

Workgroup: Network Working Group
Internet-Draft:
draft-kaliraj-idr-multinexthop-attribute-02
Published: 28 December 2021
Intended Status: Standards Track
Expires: 1 July 2022
Authors: K. Vairavakkalai M. Jeyananth
 Juniper Networks, Inc. Juniper Networks, Inc.
 G. Mishra
 Verizon Communications Inc.

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 BGP-Nexthop attribute (code 3), or inside the MP_REACH attribute (code 14).

For cases where multiple nexthops need to be advertised, BGP-Addpath is used. Though Addpath allows basic ability to advertise multiple-nexthops, it does not allow the sender to specify desired relationship between the multiple nexthops being advertised e.g., relative-preference, type of load-balancing. These are local decisions at the receiving speaker 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, in one BGP route, to ingress nodes. This document defines a new BGP attribute "MultiNexthop (MNH)" that can be used for this purpose.

This attribute can be used for both labeled and unlabeled BGP families. The MNH can be used to advertise MPLS label along with nexthop for unlabeled families (e.g. Inet Unicast, Inet6 Unicast). Such that, mechanisms at the transport layer can work uniformly on labeled and unlabeled BGP families. Service route scale can be confined closer to the service edge nodes, making the transport layer nodes light and nimble. They don't have any service route state, only have service end-point state.

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).

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 1 July 2022.

Copyright Notice

Copyright (c) 2021 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](#)
2. [Use-cases examples](#)
 - 2.1. [Optimal forwarding exit-points signaling to ingress-node](#)
 - 2.2. [Choosing a received label based on it's forwarding-semantic at advertising node](#)
 - 2.3. [Signaling desired forwarding behavior when installing MPLS Upstream labels at receiving node](#)
 - 2.4. [Load-balancing over EGBP parallel links](#)
 - 2.5. [Flowspec routes with multiple Redirect-IP nexthops](#)
 - 2.6. [Color-Only resolution nexthop](#)

- 3. [The "MultiNexthop \(MNH\)" BGP attribute encoding](#)
 - 3.1. [Operations](#)
 - 3.1.1. [BGP Capability for MNH attribute](#)
 - 3.1.2. [Scope of use, and propagation](#)
 - 3.1.3. [Interaction of MNH with Nexthop \(in attr-code 3, 14\)](#)
 - 3.1.4. [Interaction with Addpath](#)
 - 3.1.5. [Path-selection considerations](#)
 - 3.1.6. [NH-Flags U bit, denoting upstream/downstream semantics](#)
 - 3.2. [Nexthop Forwarding Semantics TLV](#)
 - 3.3. [Nexthop-Leg Descriptor TLV](#)
 - 3.4. [Nexthop Attributes Sub-TLV](#)
 - 3.4.1. [IP Address](#)
 - 3.4.2. [Labeled IP nexthop](#)
 - 3.4.3. [Transport Class ID \(Color\)](#)
 - 3.4.4. [Available Bandwidth](#)
 - 3.4.5. [Load balance factor](#)
 - 3.4.6. [Forwarding-context name](#)
 - 3.4.7. [Forwarding-context Route-Target](#)
- 4. [Error handling procedures](#)
- 5. [Scaling considerations](#)
- 6. [IANA Considerations](#)
- 7. [Security Considerations](#)
- 8. [Acknowledgements](#)
- 9. [References](#)
 - 9.1. [Normative References](#)
 - 9.2. [References](#)
- [Authors' Addresses](#)

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 attribute (code 14).

For cases where multiple nexthops need to be advertised, BGP-Addpath is used. Though Addpath allows basic ability to advertise multiple-nexthops, it does not allow the sender to specify desired relationship between the multiple nexthops being advertised e.g., relative-ordering, type of load-balancing, fast-reroute. These are local decision 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. This document defines a new BGP attribute "MultiNexthop (MNH)" that can be used for this purpose.

