

Network Working Group
Internet-Draft
Intended status: Informational
Expires: October 19, 2013

J. Dong
Z. Li
Huawei Technologies
B. Parise
Cisco Systems
April 17, 2013

A Framework for L3VPN Performance Monitoring
draft-dong-l3vpn-pm-framework-01

Abstract

The capability of BGP/MPLS IP Virtual Private Networks (L3VPN) performance monitoring (PM) is important to meet the Service Level Agreement (SLA) for the service beared. Since multipoint-to-point or multipoint-to-multipoint (MP2MP) network model applies, flow identifying is a big challenge for L3VPN PM. This document specifies the framework and mechanisms for the application of L3VPN PM.

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 October 19, 2013.

Copyright Notice

Copyright (c) 2013 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. Overview and Concepts	3
2.1. VRF-to-VRF Tunnel	3
3. Control Plane	3
3.1. VPN Membership Auto-Discovery	3
3.2. VRF-to-VRF Label Allocation	4
4. Data Plane	4
4.1. Additional Label for Ingress VRF Identification	4
4.2. Replace the VPN Label with VT Label	5
5. L3VPN Performance Monitoring	6
6. IANA Considerations	6
7. Security Considerations	6
8. Acknowledgements	6
9. References	6
9.1. Normative References	6
9.2. Informative References	7
Authors' Addresses	7

1. Introduction

Level 3 Virtual Private Network (L3VPN) [RFC4364] service is widely deployed to provide enterprise VPN, Voice over IP (VoIP), video, mobile backhaul, etc. services. Most of these services are sensitive to the packet loss and delay. The capability to measure and monitor performance metrics for packet loss, delay, as well as related metrics is essential for meeting the Service Level Agreement (SLA). This measurement capability also provides operators with greater visibility into the performance characteristics of the services in their networks, and provides diagnostic information in case of performance degradation or failure and helps for fault localization.

To perform the measurement of packet loss, delay and other metrics on a particular VPN traffic flow, the egress PE needs to identify the ingress VRF sending the VPN packets. As specified in the [I-D.zheng-l3vpn-pm-analysis], such flow identification is a big challenge for existing L3VPN.

This document specifies the framework and mechanisms for the application of performance monitoring in L3VPN.

2. Overview and Concepts

Based on the mechanisms in [RFC4364], for a particular VPN prefix, the directly connected PE allocates the same VPN label to all the remote PEs which maintain VPN Routing and Forwarding Tables (VRFs) of that VPN. Thus performance monitoring can not be performed on the egress PE, since it is not able to identify the source VRF of the received VPN packets.

As analyzed by [I-D.zheng-l3vpn-pm-analysis], to perform the packet loss or delay measurement on a specific VPN flow, it is critical for the egress PE to identify the unique VRF, i.e. to establish the Point-to-Point connection between the two VRFs. Once the Point-to-Point connection is built up, current measurement mechanisms may be applied to L3VPN. A new concept "VRF-to-VRF Tunnel" is introduced in the following section to establish such Point-to-Point connection.

2.1. VRF-to-VRF Tunnel

In order to perform performance monitoring in L3VPN, a point-to-point connection between any two VRFs of a particular VPN needs to be established. This guarantees that the egress PE could identify the ingress VRF of the received VPN traffic, thus it could measure the packet loss and delay between the ingress and egress VRFs. Such point-to-point VPN connection between an ingress VRF and an egress VRF is called "VRF-to-VRF Tunnel (VT)".

3. Control Plane

This section describes the control plane mechanisms needed for L3VPN performance monitoring.

3.1. VPN Membership Auto-Discovery

Before establishing the Point-to-Point connections between VRFs, each PE attaching a given VPN needs to know all the remote PEs that attach to the same VPN. This can be achieved by the membership auto-discovery procedure. Some mechanisms similar to the membership auto-discovery in MVPN [RFC6513] can be used.

3.2. VRF-to-VRF Label Allocation

After obtaining the VPN membership information, each PE needs to allocate MPLS labels to identify the VRF-to-VRF tunnel between the local VRF and the remote VRFs, such labels are called VT labels. For each local VRF, the egress PE SHOULD allocate different VT labels for each remote VRF in PEs belonging to the same VPN. In this way, the egress PE could identify the VPN flow received from different ingress VRFs, and the packet loss and delay measurement could be performed between each ingress VRF and the local VRF.

4. Data Plane

This section introduces two new MPLS label stack encapsulations when VT label applies.

4.1. Additional Label for Ingress VRF Identification

When a VPN data packet needs to be sent, firstly the VPN label obtained from the BGP VPN route of the destination address prefix is pushed onto the label stack. The VT label allocated by the egress VRF should then be pushed onto the label stack to identify the Point-to-Point connection between the sending and receiving VRF. Finally the MPLS tunnel label is pushed onto the label stack. The process of TTL and COS fields between the VPN label encapsulation and the tunnel label encapsulation is done according to the Pipe and Uniform Models defined by [RFC3270] and [RFC3443]. The TTL and COS value in the VPN label entry should be copied to the TTL and COS fields of the VT label encapsulation respectively. In this way, one additional label is carried in the label stack compared with L3VPN data plane in [RFC4364].

When the VPN data packet arrives at the egress PE, the outermost tunnel label is popped, then the egress PE could use the VT label to identify the ingress VRF of the packet. The process of TTL and COS fields at the egress node should be done according to the Pipe and Uniform Models defined by [RFC3270] and [RFC3443]. Since the value of the TTL and COS fields of the VPN label encapsulation and the VT label encapsulation are the same, the TTL and COS fields of the VT label encapsulation can be ignored during the course of the TTL and COS process at the egress node.

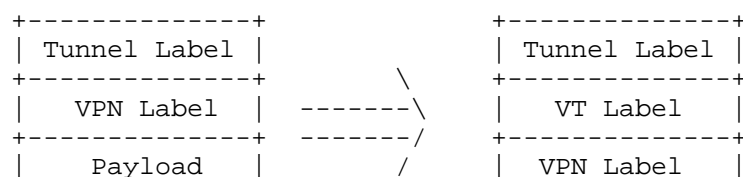




Fig.1 Additional Label for Ingress VRF Identification

4.2. Replace the VPN Label with VT Label

Since the VT label identifies the connection between the ingress VRF and egress VRF, it could also be used to identify the egress VRF table in which the VPN prefix lookup should be performed. Thus when encapsulating the VPN data packets, the ingress PE could simply replace the VPN label with the VT label, then push the tunnel label. The process of TTL and COS fields between the VPN label encapsulation and the tunnel label encapsulation is done according to the Pipe and Uniform Models defined by [RFC3270] and [RFC3443]. The TTL and COS value of the VPN label entry should be copied to the TTL and COS field of the VT label respectively. In this way the depth of the MPLS label stack is unchanged.

The encapsulation method would require the egress PE to perform VPN prefix lookup in the egress VRF table before the packet can be forwarded to a specific CE. The similar procedure is also required when per-instance VPN label allocation mechanism is used. The process of TTL and COS fields at the egress node should be done according to the Pipe and Uniform Models defined by [RFC3270] and [RFC3443]. Since the VPN label encapsulation is replaced with the VT label encapsulation, the TTL and COS fields of the VT label encapsulation should be used as those of the VPN label encapsulation during the course of the TTL and COS process at the egress node.

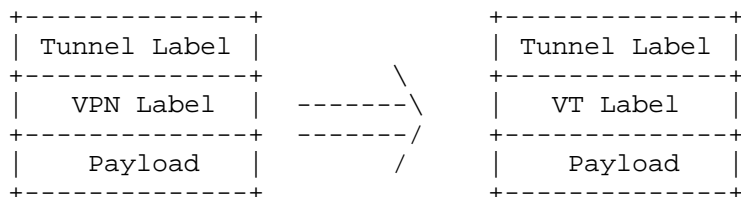


Fig.2 Replace the VPN Label with VT Label

5. L3VPN Performance Monitoring

Since the challenge of identifying the ingress VRF is resolved in section 4, the procedures for the packet loss and delay measurement as defined in [RFC6374] can be utilized for L3VPN performance monitoring. The main difference between performance monitoring of L3VPN and MPLS is the format of identifiers in the Loss Measurement (LM) and Delay Measurement (DM) messages. Specifically, for L3VPN, the source and destination addresses of the LM and DM messages should be set to the concatenation of the Route Distinguisher (RD) of the particular VRF and the IP address of the ingress and egress PE respectively.

6. IANA Considerations

This document makes no request of IANA.

Note to RFC Editor: this section may be removed on publication as an RFC.

7. Security Considerations

TBD

8. Acknowledgements

TBD

9. References

9.1. Normative References

- [I-D.zheng-l3vpn-pm-analysis]
Zheng, L. and Z. Li, "Performance Monitoring Analysis for L3VPN", draft-zheng-l3vpn-pm-analysis-00 (work in progress), October 2012.
- [RFC2119] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, March 1997.
- [RFC3270] Le Faucheur, F., Wu, L., Davie, B., Davari, S., Vaananen, P., Krishnan, R., Cheval, P., and J. Heinanen, "Multi-Protocol Label Switching (MPLS) Support of Differentiated Services", RFC 3270, May 2002.
- [RFC3443] Agarwal, P. and B. Akyol, "Time To Live (TTL) Processing in Multi-Protocol Label Switching (MPLS) Networks", RFC 3443, January 2003.

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

[RFC6374] Frost, D. and S. Bryant, "Packet Loss and Delay Measurement for MPLS Networks", RFC 6374, September 2011.

9.2. Informative References

[RFC6513] Rosen, E. and R. Aggarwal, "Multicast in MPLS/BGP IP VPNs", RFC 6513, February 2012.

Authors' Addresses

Jie Dong
Huawei Technologies
Huawei Building, No.156 Beiqing Rd.
Beijing 100095
China

Email: jie.dong@huawei.com

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

Email: lizhenbin@huawei.com

Bhavani Parise
Cisco Systems

Email: bhavani@cisco.com

L3VPN
Internet-Draft
Intended status: Standards Track
Expires: January 14, 2014

D. Rao
J. Mullooly
R. Fernando
Cisco
July 15, 2013

Layer-3 virtual network overlays based on BGP Layer-3 VPNs
draft-rao-bgp-l3vpn-virtual-network-overlays-02

Abstract

Virtual network overlays are being designed and deployed in various types of networks, including data centers. These network overlays address several requirements including flexible network virtualization and multi-tenancy, increased scale, and support for mobility of virtual machines. Such overlay networks can be used to provide both Layer-2 and Layer-3 network services to hosts at the network edge. New packet encapsulations are being defined and standardized to support these virtual networks. These encapsulations, such as VXLAN and NVGRE, are primarily based on IP and are currently defined to support a Layer-2 forwarding service.

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. However, RFC 4364 procedures use MPLS labels to provide the network virtualization capability in the data plane. With the increasing support for IP overlay encapsulations in data center devices, there is a strong preference to utilize this support to deploy Layer-3 virtual networks using the familiar policy and operational primitives of Layer-3 VPNs.

This document describes the use of BGP Layer-3 VPNs along with the new IP-based virtual network overlay encapsulations to provide a Layer-3 virtualization solution for all IP traffic, and specifies mechanisms to use the new encapsulations while continuing to leverage the BGP Layer-3 VPN control plane techniques and extensions. This mechanism provides an efficient incremental solution to support forwarding for IP traffic, irrespective of whether it is destined within or across an IP subnet boundary.

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 14, 2014.

Copyright Notice

Copyright (c) 2013 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	3
1.1. Requirements Language	4
1.2. Terminology	4
1.3. Control plane signaling requirements	4
1.4. Control plane model	5
1.5. Overlay Encapsulations	5
2. Virtual Network Identifier	5
2.1. Virtual Network Identifier Scope	6
2.1.1. Domain-scoped provisioned virtual network identifiers	6
2.1.2. Per-device scoped allocated virtual network identifiers	6
2.1.3. Global unicast table	7
2.1.4. Virtual Network Identifier Specification	7
2.2. Signaling Virtual Network Identifiers	7
2.2.1. Signaling Requirements	8
2.2.2. Signaling Specification	9
3. Overlay Encapsulation specification	9
3.1. Encapsulation for VXLAN and NVGRE	10
3.2. Encapsulation for MPLS-in-GRE	11
3.3. Multiple encapsulations	11
3.4. Gateway device encapsulation handling	11

4. Forwarding behavior	11
5. Overlay Interconnection and Interworking Scenarios	12
5.1. End-to-end overlay	12
5.2. Virtual-network overlay VPN interworking	12
5.2.1. Normalized interworking via VRF	13
5.2.2. Seamless VPN interworking	13
6. Virtual-Network Overlay Encapsulation Capability	13
6.1. Need for Capability Negotiation	13
6.2. Capability Specification	14
7. Acknowledgements	15
8. IANA Considerations	15
9. Security Considerations	15
10. References	16
10.1. Normative References	16
10.2. Informative References	16
Authors' Addresses	17

1. Introduction

Virtual network overlays are increasingly being designed and deployed in various types of networks, including data center networks. These virtual network overlays can 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 transport, such as VXLAN and NVGRE. A significant characteristic of these encapsulations is the presence of an embedded virtual network identifier field that is part of the encapsulation header. The use of these encapsulations is defined in [VXLAN] and [NVGRE] and is being currently worked on as part of the NVO3 architecture [NVO3].

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. The Layer-3 VPN BGP control plane is eminently suitable to provide a Layer-3 network virtualization solution in the data center.

However, RFC 4364 mechanisms use MPLS labels as the mechanism to provide the network virtualization capability in the data plane. An MPLS label is signaled by a device advertising a VPN-IP route. This label can identify the virtual network when the device processes packets received with that label. RFC 4364 does allow an MPLS label to be carried in an IP transport encapsulation such as MPLS-in-GRE.

This document specifies procedures to use the new IP-based virtual network overlay encapsulations such as VXLAN and NVGRE, while continuing to leverage the BGP based Layer-3 VPN control plane techniques and extensions. It also describes how virtual network

overlays based on these encapsulations can efficiently interconnect with one another and with existing MPLS based L3VPN networks.

This document describes the protocol extensions necessary to allow advertising a VPN-IP NLRI with an attached VN-ID as well as an encapsulation attribute indicating the type of encapsulation, for example, VXLAN or NVGRE.

There are aspects of the signaling of encapsulation and VN-ID that can be leveraged across different kinds of services. Hence, the generic overlay encapsulation signaling extensions are defined in [remote-next-hop]. The current document provides the necessary context of how these extensions are used with the BGP IP-VPN NLRIs.

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].

1.2. Terminology

VM: Virtual Machine

Edge device: The edge device is where end-hosts (eg. application VMs) attach to the overlay. This is where the tunnel encapsulation starts. It's called NVE in the NVO3 terminology. The NVE is equivalent to a VPN PE in the context of BGP L3VPNs.

1.3. Control plane signaling requirements

While considering the leverage of the BGP L3VPN control plane with the IP overlay technologies, the following requirements should be supported.

1. Signal VN-ID with VPN-IP routes, that can be used with IP based overlay encapsulations.
2. Support signaling of multiple encapsulations per edge device.
3. Have flexibility to support single and per-encapsulation VN-ID spaces if needed.
4. Support both device-local and domain-global VN-ID/label spaces.
5. Support per-prefix granularity for VN-ID/encapsulation.
6. Support interoperability with legacy IP-VPN PEs.

7. Be efficient in signaling to provide good scalability.

8. Minimize protocol and deployment overhead.

1.4. Control plane model

The virtual network overlay described in this document uses regular BGP peering on the edge devices for policy constrained route distribution. A typical deployment would use Route Reflectors.

It is also feasible for an alternative protocol or provisioning framework to be used to control the forwarding plane population and forwarding behavior on the edge devices, as described in [vpe-framework].

The extensions specified here are compatible with both approaches.

1.5. Overlay Encapsulations

Different tunnel encapsulations may be used to realize an overlay virtual network. Based on the encapsulation type being used, the virtual network identifier is appropriately encoded in the encapsulation header.

An overlay network may use the IP based VN-ID encapsulations such as VXLAN and NVGRE. It may also use an MPLS based encapsulation such as MPLS-in-GRE.

When VXLAN encapsulation is used, the virtual network identifier is carried as the 24-bit VNI in the VXLAN header.

When NVGRE encapsulation is used, the virtual network identifier is carried as the 24-bit VSID in the NVGRE header.

When MPLS-in-GRE is used, the regular MPLS VPN label serves as the data plane identifier for the virtual network or a specific destination.

A given overlay edge device may support a single encapsulation type or it may support multiple encapsulation types. In the latter case, it may signal the multiple encapsulations so that a receiving device can potentially use the one most suitable to it. An edge device may use the same encapsulation(s) for all routes or for a subset of routes.

2. Virtual Network Identifier

In RFC 4364 based Layer-3 VPNs, a 20-bit MPLS label is assigned to an VPN-IP route by the device that advertises the route, with itself as the BGP next-hop. This label determines the forwarding behavior in the data plane for traffic being switched as per that route. A device receiving this route will encode this label in the packet header when sending traffic to the advertiser. The advertiser will take a unique forwarding action for traffic received with this label when compared to traffic with other labels. The label may also be used at the granularity of a VPN table and drive an IP lookup in that table. This MPLS VPN label is independent of the transport encapsulation that is used to carry this traffic to this PE from other PEs across a core network. The transport encapsulation may be native MPLS or be IP (eg, MPLS-in-GRE).

On the other hand, the various IP overlay encapsulations support a virtual network identifier explicitly within their encapsulation header. A virtual network identifier is a value that at a minimum can identify a specific virtual network table in the forwarding plane, and may be used to perform an IP address lookup. It may also drive a specific forwarding action for packets destined to a particular destination address or prefix.

It is typically a 24-bit value which can support upto 16 million individual network segments or end-hosts. For instance, VXLAN defines a 24-bit VNI while NVGRE uses a 24-bit VSID that is carried in the GRE key field of the GRE header.

2.1. Virtual Network Identifier Scope

The scope of these virtual network identifiers fall into two broad categories. 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.

2.1.1. Domain-scoped provisioned virtual network identifiers

Based on the provisioning mechanism used, a virtual network identifier typically has a domain-wide scope within the network domain, where a unique value is assigned to a given virtual network or a given IP destination route at one or more edge devices.

This scope is useful in environments such as data centers where virtual networks and VMs are automatically provisioned by central orchestration systems. The system must support a very large number of VN-IDs given the scope.

2.1.2. Per-device scoped allocated virtual network identifiers

There are scenarios where it is also necessary for an identifier to have significance local to each network edge device that advertises the route. In this case, the same value may be used by different edge devices to represent different forwarding classes.

When it is locally scoped, the virtual network identifier may be dynamically allocated by the advertising device. This allocation follows the same semantics of an MPLS VPN label, and supports similar forwarding behaviors as specified in RFC 4364. The device may, for example, be a DC-WAN edge device that supports L3VPN Inter-AS Option B and use this allocation for routes received from other ASBRs.

2.1.3. 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.

2.1.4. Virtual Network Identifier Specification

The above requirements can be achieved in a simple manner by splitting the virtual network ID number space to support both domain-wide and device-local scopes.

- o Values upto 1 million (or less than 20 bits) SHOULD be treated with the same semantics as MPLS VPN labels and have significance local to the advertiser.

For future expansion, this draft stipulates that the 16 numerical values in the end of the VN-ID range be reserved for future use. These special values could be used to indicate the presence of other types of IP payloads.

- o Values greater than 1 million (greater than 20 bits) SHOULD be treated as per their original definition, ie domain-wide scoped values.

These limits are not mandatory, but are recommended defaults. As long as the provisioning system can ensure conflict-free operation, the boundary between local and domain scoped ranges can be adjusted higher or lower by configuration.

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

2.2. Signaling Virtual Network Identifiers

2.2.1. Signaling Requirements

The Introduction section listed the desirable characteristics of signaling of VN-IDs. This section elaborates on a couple of those requirements.

- o The device may support a single VN-ID space across all its supported encapsulations.

This is expected to be common deployed scenario. A given edge device will be provisioned by a single network orchestrator or controller. The device may support multiple encapsulations in order to interoperate with remote edge devices that support a different encapsulation. However, the single network orchestrator will manage the VN-ID space that will be common across multiple encapsulations on this device.

- o A device may support an independent VN-ID space per-supported encapsulation.

This is expected to be applicable mostly at network gateway devices that interconnect two different overlay domains and support the same virtual network across these domains. These border devices are likely to be managed by two different orchestrators, and hence need to support different VN-ID spaces. In this case, they typically advertise routes of one domain into another.

- o An edge device may support an independent VN-ID space per-supported encapsulation.

This is assumed to not be a common scenario, where an edge device within the domain is being shared or managed by multiple orchestration systems. However, in case this scenario must be supported, the edge device must be able to support multiple distinct VN-ID spaces. An alternative scheme would be to divide the VN-ID range among the orchestration systems.

- o It is required to support prefix-level VN-ID assignment.

Supporting prefix-level granularity is useful in various scenarios, for example, at an interworking point between DC (VXLAN) and WAN (MPLS).

2.2.2. Signaling Specification

This document specifies two options for signaling VN-IDs.

1. The VN-ID is encoded within an Virtual-Network Overlay encapsulation attribute that also contains the encapsulation type and associated parameters.

This enables the device to signal a VN-ID per encapsulation that it may support. For example, the device may use VN-ID1 for VXLAN and VN-ID2 for NVGRE. The VN-ID is encoded as a 24-bit value in each encapsulation attribute.

When multiple VN-IDs need to be signaled, one per overlay encapsulation type, then the VN-ID MUST be included in the overlay encapsulation attribute as defined in [remote-next-hop].

When MPLS-in-GRE is one of the encapsulations, there is no change from current behavior. The VPN label is encoded in the label field in the IP-VPN NLRI.

2. The VN-ID is encoded in the label field in the IP-VPN NLRI.

This option is used when a device supports a single VN-ID space across all encapsulations. The benefit of this encoding is it's efficiency of packing, even when used for per-prefix VN-ID assignment. With this option, the 24-bit VN-ID for VXLAN and NVGRE is encoded as a 3-byte label field in the IP-VPN NLRI.

When a VN-ID or VPN label is to be signaled, the value MUST be encoded in the 3-octet label field in the IP or IP-VPN NLRI.

This offers the most efficient packing of prefixes in BGP update messages. The device may still advertise multiple encapsulation types with this route, but they will all use the same VN-ID value.

An advertising device will select the suitable option as per the requirements stated above, based on configuration.

3. Overlay Encapsulation specification

Signaling the VN-ID must be coupled with signaling the appropriate overlay encapsulation type. An overlay encapsulation attribute MUST be carried with each route.

The section above specified two options of signaling VN-ID. In both options, the accompanying encapsulation attribute indicates that a 24-bit VN-ID is specified with the NLRI and must be encoded in the signaled encapsulation header.

The encapsulation attribute also indicates the accompanying parameters to be used in the packet header.

RFC 5512 defines a Tunnel Encapsulation Extended Community that can be used to signal different tunnel types. It defines an Encapsulation Sub-TLV that can be used to specify encapsulation parameters.

[remote-next-hop] specifies a Remote-Next-Hop attribute which reuses the Encapsulation Sub-TLV from RFC 5512, but adds the flexibility to signal the encapsulation attribute and parameters along with each individual route. The address specified as the remote next-hop identifies the end-point or destination of the encapsulated packets that use the dependent routes as well as the tunnel encapsulation parameters.

Hence, the Remote-Next-Hop attribute is used to signal VN-ID encapsulations. New tunnel types are defined for VXLAN and NVGRE. The format for the tunnel parameters are specified in [remote-next-hop].

3.1. Encapsulation for VXLAN and NVGRE

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.

Alternatively, an overlay egress edge 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.

3.2. Encapsulation for MPLS-in-GRE

When MPLS-in-GRE is one of the encapsulations, there is no change from current behavior. A tunnel type of [MPLS-in-GRE] as defined in RFC 5512 is used in the Remote-Next-Hop attribute.

3.3. Multiple encapsulations

A given overlay edge device MAY advertise multiple Encapsulation Sub-TLVs, in order to signal multiple encapsulations. It MAY encode a different VN-ID in each Sub-TLV as per rules specified earlier.

A receiving edge device may support one or more encapsulations that are signaled by the advertising edge device. In that case, the receiving device can select any of these encapsulations for sending traffic to the advertiser. If a receiving device supports no encapsulations that were signaled by the advertiser, then it will not send any traffic for these routes to the advertiser.

3.4. Gateway device encapsulation handling

When an intermediate gateway device changes the BGP next-hop to self before propagating a received route, it may 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. It MAY also signal a different VN-ID or VPN label than what it received, as described in the various VN-ID requirements and rules earlier.

4. Forwarding behavior

o Locally assigned virtual network identifiers

When the virtual network identifier is locally assigned, forwarding based on the identifier at the advertising device follows the semantics of an MPLS VPN label. The VN-ID may either drive an IP table lookup or provide a seamless binding to an output VN-ID or label.

o Domain-scoped provisioned virtual network identifiers

With a provisioned VN-ID, forwarding behavior at a device which is provisioned with this value is governed by the forwarding action that has been provisioned. As one example, the VN-ID may be set up to represent a specific IP VRF table on all relevant edge devices, causing incoming packets with this VN-ID to undergo an IP lookup. Alternatively, the VN-ID may be configured on only one or two edge or border devices to directly forward incoming packets to an attached end-host, without undergoing an IP lookup.

In either case, the forwarding behavior at any ingress edge device (physical or virtual) remains the same. The ingress edge device adds an encapsulation as signaled by the advertising device, and includes the VN-ID in that encapsulation header.

5. Overlay Interconnection and Interworking Scenarios

Multiple virtual network overlay domains may be inter-connected using a couple of approaches. Both these approaches may co-exist in the same data center, and be used for connectivity to different kinds of external networks.

5.1. End-to-end overlay

The IP overlay encapsulation or tunnel extends end-to-end between edge devices in different data centers.

IP routes for hosts attached to each edge device are exchanged between the overlay domains either via route exchange between BGP speakers in each overlay domain, or via an orchestration framework that manages the domains. The two networks may be located within the same ASN or may extend across ASes.

The routes are propagated to various edge device via the control plane mechanism used in the DC, along with the encapsulation and VN-ID or label to be used for sending traffic to a given destination edge device. All intermediate devices in the forwarding path between the two edge devices simply transport the IP encapsulated overlay traffic.

The tunnel endpoints, ie the edge devices need to be reachable across the ASes. This reachability may be provided by BGP peering.

5.2. Virtual-network overlay VPN interworking

The overlay encapsulation terminates at a border router such as the DC-WAN gateway device. The gateway device may re-encapsulate packets in another header when sending it onwards. This requires an interworking function which can be of multiple types.

5.2.1. Normalized interworking via VRF

The overlay based virtual network terminates into an L3VPN VRF at the DC-WAN gateway device. Internal routes of the DC as well as the external routes received from the WAN router are installed in the VRF forwarding table at the DC gateway router. The DC gateway will perform an IP lookup and forward traffic after doing the appropriate output MPLS or IP overlay/VPN encapsulation.

5.2.2. Seamless VPN interworking

In this case, the DC Gateway router performs a direct translation of VN-IDs/labels while switching packets between the DC and WAN interfaces without doing an IP lookup. The forwarding table at the DC Gateway router is set up to do a VN-ID or VPN label lookup and derive the output VN-ID or VPN label. The DC Gateway Router can act as an Inter-AS Option-B ASBR/ABR peering with other ASBRs/ABRs.

6. Virtual-Network Overlay Encapsulation Capability

6.1. Need for Capability Negotiation

When the MP-BGP NLRIs are used along with a VN-ID based encapsulation, the MPLS label field in the NLRI is either not used or is used to indicate the presence of a VN-ID that must be included in the corresponding overlay encapsulation packet header while sending data. A device that supports vanilla RFC 4364 based IP-VPNs but does not understand the extensions specified in this document may not interpret the received MP-BGP NLRI as intended, potentially causing inconsistent forwarding plane behavior. In order to avoid this situation, such devices must not receive the modified NLRIs. The presence of a capability protect against this issue and ensures interoperability with vanilla IP-VPN peers.

[RFC5492] defines a mechanism to allow two peering BGP speakers to discover if a particular capability is supported by each other and thus whether it can be used between them. This document defines a new BGP capability that can be advertised using [RFC5492] and is referred to as the Virtual-Network Overlay Encapsulation capability.

A BGP speaker MUST only advertise to a BGP peer the corresponding MP-BGP NLRIs alongwith a VN-ID if the BGP speaker has first ascertained via BGP Capability Advertisement that the BGP peer supports the Virtual-Network Overlay Encapsulation capability.

With the Virtual-Network Overlay Encapsulation Capability, a VN-capable BGP speaker will detect peers that are not capable of processing VN-ID encapsulation information received in BGP updates.

The speaker MUST not send any VPN-IP routes that contain only a VN-ID based encapsulation to such peers. If the route contains both a VN-ID encapsulation and an MPLS-in-GRE encapsulation, the speaker MAY send the route to the legacy peer with only the MPLS encapsulation information, and with the VN-ID encapsulation information removed.

If routes are advertised by a speaker via a Route Reflector (RR), then the RR MUST advertise the BGP capability for it to receive routes with VN-ID information from the speaker.

When a next-hop address needs to be passed along unchanged (e.g., by an RR), its encoding MUST NOT be changed. If a particular RR client cannot handle that encoding (as determined by the BGP Capability Advertisement), then the NLRI in question cannot be distributed to that client. The RR may, as above, send the route with only the MPLS-in-GRE encapsulation information to such legacy peers.

6.2. Capability Specification

A BGP speaker that is capable of processing VN-ID based encapsulation information in BGP updates as per this specification MUST use the Capability Advertisement procedures defined in [RFC5492] with the Virtual-Network Overlay Encapsulation Capability. The fields in the Capabilities Optional Parameter MUST be set as follows:

- o The Capability Code field MUST be set to 71, indicating the capability.
- o The Capability Length field is set to a variable value that is the length of the Capability Value field (which follows).
- o The Capability value field has the following format:

```

+-----+
| NLRI AFI - 1 (2 octets) |
+-----+
| NLRI SAFI - 1 (2 octets) |
+-----+
| ..... |
+-----+
| NLRI AFI - N (2 octets) |
+-----+
| NLRI SAFI - N (2 octets) |
+-----+

```

where:

* each NLRI AFI, NLRI SAFI pair indicates the BGP NLRI address family for which the speaker can process the VN-ID information.

- * the AFI and SAFI values are defined in the Address Family Identifier and Subsequent Address Family Identifier registries maintained by IANA.

Since this document only concerns itself with the advertisement of IP NLRI and VPN-IP NLRIs, this specification specifies the following values in the Capability Value field of the above capability:

- o NLRI AFI = 1 (IPv4), 2 (IPv6)
- o NLRI SAFI = 1, 2, 4, or 128

It is expected that if new AFI/SAFIs can use this in the future, then these AFI/SAFIs can be included in the Capability values.

7. Acknowledgements

The authors would like to acknowledge and thank Nabil Bitar, Dave Smith, Maria Napierala, Robert Raszuk, Eric Rosen, Ashok Ganesan and Luyuan Fang for their input and feedback.

8. IANA Considerations

This document defines, in Section N, a new Capability Code to indicate the Virtual-Network Overlay Encapsulation Capability in the [RFC5492] Capabilities Optional Parameter. The value for this new Capability Code is 71, which is in the range set aside for allocation using the "FCFS" policy defined in [RFC5226]. There are no additional requirements to IANA at this time. A specific VN-ID range may be reserved for future use as applications for carrying payloads different than regular IP/VPN packets emerge in future.

9. 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.

There are security considerations pertaining to IP based overlay or tunnel encapsulations which are described in the respective overlay encapsulation specifications as well as in RFC 5512.

There are certain measures that may be taken by default to increase the level of security provided at the overlay edge devices.

When an IP-based overlay encapsulation is used within a domain such as a data center, the network edge devices can enforce a default forwarding access rule to restrict the acceptance of such overlay encapsulated packets on their access interfaces attached to servers or VMs.

For example, when VXLAN is being used, an edge device may be directed to filter any VXLAN encapsulated packets (identified by the UDP port number) on their access interfaces. This rule can be further augmented by checking that the destination IP address of such VXLAN packets does not fall in the prefix range allocated to edge device addresses. Similarly, a DC edge device may be directed to not accept any VXLAN encapsulated packets on its interfaces connected to external routers, depending on the interconnectivity option being used.

10. References

10.1. 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-03 (work in progress), February 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.
- [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.

10.2. Informative References

- [I-D.fang-l3vpn-virtual-pe]
Fang, L., Ward, D., Fernando, R., Napierala, M., Bitar, N., Rao, D., Rijsman, B., and S. Ning, "BGP IP VPN Virtual PE", draft-fang-l3vpn-virtual-pe-00 (work in progress), February 2013.
- [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.vandeveldelidr-remote-next-hop]

Velde, G., Patel, K., Rao, D., Raszuk, R., and R. Bush, "BGP Remote-Next-Hop", draft-vandeveldelidr-remote-next-hop-04 (work in progress), January 2014.

[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: dhr Rao@cisco.com

John Mullooly
Cisco
New Jersey
USA

Email: jmullool@cisco.com

Rex Fernando
Cisco
San Jose
USA

Email: rex@cisco.com

l3vpn
Internet-Draft
Intended status: Standards Track
Expires: January 17, 2014

D. Rao
V. Jain
Cisco Systems
July 16, 2013

L3VPN Virtual Network Overlay Multicast
draft-drao-l3vpn-virtual-network-overlay-multicast-00

Abstract

Virtual network overlays are extremely useful for supporting multicast applications in multi-tenant data center networks, by distributing the per-tenant multicast state and forwarding actions at the network edges with minimal control plane load and simpler forwarding in the core. A virtual overlay network may use existing encapsulations such as MPLS-in-GRE or newer IP based encapsulations such as VXLAN and NVGRE.

IP multicast sources and receivers are very commonly spread out across multiple subnets. Sources and receivers may also be spread within and outside a single network domain such as a data center. Hence, a Layer-3 multicast paradigm is the most suitable and efficient approach for delivery of IP multicast traffic.

BGP based MVPNs provide a good basis for providing a solution to support IP multicast across these overlay networks. An appropriate subset of the MVPN control plane and procedures are sufficient to support the requirements, providing for a simpler model of operation. This document describes the use of BGP based MVPNs alongwith the new IP-based virtual network overlay encapsulations to provide a Layer-3 virtualization solution for IP multicast traffic, and specifies mechanisms to use the new encapsulations while continuing to leverage the BGP MVPN control plane techniques and extensions.

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 17, 2014.

Copyright Notice

Copyright (c) 2013 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. Solution Requirements	4
4. Network Topology	4
5. Solution Scenarios	5
6. Fine-grained Transit Pruning	6
7. Originating interest at a receiver edge device	6
8. Source mapping from a sender edge device	7
9. IANA Considerations	8
10. Security Considerations	8
11. Change Log	8
12. References	8
12.1. Normative References	8
12.2. Informative References	8
Authors' Addresses	8

1. Introduction

Virtual network overlays are extremely useful for supporting multicast applications in multi-tenant data center networks, by distributing the per-tenant multicast state and forwarding actions at the network edges with minimal control plane load and simpler forwarding in the core. A virtual overlay network may use existing encapsulations such as MPLS over GRE or newer IP based encapsulations such as VXLAN and NVGRE.

IP multicast sources and receivers are very commonly spread out across multiple subnets. Sources and receivers may also be spread within and outside a single network domain such as a data center. Hence, a Layer-3 multicast paradigm is the most suitable and efficient approach for delivery of multicast traffic.

To send multicast data to multiple receivers across the overlay, packets are sent on a core tree (P-tunnel) that is typically realized using an IP multicast encapsulation such as the ones mentioned above.

In order to reduce the number of multicast P-tunnels that need to be set up and maintained in the core network, aggregate core trees may be used, with multiple VPNs being supported over a given P-tunnel. The P-tunnel encapsulations such as MPLS-in-GRE, VXLAN and NVGRE support a VN-ID or VPN label in the encapsulation header which can be used to distinguish the VPN.

BGP based MVPNs provide a good basis for providing a solution to support IP multicast across these overlay networks. An appropriate subset of the MVPN control plane and procedures are sufficient to support the requirements, providing for a simpler model of operation.

This document describes the use of BGP based MVPNs alongwith the IP-based virtual network overlay encapsulations to provide a Layer-3 virtualization solution for IP multicast traffic, and specifies mechanisms to use the new encapsulations while continuing to leverage the BGP MVPN control plane techniques and extensions.

This mechanism provides an efficient incremental solution to support forwarding for IP traffic, irrespective of whether it is destined within or across an IP subnet boundary.

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. Solution Requirements

1. Support for large number of senders and receivers for a given group
2. Receivers and groups spread out across large number of PEs or edge nodes
3. Support for large number of C-multicast routes
4. Support for Multi-tenancy
5. Support for optimal multicast replication within and across subnets
6. Support for optimal pruning on transit nodes
7. Minimal control plane churn
8. Minimal state in core
9. Support for multiple overlay encapsulations
10. Support for redundancy and load-balancing

4. Network Topology

In an environment such as the data center where overlay networks are used, the IP multicast sources and receivers are hosts typically resident on servers attached to network access devices such as virtual software routers or physical access switches. These hosts may belong to different tenants who require isolation and segregation in forwarding state and traffic. At the same time, the physical transport or core network must be shared for traffic from multiple tenants.

This requires the edge devices to support multiple VPNs facing the access interfaces, with a shared overlay encapsulation towards the core.

Due to the high degree of traffic that flows among the hosts on various servers within the data center, a DC environment is likely to use uniform, densely meshed topologies, such as a spine-leaf topology, which provide high redundancy and bandwidth. The servers may be dually attached to a pair of access switches for edge redundancy.

In order to scale the core multicast, as well as to efficiently support the common case where a large number of senders and receivers are present in a VPN, shared trees are used for the overlay. Bidir multicast is the preferred mode. Also, to support a large number of VPNs efficiently, aggregate trees are used with multiple VPNs being multiplexed over a core tree. The various IP overlay encapsulations such as VXLAN, NVGRE or MPLS-in-GRE contain a VN-ID or VPN label in their header which is used as a distinguisher for a MVPN in the data plane.

5. Solution Scenarios

There are a couple of overall scenarios which are applicable for these virtual overlay networks.

One mode of operation is where all overlay multicast traffic is encapsulated within a fixed number of core trees or P-tunnels that all edge nodes or PEs can be a part of. Then both source and receiver edge devices can independently encapsulate and decapsulate overlay traffic, without requiring additional signaling among them.

In this mode, source edge devices map a given C-flow from a locally attached source onto a specific core multicast tree or P-tunnel based on a local mapping decision. Receiver edge devices join all the available P-tunnels, and hence are able to receive traffic from these source edge devices. They then discard the traffic that they do not have local receivers for, and replicate the interesting flows towards local receivers.

It is, however, desirable that all receiver edge devices do not receive all traffic and have to filter the unnecessary flows. This is especially applicable in high-bandwidth traffic environments. In order to support this ability, it is required to have signaling from a receiver edge device indicating its interest in specific C-groups.

In addition, high-bandwidth sourced traffic flows may be sent on a specific P-tunnel which is determined by the source edge device, and requires interested receiver edge devices to join that core tree. In such cases, it is also beneficial if the sourced multicast traffic is sent out into the overlay only if there are known to be receivers at other edge devices or external to the overlay.

For all such cases, it is required to have signaling from both source and receiver edge nodes. Signaling involves host group interest from receiver edge nodes as well as the C-flow to P-tunnel mapping signaling from source edge nodes.

6. Fine-grained Transit Pruning

Typically, when an overlay is used, the transit nodes in the physical topology only participate in the underlay control plane and forward all traffic based on the encapsulated packets. They are unaware of the inner header or payload. This allows them to be simpler and scale better.

One consequence of using an overlay is that multicast traffic needs to make it to the receiver edge nodes before they can be pruned in case there are no downstream receivers.

A widely deployed option to avoid redundant traffic from getting to all the receiver edge nodes is to use a larger number of core multicast trees, thereby reducing the ratio of overlay flows to each multicast tree, and allowing the receiver edge nodes to only join the multicast trees that the interesting flows map to. This has the side effect of increasing the underlay multicast state in the core, and hence the load on the underlay multicast protocols such as PIM. It also does not provide for complete pruning of multicast traffic.

However, it is possible in certain topologies to support an alternative mechanism for fine-grained pruning of multi-destination traffic.

In the uniform, meshed topologies that are used as mentioned earlier, certain transit nodes can use the receiver interest information sent by the receiver edge devices for the overlay, to filter traffic on outgoing links towards the receiver edge nodes on a per-VPN or more fine-grained basis.

This allows the core multicast trees built by PIM to be smaller in number. The transit nodes do need to run MP-BGP to obtain the overlay multicast group information sent by the various receiver nodes. The transit nodes will typically obtain this information from the RRs.

This assumes the core devices can look into the inner headers of packets to prune traffic based on either VN-ID or *G/S,Gs. In order for the join routes to be sent to the transit nodes, the appropriate RTs will be provisioned on the transit nodes.

7. Originating interest at a receiver edge device

A receiver edge device will originate shared or source tree joins on the overlay for receivers that are locally attached or downstream to it. This will be triggered by locally received IGMP or PIM joins. The receiver edge device will also typically act as a PIM first-hop or last-hop router with attached sources and receivers.

These join routes need to be propagated towards potential sources as well as the RPs. In a virtual network overlay topology, sources are attached to edge devices. So the other edge devices must receive these join routes.

For a situation where there are multiple sources which are spread out across a large number of edge devices, or for a case where the groups are Bidir groups, a shared tree join route may be propagated to all edge devices.

For high bandwidth sources, it is desirable to direct the specific group or source joins to the appropriate source leafs. More granular filtering is needed in this case - in addition to VRF, group based active source discovery and advertisement is used to control join propagation.

A receiver leaf uses C-multicast route of type Shared Tree Join (C-RP, C-G) or Source Tree Join (C-S, C-G). To be able to propagate the shared joins to all edge devices, the join routes may be originated with an RD that is specific to the originating device. This RD can be the same value as that used by unicast routes. The routes are sent with a RT that all interested edge devices may use as an import RT for this VPN, if they need to receive the join routes. Route propagation is constrained based on policy (RT) along the path.

8. Source mapping from a sender edge device

An edge device that is attached to a source may signal the active sources information on the overlay, along with the core multicast group that the edge device decides to use. Receiver edge nodes can use this information to join the core multicast trees.

This mapping is advertised by using the S-PMSI A-D routes, with the PMSI Tunnel Attribute type indicating the appropriate overlay encapsulation type - VXLAN or NVGRE. Aggregate trees will be used, with the VN-ID being signaled along with the route.

9. IANA Considerations

To be completed

10. Security Considerations

To be completed

11. Change Log

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

12.2. Informative 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-02 (work in progress), August 2012.
- [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

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

Email: dhrao@cisco.com

Vipin Jain
Cisco Systems
170 W. Tasman Drive
San Jose, CA 95124 95134
USA

Email: vipijain@cisco.com

INTERNET-DRAFT
Intended Status
Expires: August 25, 2013

Luyuan Fang
Rex Fernando
Dhananjaya Rao
Sami Boutros
Cisco

February 25, 2013

BGP IP VPN Data Center Interconnect
draft-fang-l3vpn-data-center-interconnect-01

Abstract

This document discusses solutions for inter-connection of BGP MPLS IP/VPN [RFC4364] and Data Center (DC) overlay networks. Two categories of inter-connections are discussed in this document. In the first category, Data Center overlay virtual network is built with BGP IP VPN technologies, the inter-connection of IP VPN in the Data Center to BGP IP VPN in the WAN enables end-to-end IP VPN connectivity. New Inter-AS solutions are required in certain scenarios, in addition to the existing Inter-AS Options (A, B, C) defined in [RFC4364]. In the second category, Data Centers overlay network uses non IP VPN technologies, the inter-connection of any overlay virtual network in the Data Center to BGP IP VPN in the WAN provides end user connectivity through stitching of different overlay technologies, the mapping of non IP VPN overlay to IP VPN need to be performed at the border Gateway of the two networks. The role of Software Defined Network (SDN) to assist the inter-connections is discussed.

Status of this Memo

This Internet-Draft is submitted to IETF in full conformance with the provisions of BCP 78 and BCP 79.

Internet-Drafts are working documents of the Internet Engineering Task Force (IETF), its areas, and its working groups. Note that other groups may also distribute working documents as Internet-Drafts.

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."

