Internet Engineering Task Force
Internet-Draft
Intended status: Standards Track
Expires: April 19, 2013

D. Rao
J. Mullooly
R. Fernando
Cisco

October 16, 2012

Layer-3 virtual network overlays based on BGP Layer-3 VPNs
draft-drao-bgp-l3vpn-virtual-network-overlays-00

Abstract

Virtual network overlays are being designed and deployed in various types of networks, including data center networks. These network overlays serve several purposes including flexible network virtualization, increased scale, multi-tenancy, and mobility. Such overlay networks may be used to provide both Layer-2 and Layer-3 network services to hosts at the network edge. New encapsulations are being defined and standardized to support these virtual networks. These encapsulations are primarily based on IP, such as VxLAN and NvGRE.

BGP based Layer-3 VPNs, as specified in RFC 4364, provide an industry proven and well-defined solution for supporting Layer-3 virtual network services. RFC 4364 mechanisms use MPLS labels to provide the network virtualization capability in the data plane. This document specifies a simple mechanism to use the new IP-based virtual network overlay encapsulations, while continuing to leverage the BGP based Layer-3 VPN control plane techniques and extensions.

Table of Contents

1. Introduction

Virtual network overlays are being designed and deployed in various types of networks, including data center networks. These network overlays serve several purposes including flexible network virtualization, increased scale, multi-tenancy, and mobility. Such overlay networks may be used to provide both Layer-2 and Layer-3 network services to hosts at the network edge. New encapsulations are being defined and standardized to support these virtual networks. These encapsulations are primarily based on IP, such as VxLAN and NvGRE.

BGP based Layer-3 VPNs, as specified in RFC 4364, provide an industry proven and well-defined solution for supporting Layer-3 virtual network services. RFC 4364 mechanisms use MPLS labels to provide the network virtualization capability in the data plane. This document specifies a simple mechanism to use the new IP-based virtual network overlay encapsulations, while continuing to leverage the BGP based Layer-3 VPN control plane techniques and extensions.

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. Virtual Network Identifier

In RFC 4364 L3VPNs, a 20-bit MPLS label that is assigned to a VPN route determines the forwarding behavior in the data plane for traffic following that route. These labels also serve to distinguish the packets of one VPN from another.

On the other hand, the various IP overlay encapsulations support a virtual network identifier as part of their encapsulation format. A virtual network identifier is a value that at a minimum can identify a specific virtual network in the data plane. It is typically a 24-bit value which can support upto 16 million individual network segments.

There are two useful requirements regarding the scope of these virtual network identifiers.

- Network-wide scoped virtual network identifiers

Depending on the provisioning mechanism used within a network domain such as a data center, the virtual network identifier may have a

network scope, where the same value is used to identify the specific Layer-3 virtual network across all network edge devices where this virtual network is instantiated. This network scope is useful in environments such as within the data center where networks can be automatically provisioned by central orchestration systems. Having a uniform virtual network identifier per VPN is a simple approach, while also easing network operations (i.e. troubleshooting). It also means simplifies requirements on network edge devices, both physical and virtual devices. A critical requirement for this type of approach is to have a very large amount of network identifier values given the network-wide scope.

o Locally assigned virtual network identifiers

In an alternative approach supported as per RFC 4364, the identifier has local significance to the network edge device that advertises the route. In this case, the virtual network scale impact is determined on a per node basis, versus a network basis.

When it is locally scoped, and uses the same existing semantics of a MPLS VPN label, the same forwarding behaviors as specified in RFC 4364 can be employed. It thus allows a seamless stitching together of a VPN that spans both an IP based network overlay and a MPLS VPN. This situation can occur for instance at the data center edge where the overlay network feeds into an MPLS VPN. In this case, the identifier may be dynamically allocated by the advertising device.

It is important to support both cases, and in doing so, ensure that the scope of the identifier be clear and the values not conflict with each other.

