Network Working Group
Internet-Draft
Updates: 6790 (if approved)
Intended status: Standards Track
Expires: December 13, 2019

B. Decraene
Orange
K. Kompella
Juniper Networks, Inc.
W. Henderickx
Nokia

June 11, 2019

BGP Next-Hop dependent capabilities
draft-ietf-idr-next-hop-capability-05

Abstract

RFC 5492 advertises the capabilities of the BGP peer. When the BGP peer is not the same as the BGP Next-Hop, it is useful to also be able to advertise the capability of the BGP Next-Hop, in particular to advertise forwarding plane features. This document defines a mechanism to advertise such BGP Next Hop dependent Capabilities.

This document defines a new BGP non-transitive attribute to carry Next-Hop Capabilities. This attribute is guaranteed to be deleted or updated when the BGP Next Hop is changed, in order to reflect the capabilities of the new BGP Next-Hop.

This document also defines a Next-Hop capability to advertise the ability to process the MPLS Entropy Label as an egress LSR for all NLRI advertised in the BGP UPDATE. It updates RFC 6790 with regard to this BGP signaling.

Requirements Language

The key words "MUST", "MUST NOT", "REQUIRED", "SHALL", "SHALL NOT", "SHOULD", "SHOULD NOT", "RECOMMENDED", "NOT RECOMMENDED", "MAY", and "OPTIONAL" in this document are to be interpreted as described in BCP 14 RFC 2119 [RFC2119] RFC 8179 [RFC8174] when, and only when, they appear in all capitals, as shown here.

Status of This Memo

Copyright Notice

Table of Contents

1. Introduction 3
2. BGP Next-Hop dependent Capabilities Attribute 3
  2.1. Encoding 3
  2.2. Attribute Operation 4
  2.3. Interpreting received Capability 5
  2.4. Attribute Error Handling 5
  2.5. Network operation considerations 6
3. Entropy Label Next-Hop dependent Capability 6
  3.1. Readable Label Depth 7
  3.2. Entropy Label Next-Hop Capability error handling 9
4. IANA Considerations 9
  4.1. Next-Hop Capabilities Attribute 9
  4.2. Next-Hop Capability registry 9
5. Security Considerations 9
6. Acknowledgement 10
7. References 10
  7.1. Normative References 10
  7.2. Informative References 11
Appendix A. Changes / Author Notes 12
Authors' Addresses 12

1. Introduction

[RFC5492] advertises the capabilities of the BGP peer. When the BGP peer is not the same as the BGP Next-Hop, it is useful to also be able to advertise the capability of the BGP Next-Hop, in particular to advertise forwarding plane features. This document defines a mechanism to advertise such BGP Next Hop Capabilities.

This document defines a new BGP non-transitive attribute to carry Next-Hop Capabilities. This attribute is guaranteed to be deleted or updated when the BGP Next Hop is changed, in order to reflect the capabilities of the new BGP Next-Hop. Hence it allows advertising capabilities which are dependent of the BGP Next-Hop.

This attribute advertises the capabilities of the BGP Next-Hop for the NLRI advertised in the same BGP update. A BGP Next-Hop may advertise different capabilities for different set of NLRI.

This document also defines a first application to advertise the capability to handle the MPLS Entropy Label defined in [RFC6790]. Note that RFC 6790 had originally defined a BGP attribute for this but it has been latter deprecated in [RFC7447].

2. BGP Next-Hop dependent Capabilities Attribute

2.1. Encoding

The BGP Next-Hop dependent Capabilities Attribute is an optional, non-transitive BGP Attribute, of value TBD1. The attribute consists of a set of Next-Hop Capabilities.

The inclusion of a Next-Hop Capability "X" in a BGP UPDATE message, indicates that the BGP Next-Hop, encoded in either the NEXT_HOP attribute defined in [RFC4271] or the Network Address of Next Hop field of the MP_REACH_NLRI attribute defined in [RFC4760], supports the capability "X" for the NLRI advertised in this BGP UPDATE.