This attribute can be used for both labeled and unlabeled BGP families. The MNH can be used to advertise MPLS label along with nexthop for unlabeled families (e.g. Inet Unicast, Inet6 Unicast). Such that, mechanisms at the transport layer can work uniformly on labeled and unlabeled BGP families. Service route scale can be confined closer to the service edge nodes, making the transport layer nodes light and nimble. They don't have any service route state, only have service end-point state.

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).

A new BGP capability ([[RFC3392](#)]) called "MultiNexthop (MNH)" is defined with type code: IANA TBD. This capability is used to express the ability to send and receive MNH attribute.

2. Use-cases examples

2.1. Optimal forwarding exit-points signaling 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 cost load-balancing towards exit-nodes A, B; where-as for prefix2, perform unequal-cost 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.

2.2. 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)

2.3. Signaling desired forwarding behavior when installing 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](#)]

2.4. 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 community 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.

2.5. Flowspec routes with multiple Redirect-IP nexthops

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 [[FLWSPEC-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.

2.6. Color-Only resolution nexthop

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.

3. The "MultiNexthop (MNH)" BGP attribute encoding

"MultiNexthop (MNH)" is a new BGP optional non-transitive attribute (code TBD), that can be used to convey multiple-nexthops to a BGP-speaker. This attribute describes forwarding semantics using one or more Nexthop-Forwarding-Semantics TLV.

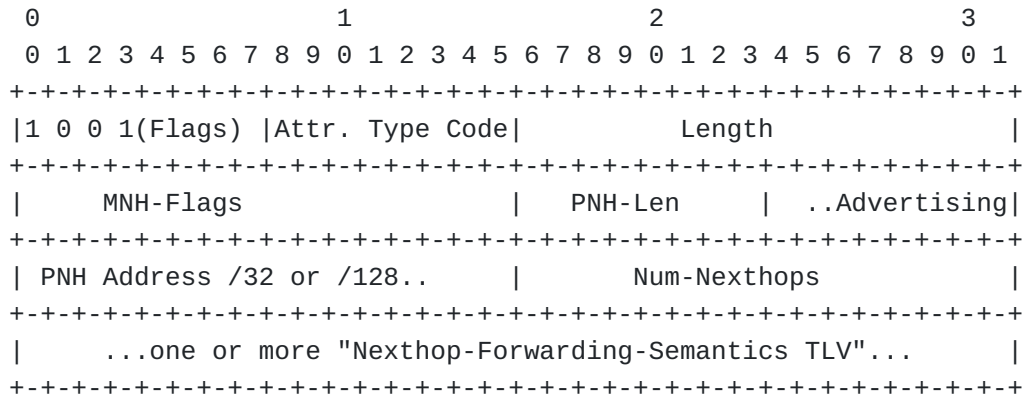


Fig 1: MultiNexthop - BGP Attribute

- Flags
 - BGP Path-attribute flags. 1001 to indicate Optional Non-Transitive, Extended-length field.
- Attr. Type Code
 - IANA TBD.
- Length
 - Two bytes field stating length of attribute value in bytes.
- MNH-Flags
 - 16 bit flag (UR..R)
 - Only one bit MSB is defined currently, others are reserved.
 - R: Reserved
 - U: 1 means the Upstream-allocation, attribute describes forwarding state desired at receiving speaker.
 - 0 means the Downstream-allocation, attribute describes forwarding state present at advertising-speaker.
- PNH-Len
 - Protocol-NH Length in bits (= 32 or 128) Advertising PNH IPv4 or
- PNH-address
 - BGP Protocol Nexthop address (Len = 32 or 128) advertised in NEXT_MP_REACH_NLRI attr. Used to sanity-check this attribute.
- Num-Nexthops
 - Number of nexthop addresses carried in the MNH.
 - >1 if ECMP or Alternate-paths.

Sec 3.2 describes the Nexthop-Forwarding-Semantics TLV.