The list of current Internet-Drafts can be accessed at

<http://www.ietf.org/lid-abstracts.html>

The list of Internet-Draft Shadow Directories can be accessed at
<http://www.ietf.org/shadow.html>

Copyright and License Notice

Copyright (c) 2013 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	3
1.1	Terminology	3
2.	Use Cases	4
2.1	Case 1: End-to-end BGP IP VPN cloud inter-connection	4
2.2	Case 2: Hybrid cloud inter-connection	4
3.	Architecture reference models	5
3.1	BGP/MPLS IP VPN Inter-AS model	5
3.2	BGP/MPLS IP VPN Gateway PE to DC vCE Model	6
3.3	Hybrid inter-connection model	7
4.	Inter-connect IP VPN between DC and WAN	7
4.1	Existing Inter-AS options and DCI gap analysis	7
4.1.1	Option A pros and cons	7
4.1.2	Option B pros and cons	8
4.1.3	Option C pros and cons	9
4.1.4	Use of RTC	9
4.2	Additional work discussion	9
5.	Inter-connect IP VPN and non-IP VPN overlay networks	10
6.	Security Considerations	10
7.	IANA Considerations	11
8.	References	11
8.1	Normative References	11
8.2	Informative References	11
	Authors' Addresses	12

1 Introduction

BGP/MPLS IP Virtual Private Networks (IP VPNs) [RFC4364] has been the most extensively deployed, and the most scalable Service Provider (SP) provisioned VPN solutions over the past decade. This document is centered around inter-connecting various Clouds/Data Centers to the BGP/MPLS IP VPNs.

With the growth of cloud services, the needs for inter-connecting Data Centers and Enterprise BGP/MPLS IP VPNs in the Wide Area Network (WAN) become important. The interests are from multiple players: Service Providers who provide BGP/MPLS IP VPNs and may provide cloud service as well; cloud providers who provide cloud services but do not provide BGP/MPLS IP VPNs in the WAN; and the Enterprise users who use both BGP/MPLS IP VPNs services and cloud services.

This document discusses use cases of the inter-connection of BGP/MPLS VPN to Data Centers, the general requirements, and the proposed solutions for the inter-connections.

1.1 Terminology

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].

Term	Definition
AS	Autonomous System
ASBR	Autonomous System Border Router
BGP	Border Gateway Protocol
CE	Customer Edge
ED	End device: where Guest OS, Host OS/Hypervisor, applications, VMs, and virtual router may reside
GRE	Generic Routing Encapsulation
Hypervisor	Virtual Machine Manager
IaaS	Infrastructure as a Service
IRS	Interface to Routing System
LTE	Long Term Evolution
MP-BGP	Multi-Protocol Border Gateway Protocol
NVGRE	Network Virtualization using GRE
PCEF	Policy Charging and Enforcement Function
P	Provider backbone router
QoS	Quality of Service
RD	Route Distinguisher
RR	Route Reflector
RT	Route Target

RTC	RT Constraint
SDN	Software Defined Network
ToR	Top-of-Rack switch
VI	Virtual Interface
vCE	virtual Customer Edge Router
VM	Virtual Machine
VN	Virtual Network
vPC	virtual Private Cloud
vPE	virtual Provider Edge Router
VPN	Virtual Private Network
vRR	virtual Route Reflector
VXLAN	Virtual eXtensible Local Area Network
WAN	Wide Area Network

2. Use Cases

2.1 Case 1: End-to-end BGP IP VPN cloud inter-connection

Many Service Providers have large deployment base of BGP/MPLS IP VPNs. They are interested in extending the same style IP VPN capabilities into their Data Centers to provide end-to-end native BGP IP VPN services to their enterprise customers. The advantage is BGP IP VPN provides better security, richer policy control and QoS support when comparing with transport through the public Internet. The technologies developed to extend IP VPN into Data Center servers or ToR are described is virtual Provider Edge (vPE) [I-D.fang-l3vpn-virtual-pe],[I-D.ietf-l3vpn-end-system]. Regardless if the WAN and DC are managed by the same administrative domain or not (other the former), inter-connecting the two VPN segments are needed.

The interested parties are Service Providers and Enterprises.

2.2 Case 2: Hybrid cloud inter-connection

Service Providers are interested in extending their cloud VPNs to provide opportunities for enterprise customers using the new services provided by other cloud providers. The cloud providers are interested to extend their customer base through the SP Enterprise IP VPN customers. The inter-connection between the SP BGP/MPLS IP VPNs and the cloud provider networks is needed to accomplished the goal. The inter-connection of different types of providers most likely not BGP/MPLS IP VPN inter-connections, as the cloud providers may use any type of technologies in their networks, virtualized or non-virtualized. The two inter-connecting networks are under the administrative domains of different commercial entities. The task can be more challenging than IP/MPLS IP VPN Inter-provider connections which had always been challenging not from technology point of view.

We now add the dimension to inter-connecting VPN to various different technologies.

3. Architecture reference models

The architecture reference models described below are focusing on the inter-connection aspects. The intra DC implementation is not in discussion, but the intra DC technology has the direct impact to Inter-DC connection. Therefore, various models are illustrated.

3.1 BGP/MPLS IP VPN Inter-AS model

The BGP/MPLS IP VPNs are implemented in both the WAN network and the Data Center. A customer VPN, for example VPNA in figure 1, consists of enterprise remote sites and VMs supporting applications in the Data Center. The IP VPN implementation is using vPE technology. The two segment of the VPNs are inter-connected through ASBRs facing each other in the respective networks.

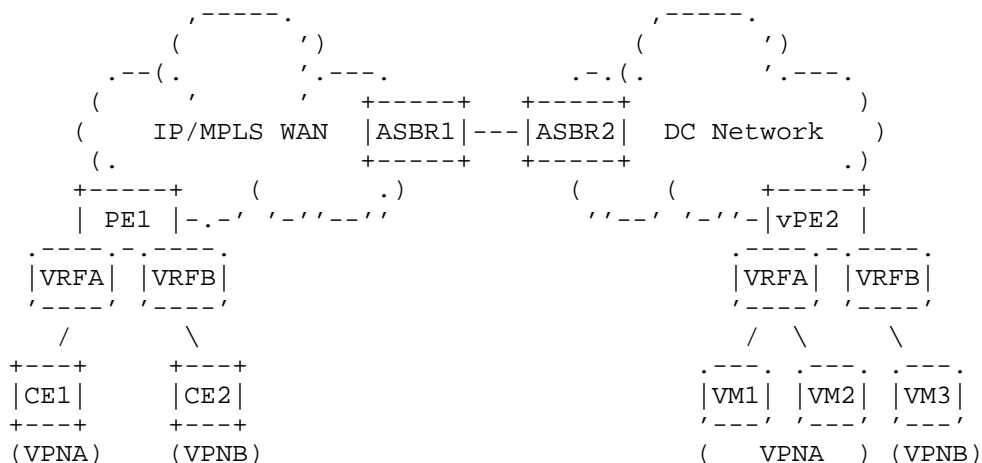


Figure 1. BGP/MPLS IP VPN Inter-Connection
with ASBR in each network

One boarding ASBR can be shared for the inter-connection of the two networks, especially if the WAN and DC belong to the same provider. Figure 2 illustrate this the shared ASBR model.

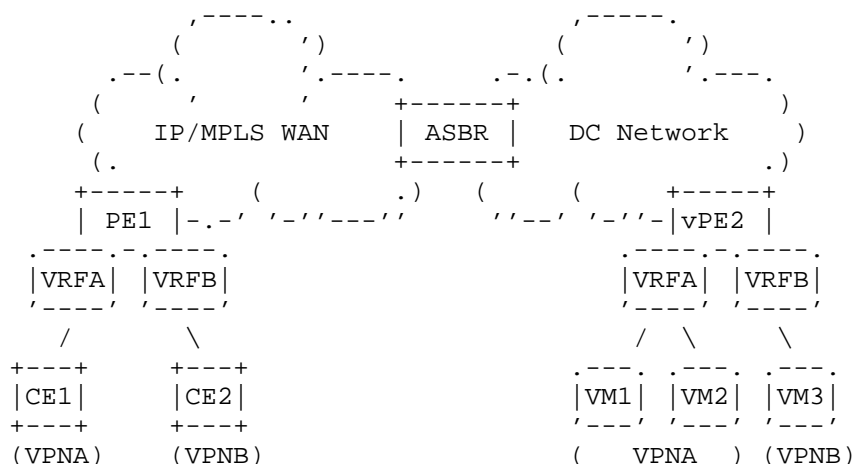


Figure 2. BGP/MPLS IP VPN Inter-Connection
with share ASBR

3.2 BGP/MPLS IP VPN Gateway PE to DC vCE Model

A simple virtual CE (vCE) [I-D.fang-l3vpn-virtual-ce] model can be used to inter-connect client containers to the DC Gateway which function as PE. This model are used SPs to provide managed services, when scale can meet the service requirement.

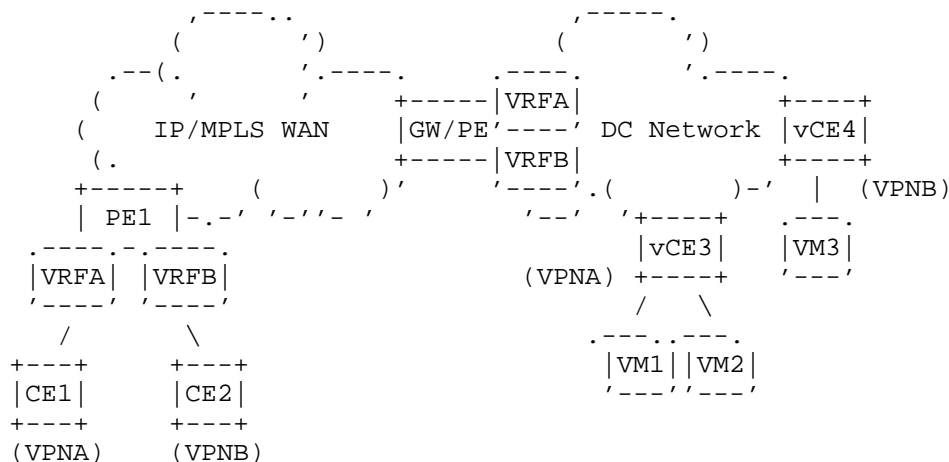


Figure 3. BGP/MPLS IP VPN GW/PE to vCEs
without BGP/MPLS IP VPN in the DC

3.3 Hybrid inter-connection model

The BGP/MPLS IP VPNs are implemented in the WAN network, and non BGP/MPLS IP VPN Overlay in DC. The connection of the two networks are outside of the technologies for Inter-AS connections for BGP IP VPNs. This model include many variations depending on the specific technologies used in the DC overlay. Figure 4 provides a general view of this inter-connecting model with ASBR on the MPLS WAN side, and the DC GW on the DC side. It is also viable to use one shared ASBR/GW for the inter-connection, especially if the WAN and the DC belong to the same provider.

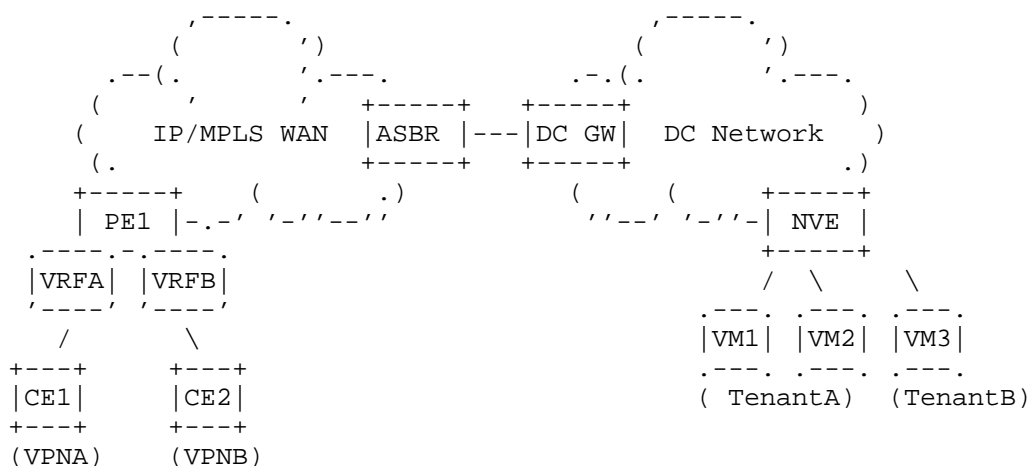


Figure 4. BGP/MPLS IP VPN Inter-Connection with non BGP/MPLS IP VPN Overlay in DC

4. Inter-connect IP VPN between DC and WAN

4.1 Existing Inter-AS options and DCI gap analysis

The inter-AS options described in [RFC4364] can be used for DC inter-connection. Option A, B, and C must be supported.

4.1.1 Option A pros and cons

In Option A: back-to-back VRF. The PE-ASBR in one AS performs MPLS or IP VPN decapsulation and transmits packets to the peer PE-ASBR in the adjacent autonomous system. The peer PE-ASBR performs MPLS or IP VPN encapsulation on the customer IPv4/IPv6 packets received, and transmits the packet through the IPv4 backbone of the autonomous system. VPN service providers exchange routes across a back-to-back VRF connection. Each VRF instance represents a separate VPN client,

and it is configured on a separate PE-ASBR interface, allowing a PE-ASBR to communicate with its peer PE-ASBR as if the peer was a CE router.

Pros: It is the most secure option among options A, B, and C. And it is the simplest model from operation perspective. Each PE-ASBR is treating the other as a CE.

Cons: Scaling limitation, because per Inter-AS VPN VRF and interface are needed on the PE-ASBR.

Option A has been used in Inter-Provider inter-connections because the security consideration and clear operational demarcation.

DCI considerationa: It is a simple way to connect DC and WAN if both sides are small scale. Scale will be the major concern for DC inter-connect if large scale support is needed. Even if the DC scale is small, there are major concerns on receiving relevant routes (potentially large number) from the WAN side.

4.1.2 Option B pros and cons

In Option B: EBGp redistribution of labeled VPN-IPv4/IPv6 routes between the neighboring ASes. ASes exchange VPN routing information (routes and labels) to establish connections. To control connections between autonomous systems, the PE routers and EBGp border edge routers maintain a label forwarding information base (LFIB). The LFIB manages the labels and routes that the PE routers and EBGp border edge routers receive during the exchange of VPN information. The autonomous systems use the following guidelines to exchange VPN routing information: Routing information: The destination network, the next hop field associated with the distributing router, a local MPLS label; An RD: route distinguisher; ASBRs are configured to change the next hop (next-hop-self) when sending VPN-IPv4 NLRI to the IBGP neighbors, the ASBRs must allocate a new label when they forward the NLRI to the IBGP neighbors.

Pros: It provides better scale than option A does, as it removes the needs of per Inter-AS VPN VRF and interface on the ASBR.

Cons: vanilla version of Option B is considered less secure in comparison with Option A, due to dynamic routing information exchange is involved. The ASBR scaling may still be issues because ASBR must maintain all VPN routes.

Option B is commonly used within single provider or for inter-provider connections.

DCI consideration: Option B is one viable option to be used in DC inter-connection. But it has the same scale concerns as other options on potential large number of routes exchange between WAN and DC.

4.1.3 Option C pros and cons

In option C: Multihop eBGP Redistribution of Labeled VPN-IPv4 Routes Between Source and Destination ASs, with eBGP Redistribution of Labeled IPv4 Routes from AS to Neighboring AS. The ASBRs need only to exchange host routes (/32 or /128) to the PE routers involved in the VPN, with the labels needed to get there. A Label Switch Path (LSP) is built from the ingress PE router in one AS to the egress PE in the other AS (using loopback addresses). VPN traffic uses this LSP to reach the other AS. From data plane is perspective, the ASBRs act as P routers, with no knowledge about the VPNs concerned. Between ASBRs the VPN traffic looks like traffic between P routers: each data packet is pre-pended with the VPN label and then with an egress-PE label. Option C can be further scaled by using route reflectors (RRs) in each AS.

Pros: It is the most scalable option among all. ASBR is no longer a bottle neck for VPN routes scaling as in Option B.

Cons: Major security issues as IGP reachability need to be exchanged between the neighboring ASes.

Option C has been used within a single SP for inter-AS connections. Using RR for VPN routes exchange is the common approach.

DCI consideration: Option C should not be used for any DCI which is between two different providers. In addition, even though ASBR is off the burden of scaling VPN routes, VRFs, VPN interfaces. The VPN routes are still exchanged between the two ASes.

4.1.4 Use of RTC

RT constraint [RFC4684] function must be used to only distribute the IP VPN routes of a VPN from one AS to another under the condition that they both support that VPN in each of the AS. This is one most important function for scaling the solution.

But all IP VPN routes are exchanged between the two ASes (e.g. WAN and DC) as long as they have support the same VPNs. The potential IP VPN routes distribution can still be very substantial in large WAN and DC deployment.

4.2 Additional work discussion

Focus is very large scale DCI. To be added.

5. Inter-connect IP VPN and non-IP VPN overlay networks

As one significant instance of the hybrid use-case described in section 2.2, a DC may support a multi-tenant virtualized service network using IP based DC overlay encapsulations such as VXLAN [I-D.mahalingam-dutt-dcops-vxlan] or NVGRE [I-D.sridharan-virtualization-nvgre]. Different deployment models may be used within the DC depending on the DC provider's functional and operational requirements.

When an IP DC overlay is terminated at the DC gGateway router and traffic directed into an MPLS IP VPN. The DC Gateway router performs MPLS encapsulation towards the WAN and IP overlay based forwarding within the DC.

In this case, the inter-connection mechanisms between the DC and the WAN may fall into two categories:

1. VRF Termination

The overlay based virtual network terminates into a BGP IP VPN VRF at the DC-WAN Gateway router. Both the internal routes of the DC as well as the external routes received from the WAN router can be installed in the VRF forwarding table at the DC gateway router. The DC gateway will perform an IP lookup and forward traffic after doing the appropriate MPLS or IP encapsulation.

The DC gateway router will peer with the WAN router using one of the existing inter-AS mechanisms described above. The DC Gateway functions as an IP-VPN ASBR with local VRFs, for example packets still undergo an IP forwarding lookup.

2. DC-VN and IP VPN Inter-working

In this case, the DC Gateway router performs a direct translation between VN-IDs and IP VPN labels while switching packets between the DC and WAN interfaces without performing an IP lookup. The forwarding table at the DC Gateway router is set up to do a VN-ID or label lookup and derive the output label or VN-ID. The DC Gateway Router acts as an Inter-AS Option B ASBR peering with other ASBRs.

6. Security Considerations

To be added.

7. IANA Considerations

None.

8. References

8.1 Normative References

- [RFC2119] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, March 1997.
- [RFC3032] Rosen, E., Tappan, D., Fedorkow, G., Rekhter, Y., Farinacci, D., Li, T., and A. Conta, "MPLS Label Stack Encoding", RFC 3032, January 2001.
- [RFC4023] Worster, T., Rekhter, Y., and E. Rosen, Ed., "Encapsulating MPLS in IP or Generic Routing Encapsulation (GRE)", RFC 4023, March 2005.
- [RFC4271] Rekhter, Y., Ed., Li, T., Ed., and S. Hares, Ed., "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.
- [RFC4684] Marques, P., Bonica, R., Fang, L., Martini, L., Raszuk, R., Patel, K., and J. Guichard, "Constrained Route Distribution for Border Gateway Protocol/MultiProtocol Label Switching (BGP/MPLS) Internet Protocol (IP) Virtual Private Networks (VPNs)", RFC 4684, November 2006.
- [RFC5036] Andersson, L., Ed., Minei, I., Ed., and B. Thomas, Ed., "LDP Specification", RFC 5036, October 2007.

8.2 Informative References

- [I-D.ietf-l3vpn-end-system] Marques, P., Fang, L., Pan, P., Shukla, A., Napierala, M., "BGP-signaled end-system IP/VPNs", draft-ietf-l3vpn-end-system-00, October 2012.
- [I-D.fang-l3vpn-virtual-pe] Fang, L., Ward, D., Fernando, R.,

Napierala, M., Bitar, N., Rao, D., Rijsman, B., So, N.,
"BGP IP VPN Virtual PE", draft-fang-l3vpn-virtual-pe-00,
Feb. 2013.

[I-D.fang-l3vpn-virtual-ce] Fang, L., Evans, J., Ward, D., Fernando,
R., Mullooly, J., So, N., Bitar., N., Napierala, M., "BGP
IP VPN Virtual PE", draft-fang-l3vpn-virtual-ce-01, Feb.
2013.

[I-D.fang-l3vpn-end-system-req] Napierala, M., and Fang, L.,
"Requirements for Extending BGP/MPLS VPNs to End-Systems",
draft-fang-l3vpn-end-system-requirements-00, Oct. 2012.

[I-D.ward-irs-framework] Atlas, A., Nadeau, T., Ward, D., "Interface
to the Routing System Framework", draft-ward-irs-
framework-00, July 2012.

[I-D.rfernando-irs-fw-req] Fernando, R., Medved, J., Ward, D., Atlas,
A., Rijsman, B., "IRS Framework Requirements", draft-
rfernando-irs-framework-requirement-00, Oct. 2012.

[I-D.mahalingam-dutt-dcops-vxlan]: Mahalingam, M, Dutt, D., et al.,
"A Framework for Overlaying Virtualized Layer 2 Networks
over Layer 3 Networks" draft-mahalingam-dutt-dcops-vxlan-
03, Feb. 2013.

[I-D.sridharan-virtualization-nvgre]: SridharanNetwork, M., et al.,
"Virtualization using Generic Routing Encapsulation",
draft-sridharan-virtualization-nvgre-02.txt, Feb. 2013.

Authors' Addresses

Luyuan Fang
Cisco
111 Wood Ave. South
Iselin, NJ 08830
Email: lufang@cisco.com

Rex Fernando
Cisco
170 W Tasman Dr
San Jose, CA
Email: rex@cisco.com

Dhananjaya Rao
Cisco
170 W Tasman Dr

INTERNET DRAFT

<BGP IP VPN DCI>

<February 25, 2013>

San Jose, CA
Email: dhr Rao@cisco.com

Sami Boutros
Cisco
170 W Tasman Dr
San Jose, CA
Email: dhr Rao@cisco.com

INTERNET-DRAFT
Intended Status: Standards track
Expires: January 15, 2014

Ning So
TATA Communications
Jim Guichard
Cisco
Wen Wang
CenturyLink
Manuel Paul
Deutsche Telekom

Luyuan Fang
David Ward
Rex Fernando
Cisco
Maria Napierala
AT&T
Nabil Bitar
Verizon
Dhananjaya Rao
Cisco
Bruno Rijsman
Juniper

July 15, 2013

BGP IP VPN Virtual PE
draft-fang-l3vpn-virtual-pe-03

Abstract

This document describes the architecture solutions for BGP/MPLS IP Virtual Private Networks (VPNs) with virtual Provider Edge (vPE) routers. It provides a functional description of the vPE control plane, the data plane, and the provisioning management process. The vPE solutions supports both Software Defined Networking (SDN) approach by allowing physical decoupling of the control and the forwarding plane of a vPE, as well as a distributed routing approach. These solutions allow vPE to be co-resident with the application virtual machines (VMs) on a single end device, such as a server, as well as on a Top-of-Rack switch (ToR), or in any network or compute device. The ability to provide end-to-end native BGP IP VPN connections between a Data Center (DC) (and/or other types of service network) applications and the Enterprise IP VPN sites is highly desirable to both Service Providers and Enterprises.

Status of this Memo

This Internet-Draft is submitted to IETF in full conformance with the provisions of BCP 78 and BCP 79.

Internet-Drafts are working documents of the Internet Engineering Task Force (IETF), its areas, and its working groups. Note that other groups may also distribute working documents as Internet-Drafts.

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."

The list of current Internet-Drafts can be accessed at
<http://www.ietf.org/lid-abstracts.html>

The list of Internet-Draft Shadow Directories can be accessed at
<http://www.ietf.org/shadow.html>

Copyright and License Notice

Copyright (c) 2013 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	4
1.1	Terminology	4
1.2	Motivation and requirements	5
2.	Virtual PE Architecture	6
2.1	Virtual PE definitions	6
2.2	vPE Architecture and Design options	7
2.2.1	vPE-F host location	7
2.2.2	vPE control plane topology	7
2.2.3	Data Center orchestration models	8
2.3	vPE Architecture reference models	8
2.3.1	vPE-F in an end-device and vPE-C in the controller	8
2.3.2	vPE-F and vPE-C on the same end-device	9
2.3.3	vPE-F and vPE-C are on the ToR	10
2.3.4	vPE-F on the ToR and vPE-C on the controller	12
2.3.5	Server view of vPE	12
3.	Control Plane	13
3.1	vPE Control Plane (vPE-C)	13

3.1.1 SDN approach	13
3.1.2 Distributed control plane	14
3.3 Use of router reflector	14
3.4 Use of Constrained Route Distribution [RFC4684]	14
4. Forwarding Plane	14
4.1 Virtual Interface	14
4.2 Virtual Provider Edge Forwarder (vPE-F)	15
4.3 Encapsulation	15
4.4 Optimal forwarding	16
4.5 Routing and Bridging Services	16
5. Addressing	17
5.1 IPv4 and IPv6 support	17
5.2 Address space separation	17
6.0 Inter-connection considerations	17
7. Management, Control, and Orchestration	19
7.1 Assumptions	19
7.2 Management/Orchestration system interfaces	19
7.3 Service VM Management	20
7.4 Orchestration and IP VPN inter-provisioning	20
7.4.1 vPE Push model	20
7.4.2 vPE Pull model	21
7. Security Considerations	22
8. IANA Considerations	22
9. References	22
9.1 Normative References	22
9.2 Informative References	23
Authors' Addresses	24

1 Introduction

Network virtualization enables multiple isolated individual networks over a shared common network infrastructure. BGP/MPLS IP Virtual Private Networks (IP VPNs) [RFC4364] have been widely deployed to provide network based IP VPNs solutions. [RFC4364] provides routing isolation among different customer VPNs and allow address overlapping among these VPNs through the implementation of per VPN Virtual Routing and Forwarding instances (VRFs) at a Service Provider Edge (PE) routers, while forwarding customer traffic over a common IP/MPLS network infrastructure.

With the advent of compute capabilities and the proliferation of virtualization in Data Center servers, a multi-tenant Data Center becomes a reality. As applications and appliances are increasingly being virtualized, support for virtual edge devices, such as virtual IP VPN PE routers, becomes feasible and a natural part of the overall virtualization solutions. There is a strong desire from Service Providers to extend their existing BGP IP VPN deployments into Data Centers to provide Virtual Private Cloud (VPC) services, as well as to support virtual network functions, including IP VPN PE functions outside of Data Centers. Scale and efficiency are crucial factors in the cloud computing environment supporting various applications and services, and in traditional service provider space.

The virtual Provider Edge (vPE) solution described in this document allows for the extension of the PE functionality of BGP/MPLS IP VPN to end devices, such as servers where the applications reside, or to the first hop routing/switching device, such as a Top of the Rack switch (ToR) in a Data Center.

The vPE solutions support both Software Defined Network (SDN) approach by allowing physical decoupling of the control and the forwarding plane of a vPE, and distributed routing approaches in the same fashion as IP VPN is achieved with the physical PEs.

1.1 Terminology

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].

Term	Definition
-----	-----
3GPP	3rd Generation Partnership Project (3GPP)
AS	Autonomous System

ASBR	Autonomous System Border Router
BGP	Border Gateway Protocol
CE	Customer Edge
ED	End device: where Guest OS, Host OS/Hypervisor, applications, VMs, and virtual router may reside
Forwarder	L3VPN forwarding function
GRE	Generic Routing Encapsulation
Hypervisor	Virtual Machine Manager
I2RS	Interface to Routing Systems
IaaS	Infrastructure as a Service
LDP	Label Distribution Protocol
LTE	Long Term Evolution
MP-BGP	Multi-Protocol Border Gateway Protocol
PCEF	Policy Charging and Enforcement Function
P	Provider backbone router
QoS	Quality of Service
RR	Route Reflector
RT	Route Target
RTC	RT Constraint
SDN	Software Defined Network
ToR	Top-of-Rack switch
VI	Virtual Interface
vCE	virtual Customer Edge Router
VM	Virtual Machine
vPC	virtual Private Cloud
vPE	virtual Provider Edge Router
vPE-C	virtual Provider Edge Control plane
vPE-F	virtual Provider Edge Forwarder
VPN	Virtual Private Network
vRR	virtual Route Reflector
WAN	Wide Area Network

1.2 Motivation and requirements

The recent rapid adoption of Cloud Services by Enterprises and the phenomenal growth of mobile IP applications accelerate the need to extend the BGP IP VPN capability into cloud service end devices. For examples, Enterprise customers' want to extend the existing IP VPN services in the WAN into the new cloud services supported by various Data Center (DC) technologies; Large Enterprises have existing L3VPN deployments and are extending them into their Data Centers; Mobile providers adopting IP VPN into their 3GPP mobile infrastructure are looking to extend the IP VPNs to their end devices of the call processing center. In addition, vPE is one of the earlier work as part of Network Function Virtualization (NfV) effort, where IP VPN PE function is one of the network functions subject to virtualization. In general, the vPE solutions can be used in cloud service development or any other environment with or without the inter-

connection to existing Enterprise BGP IP VPNs.

Key requirements for vPE solutions:

- 1) MUST support end device multi-tenancy, per tenant routing isolation and traffic separation.
- 2) MUST support large scale IP VPNs in the Data Center, upto tens of thousands of end devices and millions of VMs in the single Data Center.
- 3) MUST support end-to-end IP VPN connectivity, e.g. IP VPN can start from a Data Center end device, connect to a corresponding IP VPN in the WAN, and terminate in another Data Center end device.
- 4) MUST allow physical decoupling of IP VPN PE control plane and forwarding for network virtualization and abstraction.
- 5) MUST provide support of the control plane through a SDN controller (centralized or distributed), as well as through the traditional distributed MP-BGP approach.
- 6) MUST support VM mobility
- 7) MUST support orchestration/provisioning as a key deployment model
- 8) SHOULD be capable to support service chaining as part of the solution [I-D.rfernando-l3vpn-service-chaining], [I-D.bitar-i2rs-service-chaining].

The architecture and protocols defined in BGP/MPLS IP VPN [RFC4364] provide the foundation for virtual PE extension. Certain protocol extensions may be needed to support the virtual PE solutions.

2. Virtual PE Architecture

2.1 Virtual PE definitions

As defined in [RFC4364], an IP VPN is created by applying policies to form a subset of sites among all sites connected to the backbone network. It is a collection of "sites". A site can be considered as a set of IP systems maintaining IP inter-connectivity without direct connecting through the backbone. The typical use of L3VPM has been to inter-connect different sites of an Enterprise networks through a Service Provider's BGP IP VPNs in the WAN.

A virtual PE (vPE) is a BGP IP VPN PE software instance which may reside in any network or computing devices. The control and

forwarding components of the vPE can be decoupled, they may reside in the same physical device, or most often in different physical devices.

A virtualized Provider Edge Forwarder (vPE-F) is the forwarding element of a vPE. vPE-F can reside in an end device, such as a server in a Data Center where multiple application Virtual Machines (VMs) are supported, or a Top-of-Rack switch (ToR) which is the first hop switch from the Data Center edge. When a vPE-F is residing in a server, its connection to a co-resident VM is as the same as the PE-CE relationship in the regular BGP IP VPNs, but without routing protocols or static routing between the virtual PE and CE because the connection is internal to the device.

The vPE Control plane (vPE-C) is the control element of a vPE. When using the approach where control plane is decoupled from the physical topology, the vPE-F may be in a server and co-resident with application VMs, while one vPE-C can be in a separate device, such as an SDN Controller where control plane elements and orchestration functions are located. Alternatively, the vPE-C can reside in the same physical device as the vPE-F. In this case, it is similar to the traditional implementation of VPN PEs where, distributed MP-BGP is used for IP VPN information exchange, though the vPE is not a dedicated physical entity as it is in a physical PE implementation.

2.2 vPE Architecture and Design options

2.2.1 vPE-F host location

Option 1a. vPE-F is on an end device as co-resident with application VMs. For example, the vPE-F is on a server in a Data Center.

Option 1b. vPE-F forwarder is on a ToR or other first hop devices in a Data Center, not as co-resident with the application VMs.

Option 1c. vPE-F is located on any network or compute devices in any type of networks.

2.2.2 vPE control plane topology

Option 2a. vPE control plane is physically decoupled from the vPE-F. The control plane may be located in a controller in a separate device (a stand alone device or can be in the gateway as well) from the vPE forwarding plane.

Option 2b. vPE control plane is supported through dynamic routing protocols and located in the same physical device as the vPE-F.

2.2.3 Data Center orchestration models

Option 3a. Push model: It is a top down approach, push IP VPN provisioning state from a network management system or other centrally controlled provisioning system to the IP VPN network elements.

Option 3b. Pull model: It is a bottom-up approach, pull state information from network elements to network management/AAA based upon data plane or control plane activity.

2.3 vPE Architecture reference models

2.3.1 vPE-F in an end-device and vPE-C in the controller

Figure 1 illustrates the reference model for a vPE solution with the vPE-F in the end device co-resident with applications VMs, while the vPE-C is physically decoupled and residing on a controller.

The Data Center is connected to the IP/MPLS core via the Gateways/ASBRs. The IP VPN, e.g. VPN RED, has a single termination point within the Data Center at one of the VPE-F, and is interconnected in the WAN to other member sites which belong to the same client, and the remote ends of VPN RED can be a PE which has VPN RED attached to it, or another vPE in a different Data Center.

Note that the Data Center fabric/intermediate underlay devices in the Data Center do not participate IP VPNs, their function is the same as P routers in the IP/MPLS back bone and they do not maintain the IP VPN states, not IP VPN aware.

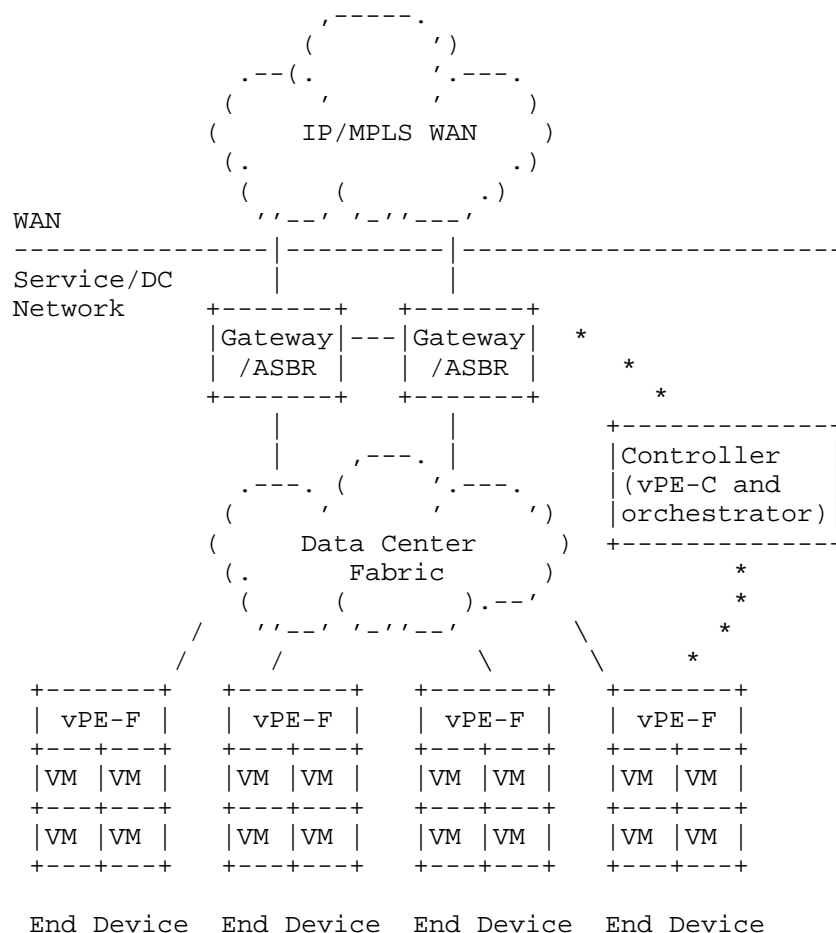


Figure 1. Virtualized Data Center with vPE at the end device and vPE-C and vPE-F physically decoupled

Note:

a) *** represents Controller logical connections to the all Gateway/ASBRs and to all vPE-F.

b) ToR is assumed included in the Data Center cloud.

2.3.2 vPE-F and vPE-C on the same end-device

In this option, vPE-F and vPE-C functionality are both resident in the end-device. The vPE functions the same as it is in a physical PE. MP-BGP is used for the VPN control plane. Virtual or physical Route

Reflectors (RR) (not shown in the diagram) can be used to assist scaling.

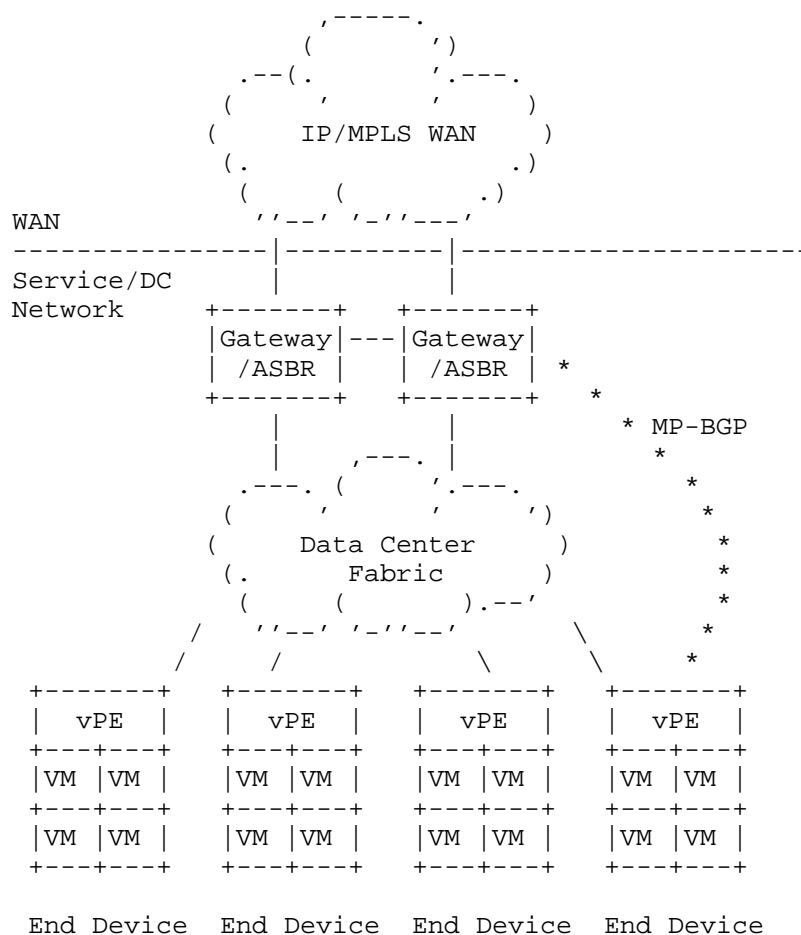


Figure 2. Virtualized Data Center with vPE at the end device, VPN control signal uses MP-BGP

Note:

a) *** represents the logical connections using MP-BGP among the Gateway/ASBRs and to the vPEs on the end devices.

b) ToR is assumed included in the Data Center cloud.

2.3.3 vPE-F and vPE-C are on the ToR

In this option, vPE functionality is the same as a physical PE. MP-BGP is used for the VPN control plane. Virtual or physical Route Reflector (RR) (not shown in the diagram) can be used to assist scaling.

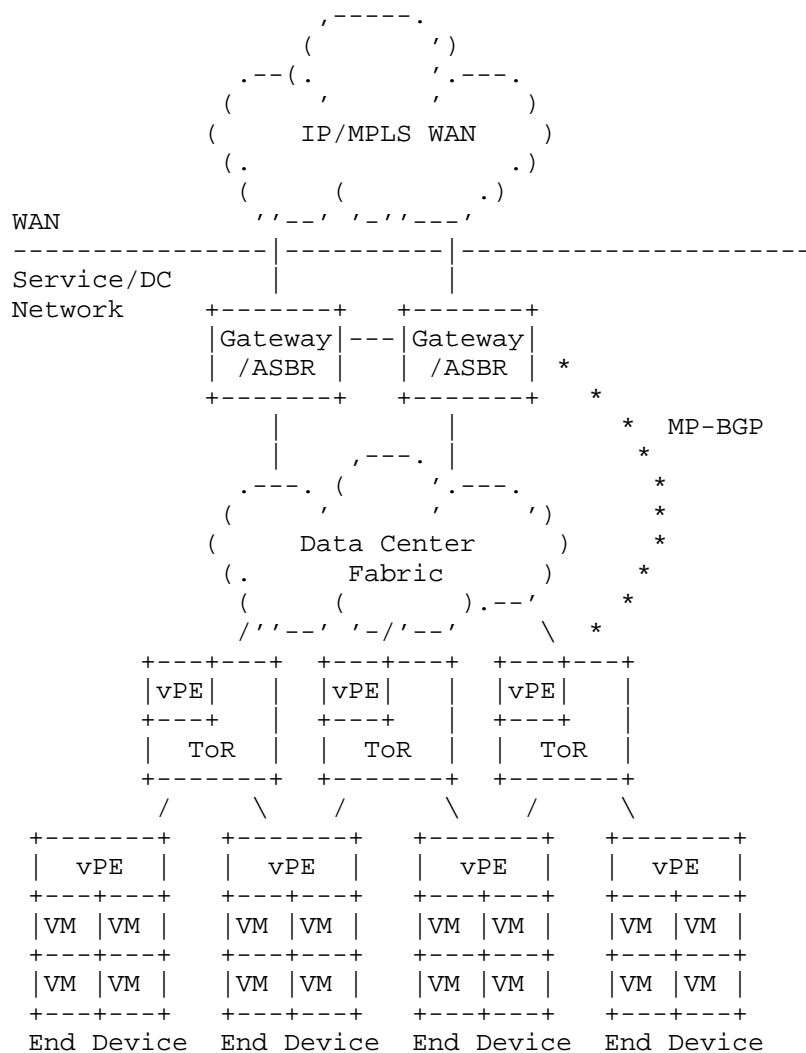


Figure 3. Virtualized Data Center with vPE at the ToP, VPN control signal uses MP-BGP

Note: *** represents the logical connections using MP-BGP among the Gateway/ASBRs and to the vPEs on the ToRs.

2.3.4 vPE-F on the ToR and vPE-C on the controller

In this option, the L3VPN termination is at the ToR, but the control plane decoupled from the data plane and resided in a controller, which can be on a stand alone device, or can be placed at the Gateway/ASBR.

2.3.5 Server view of vPE

An end device shown in Figure 4 is a virtualized server that hosts multiple VMs. The virtual PE is co-resident in the server. The vPE supports multiple VRFs, VRF Red, VRF Grn, VRF Yel, VRF Blu, etc. Each application VM is associated to a particular VRF as a member of the particular VPN. For example, VM1 is associated to VRF Red, VM2 and VM47 are associated to VRF Grn, etc. Routing isolation applies between VPNs for multi-tenancy support. For example, VM1 and VM2 cannot communicate directly in a simple intranet L3VPN topology as shown in the configuration.

The vPE connectivity relationship between vPE and the application VM is similar to the PE-to-CE relationship in a regular BGP IP VPNs. However, as the vPE and CE functions are co-resident in the same server, the connection between them is an internal implementation of the server.

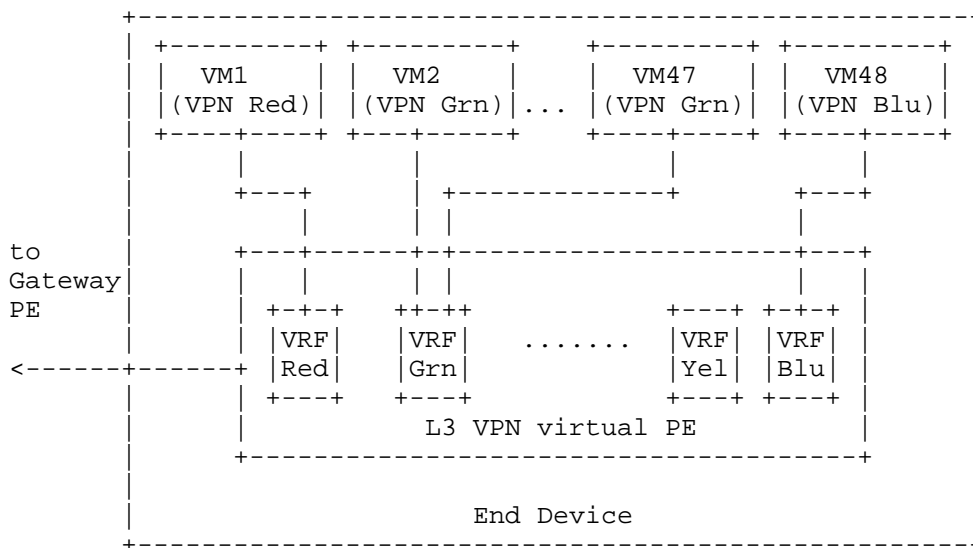


Figure 4. Server View of vPE to VM relationship

An application VM may send packets to a vPE forwarder that need to be bridged, either locally to another VM, or to a remote destination. In this case, the vPE contains a virtual bridge instance to which the application VMs (CEs) are attached.