This document does not make a distinction between these two Next-Hop fields and uses the term 'BGP Next-Hop' to refer to whichever one is used in a given BGP UPDATE message.

A Next-Hop Capability is a triple (Capability Code, Capability Length, Capability Value) aka a TLV:

```
                     +------------------------------+
                     | Capability Code (2 octets)   |
                     +------------------------------+
                     | Capability Length (2 octets) |
                     +------------------------------+
                     | Capability Value (variable)  |
                     ~                              ~
                     +------------------------------+
```

Figure 1: BGP Next-Hop Capability

Capability Code: a two-octets unsigned binary integer which indicates the type of "Next-Hop Capability" advertised and unambiguously identifies an individual capability.

Capability Length: a two-octets unsigned binary integer which indicates the length, in octets, of the Capability Value field. A length of 0 indicates that no Capability Value field is present.

Capability Value: a variable-length field. It is interpreted according to the value of the Capability Code.

BGP speakers SHOULD NOT include more than one instance of a Next-Hop capability with the same Capability Code, Capability Length, and Capability Value. Note, however, that processing of multiple instances of such capability does not require special handling, as additional instances do not change the meaning of the announced capability; thus, a BGP speaker MUST be prepared to accept such multiple instances.

BGP speakers MAY include more than one instance of a capability (as identified by the Capability Code) with non-zero Capability Length field, but with different Capability Value and either the same or different Capability Length. Processing of these capability instances is specific to the Capability Code and MUST be described in the document introducing the new capability.

2.2. Attribute Operation

The BGP Next-Hop dependent Capabilities attribute being non-transitive, as per [RFC4271], a BGP speaker which does not understand it will quietly ignore it and not pass it along to other BGP peers.

A BGP speaker that understands the BGP Next-Hop dependent Capabilities Attribute and does not change the BGP Next-Hop, SHOULD NOT change the BGP Next-Hop dependent Capabilities Attribute and SHOULD pass the attribute unchanged along to other BGP peers.

A BGP speaker that understands the BGP Next-Hop dependent Capabilities Attribute and changes the BGP Next-Hop, MUST remove or update the received BGP Next-Hop dependent Capabilities Attribute before propagating the BGP UPDATE to other BGP peers. If the capability is not removed, it MUST be updated to only advertise the capabilities of the new BGP Next-Hop for these NLRIs. An implementation MAY allow, by configuration, to not advertise some of the capabilities of a BGP Next-Hop. If a received capability is unknown, it can't be updated hence unknown capabilities MUST be removed when the BGP Next-Hop is changed.

The BGP Next-Hop Capability Code MUST reflect the capability of the router indicated in the BGP Next-Hop, for the NLRI advertised in the BGP UPDATE. If a BGP speaker sets the BGP Next-Hop to an address of a different router, it MUST NOT advertise a BGP Next-Hop Capability not supported by this router for these NLRI.

2.3. Interpreting received Capability

A BGP speaker receiving a BGP Next-Hop Capability Code that it supports behave as defined in the document defining this Capability Code. A BGP speaker receiving a BGP Next-Hop Capability Code that it does not support MUST ignore this BGP Next-Hop Capability Code. In particular, this MUST NOT be handled as an error. In both cases, the BGP speaker MUST examine the remaining BGP Next-Hop Capability Code(s) that may be present in the BGP Next-Hop Capabilities Attribute.

The presence of a Next-Hop Capability SHOULD NOT influence route selection or route preference, unless tunneling is used to reach the BGP Next-Hop or the selected route has been learnt from EBGP (i.e. the Next-Hop is in a different AS). Indeed, it is in general impossible for a node to know that all BGP routers of the Autonomous System (AS) will understand a given Next-Hop Capability; and having different routers, within an AS, use a different preference for a route, may result in forwarding loops if tunnelling is not used to reach the BGP Next-Hop.