3.1. Operations

3.1.1. BGP Capability for MNH attribute

A new BGP capability [[RFC3392](#)] called "MultiNexthop (MNH)" is defined with type code: IANA TBD. The MNH attribute MUST NOT be sent to a BGP speaker that has not advertise the MNH capability. A BGP speaker MUST ignore the MNH attribute received from a peer which has not advertised the MNH attribute.

3.1.2. 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.

Also, it is required to avoid un-intentionally leaking it to other AS on an EBGP session, via a BGP speaker that does not understand MNH attribute.

To achieve this, the attribute is defined as "optional non-transitive", and uses a new BGP capability. If a MNH-attribute is received by a PE BGP-speaker that does not understand it, the optional non-transitive nature avoids unintentionally propagating it towards EBGPeers.

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. When a RR receives the MNH-attribute from a client that supports the attribute, it propagates the attribute as-is when reflecting the route with nexthop unchanged.

When a BGP speaker receives the MNH-attribute from another speaker that did not advertise support of the attribute, the attribute is ignored.

The MNH attribute capability provides additional protection against receiving this attribute from EBGPeers, when not intended.

3.1.3. Interaction of MNH with Nexthop (in attr-code 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 NEXT_HOP attribute or MP_REACH_NLRI attribute. Any speaker that recognizes this attribute and changes the PNH while re-advertising the route **MUST** remove the MultiNexthop-Attribute in the re-advertisement. The speaker **MAY** however add a new MultiNexthop-Attribute to the re-advertisement; while doing so the speaker **MUST** record in the "Advertising-PNH" field the same next-hop address as used in NEXT_HOP field or MP_REACH_NLRI 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 next-hop address contained in NEXT_HOP field or MP_REACH_NLRI field.

3.1.4. 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".

3.1.5. Path-selection considerations

While tie breaking in the path-selection as described in RFC-4271, 9.1.2.2. step (e) viz. the "IGP cost to nexthop", consider the highest cost among the nexthop-legs present in this attribute.

3.1.6. NH-Flags U bit, denoting upstream/downstream semantics

U-bit being Set indicates that this attribute describes what the forwarding semantics of an Upstream-allocated label at the receiving-speaker should be. All other bits in NH-Flags are currently reserved, MUST be set to 0 by sender and MUST be ignored by receiver.

This attribute can be used for both labeled and unlabeled BGP families.

A MultiNexthop attribute with U=0 is called "Label Descriptor" role. A BGP speaker advertising a downstream-allocated label-route MAY add this attribute to the BGP route Update, to "describe" to the receiving speaker what the label's forwarding semantics at the sending speaker is.

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 (U-bit Set) 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, U-bit is Set. Since IP prefixes are by nature upstream allocated.

3.2. Nexthop Forwarding Semantics TLV

Each Forwarding-Semantics TLV expresses a nexthop leg's forwarding action. i.e. a "FwdAction" with an associated Nexthop. The type of

actions defined by this TLV are given below. The "Nexthop-Leg" field takes appropriate values based on the FwdAction.

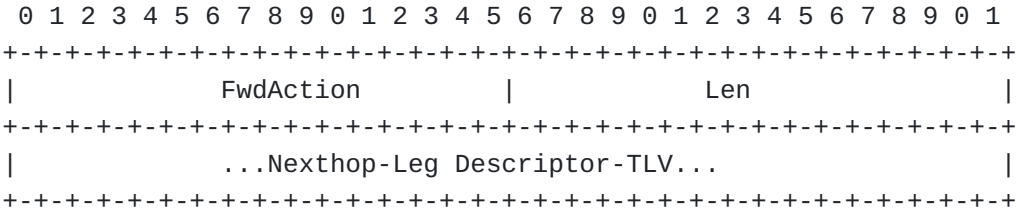


Fig 2: Nexthop Forwarding Semantics TLV

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

- Len
Length of Nexthop Forwarding Semantics TLV including all Nexthop-Leg Descriptor 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 MPLS routes only. FwdActions 1 and 6 are applicable for Flowspec routes for Redirect and Mirror actions.