It should be noted that deployment scenarios for these virtual network overlays are not constrained to the examples used above to categorize the options. For example, a virtual network overlay may extend across multiple data centers.

o Global unicast table

The overlay encapsulation can also be used to support forwarding for routes in the global or default routing table. A virtual network identifier value can be allocated for the purpose as per the above options.

2.1. Virtual Network Identifier Specification

The above requirements can be achieved in a simple manner by splitting the virtual network ID number space.

- o Values upto 1 million (or less than 20 bits) are treated exactly as MPLS labels and have significance local to the advertiser.

For future expansion, this draft stipulates that the 16 numerical values in the end of the label range, i.e. values 0xffff0 to 0xfffff, be reserved for future use. These special labels could be used to indicate the presence of other types of IP payloads.

- o Values greater than 1 million (greater than 20 bits) are treated as per their original definition.

- o A virtual network identifier value of zero is used by default to indicate the global or routing table.

It should be noted that within an administrative domain, the entire range can be used such that the values have network-wide significance. This is inline with the use of statically assigned labels today.

2.2. Identifier Scope and propagation

The virtual network identifier may be indicated by attaching to the route a new attribute. However, it is also possible to use the MPLS label field in the BGP VPN NLRIs to specify this value. The benefit of doing the latter is the reuse of existing NLRIs and label processing as is, especially keeping in mind the semantics to be supported. The specification of the identifier value in the label field is described further below.

The use of the virtual network identifier is coupled with the encapsulation used for sending traffic.

The encapsulation used may be MPLS. In this case, the identifier value should be less than 0xffff0, and will be set in the MPLS label field exactly as defined in RFC 3107. There is no change to current RFC 4364 behavior in this case.

When the encapsulation is one of the overlay encapsulation types as listed below, the virtual network identifier will be set in the 3-byte label field described in RFC 3107 as a 24-bit value, irrespective of the actual value being specified.

The value itself may fall into two ranges.

1. Less than 0xffff0 - In this case, the identifier has local significance to the network device that advertised the route.

2. Greater than 0xfffff - In this case, the identifier will have a

significance as per the original definitions, typically within a network domain that is under a common provisioning system.

From a routing perspective, if an intermediate network device changes the BGP next-hop to self before propagating the route, it will assign a new virtual network identifier and advertise it. If not, the virtual network identifier attached by the originator of the route will be carried as is.

When an intermediate network device assigns a virtual network identifier, the assigned value may be a new locally assigned value or it could still be the same network scoped value, if the route is being propagated within the domain.

2.3. Forwarding behavior

- Locally assigned virtual network identifiers

When the virtual network identifier is locally assigned, forwarding based on the identifier follows the semantics of an MPLS label. This label can serve as either an aggregate label or a per-prefix label. This allows a seamless transition out of the overlay network at an MPLS VPN edge, for example, via support of Inter-AS option B.

- Network-scoped virtual network identifiers

With the network-scoped virtual network identifier, any egress device treats the identifier as an aggregate label to lookup the appropriate forwarding table.

In both cases, the forwarding behavior at an ingress edge device, physical or virtual, does not change.

3. Overlay Encapsulation

As mentioned above, different overlay encapsulations may be used to provide an overlay virtual network.

The overlay may use proposed encapsulations such as:

1. VxLAN

2. NvGRE

Based on the encapsulation type being used, the virtual network identifier is appropriately encoded.

When VxLAN encapsulation is used, the virtual network identifier is carried as the 24-bit segment-ID in the VxLAN header.

When NvGRE encapsulation is used, the virtual network identifier is carried as the 24-bit tenant network ID in the NvGRE header.

The fact that a virtual network identifier is carried in the label field in the BGP NLRI is determined by virtue of the accompanying encapsulation attribute, that indicates an overlay encapsulation should be used.

For a given overlay edge device, the same encapsulation may be used for all routes or for selected routes.

3.1. Encapsulation specification

The overlay encapsulation attribute may be carried with each route, or it may be indirectly inferred from the route to the BGP next-hop.