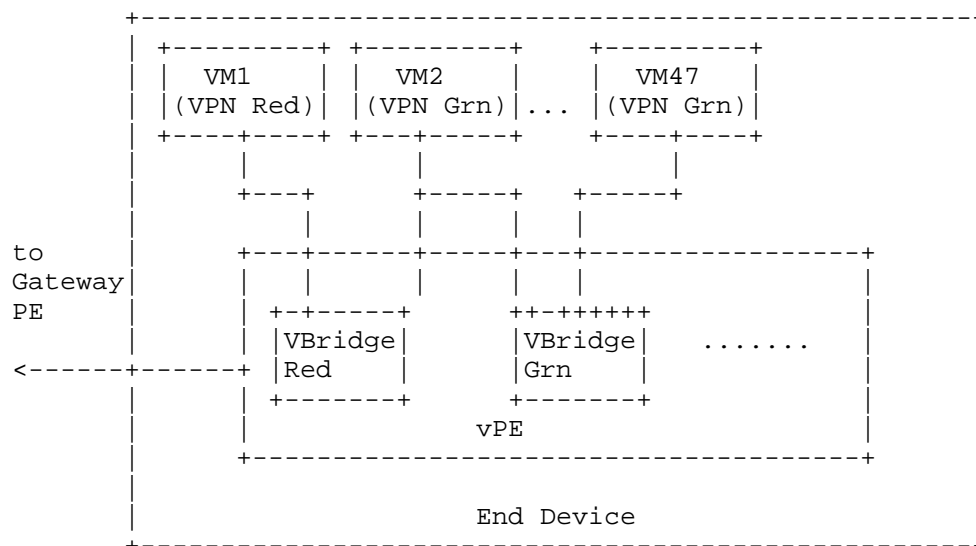


Figure 4. Bridging Service at vPE

3. Control Plane

3.1 vPE Control Plane (vPE-C)

The vPE control plane functionality MAY use a SDN controller or be distributed using MP-BGP.

3.1.1 SDN approach

This approach is appropriate when the vPE control and data planes are physically decoupled. The control plane directing the data flow may reside elsewhere, such a SDN controller. This approach requires a standard interface to the routing system. The Interface to Routing System (I2RS) is work in progress in IETF as described in [I-D.atlas-i2rs-architecture], [I-D.rfernando-irs-fw-req].

Although MP-BGP is often the de facto preferred choice between vPE and gateway-PE/ASBR, the use of extensible signaling messaging protocols MAY often be more practical in a Data Center environment. One such proposal that uses this approach is detailed in [I-D.ietf-l3vpn-end-system].

3.1.2 Distributed control plane

In the distributed control plane approach, the vPE participates in the overlay L3VPN control protocol: MP-BGP [RFC4364].

When the vPE function is on a ToR, it participates the underlay routing through IGP protocols: ISIS or OSPF.

When the vPE function is on a server, it functions as a host attached to a server.

3.3 Use of router reflector

Modern Data Centers can be very large in scale. For example, the number of VPNs routes in a very large Data Centers can surpass the scale of those in a Service Provider backbone VPN network. There may be tens of thousands of end devices in a single Data Center.

Use of Router Reflector (RR) is necessary in large-scale L3VPN networks to avoid a full iBGP mesh among all vPEs and PEs. The L3 VPN routes can be partitioned to a set of RRs, the partitioning techniques are detailed in [RFC4364].

When the RR is residing in a physical device, e.g., a server, which is partitioned to support multi-functions and applications VMs, the RR becomes a virtualized RR (vRR). Since RR's performs control plane only, a physical or virtualized server with large scale of computing power and memory can be a good candidate as host of vRRs. The vRR can also reside be in Gateway PE/ASBR, or in an end device.

3.4 Use of Constrained Route Distribution [RFC4684]

The Constrained Route Distribution [RFC4684] is a powerful tool for selective L3VPN route distribution. Using this functionality, only the BGP receivers (e.g, PE/vPE/RR/vRR/ASBRs, etc.) with the particular L3VPNs attached will receive the route update for the corresponding VPNs. It is critical to use constrained route distribution to support large-scale L3VPN developments.

4. Forwarding Plane

4.1 Virtual Interface

A Virtual Interface (VI) is an interface within an end device that is used for connection of the vPE to the application VMs in the same end device. Such application VMs are treated as CEs in the regular L3VPN's view.

4.2 Virtual Provider Edge Forwarder (vPE-F)

The Virtual Provider Edge Forwarder (vPE-F) is the forwarding component of a vPE where the tenant identifiers (for example MPLS VPN labels) are pushed/popped.

The vPE-F location options include:

- 1) Within the end device where the virtual interface and application VMs are located.
- 2) In an external device such as a Top of the Rack switch (ToR) in a Data Center into which the end device connects.

Multiple factors should be considered for the location of the vPE-F, including device capabilities, overall solution economics, QoS/firewall/NAT placement, optimal forwarding, latency and performance, operational impact, etc. There are design tradeoffs, it is worth the effort to study the traffic pattern and forwarding looking trend in your own unique Data Center as part of the exercise.

4.3 Encapsulation

There are two existing standardized encapsulation/forwarding options typically used for BGP/MPLS L3VPN.

1. MPLS label stack encoding with Label Distribution Protocol [LDP], [RFC3032][RFC5036].
2. Encapsulating MPLS packets in IP or Generic Routing Encapsulation (GRE), [RFC4023], [RFC4797].
3. Other types of encapsulation are possible. For example, VXLAN [I-D.mahalingam-dutt-dcops-vxlan], NVGRE [I-D.sridharan-virtualization-nvgre], and other modified version of these or other existing protocols.

The most common BGP/MPLS L3VPNs deployments in Service Provider networks use MPLS forwarding. This requires that an MPLS transport, e.g., Label Switched Protocol (LDP) [RFC5036] to be deployed in the network. It is proven to scale, and it comes with various security mechanisms to protect the network against attacks.

However, the Data Center environment is different than the Service Provider VPN networks or large Enterprise backbones. MPLS deployments may or may not be feasible or desirable. Two major challenges for MPLS deployments exist in this new environment: 1) the capabilities of the end devices and the transport/forwarding devices; 2) the

workforce skill set.

Encapsulating MPLS in IP or GRE tunnel [RFC4023] may often be more practical in most Data Center, and computing environments. Note that when IP encapsulations are used, the associated security considerations must be analyzed carefully.

In addition, there are new encapsulation proposals for Data Centers currently as work in progress within IETF, including several UDP based encapsulations proposals and some TCP based proposal. These overlay encapsulations can be suitable alternatives for a vPE, considering the availability and leverage of support in virtual and physical devices.

4.4 Optimal forwarding

Many large cloud service operators have reported that the traffic patterns in their Data Centers were dominated by East-West across subnet traffic (between the end device hosting different applications in different subnets) rather than North-South traffic (going in/out of the Data Center and to/from the WAN) or switched traffic within subnets. This is the primary reason that many new large scale design has moved away from traditional Layer-2 design to Layer-3, especially for the overlay networks.

When forwarding the traffic within the same VPN, the vPE SHOULD be able to provide direct communication among the VMs/application senders/receivers without the need of going through gateway devices. If it is on the same end device, the traffic should not need to leave the same device. If it is on different end device, optimal routing SHOULD be applied.

When multiple VPNs need to be accessed to the end device virtual interfaces CAN directly access multiple VPNs via using extranet VPN techniques without the need of Gateway facilitation. This is done through the use of BGP L3VPN policy control mechanisms to support this function. In addition, ECMP is a built in layer-3 mechanism and it is used for load sharing.

Optimal use of available bandwidth can be achieved by virtue of using ECMP in the underlay, as long as the encapsulation includes certain entropy in the header (e.g. VXLAN).

4.5 Routing and Bridging Services

A VPN forwarder (vPE-F) may support both IP forwarding as well as Layer-2 bridging for traffic from attached end hosts. This traffic may be between end hosts attached to the same VPN forwarder or to

different VPN forwarders.

In both cases, forwarding at a VPN forwarder takes place based on the IP or MAC route entries provisioned by the VPE controller.

When the vPE is providing a Layer-3 service to attached CEs, the VPN forwarder will have a VPN VRF instance with IP routes installed for both locally attached end-hosts and ones reachable via other VPN forwarders. The vPE may perform IP routing for all IP packets in this mode.

When the vPE provides a Layer-2 service to attached end-hosts, the VPN forwarder will have an E-VPN instance with appropriate MAC entries.

The vPE may support an Integrated Routing and Bridging service, in which case the relevant VPN forwarders will have both MAC and IP table entries installed, and will appropriately route or switch incoming packets.

The vPE controller does the necessary provisioning to support various services, as defined by an user.

5. Addressing

5.1 IPv4 and IPv6 support

IPv4 and IPv6 MUST be supported in the vPE solution.

This may present a challenge for older devices, but may not be an issue for newer forwarding devices and servers. A server is replaced much more frequently than a network router/switch and newer equipment should be capable of IPv6 support.

5.2 Address space separation

The addresses used for the IP VPN overlay in a Data Center, SHOULD be taken from separate address blocks than the ones used for the underlay infrastructure of the Data Center. This practice is to protect the Data Center infrastructure from being attacked if the attacker gains access to the tenant VPNs.

Similarity, the addresses used for the Data Center, e.g., a Data Center, SHOULD be separated from the WAN backbone addresses space.

6.0 Inter-connection considerations

The inter-connection considerations in this section are focused on

intra-DC inter-connections.

There are deployment scenarios where IP VPN may not be supported in every segment of the networks to provide end-to-end IP VPN connectivity. An IP VPN vPE may be reachable only via an intermediate inter-connecting network; interconnection may be needed in these cases.

When multiple technologies are employed in the overall solution, a clear demarcation should be preserved at the inter-connecting points. The problems encountered in one domain should not impact other domains.

From an IP VPN point of view: An IP VPN vPE that implements [RFC4364] is a component of the IP VPN network only. An IP VPN VRF on a physical PE or vPE contains IP routes only, including routes learnt over the locally attached network.

As described earlier in this document, the IP VPN vPE should ideally be located as close to the "customer" edge devices. For cases where this is not possible, simple existing "IP VPN CE connectivity" mechanisms should be used, such as static, or direct VM attachments such as described in the vCE [I-D.fang-l3vpn-virtual-ce] option below.

Consider the following scenarios when BGP MPLS VPN technology is considered as whole or partial deployment:

Scenario 1: All VPN sites (CEs/VMs) support IP connectivity. The most suited BGP solution is to use IP VPNs [RFC4364] for all sites with PE and/or vPE solutions. This is a straightforward case.

Scenario 2: Legacy layer-2 connectivity must be supported in certain sites/CEs/VMs, and the rest of the sites/CEs/VMs need only 3 connectivity.

One can consider using a combined vPE and vCE [I-D.fang-l3vpn-virtual-ce] solution to solved the problem. Use of IP VPN for all sites with IP connectivity, and a physical or virtual CE (vCE, may reside on the end device) to aggregate the Layer-2 sites which for example, are in a single container in a Data Center. The CE/vCE can be considered as inter-connecting points, where the Layer-2 network is terminated and the corresponding routes for connectivity of the L2 network are inserted into L3VPN VRFs. The Layer-2 aspect is transparent to the L3VPN in this case.

Reducing operation complicity and maintaining the robustness of the solution are the primary reasons for the recommendations.

7. Management, Control, and Orchestration

7.1 Assumptions

The discussion in this section is based on the following set of assumptions:

- The WAN and the inter-connecting Data Center, MAY be under control of separate administrative domains
- WAN Gateways/ASBRs/PEs are provisioned by existing WAN provisioning systems
- If a single Gateway/ASBR/PE connecting to the WAN on one side, and connecting to the Data Center network on the other side, then this Gateway/ASBR/PE is the demarcation point between the two networks.
- vPEs and VMS are provisioned by Data Center Orchestration systems.
- Managing IP VPNs in the WAN is not within the scope of this document except the inter-connection points.

7.2 Management/Orchestration system interfaces

The Management/Orchestration system CAN be used to communicate with both the Data Center Gateway/ASBR, and the end devices.

The Management/Orchestration system MUST support standard, programmatic interface for full-duplex, streaming state transfer in and out of the routing system at the Gateway.

The programmatic interface is currently under definition in IETF Interface to Routing Systems (I2RS)) initiative. [I-D.atlas-i2rs-architecture], and [I-D.rfernando-irs-fw-req].

Standard data modeling languages will be defined/identified in I2RS. YANG - A Data Modeling Language for the Network Configuration Protocol (NETCONF) [RFC6020] is a promising candidate currently under investigation.

To support remote access between applications running on an end device (e.g., a server) and routers in the network (e.g. the DC Gateway), a standard mechanism is expected to be identified and defined in I2RS to provide the transfer syntax, as defined by a protocol, for communication between the application and the network/routing systems. The protocol(s) SHOULD be lightweight and familiar by the computing communities. Candidate examples include ReSTful web services, JSON [RFC4627], NETCONF [RFC6241], XMPP

[RFC6120], and XML. [I-D.atlas-i2rs-architecture].

7.3 Service VM Management

Service VM Management SHOULD be hypervisor agnostic, e.g. On demand service VMs turning-up SHOULD be supported.

7.4 Orchestration and IP VPN inter-provisioning

The orchestration system

- 1) MUST support IP VPN service activation in virtualized Data Center.
- 2) MUST support automated cross-provisioning accounting correlation between the WAN IP VPN and Data Center for the same tenant.
- 3) MUST support automated cross provisioning state correlation between WAN IP VPN and Data Center for the same tenant

There are two primary approaches for IP VPN provisioning - push and pull, both CAN be used for provisioning/orchestration.

7.4.1 vPE Push model

Push model: push IP VPN provisioning from management/orchestration systems to the IP VPN network elements.

This approach supports service activation and it is commonly used in existing IP VPN Enterprise deployments. When extending existing WAN IP VPN solutions into the a Data Center, it MUST support off-line accounting correlation between the WAN IP VPN and the cloud/DC IP VPN for the tenant. The systems SHOULD be able to bind interface accounting to particular tenant. It MAY requires offline state correlation as well, for example, binding of interface state to tenant.

Provisioning the vPE solution:

- 1) Provisioning process
 - a. The WAN provisioning system periodically provides to the DC orchestration system the VPN tenant and RT context.
 - b. DC orchestration system configures vPE on a per request basis
- 2) Auto state correlation
- 3) Inter-connection options:

Inter-AS options defined in [RFC4364] may or may not be sufficient for a given inter-connection scenario. BGP IP VPN inter-connection with the Data Center is discussed in [I-D.fang-l3vpn-data-center-interconnect].

This model requires offline accounting correlation

- 1) Cloud/DC orchestration configures vPE
- 2) Orchestration initiates WAN IP VPN provisioning; passes connection IDs (e.g., of VLAN/VXLAN) and tenant context to WAN IP VPN provisioning systems.
- 3) WAN IP VPN provisioning system provisions PE VRF and policies as in typical Enterprise IP VPN provisioning processes.
- 4) Cloud/DC Orchestration system or WAN IP VPN provisioning system MUST have the knowledge of the connection topology between the DC and WAN, including the particular interfaces on core router and connecting interfaces on the DC PE and/or vPE.

In short, this approach requires off-line accounting correlation and state correlation, and requires per WAN Service Provider integration.

Dynamic BGP sessions between PE/vPE and vCE MAY be used to automate the PE provisioning in the PE-vCE model, that will remove the needs for PE configuration. Caution: This is only under the assumption that the DC provisioning system is trusted and can support dynamic establishment of PE-vCE BGP neighbor relationships, for example, the WAN network and the cloud/DC belong to the same Service Provider.

7.4.2 vPE Pull model

Pull model: pull from network elements to network management/AAA based upon data plane or control plane activity. It supports service activation. This approach is often used in broadband deployments. Dynamic accounting correlation and dynamic state correlation are supported. For example, session based accounting implicitly includes tenant context state correlation, as well as session-based state that implicitly includes tenant context. Note that the pull model is less common for vPE deployment solutions.

Provisioning process:

- 1) Cloud/DC orchestration configures vPE

2) Orchestration primes WAN IP VPN provisioning/AAA for new service, passes connection IDs (e.g., VLAN/VXLAN) and tenant context.

3) Cloud/DC ASBR detects new VLAN and sends Radius Access-Request (or Diameter Base Protocol request message [RFC6733]).

4) Radius Access-Accept (or Diameter Answer) with VRF and other policies

Auto accounting correlation and auto state correlation is supported.

7. Security Considerations

vPE solution presented a virtualized IP VPN PE model. There are potential implications to IP VPN control plane, forwarding plane, and management plane. Security considerations are currently under study, will be included in the future revisions.

8. IANA Considerations

None.

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.
- [RFC3032] Rosen, E., Tappan, D., Fedorkow, G., Rekhter, Y., Farinacci, D., Li, T., and A. Conta, "MPLS Label Stack Encoding", RFC 3032, January 2001.
- [RFC4023] Worster, T., Rekhter, Y., and E. Rosen, Ed., "Encapsulating MPLS in IP or Generic Routing Encapsulation (GRE)", RFC 4023, March 2005.
- [RFC4271] Rekhter, Y., Ed., Li, T., Ed., and S. Hares, Ed., "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.

- [RFC4684] Marques, P., Bonica, R., Fang, L., Martini, L., Raszuk, R., Patel, K., and J. Guichard, "Constrained Route Distribution for Border Gateway Protocol/MultiProtocol Label Switching (BGP/MPLS) Internet Protocol (IP) Virtual Private Networks (VPNs)", RFC 4684, November 2006.
- [RFC5036] Andersson, L., Ed., Minei, I., Ed., and B. Thomas, Ed., "LDP Specification", RFC 5036, October 2007.
- [RFC6020] Bjorklund, M., Ed., "YANG - A Data Modeling Language for the Network Configuration Protocol (NETCONF)", RFC 6020, October 2010.
- [RFC6120] Saint-Andre, P., "Extensible Messaging and Presence Protocol (XMPP): Core", RFC 6120, March 2011.
- [RFC6241] Enns, R., Ed., Bjorklund, M., Ed., Schoenwaelder, J., Ed., and A. Bierman, Ed., "Network Configuration Protocol (NETCONF)", RFC 6241, June 2011.
- [RFC6733] Fajardo, V., Ed., Arkko, J., Loughney, J., and G. Zorn, Ed., "Diameter Base Protocol", RFC 6733, October 2012.
- [I-D.ietf-l3vpn-end-system] Marques, P., Fang, L., Pan, P., Shukla, A., Napierala, M., Bitar, N., "BGP-signaled end-system IP/VPNs", draft-ietf-l3vpn-end-system-01, April, 2013.

9.2 Informative References

- [RFC4627] Crockford, D., "The application/json Media Type for JavaScript Object Notation (JSON)", RFC 4627, July 2006.
- [RFC4797] Rekhter, Y., Bonica, R., and E. Rosen, "Use of Provider Edge to Provider Edge (PE-PE) Generic Routing Encapsulation (GRE) or IP in BGP/MPLS IP Virtual Private Networks", RFC 4797, January 2007.
- [I-D.fang-l3vpn-end-system-req] Napierala, M., and Fang, L., "Requirements for Extending BGP/MPLS VPNs to End-Systems", draft-fang-l3vpn-end-system-requirements-01, Oct. 2012.
- [I-D.atlas-i2rs-architecture] Atlas A., Halpern, J., Hares, S., Ward, D., Nadeau, T., draft-atlas-i2rs-architecture-01, July 2013.

- [I-D.rfernando-irs-fw-req] Fernando, R., Medved, J., Ward, D., Atlas, A., Rijsman, B., "IRS Framework Requirements", draft-rfernando-irs-framework-requirement-00, Oct. 2012.
- [I-D.rfernando-l3vpn-service-chaining] Fernando, R., Rao, D., Fang, L., Napierala, M., So, N., draft-rfernando-l3vpn-service-chaining-02, July 15, 2013.
- [I-D.bitar-i2rs-service-chaining] Bitar, N., Geron, G., Fang, L., Krishnan, R., Leymann, N., Shah, H., Chakrabarti, S., Haddad, W., draft-bitar-i2rs-service-chaining-00, July 2013.
- [I-D.fang-l3vpn-virtual-ce] Fang, L., Evans, J., Ward, D., Fernando, R., Mullooly, J., So, N., Bitar, N., Napierala, M., "BGP IP VPN Virtual PE", draft-fang-l3vpn-virtual-ce-01, Feb. 2013.
- [I-D.fang-l3vpn-data-center-interconnect] Fang, L., Fernando, R., Rao, D., Boutros, S., BGP IP VPN Data Center Interconnect, draft-fang-l3vpn-data-center-interconnect-01, Feb. 2013.
- [I-D.mahalingam-dutt-dcops-vxlan]: Mahalingam, M, Dutt, D., et al., "A Framework for Overlaying Virtualized Layer 2 Networks over Layer 3 Networks" draft-mahalingam-dutt-dcops-vxlan-04, May 2013.
- [I-D.sridharan-virtualization-nvgre]: SridharanNetwork, M., et al., "Virtualization using Generic Routing Encapsulation", draft-sridharan-virtualization-nvgre-02.txt, July 2012.

Authors' Addresses

Luyuan Fang
Cisco
111 Wood Ave. South
Iselin, NJ 08830
Email: lufang@cisco.com

David Ward
Cisco
170 W Tasman Dr
San Jose, CA 95134
Email: wardd@cisco.com

Rex Fernando

Cisco
170 W Tasman Dr
San Jose, CA
Email: rex@cisco.com

Maria Napierala
AT&T
200 Laurel Avenue
Middletown, NJ 07748
Email: mnapierala@att.com

Nabil Bitar
Verizon
40 Sylvan Road
Waltham, MA 02145
Email: nabil.bitar@verizon.com

Dhananjaya Rao
Cisco
170 W Tasman Dr
San Jose, CA
Email: dhrao@cisco.com

Bruno Rijsman
Juniper Networks
10 Technology Park Drive
Westford, MA 01886
Email: brijsman@juniper.net

Ning So
Tata Communications
Plano, TX 75082, USA
Email: ning.so@tatacommunications.com

Jim Guichard
Cisco
Boxborough, MA 01719
Email: jguichar@cisco.com

Wen Wang
CenturyLink
2355 Dulles Corner Blvd.
Herndon, VA 20171
Email: Wen.Wang@CenturyLink.com

Manuel Paul
Deutsche Telekom
Winterfeldtstr. 21-27

10781 Berlin, Germany
Email: manuel.paul@telekom.de

L3VPN
Internet-Draft
Intended status: Standards Track
Expires: January 14, 2014

R. Kebler
P. Kurapati
Juniper Networks
July 13, 2013

Multicast Traceroute for MVPNs
draft-kebler-kurapati-l3vpn-mvpn-mtrace-00

Abstract

Mtrace is a tool used to troubleshoot issues in a network deploying Multicast service. When multicast is used within a VPN service offering, the base Mtrace specification does not detect the failures. This document specifies a method of using multicast traceroute in a network offering Multicast in VPN service.

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 14, 2014.

Copyright Notice

Copyright (c) 2013 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.

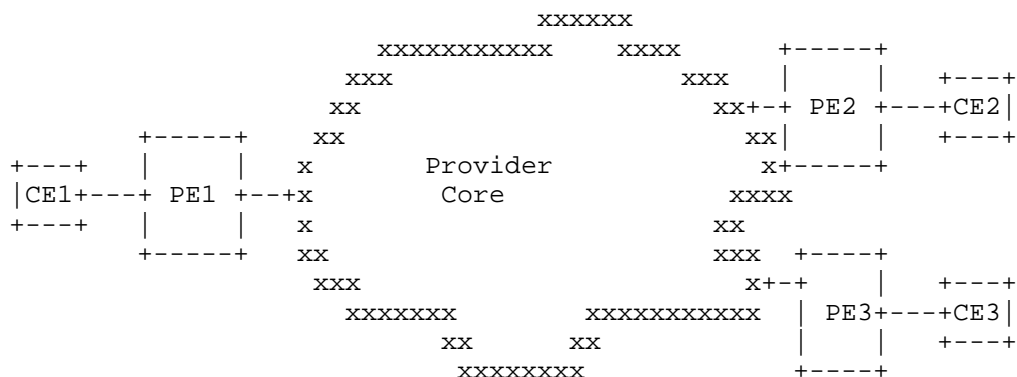
Table of Contents

1. Introduction	2
2. Overview	3
3. Protocol Details	4
3.1. Mtrace Query	4
3.2. Mtrace Request	4
3.2.1. Ingress PE Procedures	6
3.3. Downstream Requests	7
3.4. ASBR Behavior	7
3.5. Virtual Hub and Spoke	8
3.6. Inter-Area Provider Tunnels	8
3.6.1. Egress PE	8
3.6.2. ABR Behavior	9
3.7. Mtrace MVPN Procedure	9
4. Error Detection	10
4.1. MVPN Error Codes	11
5. Mtracev2 Extensions	12
5.1. New Mtracev2 TLV Type	12
5.2. MVPN Extended Query Block	12
5.3. Leaf A-D Augmented Response Block	13
5.4. PMSI Tunnel Attributes Augmented Response Block	13
6. Mtrace2 Standard Response Block considerations	13
7. IANA Considerations	14
8. Security Considerations	14
9. Acknowledgments	14
10. Normative References	14
Authors' Addresses	15

1. Introduction

The current multicast traceroute [I-D.ietf-mboned-mtrace-v2] travels up the tree hop-by-hop towards the source. This verifies the basic multicast state back to the source, but is not sufficient to verify the MVPN state. The base Mtrace specification assumes that the routers in the path are directly connected through interfaces. In the case of Multicast traffic over VPN service, the PEs who are MVPN neighbors may be separated by several router hops. The path taken by the query can be completely different from the path taken through core by the actual multicast traffic. Consider a case in the below figure, where provider tunnel between PE2 (Source) and PE1 (Receiver) is not established correctly due to incorrect MVPN state on PE2. In the current form of Mtrace, the Query would result in a successful response since there is no error detection mechanism for MVPN state available currently. Even if one can infer from the statistics of the Mtrace Response that PE2 has an issue, the existing error codes are not sufficient to identify the root cause. Also, there could be a problem sending traffic over the provider tunnel from PE2 to PE1,

but the mtrace query will not even travel over this provider tunnel. Therefore, the mtrace successful response can be misleading. This draft ensures that the Response uses same provider-tunnel that the given C-S,C-G data would traverse and returns appropriate MVPN specific error codes which would help in identifying the root cause.



MVPN topology

2. Overview

As described in the Mtracev2 specification [I-D.ietf-mboned-mtrace-v2], a Querier initiates an Mtrace Query which is sent to the Last Hop Router. Last Hop Router converts this into a Request and sends it towards the First Hop Router. This draft introduces a new "Downstream Request" mechanism to allow the First Hop Router to send the mtrace request message back on the Provider tunnel to the Last hop router. The last hop router will then change it to Response and send it to the Querier who initiated the Query. If there is any error encountered by the Last hop router or First Hop router, a Response is directly unicasted to the querier with appropriate MVPN specific error codes added. Each hop in the path of Mtrace decrements the TTL value before sending the mtrace message.

Since the Mtrace is being extended for MVPNs, the Last Hop router and First Hop router SHOULD be a Provider Edge (PE) router so that the MVPN specific error codes can be contained within the provider space. The Request will be initiated by the egress PE and will travel upstream to the ingress PE. It is assumed that the Querier knows and can reach the egress PE. A Querier and egress PE can be the same router.

For Mtrace initiated by the CEs, the specification mentioned in Mtracev2 [I-D.ietf-mboned-mtrace-v2] SHOULD be followed. If a Mtrace message is received by the PE on CE facing interfaces containing MVPN specific extensions defined in this draft, it SHOULD be discarded.

3. Protocol Details

The protocol details that follow are described in terms of mtracev2. However, the same procedures can be achieved with mtracev1. The protocol extensions needed for mtracev2 are described in Section 5 and the protocol extensions for mtracev1 and described in section 6.

3.1. Mtrace Query

A Querier willing to perform a Mtrace on a MVPN issues a Mtrace Query. The format of the Query TLV is as specified in the Mtracev2 specification [I-D.ietf-mboned-mtrace-v2]. The (C-S,C-G) to be queried is populated in the source address and group address fields of the Mtrace2 Query block. A deployment may use wild card SPMSIs as defined in [RFC6625]. For example, a (C-*,C-*) wild card SPMSI or a (C-*,ALL-PIM-ROUTERS) can be used to send messages like BSR across PEs as mentioned in section 5.3.4 MVPN specification [RFC6513]. A querier may be interested in knowing the health of such a SPMSI tunnel. In this case, the Multicast Address and Source Address fields of the Mtrace2 query can be filled with wild cards (all 1s) accordingly by the querier.

The Querier MUST add a MVPN Extended Query Block to include the RD of the C-S,C-G that it wishes to trace. When wild card SPMSIs are used, a PE could have subscribed to multiple upstream PEs for wild card SPMSIs. Hence, a query for a wild card SPMSI MUST also specify the upstream PE address that it is interested to query. The upstream PE address in the MVPN Extended Query Block MUST be filled only for wild card queries. For a regular (C-S,C-G) query, this field SHOULD be set to 0s by the querier and is ignored by the receivers.

This Query is sent to the Downstream PE (Last Hop Router) to initiate the mtrace towards the source. If a Querier does not receive a Response, it can retry sending Query messages with increasing TTL values to help diagnose where the Mtrace messages are being lost.

3.2. Mtrace Request

The PE that receives the query will lookup the (C-S,C-G) using the RD of the query to distinguish the vrf. If the RD doesn't match any VRF, PE sends a response with error code set to BAD_RD. The PE first checks the C-Mcast route that is matching (C-S,C-G) of the mtrace Query. It then finds the upstream multicast hop from the selected

C-Mcast route and unicasts the requests to the upstream multicast hop after decrementing the TTL. The Mtrace Request MUST have PMSI Tunnel Attributes Augmented Response Block populated with the PMSI attribute that the PE uses to receive the traffic for the given (C-S,C-G) traffic.

Upstream multicast hop can be same as upstream PE router in some cases, while it can be the ASBR or the BGP nexthop of the selected C-Mcast route in Inter-AS scenarios. The procedures for finding upstream multicast hop is discussed in detail under section 5.1 of MVPN specification [RFC6513].

When a wild card query is received, the PE will look for the upstream PE address in the MVPN Extended query block. The PE will then check if it has bound to the wild card SPMSI tunnel from the specified upstream PE. If it has, it will populate the Leaf A-D Augmented Response Block and PMSI Tunnel Attributes Augmented Response Block with the respective values. If the PE has not received any wild card SPMSI AD route from the specified upstream PE in the query, it should send a response with the error code set to NO_WILD_CARD_SPMSI_AD_RCVD. If the PE has received wild card SPMSI AD route from the upstream PE, but has not responded with a LEAF-AD route, it should send a response with the error code set to NO_WILD_CARD_SPMSI_LEAF_AD_SENT.

For a non-wild-card query, the upstream PE address field in the MVPN Extended query block MUST be ignored by the PEs. It MUST follow the procedure to find the upstream multicast hop as discussed earlier.

If the route does not match any MVPN-TIB state, then the PE should send a Response to the Querier with the error code set to NO_CMCAST_STATE. If the PE cannot locate the upstream PE then it should send a response to the Querier with the NO_UPSTREAM_PE error code.

From the selected UMH route, the local PE extracts the ASN of the upstream PE (as carried in the Source AS Extended Community of the route), and the source-AS field of the mtrace Query is set to that AS.

If the local and the upstream PEs are in the same AS, then the RD in the mtrace Query is set to the RD of the VPN-IP route for the source/ RP.

Section 8 of MVPN specification [RFC6513] mentions two procedures (Segmented and Non-Segmented) for handling Inter-AS scenarios. If the local and the upstream PEs are in different ASes, and if segmented Inter-AS procedure is used, then the local PE finds in its

VRF an Inter-AS I-PMSI A-D route whose Source AS field carries the ASN of the upstream PE. The RD of the found Inter-AS I-PMSI A-D route is used as the RD of the mtrace Query. If Inter-AS I-PMSI A-D route is not found, a response with error code UNKNOWN_INTER_AS is sent.

To support non-segmented inter-AS tunnels, if the local and the upstream PEs are in different ASes, the local system finds in its VRF an Intra-AS I-PMSI A-D route from the upstream PE. The Originating Router's IP Address field of that route has the same value as the one carried in the VRF Route Import of the unicast route to the address carried in the Multicast Source field. The RD of the found Intra-AS I-PMSI A-D route is used as the RD in the mtrace Query. The Source AS field in the mtrace Query is set to value of the Originating Router's IP Address field of the found Intra-AS I-PMSI A-D route.

The PE receiving Mtrace Query will check for any errors. If any error is detected it will send the error back to the Querier. Otherwise, it will change the TLV value to be an Mtrace Request, and it will add a Mtrace2 Standard Response Block. It will also add a PMSI Tunnel Attributes Augmented Response Block with the attributes of the PMSI used to receive traffic for the S,G. If a Leaf-AD route was advertised to the upstream PE for this S,G then the PE will also include a Leaf-AD Augmented Response Block with the NLRI of the associated Leaf-AD route.

3.2.1. Ingress PE Procedures

The PE that receives the Request, will check the PMSI attributes of the sender of request to see if they match the values used to send traffic for the S,G. If the values do not match, then the PE uses the appropriate pmsi error code as specified in 'MVPN Error Codes' section and sends a mtrace Response back to the Querier. Also, if a Leaf A-D Augmented Response Block is included, the PE will validate that it has received this Leaf A-D route from the router that sent the Request. If not, then this PE should change the error code to BAD_LEAF_AD and send the Response to the Querier. If the PE expects that a Leaf A-D route is needed for the downstream PE to receive traffic, but did not receive one in the mtrace Request from the sending router, then it should use a NO_LEAF_AD_RCVD error code for the mtrace Response. For a wild card SPMSI query, if the PE didn't receive LEAF AD route from the downstream PE, it should use NO_LEAF_AD_RCVD error code.

When the upstream PE receives the Request, it will check for any errors. If there are errors detected, or if the TTL expired, then the PE will change the TLV code to be a Mtrace Response and unicast the response back to the Querier.

The ingress PE will also check it has local vrf connectivity for the source/RP. If it does not have any connectivity to the source/RP then it should use the base specification error code NO_ROUTE and send an mtrace Response. Note that in a Virtual Hub and Spoke environment, it is possible for a PE to receive a mtrace Request and need to propagate it to another upstream PE. These procedures are outlined in the section "Virtual Hub and Spoke". If the PE does not expect to be receiving mtrace Responses from the mvpn core and have the route to the source located via another upstream PE, then it can use the base specification RPF_IF error code.

If the PE that receives the Request is the ingress PE that has local vrf connectivity for the source, then it will add a Standard Response Block to the mtrace message. It will not include the additional PMSI Attributes Response Block. Then it will turn the Request into a Downstream Request by changing the value of the Type field of the TLV. It will send the mtrace message on the provider tunnel used to send the S,G data traffic.

3.3. Downstream Requests

When a router receives Mtrace Downstream Request, it will determine if it has added any of the Response Blocks for this mtrace message. If it does not locate its address in the list of Response Blocks, then it will silently discard this mtrace message. Otherwise, it will set the 'D' bit in its PMSI Tunnel Attributes Augmented Response Block to indicate that this message has been received on the PMSI tunnel.

If this router is the egress PE that provided the initial Response Block, then it will change the mtrace type to a Reply and sends the Reply to the Querier (the egress PE and the Querier may be the same router). Otherwise, this router must send the Downstream PE on the PMSI that it would normally send traffic for the S,G. Before sending the Downstream Request, the router must decrement the TTL and check for TTL expiry. If the TTL has expired, then this router must send the Response to the Querier with the appropriate code.

3.4. ASBR Behavior

When an ASBR receives a mtrace Request the ASBR finds an Inter-AS I-PMSI A-D route whose RD and Source AS matches the RD and Source AS carried in the mtrace Query. If no matching route is found and the ASBR is using segmented tunnels as described in MVPN specification [RFC6513], the ASBR sends an UNKNOWN_INTER_AS error code back to the Querier. If a matching route is found, the ASBR acts as a "first hop router" and modifies the Query type to DOWNSTREAM_REQUEST. ASBR in this case MUST validate the PMSI attributes similar to the "first hop

router" and respond if there is any errors. ASBR MUST populate PMSI Tunnel Attributes Augmented Response Block with the Inter-AS provider tunnel information before sending the DOWNSTREAM_REQUEST. Note that the mtrace request does not proceed upstream as it is assumed that performing a traceroute and exposing IP addresses across AS boundaries would not be desirable with Segmented Inter-AS Provider Tunnels.

To support non-segmented inter-AS tunnels as described in [RFC6513], instead of matching the RD and Source AS carried in the mtrace Query against the RD and Source AS of an Inter-AS I-PMSI A-D route, the ASBR should match it against the RD and the Originating Router's IP Address of the Intra- AS I-PMSI A-D routes. The Next Hop field of the MP_REACH_NLRI of the found Intra-AS I-PMSI A-D route is used as the destination for the mtrace Request.

3.5. Virtual Hub and Spoke

When a Virtual-Hub (V-HUB) as described in specification [I-D.ietf-l3vpn-virtual-hub] receives a mtrace Request the S,G may be reachable via one of its vrf interfaces. In this case, the V-HUB is an ingress PE and the procedure are defined in the Section "Ingress PE Procedures". Otherwise, the C-RP/C-S of the route is reachable via some other PE. This is the case where the received route was originated by a Virtual-spoke (V-spoke) that sees the V-HUB as the "upstream PE" for the given source, but the V-HUB sees another PE as the "upstream PE" for that source. In this case, the V-HUB should check the PMSI attributes sent in the mtrace Request against the Tunnel Attributes of the Provider Tunnel used to send traffic for the S,G from the upstream PE to the V-Spoke.

The V-HUB sends a mtrace Request to its upstream PE the same way as it would if it received a mtrace Query. V-HUB MUST add PMSI Tunnel Attributes Augmented Response Block of its own before sending the mtrace Request to the upstream PE. It may also add Leaf-AD Augmented Response Block if a Leaf-AD route was advertised upstream by the V-HUB. If the RD or Source-AS of the upstream PE is different, the V-HUB updates the MVPN Extended Query Block accordingly.

3.6. Inter-Area Provider Tunnels

3.6.1. Egress PE

The egress PE does the same procedures as specified in Section "Mtrace Request" except it sends the Request upstream to the IP address determined from the Global Administrator field of the Inter-area P2MP Segmented Next-hop Extended Community as described in specification [I-D.ietf-mpls-seamless-mcast] . If the egress PE has

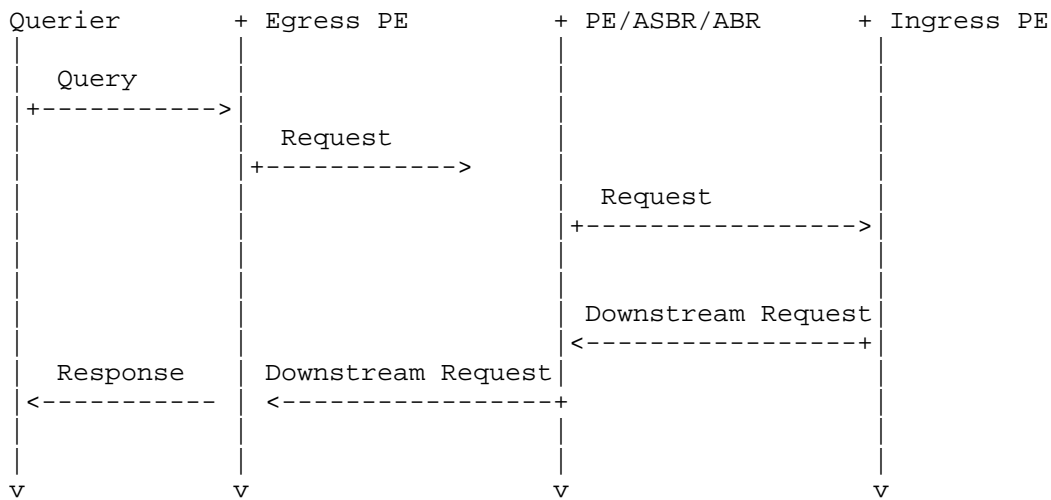
sent a Leaf-AD route then it must send a Leaf-AD Augmented Response Block with the NLRI of the Leaf A-D route.

3.6.2. ABR Behavior

ABR MUST find a S-PMSI or I-PMSI route whose NLRI has the same value as the Route Key field of the received mtrace Leaf-AD extended Query Block. If such a matching route is not found then a Response should be sent to the Querier with the NO_LEAF_AD_RCVD. If the ABR has sent a Leaf-AD route then it must add a Leaf-AD Augmented Response Block with the values of Leaf A-D route NLRI. The upstream node's IP address is the IP address determined from the Global Administrator field of the Inter-area P2MP Segmented Next-hop Extended Community.

3.7. Mtrace MVPN Procedure

In this section, we will briefly discuss the Mtrace procedure taking a working and non-working network topology.



Mtrace MVPN Procedure

The above figure depicts the path of MTRACE in working condition. MTRACE request for MVPN can traverse multiple hops when a Virtual HUB is present or when segmented P2MP inter-area tunnels are used. If no error conditions are detected the downstream request will travel the same path as the regular multicast packet for the queried mroute would flow. The last hop router/egress router will convert it into a Response and send it back to querier

Let us consider a non-working case where Mtrace is expected to be used. Taking Virtual-HUB as an example, assume that there is a data-path issue between V-HUB and Egress Spoke. The below steps take place to determine the issue between V-HUB and egress Spoke

- 1 - Querier sends the Mtrace Query towards LHR (Egress PE-Spoke).
- 2 - Egress PE sends Request to V-HUB. V-HUB realises that the first hop router is a connected spoke and sends the request to Ingress Spoke PE.
- 3 - Ingress Spoke PE sends Downstream Request to V-HUB. The same is received by V-HUB. V-HUB sets the 'D' bit in its PMSI Tunnel Attributes Augmented Response Block.
- 4 - V-HUB sends Downstream request to ingress spoke. This is never received by the ingress spoke.
- 5 - The result of first 4 steps is that querier did not receive the response. This makes the querier fall back to TTL method.
- 6 - Querier reduces the TTL and the result will show that the hop from V-HUB to ingress spoke is missing thereby pointing the issue at the right place.

4. Error Detection

All routers will check for normal multicast errors as defined in the Mtracev2 specification. In addition, they will check for errors specific to MVPNs and this specification.

All receiving routers will check the state of the Provider Tunnel used for forwarding traffic for the given S,G. The ability and manner to check if the Provider Tunnel is down depends on the Provider Tunnel type. If the Provider Tunnel is known to be down the PE will respond with a PTUNNEL_DOWN error.

In some situations the router needs to send a Leaf AD route to the upstream PE. If the upstream expects a Leaf AD route, but did not receive one from the downstream PE, then the NO_LEAF_AD_RCVD error will be sent.

The receiving router will check the values of the PMSI Tunnel attributes to see if they match the expected values for the PMSI. If an Inclusive-PMSI is used, then the router will verify that the values match those in the I-PMSI A-D route. If a Selective PMSI is used, then the Tunnel Attributes will be matched against the S-PMSI or Leaf A-D Route, depending on the Tunnel Type. If the values do not match, then a error code of the corresponding PMSI mismatch will be sent.

