Network Working Group
Internet-Draft
Intended status: Standards Track
Expires: January 5, 2015

Z. Li
L. Zhang
S. Hares
Huawei Technologies

July 4, 2014

NEXTHOP_PATH_RECORD ATTIBUTE for BGP
draft-zhang-idr-nexthop-path-record-00

Abstract

As the BGP is deployed in a single Autonomous System using converged networks such as Seamless MPLS, it is desirable for BGP to carry more IGP nexthop pathway information to help select routing more intelligently. One example of a Seamless MPLS deployment is the Mobile BackHaul (MBH) deployment with multiple IGPs Areas per ASN. This document describes a new optional transitive path attribute, NEXTHOP_PATH_RECORD ATTRIBUTE for BGP that records the next hop path which can be used by BGP network management to monitor and manage the BGP infrastructure via management interfaces (such as I2RS).

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].

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 http://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 January 5, 2015.

Copyright Notice

Copyright (c) 2014 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 (http://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

Network topologies have become more densely interconnected at the network level. Examples of this mesh topologies can be a data center or the converged carrier network that [I-D.ietf-mpls-seamless-mpls] architecture supports. In these mesh topologies, there may be multiple highly meshed IGPs connected by IBGP into a single AS. Scalability and redundancy may require exterior processes to calculate a better way through the array of next-hop pathways attached to BGP Routes of any AFI/SAFI pairing.

This document proposes a new path attribute that can record the next hop path of the route to help BGP route election and network management. In the deployment considerations section, this draft provides a deployment scenario linked to the use of I2RS and BGP Cost Community attribute.

2. Definition of NEXTHOP_PATH_RECORD ATTRIBUTE

The NEXTHOP_PATH_RECORD ATTRIBUTE is an optional transitive BGP Path Attribute. The NEXTHOP_PATH_RECORD ATTRIBUTE type is defined as below (refer to [RFC4271]):

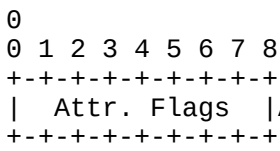

Figure 2 NEXTHOP_PATH_RECORD Type definition

Attr. Flags

SHOULD be optional transitive

Attr. Type Code

SHOULD be allocated by IANA

NEXTHOP_PATH_RECORD is composed of a sequence of next hop path segments. Each next hop path segment is represented by a triple <path segment type, path segment length, path segment value>. The format of the next hop path segment is shown in the figure 3.

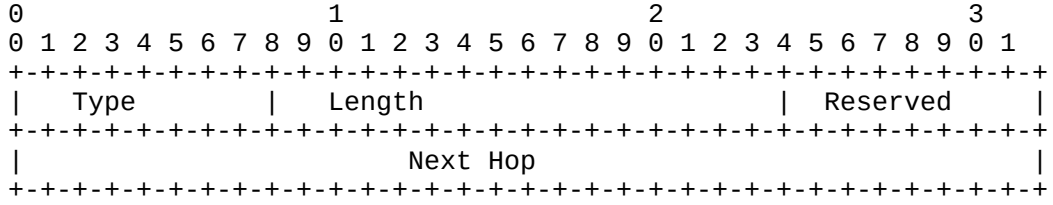

Figure 3 NH_SEQUENCE_V4 TLV

- Type: A single octet encoding the TLV Type. The Type of "NH_SEQUENCE_V4" is defined in this document, which needs to be allocated by IANA. The procedure for next hop path segment usage for IPv6 or other extensions will be described in the future revisions of this document.

Length: Two octets encoding the length in octets of the TLV, including the type and length fields. The length is encoded as an unsigned binary integer.

Reserved: A single octet that must be zero now.

NextHop: four octets encoding for the route next hop address.

3. Process of NEXTHOP_PATH_RECORD ATTRIBUTE

The NEXTHOP_PATH_RECORD attribute defined in this section is an optional transitive BGP Path Attribute as described in [RFC4271].

3.1. Creating and Modifying the NEXTHOP_PATH_RECORD Attribute

When a BGP speaker distributes a route to its BGP peer within UPDATE message, the NEXTHOP_PATH_RECORD ATTRIBUTE should be processed based on different route states:

1. If the route is originated in this BGP speaker

    * If the NEXTHOP_PATH_RECORD ATTRIBUTE is supported, the NEXTHOP_PATH_RECORD ATTRIBUTE SHOULD be originated including the BGP speaker's own next hop address in a next hop path segment. In this case, the next hop address of the originating BGP speaker will be the only entry of the next hop path segment, and this path segment will be the only segment in NEXTHOP_PATH_RECORD ATTRIBUTE.

