Interdomain Routing
Internet-Draft
Updates: 2545 (if approved)
Intended status: Standards Track
Expires: 1 December 2025

R. White
Akamai Technologies
J. Tantsura
Nvidia
D. Abraitis
Hostinger

30 May 2025

Link-Local Next Hop Capability for BGP
draft-ietf-idr-linklocal-capability-01

Abstract

BGP [RFC4271], was originally designed to provide reachability between domains and between the edges of a domain.

To support IPv6 reachability, BGP relies heavily on its Multiprotocol Extensions as defined in [RFC2545], which is crucial for enabling BGP-4 to advertise IPv6 routes alongside IPv4. This extension not only permits the exchange of IPv6 routing information but also establishes the structure for IPv6 next hop handling within BGP updates. As such, BGP assumes the next hop towards any reachable destination may not reside on the advertising speaker, but rather may either be through a router connected to the same subnet as the speaker, or through a router only reachable by traversing multiple hops through the network. [RFC2545] introduced the ability to advertise a global IPv6 address as the BGP next hop. When the advertising system is directly attached, it may include both a global and an IPv6 link-local address or only a global address, when next hop length is set to 16 bytes.

This document updates the specification to clarify the encoding of the BGP next hop when the advertising system is directly attached and only an IPv6 link-local address is available.

This clarification applies specifically to IPv6 link-local addresses and does not pertain to IPv4 link-local addresses as defined in [RFC3927].

Status of This Memo

This Internet-Draft is submitted in full conformance with the provisions of BCP 78 and BCP 79.

This Internet-Draft will expire on 1 December 2025.

Table of Contents

1. Introduction

BGP [RFC4271], was originally designed to provide reachability between domains and between the edges of a domain.

To support IPv6 reachability, BGP relies heavily on its Multiprotocol Extensions as defined in [RFC2545], which is crucial for enabling BGP-4 to advertise IPv6 routes alongside IPv4. This extension not only permits the exchange of IPv6 routing information but also establishes the structure for IPv6 next hop handling within BGP updates. As such, BGP assumes the next hop towards any reachable destination may not reside on the advertising speaker, but rather may either be through a router connected to the same subnet as the speaker, or through a router only reachable by traversing multiple hops through the network. [RFC2545] introduced the ability to advertise a global IPv6 address as the BGP next hop. When the advertising system is directly attached, it may include both a global and an IPv6 link-local address or only a global address, when next hop length is set to 16 bytes.

This document updates the specification to clarify the encoding of the BGP next hop when the advertising system is directly attached and only an IPv6 link-local address is available.

This clarification applies specifically to IPv6 link-local addresses and does not pertain to IPv4 link-local addresses as defined in [RFC3927].

BGP speakers are now often deployed on point-to-point links in networks where multihop reachability of any kind is not assumed or desired (all next hops are assumed to be the speaker reachable through a directly connected point-to-point link). This is common, for instance, in data center fabrics [RFC7938]. In these situations, a global IPv6 address is not required for the advertisement of reachability information; in fact, providing global IPv6 addresses in these kinds of networks can be detrimental to Zero Touch Provisioning (ZTP).

Such BGP deployment models require BGP to run on each link, and any ease or simplification of BGP configuration can result in simplifying orchestration and configuration management. This proposal is a step in that direction.

With this new capability, the need for a global unicast address assigned to the interfaces is eliminated.

Since IPv6 link-local addresses are not required to be globally unique, implementations must ensure that they are strictly associated with a specific interface.

2. Link-Local Next Hop Capability

The Link-Local Next Hop capability is a new BGP capability. A BGP speaker that supports capabilities advertisement [RFC5492] in an OPEN message should send this capability only when:

1. It is capable of sending IPv6 link-local address as the only next hop address for a route.

2. The implementation is capable of processing IPv6 link-local address next hops with the help of peer interface binding to come up with interface-specific next hops for its routing table.

The presence of this capability does not affect the support of global IPv6 only (16 bytes next hop) and global IPv6 combined with IPv6 link-local (32 bytes next hop), which should continue to be supported as before. The Capability Code for this capability is 77. The Capability Length field of this capability is 0.

The advantage of using this capability is that in contrast to the current situation. it can let two conforming implementations interoperate correctly without additional configuration. Existing implementations utilizing BGP next hop over an IPv6 link-local address are inconsistent, and can’t readily change their behavior without negative side effects.

A BGP speaker that is willing to use (send and receive) only IPv6 link-local addresses as next hops with a peer SHOULD advertise the Link-Local Next Hop Capability to the peer using BGP Capabilities advertisement.

The peers have the flexibility to include both IPv6 link-local and global next hops or IPv6 link-local only next hop.

In this document, all procedures described are applicable only when the capability described herein has been successfully negotiated between BGP speakers. When the capability has not been negotiated, the procedures in this document do not apply, and the resulting behavior is considered undefined and out of scope for this specification.

For example, when the capability is negotiated, a BGP speaker MUST advertise only an IPv6 link-local nexthop when using such nexthops, and the next hop length field MUST be set to 16 bytes.