If a router receives a MVPN traceroute, but does not have the proper MVPN configuration, then it will respond with a UNEXPECTED_MVPN error

4.1. MVPN Error Codes

Value	Name	Description
-----	-----	-----
0x11	PTUNNEL_DOWN	The provide tunnel for this S,G is down.
0x12	NO_LEAF_AD_RCVD	The S-PMSI has not been joined by downstream neighbor
0x13	BAD_LEAF_AD	The Leaf A-D route does not match the expected values
0x14	BAD_RD	The RD is known to not exist on this PE
0x15	UNEXPECTED_MVPN	The MVPN traceroute message is unexpected
0x16	BAD_PMSI_ATTR_FLAG	Error matching the PMSI attribute flag
0x17	BAD_PMSI_ATTR_TYPE	Error matching the PMSI attribute type
0x18	BAD_PMSI_ATTR_LABEL	Error matching the PMSI attribute label
0x19	BAD_PMSI_ATTR_ID	Error matching the PMSI attribute tunnel identifier
0x1a	UNKNOWN_INTER_AS	Could not locate the Inter-AS provider tunnel segment.
0x1b	NO_UPSTREAM_PE	No valid upstream PE or route

0x1c NO_CMCAST_STATE No C-Mcast route for the requested query

0x1d NO_WILD_CARD_SPMSI_AD_RCVD No Wild Card SPMSI AD is received from the upstream PE

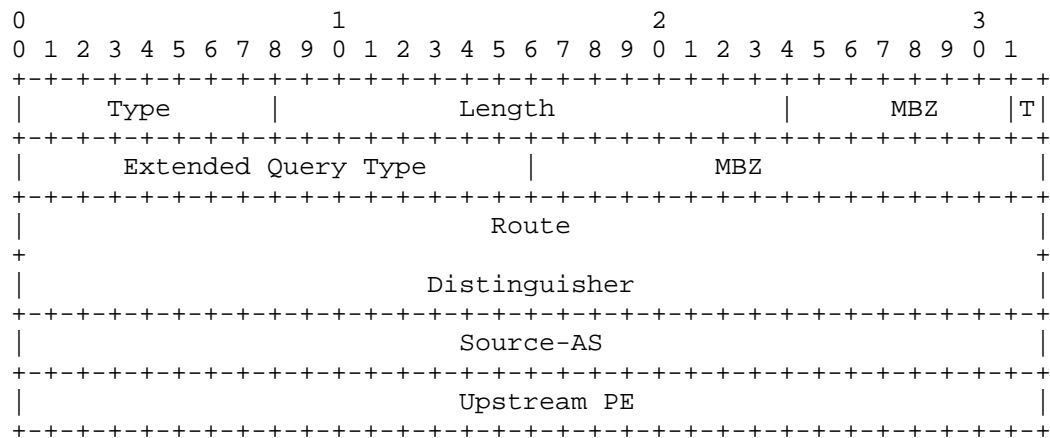
0x1e NO_WILD_CARD_SPMSI_LEAD_AD_SENT PE did not send LEAF-AD route for the wild card SPMSI

5. Mtracev2 Extensions

5.1. New Mtracev2 TLV Type

A new Mtracev2 TLV type will be created for the Mtrace2 Downstream Request.

5.2. MVPN Extended Query Block



MVPN Extended Query Block

Type: Mtrace2 Extended Query Block Type

Length: Length of the MVPN Extended Query Block

MBZ: Sent with all 0's, ignored on receipt

T bit: This bit should be 0

Extended Query Type: New type defined

MBZ: Sent with all 0's, ignored on receipt

Route Distinguisher: The RD of the S,G that should be traced

Source-AS: The Autonomous System Number (ASN) of the Source

Upstream PE: IP Address of the Upstream PE

5.3. Leaf A-D Augmented Response Block

```

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
+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+
|                                     Type                                     | Value .... |
+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+

```

Leaf A-D Augmented Response Block

MBZ: Sent with all 0's, ignored on receipt

Type: New type defined

Value: The NLRI value of the associated Leaf A-D route

5.4. PMSI Tunnel Attributes Augmented Response Block

```

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
+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+
|                                     Type                                     |D|    MBZ    | Value.. |
+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+

```

PMSI Tunnel Attributes Augmented Response Block

MBZ: Sent with all 0's, ignored on receipt

Type: New type defined

D: 'D' bit indicating that Downstream Request is received on PMSI

Value: The PMSI Tunnel Attribute as defined in RFC 6514

6. Mtrace2 Standard Response Block considerations

The PEs in the MVPN Mtrace add the Standard Response Block as defined in Mtrace2 [I-D.ietf-mboned-mtrace-v2]. For a PE, the incoming or outgoing interface can be a Tunnel. The First Hop Router (FHR) PE which is connected to the source SHOULD populate the incoming interface address with the respective interface connected to the CE. The outgoing interface address MAY be populated with 0 in this case. Other routers in the mtrace path MAY populate incoming and outgoing interface address fields as 0. 'Multicast Rtg Protocol' field MUST be populated with 0s by the Last Hop Router (LHR). First Hop Router (FHR) can populate this field with respective multicast routing protocol used towards its upstream CE. All the remaining fields of the Standard Response Block are populated as defined by the Mtrace2 [I-D.ietf-mboned-mtrace-v2] specification.

7. IANA Considerations

New TLV Type for MTRACE_MVPN_QUERY, MTRACE_MVPN_REQUEST, MTRACE_MVPN_DOWNSTREAM_REQUEST, MTRACE_MVPN_RESPONSE

8. Security Considerations

There are no security considerations for this design other than what is already in the mtracev2 specification.

9. Acknowledgments

The authors would like to thank Yakov Rekhter and Marco Rodrigues for their valuable review and feedback.

10. Normative References

[I-D.ietf-l3vpn-virtual-hub]

Jeng, H., Uttaro, J., Jalil, L., Decraene, B., Rekhter, Y., and R. Aggarwal, "Virtual Hub-and-Spoke in BGP/MPLS VPNs", draft-ietf-l3vpn-virtual-hub-08 (work in progress), July 2013.

[I-D.ietf-mboned-mtrace-v2]

Asaeda, H. and W. Lee, "Mtrace Version 2: Traceroute Facility for IP Multicast", draft-ietf-mboned-mtrace-v2-09 (work in progress), October 2012.

[I-D.ietf-mpls-seamless-mcast]

Rekhter, Y. and R. Aggarwal, "Inter-Area P2MP Segmented LSPs", draft-ietf-mpls-seamless-mcast-07 (work in progress), May 2013.

- [RFC6513] Rosen, E. and R. Aggarwal, "Multicast in MPLS/BGP IP VPNs", RFC 6513, February 2012.
- [RFC6514] Aggarwal, R., Rosen, E., Morin, T., and Y. Rekhter, "BGP Encodings and Procedures for Multicast in MPLS/BGP IP VPNs", RFC 6514, February 2012.
- [RFC6625] Rosen, E., Rekhter, Y., Hendrickx, W., and R. Qiu, "Wildcards in Multicast VPN Auto-Discovery Routes", RFC 6625, May 2012.

Authors' Addresses

Robert Kebler
Juniper Networks
10 Technology Park Drive
Westford, MA 01886
USA

Email: rkebler@juniper.net

Pavan Kurapati
Juniper Networks
1194 N. Mathilda Ave
Sunnyvale, CA 94089
USA

Email: kurapati@juniper.net

L3VPN Working Group
Internet-Draft
Intended status: Standards Track
Expires: January 10, 2014

K. Patel
Cisco Systems
Y. Rekhter
Juniper Networks
E. Rosen
Cisco Systems
July 09, 2013

BGP as an MVPN PE-CE Protocol
draft-keyupate-l3vpn-mvpn-pe-ce-00

Abstract

When a Service Provider offers BGP/MPLS IP VPN service to its customers, RFCs 6513 and 6514 describe protocols and procedures that the Service Provider can use in order to carry the customer's IP multicast traffic from one customer site to others. BGP can be used to carry customer multicast routing information from one Provider Edge (PE) router to another, but it is assumed that PIM is running on the interface between a Customer Edge (CE) router and a PE router. This document specifies protocols and procedures that, under certain conditions, allow customer multicast routing information to be carried between PE and CE via BGP. This can eliminate the need to run PIM on the PE-CE interfaces, potentially eliminating the need to run PIM on the PE routers at all.

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 10, 2014.

Copyright Notice

Copyright (c) 2013 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.

This document may contain material from IETF Documents or IETF Contributions published or made publicly available before November 10, 2008. The person(s) controlling the copyright in some of this material may not have granted the IETF Trust the right to allow modifications of such material outside the IETF Standards Process. Without obtaining an adequate license from the person(s) controlling the copyright in such materials, this document may not be modified outside the IETF Standards Process, and derivative works of it may not be created outside the IETF Standards Process, except to format it for publication as an RFC or to translate it into languages other than English.

Table of Contents

1. Introduction	2
2. C-MCAST SAFI NLRI Format	4
2.1. C-MCAST Join and Prune Routes	5
3. Exchanging C-MCAST Join Routes	7
3.1. Originating C-MCAST Join Route at the CE router	7
3.2. Receiving a C-MCAST Join Route by the CE Router	8
3.3. Originating a C-MCAST Join Route at the PE Router	9
3.4. Receiving a C-MCAST Join Route by the PE Router	10
4. Pruning Sources off the Shared Tree	11
5. Acknowledgements	12
6. IANA Considerations	12
7. Security Considerations	12
8. References	12
8.1. Normative References	12
8.2. Informative References	12
Authors' Addresses	13

1. Introduction

When a Service Provider offers BGP/MPLS IP VPN service to its customers, [RFC6513] and [RFC6514] describe protocols and procedures that the Service Provider can use in order to carry the customer's IP multicast traffic from one customer site to others. BGP can be used to carry customer multicast routing information from one Provider

Edge (PE) router to another, but it is assumed that PIM is running on the interface between a Customer Edge (CE) router and a PE router. This document specifies protocols and procedures that under certain conditions, allow customer multicast routing information to be carried between PE and CE via BGP. This can eliminate the need to run PIM on the PE-CE interfaces, potentially eliminating the need to run PIM on the PE routers at all. It is however assumed that PIM is the multicast routing protocol running at the customer sites.

This document defines a new SAFI known as Customer-Multicast (C-MCAST) SAFI. This SAFI is used to carry customer multicast routing information from CE to PE (and vice versa). The use of this SAFI is only defined when the AFI is either IPv4 or IPv6. The procedures of this document are applicable only if BGP is being used as the PE-PE control protocol for carrying customer multicast routing. It is presupposed that if these procedures are being used on any interface of a given VRF, then PIM is NOT enabled on that interface or on any other VRF interface of that same VRF. It is also assumed that if a CE is using BGP C-MCAST on its interface to one PE, then it is using BGP C-MCAST on its interfaces to all the PEs to which it is connected, and that PIM is not enabled on any of these interfaces.

Throughout this document, we will use the terms "MCAST-VPN route" and "C-MCAST route" to mean routes that have the corresponding SAFI.

This document assumes that a CE and a PE exchange C-MCAST routes over a direct BGP session (i.e., the C-MCAST routes do not pass through a route reflector or other third party on their way from CE to PE, or vice versa).

The NLRI format of a C-MCAST route is modeled on the NLRI format of the MCAST-VPN SAFI, as defined in [RFC6514]. However, since the C-MCAST routes are always exchanged in the context of a particular VPN, Route Distinguishers (RDs) are not used. Also, C-MCAST routes are never distributed from one PE to another.

Where the procedures of this document require a PE or a CE to determine the upstream neighbor for a particular multicast (*,G) or (S,G) state, the upstream neighbor is determined using the procedures of [RFC4601], rather than the procedures of [RFC6513] or [RFC6514]. That is, the upstream neighbor selected for a particular (*,G) or (S,G) is the same as it would be if PIM were being used between the PE and CE. This includes support for non-congruent multicast topologies. The upstream neighbor address determined through these procedures MUST be the same address that the upstream neighbor puts in the next hop field of BGP Updates that it sends to the "downstream" PE or CE. Note that neither the VRF Route Import

Extended Community nor the Source AS Extended Community, defined in [RFC6514], are used.

When following the procedures of this document, customer IP multicast traffic is sent natively from CE to PE, and vice versa. There is no encapsulation or tunneling of the multicast traffic. Therefore there is no need for C-MCAST routes corresponding to the I-PMSI A-D routes or S-PMSI A-D routes or [RFC6514]. Similarly, there is no need for the PMSI Tunnel attribute of [RFC6514].

This document does not provide a mechanism corresponding to the "PIM Assert" mechanism of [RFC4601]. It is assumed either that the PE-CE interfaces are point-to-point interfaces, or else that the multicast procedures in the PE and CE can determine, by examination of the layer 2 headers, the node from which a given multicast data packet was received.

Support for "Dense Mode Multicast" (PIM-DM) is outside the scope of this document.

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 [RFC2119].

2. C-MCAST SAFI NLRI Format

The BGP Multiprotocol Extensions [RFC4760] allow BGP to carry routes from multiple different "AFI/SAFIs". This document defines a new a new SAFI known as a C-MCAST SAFI with a value to be assigned by the IANA. This SAFI is used along with the AFI of IPv4 (1) or IPv6 (2).

The C-MCAST NLRI defined below is carried in the BGP UPDATE messages [RFC4271] using the BGP multiprotocol extensions [RFC4760] with a AFI of IPv4 (1) or IPv6 (2) assigned by IANA and a C-MCAST SAFI with a value to be assigned by the IANA.

The Next hop field of MP_REACH_NLRI attribute SHALL be interpreted as an IPv4 address whenever the length of the Next Hop address is 4 octets, and as an IPv6 address whenever the length of the Next Hop is address is 16 octets.

The NLRI field in the MP_REACH_NLRI and MP_UNREACH_NLRI is a prefix with a maximum length of 12 octets for IPv4 AFI and 36 octets for IPv6 AFI. The following is the format of the C-MCAST NLRI:

+-----+

	Route Type (1 octet)	
+-----+		
	Length (1 octet)	
+-----+		
	Route Type specific (variable)	
+-----+		

The Route Type field defines encoding of the rest of the C-MCAST NLRI. (Route Type specific C-MCAST NLRI).

The Length field indicates the length in octets of the Route Type specific field of C-MCAST NLRI.

This document defines the following Route Types:

- 1 - Shared Tree Join Route
- 2 - Source Tree Join Route
- 4 - Source Prune A-D Route

The encodings and procedures for these route types are described in the subsequent sections. The NLRI encodings are modeled after those of [RFC6514], in order to facilitate implementation in generation of MCAST-VPN routes.

In order for two BGP speakers to exchange C-MCAST NLRI, they must use BGP Capabilities Advertisement [RFC5492] to ensure that they both are capable of properly processing the C-MCAST NLRI. This is done as specified in [RFC4760], by using a capability code 1 (multiprotocol BGP) with an AFI of IPv4 (1) or IPv6 (2) and a SAFI of C-MCAST with a value to be assigned by IANA.

2.1. C-MCAST Join and Prune Routes

The "route type specific" part of the C-MCAST NLRI is the same for the Shared Tree Join, Source Tree Join, and Source Prune A-D Route Types.

```

+-----+
| Multicast Source Length (1 octet) |
+-----+
|   Multicast Source (variable)   |
+-----+
| Multicast Group Length (1 octet) |
+-----+
|   Multicast Group (variable)   |
+-----+

```

The value of the AFI field in the MP_REACH_NLRI/MP_UNREACH_NLRI that carries the C-MCAST NLRI determines whether the multicast source and multicast group addresses fields are IPv4 addresses (AFI 1) or IPv6 addresses (AFI 2). The length field of IPv4 source and group addresses is 32 bits and the length field of IPv6 addresses is 128 bits.

Use of other values of the Multicast Source Length and Multicast Group Length fields is outside the scope of this document. If other values occur, or if the NLRI length is not as expected, given the AFI value, the NLRI should be considered to be malformed. An implementation SHOULD treat such an UPDATE as though the NLRI has been withdrawn, SHOULD log an error.

In a Source Tree Join route, the Multicast Source field and the Multicast Group field identify an (S,G) multicast flow, where the IP address of S appears in the Multicast Source field and the IP address of G appears in the Multicast Group field.

In a Shared Tree Join route, the Multicast Source field contains the Rendezvous Point (RP) address that is associated, at the originator of the UPDATE, with the multicast group whose address appears in the Multicast Group field.

In the Source Prune route, the Multicast Source field contains the address of a source that is to be pruned from the shared tree associated with the multicast group whose address appears in the Multicast Group field.

Usage of C-MCAST Join routes is described in Section 3. Usage of C-MCAST Source Prune routes is described in Section 4.

3. Exchanging C-MCAST Join Routes

Multicast Routing Information is exchanged between a CE router and a PE router using C-MCAST NLRI's. If the procedures of this draft are followed, the CE router and the PE router MUST NOT exchange PIM messages with each other. The C-MCAST routes are originated and propagated as follows.

3.1. Originating C-MCAST Join Route at the CE router

Whenever a) PIM instance on a particular CE router creates a new (S,G) or (*,G) state and b) the selected upstream neighbor address for that state is a PE router to which the CE router is directly connected and with which the CE has a BGP session on which the C-MCAST SAFI is enabled, then the CE router MUST originate a C-MCAST Source Tree Join route or a C-MCAST Shared Tree Join route. The CE router uses the procedures of [RFC4601] to determine the upstream neighbor (also called the "RPF neighbor"). Note that, unlike [RFC6513], the upstream nexthop selection procedures defined in this document are not based on the VRF Route Import Extended community [RFC6513].

The Route Type field of the NLRI is set to "Source Tree Join" if the corresponding state is (S,G), and to "Shared Tree Join" if the corresponding state is (*,G).

The Multicast Source field of the NLRI of a C-MCAST Source Tree Join route MUST be set to the source address of the corresponding (S,G). The Multicast Source field of the NLRI of a C-MCAST Shared Tree Join route MUST be set to the address of the RP that is mapped to the corresponding (*,G) state.

The Multicast Group field of the NLRI of either kind of C-MCAST Join route MUST be set to the multicast group address of the (S,G) or (*,G) state.

The CE router MUST put its own IP address in the Next Hop field of either type of C-MCAST Join route.

The CE router MUST attach a particular RT to the C-MCAST Join route. This RT MUST be either an IP Address Specific RT or an IPv6 Address Specific RT, depending upon whether the address of the upstream neighbor is an IPv4 address or an IPv6 address. In either case, the address of the upstream neighbor is placed in the Global Administrator field of the RT, and the Local Administrator field is set to 0. (Compare to the "C-multicast Import RT" described in section 7 of [RFC6514].) The "address of the upstream neighbor" MUST be the address that the upstream neighbor puts in the Next Hop field of BGP UPDATES that it sends to the CE router.

The CE MUST send the C-MCAST Join route on the BGP session to the PE that it has selected as the upstream neighbor for the (*,G) or (S,G) identified in the NLRI of the route. Although not required for correctness, it is RECOMMENDED that the C-MCAST Join not be sent on other BGP sessions. How this distribution constraint is met is a local matter. Policy based filtering based on the RT is one possible way to meet the constraint. Use of [RFC4684] is another.

The C-MCAST Shared Tree Join is used to join the shared tree of an "Any Source Multicast" (ASM) group. The C-MCAST Source Tree Join is used in both ASM mode and in "Single Source Multicast" (SSM) mode to join a source tree. The C-MCAST Shared Tree Join can be used to join a BIDIR-PIM [RFC5015] multicast group, as long as the interface between the PE and the CE is a point-to-point interface.

If a PIM instance on a particular CE router deletes its (S,G) or a (*,G) entry, and if there is a currently originated corresponding C-MCAST Join route, then that route MUST be withdrawn. The C-MCAST route is withdrawn using the MP_UNREACH_NLRI attribute. If the upstream neighbor of a (S,G) or (*,G) route changes, and if there is a currently originated corresponding C-MCAST Join route, then the RT MUST be changed to identify the new upstream neighbor, and the route MUST be advertised on the BGP session to that neighbor.

3.2. Receiving a C-MCAST Join Route by the CE Router

When a CE router receives a C-MCAST route from its PE router, it checks to see if the route carries an IP Address Specific Route Target Extended Community whose Global Administrator field contains the address that the CE puts in the Next Hop field of the BGP UPDATES it sends to the PE router. If there is no such Route Target, the received route does not impact the CE's multicast state. Otherwise, the received route is used to create or modify the corresponding (S,G) or (*,G) state in the multicast forwarding information base (FIB). The CE-to-PE interface is stored in the outgoing interface list of the corresponding (S,G) or a (*,G) state. If the C-MCAST route is received with the Route Type of Shared Tree Join, then the CE router attaches RP address to the corresponding (*,G) state.

If the CE router is connected to the multiple PE routers, it is possible that it will receive C-MCAST routes with the same NLRI from multiple PE routers. In this case, the CE router adds multiple CE-to-PE interfaces to its outgoing interface list, one for each PE router from which the given route was received. (Note that a particular CE-to-PE interface is added only if the route from the corresponding PE identifies the CE in its IP Address Specific RT.)

When the last C-MCAST route for a given (S,G) or a (*,G) is withdrawn, resulting in a state where BGP C-MCAST SAFI has no route for a given (S,G) or a (*,G) state, the CE router MUST remove all the interfaces learnt via BGP from the outgoing interface list. If this results in an empty outgoing interface list then the CE router using PIM procedures MUST prune itself off the corresponding (S,G) or a (*,G) tree.

3.3. Originating a C-MCAST Join Route at the PE Router

The C-MCAST routes on a PE router are originated in BGP as a result of updates in the (C-S,C-G) or (C-*,C-G) state in the associated VRF of a PE router. These states are created by BGP routes learnt from other PEs via the BGP MCAST-VPN SAFI [RFC6514], or from CEs via the C-MCAST SAFI. Whenever a) a particular VRF on a particular PE router creates a new (C-S,C-G) or a (C-*,C-G) state, b) the selected upstream neighbor address identifies a CE, and c) the PE has a direct BGP session to the CE, on which C-MCAST SAFI is enabled, then the PE router MUST originate a C-MCAST Join route from the given VRF. The PE router finds the upstream neighbor address based on the procedures of [RFC6513].

The Route Type field of the NLRI is set to "Source Tree Join" if the corresponding state is (S,G), and to "Shared Tree Join" if the corresponding state is (*,G).

The Multicast Source field of the NLRI of a C-MCAST Source Tree Join route MUST be set to the source address of the corresponding (S,G). The multicast source address field of the NLRI of a C-MCAST Shared Tree Join route MUST be set to the address of the RP that is mapped to the corresponding (*,G) state.

The Multicast Group field of the NLRI of either kind of C-MCAST Join route MUST be set to the multicast group address of the (S,G) or (*,G) state.

The PE router must put its own IP address in the Next Hop field of the C-MCAST Join route.

The PE router MUST attach a particular RT to the C-MCAST Join route. This RT MUST be either an IP Address Specific RT or an IPv6 Address Specific RT, depending upon whether the address of the upstream neighbor is an IPv4 address or an IPv6 address. In either case, the address of the upstream neighbor is placed in the Global Administrator field of the RT, and the Local Administrator field is set to 0. (Compare to the "C-multicast Import RT" described in section 7 of [RFC6514].) The "address of the upstream neighbor" MUST be the address that the upstream neighbor puts in the Next Hop field of BGP UPDATES that it sends to the PE router.

The PE MUST send the C-MCAST Join route on the BGP session to the CE that it has selected as the upstream neighbor for the (*,G) or (S,G) identified in the NLRI of the route. Although not required for correctness, it is RECOMMENDED that the C-MCAST Join not be sent on other BGP sessions. How this distribution constraint is met is a local matter. Policy based filtering based on the RT is one possible way to meet the constraint. Use of [RFC4684] is another. Of course, a C-MCAST route originated by a particular PE is originated in the context of a particular VRF, and MUST NOT be advertised to a particular CE unless the interface between that PE and that CE is associated with that VRF.

The C-MCAST Shared Tree Join is used to join the shared tree of an "Any Source Multicast" (ASM) group. The C-MCAST Source Tree Join is used in both ASM mode and in "Single Source Multicast" (SSM) mode to join a source tree. The C-MCAST Shared Tree Join can be used to join a BIDIR-PIM [RFC5015] multicast group, as long as the interface between the PE and the CE is a point-to-point interface.

If a PE router deletes its (S,G) or a (*,G) entry in the context of a particular VRF, and if there is a currently originated corresponding C-MCAST Join route from that VRF, then that route MUST be withdrawn. The C-MCAST route is withdrawn using the MP_UNREACH_NLRI attribute. If the upstream neighbor of a (S,G) or (*,G) route changes, and if there is a currently originated corresponding C-MCAST Join route, then the RT MUST be changed to identify the new upstream neighbor.

3.4. Receiving a C-MCAST Join Route by the PE Router

When a PE router receives a C-MCAST Join route from a CE router, it checks to see if the route carries an IP Address Specific Route Target Extended Community whose Global Administrator field contains the address that the PE puts in the Next Hop field of the BGP UPDATES it sends to the CE router. If there is no such Route Target, the received route does not impact the PE's multicast state. Otherwise, the received route is used to create or modify the corresponding (S,G) or (*,G) state in the MVPN-TIB ([RFC6514]). The PE-to-CE

interface is stored in the outgoing interface list of the corresponding (S,G) or a (*,G) state. If the C-MCAST route is received with the Route Type of Shared Tree Join, then the CE router attaches RP address to the corresponding (*,G) state.

If the PE router is connected to the multiple CE routers, it is possible that it will receive C-MCAST routes with the same NLRI from multiple CE routers. In this case, the PE router adds multiple PE-to-CE interfaces to its outgoing interface list, one for each CE router from which the given route was received. (Note that a particular PE-to-CE interface is added only if the route from the corresponding CE identifies the PE in its IP Address Specific RT.)

When the last C-MCAST route for a given (S,G) or a (*,G) is withdrawn, resulting in a state where BGP C-MCAST SAFI has no route for a given (S,G) or a (*,G) state, the withdrawal of the route has the (S,G) or a (*,G) prune semantics. The corresponding MCAST-VPN route is withdrawn using the MP_UNREACH_NLRI attribute.

4. Pruning Sources off the Shared Tree

Suppose a router, say R1, has originated a C-MCAST Source Tree Join A-D route for (*,G), and has identified another router, say R2, in that route's RT. Under certain conditions, R1 may need to prune a particular source, say S1, off the (*,G) tree. To do so, R1 originates a C-MCAST Prune Source A-D route whose NLRI contains S1 in the multicast source field and G in the multicast group field. This route MUST have the same IP Address Specific RT (identifying R2), and the same Next Hop field, as the corresponding C-MCAST Source Tree Join A-D route. This route MUST be sent on the BGP session to R2. It is RECOMMENDED that it not be sent on other BGP sessions. Note that either R1 is a CE and R2 is a PE, or vice versa. The procedures are the same in either case. When R1 no longer needs to prune S1 from the (*,G) tree, R1 MUST withdraw the (S1,G) Source Prune route. If R1 changes the RT on the (*,G) Shared Tree Join route, it MUST change the RT on all the corresponding (S,G) Source Prune routes. If R1 withdraws the (*,G) Shared Tree Join route, it MUST also withdraw all the corresponding (S,G) Source Prune routes. When a router withdraws a Source Tree Join route for (*,G), it MUST withdraw all its Source Prune (S,G) routes. A received Source Prune (S,G) route does not impact a router's multicast state unless there is a the router has also received a Shared Tree Join (*,G) route over the same BGP session. However, a router should handle the case where one or more Source Prune routes are received before the corresponding Shared Tree Join route. Further details will be provided in future revisions of this document.

5. Acknowledgements

The authors would like to thank for his review and comments.

6. IANA Considerations

This document define a new SAFI called "C-MCAST SAFI". IANA is requested to allocate a code point for C-MCAST SAFI.

7. Security Considerations

This extension to BGP does not change the underlying security issues inherent in the existing [RFC4271] and [RFC6514].

8. References

8.1. Normative References

- [RFC2119] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, March 1997.
- [RFC4271] Rekhter, Y., Li, T., and S. Hares, "A Border Gateway Protocol 4 (BGP-4)", RFC 4271, January 2006.
- [RFC4601] Fenner, B., Handley, M., Holbrook, H., and I. Kouvelas, "Protocol Independent Multicast - Sparse Mode (PIM-SM): Protocol Specification (Revised)", RFC 4601, August 2006.
- [RFC4760] Bates, T., Chandra, R., Katz, D., and Y. Rekhter, "Multiprotocol Extensions for BGP-4", RFC 4760, January 2007.
- [RFC5015] Handley, M., Kouvelas, I., Speakman, T., and L. Vicisano, "Bidirectional Protocol Independent Multicast (BIDIR-PIM)", RFC 5015, October 2007.
- [RFC5492] Scudder, J. and R. Chandra, "Capabilities Advertisement with BGP-4", RFC 5492, February 2009.
- [RFC6513] Rosen, E. and R. Aggarwal, "Multicast in MPLS/BGP IP VPNs", RFC 6513, February 2012.
- [RFC6514] Aggarwal, R., Rosen, E., Morin, T., and Y. Rekhter, "BGP Encodings and Procedures for Multicast in MPLS/BGP IP VPNs", RFC 6514, February 2012.

8.2. Informative References

[RFC4684] Marques, P., Bonica, R., Fang, L., Martini, L., Raszuk, R., Patel, K., and J. Guichard, "Constrained Route Distribution for Border Gateway Protocol/MultiProtocol Label Switching (BGP/MPLS) Internet Protocol (IP) Virtual Private Networks (VPNs)", RFC 4684, November 2006.

Authors' Addresses

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

Email: keyupate@cisco.com

Yakov Rekhter
Juniper Networks
1194 North Mathilda Avenue
Sunnyvale, CA 94089
USA

Email: yakov@juniper.net

Eric Rosen
Cisco Systems
1414 Massachusetts Avenue
Boxborough, MA 01719
USA

Email: erosen@cisco.com

Network Working Group
Internet-Draft
Intended status: Standards Track
Expires: January 13, 2014

Z. Li
S. Zhuang
H. Ni
Huawei Technologies
July 12, 2013

Connecting IPv6 Multicast Islands over IPv4 MPLS Using IPv6 Provider
Edge Routers (6PE)
draft-li-idr-mcast-6pe-00

Abstract

This document defines a new Network Layer Reachability Information (NLRI), called as the MCAST-6PE NLRI. The MCAST-6PE NLRI is used to interconnect IPv6 C-Multicast islands over a Multiprotocol Label Switching (MPLS)-enabled IPv4 cloud. This approach relies on IPv6 Provider Edge routers (6PE), which can exchange the IPv6 C-Multicast reachability information transparently over the core using the Multiprotocol Border Gateway Protocol (MP-BGP) over IPv4. This document describes the BGP encodings and procedures for exchanging the information elements required by IPv6 Multicast in 6PE. MPLS-based Service Providers may use the 6PE Multicast mechanism to provide IPv6 Multicast service for customers.

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 13, 2014.

Copyright Notice

Copyright (c) 2013 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	3
2. Terminology	3
3. MCAST-6PE NLRI	4
3.1. Intra-AS 6PE I-PMSI A-D Route	5
3.2. Inter-AS 6PE I-PMSI A-D route	6
3.3. 6PE S-PMSI A-D Route	6
3.4. 6PE Leaf A-D Route	7
3.5. 6PE Source Active A-D Route	7
3.6. 6PE C-Multicast Route	8
4. PMSI Tunnel Attribute	9
5. Source AS Extended Community	9
6. Route Import Extended Community	9
7. PE Distinguisher Labels Attribute	10
8. Operations	10
8.1. BGP-Based MCAST-6PE Membership Auto-Discovery	10
8.1.1. Intra-AS Operations	10
8.1.2. Inter-AS Operations	11
8.2. PE-PE Transmission of IPv6 C-Multicast Routing	11
8.2.1. Selecting the Upstream Multicast Hop (UMH)	11
8.2.2. Signaling P-tunnel	12
8.2.3. Use of BGP for Carrying IPv6 C-Multicast Routing	12
8.2.4. Propagating IPv6 C-Multicast Routes by an ASBR	14
8.3. Using 6PE S-PMSI A-D Routes to Bind C-Trees to P-Tunnels	14
9. IANA Considerations	14
10. Security Considerations	14
11. References	14
11.1. Normative References	14
11.2. Informative References	15
Authors' Addresses	15

1. Introduction

6PE[RFC4798] defines a mechanism to interconnect IPv6 islands over an MPLS-enabled IPv4 cloud using the IPv6 Provider Edge routers (6PE) approach. In this document an 'IPv6 island' is a network running native IPv6 as per [RFC2460]. A typical example of an IPv6 island would be a customer's IPv6 site connected via its IPv6 Customer Edge (CE) router to one (or more) Dual Stack Provider Edge router(s) of a Service Provider. These IPv6 Provider Edge routers (6PE) are connected to an IPv4 MPLS core network. This document defines a new Network Layer Reachability Information (NLRI), called as the MCAST-6PE NLRI. The MCAST-6PE NLRI is used to interconnect IPv6 C-Multicast islands over a Multiprotocol Label Switching (MPLS)-enabled IPv4 cloud. This approach relies on IPv6 Provider Edge routers (6PE), which can exchange the IPv6 C-Multicast reachability information transparently over the core using the Multiprotocol Border Gateway Protocol (MP-BGP) over IPv4. This document describes the BGP encodings and procedures for exchanging the information elements required by IPv6 Multicast in 6PE. MPLS-based Service Providers may use the 6PE Multicast mechanism to provide IPv6 Multicast service for customers.

2. Terminology

This document uses terminology from [RFC4798], [RFC6513], [RFC6514].

Term Definition

6PE: IPv6 Provider Edge routers

A-D: auto-discovery

BGP: Border Gateway Protocol

CE: customer edge

C-G: customer multicast group address

C-join: customer join message

C-multicast: customer multicast

C-PIM: customer PIM

C-RP: customer rendezvous point

C-RPT: customer RP Tree

C-S: customer multicast source address

I-PMSI: inclusive PMSI

LSP: label switched path

MCAST: multicast

mLDP: multipoint Label Distribution Protocol

MP2MP: multipoint to multipoint

MVPN: multicast VPN

NG MVPN: next-generation multicast VPN

NLRI: Network Layer Reachability Information

OIL: outgoing interface list

P2MP: point to multipoint PE: provider edge

PIM: Protocol Independent Multicast

PMSI: Provider Multicast Service Interface

P-group: Provider multicast group

P-join: Provider join message

P-PIM: Provider PIM

P-RP: Provider Rendezvous Point

SAFI: Subsequent Address Family Identifier

S-PMSI: Selective PMSI

UMH: Upstream Multicast Hop

3. MCAST-6PE NLRI

This document defines a new BGP NLRI, called as the MCAST-6PE NLRI. Following is the format of the MCAST-6PE NLRI:

	Route Type (1 octet)	
	Length (1 octet)	
	Route Type specific (variable)	

The Route Type field defines the encoding of the rest of MCAST-6PE NLRI (Route Type specific MCAST-6PE NLRI). The Length field indicates the length in octets of the Route Type specific field of the MCAST-6PE NLRI. This document defines the following Route Types for A-D routes:

- + 1 - Intra-AS 6PE I-PMSI A-D route;
- + 2 - Inter-AS 6PE I-PMSI A-D route;
- + 3 - 6PE S-PMSI A-D route;
- + 4 - 6PE Leaf A-D route;
- + 5 - 6PE Source Active A-D route.

This document defines the following Route Types for IPv6 C-multicast routes:

- + 6 - 6PE Shared Tree Join route;
- + 7 - 6PE Source Tree Join route;

The MCAST-6PE NLRI is carried in BGP using BGP Multiprotocol Extensions [RFC4760] with an AFI of 2 (IPv6 AFI), and a SAFI of MCAST-6PE [To be assigned by IANA]. The NLRI field in the MP_REACH_NLRI / MP_UNREACH_NLRI attribute contains the MCAST-6PE NLRI (encoded as specified above). The following sections describe the format of the Route Type specific MCAST-6PE NLRI for various Route Types defined in this document.

3.1. Intra-AS 6PE I-PMSI A-D Route

An Intra-AS 6PE I-PMSI A-D Route Type specific MCAST-6PE NLRI consists of the following:

```

+-----+
|   Originating Router's IP Addr   |
+-----+

```

Originating Router's IP Addr field set to the IP address of the MCAST 6PE router originating this route, which is typically the primary loopback address of the MCAST 6PE router.

All MCAST 6PE routers create and advertise a Type 1 intra-AS 6PE I-PMSI A-D route for IPv6 MCAST service to which they are connected.

3.2. Inter-AS 6PE I-PMSI A-D route

An Inter-AS 6PE I-PMSI A-D Route Type specific MCAST-6PE NLRI consists of the following:

```

+-----+
|   Source AS (4 octets)           |
+-----+

```

The Source AS contains an Autonomous System Number (ASN), 4 octets.

Two-octet ASNs are encoded in the two low-order octets of the Source AS field, with the two high-order octets set to zero.

Type 2 routes are used for MCAST 6PE membership discovery between MCAST-6PE routers that belong to different ASes.

3.3. 6PE S-PMSI A-D Route

A 6PE S-PMSI A-D Route Type specific MCAST-6PE NLRI consists of the following:

```

+-----+
| Multicast Source Length (1 octet) |
+-----+
| Multicast Source (variable)       |
+-----+
| Multicast Group Length (1 octet)  |
+-----+
| Multicast Group (variable)        |
+-----+
| Originating Router's IP Addr      |
+-----+

```

For MCAST-6PE, the Multicast Source field contains the C-S address i.e. the address of the multicast source, which is an IPv6 address, then the value of the Multicast Source Length field is 128 bits.

For MCAST-6PE, the Multicast Group field contains the C-G address i.e. the address of the multicast group, which is an IPv6 address, then the value of the Multicast Group Length field is 128 bits.

The Originating Router's IP Addr field set to the IP address of the MCAST-6PE router originating this route, which is typically the primary loopback address of the MCAST-6PE router.

A sender MCAST-6PE that initiates a selective P-tunnel is required to originate a Type 3 6PE S-PMSI A-D route with the appropriate PMSI attribute.

3.4. 6PE Leaf A-D Route

A 6PE Leaf A-D Route Type specific MCAST-6PE NLRI consists of the following:

```

+-----+
|           Route Key (variable)           |
+-----+
|           Originating Router's IP Addr           |
+-----+

```

The Route Key field contains the original Type 3 route received. The Originating Router's IP Addr field set to the IP address of the MCAST-6PE originating the 6PE leaf A-D route, typically the primary loopback address.

A 6PE Leaf A-D routes may be originated as a result of processing a received Inter-AS 6PE I-PMSI A-D route [Type 2] or 6PE S-PMSI A-D route [Type 3]. A 6PE Leaf A-D route is originated in these situations only if the received route has a PMSI Tunnel attribute whose "Leaf Information Required" bit is set to 1.

Typically a receiver MCAST-PE router responds to a Type 3 route by originating a Type 4 6PE leaf A-D route if it has local receivers interested in the traffic transmitted on the selective P-tunnel. The Type 4 route informs the sender MCAST-6PE of the leaf MCAST-6PE routers.

3.5. 6PE Source Active A-D Route

A 6PE Source Active A-D Route Type specific MCAST-6PE NLRI consists of the following:

```

+-----+
| Multicast Source Length (1 octet) |
+-----+
| Multicast Source (variable)       |
+-----+
| Multicast Group Length (1 octet)  |
+-----+
| Multicast Group (variable)        |
+-----+

```

For MCAST-6PE, the Multicast Source field contains the C-S address i.e. the address of the multicast source, which is an IPv6 address, then the value of the Multicast Source Length field is 128 bits.

For MCAST-6PE, the Multicast Group field contains the C-G address i.e. the address of the multicast group, which is an IPv6 address, then the value of the Multicast Source Length field is 128 bits.

Type 5 6PE Source Active A-D routes carry information about active IPv6 Multicast sources and the groups to which they are transmitting data. These routes can be generated by any MCAST-6PE router that becomes aware of an active source.

3.6. 6PE C-Multicast Route

A 6PE Shared Tree Join Route and a 6PE Source Tree Join Route Type specific MCAST-6PE NLRI consists of the following:

```

+-----+
| Source AS (4 octets)              |
+-----+
| Multicast Source Length (1 octet) |
+-----+
| Multicast Source (variable)       |
+-----+
| Multicast Group Length (1 octet)  |
+-----+
| Multicast Group (variable)        |
+-----+

```

The Source AS contains an ASN, 4 octets. Two-octet ASNs are encoded in the low-order two octets of the Source AS field.

For MCAST-6PE, the Multicast Source field contains an IPv6 address, then the value of the Multicast Source Length field is 128 bits. For a 6PE Shared Tree Join Route, the Multicast Source field contains the C-RP address; for a 6PE Source Tree Join Route, the Multicast Source field contains the C-S address.

For MCAST-6PE, the Multicast Group field contains an IPv6 address, then the value of the Multicast Group Length field is 128 bits. The Multicast Group field contains the C-G address.

The 6PE C-Multicast Routes exchange between MCAST-6PE routers refers to the propagation of C-joins from receiver MCAST-6PEs to the sender MCAST-6PEs.

In a 6PE MCAST Network, IPv6 C-joins received by MCAST-6PE Router from the CEs are encoded as BGP 6PE C-Multicast Routes and advertised via 6PE C-Multicast Routes towards the sender MCAST-6PEs. Two types of 6PE C-Multicast Routes are specified. The Type 6 6PE C-Multicast Routes are used in representing information contained in a shared tree (C-*, C-G) join. The Type 7 6PE C-Multicast Routes are used in representing information contained in a source tree (C-S, C-G) join.

4. PMSI Tunnel Attribute

The usage of PMSI Tunnel Attribute is described in [RFC6514].

5. Source AS Extended Community

The Source AS is an AS-specific Extended Community, of an extended type, and is transitive across AS boundaries [RFC4360]. The Global Administrator field of this Community MUST be set to the ASN of the MCAST-6PE router. The Local Administrator field of this Community MUST be set to 0.

The usage of a received Source AS Extended Community in MCAST 6PE is the same as described in [RFC6514].

6. Route Import Extended Community

This document defines a new BGP Extended Community called "Route Import", type value is to be assigned by IANA. The Route Import Extended Community is an IP-address-specific extended community that is used for importing IPv6 C-Multicast routes in the active sender MCAST-6PE router's MCAST-6PE routing table to which the source is attached. For MCAST-6PE Network case, for constructing IPv6 C-Multicast Import RT, the Local Administrator is set to 0 and the Global Administrator field MUST be set to an IP address of the MCAST-6PE router.

7. PE Distinguisher Labels Attribute

The usage of PE Distinguisher Labels Attribute is described in [RFC6513].

8. Operations

8.1. BGP-Based MCAST-6PE Membership Auto-Discovery

This section specifies procedures for the auto-discovery of MCAST-6PE memberships and the distribution of information used to instantiate I-PMSIs.

There are two MCAST-6PE auto-discovery mechanisms, dubbed "intra- AS" and "inter-AS" respectively. The intra-AS mechanisms provide auto-discovery within a single AS. The inter-AS mechanisms provide auto-discovery across multiple ASes when segmented inter-AS tunnels are being used.

BGP-Based MCAST-6PE Membership Auto-Discovery is done by means of a new address family, the MCAST-6PE address family. Any PE that attaches to a MCAST-6PE service MUST issue a BGP Update message containing a NLRI in this address family, along with a specific set of attributes.

8.1.1. Intra-AS Operations

This section describes exchanges of Type 1 Intra-AS 6PE I-PMSI A-D routes originated/received by PEs within the same AS.

To participate in the MCAST-6PE auto-discovery, a PE router that provides MCAST 6PE service MUST originate an Intra-AS 6PE I-PMSI A-D route and advertises this route in IBGP. The route is constructed as follows.

The route carries a single MCAST-6PE NLRI with the Originating Router's IP Addr field set to the IP address of the MCAST 6PE router originating this route. Note that the <Originating Router's IP Addr> uniquely identifies a given MCAST-6PE router.

The route carries the PMSI Tunnel attribute if and only if an I-PMSI is used for the MCAST-6PE (the conditions under which an I-PMSI is used can be found in [RFC6513]). Depending on the technology used for the P-tunnel for the MCAST-6PE on the PE, the PMSI Tunnel attribute of the Intra-AS 6PE I-PMSI A-D route is the same as described in [RFC6514].

The Next Hop field of the MP_REACH_NLRI attribute of the route MUST be set to the same IP address as the one carried in the Originating Router's IP Addr field.

When PE-PE Type 1 intra-AS 6PE I-PMSI A-D routes are exchanged among all provider routers, every PE can know the MCAST-6PE neighbors to itself.

8.1.2. Inter-AS Operations

This section applies only to the case where segmented inter-AS tunnels are used.

Type 2 routes are used for MCAST 6PE membership discovery between MCAST-6PE routers that belong to different ASs.

