Network Working Group
Internet-Draft
Intended status: Experimental
Expires: December 13, 2021

K. Vairavakkalai
M. Jeyananth
Juniper Networks, Inc.

June 11, 2021

BGP MultiNexthop attribute
draft-kaliraj-idr-multinexthop-attribute-01

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 path-selection between the various additional-paths, which may tie-break on some arbitrary step like Router-Id.

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" that can be used for this purpose.

This attribute can be used for both labeled and unlabled BGP families. For labeled-families, it is used for a different purpose in "downstream allocation" case than "upstream allocation" scenarios.

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/.

Copyright Notice

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 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 upstream node based on path-selection between the various additional-paths, which may tie-break on some arbitrary step like Router-Id.

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" that can be used for this purpose.

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, 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 draft [MPLS-NAMESPACES]

3. The "MultiNexthop" BGP attribute encoding

"MultiNexthop" is a new BGP optional-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.

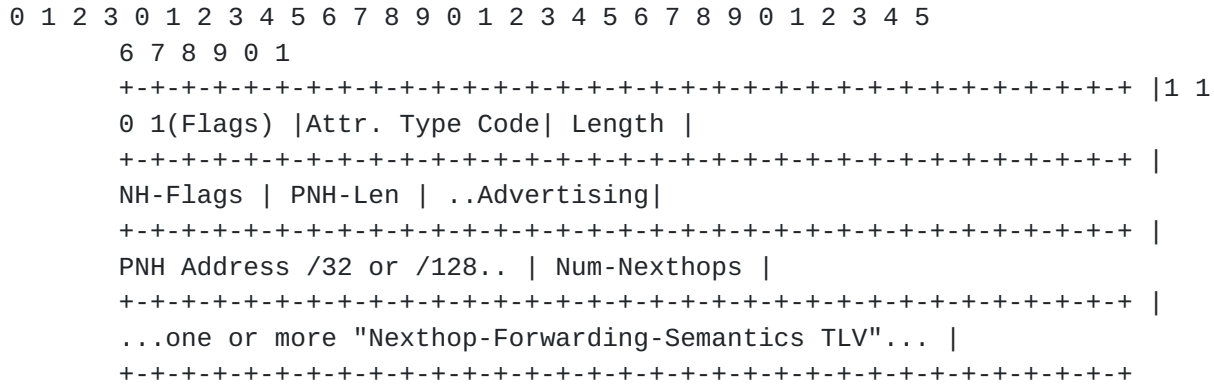


Fig 1: MultiNexthop - BGP Attribute

Flags BGP Path-attribute flags. 1101 to indicate Optional Transitive, Extended-length field Length Two bytes field stating length of attribute value in bytes NH-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 U: 0 means the Downstream-allocation, attribute describes forwarding state present at advertising-speaker PNH-Len NH-Length in bits (= 32 or 128) Advertising PNH IPv4 or IPv6 PNH-address (Len = 32 or 128) advertised in NEXT_HOP or MP_REACH_NLRI attr. Used to sanity-check this attribute Num-Nexthops >1 if ECMP or Alternate-paths

Sec 3.2 describes the Nexthop-Forwarding-Semantics TLV.

3.1. Operations

3.1.1. Interaction 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 MultiNexthop-attribute SHOULD ignore the attribute 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.

3.1.2. Interaction with Addpath

A RR advertising ADD_PATHs should use the MultiNexthop attribute when comparing with next-hop of other contributing paths and arriving on set of paths to advertise to Addpath receivers.

3.1.3. 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.

3.1.4. 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 unlabled BGP families.

A MultiNexthop attribute with U=0 is called "Label-Nexthop-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. Like, this may help in better debugging and manageability, or enable the label-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, today there is no way to signal to the receiving-speaker what the forwarding-semantic for the label should be. 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.

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.

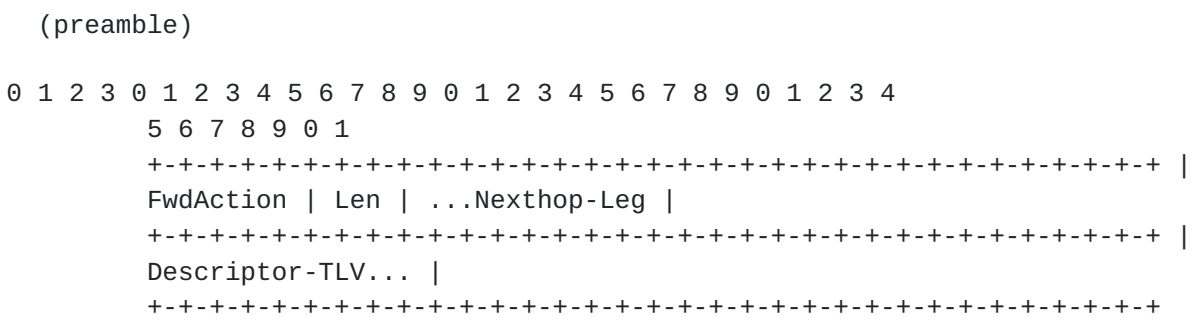


Fig 2: Nexthop Forwarding Semantics TLV

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

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.