Implementers are encouraged to consult the Appendix for currently known interoperability concerns or incompatibilities when this capability is absent or inconsistently implemented.

3. Changes to BGP Next Hop Attribute to Support Link-Local on Point-to-Point

[RFC2545], section 2, notes IPv6 link-local addresses are not generally suitable for use in the Next Hop field of the MP_REACH_NLRI. In order to support the many uses of IPv6 link-local addresses, however, [RFC2545] constructs the Next Hop field in IPv6 route advertisements by setting the length of the field to 32, and including both an IPv6 link-local and global address in the resulting enlarged field. In this way, the receiving BGP speaker can use the global IPv6 address to build local forwarding information, and the IPv6 link-local address for ICMPv6 redirects, etc. [RFC2545] does not, however, provide an explanation for situations where there is only an IPv6 link-local address in the Next Hop field of the MP_REACH_NLRI. The result is each implementation that supports IPv6 link-local peering along with forwarding to an IPv6 link-local address has implemented the construction of the Next Hop field in the MP_REACH_NLRI when there is only an IPv6 link-local address available in slightly different ways.

If an implementation intends to send a single IPv6 link-local forwarding address in the Next Hop field of the MP_REACH_NLRI, it MUST set the length of the Next Hop field to 16 and include only the IPv6 link-local address in the Next Hop field.

If an implementation intends to send both an IPv6 link-local and global forwarding address in the Next Hop field of the MP_REACH_NLRI, it MUST set the length of the Next Hop field to 32 and include both the IPv6 link-local and global forwarding addresses in the Next Hop field. If both the IPv6 link-local and global forwarding addresses are carried in the Next Hop Field, the speaker SHOULD provide a local configuration option to determine which address should be preferred for forwarding.

For iBGP peers configured as a route-reflector [RFC4456], when route-reflector isn’t configured to be in the data-path, the proposed IPv6 link-local (only) next hops MUST NOT be reflected.

A single (only) IPv6 link-local next hop address needs to always be reset as next hop self when passed to another link.

4. Receiver Processing of IPv6 Link-Local Forwarding Addresses

On receiving an MP_REACH_NLRI with a Next Hop length of 16, implementations SHOULD form the forwarding information using the IPv6 next hop contained in the Next Hop field, regardless of whether it is a link-local or globally reachable IPv6 address.

5. Error handling

A BGP speaker receiving an MP_REACH_NLRI with the length of the Next Hop Field set to 32, where the update contains anything other than an IPv6 link-local address and a global address, SHOULD consider this a malformed UPDATE message, and proceed as described in the following paragraphs. In order to support backward compatibility with existing implementations, an implementation MAY ignore a second IPv6 link-local address or 0::0/0 included with an IPv6 link-local address when the length of the Next Hop Field is set to 32; in this case, the implementation SHOULD report the existence of this additional information so the operator can correct the sending BGP implementation.

And if a BGP speaker receiving an MP_REACH_NLRI with the length of the Next Hop Field set to 16, where the update contains an IPv6 link-local address, SHOULD consider this a malformed UPDATE message, and handle the malformed UPDATE message using the approach of "treat-as-withdraw", as described in section 7.3 of [RFC7606].

If the Next Hop field is malformed, the implementation MUST handle the malformed UPDATE message using the approach of "treat-as-withdraw", as described in section 7.3 of [RFC7606].

If the Next Hop field is properly formed, but the IPv6 link-local next hop is not reachable (as determined by an examination of the IPv6 neighbor table), the implementation MAY handle the malformed UPDATE message using the approach of "treat-as-withdraw", as described in section 7.3 of [RFC7606] (see the note above on checking the local neighbor table for the correctness of the next hop).

6. Acknowledgements

The authors would like to thank Vipin Kumar, Dinesh Dutt, Donald Sharp, Jeff Haas, and Brian Carpenter for their contributions to this draft.

This document builds on prior work exploring the use of IPv6 link-local addresses as BGP next hops. Notably, [I-D.kumar-idr-link-local-nexthop] and [I-D.kato-bgp-ipv6-link-local] identified operational limitations and proposed mechanisms to enable

link-local next hop propagation in BGP. These drafts laid the groundwork for defining a standardized capability-based approach, as presented in this document, to ensure interoperable signaling and safe deployment of link-local next hops across BGP sessions.

7. IANA Considerations

IANA has assigned capability number 77 for the Link-Local Next Hop Capability described in this document. This registration is in the BGP Capability Codes registry.

Value	Description
77	Link-Local Next Hop Capability

Table 1: Link-Local Next Hop Capability

8. Security Considerations

The mechanism described in this draft can be used as a component of ZTP for building BGP peering across point-to-point links. This method, then, can be used by an attacker to form a peering session with a BGP speaker, ultimately advertising incorrect routing information into a routing domain in order to misdirect traffic or cause a denial of service. By using IPv6 link-local addresses, the attacker would be able to forego the use of a valid IPv6 address within the domain, making such an attack easier.

