Workgroup: Internet Engineering Task Force
Internet-Draft:
draft-ietf-idr-entropy-label-02
Updates: 6790, 7447 (if approved)
Published: 21 December 2022
Intended Status: Standards Track
Expires: 24 June 2023
Authors: B. Decraene, Ed. J. G. Scudder, Ed.
Orange Juniper Networks
W. Henderickx K. Kompella S. Mohanty
Nokia Juniper Networks Cisco Systems
J. Uttaro B. Wen
AT&T Comcast

BGP Router Capabilities Attribute

Abstract

RFC 5492 allows a BGP speaker to advertise its capabilities to its peer. When a route is propagated beyond the immediate peer, it is useful to allow certain capabilities to be conveyed further. In particular, it may be useful to advertise forwarding plane features.

This specification defines a new BGP transitive attribute to carry such capability information, the "Router Capabilities Attribute," or RCA.

This specification also defines an RCA capability that can be used 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 and RFC 7447 concerning this BGP signaling.

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 24 June 2023.

Table of Contents

1. Introduction

[RFC5492] allows a BGP speaker to advertise its capabilities to its peer. When a route is propagated beyond the immediate peer, it is useful to allow certain capabilities to be conveyed further. In particular, it may be useful to advertise forwarding plane features.

This specification defines a new BGP optional transitive attribute to carry such capability information, the "Router Capabilities Attribute", or RCA. (This somewhat ponderous name is regrettable but is needed in order to be descriptive while still distinguishing it from RFC 5492 BGP Capabilities.)

Since the RCA is intended chiefly for conveying information about forwarding plane features, it needs to be regenerated whenever the BGP route's next hop is changed. Since owing to the properties of BGP transitive attributes this can't be guaranteed (an intermediate router that doesn't implement this specification would be expected to propagate the RCA as opaque data), the RCA identifies itself with the next hop of its originator. If the RCA passes through a router that changes the next hop without regenerating the RCA, they will fail to match when later examined, and the recipient can act accordingly. This scheme allows RCA support to be introduced into a network incrementally. Complete details are provided in Section 2.

This specification also defines an RCA to advertise the ability to process the MPLS Entropy Label as an egress LSR for all NLRI advertised in the BGP UPDATE. It updates [RFC6790] and [RFC7447] with regard to this BGP signaling, this is further discussed in Section 3.

1.1. 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 [RFC2119] [RFC8174] when, and only when, they appear in all capitals, as shown here.

2. BGP Router Capabilities Attribute

2.1. Encoding

The BGP Router Capabilities attribute (RCA attribute, or just RCA) is an optional, transitive BGP path attribute with type code 39. The RCA has as its data a network layer address, representing the next hop of the route the RCA accompanies. The RCA signals potentially useful optimizations, so it is desirable to make it transitive; the next hop data is to ensure correctness if it traverses BGP speakers that do not understand the RCA.

The Attribute Data field of the RCA attribute is encoded as a header portion that identifies the originator of the attribute, followed by one or more capability TLVs.

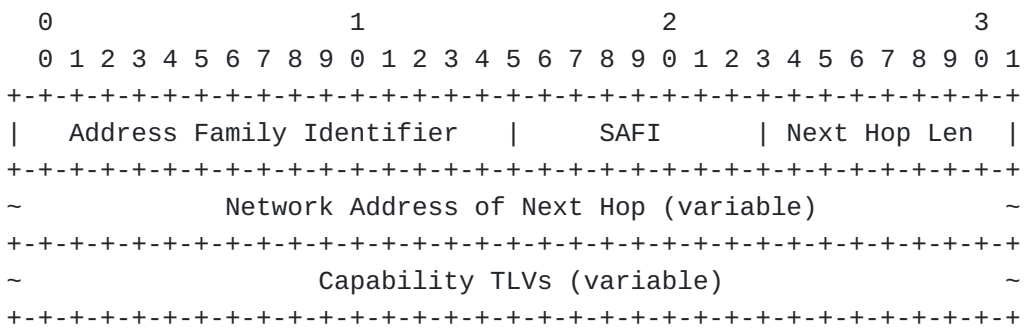

Figure 1: RCA Format

The meanings of the header fields (Address Family Identifier, SAFI or Subsequent Address Family Identifier, Length of Next Hop, and Network Address of Next Hop) are as given in Section 3 of [RFC4760].

