

IDR
Internet-Draft
Intended status: Standards Track
Expires: January 1, 2015

G. Van de Velde
K. Patel
D. Rao
Cisco Systems
R. Raszuk
NTT MCL Inc.
R. Bush
Internet Initiative Japan
June 30, 2014

BGP Remote-Next-Hop
draft-vandavelde-idr-remote-next-hop-07

Abstract

The BGP Remote-Next-Hop is an optional transitive attribute intended to facilitate automatic tunneling across an AS on a per address family basis. The attribute carries one or more tunnel end-points for a NLRI. Additionally, tunnel encapsulation information is communicated to successfully setup these tunnels.

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 1, 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	2
2.	Requirements Language	3
3.	Tunnel Encapsulation attribute versus BGP Remote-Next-Hop attribute	3
4.	BGP Remote-Next-Hop attribute TLV Format	4
4.1.	Encapsulation sub-TLVs for virtual network overlays . . .	5
4.1.1.	Encapsulation sub-TLV for VXLAN	6
4.1.2.	Encapsulation sub-TLV for NVGRE	7
4.1.3.	Encapsulation sub-TLV for GTP	8
5.	Use Case scenarios	8
5.1.	Stateless user-plane architecture for virtualized EPC (vEPC)	8
5.2.	Stateless User-plane Architecture for virtual Packet Edge	9
5.3.	Multi-homing for IPv6	9
5.4.	Dynamic Network Overlay Infrastructure	10
5.5.	The Tunnel end-point is NOT the originating BGP speaker .	10
5.6.	Networks that do not support BGP Remote-Next-Hop attribute	10
5.7.	Networks that do support BGP Remote-Next-Hop attribute .	10
6.	BGP Remote-Next-Hop Community	10
7.	IANA Considerations	10
8.	Security Considerations	11
8.1.	Protecting the validity of the BGP Remote-Next-Hop attribute	11
9.	Privacy Considerations	11
10.	Acknowledgements	11
11.	Change Log	11
12.	References	12
12.1.	Normative References	12
12.2.	Informative References	12
	Authors' Addresses	13

[1.](#) Introduction

[RFC5512] defines an attribute attached to an NLRI to signal tunnel end-point encapsulation information between two BGP speakers for a single tunnel. It assumes that the exchanged tunnel endpoint is the NLRI.

This document defines a new BGP transitive attribute known as a Remote-Next-Hop BGP attribute for Intra-AS and Inter-AS usage which removes the assumption of both a single tunnel and that the exchanged NLRI is the tunnel endpoint.

The tunnel endpoint information and the tunnel encapsulation information is carried within a Remote-Next-Hop BGP attribute. This attribute can be added to any BGP NLRI. This way the Address Family (AF) of the NLRI exchanged is decoupled from the tunnel SAFI address-family defined in [[RFC5512](#)].

2. Requirements Language

The key words "MUST", "MUST NOT", "REQUIRED", "SHALL", "SHALL NOT", "SHOULD", "SHOULD NOT", "RECOMMENDED", "MAY", and "OPTIONAL" are to be interpreted as described in [[RFC2119](#)] only when they appear in all upper case. They may also appear in lower or mixed case as English words, without any normative meaning.

3. Tunnel Encapsulation attribute versus BGP Remote-Next-Hop attribute

The Tunnel Encapsulation attribute [[RFC5512](#)] is based on the principle that the tunnel end-point is the BGP speaker originating the update and is inserted as the NLRI in the exchange, with the consequence that it is impossible to set the endpoint to an arbitrary address. It is also assumed that there is only a single tunnel between endpoints.

There are use cases where it is desired that the tunnel end-point address should be a different address, or set of addresses, than the originating BGP speaker. It is also useful to be able to signal different encapsulation parameters for different prefixes with the same remote tunnel end-point. The BGP Remote-Next-Hop attribute provides the ability to have one or more different tunnel end-point addresses from IPv4, IPv6 and/or other address-families, and be able to signal next-hop encapsulation parameters along with any prefix.