    * If the NEXTHOP_PATH_RECORD ATTRIBUTE is not supported, the route will be distributed without NEXTHOP_PATH_RECORD ATTRIBUTE.

2. if the route is received from one BGP speaker's UPDATE message

    * If the NEXTHOP_PATH_RECORD ATTRIBUTE is NULL and the local BGP speaker supports NEXTHOP_PATH_RECORD ATTRIBUTE, when the route is propagated to another IBGP speaker with next hop self (NHS), the NEXTHOP_PATH_RECORD ATTRIBUTE SHOULD be originated including the BGP speaker's own next hop address in a next hop path segment. In this case, the next hop address of this BGP speaker will be the only entry to the next hop path segment, and this path segment will be the only segment in NEXTHOP_PATH_RECORD ATTRIBUTE

    * If the NEXTHOP_PATH_RECORD ATTRIBUTE is non-NULL and the local BGP speaker support NEXTHOP_PATH_RECORD ATTRIBUTE, when the route is propagated to another IBGP speaker with next hop self (NHS), the BGP speaker MUST appends its own next hop address as the last one of the next hop path segments.

    * If the NEXTHOP_PATH_RECORD ATTRIBUTE is NULL and the local BGP speaker support NEXTHOP_PATH_RECORD ATTRIBUTE, when the route is propagated to another BGP speaker without changing the next

hop by the BGP speaker, the BGP speaker MUST NOT originate the NEXTHOP_PATH_RECORD ATTRIBUTE.

* If the NEXTHOP_PATH_RECORD ATTRIBUTE is non-NULL and the local BGP speaker support NEXTHOP_PATH_RECORD ATTRIBUTE, when the route is propagated to another BGP speaker without changing the next hop by the BGP speaker, the BGP speaker MUST NOT change the next hop path sequence.

* If the BGP speaker does not support NEXTHOP_PATH_RECORD ATTRIBUTE, it SHOULD keep the NEXTHOP_PATH_RECORD ATTRIBUTE unchanged whether the route is distribute with next hop self or not.

4. Deployment considerations

Two deployment examples are given to demonstrate how the nexthop information can be deployed in existing BGP technologies to monitor and tune the IBGP cloud to provide better operation. maintenance. This attribute has records information.

4.1. Customized Best Path Selection

The next_hop information gathered on an IBGP or EBP route could be used by off-line decision processing to select paths, and re-inserted as policy to affect the decision making via I2RS. The I2RS BGP use case draft [I-D.keyupate-i2rs-bgp-usecases] describes this its section on customized best path selection (section 4.1) which uses the BGP feature [I-D.ietf-idr-custom-decision] to set a custom decision community.

4.2. Use in Seamless MPLS case with PE-RR

In a Seamless MPLS network [I-D.ietf-mpls-seamless-mpls], the Area Border Routers (ABRs) which run IBGP may act RR-clients or be part of RR mesh as described in section 5.1.7. Seamless MPLS places restrictions on the BGP NEXT_HOP to make Seamless MPLS work in the general case. With the transmittal of the next-hop-path attribute offline calculation can insert a better pathway decision using the BGP customer. A sample description of in a seamless MPLS is included below.