If an ASBR is configured to support MCAST 6PE service, the ASBR MUST participate in the intra-AS MCAST 6PE auto-discovery procedures, for that MCAST 6PE within the ASBR's own AS, as specified in Section 8.1.1; "Intra-AS Operations".

A Type 2 Inter-AS 6PE I-PMSI A-D route for MCAST 6PE originated by an ASBR within a given AS is propagated via BGP to other ASes.

The route carries a single MCAST-6PE NLRI with the Source AS field set to the ASBR's own AS.

When re-advertising an Inter-AS 6PE I-PMSI A-D route, the ASBR MUST set the Next Hop field of the MP_REACH_NLRI attribute to a routable IP address of the ASBR.

8.2. PE-PE Transmission of IPv6 C-Multicast Routing

IPv6 C-Multicast Routing Information is exchanged among PEs by using 6PE C-multicast routes that are carried using an MCAST-6PE NLRI. These routes are originated and propagated as follows.

8.2.1. Selecting the Upstream Multicast Hop (UMH)

Section 5.1 of [RFC6513] describes the method of Selecting the Upstream Multicast Hop (UMH). Constructing the C-Multicast Import RT as specified in Section 7 of [RFC6514].

For a PE as the MCAST-6PE sender, issues the UMH route through a 6PE UCAST route carrying Route Import Extended Community and Source AS Extended Community.

The Route Import Extended Community is an IP-address-specific extended community that is used for importing IPv6 C-Multicast routes in the active sender MCAST-6PE's MCAST-6PE routing table to which the source is attached. For MCAST-6PE Network case, for constructing IPv6 C-Multicast Import RT, the Local Administrator is set to 0 and the Global Administrator field MUST be set to an IP address of the MCAST-6PE router.

8.2.2. Signaling P-tunnel

The PMSI tunnel attribute carries information about the P-tunnel. In a MCAST-6PE Network, the sender PE router sets up the P-tunnel, and therefore is responsible for originating the PMSI tunnel attribute. The PMSI tunnel attribute can be attached to Type 1, Type 2, and Type 3 routes.

The MCAST-6PE sender router, attaches a PMSI tunnel attribute to Type 1 Intra-AS 6PE I-PMSI A-D Route, begins to signal P-tunnel for MCAST-6PE Network.

MCAST-6PE sender sends Type 1 route to other PEs, when other PEs receive the Type 1 route with PMSI tunnel attribute from MCAST-6PE sender, then join the P-tunnel.

8.2.3. Use of BGP for Carrying IPv6 C-Multicast Routing

Part of the procedures for constructing MCAST-6PE NLRI depends on the multicast routing protocol between CE and PE (C-multicast protocol).

8.2.3.1. PIM as the C-Multicast Protocol

Whenever (a) a C-PIM instance on a particular PE creates a new (C-S,C-G) state, and (b) the selected upstream PE for C-S (see [RFC6513]) is not the local PE, then the local PE MUST originate a C-multicast route of type Source Tree Join. The Multicast Source field in the MCAST-6PE NLRI of the route is set to C-S; the Multicast Group field is set of C-G.

This C-multicast route is said to "correspond" to the C-PIM (C-S,C-G) state.

The semantics of the route are such that the PE has one or more receivers for (C-S,C-G) in the sites connected to the PE (the route has the (C-S,C-G) Join semantics).

Whenever a C-PIM instance on a particular PE deletes a (C-S,C-G) state, the corresponding C-multicast route MUST be withdrawn. (The withdrawal of the route has the (C-S,C-G) Prune semantics). The

MCAST-6PE NLRI of the withdrawn route is carried in the MP_UNREACH_NLRI attribute.

8.2.3.1.1. Source Tree Join (C-S, C-G)

When receiver PE receives a source tree join (C-S, C-G) from CE, it does a route look up for C-S. If there is more than one route, the receiver PE chooses a single forwarder PE. The procedures used for choosing a single forwarder are outlined in [RFC6514]. When the C-S route has been selected, the receiver PE will originate a Type 7 route, carrying Route Import attribute extracting from the C-S route, and sends this Type 7 route to other PEs.

When sender PE receives a Type 7 route, if RT-Import of this route belongs to itself, it translates this Type 7 route back into a C-join message and sends it to its CE.

8.2.3.1.2. Shared Tree Join (C-*, C-G)

When receiver PE receives a shared tree join (C-*, C-G) from CE, it does a route look up for C-RP. If there is more than one route, the receiver PE chooses a single forwarder PE. The procedures used for choosing a single forwarder are outlined in [RFC6514].

When the C-RP route has been selected, the receiver PE will create a Type 6 route. If this PE has not received a Type 5 route, it will not advertise it.

When source connected to CE is active, register message is sent to the sender PE. The sender PE originates a Type 5 route, and sends to other MCAST-6PE routers.

When receiver PE receives the Type 5 route from the remote PE, it will originate a Type 7 route based on Type 5 and Type 6, then it sends the Type 7 route carrying Route Import attribute extracting from the C-RP route, and sends this Type 7 route to other PEs.

When sender PE receives the Type 7 routes, compares local RT-Import to RT received with Type 7 routes. If match, it imports the Type 7 routes, then translates the Type 7 route back into a C-join message and passes the C-join messages to CE.

8.2.3.2. mLDP as the C-Multicast Protocol

The construction of the MCAST-6PE NLRI of C-multicast routes for the case where the C-multicast protocol is mLDP [mLDP] is described in [RFC6514].

8.2.4. Propagating IPv6 C-Multicast Routes by an ASBR

The mechanisms for IPv6 C-Multicast Routes by an ASBR are the same as the MVPN case described in section 11.2 of [RFC6514].

8.3. Using 6PE S-PMSI A-D Routes to Bind C-Trees to P-Tunnels

BGP-based procedures for using 6PE S-PMSIs A-D routes to bind (C-S,C-G) trees to P-tunnels are the same as the MVPN case described in section 12 of [RFC6514].

9. IANA Considerations

This document defines a new BGP Extended Community called "Route Import" (Type value is to be assigned by IANA). This Community is IP address specific, of an extended type, and is transitive.

This document defines a new NLRI, called as MCAST-6PE NLRI, to be carried in BGP using multiprotocol extensions. It requires assignment of a new SAFI. This is to be assigned by IANA.

10. Security Considerations

This document raises no new security issues. Security considerations for the base protocol are covered in [RFC6513] and [RFC6514].

11. References

11.1. Normative References

- [RFC2119] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, March 1997.
- [RFC4271] Rekhter, Y., Li, T., and S. Hares, "A Border Gateway Protocol 4 (BGP-4)", RFC 4271, January 2006.
- [RFC4360] Sangli, S., Tappan, D., and Y. Rekhter, "BGP Extended Communities Attribute", RFC 4360, February 2006.
- [RFC4364] Rosen, E. and Y. Rekhter, "BGP/MPLS IP Virtual Private Networks (VPNs)", RFC 4364, February 2006.
- [RFC4760] Bates, T., Chandra, R., Katz, D., and Y. Rekhter, "Multiprotocol Extensions for BGP-4", RFC 4760, January 2007.

- [RFC4798] De Clercq, J., Ooms, D., Prevost, S., and F. Le Faucheur, "Connecting IPv6 Islands over IPv4 MPLS Using IPv6 Provider Edge Routers (6PE)", RFC 4798, February 2007.
- [RFC6513] Rosen, E. and R. Aggarwal, "Multicast in MPLS/BGP IP VPNs", RFC 6513, February 2012.
- [RFC6514] Aggarwal, R., Rosen, E., Morin, T., and Y. Rekhter, "BGP Encodings and Procedures for Multicast in MPLS/BGP IP VPNs", RFC 6514, February 2012.

11.2. Informative References

- [RFC3031] Rosen, E., Viswanathan, A., and R. Callon, "Multiprotocol Label Switching Architecture", RFC 3031, January 2001.
- [RFC4272] Murphy, S., "BGP Security Vulnerabilities Analysis", RFC 4272, January 2006.
- [RFC4610] Farinacci, D. and Y. Cai, "Anycast-RP Using Protocol Independent Multicast (PIM)", RFC 4610, August 2006.
- [RFC5331] Aggarwal, R., Rekhter, Y., and E. Rosen, "MPLS Upstream Label Assignment and Context-Specific Label Space", RFC 5331, August 2008.
- [RFC6388] Wijnands, IJ., Minei, I., Kompella, K., and B. Thomas, "Label Distribution Protocol Extensions for Point-to-Multipoint and Multipoint-to-Multipoint Label Switched Paths", RFC 6388, November 2011.

Authors' Addresses

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

Email: lizhenbin@huawei.com

Shunwan Zhuang
Huawei Technologies
Huawei Bld., No.156 Beiqing Rd.
Beijing 100095
China

Email: zhuangshunwan@huawei.com

Hui Ni
Huawei Technologies
Huawei Bld., No.156 Beiqing Rd.
Beijing 100095
China

Email: nihui@huawei.com

Network Working Group
Internet-Draft
Intended status: Standards Track
Expires: January 16, 2014

Z. Li
H. Ni
Huawei Technologies
July 15, 2013

Role-Based State Advertisement for Multicast in MPLS/BGP IP VPNs
draft-li-l3vpn-mvpn-role-state-ad-00

Abstract

The document defines a new type of Intra-AS I-PMSI A-D route to advertise the role and corresponding primary/backup state for Multicast in MPLS/BGP IP VPNs. The role-based state advertisement can help optimization of process in MPLS/BGP Multicast VPN to reduce unnecessary traffic replication and facilitate service provision.

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 16, 2014.

Copyright Notice

Copyright (c) 2013 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. Terminology	3
3. Motivations	3
3.1. Easing Provision of mLDP P2MP LSP	3
3.2. Reducing Unnecessary Traffic Replication	3
3.3. Local Protection of Egress Nodes	4
4. Protocol Extensions	4
4.1. Role-Based Intra-AS I-PMSI A-D route	4
5. Operations	5
5.1. Advertisement of Role-base State for Root Nodes	6
5.2. Advertisement of Role-base State for Leaf Nodes	6
6. IANA Considerations	7
7. Security Considerations	7
8. Normative References	7
Authors' Addresses	8

1. Introduction

[RFC6513] defines the protocols and procedures for multicast in the BGP/MPLS IP VPN (Virtual Private Network) and [RFC6514] describes the BGP encodings and procedures for exchanging the information elements required by Multicast in MPLS/BGP IP VPNs.

In MPLS/BGP Multicast VPN, there is close relation between the multicast service and the tunnel which bears the multicast service. The tunnel can be triggered to setup automatically after the auto-discovery of the leaf PEs. This can facilitate the provision of the tunnels for multicast. Or else, it will take much effort for the provision work which is troublesome and error-prone. Based on the thinking, it is desirable that more information can be advertised along with the auto-discovery route which can optimize the provision of MPLS/BGP Multicast VPN.

This document identifies the requirements to advertise the role and corresponding state for Multicast in MPLS/BGP IP VPNs and defines a new type of Intra-AS I-PMSI A-D route to achieve the object. The role-based state advertisement can help optimization of multicast process to reduce unnecessary traffic replication and facilitate service provision in MPLS/BGP Multicast VPN .

2. Terminology

CE: Customer Edge

PE: Provider Edge

MVPN: Multicast VPN

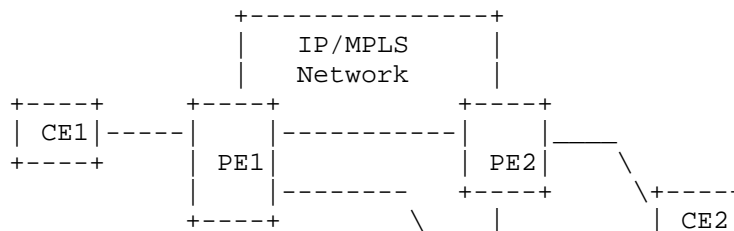
3. Motivations

3.1. Easing Provision of mLDP P2MP LSP

mLDP P2MP LSP can be used to bear multicast service in MPLS/BGP MVPN. It needs to send label mappings to the root to set up the P2MP LSP. So it has to specify the root node address on all leaf nodes for a specific P2MP LSP. In MPLS/BGP MVPN, if the root/leaf role information of the PE in the MVPN can be advertised in the auto-discovery route, the leaf PE can directly trigger mLDP to send label mapping to the ingress PE of the MVPN without explicitly specifying the root address. It can facilitate the provision of the MVPN when mLDP P2MP LSP is used.

3.2. Reducing Unnecessary Traffic Replication

There exist multi-homing scenarios in MPLS/BGP MVPN. As shown in the figure 1, CE2 multi-homes to two PEs (PE2 and PE3). We assume PE1 is the ingress PE and PE2/PE3 are the egress PEs for a specific MPLS/BGP MVPN. If C-join is always sent from the CE2 to the PE2 for the MVPN, the multicast traffic sent from the PE1 to PE3 will be always dropped since there is not (C-S, C-G) entry in the MVPN on PE3. If PE1 can learn that the remote PE would not forward the multicast traffic to any CE, the bandwidth can be saved in the network for PE1 can stop to setup the ingress replication tunnel or P2MP LSPs to the remote PE or stop to replicate the unnecessary traffic to the tunnels to the remote PE. In order to achieve the object, the primary or backup state for the leaf PE can be advertised. As to a PE in a specific MVPN, the primary state means that it needs to forward the multicast traffic to the CE. The backup state means it would not forward the multicast traffic to any CE.



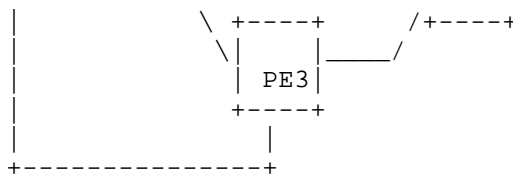


Figure 1 Multi-homing Network for Multicast in MPLS/BGP IP VPN

3.3. Local Protection of Egress Nodes

[I-D.chen-mpls-p2mp-ingress-protection] and [I-D.chen-mpls-p2mp-egress-protection] proposes mechanisms for locally protecting ingress and egress nodes of MPLS TE P2MP LSPs. In the mechanism for the local protection of egress nodes, the backup egress node needs to be designated for the primary egress node for a P2MP LSP. The previous hop node of the primary egress node sets up a backup Sub-LSP from itself to the backup egress node after receiving the information about the backup egress node. The provision of the local protection mechanism of egress nodes in P2MP LSPs can be facilitated in MPLS/BGP MVPN by advertising the primary/backup state and the protected egress node address with the auto-discovery route. When the ingress PE of the MVPN learns which egress PE can be used as the backup node to protect the primary egress node, it can directly trigger to set up the P2MP LSP with local protection of egress nodes. The method saves much provision effort since it need not statically designate the protection between the backup egress node and the primary egress node for a P2MP LSP.

4. Protocol Extensions

A new type of Intra-AS I-PMSI A-D route is defined for the MCAST-VPN NLRI. It is called as Role-based Intra-AS I-PMSI A-D route. Route Type for the A-D route is to be defined.

4.1. Role-Based Intra-AS I-PMSI A-D route

A Role-based Intra-AS I-PMSI A-D route type specific MCAST-VPN NLRI consists of the following:

```

+-----+
|      RD      (8 octets)      |
+-----+
|  Originating Router's IP Addr  |
+-----+
|R|RS|L|LS|      Reserved      |
+-----+
|Protected Root's IP Addr(Optional) |
+-----+

```

```
+-----+
|Protected Leaf's IP Addr(Optional) |
+-----+
```

The Route Distinguisher (RD) is encoded as described in [RFC4364].

Originating Router's IP Addr is to advertising the originating PE's IPv4/IPv6 address.

R field is one bit to identify if the PE is used as the root node in the MVPN.

RS field is two bits to identify the primary/backup state if the PE is used as the root node. There are three value for the RS field:

-- 0 means the PE is used as the primary root node.

-- 1 means the PE is used as the backup root node. But the protected root node's IP address does not exist.

-- 2 means the PE is used as the backup root node and there exists the protected root node's IP address.

L field is one bit to identify if the PE is used as the leaf node in the MVPN.

LS field is two bits to identify the primary/backup state if the PE is used as the leaf node. There are three value for the LS field:

-- 0 means the PE is used as the primary leaf node.

-- 1 means the PE is used as the backup leaf node. But the protected leaf node's IP address does not exist.

-- 2 means the PE is used as the backup leaf node and there exists the protected leaf node's IP address.

Protected Root's IP Addr is an optional field. It specifies the IPv4 /IPv6 address of the protected root node. The field exists only when the value of the RS field is 2.

Protected Leaf's IP Addr is an optional field. It specifies the IPv4 /IPv6 address of the protected leaf node. The field exists only when the value of the LS field is 2.

5. Operations

5.1. Advertisement of Role-base State for Root Nodes

The Role-Based Intra-AS I-PMSI A-D route can be used to advertise the role and corresponding primary/backup state for root nodes in MPLS/BGP MVPN.

If a PE is specified as the root PE for a specific MVPN, it MUST set the R bit as 1 in the A-D route. Otherwise, the R bit MUST NOT be set.

If the root PE is used as the backup ingress node to protect the primary root PE, the RS field in the A-D route MUST set as 1 when there is no determined root node to be protected. In this case the primary root node protected by the backup root node can be calculated by all nodes according to some uniform algorithms which is out of the scope of this document. When the RS field in the A-D route is set as 1, the Protected Root's IP Addr field MUST NOT exist in the A-D route.

If the root PE is used as the backup ingress node to protect the primary root PE, the RS field in the A-D route MUST set as 2 when there is a determined root node to be protected. In this case the Protected Root's IP Addr field MUST exist in the A-D route which will specify the IPv4/IPv6 address of the protected root node.

If the R bit is set as 1 and the RS field is set as 0 in the A-D route, the Protected Root's IP Addr field MUST NOT exist in the A-D route. This means the PE is used as the primary root PE.

If the R bit in the A-D route is set as 0, the RS field MUST be ignored and the Protected Root's IP Addr field MUST NOT exist in the A-D route.

5.2. Advertisement of Role-base State for Leaf Nodes

The Role-Based Intra-AS I-PMSI A-D route can be used to advertise the role and corresponding primary/backup state for leaf nodes in MPLS/BGP MVPN.

If a PE is specified as the leaf PE for a specific MVPN, it MUST set the L bit as 1 in the A-D route. Otherwise, the L bit MUST NOT be set.

If the leaf PE is used as the backup egress node to protect the primary leaf PE, the LS field in the A-D route MUST set as 1 when there is no determined leaf node to be protected. In this case the primary leaf node protected by the backup leaf node can be calculated by all nodes according to some uniform algorithms which is out of the

scope of this document. When the LS field in the A-D route is set as 1, the Protected Leaf's IP Addr field MUST NOT exist in the A-D route.

If the leaf PE is used as the backup egress node to protect the primary leaf PE, the LS field in the A-D route MUST set as 2 when there is a determined leaf node to be protected. In this case the Protected Leaf's IP Addr field MUST exist in the A-D route which will specify the IPv4/IPv6 address of the protected leaf node.

If the L bit is set as 1 and the LS field is set as 0 in the A-D route, the Protected Leaf's IP Addr field MUST NOT exist in the A-D route. This means the PE is used as the primary leaf PE.

If the L bit in the A-D route is set as 0, the LS field MUST be ignored and the Protected Leaf's IP Addr field MUST NOT exist in the A-D route.

6. IANA Considerations

This document defines a new type of A-D route for MCAST-VPN NLRI. The type is to be assigned by IANA.

7. Security Considerations

There are no additional security aspects beyond those specified in [RFC6513] and [RFC6514].

8. Normative References

- [I-D.chen-mppls-p2mp-egress-protection]
Chen, H., Ning, S., Liu, A., Xu, F., Toy, M., and L. Liu,
"Extensions to RSVP-TE for P2MP LSP Egress Local
Protection", draft-chen-mppls-p2mp-egress-protection-09
(work in progress), May 2013.
- [I-D.chen-mppls-p2mp-ingress-protection]
Chen, H., Ning, S., Liu, A., and L. Liu, "Extensions to
RSVP-TE for P2MP LSP Ingress Local Protection", draft-
chen-mppls-p2mp-ingress-protection-08 (work in progress),
February 2013.
- [RFC2119] Bradner, S., "Key words for use in RFCs to Indicate
Requirement Levels", BCP 14, RFC 2119, March 1997.
- [RFC6513] Rosen, E. and R. Aggarwal, "Multicast in MPLS/BGP IP
VPNs", RFC 6513, February 2012.

[RFC6514] Aggarwal, R., Rosen, E., Morin, T., and Y. Rekhter, "BGP Encodings and Procedures for Multicast in MPLS/BGP IP VPNs", RFC 6514, February 2012.

Authors' Addresses

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

Email: lizhenbin@huawei.com

Hui Ni
Huawei Technologies
Huawei Bld., No.156 Beiqing Rd.
Beijing 100095
China

Email: nihui@huawei.com

Network Working Group
Internet-Draft
Intended status: Informational
Expires: October 19, 2013

Z. Li
Z. Zhuang
J. Dong
Huawei Technologies
April 17, 2013

A Framework for Service-Driven Co-Routed MPLS Traffic Engineering LSPs
draft-li-mpls-serv-driven-co-lsp-fmwk-01

Abstract

MPLS TE has been widely deployed to provide traffic engineering and traffic protection. The complexity in configuration has much effect on the MPLS TE deployment in the large-scale network. The document identifies the configuration issues for MPLS TE deployment and proposes a new mechanism, the service-driven mechanism, by which the setup of co-routed MPLS TE LSPs is triggered by the bidirectional service. Then the document provides the framework for setting up service-driven co-routed MPLS Traffic-Engineered Label-Switched Paths (TE LSPs) for L2VPN and L3VPN.

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 October 19, 2013.

Copyright Notice

Copyright (c) 2013 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. Terminology	3
3. Problem Statement	4
3.1. Massive Configuration Issue of TE LSPs	4
3.2. Return Path Issue of BFD for MPLS LSPs	5
3.3. Upgrading Issue of Co-routed Bidirectional LSP	6
4. Framework and Procedures	6
4.1. Service-Driven Co-Routed Unidirectional LSPs for L2VPN	7
4.1.1. Framework	7
4.1.2. Procedures	7
4.2. Service-Driven Co-Routed Unidirectional LSPs for L3VPN	9
4.2.1. Framework	9
4.2.2. Procedures	10
5. IANA Considerations	12
6. Security Considerations	12
7. References	12
7.1. Normative References	12
7.2. Informative References	13
Authors' Addresses	15

1. Introduction

Multiprotocol Label Switching (MPLS) traffic engineering (TE) effectively schedules, allocates, and uses existing network resources to provide bandwidth guarantee and traffic protection. MPLS TE establishes label switched paths (LSPs) satisfying specific traffic engineering attributes [RFC2702]. MPLS TE is being widely deployed to support packet-based services. Since rich set of traffic engineering attributes have to be specified for each LSP and a great deal of configuration has to be done as the number of MPLS TE LSPs increases, a scalable and simple solution is required to implement TE on a large-scale network and reduce the complexity in operation and management of TE LSPs.

LDP LSP setup is topology-driven which is a scalable way to adapt to the large-scale network. The similar way cannot be used for MPLS TE since the traffic engineering attributes should be specified for the MPLS TE tunnel which is not necessary for LDP LSP. On the other hand, MPLS TE LSP is always setup to bear specific services such as L3VPN and L2VPN. That is, MPLS TE LSPs will not be setup aimlessly which is always inevitable for MPLS topology-driven LSP if there is no policy exerted on it. So it seems a natural way to combine the MPLS TE LSP setup with the service it beared. The MPLS TE LSP setup can be triggered automatically by the service instead of explicitly configuring each tunnel and its traffic engineering attributes. We call this method as service-driven comparing to topology-driven. In addition the service-driven method has much advantage in the process of setting up co-routed TE LSPs. The service beared by MPLS TE LSPs is always bi-directional. The characteristic can be utilized to setup the forward MPLS TE LSP and the co-routed reverse MPLS TE LSP.

This document describes the framework of automatically setting up co-routed TE LSPs, in which the co-routed MPLS TE LSPs are automatically setting up on demand of the services, e.g. VPNs. This mechanism facilitates the provisioning of services and the TE LSPs.

2. Terminology

This document uses terminology from the MPLS architecture document [RFC3031], the RSVP-TE protocol specification [RFC3209] which inherits from the RSVP specification [RFC2205] and the Provider Provisioned VPN terminology document [RFC4026].

The document introduces two new concepts by which VPN PEs can be generally categorized into two types:

1. Active PE: the PE which primarily triggers the set up of the LSPs and informs the remote PE;
2. Passive PE: the PE which secondarily complies with the active PE's suggestion to set up LSPs.

In this document, the terminology of "tunnel" is identical to the "TE Tunnel" defined in Section 2.1 of [RFC3209], which is uniquely identified by a SESSION object that includes Tunnel end point address, Tunnel ID and Extended Tunnel ID. The terminology "LSP" is identical to the "LSP tunnel" defined in Section 2.1 of [RFC3209], which is uniquely identified by the SESSION object together with SENDER_TEMPLATE (or FILTER_SPEC) object that consists of LSP ID and Tunnel end point address.

3. Problem Statement

3.1. Massive Configuration Issue of TE LSPs

It is a common deployment scenario to set up MPLS TE LSPs among a set of Label Switching Routers (LSR). Such deployment may require the configuration of a potentially large number of TE tunnels.

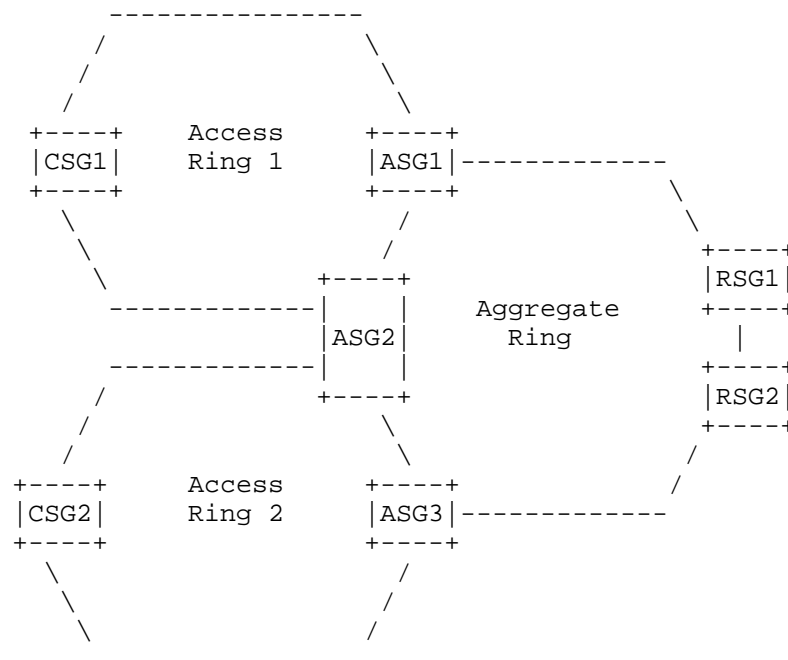


Figure 1 Mobile Backhaul Network

Figure 1 shows an example of the mobile backhaul network. Mobile multimedia devices such as smartphones are ubiquitous now which runs a wide variety of bandwidth-intensive applications and causes unprecedented growth in mobile data traffic. In order to cope with the growth, more cell sites are introduced into the network: more LTE eNodeBs and associated Cell Site Gateways(CSGs) are added in the networks. This causes the network scale expands fast and more and more MPLS TE tunnels need setup between Cell Site Gateways(CSGs) connects the eNodeBs and RNC site gateways(RSGs) connects the RNCs. Typically, we assume that:

1. There are 1,000 CSGs need to connect to one RSG.

2. There are three types of bi-directional services beared between one CSG and one RSG. Each type of service needs one VPN and one TE tunnel.

3. There are 10 command lines to configure necessary attributes for each TE tunnel and at least one command line to bind one VPN to one TE tunnel.

Then there are at least 66,000 command lines to configure MPLS TE tunnels and VPN and TE tunnel bindings. It is truly a huge configuration work. The operation is not only time consuming but also prone to mis-configuration for Service Providers. Hence, a mechanism to set up MPLS TE tunnels automatically is desirable which can significantly reduce the complexity of MPLS TE configuration.

3.2. Return Path Issue of BFD for MPLS LSPs

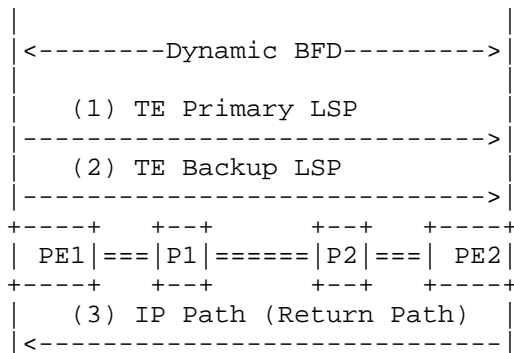


Figure 2: BFD for TE LSPs Scenario

As shown in Figure 2, BFD for MPLS LSPs ([RFC5884]) is used to detect the possible failure fast which can trigger traffic switch between the primary LSP and the backup LSP. When BFD for MPLS LSPs is deployed, the return path may take IP path which is different from the forward path. The failure that happens in the return path may cause wrong traffic switch.

In order to solve the return path issue of BFD for MPLS LSPs, it has to be guaranteed that the forward path and the return path must be co-routed. For MPLS TE LSPs explicit path has to be configured for the forward LSP and the return LSP. In addition, there is at least one configuration to bind the return BFD traffic corresponding to the forward BFD traffic to the right return MPLS TE tunnel at the ingress node or the egress node. This will deteriorate the configuration work described above. In addition, if the forward path changes, the

return path may not change accordingly owing to statically binding the forward path and the return path. It will cause that the return path issue of BFD for MPLS LSPs happens again.

3.3. Upgrading Issue of Co-routed Bidirectional LSP

The co-routed bidirectional LSP is defined in [RFC3945]. If co-routed bidirectional LSP is used, the return path is not necessary to configure and the return path issue of BFD for LSPs can be solved naturally. This can simplify operation and management for Service Providers to some extent. But as to the example described in sec 3.2 it is still necessary to configure each tunnel and configure explicit binding of each tunnel and VPN pair. The configuration work is still a little complex. In addition, the unidirectional LSPs have been deployed widely and it is difficult for the service providers to upgrade all possible routers to support co-routed bidirectional LSPs.

4. Framework and Procedures

MPLS TE LSPs depend heavily on manual configuration. So some auto configuration method (e.g. auto mesh [RFC4972]) has been proposed. This document proposes a new mechanism, the service-driven mechanism, to reduce the operation cost of MPLS TE networks.

It is well known that LDP LSP setup is topology-driven which is a scalable way to adapt to the large-scale network. The similar way cannot be used for MPLS TE since the traffic engineering attributes has to be specified for the MPLS TE tunnel. On the other hand, MPLS TE LSP is always setup to bear specific services such as L3VPN and L2VPN. That is, MPLS TE LSPs will not be setup aimlessly which is always inevitable for MPLS topology-driven LSP if there is no policy exerted on it. So it is a natural way to trigger MPLS TE LSP setup by the service instead of explicitly configuring each tunnel. We call this method as service-driven comparing to topology-driven. BGP-based MVPN ([RFC6513] and [RFC6514]) provides an example of service-driven tunnel which can trigger P2MP TE tunnel setup after MVPN membership auto-discovery.

The service-driven method also has much advantage in the process of setting up co-routed TE LSPs. The service beared by MPLS TE LSPs is always bi-directional. The characteristic can be utilized to setup the forward MPLS TE LSP and the co-routed reverse MPLS TE LSP. This section describes the framework and procedures of setting up the co-routed MPLS TE LSPs. In this method, MPLS TE LSPs can be set up on demand which can reduce the manual configuration. The signaling of the service will advertise the tunnel information between the active PE and the passive PE. The PE on the passive side can set up the reverse LSP based on RRO information of the LSP from the active PE to

the passive PE. Thus the path of the reverse LSP can be co-routed with the path of the LSP from the active PE to the passive PE.

Service-driven co-routed MPLS TE LSP has following advantages:

- 1) Setup LSP on demand and save massive configuration effort.
- 2) Reuse current mechanism instead of whole network upgrading.

4.1. Service-Driven Co-Routed Unidirectional LSPs for L2VPN

4.1.1. Framework

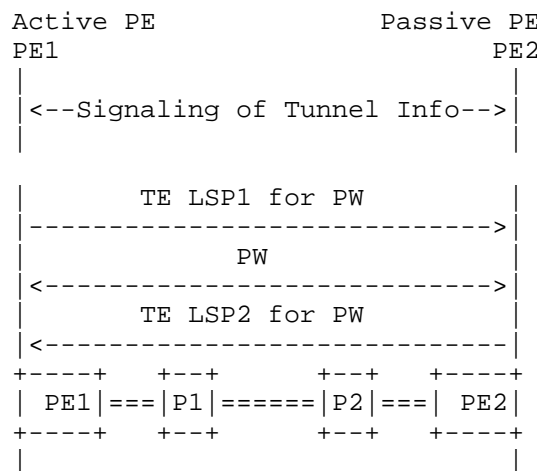


Figure 3: Framework of L2VPN driven TE LSP

L2VPN, as defined in [RFC4664], is a proven and widely deployed technology. Figure 3 shows a framework for co-routed MPLS TE LSPs driven by L2VPN service. L2VPN is provisioned on PEs and the PW is setup. A pair of PEs for a specific PW will be identified as the active PE and the passive PE respectively. The active PE triggers the set up of the primary LSP to the passive PE and advertises the tunnel information to the passive PE. According to the information advertised by the active PE, the passive PE will set up the reverse MPLS TE LSP which is co-routed with the LSP from the active PE to the passive PE LSP.

4.1.2. Procedures



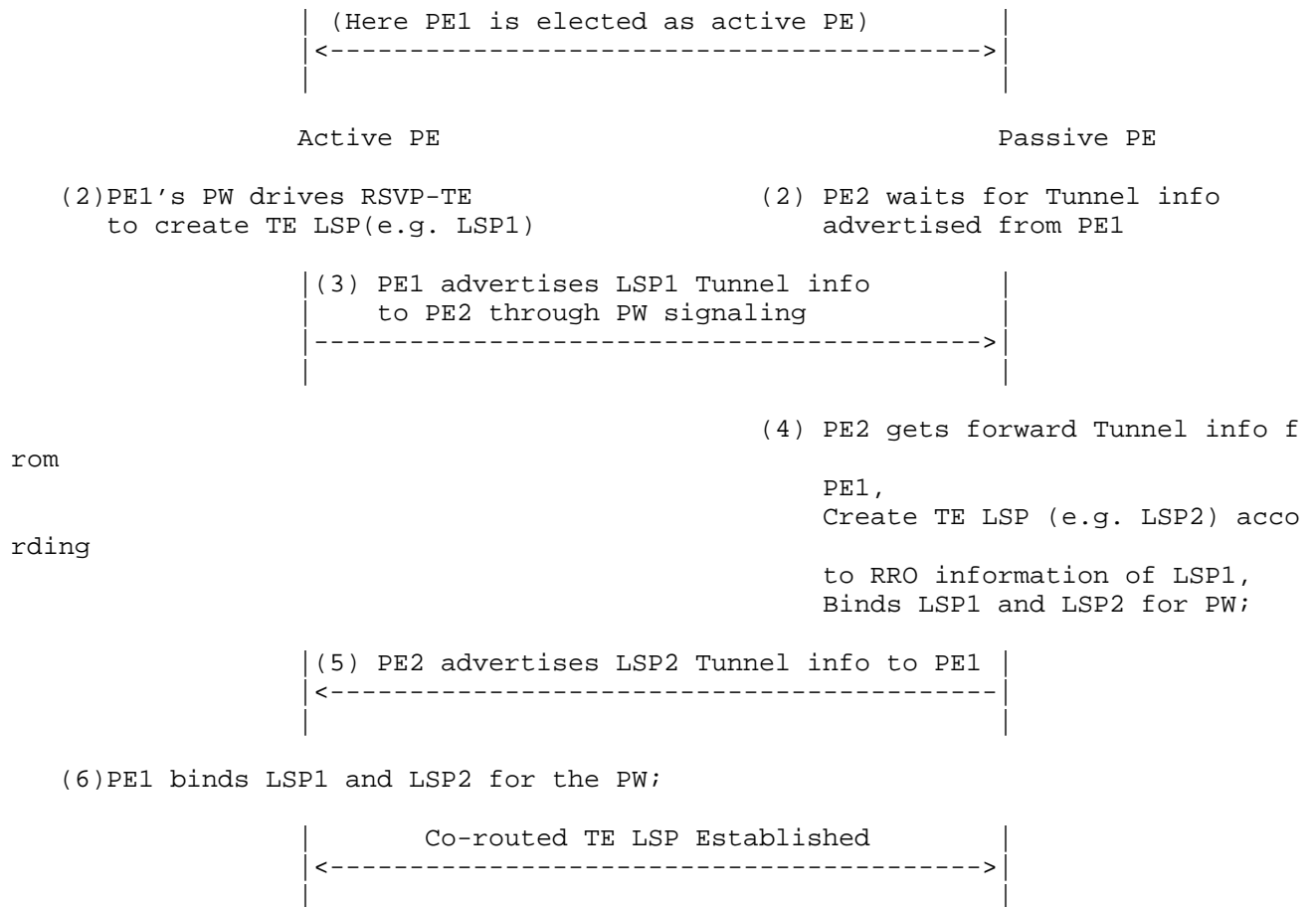


Figure 4: Signaling Procedures of L2VPN driven co-routed TE LSPs

Figure 4 shows the detailed procedures for L2VPN driven co-routed MPLS TE LSPs. Through the above procedure, the co-routed MPLS TE LSPs driven by the PW are established.

4.1.2.1. Active/Passive Role Election

The active and passive role of PEs can be determined through manual configuration or dynamic election between two PEs. If the dynamic election method is used, LSR IDs of a pair of PEs between which PW is setup are compared as unsigned integers and the PE with the larger value of LSR ID assumes the active role.

4.1.2.2. Signaling Tunnel Information

In the service-driven co-routed MPLS TE framework for L2VPN, the tunnel information need to be advertised between the active PE and the passive PE. The passive PE uses the tunnel information to get corresponding MPLS TE tunnel and RRO information which is used to setup the reverse co-routed MPLS TE LSP. [I-D.ietf-pwe3-mpls-tp-pw-over-bidir-lsp] defines how the bidirectional Tunnel/LSP identifier information is advertised between a pair of PEs for PW. The similar mechanism can be reused for advertising MPLS TE tunnel/LSP identifier information for service-driven MPLS TE LSPs for L2VPN.

4.1.2.3. Operation

Step 1: Active/passive role election through signaling between PEs of a PW. In this case, assume PE1 as active PE and PE2 as passive PE after election;

Step 2: As the active role, the PW service on PE1 drives RSVP-TE to create TE LSP(e.g. LSP1), as the passive role, PE2 waits for tunnel information advertised by PE1;

Step 3: PE1 advertises tunnel information related with LSP1 to PE2;

Step 4: PE2 gets tunnel information from PE1 and creates TE LSP (e.g. LSP2) according to RRO information derived from LSP1. PE2 binds LSP1 and LSP2 for PW;

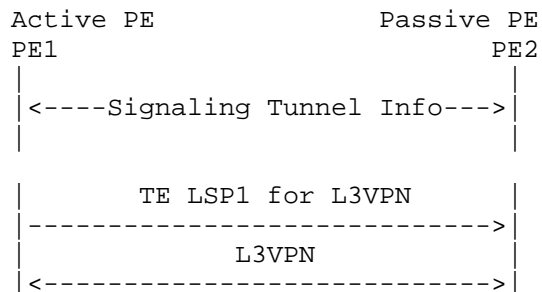
Step 5: PE2 advertises tunnel information related with LSP2 to PE1;

Step 6: PE1 binds LSP1 and LSP2 for PW.

Through the above procedure, the co-routed MPLS TE LSPs driven by the PW are established.

4.2. Service-Driven Co-Routed Unidirectional LSPs for L3VPN

4.2.1. Framework



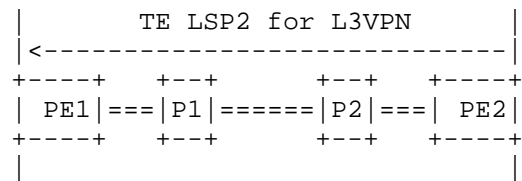
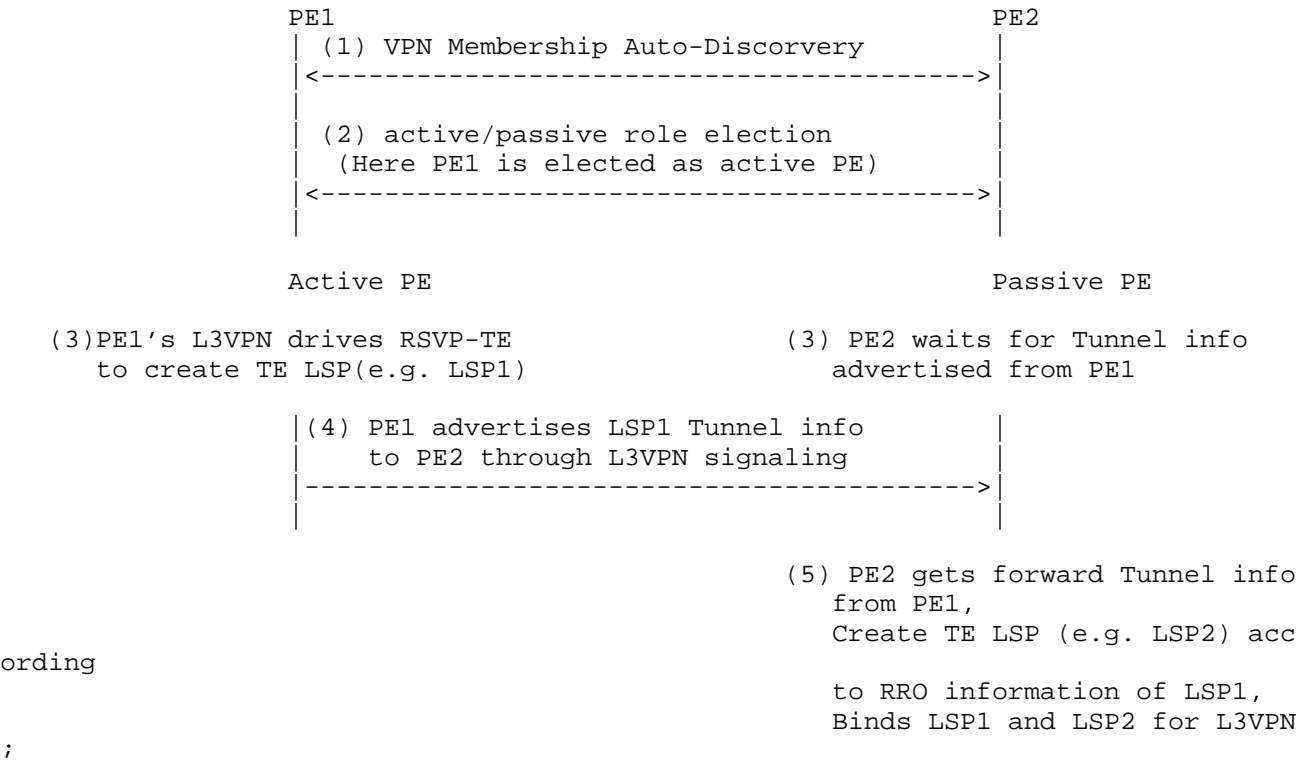


Figure 5: Framework of L3VPN driven TE Tunnel

L3VPN services are provided by [RFC4110]. Figure 5 shows a framework for co-routed MPLS TE LSPs driven by L3VPN. L3VPN is provisioned on PEs and VPN membership is discovered.

A pair of PEs for a specific L3VPN are identified as the active PE and the passive PE respectively. The active PE initiates the set up of the primary LSP to the passive PE and advertises the tunnel information to the passive PE. According to the information advertised by the active PE, the passive PE will set up the reverse MPLS TE LSP which is co-routed with the forward LSP from the active PE to the passive PE.