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 explained in [MPLS-NAMESPACES] when Upstream-label-allocation is in use.

The "Pop-And-Lookup" action may result in a MPLS-lookup or an upper-layer (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.

(preamble)

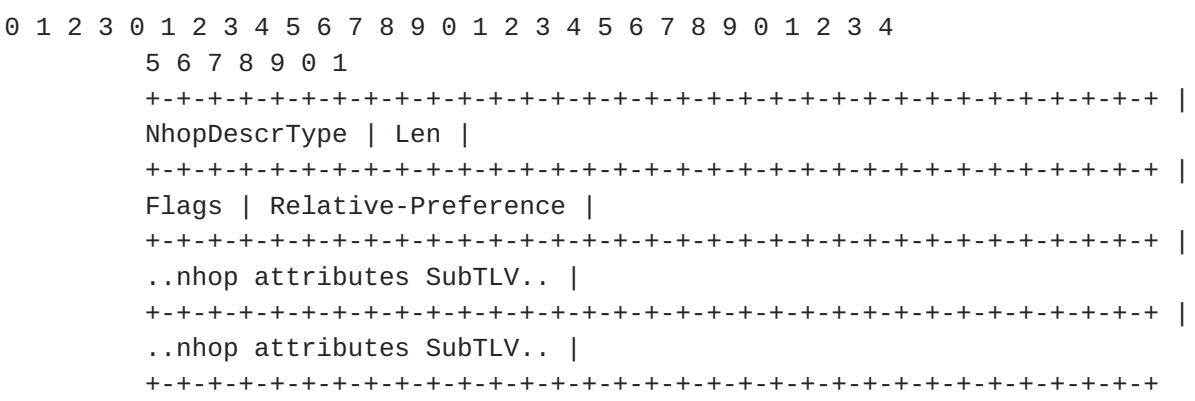


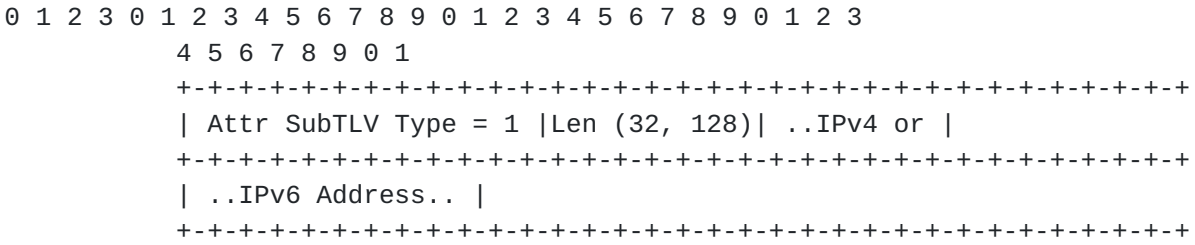
Fig 3: Nexthop Descriptor TLV

NhopDescrType Meaning 1 IPv4-nexthop 2 IPv6-nexthop 3 Labeled-IP-Nexthop 4 Forwarding-Context-Nexthop Len Length of Nexthop-Descriptor-TLV including Flags, Relative-Weight and all SubTLVs Flags Must send zero. Must ignore on receive. Relative-Preference Unsigned 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-Nexthop 3 Bandwidth 4 Load-Balance-Factor 5 Forwarding-context Name 6 Fprwarding-context Route-Distinguisher

