

Interdomain Routing
Internet-Draft
Intended status: Standards Track
Expires: 11 May 2022

R.W. White
Juniper Networks
D.S. Sharp
Cumulus Networks
D.D. Dutt
Stardust Consulting
B.S. Sadhu
VMWare
J.T. Tantsura
Apstra, Inc.
D.A. Abraitis
Hostinger
November 2021

Link Local Next Hop Handling for BGP
draft-white-linklocalnh-02

Abstract

BGP, described in [[RFC4271](#)], was originally designed to provide reachability between domains and between the edges of a domain. 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. Because of this, BGP does not recognize the use of IPv6 link local addresses, as described in [[RFC4291](#)], as a valid next hop for forwarding purposes.

However, 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. 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).

This draft standardizes the operation of BGP over a point-to-point link using link local IPv6 addressing only.

Status of This Memo

This Internet-Draft is submitted in full conformance with the provisions of [BCP 78](#) and [BCP 79](#).

Internet-Draft Link Local Next Hop Handling for BGP November 2021

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 5 May 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 Simplified BSD License text as described in Section 4.e of the [Trust Legal Provisions](#) and are provided without warranty as described in the Simplified BSD License.

Table of Contents

1.	Introduction	3
2.	Changes to BGP Next Hop Attribute to Support Link Local on Point-to-Point	3
3.	Receiver Processing of IPv6 Link Local Forwarding Addresses	4
4.	Error handling	4
5.	Acknowledgements	5
6.	IANA Considerations	5
7.	Security Considerations	5
8.	References	6
8.1.	Normative References	6
8.2.	Informative References	7
	Authors' Addresses	7

1. Introduction

BGP, described in [\[RFC4271\]](#), was originally designed to provide reachability between domains and between the edges of a domain. 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. Because of this, BGP does not recognize the use of IPv6 link local addresses, as described in [\[RFC4271\]](#), as a valid next hop for forwarding purposes.

However, 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. 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).

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

[\[RFC2545\]](#), [section 2](#), notes link local IPv6 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 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 a link local and global IPv6 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 link local address for ICMPv6 redirects, etc. [\[RFC2545\]](#) does not, however, provide an explanation for situations where there is only a link local IPv6 address in the Next Hop field of the MP_REACH_NLRI. The result is each implementation that supports link local peering

along with forwarding to a link local address has implemented the construction of the Next Hop field in the MP_REACH_NLRI when there is only a link local address available in slightly different ways.

If an implementation intends to send a single 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 a link local and global IPv6 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 IPv6 forwarding addresses in the Next Hop field. If both link local and global IPv6 forwarding addresses are carried in the Next Hop Field, the speaker SHOULD provide a local configuration option to determine which address is preferred for forwarding.

[3.](#) 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.

Implementations MAY check the validity of any IPv6 link local address used to calculate forwarding information by insuring the address is in the local neighbor table for the interface on which the BGP update was received (or through which the BGP speaker from which the update was received is reachable). There MUST be a configuration option to enable/disable this check.

Note: It is possible that checking the IPv6 neighbor table for the existence or validity of a link local next hop may make instances where a link is being overwhelmed through some form of Denial of Service (DoS) attack worse than they would otherwise be. If the IPv6 neighbor cache is overrun in a way that causes the link local address being used for BGP peering to be removed from the table, which is possible through an on-link DoS attack, any fresh BGP update will cause the entire peering session to fail if the implementation is checking the validity of link local next hops as described above.

Operators should carefully assess the use of validation against the local IPv6 neighbor table to determine if it is appropriate for any particular peering session.

4. 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 a link local IPv6 address and a global IPv6 address, SHOULD consider this a malformed UPDATE message, and proceed as described in the following paragraphs. In order to support backwards compatibility with existing implementations, an implementation MAY ignore a second link local IPv6 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.

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 \[RFC7606\]](#). It MAY send a NOTIFICATION message as described in [section 4 of \[RFC4271\]](#), using the UPDATE error message code (8 - Invalid NEXT_HOP Attribute) indicating there is an invalid NEXT_HOP field

If the Next Hop field is properly formed, but the 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 note above on checking the local neighbor table for the correctness of the next hop). The implementation MAY send a NOTIFICATION message as described in [section 4 of \[RFC4271\]](#) using the UPDATE error message code (TBA), indicating a link local address was included in the MP_REACH_NLRI, but the link local address included cannot be reached. As this could indicate a security breach of some type (see the security considerations section below), the operator SHOULD have a local configuration option to terminate the peering session until manual intervention is initiated.

5. Acknowledgements

The authors would like to thank Don Slice, Jeff Haas, and John Scudder for their contributions to this draft.

6. IANA Considerations

This memo requests IANA assign a number from the "Error Subcodes" registry defined in the IANA Considerations section in [[RFC4271](#)]. This allocation will be for a new UPDATE error subcode, code (TBA), with a value of "Unreachable Link Local Address."

7. Security Considerations

The mechanism described in this draft can be used as a component of ZTP for building BGP adjacencies 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 link local IPv6 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 described in [[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 are not present on BGP speakers 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.

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>>.
- [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>>.
- [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>>.
- [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>>.

8.2. Informative References

- [RFC2629] Rose, M., "Writing I-Ds and RFCs using XML", [RFC 2629](#),

DOI 10.17487/RFC2629, June 1999,
<<https://www.rfc-editor.org/info/rfc2629>>.

[RFC3552] Rescorla, E. and B. Korver, "Guidelines for Writing RFC Text on Security Considerations", [BCP 72](#), [RFC 3552](#), DOI 10.17487/RFC3552, July 2003,
<<https://www.rfc-editor.org/info/rfc3552>>.

Authors' Addresses

Russ White
Juniper Networks

Email: russ@riw.us

Donald Sharp
Cumulus Networks

Email: sharpd@cumulusnetworks.com

Dinesh Dutt
Stardust Consulting

Email: ddutt@hobbesdutt.com

Biswajit Sadhu
VMWare

Email: biswajit.sadhu@gmail.com

Jeff Tantsura
Apstra, Inc.

Email: jefftant.ietf@gmail.com

Hostinger

Email: donatas.abraitis@gmail.com