4.2.2. Procedures



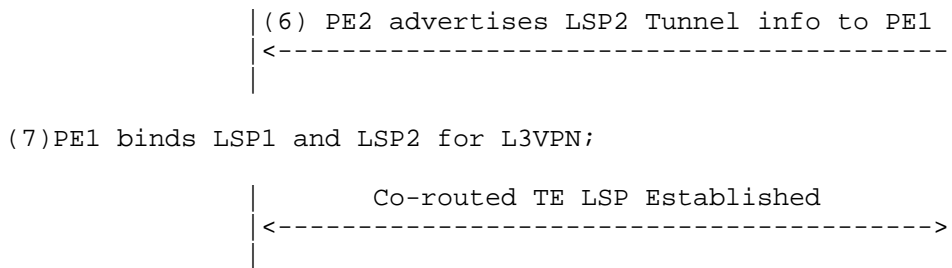


Figure 6: Signaling Procedures of L3VPN driven co-routed TE LSPs

Figure 6 shows the detailed procedures for L3VPN to drive the set up of co-routed MPLS TE LSPs. Through the above procedure, the co-routed MPLS TE LSPs driven by the L3VPN are established.

4.2.2.1. VPN Membership Auto-Discovery

In order to set up co-routed MPLS TE LSPs in L3VPN, a point-to-point connection between any two VRFs of a particular VPN needs to be established. VPN membership auto-discovery should be done firstly and the mechanism defined in [I-D.dong-l3vpn-pm-framework] can be used.

4.2.2.2. Active/Passive Role Election

After obtaining the VPN membership information via VPN membership auto-discovery process, we can identify a pair of VPN members.

The active and passive role of PEs can be determined through manual configuration or dynamic election between two PEs. If the dynamic election method is used, LSR IDs of a pair of PEs corresponding to the pair of VPN members are compared as unsigned integers and the PE with the larger value of LSR ID assumes the active role.

4.2.2.3. Signaling Tunnel Information

In the service-driven co-routed MPLS TE framework for L3VPN, the tunnel information need to be advertised between the active PE and the passive PE. The passive PE uses the tunnel information to get corresponding MPLS TE tunnel and RRO information which is used to setup the reverse co-routed MPLS TE LSP. MP-BGP signaling need extensions to advertise the MPLS TE tunnel/LSP identifier information for service-driven MPLS TE LSPs for L3VPN.

4.2.2.4. Operation

Step 1: VPN membership auto-discovery process is done through signaling to identify a pair of VPN members;

Step 2: Active/passive role election through signaling between a pair of PEs of a L3VPN. In this case, assume PE1 as active PE and PE2 as passive PE after election;

Step 3: As the active role, L3VPN service on PE1 drives RSVP-TE to create TE LSP(e.g. LSP1), as the passive role, PE2 waits for tunnel information advertised by PE1;

Step 4: PE1 advertises tunnel information related with LSP1 to PE2;

Step 5: PE2 gets tunnel information from PE1 and creates TE LSP (e.g. LSP2) according to RRO information derived from LSP1. PE2 binds LSP1 and LSP2 for L3VPN;

Step 6: PE2 advertises tunnel information related with LSP2 to PE1;

Step 7: PE1 binds LSP1 and LSP2 for L3VPN.

Through the above procedure, the co-routed MPLS TE LSPs driven by the L3VPN are established.

5. IANA Considerations

This document makes no request of IANA.

6. Security Considerations

This document does not change the security properties of L2VPN & L3VPN.

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.
- [RFC3032] Rosen, E., Tappan, D., Fedorkow, G., Rekhter, Y., Farinacci, D., Li, T., and A. Conta, "MPLS Label Stack Encoding", RFC 3032, January 2001.
- [RFC3209] Awduche, D., Berger, L., Gan, D., Li, T., Srinivasan, V., and G. Swallow, "RSVP-TE: Extensions to RSVP for LSP Tunnels", RFC 3209, December 2001.

- [RFC4447] Martini, L., Rosen, E., El-Aawar, N., Smith, T., and G. Heron, "Pseudowire Setup and Maintenance Using the Label Distribution Protocol (LDP)", RFC 4447, April 2006.
- [RFC6370] Bocci, M., Swallow, G., and E. Gray, "MPLS Transport Profile (MPLS-TP) Identifiers", RFC 6370, September 2011.

7.2. Informative References

- [I-D.dong-l3vpn-pm-framework]
Dong, J. and Z. Li, "A Framework for L3VPN Performance Monitoring", draft-dong-l3vpn-pm-framework-00 (work in progress), October 2012.
- [I-D.ietf-mpls-return-path-specified-lsp-ping]
Chen, M., Cao, W., Ning, S., JOUNAY, F., and S. DeLord, "Return Path Specified LSP Ping", draft-ietf-mpls-return-path-specified-lsp-ping-11 (work in progress), October 2012.
- [I-D.ietf-pwe3-mpls-tp-pw-over-bidir-lsp]
Chen, M., Cao, W., Takacs, A., and P. Pan, "LDP extensions for Pseudowire Binding to LSP Tunnels", draft-ietf-pwe3-mpls-tp-pw-over-bidir-lsp-00 (work in progress), December 2012.
- [I-D.zheng-l3vpn-pm-analysis]
Zheng, L. and Z. Li, "Performance Monitoring Analysis for L3VPN", draft-zheng-l3vpn-pm-analysis-00 (work in progress), October 2012.
- [RFC2205] Braden, B., Zhang, L., Berson, S., Herzog, S., and S. Jamin, "Resource ReSerVation Protocol (RSVP) -- Version 1 Functional Specification", RFC 2205, September 1997.
- [RFC2702] Awduche, D., Malcolm, J., Agogbua, J., O'Dell, M., and J. McManus, "Requirements for Traffic Engineering Over MPLS", RFC 2702, September 1999.
- [RFC3031] Rosen, E., Viswanathan, A., and R. Callon, "Multiprotocol Label Switching Architecture", RFC 3031, January 2001.
- [RFC3471] Berger, L., "Generalized Multi-Protocol Label Switching (GMPLS) Signaling Functional Description", RFC 3471, January 2003.

- [RFC3473] Berger, L., "Generalized Multi-Protocol Label Switching (GMPLS) Signaling Resource ReserVation Protocol-Traffic Engineering (RSVP-TE) Extensions", RFC 3473, January 2003.
- [RFC3945] Mannie, E., "Generalized Multi-Protocol Label Switching (GMPLS) Architecture", RFC 3945, October 2004.
- [RFC4026] Andersson, L. and T. Madsen, "Provider Provisioned Virtual Private Network (VPN) Terminology", RFC 4026, March 2005.
- [RFC4110] Callon, R. and M. Suzuki, "A Framework for Layer 3 Provider-Provisioned Virtual Private Networks (PPVPNs)", RFC 4110, July 2005.
- [RFC4271] Rekhter, Y., Li, T., and S. Hares, "A Border Gateway Protocol 4 (BGP-4)", RFC 4271, January 2006.
- [RFC4664] Andersson, L. and E. Rosen, "Framework for Layer 2 Virtual Private Networks (L2VPNs)", RFC 4664, September 2006.
- [RFC4972] Vasseur, JP., Leroux, JL., Yasukawa, S., Previdi, S., Psenak, P., and P. Mabbey, "Routing Extensions for Discovery of Multiprotocol (MPLS) Label Switch Router (LSR) Traffic Engineering (TE) Mesh Membership", RFC 4972, July 2007.
- [RFC5036] Andersson, L., Minei, I., and B. Thomas, "LDP Specification", RFC 5036, October 2007.
- [RFC5880] Katz, D. and D. Ward, "Bidirectional Forwarding Detection (BFD)", RFC 5880, June 2010.
- [RFC5884] Aggarwal, R., Kompella, K., Nadeau, T., and G. Swallow, "Bidirectional Forwarding Detection (BFD) for MPLS Label Switched Paths (LSPs)", RFC 5884, June 2010.
- [RFC6513] Rosen, E. and R. Aggarwal, "Multicast in MPLS/BGP IP VPNs", RFC 6513, February 2012.
- [RFC6514] Aggarwal, R., Rosen, E., Morin, T., and Y. Rekhter, "BGP Encodings and Procedures for Multicast in MPLS/BGP IP VPNs", RFC 6514, February 2012.

Authors' Addresses

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

Email: lizhenbin@huawei.com

Shunwan Zhuang
Huawei Technologies
Huawei Bld., No.156 Beiqing Rd.
Beijing 100095
China

Email: zhuangshunwan@huawei.com

Jie Dong
Huawei Technologies
Huawei Bld., No.156 Beiqing Rd.
Beijing 100095
China

Email: jie.dong@huawei.com

Network Working Group
Internet-Draft
Expires: January 16, 2014

T. Morin, Ed.
S. Litkowski
Orange
K. Patel
Cisco Systems
J. Zhang
R. Kebler
Juniper Networks
July 15, 2013

Multicast state damping
draft-morin-multicast-damping-00

Abstract

This document describes procedures to damp multicast routing state changes and prevent the churn due to the multicast dynamicity at the edge of a network. The procedures described in this document help avoid uncontrolled control plane load increase on the core routing infrastructure. New procedures are proposed inspired from BGP unicast route damping principles, but adapted to multicast. They cover multicast and multicast in VPNs contexts.

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 16, 2014.

Copyright Notice

Copyright (c) 2013 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	3
2. Terminology	3
3. Overview	3
4. Existing mechanisms	4
4.1. Rate-limiting of multicast control traffic	4
4.2. Existing PIM, IGMP and MLD timers	4
4.3. BGP Route Damping	5
5. Procedures for multicast state damping	6
6. Procedures for multicast in IP VPNs	8
6.1. Damping P-tunnel change events	9
7. Procedures for Ethernet VPNs	10
8. Operational considerations	10
8.1. Enabling and configuring multicast damping	10
8.2. Troubleshooting and monitoring	11
8.3. Maximum values for exponential decay and thresholds parameters	11
8.4. Default values	11
9. IANA Considerations	11
10. Security Considerations	11
11. Acknowledgements	12
12. References	12
12.1. Normative References	12
12.2. Informative References	13
Authors' Addresses	13

1. Introduction

When multicast receivers join and leave a said multicast group or channel at the edge of a network through multicast membership control protocols (IGMP, MLD), multicast routing protocols (e.g. PIM-SM, or mVPN) adjust multicast routing states accordingly to forward or prune multicast traffic to these receivers.

Mechanisms need to be put in place to ensure that the load put on the control plane of core routers remains under control regardless of the frequency at which multicast memberships changes are made by end hosts. By nature multicast memberships change based on the behavior of multicast applications running on end hosts, hence the frequency of membership changes can legitimately be much higher than the typical churn of unicast routing states.

This document describes procedures aimed at protecting the control plane of the core network infrastructure (more specifically edge routers, core routers and in the case of multicast in VPN contexts BGP Route Reflectors) while at the same time avoiding negative effects on the service provided, although at the expense of a minimal increase in average of bandwidth use in the network.

The base principle is described in Section 3. Existing mechanisms that could be relied upon are discussed in Section 4. Section 5 details the proposed procedures.

Sections 6 and 7 provide more specific details related to multicast in VPNs contexts.

Finally, Section 8 discusses operational considerations related to the proposed mechanism.

2. Terminology

TBC

3. Overview

The procedures described in this document allows the network operator to configure multicast routers so that they can delay the propagation of multicast state prune messages, when faced with a rate of multicast state dynamicity exceeding a certain configurable threshold. Assuming that the number of multicast states that can be created by a receiver is bounded, delaying the propagation of multicast state pruning results in setting up an upper bound to the

average frequency at which the router will send state updates to an upstream router.

From the point of view of a downstream router, this approach has no impact: the multicast routing states changes that it solicits to its upstream router will be honored without any additional delay. Indeed the propagation of joins is not impacted by the proposed defined procedures, and having the upstream router delay state prune propagation to its own upstream does not affect what traffic is sent to the downstream router. In particular, the amount of bandwidth used on the link downstream to a router applying this damping technique is not increased.

This approach increases the average bandwidth utilization on a link upstream to a router applying this technique: indeed, the bandwidth of a said multicast flow will be used for a longer time than if no damping was applied. That said, it is expected that this technique will allow to meet the goals of protecting the multicast routing infrastructure control plane without a significant average increase of bandwidth; for instance, damping events happening at a frequency higher than one event per X second, can be done without increasing the time during which a multicast flow is present on a link of more than X second.

To be practical, such a mechanism requires configurability, in particular, needs to offer means to control when damping is triggered and allow delaying Pruning for a longer period of time the more activity there is on a multicast state.

Note that the issues related to control plane load due to the dynamicity of multicast sources coming and going in the context of ASM multicast, are out of the scope of this document.

4. Existing mechanisms

4.1. Rate-limiting of multicast control traffic

[RFC4609] examines multicast security threats and among other things the risk described in Section 1. A mechanism relying on rate-limiting PIM messages is proposed in section 5.3.3 [RFC4609], but has the identified drawbacks of impacting the service delivered and having side-effects on legitimate users.

4.2. Existing PIM, IGMP and MLD timers

In the context of PIM multicast routing protocols (), a mechanism exists that in some context may offer a form of de facto damping

mechanism for multicast states. Indeed, when active, the prune override mechanism consist in having a PIM upstream router delay for a certain time [prune override interval] before taking into account a PIM Prune message sent by a downstream neighbor. This mechanism has not been designed specifically for the purpose of damping multicast state, but as a means to allow PIM to operate on multi-access networks. See [RFC4601] section 4.3.3.

However, when active, this mechanism will prevent a downstream router to produce multicast routing protocol messages for a said multicast state that would result in the upstream router to send, to its own upstream, multicast routing protocol messages at a rate higher than $1/[\text{prune override interval}]$.

Similarly, the IGMP and MLD multicast membership control protocols can provide under the right conditions a similar behavior.

These mechanisms are not considered suitable to meet the goals spelled out in Section 1, the main reasons being that:

- o when enabled these mechanisms require additional bandwidth on the local link on which the effect of a Prune is delayed
- o to be active, they may require disabling features that may otherwise be required or useful; one typical example is explicit tracking for IGMP/MLD or PIM
- o on certain implementation, would require disabling behavior that cannot be turned off
- o do not provide a suitable level of configurability
- o do not provide a way to discriminate between multicast flows based on an averaged estimation of their recent past dynamicity

4.3. BGP Route Damping

The procedures defined in [RFC2439] for BGP route flap damping are useful for operators who want to control the impact of unicast route churn on the routing infrastructure, and offer a standardized set of parameters to control damping.

These procedures are not directly relevant in a multicast context, for the following reasons:

- o they are not specified for multicast routing protocol in general

- o even in contexts where BGP routes are used to carry multicast routing states (e.g. [RFC6514]), these procedures do not allow to implement the principle described in this document, the main reason being that a damped route becomes suppressed, while the target behavior would be to keep advertising when damping is triggered on a multicast route

However, the set of parameters standardized to control the thresholds of the exponential decay mechanism can be relevantly reused. This is the approach proposed for the procedures described in this document (Section 5). Motivations for doing so is to help the network operator deploy this feature based on consistent configuration parameter, and obtain predictable results, without the drawbacks of exposed in Section 4.1 and Section 4.2.

5. Procedures for multicast state damping

This section describes procedures for multicast state damping satisfying the goals spelled out in Section 1. This section spells out procedures for (S,G) states in the PIM-SM protocol ([RFC4601] ; they apply unchanged for such states created based on multicast group management protocols (IGMP [RFC3376], MLD [RFC3810]) on downstream interfaces. How these procedures apply for any-source multicast (ASM) routing state will be covered in a further revision.

The following notions introduced in [RFC2439] are reused in these procedures:

figure-of-merit **a number reflecting the current estimation of past recent activity of an (S,G) multicast routing state, which evolves based on routing events related to this state and based an exponential decay algorithm ; the activation or inactivation of damping on the state is based on this number

cutoff-threshold parameter value of the *figure-of-merit* over which damping is applied (configurable value)

reuse-threshold parameter value of the *figure-of-merit* under which damping stops being applied (configurable value)

decay-half-life parameter period of time used to control how fast is the exponential decay of the *figure-of-merit* (configurable value)

Additionally to these values a configurable "*increment-factor*" parameter is introduced, that controls by how much the figure-of-merit is incremented on multicast state update events.

Section 8.4 will propose default values for all these parameters.

On reception of updated multicast membership or routing information on a downstream interface I for a said (S,G) state, that results in a change of the state of the PIM downstream state machine (see section 4.5.3 of [RFC4601]), a router implementing these procedures MUST:

- o apply unchanged procedures for everything relating to what multicast traffic ends up traffic being sent on downstream interfaces, including interface I
- o increasing the **figure-of-merit** for the (S,G) by the **increment-factor** (updating the **figure-of-merit** based on the decay algorithm must be done prior to this increment)
- o update the damping state for the (S,G) state: damping becomes active on the state if the recomputed **figure-of-merit** is above the configured **cutoff-threshold**
- o update the upstream state machine for (S,G) as per section 4.5.7 of [RFC4601], with the following change : if the state machine transitions to NotJoined state because of the reception of a PIM or IGMP/MLD message on a downstream interface (i.e. in the terminology of [RFC4601] *inheritedolist(S,G)* becomes NULL), and if damping is active on the state, the router SHOULD NOT send the resulting Prune(S,G) message to its upstream neighbor ; this message MUST be sent when the damping state becomes, i.e. inactive when **figure-of-merit** decays to a value below the configured **reuse-threshold**

Same techniques as the ones described in [RFC2439] can be applied to determine when the figure-of-merit value is recomputed based on the exponential decay algorithm and the configured **decay-half-life**. Given the specificity of multicast applications, it is REQUIRED for the implementation to let the operator configure the **decay-half-life** in seconds, rather than in minutes. When the recomputation is done periodically, the period should be low enough to not significantly delay the inactivation of damping on a multicast state beyond what the operator wanted to configure (i.e. for a half-life of 10s, recomputing the **figure-of-merit** each minute would result in a multicast state to remained damped for a time longer than what the parameters are supposed to command).

When a (S,G) state expires, its associated **figure-of-merit** and damping state are removed as well.

These procedures do interact with PIM procedures related to refreshes

or expiration of multicast routing states. Indeed, PIM Prune messages triggered by the expiration of the (S,G) keep-alive timer, are not suppressed or delayed (see Section 8.3 for a discussion on why this specific aspect is not expected to impede the efficiency of damping procedures), and the reception of Join messages not causing transition of state on the downstream interface does not lead to incrementing the *figure-of-merit*.

Note that these procedures do not impact the PIM assert mechanism, in particular PIM Prune messages triggered by a change of the PIM assert winner on the upstream interface, are not suppressed or delayed.

Note also that no action is triggered based on the reception of PIM Prune messages (or corresponding IGMP/MLD messages) that relate to non-existing (S,G) state, in particular, no *figure-of-merit* or damping state is created in this case.

6. Procedures for multicast in IP VPNs

In VPN contexts, providing isolation between customers of a shared infrastructure is a core requirement resulting in even stringent expectations with regards to risks of denial of service attacks. Procedures for multicast support in IP VPNs are described in [RFC6513] and [RFC6514] and section 16 of [RFC6514] specifically spells out the need for damping the activity of C-multicast and Leaf Auto-discovery route.

The procedures described in Section 5 can be applied in the VRF PIM-SM implementation (in the "C-PIM instance"), with the corresponding action to suppressing the emission of a Prune(S,G) message being to not withdraw the C-multicast Source Tree Join (C-S,C-G) BGP route. Implementation of [RFC6513] relying on the use of PIM to carry C-multicast routing information MUST support this technique.

In the context of [RFC6514] where BGP is used to distribute C-multicast routing information, an additional option consists in applying damping at the level of the BGP implementation based on existing BGP damping mechanism, applied to C-multicast Source Tree Join routes and Shared Tree Join routes (and also Leaf A-D routes - see Section 6.1), and modified to provide the same effect of procedures described in Section 5 along the following guidelines:

- o not withdrawing (instead of not advertising) damped routes
- o providing means to configure the half-life in seconds if that option is not already available

- o using parameters for the exponential decay that are specific to multicast, based on default values and multicast specific configuration

Note that in a context where BGP Route Reflectors are used, it can be considered useful to also be able to apply damping on RRs. Additionally, for mVPN Inter-AS deployments, it can be needed to protect one AS from the dynamicity of multicast VPN routing events from other ASes. In that perspective, it is RECOMMENDED for implementations to support damping mVPN C-multicast routes directly into BGP, without relying on the PIM-SM state machine.

The choice to implement damping based on BGP routes or the procedures described in Section 5, is up to the implementor, but at least one of the two MUST be implemented; keeping in mind that in contexts where damping on RRs and ASBRs the BGP approach is RECOMMENDED.

Note well that damping SHOULD NOT be applied to BGP routes of the following sub-types: "Intra-AS I-PMSI A-D Route", "Inter-AS I-PMSI A-D Route", "S-PMSI A-D Route", and "Source Active A-D Route".

The following sub-sections describe additional procedures providing coverage against harmful effects of high multicast membership state dynamicity specific to mVPNs, and preserving the goals spelled out in Section 1.

6.1. Damping P-tunnel change events

When selective P-tunnels are used (see section 7 of [RFC6513]), the effect of updating the upstream state machine for a said (C-S,C-G) state on a PE connected to multicast receivers, is not only to generate activity to propagate C-multicast routing information to the source connected PE, but also to possibly trigger changes related to the P-tunnels carrying (C-S,C-G) traffic. Protecting the provider network for an excessive amount of change in the state of P-tunnels is required, and this section details how it can be done.

A PE implementing these procedures for mVPN MUST damp Leaf A-D routes, in the same manner as it would for C-multicast routes (see Section 6).

A PE implementing these procedures for mVPN MUST damp the activity related to removing itself from a P-tunnel. Possible ways to do so depend on the type of P-tunnel, and local implementation details are left up to the implementor.

The following is proposed as example of how the above can be achieved.

- o For P-tunnels implemented with the PIM protocol, this consists in applying multicast state damping techniques describe in Section 5 to the P-PIM instance, at least for (S,G) states corresponding to P-tunnels.
- o For P-tunnels implemented with the mLDP protocol, this consists in applying damping techniques completely similar as the one described in Section 5, but generalized to apply to mLDP states
- o For root-initiated P-tunnel (P-tunnels implemented with the P2MP RSVP-TE, or relying on ingress replication), no particular action needs to be implemented to damp P-tunnels membership as soon as the activity of Leaf A-D route is damped
- o Another possibility is to base the decision to join or not join the P-tunnel to which a said (C-S,C-G) is bound, and to advertise or not advertise a Leaf A-D route related to (C-S,C-G), based on whether or not a C-multicast Source Tree Join route is being advertised for (C-S,C-G), rather than by relying on the state of the C-PIM Upstream state machine for (C-S,C-G)

7. Procedures for Ethernet VPNs

Specifications exists to support or optimize multicast and broadcast in the context of Ethernet VPNs ([I-D.ietf-l2vpn-vpls-mcast], [I-D.ietf-l2vpn-evpn]). The said specifications make use of S-PMSI and P-tunnels and for this reason, an implementation of these procedures MUST follow the procedures described in Section 6.1.

8. Operational considerations

8.1. Enabling and configuring multicast damping

In the context of flat multicast routing, it is proposed that enabling this multicast damping mechanism at the edge of a network providing a multicast service, for instance at receiver-facing routers or in ASBRs, will be sufficient to address the targeted issue. Additionally, these procedures can be enabled on core routers as well.

In the context of multicast VPNs, these procedures would be enabled on PE routers. Additionally in the case of C-multicast routing based on BGP extensions ([RFC6514]) these procedures can be enabled on ASBRs, and possibly Route Reflectors as well.

8.2. Troubleshooting and monitoring

Implementing the damping mechanisms described in this document should be complemented by appropriate tools to observe and troubleshoot damping activity.

More specifically it is RECOMMENDED to complement the existing interface providing information on multicast states with information on eventual damping of corresponding states (e.g. MRIB states). In the case of mVPN this applies also to information on P-tunnels damping, and when BGP is used for C-multicast routing propagation, to BGP C-multicast routes.

8.3. Maximum values for exponential decay and thresholds parameters

[TBC]

8.4. Default values

[TBC]

9. IANA Considerations

This document makes no request of IANA.

Note to RFC Editor: this section may be removed on publication as an RFC.

10. Security Considerations

The procedures defined in this document do not introduce additional security issues not already present in the contexts addressed, and actually aim at addressing some of the identified risks without introducing as much denial of service risk as some of the mechanisms already defined.

The protection provided relates to the control plane of the multicast routing protocols, including the components implementing the routing protocols and the components responsible for updating the multicast forwarding plane.

The procedures describe are meant to provide some level of protection for the router on which they are enabled by reducing the amount of routing state updates that it needs to send to its upstream neighbor or peers, but do not provide any reduction of the control plane load related to processing routing information from downstream neighbors.

Protecting routers from an increase in control plane load due to activity on downstream interfaces toward core routers (or in the context of BGP-based mVPN C-multicast routing, BGP peers) shall rely upon the activation of damping on corresponding downstream neighbors (or BGP peers) and/or at the edge of the network. Protecting routers from an increase in control plane load due to activity on customer-facing downstream interfaces or downstream interfaces to routers in another administrative domain, is out of the scope of this document and should rely upon already defined mechanisms (see [RFC4609]).

To be effective the procedures described here must be complemented by configuration limiting the number of multicast states that can be created on a multicast router through protocol interactions with multicast receivers, neighbor routers in adjacent ASes, or in multicast VPN contexts with multicast CEs. Note well that the two mechanism may interact: state for which Prune has been requested may still remain taken into account for some time if damping has been triggered and hence result in otherwise acceptable new state from being successfully created.

Additionally, it is worth noting that these procedures are not meant to protect against peaks of control plane load, but only address averaged load. For instance, assuming a set of multicast states submitted to the same Join/Prune events, damping can prevent more than a certain number of Join/Prune messages to be sent upstream in the period of time that elapses between the reception of Join/Prune messages triggering the activation of damping on these states and when damping becomes inactive after decay.

11. Acknowledgements

We would like to thank Bruno Decreane, Jeff Haas and Lenny Giuliano for discussions that helped shape this proposal. We would also like to thank Yakov Rekhter and Eric Rosen for their reviews and helpful comments. Thanks to Wim Henderickx for his comments and support of this proposal.

12. References

12.1. Normative References

[I-D.ietf-l2vpn-evpn]

Sajassi, A., Aggarwal, R., Henderickx, W., Balus, F., Isaac, A., and J. Uttaro, "BGP MPLS Based Ethernet VPN", draft-ietf-l2vpn-evpn-04 (work in progress), July 2013.

- [I-D.ietf-l2vpn-vpls-mcast]
Aggarwal, R., Rekhter, Y., Kamite, Y., and L. Fang,
"Multicast in VPLS", draft-ietf-l2vpn-vpls-mcast-14 (work
in progress), July 2013.
- [RFC2119] Bradner, S., "Key words for use in RFCs to Indicate
Requirement Levels", BCP 14, RFC 2119, March 1997.
- [RFC2439] Villamizar, C., Chandra, R., and R. Govindan, "BGP Route
Flap Damping", RFC 2439, November 1998.
- [RFC3376] Cain, B., Deering, S., Kouvelas, I., Fenner, B., and A.
Thyagarajan, "Internet Group Management Protocol, Version
3", RFC 3376, October 2002.
- [RFC3810] Vida, R. and L. Costa, "Multicast Listener Discovery
Version 2 (MLDv2) for IPv6", RFC 3810, June 2004.
- [RFC4601] Fenner, B., Handley, M., Holbrook, H., and I. Kouvelas,
"Protocol Independent Multicast - Sparse Mode (PIM-SM):
Protocol Specification (Revised)", RFC 4601, August 2006.
- [RFC6513] Rosen, E. and R. Aggarwal, "Multicast in MPLS/BGP IP
VPNs", RFC 6513, February 2012.
- [RFC6514] Aggarwal, R., Rosen, E., Morin, T., and Y. Rekhter, "BGP
Encodings and Procedures for Multicast in MPLS/BGP IP
VPNs", RFC 6514, February 2012.

12.2. Informative References

- [RFC4609] Savola, P., Lehtonen, R., and D. Meyer, "Protocol
Independent Multicast - Sparse Mode (PIM-SM) Multicast
Routing Security Issues and Enhancements", RFC 4609,
October 2006.

Authors' Addresses

Thomas Morin (editor)
Orange
2, avenue Pierre Marzin
Lannion 22307
France

Email: thomas.morin@orange.com

Stephane Litkowski
Orange
France

Email: stephane.litkowski@orange.com

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

Email: keyupate@cisco.com

Jeffrey (Zhaohui) Zhang
Juniper Networks Inc.
10 Technology Park Drive
Westford, MA 01886
USA

Email: zzhang@juniper.net

Robert Kebler
Juniper Networks Inc.
10 Technology Park Drive
Westford, MA 01886
USA

Email: rkebler@juniper.net

Network Working Group
Internet Draft
Intended status: Standards Track
Expires: January 2014

Hui Ni
ShunWan Zhuang
Zhenbin Li
Huawei
July 8, 2013

BGP Extensions for Service-Driven Co-Routed MPLS Traffic Engineering
LSP
draft-ni-l3vpn-bgp-ext-sd-co-lsp-00.txt

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), its areas, and its working groups. Note that other groups may also distribute working documents as Internet-Drafts.

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."

The list of current Internet-Drafts can be accessed at
<http://www.ietf.org/ietf/lid-abstracts.txt>

The list of Internet-Draft Shadow Directories can be accessed at
<http://www.ietf.org/shadow.html>

This Internet-Draft will expire on January 8, 2014.

Copyright Notice

Copyright (c) 2013 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.

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.

Abstract

In some large scale L3VPN deployment scenarios like mobile backhaul network, it is required that tunnels between two PEs could be setup automatically driven by L3VPN service to reduce manual configuration effort. Moreover the tunnels must be setup co-routed for the goodness of performance monitoring and uniform protection behavior for link failure on two directions. This is described in [I-D.li-mppls-serv-driven-co-lsp-fmwk]. This document introduces a new BGP VT route based on [I-D.ni-bgp-ext-l3vpn-pm]. The route is utilized by on side PE to advertise Tunnel ID to other side PE, so inverse direction co-routed LSPs can be setup based on path information of member LSPs in the first tunnel.

Table of Contents

1. Introduction	3
2. Conventions used in this document.....	3
3. Terminologies	3
4. VT Tunnel-ID Signal Route Definition	4
5. Operations	6
5.1. Active/Passive PE Selection.....	6
5.2. VPN Membership Auto Discovery.....	7
5.3. Active PE Advertise VT Tunnel-ID Signal Route	7
5.4. Passive PE advertise Tunnel ID to Active PE	7
5.5. VT Tunnel-ID Signal Route Application	8
6. VT Tunnel-ID Signal Route Selection Consideration	8
7. Deployment Consideration.....	9
8. Security Considerations.....	9
9. Normative References.....	9
10. Informative References.....	10
11. Acknowledgments	10

1. Introduction

In some large scale L3VPN deployment scenarios like mobile backhaul network, it is required that LSPs between two PEs could be setup automatically driven by L3VPN service to reduce manual configuration effort and the LSPs must be setup co-routed for the goodness of performance monitoring and uniform protection behavior under link failure, which is in detail described in [I-D.li-mpls-serv-driven-co-lsp-fmwk].

This document introduces a new BGP Tunnel-ID Signal Route under VT SAFI([I-D.ni-bgp-ext-l3vpn-pm]). After VPN-membership discovered one PE can use the route to proactively advertise one direction MPLS TE Tunnel ID to remote PE, the latter can setup inverse direction co-routed LSPs based on path information of member LSPs of the first tunnel. For RSVP-TE type LSP, the path information could be got from RRO naturally.

The extension is based on VT SAFI defined in [I-D.ni-bgp-ext-l3vpn-pm].

Type 1(VT VPN Membership A-D Route) is utilized to support VPN-membership auto-discovery between a pair of VRFs.

A new Type 3 VT Route, namely VT Tunnel-ID Signal Route, is defined in this document.

2. Conventions used in this document

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].

3. Terminologies

This document uses the terminologies defined in [RFC3031], [RFC3209], [RFC4026] and [I-D.ni-bgp-ext-l3vpn-pm].

ERT: Export Route Target

IRT: Import Route Target

LSP: Label Switched Path

LSR: Label Switching Router

MPLS: Multi Protocol Label Switching

PE: Provider Edge Router in BGP/MPLS VPN

RRO: Record Route Object

RSVP-TE: Traffic Engineering Extensions of RSVP

VT: VRF-to-VRF Tunnel

4. VT Tunnel-ID Signal Route Definition

VT Tunnel-ID Signal Route defined as Type 3 Route under BGP VT NLRI

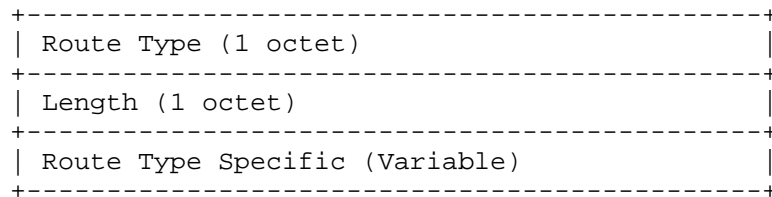


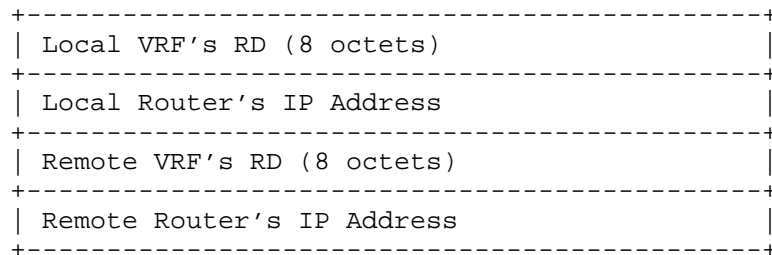
Figure 1 IPv4 VT-Family NLRI

a)Route Type:

Type 3: indicates VT Tunnel-ID Signal Route

b)Length indicates Route Type Specific field's length in octets

c)Route Type specific contains VT Tunnel-ID Signal Route information, encoded as following diagram.



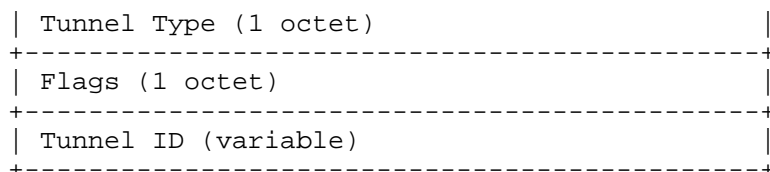


Figure 2 VT VRF-to-VRF Tunnel-ID Signal Route Format

- a) Local VRF's RD Route Distinguisher of one VRF on advertising PE encoded as [RFC4364] definition.
- b) Local Router's IP Address: Advertising PE's IPv4/IPv6 address.
- c) Remote VRF's RD Route Distinguisher of one VRF on Receiving PE encoded as [RFC4364] definition.
- d) Remote Router's IP Address: Receiving PE's IPv4/IPv6 address.
- e) Tunnel Type: Indicates type of tunnel

Type 0: RESERVED

Type 1: RSVP-TE Tunnel

Type other: To be defined later if necessary

- f) Flags: 8-bits Flags

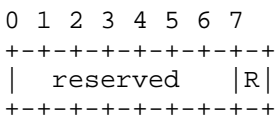


Figure 3 Tunnel Flags Format

The high-order 7 bits are RESERVED and MUST be filled with ZERO, the lowest R bit is defined to indicate Tunnel's direction:

0-Passive: indicate the tunnel is setup by Passive PE

1-Active: indicate the tunnel is setup by Active PE

- g) Tunnel ID Tunnel Identifier defined according to specific Tunnel type.

For RSVP-TE type tunnel defined in this document, Tunnel Identifier is < tunnel end point address, Reserved, Tunnel ID, Extended Tunnel ID> as carried in the RSVP-TE LSP's SESSION Object [RFC3209].

Following diagram describes RSVP-TE IPv4 Tunnel ID as example.

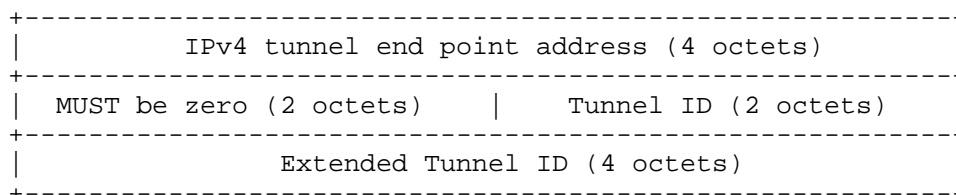


Figure 4 RSVP-TE Tunnel Identifier Format

- a) IPv4 tunnel end point address: 4 octets IPv4 address of Egress PE.
- b) Tunnel ID: 2 octets Tunnel Identifier allocated by Ingress PE which remains constant over the life of the RSVP-TE tunnel.
- c) Extended Tunnel ID: 4 octets Extended Tunnel Identifier which also remains constant over the life of the RSVP-TE Tunnel, MUST be filled with ZERO or IPv4 Address of Ingress PE.

<Local VRF's RD, Local Router's IP Address, Remote VRF's RD, Remote Router's IP Address> which indicates a pair of VRFs is defined as the Prefix of VT Tunnel-ID Signal Route.

5. Operations

5.1. Active/Passive PE Selection

For a pair of PEs which are going to exchange VT Tunnel-ID Signal Routes need to select Active PE/Passive PE first. The selection SHOULD be done by two ways:

- a) Locally decided by manually configuration. That is one PE is configured as Active PE and other PE is configured as Passive PE.
- b) Doing auto-selection in BGP's Capability Negotiation phase. Router IDs of the two PEs are compared as unsigned integers, the PE with larger Router ID value MUST be selected as Active PE and the PE with smaller Router ID value is selected as Passive PE.

5.2. VPN Membership Auto Discovery

Following the detailed description in [I-D.ni-bgp-ext-l3vpn-pm], all PEs firstly need to send VT VPN A-D Routes to do VPN membership discovery.

5.3. Active PE Advertise VT Tunnel-ID Signal Route

Once found a pair of VRFs between Active PE and Passive PE belongs to the same VPN, Active PE MUST initialize setup a RSVP-TE tunnel to Passive PE for the pair of VRFs. Once Tunnel ID is allocated locally, Active PE MUST advertise the RSVP TE tunnel info to Passive PE by using VT Tunnel-ID Signal Route:

Local VRF's RD: MUST be filled with Route Distinguisher value of the Local VRF.

Local Router's IP Address: MUST be filled with Active PE's IPv4/IPv6 Address.

Remote VRF's RD: MUST be filled with Route Distinguisher value of the remote VRF on Passive PE which belongs to same L3VPN with Local VRF.

Remote Router's IP Address: MUST be filled with Passive PE's IPv4/IPv6 Address.

Tunnel Type: MUST be filled with a valid Tunnel Type value, for example filled with value <1> which indicates RSVP-TE Tunnel.

Flags: the high-order 7 bits MUST be filled with ZERO and "R" bit flag MUST be filled with value <1> which indicates the tunnel is initialized by Active PE.

Tunnel ID: <IP tunnel end point address> field MUST be filled with IP address of Passive PE, <Tunnel ID> field MUST be filled with the Tunnel Identifier value allocated by Active PE, <Extended Tunnel ID> filed MUST be filled with ZERO or IP address of Active PE.

Note that if tunnel goes down, the responding VT Tunnel-ID Signal Route Withdrawal message SHOULD NOT be sent out.

5.4. Passive PE advertise Tunnel ID to Active PE

After receiving VT Tunnel-ID Signal Route from Active PE, Passive PE gets all LSP's path information under the Tunnel and setup inverse direction RSVP-TE LSPs following same path.

Once Tunnel ID is allocated locally, the Tunnel ID information MUST be advertised from Passive PE to Active PE through VT Tunnel-ID Signal Route:

Local VRF's RD: MUST be filled with Route Distinguisher value of the Local VRF on Passive PE.

Local Router's IP Address: MUST be filled with Passive PE's IPv4/IPv6 Address.

Remote VRF's RD: MUST be filled with Route Distinguisher value of the remote VRF on Active PE which belongs to same L3VPN with Local VRF.

Remote Router's IP Address: MUST be filled with Active PE's IPv4/IPv6 Address.

Tunnel Type: MUST be filled with a valid Tunnel Type value, usually the tunnel Type SHOULD be same with previous Tunnel Type setup by Active PE.

Flags: Filled "R" Flag with value "0" which indicates the tunnel is setup by Passive PE.

Tunnel ID: "IPv4 tunnel end point address" field MUST be filled with IP address of Active PE, "Tunnel ID" field MUST be filled with the Tunnel Identifier value allocated by Passive PE, "Extended Tunnel ID" MUST be filled with ZERO or IP Address of Passive PE.

5.5. VT Tunnel-ID Signal Route Application

When Active PE and Passive PE learnt VT Tunnel-ID Signal Routes from each other, both PEs need binding two uni-direction Tunnels together for the L3VPN. It means L3VPN traffic is transit over the tunnels and that if Active PE changes LSP in Active Tunnel, Passive PE MUST detect it and manipulate the reverse direction LSP in Passive Tunnel accordingly through RSVP protocol.

6. VT Tunnel-ID Signal Route Selection Consideration

VT Tunnel-ID Signal Route SHOULD follow the BGP best route selection procedure described in [RFC4271].

VT Tunnel-ID Signal Route MUST be advertised ONLY to the Peer from which the best VT VPN A-D route is received. VT VPN A-D route is the one which carries the same Remote VRF's RD value and Remote PE's IP address

If Peer receives VT Tunnel-ID Signal Route originated from itself, the route MUST be ignored.

7. Deployment Consideration

This document currently supports deploying VT Tunnel-ID Signal Route in 2 manners:

Inner-AS L3VPN: with Full-mesh IBGP sessions or Router Reflectors.

Inter-AS L3VPN: with Option A (VRF-to-VRF)[RFC4364].

How to support Inter-AS L3VPN Option B(MP-EBGP) [RFC4364] and Option-C (Multi-Hop MP-EBGP) [RFC4364] will be described in this draft's future version.

8. Security Considerations

This extension to BGP does not change the underlying security issues.

9. Normative References

- [1] [I-D.li-mpls-serv-driven-co-lsp-fmwk] Z. Li, "A Framework for Service-Driven Co-Routed MPLS Traffic Engineering LSPs", April 2013.
- [2] [I-D.ni-bgp-ext-l3vpn-pm] H. Ni, "BGP Extension For L3VPN Performance Monitoring", June 2013.
- [3] [RFC3209] D. Awduche, "RSVP-TE: Extensions to RSVP for LSP Tunnels", RFC 3209, December 2001.
- [4] [RFC4026] L. Andersson, "Provider Provisioned Virtual Private Network (VPN) Terminology", RFC4026, March 2005.
- [5] [RFC4271] Y. Rekhter, "A Border Gateway Protocol 4 (BGP-4)", RFC 4271, January 2006.
- [6] [RFC4364] E. Rosen, "BGP/MPLS IP Virtual Private Networks (VPNs)", RFC 4364, February 2006.
- [7] [RFC4456] T. Bates, "BGP Route Reflection: An Alternative to Full Mesh Internal BGP (IBGP)", BCP 14, RFC 4456, April 2006.

10. Informative References

N.A

11. Acknowledgments

Authors' Addresses

Hui Ni
Huawei Technologies
Huawei Building, No.156 Beiqing Rd.
Beijing 100095
China
Email: nihui@huawei.com

Shunwan Zhuang
Huawei Technologies
Huawei Building, No.156 Beiqing Rd.
Beijing 100095
China
Email: zhuangshunwan@huawei.com

Zhenbin Li
Huawei Technologies
Huawei Building, No.156 Beiqing Rd.
Beijing 100095
China
Email: lizhenbin@huawei.com

Network Working Group
Internet Draft
Intended status: Standards Track
Expires: January 2014

Hui Ni
Shunwan Zhuang
Zhenbin Li
Huawei
July 8, 2013

BGP Extension For L3VPN Performance Monitoring
draft-ni-l3vpn-pm-bgp-ext-00.txt

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), its areas, and its working groups. Note that other groups may also distribute working documents as Internet-Drafts.

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."

The list of current Internet-Drafts can be accessed at <http://www.ietf.org/ietf/lid-abstracts.txt>

The list of Internet-Draft Shadow Directories can be accessed at <http://www.ietf.org/shadow.html>

This Internet-Draft will expire on January 8, 2014.

Copyright Notice

Copyright (c) 2013 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.

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.

Abstract

This document describes a new VT address family in BGP to exchange information required for apply performance monitoring in MPLS/BGP VPN, as described in [I-D.dong-l3vpn-pm-framework].