In turn, each Capability is a triple (Capability Code, Capability Length, Capability Value):

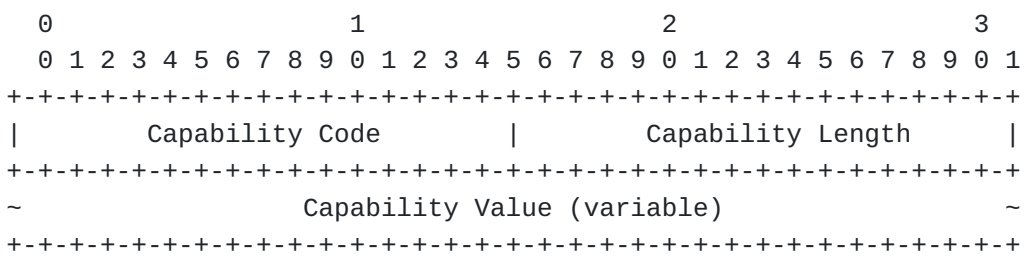

Figure 2: Capability TLV Format

Capability Code: a two-octet unsigned binary integer that indicates the type of Capability advertised and unambiguously identifies an individual capability.

Capability Length: a two-octet unsigned binary integer that 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.

A BGP speaker **MUST NOT** include more than one instance of a capability with the same Capability Code, Capability Length, and Capability Value. Note, however, that processing multiple instances of such a 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 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.

Capability TLVs **MUST** be placed in the RCA in increasing order of Capability Code. (In the event of multiple instances of a capability with the same Capability Code as discussed above, no further sorting order is defined here.) Although the major sorting order is mandated, an implementation **MAY** elect to be prepared to consume capabilities in any order, for robustness reasons.

2.2. Sending the RCA

Suppose a BGP speaker S has a route R it wishes to advertise with next hop N to its peer.

If S is originating R into BGP, it **MAY** include an RCA attribute with it, that carries capability TLVs that describe aspects of R. S **MUST** set the header portion of the RCA to be equal to N, using the encoding given above.

If S has received R from some other BGP speaker, two possibilities exist. First, S could be propagating R without changing N. In that case, S need take no special action, it **SHOULD** simply propagate the RCA unchanged unless specifically configured otherwise. Indeed, we observe that this is no different from the default action a BGP speaker takes with an unrecognized optional transitive attribute -- it is treated as opaque data and propagated.

Second, S could be changing R in some way, and in particular, it could be changing N. If S has changed N it **MUST NOT** propagate the RCA unchanged. It **MAY** include a newly-constructed RCA attribute with R, constructed as described above in the "originating R into BGP" case. Any given capability TLV carried by the newly-constructed RCA attribute might use information from the received RCA attribute as input to its construction; the details of this are specific to the definition of each capability.

The RCA **MAY** be sent by default to IBGP peers. It **MUST NOT** be sent by default to peers not under the administrative control of the local network administrator (so, generally, to EBGP peers).

We note that due to the nature of BGP optional transitive path attributes, any BGP speaker that does not implement this specification will propagate the RCA, the requirements of this section notwithstanding. Such a speaker will not update the RCA, however.

2.3. Receiving the RCA

By default, the RCA **MUST NOT** be accepted from peers not under the administrative control of the local network administrator (so, generally, from EBGP peers); if received it **MUST** be discarded without further processing, except that the contents **MAY** be logged. An implementation **MAY** enable RCA processing by default from peers under the administrative control of the local network administrator (so, generally, from IBGP peers). An implementation **SHOULD** provide the ability to modify these default settings by configuration.

When a BGP speaker receives a BGP route that includes the RCA, it **MUST** compare the address given in the header portion of the RCA to the next hop of the BGP route. If the two are equal, the RCA may be further processed. If the two are not equal, it means some intermediate BGP speaker that handled the route in transit both does not support RCA, and changed the next hop of the route. In this case, the contents of the RCA cannot be used, and the RCA **MUST** be discarded without further processing, except that the contents **MAY** be logged.

A BGP speaker receiving a Capability Code that it supports behaves as defined in the document defining the Capability Code. A BGP speaker receiving a Capability Code that it does not support **MUST** ignore that Capability Code. In particular, it **MUST NOT** be handled as an error.

The presence of a 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 learned from External BGP (that is, the next hop is in a different Autonomous System). Indeed, it is in general impossible for a node to know that all BGP routers of the Autonomous System (AS) will understand a given capability, and if different routers within an AS were to use a different preference for a route, forwarding loops could result unless tunneling is used to reach the BGP next hop.

2.4. Attribute Error Handling

An RCA is considered malformed if the length of the attribute is inconsistent with the lengths of the contained capability TLVs.