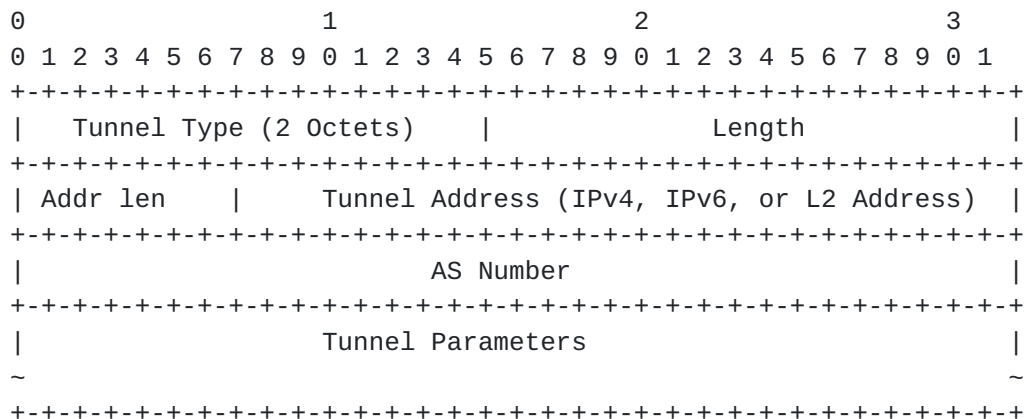
The sub-TLVs from the Tunnel Encapsulation Attribute [[RFC5512](#)] are reused for the BGP Next-Hop-Attribute.

Due to the intrinsic nature of both attributes, the tunnel encapsulation end-point assumes that the tunnel end-point is both the NLRI exchanged and the originating router, while the BGP Remote-Next-Hop attribute is inserted for an exchanged NLRI by adding a set of tunnel end-points and hence these two attributes are mutually exclusive.

4. BGP Remote-Next-Hop attribute TLV Format

This attribute is an optional transitive attribute [[RFC1771](#)].

The BGP Remote-Next-Hop attribute is composed of a set of Type-Length-Value (TLV) encodings. The type code of the attribute is (IANA to assign). Each TLV contains information corresponding to a particular tunnel technology and tunnel end-point address. The TLV is structured as follows:



Tunnel Type (2 octets): identifies the type of tunneling technology being signaled. This document specifies the following types:

- L2TPv3 over IP [[RFC3931](#)]: Tunnel Type = 1
- GRE [[RFC2784](#)]: Tunnel Type = 2
- IP in IP [[RFC2003](#)] [[RFC4213](#)]: Tunnel Type = 7

This document also defines the following types:

- VXLAN: Tunnel Type = 8
- NVGRE: Tunnel Type = 9
- GTP: Tunnel Type = 10
- MPLS-in-GRE: Tunnel Type = 11

Unknown types MUST be ignored and skipped upon receipt.

Length (2 octets): the total number of octets of the value field.

Tunnel Address Length - Addr len (1 octet): Length of Tunnel Address. Set to 4 bytes for an IPv4 address, 16 bytes for an IPv6 address or 8 bytes for a MAC address.

AS Number - The AS number originating the BGP Remote-Next-Hop attribute and is either a 2-byte AS or 4-Byte AS number

Tunnel Parameter - (variable): comprised of multiple sub-TLVs. Each sub-TLV consists of three fields: a 1-octet type, 1-octet length, and zero or more octets of value. The sub-TLV definitions and the sub-TLV data are described in depth in [[RFC5512](#)].

[4.1.](#) Encapsulation sub-TLVs for virtual network overlays

A VN-ID may need to be signaled along with the encapsulation types for DC overlay encapsulations such as [VXLAN] and [NVGRE]. The VN-ID when present in the encapsulation sub-TLV for an overlay

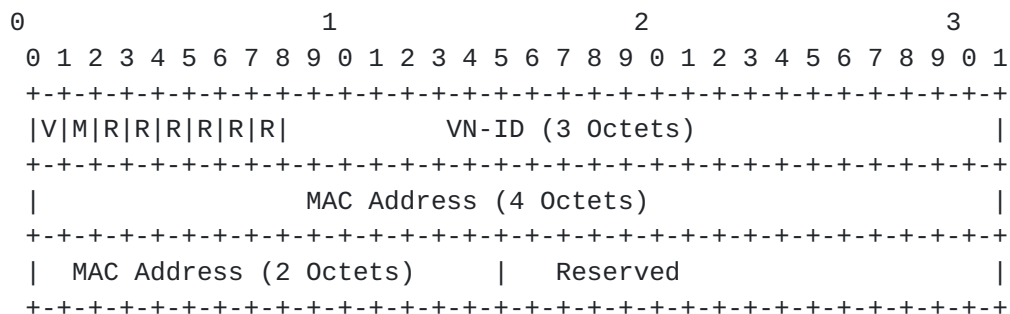
4.1.1. Encapsulation sub-TLV for VXLAN

[illegible]

The MAC address is local to the device advertising the route, and should be included as the destination MAC address in the inner Ethernet header immediately following the outer VXLAN header, in the packets destined to the advertiser.