The Tunnel Encapsulation Extended community defined in RFC 5512 can be used to convey this information. [remote-next-hop] specifies an alternative mechanism to carry this information along with each route. The address specified as the remote next-hop identifies the end-point or destination of the encapsulated packets that use the dependent routes.

A single encapsulation may be used on a given device. In this case, the encapsulation may be specified for a given next-hop and inherited by all routes sent with that next-hop (RFC 5512).

When VxLAN and NvGRE encapsulations are used, the header by definition contains an Ethernet MAC address within the overlay header. When these encapsulations are used for Layer-3 as specified in this document, the MAC addresses are not relevant. A single well-known MAC address may be specified for the purpose of deterministically driving a Layer-3 lookup based on the inner IP or IPv6 address.

However, an overlay egress edge device device may choose to specify a MAC address as part of the encapsulation header in its route advertisement. In this case, any ingress edge device sending traffic as per this route must use the above specified MAC address as the destination MAC address in the header. The egress device may use this address to drive the Layer-3 table lookup or for other purposes.

When an intermediate device changes the BGP next-hop to self before propagating a received route, it will also need to advertise a new overlay encapsulation attribute with the local address as the

endpoint. While doing so, it may use an overlay encapsulation type that is different from the received route.

4. Acknowledgements

The authors would like to acknowledge and thank Dave Smith, Maria Napierala, Ashok Ganesan and Luyuan Fang for their input and feedback.

5. IANA Considerations

The virtual network identifier values 0xffff0 to 0xfffff should be allocated by IANA as applications for carrying payloads different than regular IP/VPN packets emerge in future.

6. Security Considerations

This draft does not add any additional security implications to the BGP/L3VPN control plane. All existing authentication and security mechanisms for BGP apply here.

The security considerations pertaining to the various IP overlay encapsulations referenced here are described in the respective overlay encapsulation specifications.

7. References

7.1. Normative References

[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-00 (work in progress), August 2011.

[I-D.sridharan-virtualization-nvgre]
Sridhavan, M., Duda, K., Ganga, I., Greenberg, A., Lin, G., Pearson, M., Thaler, P., Tumuluri, C., and Y. Wang, "NVGRE: Network Virtualization using Generic Routing Encapsulation", draft-sridharan-virtualization-nvgre-00 (work in progress), September 2011.

[RFC2119] Bradner, S., "Key words for use in RFCs to Indicate

Requirement Levels", BCP 14, RFC 2119, March 1997.

[min_ref] authSurName, authInitials., "Minimal Reference", 2006.

7.2. Informative References

[I-D.narten-iana-considerations-rfc2434bis]
Narten, T. and H. Alvestrand, "Guidelines for Writing an IANA Considerations Section in RFCs", draft-narten-iana-considerations-rfc2434bis-09 (work in progress), March 2008.

[I-D.vandevelde-idr-remote-next-hop]
Velde, G., Patel, K., Raszuk, R., and R. Bush, "BGP Remote-Next-Hop", draft-vandevelde-idr-remote-next-hop-01 (work in progress), July 2012.

[RFC2629] Rose, M., "Writing I-Ds and RFCs using XML", RFC 2629, June 1999.

[RFC3107] Rekhter, Y. and E. Rosen, "Carrying Label Information in BGP-4", RFC 3107, May 2001.

[RFC3552] Rescorla, E. and B. Korver, "Guidelines for Writing RFC Text on Security Considerations", BCP 72, RFC 3552, July 2003.

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

[RFC4364] Rosen, E. and Y. Rekhter, "BGP/MPLS IP Virtual Private Networks (VPNs)", RFC 4364, February 2006.

[RFC5512] Mohapatra, P. and E. Rosen, "The BGP Encapsulation Subsequent Address Family Identifier (SAFI) and the BGP Tunnel Encapsulation Attribute", RFC 5512, April 2009.

Authors' Addresses

Dhananjaya Rao
Cisco
San Jose,
USA

Email: dhrao@cisco.com

John Mullooly
Cisco
New Jersey,
USA

Email: jmullool@cisco.com

Rex Fernando
Cisco
San Jose,
USA

Email: rex@cisco.com