Operators SHOULD carefully consider security when deploying link-local addresses for BGP peering. Operators SHOULD filter traffic on links where BGP peering is not intended to occur to prevent speakers from accepting BGP session requests, as well as other mechanisms described in [RFC7454].

Operators MAY also use some form of cryptographic validation on links within the network to prevent unauthorized devices from forming BGP peering sessions. Authentication, such as the TCP authentication [RFC5925], may provide some relief if it is present and correctly configured. However, the distribution and management of keys in an environment where global addresses on BGP speakers are not present may be challenging.

Operators also MAY instruct a BGP peer which has received an UPDATE with an unreachable NEXT_HOP to disable the peering session over which the invalid NEXT_HOP was received pending manual intervention.

9. References

9.1. Normative References

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

[RFC3927] Cheshire, S., Aboba, B., and E. Guttman, "Dynamic Configuration of IPv4 Link-Local Addresses", RFC 3927, DOI 10.17487/RFC3927, May 2005, <https://www.rfc-editor.org/info/rfc3927>.

[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>.

[RFC4291] Hinden, R. and S. Deering, "IP Version 6 Addressing Architecture", RFC 4291, DOI 10.17487/RFC4291, February 2006, <https://www.rfc-editor.org/info/rfc4291>.

[RFC4456] Bates, T., Chen, E., and R. Chandra, "BGP Route Reflection: An Alternative to Full Mesh Internal BGP (IBGP)", RFC 4456, DOI 10.17487/RFC4456, April 2006, <https://www.rfc-editor.org/info/rfc4456>.

[RFC5309] Shen, N., Ed. and A. Zinin, Ed., "Point-to-Point Operation over LAN in Link State Routing Protocols", RFC 5309, DOI 10.17487/RFC5309, October 2008, <https://www.rfc-editor.org/info/rfc5309>.

[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>.

[RFC5925] Touch, J., Mankin, A., and R. Bonica, "The TCP Authentication Option", RFC 5925, DOI 10.17487/RFC5925, June 2010, <https://www.rfc-editor.org/info/rfc5925>.

[RFC7454] Durand, J., Pepelnjak, I., and G. Doering, "BGP Operations and Security", BCP 194, RFC 7454, DOI 10.17487/RFC7454, February 2015, <https://www.rfc-editor.org/info/rfc7454>.

[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>.

[RFC7938] Lapukhov, P., Premji, A., and J. Mitchell, Ed., "Use of BGP for Routing in Large-Scale Data Centers", RFC 7938, DOI 10.17487/RFC7938, August 2016, <https://www.rfc-editor.org/info/rfc7938>.

[RFC7942] Sheffer, Y. and A. Farrel, "Improving Awareness of Running Code: The Implementation Status Section", BCP 205, RFC 7942, DOI 10.17487/RFC7942, July 2016, <https://www.rfc-editor.org/info/rfc7942>.

[RFC8950] Litkowski, S., Agrawal, S., Ananthamurthy, K., and K. Patel, "Advertising IPv4 Network Layer Reachability Information (NLRI) with an IPv6 Next Hop", RFC 8950, DOI 10.17487/RFC8950, November 2020, <https://www.rfc-editor.org/info/rfc8950>.

9.2. Informative References

[I-D.kato-bgp-ipv6-link-local] 明洋, K. and B. Manning, "BGP4+ Peering Using IPv6 Link-local Address", Work in Progress, Internet-Draft, draft-kato-bgp-ipv6-link-local-00, 24 September 2001, <https://datatracker.ietf.org/doc/html/draft-kato-bgp-ipv6-link-local-00>.

[I-D.kumar-idr-link-local-nexthop] Kumar, V., Mohapatra, P., Dutt, D., and M. Valentine, "BGP Link-Local Next Hop Capability", Work in Progress, Internet-Draft, draft-kumar-idr-link-local-nexthop-02, 13 November 2014, <https://datatracker.ietf.org/doc/html/draft-kumar-idr-link-local-nexthop-02>.

Appendix A. Inconsistency Reports

According to [RFC7942], "This will allow reviewers and working groups to assign due consideration to documents that have the benefit of running code, which may serve as evidence of valuable experimentation and feedback that have made the implemented protocols more mature".

FRRouting (https://github.com/frrouting/frr/commit/606fdbb1fab98bac305dca3d19eb38b140b7c3e6) IPv6 next-hop handling when GUA/LL is set to ::/LL.

Bird (https://gitlab.nic.cz/labs/bird/-/commit/17de3a023f7bde293892b41bfafe5740c8553fc8) handling LL/LL case.

Appendix B. Implementation Report

According to [RFC7942], "This will allow reviewers and working groups to assign due consideration to documents that have the benefit of running code, which may serve as evidence of valuable experimentation and feedback that have made the implemented protocols more mature".

FRRouting (https://github.com/FRRouting/frr/pull/17871) implementation.

Authors' Addresses

Russ White
Akamai Technologies
Email: russ@riw.us

Jeff Tantsura
Nvidia
Email: jefftant.ietf@gmail.com

Donatas Abraitis
Hostinger
Email: donatas.abraitis@gmail.com