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 MPLS-label and forward the payload towards the Nexthop IP-address specified in the sub-TLV, using appropriate encapsulation to reach the Nexthop.

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 section 5.

3.3. Nexthop-Leg Descriptor TLV

The Nexthop-Leg Descriptor TLV describes various attributes of the Nexthop-legs that the FwdAction is associated with.

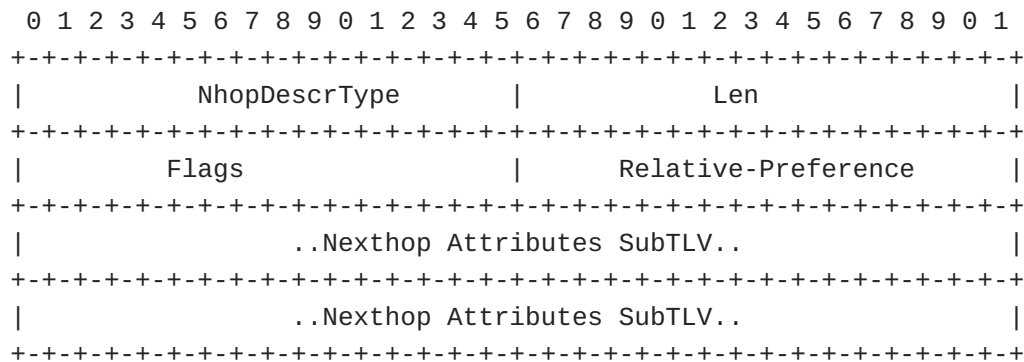


Fig 3: Nexthop-Leg Descriptor TLV

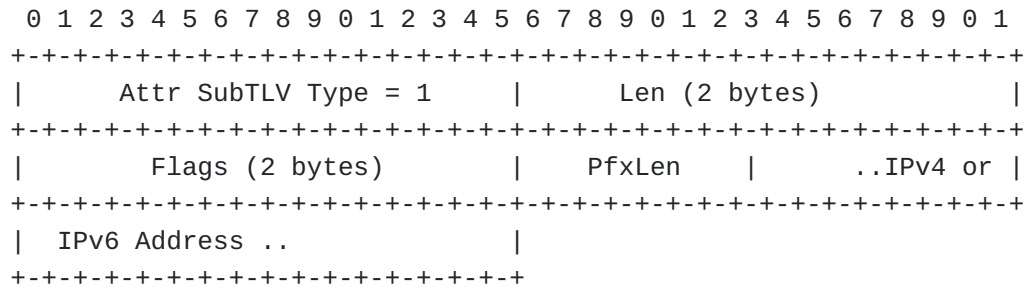
NhopDescrType	Meaning
-----	-----
1	IPv4-nexthop
2	IPv6-nexthop
3	Labeled-IP-Nexthop
4	Forwarding-Context-Nexthop

- Len (2 octets)
Length in bytes of Nexthop-Leg Descriptor TLV, including Flags, Relative-Preference, and Nexthop Attributes SubTLVs.
- Flags
2 octets. Must send zero. Must ignore on receive.
- Relative-Preference
Unsigned 2 octet integer specifying relative order or preference, to use in FIB. Use in FIB all usable legs with lowest relative-weight. If multiple legs exist with that weight, form ECMP

3.4. Nexthop Attributes Sub-TLV

SubTLV type	Meaning
-----	-----
1	IP-Address
2	Labeled-IP-Nexthop
3	Transport Class ID (Color)
4	Bandwidth
5	Load-Balance-Factor
6	Forwarding-context Name
7	Forwarding-context Route-Target