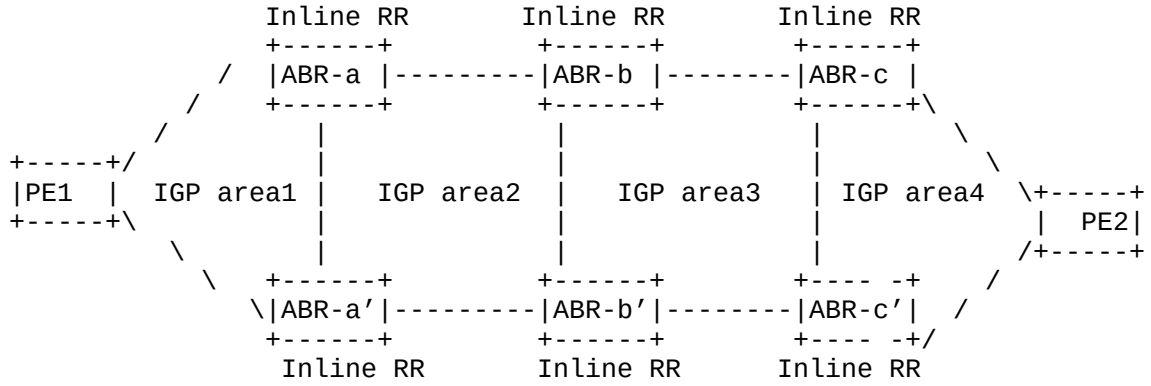

Figure 1 Seamless MPLS Network with Multiple IGP Areas

Just like Figure 1 shown, PE1 and PE2 are BGP VPN service end-point. IBGP peers runs contiguously between ABRs in different IGP areas, and each ABR works as inline RR. When labeled BGP routes or BGP VPN routes originated from PE1 is distributed to the other service end-point PE2, the route can be reflected by the ABRs one by one with next hop self (NHS).

The inline RR will distribute the route to all of the IBGP peers except the IBGP peer from which the route was received. As a result, an ABR may receive routes of the same prefix from different IBGP peers with different next hop. Traditionally the BGP RR should select the best route to reflect to other IBGP peers. But in this network the route selection process will be more complex which needs to introduce complex route policy.

The NEXTHOP_PATH ATTRIBUTE can optionally collect information on the pathway the routes are taking through the IBGP mesh. This information may aid in monitoring paths in this complex path or in offline processing that reduces complex policy to simple instantiation of community policy or the BGP Custom Cost community to allow specialized pathways through the MPLS mesh.

5. IANA Considerations

IANA need to assign the codepoint in the "BGP Path Attributes" registry to the NEXTHOP_PATH_RECORD ATTRIBUTE.

IANA shall create a registry for "next hop path segment". The type field consists of a single octet, with possible values from 0 to 255. The allocation policy for this field is to be "Standards Action with Early Allocation". A new Type should be defined as "NH_SEQUENCE_V4".

6. Security Considerations

Note that, the NEXTHOP_PATH_RECORD ATTRIBUTE is defined as a optional transitive BGP Path attribute. Both the IBGP and EBGP speaker can use this attribute. When an ASBR propagates the route receive from a IBGP peer to an EBGP peer, the NEXTHOP_PATH_RECORD ATTRIBUTE will be distribute to the EBGP Speaker which may be controlled by other Service Provider. If the EBGP speaker can support the NEXTHOP_PATH_RECORD ATTRIBUTE, it can parse the NEXTHOP_PATH_RECORD ATTRIBUTE to get the inner network architecture of the other network.

BGP requires the use of TCP-MD5 TCP-MD5 [RFC2385] or TCP-AO [RFC5925]). Use of encryption will prevent unauthorized view of the NEXTHOP_PATH_RECORD attribute. For those not supporting the required TCP-MD5 or TCP-AO, the NEXTHOP_PATH_RECORD ATTRIBUTE capability SHOULD disabled for specific BGP speaker to prevent this attack.

7. References

7.1. Normative References

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

[RFC2385] Heffernan, A., "Protection of BGP Sessions via the TCP MD5 Signature Option", RFC 2385, August 1998.

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

[RFC5925] Touch, J., Mankin, A., and R. Bonica, "The TCP Authentication Option", RFC 5925, June 2010.

7.2. Informative References

[I-D.ietf-idr-custom-decision] Retana, A. and R. White, "BGP Custom Decision Process", draft-ietf-idr-custom-decision-04 (work in progress), November 2013.

[I-D.ietf-mpls-seamless-mpls] Leymann, N., Decraene, B., Filsfils, C., Konstantynowicz, M., and D. Steinberg, "Seamless MPLS Architecture", draft-ietf-mpls-seamless-mpls-07 (work in progress), June 2014.

[I-D.keyupate-i2rs-bgp-usecases]
Patel, K., Fernando, R., Gredler, H., Amante, S., White, R., and S. Hares, "Use Cases for an Interface to BGP Protocol", draft-keyupate-i2rs-bgp-usecases-03 (work in progress), June 2014.

Authors' Addresses

Zhenbin Li
Huawei Technologies
Huawei Bld., No.156 Beiqing Rd.
Beijing 100095
China

Email: lizhenbin@huawei.com

Li Zhang
Huawei Technologies
Huawei Bld., No.156 Beiqing Rd.
Beijing 100095
China

Email: monica.zhangli@huawei.com

Susan Hares
Huawei Technologies
7453 Hickory Hill
Saline, MI 48176
USA

Email: shares@ndzh.com