Table of Contents

1. Introduction	2
2. Conventions used in this document.....	3
3. Terminologies	3
4. The New VT Sub Address Family.....	3
4.1. VT Sub Address Family.....	4
4.2. VT NLRI	4
4.2.1. VT VPN-membership A-D Route	5
4.2.2. VT Labeled Route.....	5
5. Operations	6
5.1. VRF-to-VRF VPN Membership Auto Discovery	6
5.2. VRF-to-VRF Labeled Route Exchange	7
5.2.1. VT Labeled Route Update.....	7
5.2.2. VT Labeled Route Withdrawal	8
5.3. VRF-to-VRF Labeled Route Application	8
6. VT Route Selection Consideration.....	8
7. Deployment Consideration.....	9
8. Security Considerations.....	9
9. IANA Considerations	9
9.1. Normative References.....	9
9.2. Informative References.....	10
10. Acknowledgments	10

1. Introduction

This document describes the BGP encodings and procedures for exchanging the information elements required by applying traffic performance monitoring in MPLS/BGP VPN, as specified in [ID.draft-dong-l3vpn-pm-framework-01].

Current BGP Labeled VPN Route exchange procedure combines VRF VPN-membership Auto-Discovery and L3VPN Label allocation together. While applying PM for L3VPN needs BGP extended to support VPN membership Auto-Discovery and L3VPN Label allocation in a VRF-to-VRF manner. To achieve this, a new Sub address family, called VRF-to-VRF Tunnel(VT) Subsequent Address Family, is introduced.

This document defines two kinds of routes for VT NLRI:

VPN-Membership A-D Route: for the use of doing VRF VPN membership auto-discovery in VRF-to-VRF manner

VT Labeled Route: for the use of allocating VT Label from Local VRF to Remote VRF to setup VRF-to-VRF Tunnel between the pair of VRFs.

2. Conventions used in this document

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].

3. Terminologies

This document uses the terminologies defined in [RFC4026]:

ERT: Export Route Target

IRT: Import Route Target

PE: Provider Edge

RD: Route Distinguisher

VRF: Virtual Routing and Forwarding

VT: VRF-to-VRF Tunnel

4. The New VT Sub Address Family

The BGP Multiprotocol Extensions [RFC4760] allow BGP to carry routes from multiple "address families". In this document a new Subsequent Address Family is introduced, called "VT Sub Address Family".

4.1. VT Sub Address Family

VT Address Family uses AFI 1/2 to present IPv4/IPv6 Address Family and a specific VT_SAFI(TBD) to present VT Subsequent Address Family.

VT MP_REACH_NLRI and MP_UNREACH_NLRI are formatted as described in [RFC4760]

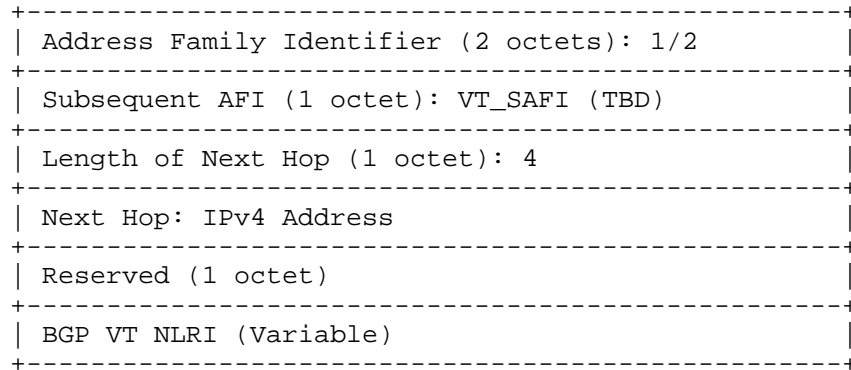


Figure 1 VT MP_REACH_NLRI

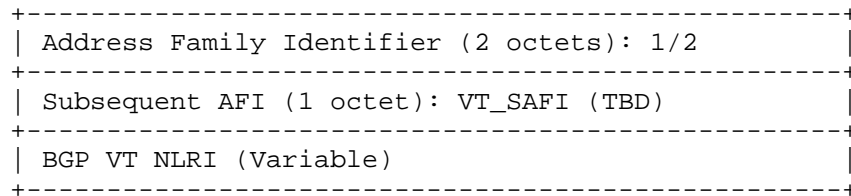
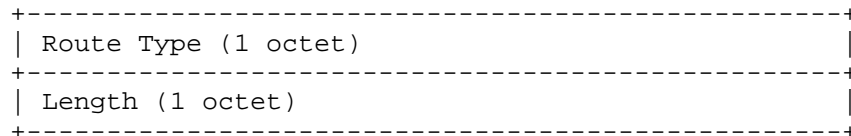


Figure 2 VT MP_UNREACH_NLRI

4.2. VT NLRI

BGP VT NLRI has format as depicted in following diagram



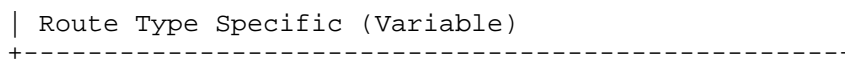


Figure 3 IPv4 VT-Family NLRI

Route Type indicates type of route under VT SAFI.

Type 1: VT VPN membership A-D Route

Type 2: VT Labeled Route

Length defines Route Type specific routes length in octets

Route Type specific route information field, encoded according to Route Type definition.

4.2.1. VT VPN-membership A-D Route

VT VPN membership A-D Route, concisely named as VT A-D Route hereafter, is utilized for VRF-to-VRF VPN Membership Auto-Discovery between PEs.

Its format is defined as following diagram:

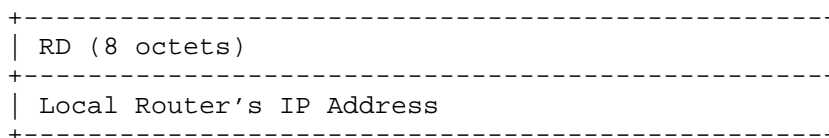


Figure 4 VT VPN Membership A-D Route

a)RD RD of one VRF on advertising PE, encoded as described in [RFC4364].

b)Local Router's IP Address Advertising PE's IPv4/IPv6 address

<RD, Local Router's IP Address> is defined as Prefix of VT A-D Route.

4.2.2. VT Labeled Route

VT Labeled Route is utilized for VRF-To-VRF Label(s) allocation and advertisement, its format is defined as following diagram.

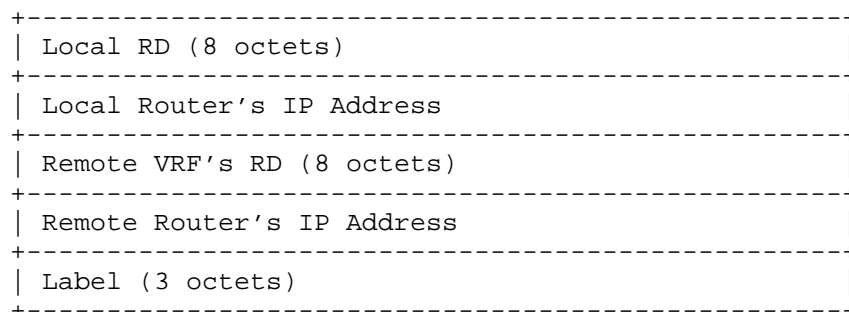


Figure 5 VT Labeled Route Format

a)Local RD Route Distinguisher value of one VRF on advertising PE, encoded as described in [RFC4364].

b)Local Router's IP Address Advertising PE's IPv4/IPv6 address.

c)Remote VRF's RD Route Distinguisher value of Remote VRF encoded as described in [RFC4364].

d)Remote Router's IP Address: Remote PE's IPv4/IPv6 address.

e)Label The Label field carries one or more labels that corresponds to the stack of labels [RFC3032]. Each label is encoded as 3 octets, where the high-order 20 bits contain the label value, and the low-order bit contains "Bottom of Stack" as defined in [RFC3032].

<Local RD, Local Router's IP Address, Remote VRF's RD, Remote Router's IP Address> which indicates a pair of VRFs is defined as the Prefix of VT Labeled Route.

5. Operations

5.1. VRF-to-VRF VPN Membership Auto Discovery

For every PE, it needs to process all its VRF configured and generate one VT A-D Route for each VRF respectively.

RD field MUST be filled with the VRF's RD value.

Local Router's IP Address field MUST filled with the Advertising Router's IP address.

The VT A-D Route MUST carry all IRTs of the VRF in BGP Update's Ext-Community Path Attribute, route importing request of one VRF is described by its corresponding VT A-D route. In contrast VPN Labeled Routes carry ERTs in BGP Update's Ext-Community Path Attribute.

If a VRF is created, then its corresponding VT A-D Route MUST be generated and advertised.

If the VRF whose VT A-D Route has been advertised is deleted, then the VT A-D Route Withdrawal message MUST be generated and advertised.

If IRT of the VRF whose VT A-D Route has been advertised is changed, then a VT A-D Route Update with same Prefix and latest IRTs MUST be advertised.

When receiving PE receives VT A-D Route, VPN relationship matching MUST be checked between IRTs in VT A-D Route and ERTs of each Local VRF, this process is called VRF-to-VRF VPN membership Auto Discovery.

Either finding one VRF-to-VRF VPN membership newly formed or released, receiving PE MUST proceed to the VT Labeled Route processing described in next section.

5.2. VRF-to-VRF Labeled Route Exchange

5.2.1. VT Labeled Route Update

If Receiving PE finds one new VRF-to-VRF VPN membership formed, it MUST allocate one VT MPLS Label for the VRF-to-VRF VPN membership and the label is advertised to the Remote VRF by VT Labeled Route.

Local RD MUST filled with RD value of the Local VRF which is found belong to the same VPN with Remote VRF.

Local Router's IP Address Must filled with the advertising PE's IPv4/IPv6 address.

Remote VRF's RD MUST filled with RD value of the Remote VRF which belongs to a same VPN with the Local VRF.

Remote Router's IP Address: Remote PE's IPv4/IPv6 address.

Label: MUST be filled with one or more MPLS Labels allocated by advertising PE for the pair of VRFs.

Only both sides of a pair of VRFs learnt each other's VT Labeled Route advertisement, the VRF-to-VRF tunnel between the pair of VRFs is considered setup.

5.2.2. VT Labeled Route Withdrawal

If receiving PE finds one existing VRF-to-VRF VPN membership released then it MUST send out the VT Labeled Route Withdrawal message, then release the MPLS Label(s) allocated.

Local RD MUST be filled with RD value of the Local VRF.

Local Router's IP Address MUST be filled with the advertising PE's IPv4/IPv6 address.

Remote VRF's RD MUST be filled with RD value of the Remote VRF.

Remote Router's IP Address: MUST be filled with Remote PE's IPv4/IPv6 address.

Label: MUST be filled with ZERO or the MPLS Labels value allocated for the VT Labeled Route.

5.3. VRF-to-VRF Labeled Route Application

To achieve the goal of converting normal L3VPN MP2P forwarding model into P2P model which is required in [ID.draft-zheng-l3vpn-pm-analysis-01], after VPNv4 routes received, Receiving PE MUST apply VT Labels when downloading VPNv4 Route into Data Plan which is in detail described in [I-D.dong-l3vpn-pm-framework].

Between a pair of PEs both support VT capability, It COULD be an implementation option that VPNv4 Routes from a remote VRF WOULD NOT be downloaded into a Local VRF's Forwarding Plan until a VT Labeled route received from same Remote VRF for the Local VRF.

If VT Labeled Route withdrawal message is received, receiving PE MUST delete VT Labels from Forwarding Plane and VPNv4 Routes MUST be kept on Forwarding Plane with original VPNv4 Label as inner Label.

6. VT Route Selection Consideration

When receiving and processing VT A-D Route, the BGP best route selection procedure described in [RFC4271] MUST be followed.

When receiving and processing VT Labeled Route, the BGP best route selection procedure described in [RFC4271] COULD be followed.

Especially VT Labeled Route MUST be advertised ONLY to the BGP peer from which the best VT A-D route is received, the VT A-D route contains the Remote VRF's RD and Remote PE's IP address.

If a Peer receives VT A-D or VT Labeled Route originated from itself, the route MUST be ignored.

7. Deployment Consideration

This document currently supports deploying VT SAFI in following two manners:

- a) Inner-AS L3VPN with Full-mesh IBGP sessions or Router Reflectors.
- b) Inter-AS L3VPN with Option A(VRF-to-VRF)[RFC4364].

How to support Inter-AS L3VPN Option B(MP-EBGP) and Option-C [RFC4364] will be described in this draft's future version.

8. Security Considerations

This extension to BGP does not change the underlying security issues.

9. IANA Considerations

A new SAFI value to present VT Subsequent Address Family is required and to be allocated by IANA.

9.1. Normative References

- [1] [I-D.dong-l3vpn-pm-framework] J. Dong, Z. Li, B. Parise, "A Framework for L3VPN Performance Monitoring", draft-dong-l3vpn-pm-framework-01.txt.
- [2] [RFC3032] E. Rosen, "MPLS Label Stack Encoding", RFC 3032, January 2001.
- [3] [RFC4026] L. Andersson, "Provider Provisioned Virtual Private Network (VPN) Terminology", RFC4026, March 2005.
- [4] [RFC4271] Y. Rekhter, "A Border Gateway Protocol 4 (BGP-4)", RFC 4271, January 2006.

- [5] [RFC4364] E. Rosen, "BGP/MPLS IP Virtual Private Networks (VPNs)", RFC 4364, February 2006.
- [6] [RFC4456] T. Bates, "BGP Route Reflection: An Alternative to Full Mesh Internal BGP (IBGP)", RFC 4456, April 2006.
- [7] [RFC4760] T. Bates, "Multiprotocol Extensions for BGP-4", RFC 4760, January 2007.

9.2. Informative References

- [1] [ID.draft-zheng-l3vpn-pm-analysis-01] L. Zheng, Z. Li, B. Parise, "Performance Monitoring Analysis for L3VPN", draft-zheng-l3vpn-pm-analysis-01.txt.

10. Acknowledgments

Authors' Addresses

Hui Ni
Huawei Technologies
Huawei Building, No.156 Beiqing Rd.
Beijing 100095
China
Email: nihui@huawei.com

Shunwan Zhuang
Huawei Technologies
Huawei Building, No.156 Beiqing Rd.
Beijing 100095
China
Email: zhuangshunwan@huawei.com

Zhenbin Li
Huawei Technologies
Huawei Building, No.156 Beiqing Rd.
Beijing 100095
China
Email: lizhenbin@huawei.com

Network working group
Internet Draft
Category: Informational

X. Xu
Huawei Technologies

S. Hares

Y. Fan
China Telecom

C. Jacquenet
France Telecom

Expires: January 2014

July 15, 2013

Virtual Subnet: A L3VPN-based Subnet Extension Solution

draft-xu-l3vpn-virtual-subnet-00

Abstract

This document describes a Layer3 Virtual Private Network (L3VPN)-based subnet extension solution referred to as Virtual Subnet, which can be used as a kind of Layer3 network virtualization overlay approach for data center interconnect.

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), its areas, and its working groups. Note that other groups may also distribute working documents as Internet-Drafts.

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."

The list of current Internet-Drafts can be accessed at <http://www.ietf.org/ietf/lid-abstracts.txt>.

The list of Internet-Draft Shadow Directories can be accessed at <http://www.ietf.org/shadow.html>.

This Internet-Draft will expire on January 15, 2014.

Copyright Notice

Copyright (c) 2013 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.

Conventions used in this document

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].

Table of Contents

1. Introduction	4
2. Terminology	6
3. Solution Description.....	6
3.1. Unicast	6
3.1.1. Intra-subnet Unicast	6
3.1.2. Inter-subnet Unicast	7
3.2. Multicast	9
3.3. CE Host Discovery	9
3.4. ARP/ND Proxy	10
3.5. CE Host Mobility	10
3.6. Forwarding Table Scalability	10
3.6.1. MAC Table Reduction on Data Center Switches	10
3.6.2. PE Router FIB Reduction	11
3.6.3. PE Router RIB Reduction	12
3.7. ARP/ND Cache Table Scalability on Default Gateways	14
3.8. ARP/ND and Unknown Uncast Flood Avoidance	14
3.9. Path Optimization	14
4. Considerations for Non-IP traffic	15
5. Security Considerations	15
6. IANA Considerations	15
7. Acknowledgements	15
8. References	15

8.1. Normative References	15
8.2. Informative References	15
Authors' Addresses	16

1. Introduction

For business continuity purposes, Virtual Machine (VM) migration across data centers is commonly used in those situations such as data center maintenance, data center migration, data center consolidation, data center expansion, and data center disaster avoidance. It's generally admitted that IP renumbering of servers (i.e., VMs) after the migration is usually complex and costly at the risk of extending the business downtime during the process of migration. To allow the migration of a VM from one data center to another without IP renumbering, the subnet on which the VM resides needs to be extended across these data centers.

In Infrastructure-as-a-Service (IaaS) cloud data center environments, to achieve subnet extension across multiple data centers in a scalable way, the following requirements SHOULD be considered for any data center interconnect solution:

1) VPN Instance Space Scalability

In a modern cloud data center environment, thousands or even tens of thousands of tenants could be hosted over a shared network infrastructure. For security and performance isolation purposes, these tenants need to be isolated from one another. Hence, the data center interconnect solution SHOULD be capable of providing a large enough Virtual Private Network (VPN) instance space for tenant isolation.

2) Forwarding Table Scalability

With the development of server virtualization technologies, a single cloud data center containing millions of VMs is not uncommon. This number already implies a big challenge for data center switches, especially for core/aggregation switches, from the perspective of forwarding table scalability. Provided that multiple data centers of such scale were interconnected at layer2, this challenge would be even worse. Hence an ideal data center interconnect solution SHOULD prevent the forwarding table size of data center switches from growing by folds as the number of data centers to be interconnected increases. Furthermore, if any kind of L2VPN or L3VPN technologies is used for interconnecting data centers, the scale of forwarding tables on PE routers SHOULD be taken into consideration as well.

3) ARP/ND Cache Table Scalability on Default Gateways

[RFC6820] notes that the Address Resolution Protocol (ARP)/Neighbor Discovery (ND) cache tables maintained by data center default gateways in cloud data centers can raise both scalability and security issues. Therefore, an ideal data center interconnect solution SHOULD prevent the ARP/ND cache table size from growing by multiples as the number of data centers to be connected increases.

4) ARP/ND and Unknown Unicast Flood Suppression or Avoidance

It's well-known that the flooding of Address Resolution Protocol (ARP)/Neighbor Discovery (ND) broadcast/multicast and unknown unicast traffic within a large Layer2 network are likely to affect performances of networks and hosts. As multiple data centers each containing millions of VMs are interconnected together across the Wide Area Network (WAN) at layer2, the impact of flooding as mentioned above will become even worse. As such, it becomes increasingly desirable for data center operators to suppress or even avoid the flooding of ARP/ND broadcast/multicast and unknown unicast traffic across data centers.

5) Path Optimization

A subnet usually indicates a location in the network. However, when a subnet has been extended across multiple geographically dispersed data center locations, the location semantics of such subnet is not retained any longer. As a result, the traffic from a cloud user (i.e., a VPN user) which is destined for a given server located at one data center location of such extended subnet may arrive at another data center location firstly according to the subnet route, and then be forwarded to the location where the service is actually located. This suboptimal routing would obviously result in the unnecessary consumption of the bandwidth resources which are intended for data center interconnection. Furthermore, in the case where the traditional VPLS technology [RFC4761, RFC4762] is used for data center interconnect and default gateways of different data center locations are configured within the same virtual router redundancy group, the returning traffic from that server to the cloud user may be forwarded at layer2 to a default gateway located at one of the remote data center premises, rather than the one placed at the local data center location. This suboptimal routing would also unnecessarily consume the bandwidth resources which are intended for data center interconnect.

This document describes a L3VPN-based subnet extension solution referred to as Virtual Subnet (VS), which can meet all of the

requirements of cloud data center interconnect as described above. Since VS mainly reuses existing technologies including BGP/MPLS IP VPN [RFC4364] and ARP/ND proxy [RFC925][RFC1027][RFC4389], it allows those service providers offering IaaS public cloud services to interconnect their geographically dispersed data centers in a much scalable way, and more importantly, data center interconnection design can rely upon their existing MPLS/BGP IP VPN infrastructures and their experiences in the delivery and the operation of MPLS/BGP IP VPN services.

Although Virtual Subnet is described as a data center interconnection solution in this document, there is no reason to assume that this technology couldn't be used within data centers.

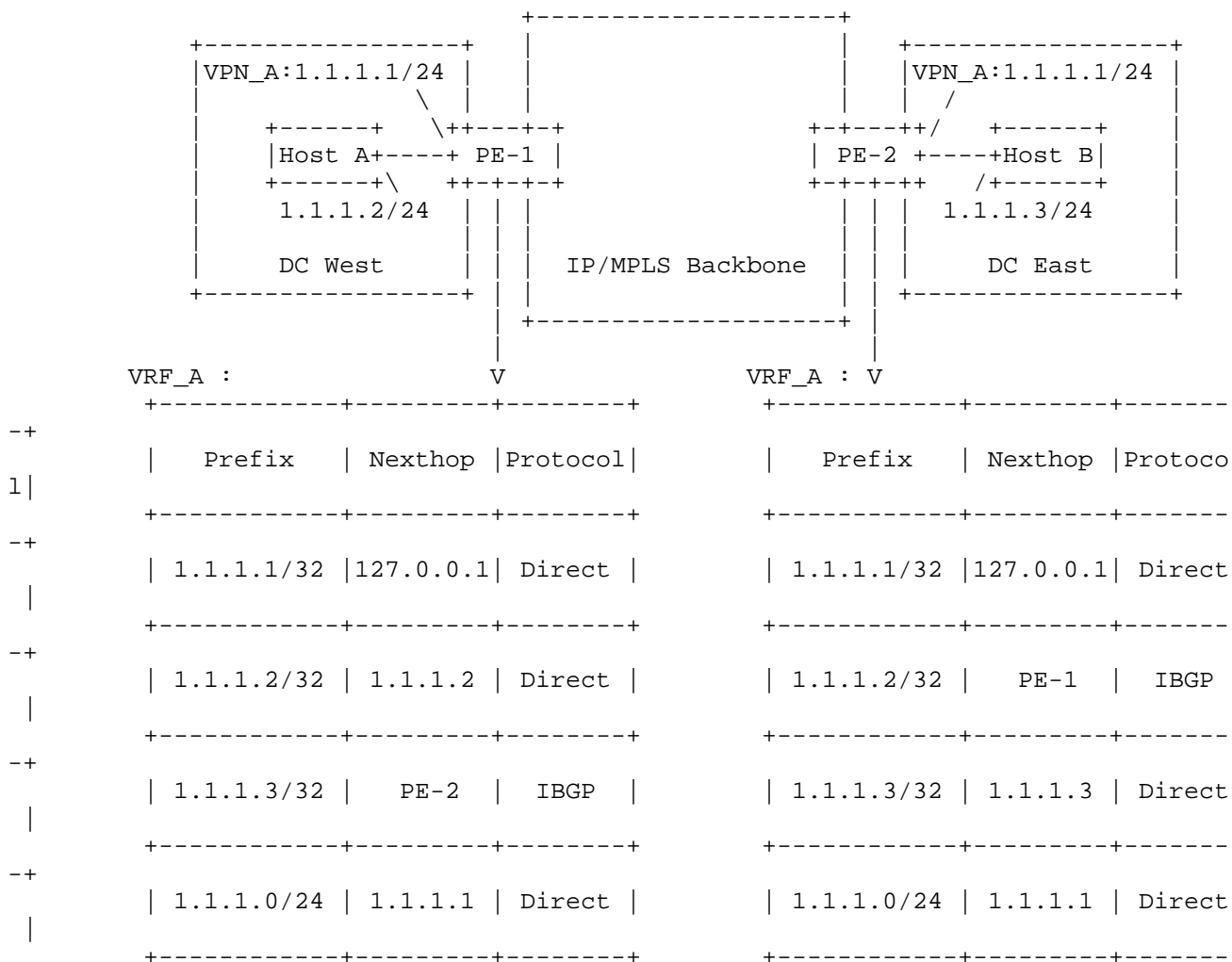
2. Terminology

This memo makes use of the terms defined in [RFC4364], [RFC2338] [MVPN] and [VA-AUTO].

3. Solution Description

3.1. Unicast

3.1.1. Intra-subnet Unicast



-+

Figure 1: Intra-subnet Unicast Example

Xu, et al.

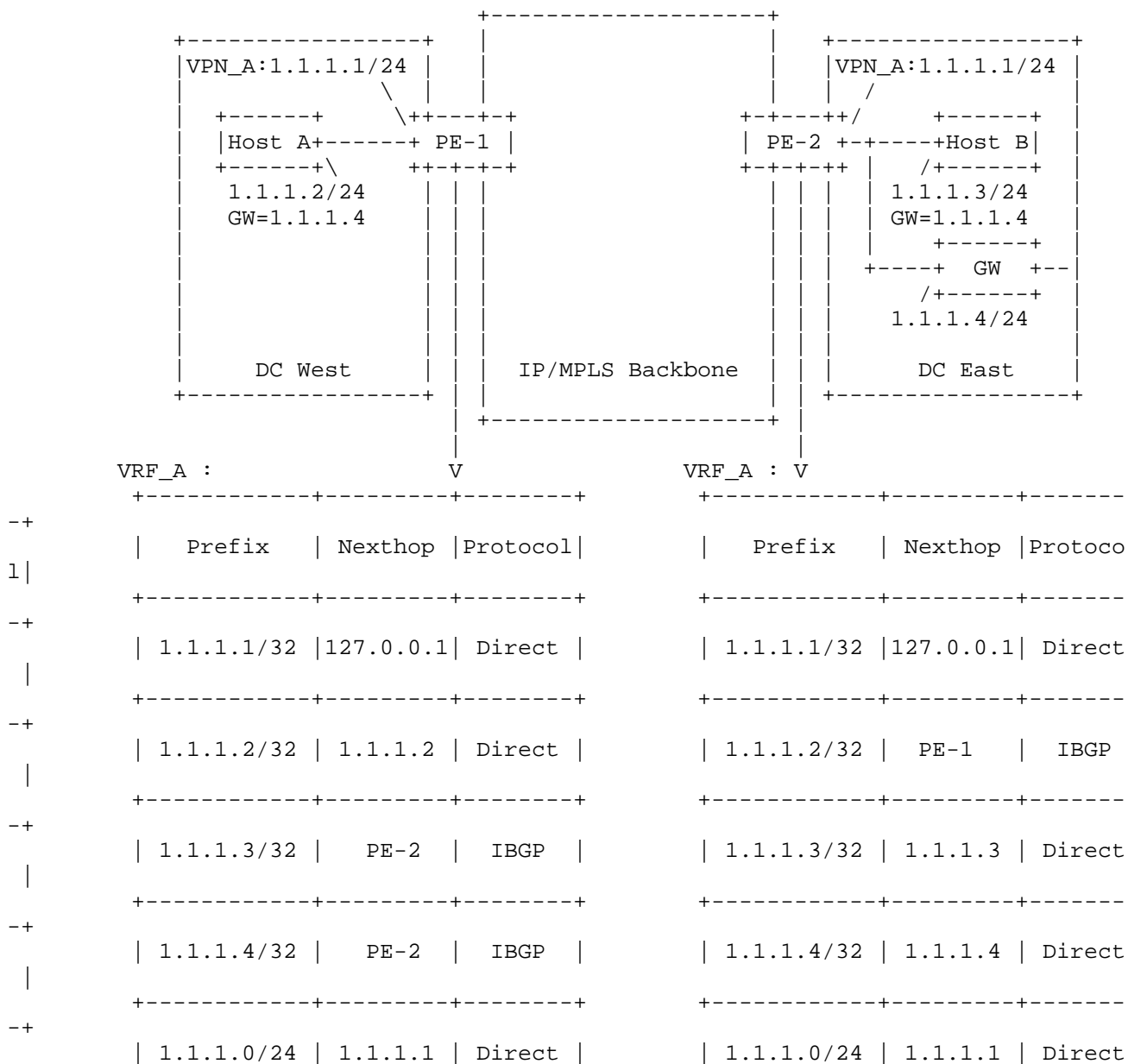
Expires January 15, 2014

[Page 6]

As shown in Figure 1, two CE hosts (i.e., Hosts A and B) belonging to the same subnet (i.e., 1.1.1.0/24) are located at different data centers (i.e., DC West and DC East) respectively. PE routers (i.e., PE-1 and PE-2) which are used for interconnecting these two data centers create host routes for their local CE hosts respectively and then advertise them via L3VPN signaling. Meanwhile, ARP proxy is enabled on VRF attachment circuits of these PE routers.

Now assume host A sends an ARP request for host B before communicating with host B. Upon receiving the ARP request, PE-1 acting as an ARP proxy returns its own MAC address as a response. Host A then sends IP packets for host B to PE-1. PE-1 tunnels such packets towards PE-2 which in turn forwards them to host B. Thus, hosts A and B can communicate with each other as if they were located within the same subnet.

3.1.2. Inter-subnet Unicast



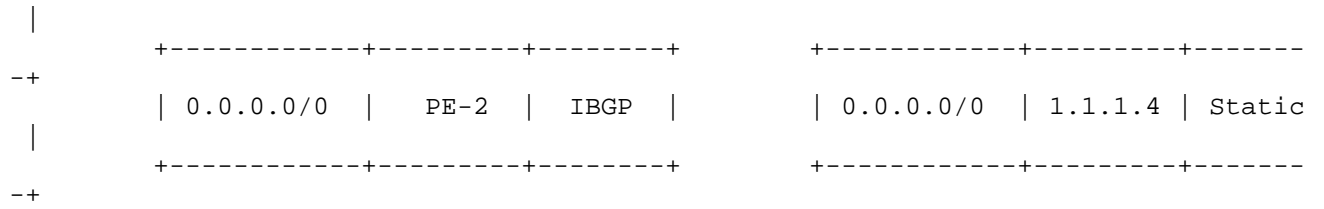
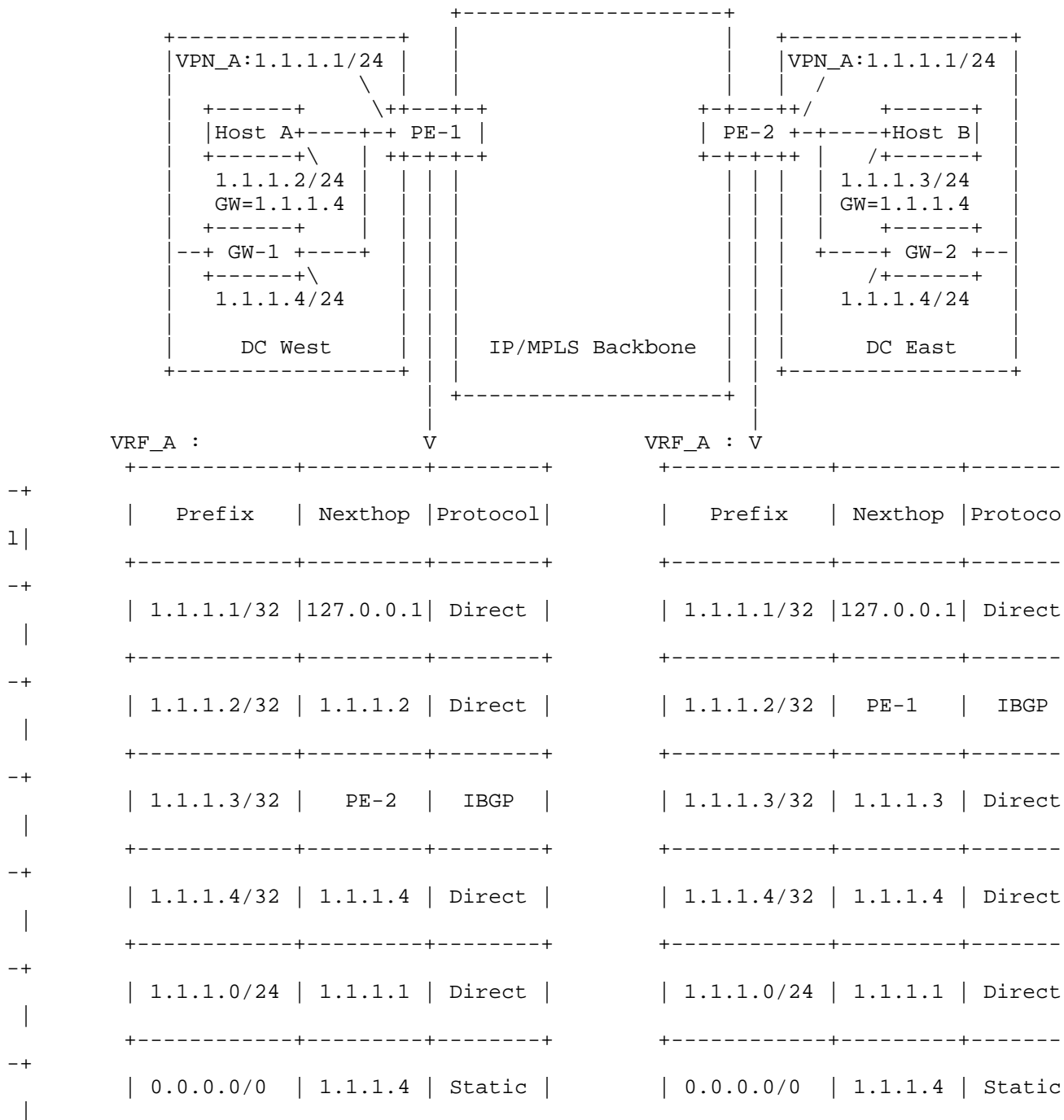


Figure 2: Inter-subnet Unicast Example (1)

As shown in Figure 2, only one data center (i.e., DC East) is deployed with a default gateway (i.e., GW). PE-2 which is connected to GW would either be configured with or learn from GW a default route with next-hop being pointed to GW. Meanwhile, this route is distributed to other PE routers (i.e., PE-1) as per normal [RFC4364] operation. Assume host A sends an ARP request for its default gateway (i.e., 1.1.1.4) prior to communicating with a destination host outside of its subnet. Upon receiving this ARP request, PE-1 acting as an ARP proxy returns its own MAC address as a response. Host A then sends a packet for Host B to PE-1. PE-1 tunnels such packet towards PE-2 according to the default route learnt from PE-2, which in turn forwards that packet to GW.



+-----+-----+-----+ +-----+-----+-----+
-+

Figure 3: Inter-subnet Unicast Example (2)

As shown in Figure 3, in the case where each data center is deployed with a default gateway, CE hosts will get ARP responses directly from their local default gateways, rather than from their local PE routers when sending ARP requests for their default gateways.

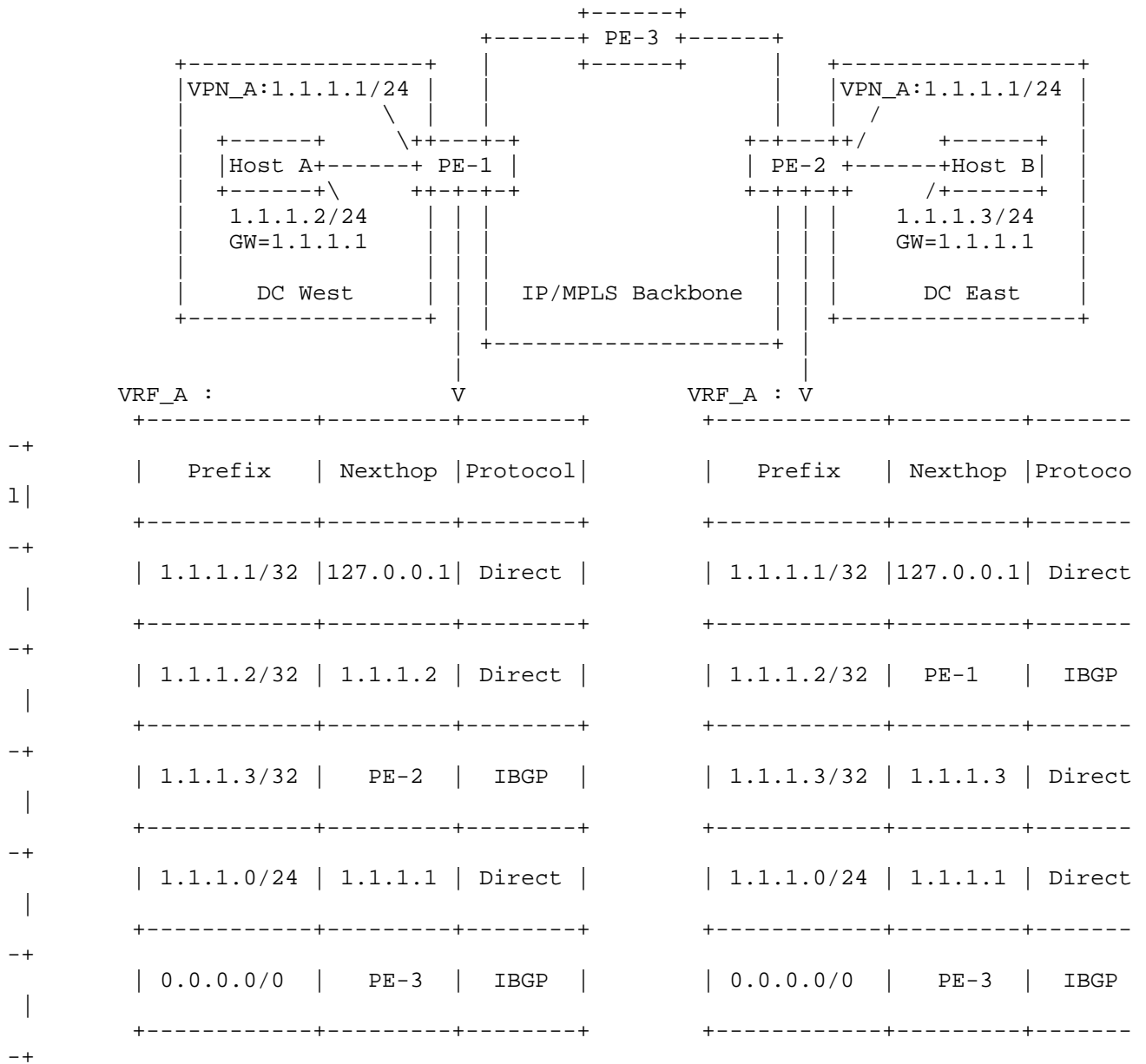


Figure 4: Inter-subnet Unicast Example (3)

Alternatively, as shown in Figure 4, PE routers themselves could be directly configured as default gateways of their locally connected CE hosts as long as these PE routers have routes for outside networks.

3.2. Multicast

To support IP multicast between CE hosts of the same virtual subnet, MVPN technology [MVPN] could be directly reused. For example, PE routers attached to a given VPN join a default provider multicast distribution tree which is dedicated for that VPN. Ingress PE routers, upon receiving multicast packets from their local CE hosts, forward them towards remote PE routers through the corresponding default provider multicast distribution tree.

More details about how to support multicast and broadcast in VS will be explored in a later version of this document.

3.3. CE Host Discovery

PE routers SHOULD be able to discover their local CE hosts and keep the list of these hosts up to date in a timely manner so as to ensure

the availability and accuracy of the corresponding host routes originated from them. PE routers could accomplish local CE host discovery by some traditional host discovery mechanisms using ARP or ND protocols. Furthermore, Link Layer Discovery Protocol (LLDP) described in [802.1AB] or VSI Discovery and Configuration Protocol (VDP) described in [802.1Qbg], or even interaction with the data center orchestration system could also be considered as a means to dynamically discover local CE hosts.

3.4. ARP/ND Proxy

Acting as ARP or ND proxies, PE routers SHOULD only respond to an ARP request or Neighbor Solicitation (NS) message for the target host when there is a corresponding host route in the associated VRF and the outgoing interface of that route is different from the one over which the ARP request or the NS message arrived.

In the scenario where a given VPN site (i.e., a data center) is multi-homed to more than one PE router via an Ethernet switch or an Ethernet network, Virtual Router Redundancy Protocol (VRRP) [RFC5798] is usually enabled on these PE routers. In this case, only the PE router being elected as the VRRP Master is allowed to perform the ARP/ND proxy function.

3.5. CE Host Mobility

During the VM migration process, the PE router to which the moving VM is now attached would create a host route for that CE host upon receiving a notification message of VM attachment while the PE router to which the moving VM was previously attached would withdraw the corresponding host route when receiving a notification message of VM detachment. Meanwhile, the latter PE router could optionally broadcast a gratuitous ARP/ND message on behalf of that CE host with source MAC address being one of its own. In the way, the ARP/ND entry of that moved CE host which has been cached on any local CE host would be updated accordingly.

3.6. Forwarding Table Scalability

3.6.1. MAC Table Reduction on Data Center Switches

In a VS environment, the MAC learning domain associated with a given virtual subnet which has been extended across multiple data centers is partitioned into segments and each segment is confined within a single data center. Therefore data center switches only need to learn local MAC addresses, rather than learning both local and remote MAC addresses.

3.6.2. PE Router FIB Reduction

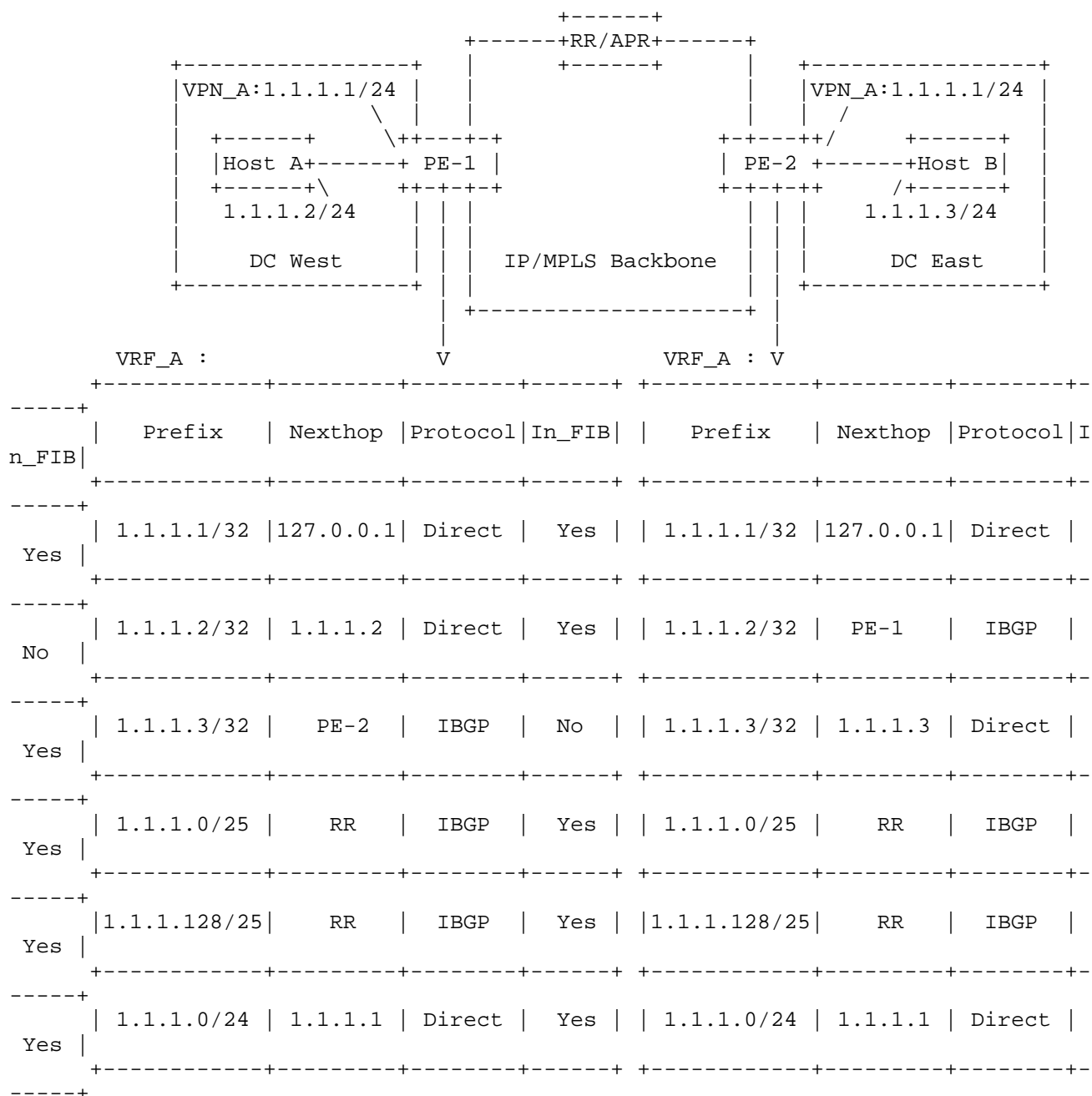


Figure 5: FIB Reduction Example

To reduce the FIB size of PE routers, Virtual Aggregation (VA) [VA-AUTO] technology can be used. Take the VPN instance A shown in Figure 5 as an example, the procedures of FIB reduction are as follows:

- 1) Multiple more specific prefixes (e.g., 1.1.1.0/25 and 1.1.1.128/25) corresponding to the prefix of virtual subnet (i.e., 1.1.1.0/24) are configured as Virtual Prefixes (VPs) and a Route-Reflector (RR) is configured as an Aggregation Point Router (APR) for these VPs. PE routers as RR clients advertise host routes for their own local CE hosts to the RR which in turn, as an APR, installs those host routes into its FIB and then attach the "can-suppress" tag to those

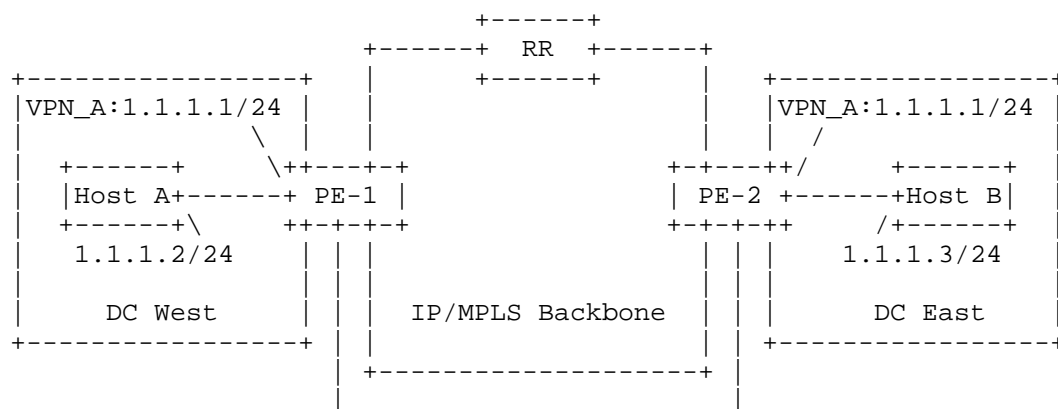
host routes before reflecting them to its clients.