3.4.1. IP Address

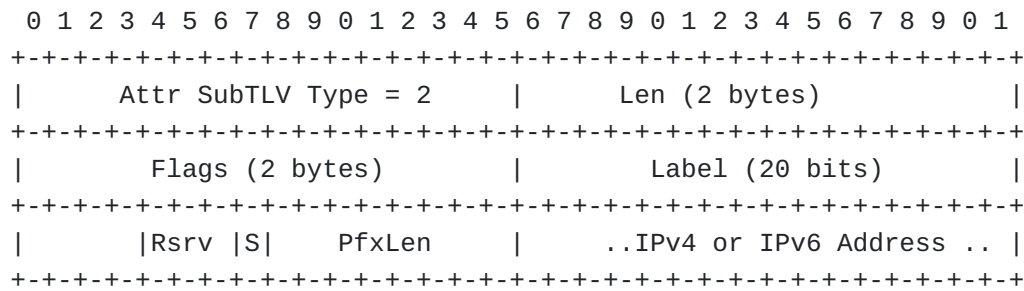


- Len (2 octets)
Length in bytes of remaining portion of SubTLV.
- Flags
2 octets. Must send zero. Must ignore on receive.
- PfxLen (1 octet)
Length in bits of Nexthop IP-address (32 or 128)
- IPv4 or IPv6 Address
Remaining bytes in sub-TLV are the 32 bit or 128 bit Nexthop address

Fig 4: IP-Address attribute sub-TLV

This sub-TLV would be valid with Nexthop-Forwarding-Semantics TLV with FwdAction of Pop-And-Forward or Forward.

3.4.2. Labeled IP nexthop



- Len (2 octets)
Length in bytes of remaining portion of SubTLV.
- Flags (2 octets):
ELC (MSB bit): indicates if this egress NH is Entropy Label Capable
Remaining bits are Reserved. Must send zero. Must ignore on receive
- Label:
The Label field is a 20-bit field containing an MPLS label value
(see [RFC3032]).
- Rsrv:
This 3-bit field SHOULD be set to zero on transmission and
MUST be ignored on reception.
- S:
This 1-bit field MUST be set to one on last label being pushed.
- PfxLen (1 octet)
Length in bits of Nexthop IP-address (32 or 128)
- IPv4 or IPv6 Address
Remaining bytes in sub-TLV are the 32 bit or 128 bit Nexthop address

Fig 5: "Labeled nexthop" attribute sub-TLV

This sub-TLV would be valid with Nexthop-Leg Forwarding-Semantics
TLV with FwdAction of Swap or Push.

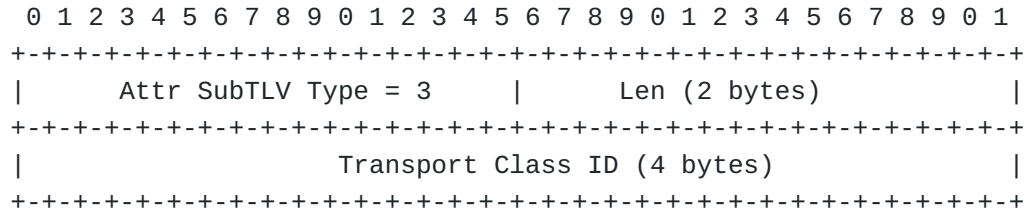
3.4.3. Transport Class ID (Color)

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 attribute
(code 14). With MNH, the Transport Class can be specified per
Nexthop-Leg TLV. It is applied to the IP-address encoded in the

Nexthop Attribute Sub-TLVs of type "IP Address", "Labeled IP nexthop".