4.1.2. Encapsulation sub-TLV for NVGRE

This document defines a new encapsulation sub-TLV format, defined in [RFC5512], for NVGRE tunnels. When the tunnel type is NVGRE, the following is the structure of the value field in the encapsulation sub-TLV:



V: When set to 1, it indicates that a valid VN-ID is present in the encapsulation sub-TLV.

M: When set to 1, it indicates that a valid MAC Address is present in the encapsulation sub-TLV.

R: The remaining bits in the 8-bit flags field are reserved for further use. They MUST be set to 0 on transmit and MUST be ignored on receipt.

VN-ID: Contains a 24-bit VN-ID value, if the 'V' flag bit is set. If the 'V' flag is not set, it SHOULD be set to zero and MUST be ignored on receipt.

The VN-ID value is filled in the VSID field in the NVGRE packet header as defined in [NVGRE].

MAC Address: Contains an Ethernet MAC address if the 'M' flag bit is set. If the 'M' flag is not set, it SHOULD set to all zeroes and MUST be ignored on receipt.

The MAC address is local to the device advertising the route, and should be included as the destination MAC address in the inner Ethernet header immediately following the outer NVGRE header, in the packets destined to the advertiser.

4.1.3. Encapsulation sub-TLV for GTP

This document defines a new encapsulation sub-TLV format, defined in [RFC5512], for GTP tunnels. When the tunnel type is GTP, the following is the structure of the value field in the encapsulation sub-TLV:

```

0      1      2      3
0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1
+--+--+--+--+--+--+--+--+--+--+--+--+--+--+--+--+--+--+--+--+--+
|                               Local TEID (4 Octets)                               |
+--+--+--+--+--+--+--+--+--+--+--+--+--+--+--+--+--+--+--+--+--+
|      Local Endpoint Address (4/16 Octets (IPv4/IPv6))      |
.
+--+--+--+--+--+--+--+--+--+--+--+--+--+--+--+--+--+--+--+--+--+

```

Local TEID: Contains a 32-bit Tunnel Endpoint Identifier of a GTP tunnel assigned by EPC that is used to distinguish different connections in received packets within the tunnel.

Local Endpoint Address: Indicates a 4-octets IPv4 address or 16-octets IPv6 address as a local endpoint address of GTP tunnel.

Local Endpoint Address element makes a tunnel endpoint router allow to have multiple Local TEID spaces. Received GTP packets are identified which tunnel connection by combination of Local Endpoint Address and Local TEID.

5. Use Case scenarios

This section provides a short overview of some use-cases for the BGP Remote-Next-Hop attribute. Use of the BGP Remote-Next-Hop is not limited to the examples in this section.

5.1. Stateless user-plane architecture for virtualized EPC (vEPC)

The full usage case of BGP remote-next-hop regarding vEPC can be found in [vEPC], while [RFC6459] documents IPv6 in 3GPP EPS.

3GPP introduces Evolved Packet Core (EPC) that is fully IP based mobile system for LTE and -advanced in their Release-8 specification and beyond. Operators are now deploying EPC for LTE services and encounter rapid LTE traffic growth. There are various activities to offload mobile traffic in 3GPP and IETF such as LIPA, SIPTO and DMM. The concept is similar that traffic of OTT (Over The Top) application

is offloaded at entity that is closer to the mobile node (ex. eNodeB or closer anchor).

5.2. Stateless User-plane Architecture for virtual Packet Edge

With the emergence of the NfV technologies, different architectures are proposed for virtualised Services. These functions will normally run in the datacenter. BGP remote-next-hop can be used to inject traffic into the virtualised services running in the datacenter for a optimized, simple and clean routing architecture. BGP Remote Next Hop can simplify the orchestration or provisioning layer by signalling the tunnel endpoint (virtual provider edge router) and the encapsulation protocol.