A BGP UPDATE message with a malformed RCA **SHALL** be handled using the approach of "attribute discard" defined in [RFC7606].

Unknown Capability Codes **MUST NOT** be considered to be an error.

A document that specifies a new RCA Capability should provide specifics regarding what constitutes an error for that RCA Capability.

If a capability TLV is malformed, that capability TLV **MUST** be ignored and removed. Other capability TLVs **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, such as with anycast nodes as described in [RFC4786], a BGP speaker **MUST NOT** advertise a given capability unless all nodes sharing this same IP address support this capability. The network operator operating those anycast nodes is responsible for ensuring that an anycast node does not advertise a capability not supported by all nodes sharing this anycast address. The means for accomplishing this are beyond the scope of this document.

3. Entropy Label Capability (ELCv3)

When BGP [RFC4271] is used for distributing labeled Network Layer Reachability Information (NLRI) as described in, for example, [RFC8277], the route may include the ELCv3 as part of the RCA. The inclusion of this capability with a route indicates that the egress of the associated Label Switched Path (LSP) can process entropy labels as an egress Label Switched Router (LSR) for that route -- see Section 4.2 of [RFC6790]. Below, we refer to this for brevity as being "EL-capable."

For historical reasons, this capability is referred to as "ELCv3", to distinguish it from the prior Entropy Label Capability (ELC) defined in [RFC6790] and deprecated in [RFC7447], and the ELCv2 described in [I-D.scudder-bgp-entropy-label].

3.1. Encoding

The ELCv3 has capability code 1, capability length 0, and carries no value:

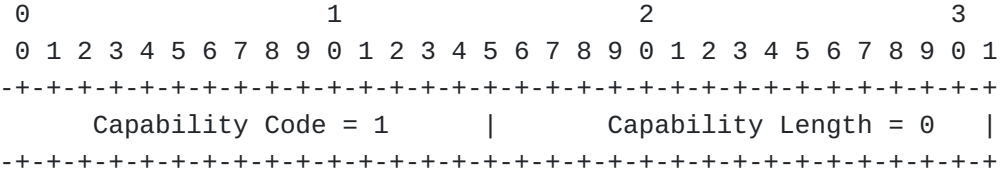

Figure 3: ELCv3 TLV Format

3.2. Sending the ELCv3

When a BGP speaker S has a route R it wishes to advertise with next hop N to its peer, it **MUST NOT** include the ELCv3 capability except

if it knows that the egress of the associated LSP L is EL-capable. Specifically, this will be true if S:

- Is itself the egress, and knows itself to be EL-capable, or
- Is re-advertising a BGP route it received with a valid ELCv3 capability, and is not changing the value of N, or
- Is re-advertising a BGP route it received with a valid ELCv3 capability, and is changing the value of N, and knows (for example, through configuration) that the router represented by N is either the LSP egress and is EL-capable, or that it will simply swap labels without popping the entire label stack and processing the label below, as with a transit LSR, or
- Is redistributing a route learned from another protocol, and that other protocol conveyed the knowledge that the egress of L was EL-capable (for example, this might be known through the LDP ELC TLV, Section 5.1 of [RFC6790]).

The ELCv3 **MAY** be advertised with routes that are labeled, such as those using SAFI 4 [RFC8277]. It **MUST NOT** be advertised with unlabeled routes.

3.3. Receiving the ELCv3

(Below, we assume that "includes the ELCv3" implies that the containing RCA has passed the checks specified in Section 2.3. If it had not passed, then the RCA would have been discarded and the ELCv3 would be deemed not to have been included.)

When a BGP speaker receives an unlabeled route that includes the ELCv3, it **MUST** discard the ELCv3.

When a BGP speaker receives a labeled route that includes the ELCv3, that indicates the LSP supports entropy labels, which implies that the receiving BGP speaker, if acting as ingress, **MAY** insert an entropy label as per Section 4.2 of [RFC6790].

3.4. ELCv3 Error Handling

The ELCv3 is considered malformed and must be disregarded if its length is other than zero.

4. IANA Considerations

IANA has made a temporary allocation in the BGP Path Attributes registry of the Border Gateway Protocol (BGP) Parameters group. IANA is requested to make this allocation permanent.