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

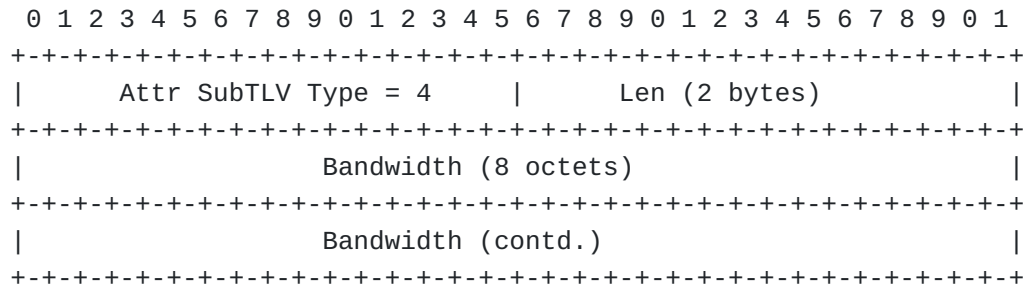


- Len (2 octets)
Length in bytes of remaining portion of SubTLV.
- Transport Class ID (Color):
This is a 32 bit identifier, associated with the Nexthop address.
The Nexthop specified in "IP-address or Labeled Nexthop" TLVs are resolved over tunnels of this color.
Defined in [BGP-CT] [draft-kaliraj-idr-bgp-classful-transport-planes]

Fig 6: "Transport Class ID (Color)" attribute sub-TLV

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

3.4.4. Available Bandwidth



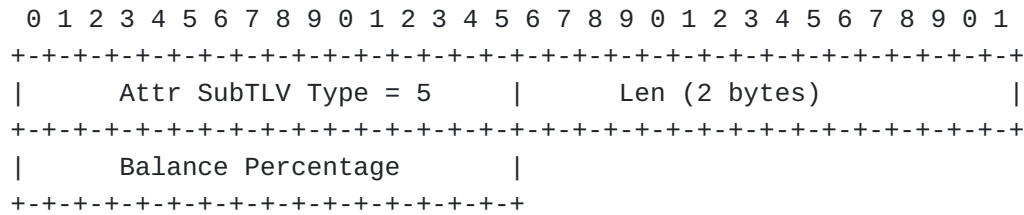
- Len (2 octets)
Length in bytes of remaining portion of SubTLV.
- Bandwidth
The bandwidth of the link expressed as 8 octets, units being bits per second.

Fig 6: "Bandwidth" attribute sub-TLV

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

This sub-TLV would also be valid in a Label-Descriptor-attribute whose U-bit is reset.

3.4.5. Load balance factor



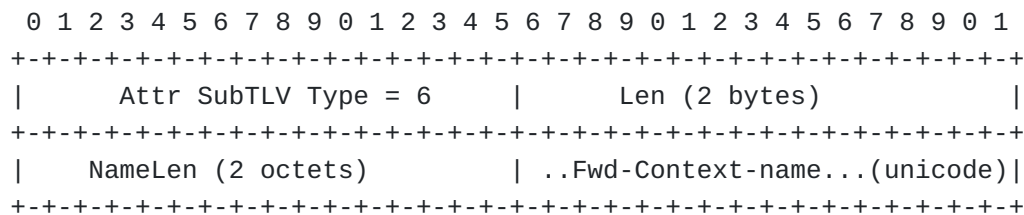
- Len (2 octets)
Length in bytes of remaining portion of SubTLV.
- Balance Percentage:
This is the explicit "balance percentage" requested by the sender, for unequal load-balancing over these Nexthop-Descriptor-TLV legs. This balance percentage would override the implicit balance-percentage calculated using "Bandwidth" attribute sub-TLV.

Fig 7: "Load-Balance-Factor" attribute sub-TLV

This sub-TLV would be valid with Nexthop-Forwarding-Semantics 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

3.4.6. Forwarding-context name



- Len (2 octets)
Length in bytes of remaining portion of SubTLV.
- NameLen (2 octets)
Length in bytes of Fwd-Context-Name
- Forwarding Context Name:
Name of forwarding context (e.g. VRF-name) where lookup should happen

Fig 8: Forwarding-Context name attribute sub-TLV

This sub-TLV would be valid with Nexthop-Forwarding-Semantics TLV with FwdAction of Pop-And-Lookup. Ref: usecase 2.3. The Forwarding-context-name identifies the forwarding-context (for e.g. the VRF-name) where the lookup should happen after pop label.