If this is used together with orchestrated traffic steering mechanisms (i.e. BGP Flowspec) , it is possible to differentiate at application level, and forward each different traffic types towards the desired destination.

5.3. Multi-homing for IPv6

When an end-user IPv6 network is multi-homed to the Internet, it may be assigned more than a single prefix originated by various upstream ASs. Each AS prefers to only announce a supernet of all its assigned IPv6 prefixes, unlike IPv4 where the AS announced the end-users assigned prefix. The goal of this BGP policy behaviour is to keep the number of entries in the IPv6 global BGP table to a minimum, it also it also results in well known resiliency improvements.

For example, if an end-user IPv6 is peering with 2 different Service providers AS1 and AS2. In this case the IPV6 end-user will have at least one prefix assigned from each of these service providers. The devices at the IPv6 end-user will each receive an address from these prefixes. The devices will in most cases, when building IPv6 sessions (TCP, etc...), do so with only a single IPv6 address. The decision which IPv6 address the device will use is documented in [[RFC3484](#)].

If one of the links between the end-user and one the neighboring AS's fails, a consequence will be that a set of sessions need to be reset, or that a section of the end-user network becomes unreachable.

With usage of the BGP-remote-Next-Hop attribute the service provider can tunnel that packet towards an alternate BGP Remote-Next-Hop at the end-users alternate provider and restore the network connectivity even though the local link towards the end-user is broken.

5.4. Dynamic Network Overlay Infrastructure

The BGP Remote-Next-Hop extension allows signaling tunnel encapsulations needed to build and dynamically create an overlay tunneled network with traffic isolation and virtual private networks.

5.5. The Tunnel end-point is NOT the originating BGP speaker

Note that, in each network environment, the originating router is the preferred tunnel end-point server. It may be that the network administrator has deployed an independent set of tunnel end-point servers across their network, which may or may not speak BGP. The BGP Remote-Next-Hop attribute provides the ability to signal this via BGP.

5.6. Networks that do not support BGP Remote-Next-Hop attribute

If a device does not support this attribute, and receives this attribute, then normal NLRI BGP forwarding is used as the attribute is optional and transitive.

5.7. Networks that do support BGP Remote-Next-Hop attribute

If a BGP speaker does understand this attribute, and receives this attribute, then the BGP speaker MAY, by configuration, skip use or not use the information within this attribute.

6. BGP Remote-Next-Hop Community

place-holder for an BGP extension to signal valid prefixes allowed to be considered as tunnel end-points. To be completed.

7. IANA Considerations

This document defines a new BGP attribute known as a BGP Remote-Next-Hop attribute. We request IANA to allocate a new attribute code from the "BGP Path Attributes" registry with a symbolic name "Remote-Next-Hop" attribute.

We also request IANA to allocate four new BGP Tunnel Types from the "BGP Tunnel Encapsulation Attribute Tunnel Types" registry with the following symbolic names: "VXLAN" with Tunnel type 8, "NVGRE" with Tunnel type 9, "GTP" with Tunnel type 10, and "MPLS-in-GRE" with Tunnel type 11.

8. Security Considerations

This technology could be used as technology as man in the middle attack, however with existing RPKI validation for BGP that risk is reduced.

The distribution of Tunnel end-point address information can result in potential DoS attacks if the information is sent by malicious organisations. Therefore is it strongly recommended to install traffic filters, IDSs and IPSS at the perimeter of the tunneled network infrastructure.

8.1. Protecting the validity of the BGP Remote-Next-Hop attribute

It is possible to inject a rogue BGP Remote-Next-Hop attribute to an NLRI resulting in Monkey-In-The-Middle attack (MITM). To avoid this type of MITM attack, it is strongly recommended to use a technology a mechanism to verify that for NLRI it is the expected BGP Remote-Next-Hop. We anticipate that this can be done with an expansion of RPKI-Based origin validation, see [[I-D.ietf-sidr-pfx-validate](#)].

This does not avoid the fact that rogue AS numbers may be inserted or injected into the AS-Path. To achieve protection against that threat BGP Path Validation should be used, see [[I-D.ietf-sidr-bgpsec-overview](#)].

9. Privacy Considerations

This proposal may introduce privacy issues, however with BGP security mechanisms in place they should be prevented.