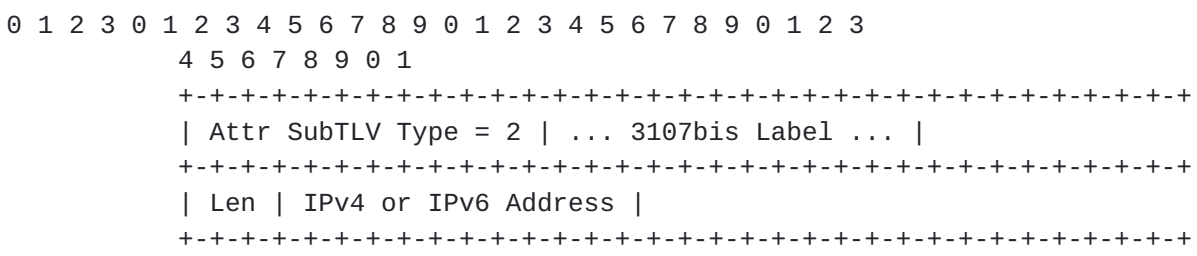
3.4.1. IP Address



IP-Address attribute sub-TLV

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

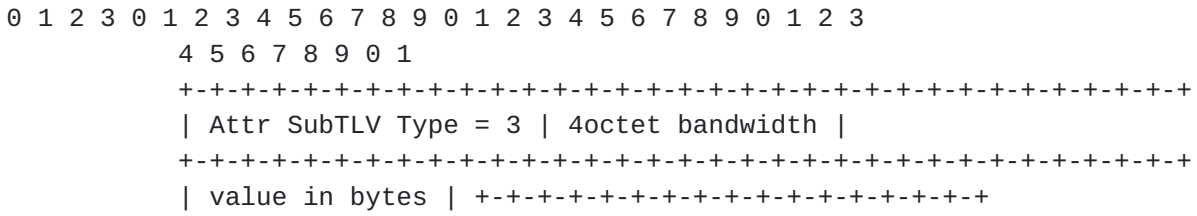
3.4.2. Labeled IP nexthop



"Labeled nexthop" attribute sub-TLV

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

3.4.3. Available Bandwidth



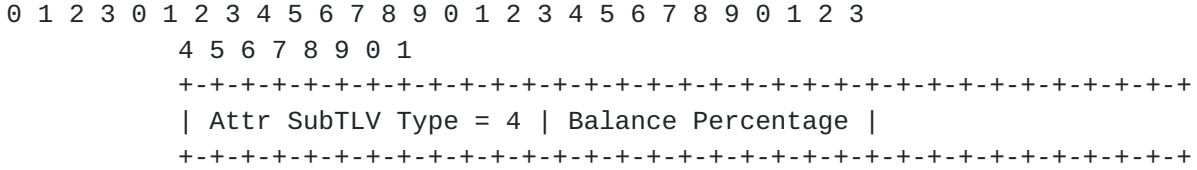
3.3.6. "Bandwidth" attribute sub-TLV

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

The bandwidth of the link is expressed as 4 octets in IEEE floating point format, units being bytes (not bits!) per second

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

3.4.4. Load balance factor

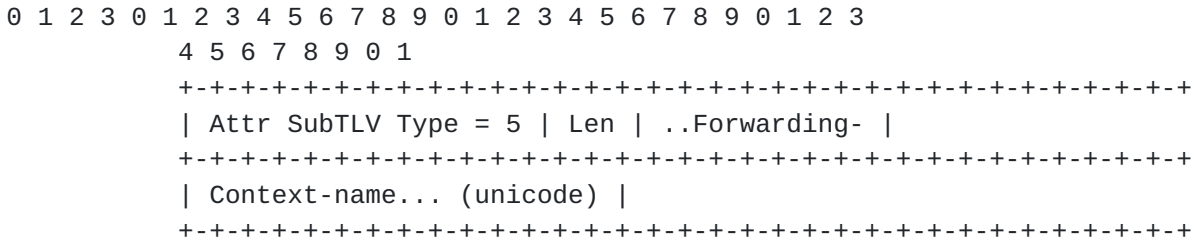


"Load-Balance-Factor" attribute sub-TLV

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

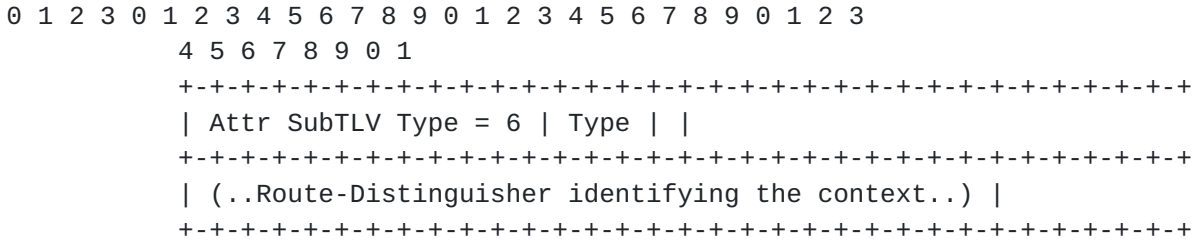
3.4.5. Forwarding-context name



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 Fowarding-context-name identfies the forwarding-context (for e.g. the VRF-name) where the lookup should happen after pop label.

3.4.6. Forwarding-context Route-Distinguisher



"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 RD uniquely identfies 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 possibe.

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 keeping the route hidden and not act on it, and MUST NOT tear down the BGP session.

5. IANA Considerations

This document makes request to IANA to allocate the following codes.

1. Multi-Nexthop-Descriptor BGP-attribute: A new BGP attribute code TBD.

2. "FwdAction" type as defined in 3.1.

3. Nexthop-Leg Descriptor TLV:"NhopDescrType" as defined in 3.2.

4. "Nexthop Attributes Sub-TLV type" as defined in 3.3.

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

6. Security Considerations

Like any other optional transitive BGP attribute, it is possible that this attribute gets propagated thru speakers that don't understand this attribute and an error detected by a speaker multiple hops away. This is mitigated by requiring the receiving speaker to remove this attribute when doing nexthop-self. And following the error handling procedures described above.

7. Acknowledgements

8. References

8.1. Normative References

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

8.2. References

[MPLS-NAMESPACES] Vairavakkalai, K., "BGP signalled MPLS-namespaces (draft-kaliraj-bess-bgp-signaled-private-mpls-labels-01)".

Authors' Addresses

Kaliraj Vairavakkalai
Juniper Networks, Inc.
1194 N. Mathilda Ave.
Sunnyvale, CA 94089
US

Email: kaliraj@juniper.net

Minto Jeyananth
Juniper Networks, Inc.
1194 N. Mathilda Ave.
Sunnyvale, CA 94089
US

Email: minto@juniper.net