Network Working Group
Internet-Draft
Intended status: Standards Track
Expires: May 17, 2015

V. Kumar
Cumulus Networks
P. Mohapatra
Sproute Networks
D. Dutt
Cumulus Networks
M. Valentine
Goldman Sachs
November 13, 2014

BGP Link-Local Next Hop Capability
draft-kumar-idr-link-local-nexthop-02.txt

Abstract

This document proposes a new BGP capability to allow route resolution over IPv6 link-local next hop. It eliminates the requirement of assigning a global IPv6 address for the next hop.

Status of This Memo

This Internet-Draft will expire on May 17, 2015.

Copyright Notice

Table of Contents

1. Introduction

BGP [RFC4271] implementations support peering over link-local IPv6 addresses [RFC4291]. However, for the prefixes advertised over such a peering the resulting next hop attribute and route installation is still dependent on the Next Hop carrying a global IPv6 address. For the deployments where next hops need not have a scope beyond the peering link, the configuration can be simplified by lifting the requirement that the Next Hop field carry a global IPv6 address.

While the current proposal has no dependency on the link-local peering (e.g. link-local next hops could be used over ipv4 peering too), the use case with link-local peering offers clear advantages. Link-local peering already mandates an interface to be attached explicitly with the neighbor configuration. With the negotiation of the proposed capability, a BGP speaker sends link-local addresses as

the only IPv6 next hop address. Correspondingly, the receiving peer resolves the routes in the context of the peering interface.

Many large modern data-center networks that are based on topologies such as CLOS tend to be rather symmetric, and the BGP deployment in such networks do not require next hops to have relevance across peerings. 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 the requirement of any global interface address being removed by this new capability, BGP neighbor configuration can be further simplified by making it (look) address-family independent. E.g BGP can just take interface name for the peer config and link-local IPv6 address of the peer can be learned via a discovery protocol running on the link or by an out-of-band tool. In essence, link-local next hop in combination with [RFC5549] makes it possible to achieve an unnumbered interface-like solution [RFC5309] in BGP.

1.1. 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].

2. Link-Local Next Hop Capability

The LINK-LOCAL-ONLY-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 link-local IPv6 address as the only next hop address for a route.

2. The implementation is capable of processing 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 link-local IPv6 (32 bytes next hop), which should continue to be supported as before.

The Capability Code for this capability is specified in the IANA Considerations section of this document. The Capability Length field of this capability is 0.

3. Constructing the Next Hop field

Section 3 of [RFC2545] standardizes IPv6 next-hop construction. Here we suggest modifications required for link-local next hop construction.

A BGP speaker shall advertise to its peer in the Network Address of Next Hop field the link-local IPv6 address of the next hop.

The value of the Length of Next Hop Network Address field on a MP_REACH_NLRI attribute shall be set to 16.

For iBGP peers configured as a route-reflector, when route-reflector isn't configured to be in the data-path, the proposed link-local (only) next hops MUST not be reflected.

In general, implementations should not relay the link-local only next hop. Implementations supporting this capability should provide a way to handle the relay of link-local only next hops over point-to-point links (route-reflector and EBGP-to-IBGP cases) by either:

- an implicit next-hop-self.
- providing a configuration to enable next-hop-self. In this case, the link-local next hop MUST not be relayed, if this knob is not enabled.

Note: On a route-reflector, when source of link-local only next hop and route-reflector client are on the same broadcast segment, then implicit next-hop-self should not be done. Same goes for eBGP to iBGP scenarios.

4. Operation

A BGP speaker that is willing to use (send and receive) only link-local addresses as next hops with a peer SHOULD advertise the LINK-LOCAL-ONLY-NEXT-HOP Capability to the peer using BGP Capabilities advertisement.

[draft-kato] recommended implementations to ignore the ipv6 global next hop if it didn't match any of the link's global addresses. The proposal has the following limitations:

- It results in poor error handling, specifically for next hop validation.
- It does not allow the sender to set a global next hop value that is _not_ one of the assigned prefixes on the link.

- It does not specify the behavior for IBGP sessions.
- A global next hop field has to be always present in the UPDATE messages.

We formalize this idea with the proposed new capability, so that the peers have the flexibility to include both link-local and global next hops or link-local only next hop. The error handling of messages is not compromised.

5. Deployment Considerations

The usage of this capability is restricted to the cases where the scope of the next hop is limited to the peering interface. This restriction comes from the fact that link-local IPv6 addresses are link-scoped, therefore link-local address of the one peer can not be used as next hop if its to be carried with the updates over another peer.

6. Acknowledgments

We would like to thank Daniel Walton for his comments and suggestions.

7. IANA Considerations

This document defines a new link-local next hop capability. IANA is requested to assign a capability number to the same.

8. Security Considerations

There are no additional security risks introduced by this design.

9. References

9.1. Normative References

[RFC2119] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, March 1997.

[RFC2545] Marques, P. and F. Dupont, "Use of BGP-4 Multiprotocol Extensions for IPv6 Inter-Domain Routing", RFC 2545, March 1999.

[RFC4271] Rekhter, Y., Li, T., and S. Hares, "A Border Gateway Protocol 4 (BGP-4)", RFC 4271, January 2006.

[RFC4291] Hinden, R. and S. Deering, "IP Version 6 Addressing Architecture", RFC 4291, February 2006.

[RFC5309] Shen, N. and A. Zinin, "Point-to-Point Operation over LAN in Link State Routing Protocols", RFC 5309, October 2008.

[RFC5492] Scudder, J. and R. Chandra, "Capabilities Advertisement with BGP-4", RFC 5492, February 2009.

[RFC5549] Le Faucheur, F. and E. Rosen, "Advertising IPv4 Network Layer Reachability Information with an IPv6 Next Hop", RFC 5549, May 2009.

9.2. Informational References

[draft-kato] "http://tools.ietf.org/html/draft-kato-bgp-ipv6-link-local-00", September 2001.

Authors' Addresses

Vipin Kumar
Cumulus Networks
185 E. Dana Street
Mountain View, CA 94041
USA

Email: vipin@cumulusnetworks.com

Pradosh Mohapatra
Sproute Networks

Email: mpradosh@yahoo.com

Dinesh Dutt
Cumulus Networks
185 E. Dana Street
Mountain View, CA 94041
USA

Email: ddutt@cumulusnetworks.com

Mike Valentine
Goldman Sachs
30 Hudson St
Jersey City, NY 07302
USA

Email: michael.j.valentine@gs.com