10. Acknowledgements

The authors would like to thanks Satoru Matsushima, Ryuji Wakikawa and Miya Kohno for their usefull vEPC discussions. Istvan Kakonyi provided insight in the vPE use case scenario.

11. Change Log

Initial Version: 16 May 2012

Hacked for -01: 17 July 2012

Hacked for -05: 07 January 2014

Hacked for -07: 30 June 2014

12. References

12.1. Normative References

- [RFC1771] Rekhter, Y. and T. Li, "A Border Gateway Protocol 4 (BGP-4)", [RFC 1771](#), March 1995.
- [RFC2119] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", [BCP 14](#), [RFC 2119](#), March 1997.
- [RFC2784] Farinacci, D., Li, T., Hanks, S., Meyer, D., and P. Traina, "Generic Routing Encapsulation (GRE)", [RFC 2784](#), March 2000.
- [RFC3484] Draves, R., "Default Address Selection for Internet Protocol version 6 (IPv6)", [RFC 3484](#), February 2003.
- [RFC3931] Lau, J., Townsley, M., and I. Goyret, "Layer Two Tunneling Protocol - Version 3 (L2TPv3)", [RFC 3931](#), March 2005.
- [RFC4213] Nordmark, E. and R. Gilligan, "Basic Transition Mechanisms for IPv6 Hosts and Routers", [RFC 4213](#), October 2005.
- [RFC5512] Mohapatra, P. and E. Rosen, "The BGP Encapsulation Subsequent Address Family Identifier (SAFI) and the BGP Tunnel Encapsulation Attribute", [RFC 5512](#), April 2009.
- [RFC6459] Korhonen, J., Soininen, J., Patil, B., Savolainen, T., Bajko, G., and K. Iisakkila, "IPv6 in 3rd Generation Partnership Project (3GPP) Evolved Packet System (EPS)", [RFC 6459](#), January 2012.

12.2. Informative References

- [I-D.ietf-sidr-bgpsec-overview]
Lepinski, M. and S. Turner, "An Overview of BGPSEC", [draft-ietf-sidr-bgpsec-overview-04](#) (work in progress), December 2013.
- [I-D.ietf-sidr-pfx-validate]
Mohapatra, P., Scudder, J., Ward, D., Bush, R., and R. Austein, "BGP Prefix Origin Validation", [draft-ietf-sidr-pfx-validate-10](#) (work in progress), October 2012.

[I-D.mahalingam-dutt-dcops-vxlan]

Mahalingam, M., Dutt, D., Duda, K., Agarwal, P., Kreeger, L., Sridhar, T., Bursell, M., and C. Wright, "VXLAN: A Framework for Overlaying Virtualized Layer 2 Networks over Layer 3 Networks", [draft-mahalingam-dutt-dcops-vxlan-02](#) (work in progress), August 2012.

[I-D.matsushima-stateless-uplane-vepc]

Matsushima, S. and R. Wakikawa, "Stateless user-plane architecture for virtualized EPC (VEPC)", [draft-matsushima-stateless-uplane-vepc-01](#) (work in progress), July 2013.

[I-D.sridharan-virtualization-nvgre]

Sridharan, M., Greenberg, A., Venkataramaiah, N., Wang, Y., Duda, K., Ganga, I., Lin, G., Pearson, M., Thaler, P., and C. Tumuluri, "NVGRE: Network Virtualization using Generic Routing Encapsulation", [draft-sridharan-virtualization-nvgre-02](#) (work in progress), February 2013.

Authors' Addresses

Gunter Van de Velde
Cisco Systems
De Kleetlaan 6a
Diegem 1831
Belgium

Phone: +32 2704 5473
Email: gvandeve@cisco.com

Keyur Patel
Cisco Systems
170 W. Tasman Drive
San Jose, CA 95124 95134
USA

Email: keyupate@cisco.com

Dhananjaya Rao
Cisco Systems
170 W. Tasman Drive
San Jose, CA 95124 95134
USA

Email: dhrao@cisco.com

Robert Raszuk
NTT MCL Inc.
101 S Ellsworth Avenue Suite 350
San Mateo, CA 94401
US

Email: robert@raszuk.net

Randy Bush
Internet Initiative Japan
5147 Crystal Springs
Bainbridge Island, Washington 98110
US

Email: randy@psg.com