Value	Code	Reference
39	BGP Router Capabilities (RCA)	(this doc)

Table 1

IANA is requested to create a new registry called "BGP Router Capability Codes" within the Border Gateway Protocol (BGP) Parameters group. The registry's allocation policy is First Come, First Served. It is seeded with the following values:

Value	Description	Reference	Change Controller
0	reserved	(this doc)	IETF
1	ELCv3	(this doc)	IETF
65500 - 65534	reserved for experimental use	(this doc)	IETF
65535	reserved	(this doc)	IETF

Table 2

5. Security Considerations

5.1. Considerations for the RCA

The header portion of the RCA contains the next hop the attribute's originator included when sending it. This will typically be an IP address of the router in question. This may be an infrastructure address the network operator does not intend to announce beyond the border of its Autonomous System, and it may even be considered in some weak sense, confidential information. Although the desired operation of the protocol is for the attribute's propagation scope to be limited to the network operator's own Autonomous System, this can't be guaranteed in all cases -- if a border router doesn't implement this specification, the attribute, like all BGP optional transitive attributes, will propagate to neighboring Autonomous Systems. So, sometimes this information could leak beyond its intended scope. (Note that it will only propagate as far as the first router that does support this specification, at which point it will be discarded per Section 2.3.)

If the attribute leaks beyond its intended scope, capabilities within it would potentially be exposed. Specifications for individual capabilities should consider the consequences of such unintended exposure.

5.2. Considerations for the ELCv3 Capability

Insertion of an ELCv3 by an attacker could cause forwarding to fail. Deletion of an ELCv3 by an attacker could cause one path in the network to be overutilized and another to be underutilized. However, we note that an attacker able to accomplish either of these (below, an "on-path attacker") could equally insert or remove any other BGP path attribute or message. The former attack described above denies service for a given route, which can be accomplished by an on-path attacker in any number of ways even absent ELCv3. The latter attack defeats an optimization but nothing more; it seems dubious that an attacker would go to the trouble of doing so rather than launching some more damaging attack.

6. References

6.1. 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>.

[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>.

[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>.

[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>.

6.2. Informative References

[I-D.ietf-idr-next-hop-capability] Decraene, B., Kompella, K., and W. Henderickx, "BGP Next-Hop dependent capabilities", Work in Progress, Internet-Draft, draft-ietf-idr-next-hop-capability-08, 8 June 2022, <https://www.ietf.org/archive/id/draft-ietf-idr-next-hop-capability-08.txt>.

[I-D.scudder-bgp-entropy-label] Scudder, J. and K. Kompella, "BGP Entropy Label Capability, Version 2", Work in Progress, Internet-Draft, draft-scudder-bgp-entropy-label-00, 28 April 2022, <https://www.ietf.org/archive/id/draft-scudder-bgp-entropy-label-00.txt>.

[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>.

[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>.

Appendix A. Other Means of Signaling EL Capability

A router that supports this specification could also have other means to know that an egress is EL-capable, for example, it could support ELCv2 [I-D.scudder-bgp-entropy-label], or it could know through configuration. If a router learns through any means that an egress is EL-capable, it **MAY** treat the egress as EL-capable. For example, reception of a valid ELCv2 would be sufficient (even if a valid ELCv3 is not received), and similarly, reception of a valid ELCv3 would be sufficient (even if a valid ELCv2 is not received). The details of which methods are accepted for signaling EL capability are beyond the scope of this specification but **SHOULD** be configurable by the user.

Acknowledgements

This specification derives from two earlier documents, [I-D.ietf-idr-next-hop-capability] and [I-D.scudder-bgp-entropy-label].

[I-D.ietf-idr-next-hop-capability] included the following acknowledgements:

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.

[I-D.scudder-bgp-entropy-label] included the following acknowledgements:

Thanks to Swadesh Agrawal, Alia Atlas, Bruno Decraene, Martin Djernaes, John Drake, Adrian Farrell, Keyur Patel, Toby Rees, and Ravi Singh, for their discussion of this issue.

Contributors

Serge Krier
Cisco Systems

Email: sekrier@cisco.com

Kevin Wang
Juniper Networks

Email: kfwang@juniper.net

Authors' Addresses

Bruno Decraene (editor)
Orange

Email: bruno.decraene@orange.com

John G. Scudder (editor)
Juniper Networks

Email: jgs@juniper.net

Wim Henderickx

Nokia

Email: wim.henderickx@nokia.com

Kireeti Kompella
Juniper Networks

Email: kireeti@juniper.net

Satya Mohanty
Cisco Systems

Email: satyamoh@cisco.com

James Uttaro
AT&T

Email: ju1738@att.com

Bin Wen
Comcast

Email: Bin_Wen@comcast.com