE2. And, for the same service prefix, a distinct label may be advertised to different CEs, such that incoming traffic from different CEs to the same service prefix can be diverted to a distinct devices in the network for further processing. This provides Ingress Peer Engineering control to the network.

PE1 installs a MPLS FIB route for PVL1 with next hop as "Swap VL1, Push TL1 towards PE2". TL1 is the BGP CT label received for the tuple 'PE2, TC1'. In forwarding, when MPLS packet with label PVL1 is received from CE1, PVL1 Swaps to label VL1 and pushes the BGP CT label TL1. PE1 advertises the label "PVL1" in the MNH to CE1. PE1 forwards based on MPLS label without performing any IP lookup. This allows for PE1 to be a low IP FIB device and still support CBF by using MPLS Label inferred PHB. The number of MPLS Labels consumed at PE1 for this approach will be proportional to the number of Service functions and Intents that are exposed to CE1.

A BGP MultiNexthop Attribute is attached to a AFI/SAFI 1/1 service route to convey the MPLS Label information to CE1. In the preceding example, the MNH contains two Next hop Legs, described by two Forwarding Instruction TLVs. Each Next hop Leg contains PE1's peering self address in Endpoint Identifier TLV (Section 5.4.1), the color Gold or Bronze encoded in the Transport class ID TLV (Figure 13), and associated MPLS Label "PVL1" or "PVL2" encoded in the Payload Encapsulation Info TLV (Section 5.4.3.1, Section 5.4.3). This allows the CE to discover what transport classes exist in the provider network, and which MPLS Label to encode so that traffic is forwarded using the desired transport class.

A.10.2.1. Secure MPLS Forwarding on Inter-AS Link

The MPLS enabled PE-CE attachment circuit is considered connecting to an untrusted domain. Such interfaces can be secured against MPLS label spoofing by a walled garden approach using "MPLS context tables".

The PE1-CE1 interface can be confined to a specific MPLS context table "A" corresponding to the BGP peer. Such that only the routes for labels advertised to CE1 are installed in MPLS context table "A".

This ensures that if CE1 sends MPLS packet with a label that was not advertised to the CE1, the packet will be dropped.

Furthermore, the routes for labels PVL1, PVL2 installed in MPLS context table "A" can match on 'Bottom of stack' bit being 'one', ensuring a MPLS packet is accepted from CE1 only if it has no more than one label in the label stack.

However, the PE itself may not be able to perform any checks based on inner payload in the MPLS packet since it performs label swap forwarding. Such inner payload based checks may be offloaded to a downstream node that forwards and processes inner payload, e.g. a IP FIB router. These security aspects should be considered when using MPLS enabled CE devices.

A.11. 4PE - Signal MPLS Label for IPv4 Unicast routes

This section describes how MNH can be used to signal MPLS explicit null label in AFI/SAFI: 1/1 routes in a pure IPv6 core environment, to achieve 4PE.

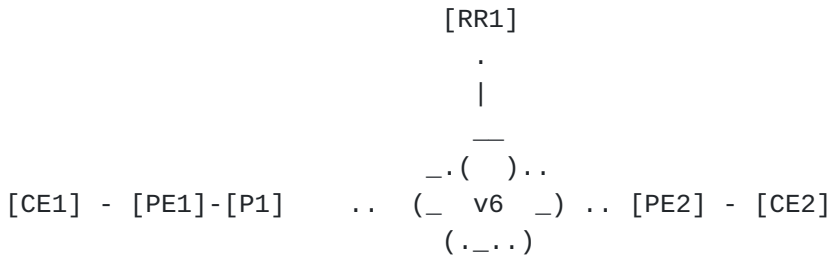


Figure 25: 4PE Network with Pure IPv6 Core

Figure 25 shows a 4PE network with pure IPv6 core, PE1 is the egress PE connected to penultimate hop node P1. PE1 to PE2 have some IPv6 core tunneling protocol like LDPv6. When PE1 has advertised Implicit Null label in LDPv6, some implementations of P1 may not be able to forward the inner IPv4 payload to PE1.

To solve this problem, PE1 needs to signal IPv4 Explicit NULL Label (Special Label 0) to PE2. PE2 will push this IPv4 Explicit NULL Label received in the MNH on the AFI/SAFI:1/1 route. Such that P1 does a MPLS Label swap operation and does not need to look into inner payload.

MNH can be used by PE1 on a AFI/SAFI: 1/1 route, to advertise the IPv4 Explicit Null label for the IPv4 Unicast service route. MPLS Label is encoded in the Payload Encapsulation Info TLV (Section 5.4.3.1, Section 5.4.3).

This allows the network to provide clear separation of service and transport routes, and not overloading AFI/SAFI: 1/4 to carry the IPv4 service routes. Not mixing service and transport routes improves security and manageability aspects of the network.

An egress PE may not need to advertise IPv4 Explicit Null label for the IPv4 service route, if it does UHP label in LDPv6. This model using MNH provides a homogenous service layer (AFI/SAFI: 1/1) that accomodates differences in requirement of different PE and P routers. Only the PEs which are connected to P nodes that cannot handle the PHP situation need to advertise Label using MNH. The service layer is kept consistent in the network, and can seamlessly extend to multiple domains without needing redistribution between AFI/SAFIs.

Not mixing service and transport routes improves security and manageability aspects of the network.

Authors' Addresses

Kaliraj Vairavakkalai (editor)
Juniper Networks, Inc.
1133 Innovation Way,
Sunnyvale, CA 94089
United States of America

Email: kaliraj@juniper.net

Minto Jeyananth
Juniper Networks, Inc.
1133 Innovation Way,
Sunnyvale, CA 94089
United States of America

Email: minto@juniper.net

Mohan Nanduri
Microsoft
1 Microsoft Way,
Redmond, WA 98052
United States of America

Email: mohannanduri@microsoft.com

Avinash Reddy
AT&T
3400 W Plano Pkwy,
Plano, TX 75075
United States of America

Email: ar977m@att.com