2.4. Attribute Error Handling

A BGP Next-Hop dependent Capabilities Attribute is considered malformed if the length of the Attribute is not equal to the sum of all (BGP Next-Hop dependent Capability Length +4) of the capabilities carried in this attribute. Note that "4" is the length of the fields "Type" and "Length" of each BGP Next Hop dependent Capability, as the capability length only account for the length of the Value field.

A BGP UPDATE message with a malformed BGP Next-Hop dependent Capabilities Attribute SHALL be handled using the approach of "attribute discard" defined in [RFC7606].

Unknown Next-Hop Capabilities Codes MUST NOT be considered as an error.

A document that specifies a new Next-Hop Capability SHOULD provide specifics regarding what constitutes an error for that Next-Hop Capability.

If a Next-Hop dependent Capability is malformed, this Capability MUST be ignored and removed. Others Next-Hop Capabilities MUST be processed as usual.

2.5. Network operation considerations

In the corner case where multiple nodes use the same IP address as their BGP Next-Hop, aka anycast nodes as described in [RFC4786], a BGP speaker MUST NOT advertise a given Next-Hop Capability unless all nodes sharing this same IP address support this Next-Hop Capability. The network operator operating those anycast nodes is responsible for enforcing that an anycast node does not advertise a BGP Next-Hop capability not supported by all nodes advertising this anycast address. This can be performed by using anycast nodes sharing the same capabilities or by filtering the BGP Next-Hop Capabilities which are not shared by all anycast nodes.

For security considerations, a network operator may want to filter the BGP Next-Hop capabilities advertised from or to external Autonomous Systems on a per capability, capability type or attribute basis.

3. Entropy Label Next-Hop dependent Capability

The Entropy Label Next-Hop Capability has type code 1 and a length of 0 or 1 octet.

The inclusion of the "Entropy Label" Next-Hop Capability indicates that the BGP Next-Hop can be sent packets, for all routes indicated in the NRLI, with a MPLS entropy label (ELI, EL) added immediately after the label stack advertised with the NLRI.

On the receiving side, suppose BGP speaker S has determined that packet P is to be forwarded according to BGP route R, where R is a route of one of the labeled address families. And suppose that L is the label stack embedded in the NLRI of route R. Then to forward packet P according to route R, S either replaces P's top label with

L, or else pushes L onto the MPLS label stack. If the EL-Capability is advertised in the BGP UPDATE advertising this route R, S knows that it may safely place the ELI and an EL on the label stack immediately beneath L.

A BGP speaker S that sends an UPDATE with the BGP Next-Hop "NH" MAY include the Entropy Label Next-Hop Capability only if the NLRI are labelled and for all the NLRI in the BGP UPDATE, either of the following is true:

- o Egress case: NH is the egress of the LSP advertised with the NLRI and its capable of handling the ELI during the lookup of the MPLS top label.

- o Transit LSR case: NH is a transit LSR for the LSP advertised with the NLRI (i.e. NH swaps one of the label advertised in the NLRI) and next downstream BGP Next-Hop(s) has(have) advertised the Entropy Label Next-Hop Capability (or a similar capability signalled by protocol P if the route is redistributed, by NH, from protocol P into BGP).

3.1. Readable Label Depth

When stacked LSPs are used and a LSR nests LSP(s) inside this BGP signalled LSP, its useful for the ingress LSRs to know how many labels the BGP Next-Hop and its downstream LSR(s) may read when load-balancing based on the Entropy Label. In other words, how many labels the ingress LER may push, before pushing an entropy label that will be seen by the BGP Next-Hop and its downstream LSR(s).

This maximum number of labels is called the Readable Label Depth (RLD) of the LSP(s). It is related, yet different, to the RLD of an node which is defined in [I-D.ietf-mpls-spring-entropy-label]

The RLD of the LSP(s) advertised in the NLRI, may be advertised in the first octet of value field of the Entropy Label Next-Hop Capability. This value field is optional. If present, the value field is a one-octet unsigned binary integer which indicates the maximum Readable Label Depth (RLD) of the LSP(s) advertised in the NLRI. In other words, this is the maximum number of MPLS labels that may be pushed by the ingress, before pushing the ELI, EL labels, where the BGP Next-Hop and its downstream LSR(s) are capable of performing load-balancing based on the entropy label.