3.4.7. Forwarding-context Route-Target

```

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
+--+--+--+--+--+--+--+--+--+--+--+--+--+--+--+--+--+--+--+--+--+
|      Attr SubTLV Type = 7      |      Len (2 bytes)      |
+--+--+--+--+--+--+--+--+--+--+--+--+--+--+--+--+--+--+--+--+--+
|      Type (2 octets)            |  ...Route Target... (8 octets)|
+--+--+--+--+--+--+--+--+--+--+--+--+--+--+--+--+--+--+--+--+--+
|      ..Route Target... (continued)      |
+--+--+--+--+--+--+--+--+--+--+--+--+--+--+--+--+--+--+--+--+--+
|  ...Route Target... (8 octets)  |
+--+--+--+--+--+--+--+--+--+--+--+--+--+--+--+--+--+--+--+--+--+

```

- Len (2 octets)
Length in bytes of remaining portion of SubTLV.
- Type:
value of 1 indicates Route Target follows.
- Route Target:
Import Route Target of the forwarding context
(e.g. VRF-name) where lookup should happen.

Fig 9: "Route-Target identifying the Forwarding-Context" attribute sub-TLV

This sub-TLV would be valid with Nexthop-Forwarding-Semantics TLV with FwdAction of Pop-And-Lookup. Ref: usecase 2.3. The Route Target identifies the forwarding-context (for e.g. VRF) where the lookup should happen after pop label.

If any of these sub-TLVs or FwdAction combinations are unrecognized or unsupported by a receiving speaker, it is considered a semantic error for that speaker, and in such case error-handling procedures described in section 4 should be followed.

4. Error handling procedures

When U-bit is Reset, 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 possible.

When U-bit is Set, 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 with U-bit Reset 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 with U-bit Reset 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-Nexthop" value in a MNH attribute (U-bit Set) 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.

5. 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.

6. IANA Considerations

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

1. MultiNexthop (MNH) BGP-attribute: A new BGP attribute code TBD.

This document makes request to IANA to allocate the following sub registries for MNH attribute:.

1. "FwdAction" type as defined in 3.1.
2. Nexthop-Leg Descriptor TLV:"NhopDescrType" as defined in 3.2.
3. "Nexthop Attributes Sub-TLV type" as defined in 3.3.

This document makes request to IANA to allocate a BGP capability code TBD for MNH attribute:.

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

7. Security Considerations

The 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.

8. Acknowledgements

Thanks to Robert Raszuk, Gyan Mishra, Ron Bonica for the review, discussions and input to the draft.

9. References

9.1. Normative References

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

[RFC3392] Chandra, R. and J. Scudder, "Capabilities Advertisement with BGP-4", RFC 3392, DOI 10.17487/RFC3392, November 2002, <<https://www.rfc-editor.org/info/rfc3392>>.

9.2. 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](https://datatracker.ietf.org/doc/html/draft-ietf-idr-add-paths-guidelines-08#section-2)>.

[BGP-CT] Vairavakkalai, Ed., "BGP Classful Transport Planes", 25 August 2021, <<https://datatracker.ietf.org/doc/draft-kaliraj-idr-bgp-classful-transport-planes/12/>>.

[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 signalled MPLS-namespaces", 28 December 2021, <<https://datatracker.ietf.org/doc/html/draft-kaliraj-bess-bgp-sig-private-mpls-labels-04>>.

[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>>.

Authors' Addresses

Kaliraj Vairavakkalai
Juniper Networks, Inc.
1194 N. Mathilda Ave.
Sunnyvale, CA 94089
United States of America

Email: kaliraj@juniper.net

Minto Jeyananth
Juniper Networks, Inc.
1194 N. Mathilda Ave.
Sunnyvale, CA 94089
United States of America

Email: minto@juniper.net

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

Email: gyan.s.mishra@verizon.com