S SHOULD advertise a RLD of:

- o If S is the egress of the LSP(s) advertised in the NLRI: its own local RLD;

- o If S is propagating in BGP a route received in BGP: the minimum of:

  - * its own node RLD;

  - * the RLD of the LSP from itself to the BGP NEXT_HOP of its received route minus (Number of Labels in the received NLRI - Number of Labels in the sent NLRI);

  - * 0 if a RLD is not present in its received routes or the RLD in the received BGP route minus (Number of Labels in the received NLRI - Number of Labels in the sent NLRI).

  - * Note that the first term represents the limitation of the new BGP NEXT_HOP (S), the second term the contribution of the LSR(s) between the new BGP NEXT_HOP (S) and the old (received) BGP NEXT_HOP (S'), the third term represent the contribution from the old BGP NEXT_HOP (S') toward the egress.

- o If S is propagating in BGP a route received in protocol X: the minimum of:

  - * its own node RLD;

  - * the minimum of thg RLD(s) in the received protocol X to reach the NLRI(s).

255 is a reserved value.

Note that the local RLD is meant as a node value. If a router has multiple line cards with different capabilities, the router SHOULD advertise the smallest one. However, a router MAY choose to only consider the line cards that may be used by the BGP routers receiving the ELC. e.g. if the ELC is advertised over an EBGP session with peer A, a router MAY consider only the line cards connected to peer A.

Advertisement of the RLD is optional. When used, changes in IGP routing may trigger BGP re-advertisement and hence will increase BGP churn. If the RLD is decreased, it SHOULD be readvertised immediatly. If the RLD is increased is MAY NOT be readvertised immediatly. We note however that labelled BGP routes are typically not advertised outside of an administrative domain hence the churn would be limited to this administrative domain.

3.2. Entropy Label Next-Hop Capability error handling

If the Entropy Label Next-Hop Capability is present more than once, it MUST be considered as received once with a length of 0.

If the Entropy Label Next-Hop Capability is received with a length other than 0 or 1, it is not considered malformed, and its semantics are exactly the same as if it had a length of 1. In other words, additional octets MUST be ignored. This allows for the graceful addition of future extensions.

4. IANA Considerations

4.1. Next-Hop Capabilities Attribute

IANA is requested to allocate a new Path Attribute, called "Next-Hop Capabilities", type Code TBD1, from the "BGP Path Attributes" registry.

4.2. Next-Hop Capability registry

The IANA is requested to create and maintain a registry entitled "BGP Next-Hop Capabilities".

The registration policies [RFC8126] for this registry are:

0	Reserved
1-63	IETF Review
64-65534	First Come First Served
65535	Reserved

IANA is requested to make the following initial assignments:

Registry Name: Next-Hop Capability.

Value	Meaning	Reference
0	Reserved (not to be allocated)	This document
1	Entropy Label	This document
2-65534	Unassigned	
65535	Reserved for future registry extension	This document

5. Security Considerations

This document does not introduce new security vulnerabilities in BGP. Specifically, an operator who is relying on the information carried in BGP must have a transitive trust relationship back to the source of the information. Specifying the mechanism(s) to provide such a

relationship is beyond the scope of this document. Please refer to the Security Considerations section of [RFC4271] for security mechanisms applicable to BGP.

As this attribute is removed when the BGP Next-Hop is changed, the source of the information is the router which IP address is indicated in the BGP Next-Hop. Such Next-Hop is typically either within the AS when a BGP Next-Hop Self policy is configured, or in the neighboring AS with which an interconnection and an EBGP has been established. If this neighboring AS is not trusted with regards to information carried in the BGP attribute, or carried in a specific capability, this attribute or specific capability should be removed when received. Note that in some cases, this Next-Hop may advertise information based on information it has received from its own downstream BGP Next-Hop, hence the transitive trust relationship. If the underlying transport between both ASes is not trusted, BGP transport should be protected for integrity and authentification.

The advertisement of BGP Next-Hop capabilities to EBGP peers may disclose, to the peer AS, some capabilities of the BGP node and may help fingerprinting its hardware model and software version. This may be mitigated by filtering the capability advertised to EBGP peers.

Security of the Entropy Label capability advertisement is unchanged compared to [RFC6790] which originally defined this signaling.

6. Acknowledgement

The Entropy Label Next-Hop Capability defined in this document is based on the ELC BGP attribute defined in section 5.2 of [RFC6790].

The authors wish to thank John Scudder for the discussions on this topic and Eric Rosen for his in-depth review of this document.

The authors wish to thank Jie Dong and Robert Raszuk for their review and comments.

7. References

7.1. Normative References

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

[RFC4271] Rekhter, Y., Ed., Li, T., Ed., and S. Hares, Ed., "A Border Gateway Protocol 4 (BGP-4)", RFC 4271, DOI 10.17487/RFC4271, January 2006, <https://www.rfc-editor.org/info/rfc4271>.

[RFC4760] Bates, T., Chandra, R., Katz, D., and Y. Rekhter, "Multiprotocol Extensions for BGP-4", RFC 4760, DOI 10.17487/RFC4760, January 2007, <https://www.rfc-editor.org/info/rfc4760>.

[RFC6790] Kompella, K., Drake, J., Amante, S., Henderickx, W., and L. Yong, "The Use of Entropy Labels in MPLS Forwarding", RFC 6790, DOI 10.17487/RFC6790, November 2012, <https://www.rfc-editor.org/info/rfc6790>.

[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>.

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

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

7.2. Informative References

[I-D.ietf-mpls-spring-entropy-label]
Kini, S., Kompella, K., Sivabalan, S., Litkowski, S., Shakir, R., and J. Tantsura, "Entropy label for SPRING tunnels", draft-ietf-mpls-spring-entropy-label-12 (work in progress), July 2018.

[RFC4786] Abley, J. and K. Lindqvist, "Operation of Anycast Services", BCP 126, RFC 4786, DOI 10.17487/RFC4786, December 2006, <https://www.rfc-editor.org/info/rfc4786>.

[RFC5492] Scudder, J. and R. Chandra, "Capabilities Advertisement with BGP-4", RFC 5492, DOI 10.17487/RFC5492, February 2009, <https://www.rfc-editor.org/info/rfc5492>.

[RFC7447] Scudder, J. and K. Kompella, "Deprecation of BGP Entropy Label Capability Attribute", RFC 7447, DOI 10.17487/RFC7447, February 2015, <https://www.rfc-editor.org/info/rfc7447>.

Appendix A. Changes / Author Notes

[RFC Editor: Please remove this section before publication]

Changes -01:

- o Capability code and length encoded over 2 octets (from one). IANA registry is now mainly FCFS.

- o Addition of section "Network operation considerations", in particular to discuss anycast nodes.

- o Enhanced Security consideration (capability advertised to external ASes).

- o Editorial changes and typo corrections.

Changes -02: No change. Refresh only.

Changes -03: Addition of the optional Readable Label Depth.

Changes -04: Update to security section, following discussion on the IDR mailing list.

Changes -05: No change. Refresh only.

Authors' Addresses

Bruno Decraene
Orange

Email: bruno.decraene@orange.com

Kireeti Kompella
Juniper Networks, Inc.
1194 N. Mathilda Avenue
Sunnyvale, CA 94089
USA

Email: kireeti.kompella@gmail.com

Wim Henderickx
Nokia
Copernicuslaan 50
Antwerp 2018, CA 95134
Belgium

Email: wim.henderickx@nokia.com