- 2) Those host routes which have been attached with the "can suppress" tag would not be installed into FIBs by clients who are VA-aware since they are not APRs for those host routes. In addition, the RR as an APR would advertise the corresponding VP routes to all of its

clients, and those of which who are VA-aware in turn would install these VP routes into their FIBs.

- 3) Upon receiving a packet from a local CE host, if no matching host route found, the ingress PE router will forward the packet to the RR according to one of the VP routes learnt from the RR, which in turn forwards the packet to the relevant egress PE router according to the host route learnt from that egress PE router. In a word, the FIB table size of PE routers can be greatly reduced at the cost of path stretch. Note that in the case where the RR is not available for transferring L3VPN traffic between PE routers for some reason (e.g., the RR is implemented on a server, rather than a router), the APR function could actually be performed by a given PE router other than the RR as long as that PE router has installed all host routes belonging to the virtual subnet into its FIB. Thus, the RR only needs to attach a "can-suppress" tag to the host routes learnt from its clients before reflecting them to the other clients. Furthermore, PE routers themselves could directly attach the "can-suppress" tag to those host routes for their local CE hosts before distributing them to remote peers as well.
- 4) Provided a given local CE host sends an ARP request for a remote CE host, the PE router that receives such request will install the host route for that remote CE host into its FIB, in case there is a host route for that CE host in its RIB and has not yet been installed into the FIB. Therefore, the subsequent packets destined for that remote CE host will be forwarded directly to the egress PE router. To save the FIB space, FIB entries corresponding to remote host routes which have been attached with "can-suppress" tags would expire if they have not been used for forwarding packets for a certain period of time.

3.6.3. PE Router RIB Reduction



Internet-Draft	Virtual Subnet			July 2013			
VRF_A :	V			VRF_A : V			
	+	+	+	+	+	+	+
1		Prefix	Nexthop Protocol		Prefix	Nexthop Protoco	
	+	+	+	+	+	+	+
		1.1.1.1/32	127.0.0.1 Direct		1.1.1.1/32	127.0.0.1 Direct	
	+	+	+	+	+	+	+
		1.1.1.2/32	1.1.1.2 Direct		1.1.1.3/32	1.1.1.3 Direct	
	+	+	+	+	+	+	+
		1.1.1.0/25	RR IBGP		1.1.1.0/25	RR IBGP	
	+	+	+	+	+	+	+
		1.1.1.128/25	RR IBGP		1.1.1.128/25	RR IBGP	
	+	+	+	+	+	+	+
		1.1.1.0/24	1.1.1.1 Direct		1.1.1.0/24	1.1.1.1 Direct	
	+	+	+	+	+	+	+

Figure 6: RIB Reduction Example

To reduce the RIB size of PE routers, BGP Outbound Route Filtering (ORF) mechanism is used to realize on-demand route announcement. Take the VPN instance A shown in Figure 6 as an example, the procedures of RIB reduction are as follows:

- 1) PE routers as RR clients advertise host routes for their local CE hosts to a RR which however doesn't reflect these host routes by default unless it receives explicit ORF requests for them from its clients. The RR is configured with routes for more specific subnets (e.g., 1.1.1.0/25 and 1.1.1.128/25) corresponding to the virtual subnet (i.e., 1.1.1.0/24) with next-hop being pointed to Null0 and then advertises these routes to its clients via BGP.
- 2) Upon receiving a packet from a local CE host, if no matching host route found, the ingress PE router will forward the packet to the RR according to one of the subnet routes learnt from the RR, which in turn forwards the packet to the relevant egress PE router according to the host route learnt from that egress PE router. In a word, the RIB table size of PE routers can be greatly reduced at the cost of path stretch.
- 3) Just as the approach mentioned in section 3.6.2, in the case where the RR is not available for transferring L3VPN traffic between PE routers for some reason, a PE router other than the RR could advertise the more specific subnet routes as long as that PE router has installed all host routes belonging to that virtual subnet into its FIB.
- 4) Provided a given local CE host sends an ARP request for a remote CE host, the ingress PE router that receives such request will request the corresponding host route from its RR by using the ORF

mechanism (e.g., a group ORF containing Route-Target (RT) and prefix information) in case there is no host route for that CE host in its RIB yet. Once the host route for the remote CE host is

learned from the RR, the subsequent packets destined for that CE host would be forwarded directly to the egress PE router. Note that the RIB entries of remote host routes could expire if they have not been used for forwarding packets for a certain period of time. Once the expiration time for a given RIB entry is approaching, the PE router would notify its RR not to pass the updates for corresponding host route by using the ORF mechanism.

3.7. ARP/ND Cache Table Scalability on Default Gateways

In case where data center default gateway functions are implemented on PE routers of the VS as shown in Figure 4, since the ARP/ND cache table on each PE router only needs to contain ARP/ND entries of local CE hosts, the ARP/ND cache table size will not grow as the number of data centers to be connected increases.

3.8. ARP/ND and Unknown Unicast Flood Avoidance

In VS, the flooding domain associated with a given virtual subnet that has been extended across multiple data centers, has been partitioned into segments and each segment is confined within a single data center. Therefore, the performance impact on networks and servers caused by the flooding of ARP/ND broadcast/multicast and unknown unicast traffic is alleviated.

3.9. Path Optimization

Take the scenario shown in Figure 4 as an example, to optimize the forwarding path for traffic between cloud users and cloud data centers, PE routers located at cloud data centers (i.e., PE-1 and PE-2), which are also data center default gateways, propagate host routes for their local CE hosts respectively to remote PE routers which are attached to cloud user sites (i.e., PE-3).

As such, traffic from cloud user sites to a given server on the virtual subnet which has been extended across data centers would be forwarded directly to the data center location where that server resides, since traffic is now forwarded according to the host route for that server, rather than the subnet route.

Furthermore, for traffic coming from cloud data centers and forwarded to cloud user sites, each PE router acting as a default gateway would forward the traffic received from its local CE hosts according to the best-match route in the corresponding VRF. As a result, traffic from data centers to cloud user sites is forwarded along the optimal path as well.

4. Considerations for Non-IP traffic

Although most traffic within and across data centers is IP traffic, there may still be a few legacy clustering applications which rely on non-IP communications (e.g., heartbeat messages between cluster nodes). To support those few non-IP traffic (if present) in the Virtual Subnet solution, the approach following the idea of "route all IP traffic, bridge non-IP traffic" could be considered as an enhancement to the original Virtual Subnet solution.

Note that more and more cluster vendors are offering clustering applications based on Layer 3 interconnection.

5. Security Considerations

This document doesn't introduce additional security risk to BGP/MPLS L3VPN, nor does it provide any additional security feature for BGP/MPLS L3VPN.

6. IANA Considerations

There is no requirement for any IANA action.

7. Acknowledgements

Thanks to Dino Farinacci, Himanshu Shah, Nabil Bitar, Giles Heron, Ronald Bonica, Monique Morrow, Rajiv Asati and Eric Osborne for their valuable comments and suggestions on this document.

8. References

8.1. Normative References

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

8.2. Informative References

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

[MVPN] Rosen. E and Aggarwal. R, "Multicast in MPLS/BGP IP VPNs", draft-ietf-l3vpn-2547bis-mcast-10.txt, Work in Progress, January 2010.

- [VA-AUTO] Francis, P., Xu, X., Ballani, H., Jen, D., Raszuk, R., and L. Zhang, "Auto-Configuration in Virtual Aggregation", draft-ietf-grow-va-auto-05.txt, Work in Progress, December 2011.
- [RFC925] Postel, J., "Multi-LAN Address Resolution", RFC-925, USC Information Sciences Institute, October 1984.
- [RFC1027] Smoot Carl-Mitchell, John S. Quarterman, "Using ARP to Implement Transparent Subnet Gateways", RFC 1027, October 1987.
- [RFC4389] D. Thaler, M. Talwar, and C. Patel, "Neighbor Discovery Proxies (ND Proxy) ", RFC 4389, April 2006.
- [RFC5798] S. Nadas., "Virtual Router Redundancy Protocol", RFC 5798, March 2010.
- [RFC4761] Kompella, K. and Y. Rekhter, "Virtual Private LAN Service (VPLS) Using BGP for Auto-Discovery and Signaling", RFC 4761, January 2007.
- [RFC4762] Lasserre, M. and V. Kompella, "Virtual Private LAN Service (VPLS) Using Label Distribution Protocol (LDP) Signaling", RFC 4762, January 2007.
- [802.1AB] IEEE Standard 802.1AB-2009, "Station and Media Access Control Connectivity Discovery", September 17, 2009.
- [802.1Qbg] IEEE Draft Standard P802.1Qbg/D2.0, "Virtual Bridged Local Area Networks -Amendment XX: Edge Virtual Bridging", Work in Progress, December 1, 2011.
- [RFC6820] Narten, T., Karir, M., and I. Foo, "Problem Statement for ARMD", RFC 6820, January 2013.

Authors' Addresses

Xiaohu Xu
Huawei Technologies,
Beijing, China.
Phone: +86 10 60610041
Email: xuxiaohu@huawei.com

Susan Hares
Email: shares@ndzh.com

Internet-Draft

Virtual Subnet

July 2013

Yongbing Fan
Guangzhou Institute, China Telecom
Guangzhou, China.
Phone: +86 20 38639121
Email: fanyb@gsta.com

Christian Jacquenet
France Telecom
Rennes
France
Email: christian.jacquenet@orange.com

INTERNET-DRAFT
Intended Status: Proposed Standard
Expires: January 13, 2014

Mingui Zhang
Peng Zhou
Huawei
July 12, 2013

Label Sharing for Fast PE Protection
draft-zhang-l3vpn-label-sharing-00.txt

Abstract

This document describes a method to be used by Service Providers to provide fast protection of VPN connections for a CE. Egress PEs in a redundant group always assign the same label for VPN routes from a VRF. These egress PEs create a BGP virtual Next Hop (vNH) in the domain of the IP/MPLS backbone network as an agent of the CE router. Primary and backup tunnels terminated at the vNH are set up by the BGP/MPLS IP VPN based on IGP FRR. If the primary egress PE fails, the backup egress PEs can recognize the "shared" VPN route label and deliver the failure affected packets accordingly.

Status of this Memo

This Internet-Draft is submitted to IETF in full conformance with the provisions of BCP 78 and BCP 79.

Internet-Drafts are working documents of the Internet Engineering Task Force (IETF), its areas, and its working groups. Note that other groups may also distribute working documents as Internet-Drafts.

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."

The list of current Internet-Drafts can be accessed at
<http://www.ietf.org/lid-abstracts.html>

The list of Internet-Draft Shadow Directories can be accessed at
<http://www.ietf.org/shadow.html>

Copyright and License Notice

Copyright (c) 2013 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	3
1.1. Conventions used in this document	3
1.2. Terminology	3
2. The Label Sharing Method	3
2.1. The Virtual Next Hop	4
2.1.1. Generating OSPF LSAs	5
2.1.2. Generating ISIS LSPs	7
2.2. Link Costs Set Up for IGP FRR	9
2.3 Label Assignment and Processing	10
2.3.1. The VPN Route Label	10
2.3.2. The Tunnel Label	10
3. Security Considerations	10
4. IANA Considerations	11
5. References	11
5.1. Normative References	11
5.2. Informative References	11
Author's Addresses	12

1. Introduction

For the sake of reliability, ISPs usually connect one CE to multiple PEs. When the primary egress PE fails, a backup egress PE continues to offer VPN connectivity to the CE. If local repair is performed by the upstream neighbor of the primary egress PE on the data path, it's possible to achieve 50msec switchover.

VPN routes learnt from CEs are distributed by egress PEs to ingress PEs that need to know these VPN routes. Egress PEs in a redundant group (RG) MUST allocate the same VPN route label for routes of the same VPN. When the primary egress PE fails, data packets are redirected to a backup egress PE by the PLR router, the backup PE can recognize the VPN route label in these data packets and deliver them correctly. The method developed in this document is so called "Label Sharing for Fast PE Protection". This method requires only software update on egress PE routers while their data plane remains unchanged.

This document supposes BGP/MPLS IP VPN is deployed on the backbone and Label Distribution Protocol (LDP) is used as the tunneling technology. Through generating virtual LSAs/LSPs in OSPF/ISIS, egress PEs in an RG create a virtual router (the vNH) in the IP/MPLS backbone to represent the CE router. When the VPN route is distributed, those egress PEs use vNH as the "BGP next hop". The vNH will be treated as the egress point of the tunnel by other routers. Metrics for the virtual links attached to the vNH are set up in a way that the IGP FRR mechanism defined in [LFA] can be leveraged to achieve local protection.

1.1. Conventions used in this document

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].

1.2. Terminology

VRF: Virtual Routing and Forwarding table
FRR: Fast ReRouting
PLR: Point of Local Repair
LFA: loop-free alternate

2. The Label Sharing Method

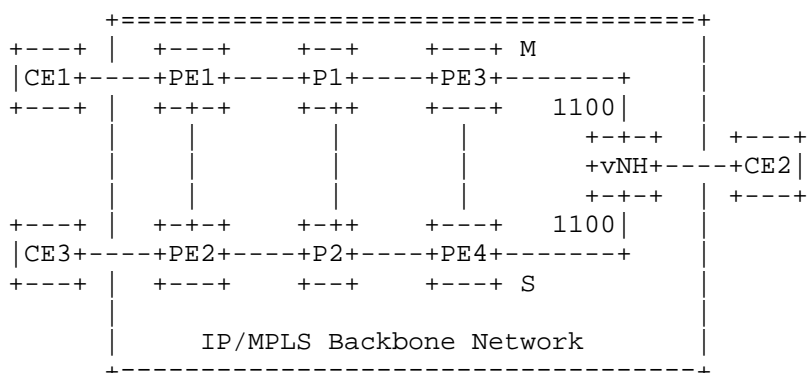


Figure 2.1: Egress PE routers share the same VPN route label.

A CE router is usually connected to multiple PE routers of the IP/MPLS backbone network for the sake of reliability. Figure 2.1 shows such a scenario. In this document, PE1 and PE2 are defined as ingress routers and PE3 and PE4 are defined as egress routers. Suppose PE3 is the primary PE while PE4 is the backup egress PE. In this document, we suppose there are two PEs in one RG. It's possible to expand the method to support more than two PEs in one RG, though it is out the scope of this document.

Those egress PE routers may discover each other as in the same RG from the CE routes learning process which can be a dynamic routing algorithm or a static routing configuration [RFC4364].

2.1. The Virtual Next Hop

Egress PEs create a vNH router in IGP to represent the set of CEs dual-homed to the same egress PEs in the Service Provider's backbone. The PE with the highest priority in the RG determines the loopback IP address for the vNH. This loopback IP address can be configured manually or automatically. The SystemID of the vNH under ISIS is composed based on this loopback IP address. The router LSA/LSP for the vNH is generated by the egress PE with the highest priority. This router LSA/LSP also includes the the outgoing links of the vNH. For the incoming links of the vNH, all egress PEs need include these P2P adjacencies in their router LSAs/LSPs.

Egress PEs may create multiple vNHs for one CE. Then multiple tunnels can be set up from ingress PEs to the vNHs. Ingress PEs can choose from these tunnel routes to achieve load balance for the CE.

The overload mode MUST be set so that the rest routers in the network will not route transit traffic through the vNH. In OSPF, the overload

LS type
1

Link State ID
Same as the Advertising Router

Advertising Router
The Router ID of the vNH.

LS sequence number
As defined in [RFC2328].

LS checksum
As defined and computed in [RFC2328].

length
The length in bytes of the LSA. This includes the 20 byte LSA header. (As defined and computed in [RFC2328].)

VEB
As defined in [RFC2328], set its value to 000.

#links
The number of router links described in this LSA. It equals to the number of Egress PEs in the RG.

The following fields are used to describe each router link connected to an egress PE. Each router link is typed as Type 1 Point-to-point connection to another router.

Link ID
The Router ID of one of the egress PEs in the RG.

Link Data
It specifies the interface's MIB-II [RFC1213] ifIndex value. It ranges between 1 and the value of ifNumber. The ifNumber equals to the number of the PEs in the RG. The PE with the highest priority sorts the PEs according to their unsigned integer Router ID in the ascend order and assigns the ifIndex for each.

Type
Value 1 is used, indicating the router link is a point-to-point connection to another router.

TOS
This field is set to 0 for this version.

Metric

It is set to 0xFFFF.

The fields used here to describe the virtual router links are also included in the Router-LSA of each egress PEs. The Link ID is replaced with the Router ID of the vNH. The Link Data specifies the interface's MIB-II [RFC1213] ifIndex value. The "Metric" field is set as defined in Section 2.2.

2.1.2. Generating ISIS LSPs

The primary egress PE generates the following level 1 LSP to describe the vNH node.

	No. of octets
+-----+ Intradomain Routeing Protocol Discriminator	1
+-----+ Length Indicator	1
+-----+ Version/Protocol ID Extension	1
+-----+ ID Length	1
+-----+ R R R PDU Type	1
+-----+ Version	1
+-----+ Reserved	1
+-----+ Maximum Area Address	1
+-----+ PDU Length	2
+-----+ Remaining Lifetime	2
+-----+ LSP ID	ID Length + 2
+-----+ Sequence Number	4
+-----+ Checksum	2
+-----+ P ATT LSPDBOL IS Type	1
+-----+ : Variable Length Fields :	Variable
+-----+	

Intradomain Routeing Protocol Discriminator - 0x83 (as defined in [ISIS])

Length Indicator - Length of the Fixed Header in octets

Version/Protocol ID Extension - 1

ID Length - As defined in [ISIS]

PDU Type (bits 1 through 5) - 18

Version - 1

Reserved - transmitted as zero, ignored on receipt

Maximum Area Address - same as the primary egress PE

PDU Length - Entire Length of this PDU, in octets, including the header.

Remaining Lifetime - Number of seconds before this LSP is considered expired. (Set to 0x384 by default.)

LSP ID - the system ID of the source of the LSP. It is structured as follows:

+-----+	
Source ID	6
+-----+	
Pseudonode ID	1
+-----+	
LSP Number	1
+-----+	

Source ID - SystemID of the vNH

Pseudonode ID - Transmitted as zero

LSP Number - Fragment number

Sequence Number - sequence number of this LSP (as defined in [ISIS])

Checksum - As defined and computed in [ISIS]

P - Bit 8 - 0

ATT - Bit 7-4 - 0

LSDBOL - Bit 3 - 1

IS Type - Bit 1 and 2 - bit 1 set, indicating the vNH is a Level 1 Intermediate System

In the Variable Length Field, each link outgoing from the vNH to an egress PE is depicted by a Type #22 Extended Intermediate System Neighbors TLV [RFC5305]. The egress PE is identified by the 6 octets SystemID plus one octet of all-zero pseudonode number. The 3 octets metric is set as that in Section 2.2. None sub-TLVs is used by this version, therefore the value of the one octet length of sub-TLVs is 0. The Type #22 TLV requires 11 octets.

The Type #22 TLV is also included in the LSP of each egress PE to depict the incoming link of the vNH. Only the 6 octets SystemID is replaced with the SystemID of the vNH.

2.2. Link Costs Set Up for IGP FRR

Tunnel LSPs are set up based on IGP routes through LDP signaling. If the IGP costs for the links between egress PEs and the vNH can be set up in a way that one egress PE appears on the primary path while other PE(s) appears on the backup path, the PLR can make use of the multiple egress PEs to achieve fast failure protection. Suppose [LFA] is being used as the IGP FRR mechanism, the link weights can be set up according to the following rule.

1. This document supposes bidirectional link weights are being used. Assume the weight for the link between PE3 and vNH is "M" and the weight for the link between PE4 and vNH is "S". The weight for the link between PE3 and PE4 is C34.

2. Px is a neighbor of PE3. This Px will act as the PLR. Suppose Pxy is Px's neighbor with the shortest path to PE4, after PE3 is removed from the topology. The cost of this path is Sxy4.

3. Add PE3 back to the topology. The cost of the path from Pxy to PE3 is Sxy3.

4. "M" and "S" can be set up as long as the following two equations hold.

$$\text{eq1: } Sxy4+S < Sxy3+M$$

$$\text{eq2: } C34+S > M$$

Although this document designs the method based on [LFA] which is widely deployed, other IGP FRR mechanisms can also be utilized to

achieve the protection. For example, [MRT] is applicable regardless of how the link weights are set up.

2.3 Label Assignment and Processing

2.3.1. The VPN Route Label

Egress PEs use BGP to distribute to ingress PEs the routes that they have learnt from CEs [RFC4364]. When egress PEs distribute the routes of the VPN that the CE is in, they MUST assign the same "VPN route label" for one VPN (per VRF label assignment). This label will become the first label of a data packet. The IP address of the vNH is used as the "BGP next hop". For example, in Figure 2.1, both PE3 and PE4 use 1100 as the VPN route label for the routes learnt from CE2.

Suppose PE3 fails and the packet with VPN route label 1100 is redirected to PE4, PE4 recognizes 1100 as the VPN route label it assigned for the VPN that the CE is in. As specified in Section 5 of [RFC4364], PE4 will be able to determine, the attachment circuit over which the packet should be transmitted (to the CE) as well as the data link layer header for that interface. It need to lookup the packet's destination address in the VRF identified by the VPN route label 1100.

When we speak of a PE fails, it may also means that a link to the PE on the primary tunnel fails. In general, we can say that a primary PE fails means that this PE becomes unreachable via its upstream neighbor on the primary tunnel.

The shared label may be manually configured or negotiated through signaling between egress PEs. In [LS-ICCP], application TLVs are defined for [ICCP] to achieve such kind of signaling.

2.3.2. The Tunnel Label

This document supposes Label Distribution Protocol is being used as the tunneling technology. The LDP LSP tunnel follows a IGP route from ingress PEs to the vNH. The backup path to vNH can be calculated according to IGP FRR mechanism, such as [MRT] and [LFA].

The ingress PE tunnels the data packet through the backbone network using the "tunnel label" as the second entry of the label stack. The "VPN route label" is not visible again until the MPLS packet reaches the egress PE. The egress PE need pop the second label and deliver the packet according to the "VPN route label".

3. Security Considerations

This document raises no new security issues.

4. IANA Considerations

No requirements for IANA.

5. References

5.1. Normative References

- [LFA] Filsfils, C., Ed., Francois, P., Ed., Shand, M., Decraene, B., Uttaro, J., Leymann, N., and M. Horneffer, "Loop-Free Alternate (LFA) Applicability in Service Provider (SP) Networks", RFC 6571, June 2012.
- [RFC4364] Rosen, E. and Y. Rekhter, "BGP/MPLS IP Virtual Private Networks (VPNs)", RFC 4364, February 2006.
- [ICCP] L. Martini, S. Salam, et al, "Inter-Chassis Communication Protocol for L2VPN PE Redundancy", draft-ietf-pwe3-iccp-11.txt, work in progress.
- [ISIS] ISO, "Intermediate system to Intermediate system routing information exchange protocol for use in conjunction with the Protocol for providing the Connectionless-mode Network Service (ISO 8473)," ISO/IEC 10589:2002.
- [RFC2328] Moy, J., "OSPF Version 2", STD 54, RFC 2328, April 1998.
- [RFC1213] McCloghrie, K. and M. Rose, "Management Information Base for Network Management of TCP/IP-based internets:MIB-II", STD 17, RFC 1213, March 1991.
- [LS-ICCP] M. Zhang, P. Zhou, "ICCP Application TLVs for VPN Route Label Sharing", draft-zhang-pwe3-iccp-label-sharing-00.txt, work in progress

5.2. Informative References

- [MRT] A. Atlas, Ed., R. Kebler, et al, "An Architecture for IP/LDP Fast-Reroute Using Maximally Redundant Trees", draft-ietf-rtgwg-mrt-frr-architecture-02.txt, work in progress.

Author's Addresses

Mingui Zhang
Huawei Technologies Co., Ltd
Huawei Building, No.156 Beiqing Rd.
Z-park, Shi-Chuang-Ke-Ji-Shi-Fan-Yuan, Hai-Dian District,
Beijing 100095 P.R. China

Email: zhangmingui@huawei.com

Peng Zhou
Huawei Technologies Co., Ltd
Huawei Building, No.156 Beiqing Rd.
Z-park, Shi-Chuang-Ke-Ji-Shi-Fan-Yuan, Hai-Dian District,
Beijing 100095 P.R. China

Email: Jewpon.zhou@huawei.com

Network Working Group
Internet-Draft
Intended status: Standards Track
Expires: October 19, 2013

L. Zheng
Z. Li
Huawei Technologies
B. Parise
Cisco Systems
April 17, 2013

Performance Monitoring Analysis for L3VPN
draft-zheng-l3vpn-pm-analysis-01

Abstract

To perform the measurement of packet loss, delay and other metrics on a particular VPN flow, the egress PE need to tell to which specific ingress VRF a packet belongs to. But for L3VPN, multipoint-to-point or multipoint-to-multipoint (MP2MP) network model applies, flow identifying is a big challenge. This document summarizes the current performance monitoring mechanisms for MPLS networks, and analyzes the challenge for L3VPN performance monitoring. This document also discuss the key points need to be taken in consideration when designing L3VPN performance monitoring mechanisms.

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 October 19, 2013.

Copyright Notice

Copyright (c) 2013 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. Overview of current mechanisms for MPLS networks	3
2.1. Packet Loss and Delay Measurement for MPLS Networks	3
2.2. Profile for MPLS-based Transport Networks	3
3. Challenge for L3VPN Performance Monitoring	4
4. Design Consideration	5
4.1. P2P Connection	5
4.2. Hierarchy L3VPN	6
4.3. Control Plane	6
4.4. Data Plane	6
4.5. MPLS OAM	6
4.6. QoS	6
5. Manageability Consideration	7
6. Security Considerations	7
7. IANA Considerations	7
8. Acknowledgements	7
9. References	7
9.1. Normative References	7
9.2. Informative References	7
Authors' Addresses	8

1. Introduction

Level 3 Virtual Private Network (L3VPN) [RFC4364] service is widely deployed in the production network. It is deployed to provide enterprise interconnection, Voice over IP (VoIP), video, mobile, etc. services. Most of these services are sensitive to the packet loss and delay. The capability to measure and monitor performance metrics for packet loss, delay, as well as related metrics is essential for SLA. The requirement for SLA measurement for MPLS networks has been documented in [RFC4377].

One popular deployment of L3VPN nowadays is in mobile backhaul networks. When deploying MPLS-TP in mobile backhaul network, due to the scalability issue with PW, L3VPN is used either for end-to-end service delivery, or L2VPN and L3VPN hybrid networking. The measurement capability of L3VPN provides operators with greater visibility into the performance characteristics of their networks, and provides diagnostic information in case of performance degradation or failure and helps for fault localization.

To perform the measurement of packet loss, delay and other metrics on a particular VPN flow, the egress PE need to tell to which specific ingress VRF a packet belongs. But for L3VPN, multipoint-to-point or multipoint-to-multipoint (MP2MP) network model applies, flow identifying is a big challenge. This document summarizes the current performance monitoring mechanisms for MPLS networks, and analyzes the challenge for L3VPN performance monitoring. This document also discuss the key points need to be taken in consideration when designing L3VPN performance monitoring mechanisms.

2. Overview of current mechanisms for MPLS networks

2.1. Packet Loss and Delay Measurement for MPLS Networks

[RFC6374] defines procedure and protocol mechanisms to enable the efficient and accurate measurement of packet loss, delay, as well as related metrics in MPLS networks.

The LM protocol can perform two distinct kinds of loss measurement. In inferred mode, it can measure the loss of specially generated test packets (in order to infer the approximate data-plane loss level). In direct mode, it can directly measure data-plane packet loss. Direct measurement provides perfect loss accounting, but may require specialized hardware support and is only applicable to some LSP types. Inferred measurement provides only approximate loss accounting but is generally applicable. There should also be inferred mode and direct mode for LM in the L3VPN environment. This document focuses on the direct mode. The inferred mode is out of scope of the document.

The LM and DM protocols are initiated from a single node, the querier. A query message may be received either by a single node or by multiple nodes; i.e. these protocols provide point-to-point or point-to-multipoint measurement capabilities.

2.2. Profile for MPLS-based Transport Networks

Procedures for the measurement of packet loss, delay, and throughput in MPLS networks are defined in [RFC6374]. [RFC6375] describes a

profile, i.e. a simplified subset, of procedures that suffices to meet the specific requirements of MPLS-based transport networks [RFC5921] as defined in [RFC5860]. This profile is presented for the convenience of implementers who are concerned exclusively with the transport network context.

LM session is externally configured and the values of several protocol parameters can be fixed in advance at the endpoints involved in the session, so that inspection or negotiation of these parameters is not required.

3. Challenge for L3VPN Performance Monitoring

To perform the measurement of packet loss, delay and other metrics on a particular VPN flow, the egress PE need to tell to which specific ingress VRF a packet belongs.

The above mentioned existing mechanisms for MPLS networks provide either point-to-point or point-to-multipoint measurement capabilities. For a specific receiver, it could easily identify a specific flow by the label stack information, when LDP is not used and Penultimate Hop Pop (PHP) function is disabled.

But in the case of L3VPN, multipoint-to-point or multipoint-to-multipoint (MP2MP) network model applies, it makes the flow identifying a big challenge for packets loss and delay measurement. According to the label allocation mechanisms of L3VPN, a private label itself cannot uniquely identify a specific VPN flow. That is, when the egress PE allocates VPN label for a specific prefix of a VPN, the same label will be advertised to all its peers. Given a VPN flow, the egress PE cannot tell which ingress VRF is from based on the private label it carries. As a result, it's not feasible to perform the loss or delay measurement on this flow.

In L3VPN the LSPs may be merged at any intermediate nodes along the LSP (e.g., Label Distribution Protocol (LDP) [RFC5036] based LSP). The egress PE cannot derive a unique identifier of the source PE from label stack. The tunnel label cannot help for flow identification due to the LSP merge.

In L3VPN, the ingress PE could be identified by the tunnel label when TE LSP applies [RFC3209], but the egress PE cannot tell to which specific VRF a packet belongs when extranet (If the various sites in a VPN are owned by different enterprises) exist on ingress PE. Figure 1 shows an example of extranet. In Figure1, Site A,B,C,D all belong to the same VPN-A, but Site C and Site D does not belong to the same enterprise (Site D also belongs to a VPN-B), so different VRFs are maintained for each site on PE3. PE1 assign the same label

L for prefix 10.0.0.1 to PE3 of VPN-A, when it receives the VPN-A flow from PE3, it can not tell the flow is from either VRFC or VRFD by the label stack.

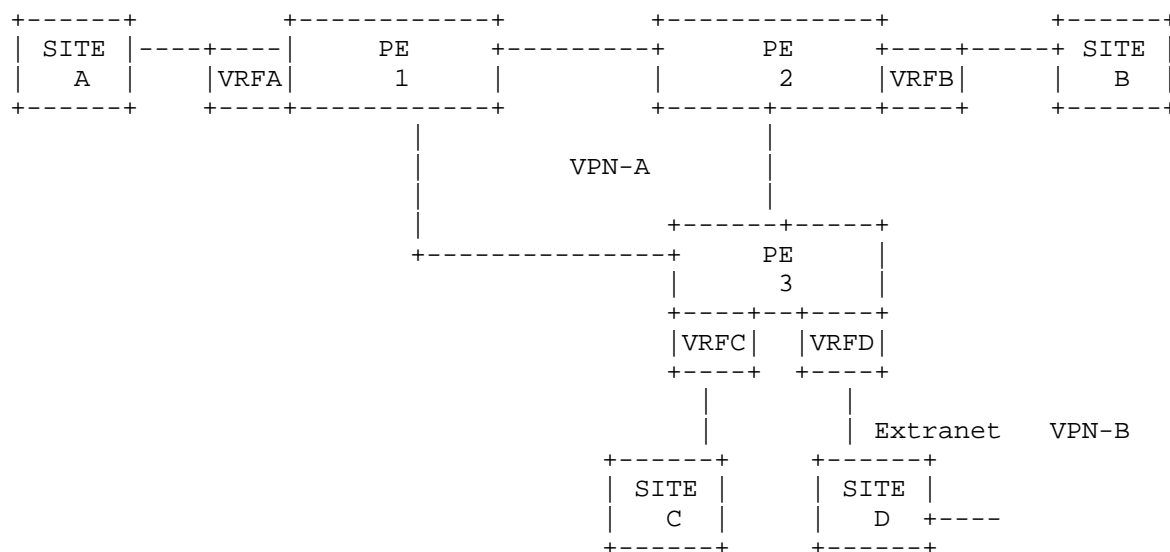


Figure1: Extranet on Ingress PE

The current label allocation mechanism of L3VPN makes the flow identification a big challenge for L3VPN performance monitoring, as a result the current performance monitoring mechanisms for MPLS networks cannot be applied to L3VPN networks. Extension or alteration to current label allocation mechanism is needed to solve the problem.

4. Design Consideration

This section discuss the key points need to be taken in consideration when designing L3VPN performance monitoring mechanism.

4.1. P2P Connection

As analyzed above, to perform the packet loss or delay measurement on a specific VPN flow, it is critical for the egress PE to identify the unique ingress VRF, i.e. to establish the Point-to-Point connection between the two VRFs. Current allocation mechanism may need extension or alteration to help build up the Point-to-Point connection. Once the Point-to-Point connection is built up, current measurement mechanisms may be applied to L3VPN .

Conditions like Penultimate Hop Popping (PHP), Equal-Cost Multi-Path (ECMP) load-balancing and BGP multi-path may make it infeasible for receiving PE to identify the ingress PE. These conditions are out of scope of the mechanism design.

4.2. Hierarchy L3VPN

There are flexible hierarchy L3VPN deployment scenarios such as inter-AS, carrier's carrier, etc[RFC4364]. The mechanism design should take into account these scenarios. Since the document focuses on MPLS transport network which seldom introduces the complex L3VPN scenarios, it is for further research if more complex mechanisms for these hierarchy L3VPN scenarios besides reusing the mechanism for the simple L3VPN scenarios has to be introduced.

4.3. Control Plane

In L3VPN, BGP is used to distribute a particular route, as well as an MPLS label that is mapped to that route [RFC4364]. The label mapping information for a particular route is piggybacked in the same BGP Update message that is used to distribute the route itself. In order to setup the Point-to-Point connection between ingress and egress VRFs the current label distribution mechanism may be altered. For compatibility, this alteration SHOULD NOT change the current label distribution mechanism dramatically.

4.4. Data Plane

Same as for control plane, for compatibility reason, the data plane should as far as be compatible with the current L3VPN forwarding procedure.

4.5. MPLS OAM

[RFC6374], [RFC6375] defines procedure and protocol mechanisms to enable the measurement of packet loss, delay, as well as related metrics in MPLS networks. These mechanisms SHOULD be reasonably reused in L3VPN networks. The addressing of source and destination of Loss Measurement (LM) and Delay Measurement (DM) messages may need to be changed to identify the measured VRF.

4.6. QoS

To perform the packet loss or delay measurement in L3VPN network, either proactive or on-demand, SHOULD NOT impact the customer QoS experience.

5. Manageability Consideration

[RFC6374] describes manageability consideration of packet loss and delay measurement for MPLS network. The defined mechanisms should be reused for L3VPN PM.

6. Security Considerations

This document does not change the security properties of L3VPN.

7. IANA Considerations

This document makes no request to IANA.

8. Acknowledgements

The authors would like to thank XXX for their valuable comments.

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.

9.2. Informative References

[RFC3209] Awduche, D., Berger, L., Gan, D., Li, T., Srinivasan, V., and G. Swallow, "RSVP-TE: Extensions to RSVP for LSP Tunnels", RFC 3209, December 2001.

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

[RFC4377] Nadeau, T., Morrow, M., Swallow, G., Allan, D., and S. Matsushima, "Operations and Management (OAM) Requirements for Multi-Protocol Label Switched (MPLS) Networks", RFC 4377, February 2006.

[RFC5036] Andersson, L., Minei, I., and B. Thomas, "LDP Specification", RFC 5036, October 2007.

[RFC5860] Vigoureux, M., Ward, D., and M. Betts, "Requirements for Operations, Administration, and Maintenance (OAM) in MPLS Transport Networks", RFC 5860, May 2010.

- [RFC5921] Bocci, M., Bryant, S., Frost, D., Levrau, L., and L. Berger, "A Framework for MPLS in Transport Networks", RFC 5921, July 2010.
- [RFC6374] Frost, D. and S. Bryant, "Packet Loss and Delay Measurement for MPLS Networks", RFC 6374, September 2011.
- [RFC6375] Frost, D. and S. Bryant, "A Packet Loss and Delay Measurement Profile for MPLS-Based Transport Networks", RFC 6375, September 2011.

Authors' Addresses

Lianshu Zheng
Huawei Technologies
Huawei Building, No.156 Beiqing Rd.
Beijing 100095
China

Email: vero.zheng@huawei.com

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

Email: lizhenbin@huawei.com

Bhavani Parise
Cisco Systems

Email: bhavani@cisco.com

Network Working Group
Internet-Draft
Updates: 6514 (if approved)
Intended status: Standards Track
Expires: December 5, 2013

Zhang
Rekhter
Juniper Networks
Dolganow
Alcatel-Lucent
June 3, 2013

Simulating "Partial Mesh of MP2MP P-Tunnels" with Ingress Replication
draft-zzhang-l3vpn-mvpn-bidir-ingress-replication-00.txt

Abstract

RFC 6513 described a method to support bidirectional C-flow using "Partial Mesh of MP2MP P-Tunnels". This document describes how partial mesh of MP2MP P-Tunnels can be simulated with Ingress Replication, instead of a real MP2MP tunnel.

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 December 5, 2013.

Copyright Notice

Copyright (c) 2013 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	3
1.1. Terminology	3
2. Requirements Language	5
3. Operation	6
3.1. Control State	6
3.2. Forwarding State	7
4. Security Considerations	9
5. IANA Considerations	10
6. Acknowledgements	11
7. Normative References	12
Authors' Addresses	13

1. Introduction

Section 11.2 of RFC 6513, "Partitioned Sets of PEs", describes two methods of carrying bidirectional C-flow traffic over a provider core without using the core as RPL or requiring Designated Forwarder election.

With these two methods, all PEs of a particular VPN are separated into partitions, with each partition being all the PEs that elect the same PE as the UMH wrt the C-RPA. A PE must discard bidirectional C-flow traffic from PEs that are not in the same partition as the PE itself.

In particular, Section 11.2.3 of RFC 6513, "Partial Mesh of MP2MP P-Tunnels", guarantees the above discard behavior without using an extra PE Distinguisher label by having all PEs in the same partition join a single MP2MP tunnel dedicated to that partition and use it to transmit traffic. All traffic arriving on the tunnel will be from PEs in the same partition, so it will be always accepted.

RFC 6514 specifies BGP encodings and procedures used to implement MVPN as specified in RFC 6513, while the details related to MP2MP tunnels are specified in [draft-ietf-l3vpn-mvpn-bidir-05].

[draft-ietf-l3vpn-mvpn-bidir-05] assumes that an MP2MP P-tunnel is realized either via PIM-Bidir, or via MP2MP mLDP. Each of them would require signaling and state not just on PEs, but on the P routers as well. This document describes how the MP2MP tunnel can be simulated with a mesh of P2P or MP2P LSPs, i.e. Ingress Replication. The advantage is that existing P2P/MP2P LSPs created for unicast can be used for multicast as well w/o introducing additional signaling or state in the core. While there may be concerns with traffic replication in the core, in many situations the traffic could be low-rate and/or sporadic and the advantage of signaling and state savings will outweigh the concerns with traffic replication, making Ingress Replication an applicable and attractive alternative.

This documentation specifies the BGP signaling and procedures used to simulate "Partial Mesh of MP2MP P-Tunnels" with Ingress Replication.

1.1. Terminology

This document uses terminology from [RFC6513], [RFC6514], and [draft-ietf-l3vpn-mvpn-bidir-05]. In particular, the following new term is defined:

- o C-G-BIDIR: A C-G where G is a Bidir-PIM group.

2. 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 [RFC2119].

3. Operation

3.1. Control State

If a PE, say PEX, is connected to a site of a given VPN, and that site hosts the C-RPA for some Bidir-PIM groups, i.e., the route to the C-RPA is through a local PE-CE interface, then PEX MUST advertise a (C-*,C-BIDIR) S-PMSI A-D route, regardless of whether it has any local Bidir-PIM join states corresponding to the C-RPA learned from its CEs. It MAY also advertise a (C-*,C-G-BIDIR) S-PMSI A-D route, just like how any other S-PMSI A-D routes are triggered (e.g, when the (C-*,C-G-BIDIR) traffic rate goes above a threshold). Here the C-G-BIDIR refers to a C-G where G is a Bidir-PIM group, and the corresponding C-RPA is in the site that the PEX connects to.

The S-PMSI A-D routes include a Provider Tunnel Attribute (PTA) with tunnel type set to Ingress Replication, with Leaf Information Required flag set, and with a downstream allocated MPLS label that other PEs in the same partition MUST use when sending relevant C-bidir flows to this PE.

If some other PE, PEy, receives and imports into one of its VRFs such a (C-*,C-BIDIR) S-PMSI A-D route, and the VRF has any local Bidir-PIM join state that PEy has received from its CEs, and if PEy chooses PEX as its UMH wrt the C-RPA for those states, PEy MUST advertise a Leaf A-D route in response. Or, if PEy has received and imported into one of its VRFs a (C-*,C-BIDIR) S-PMSI A-D route from PEX before, then upon receiving in the VRF any local Bidir-PIM join state from its CEs with PEX being the UMH for those states' C-RPA, PEy MUST advertise a Leaf A-D route.

The encoding of the Leaf A-D route is as specified in RFC 6514, except that the Route Targets are set to the same value as in the corresponding S-PMSI A-D route so that the Leaf A-D route will be imported by all VRFs that import the corresponding S-PMSI A-D route. This is irrespective of whether from a receiving PE, PEz's perspective PEX (originator of the S-PMSI A-D route) is the UMH PE or not. The label in the PTA of the Leaf A-D route originated by PEy MUST be allocated specifically for PEX, so that when traffic arrives with that label, the traffic can associated with the partition (represented by the PEX).

With PEy advertising Leaf A-D route only if it chooses the originator of the S-PMSI A-D route as its UMH, it won't receive traffic from PEs in other partitions, so the label is actually useful only when PEy switches to a different UMH - it will stop accepting traffic before sending PEs stop sending it traffic (upon the receipt of its Leaf A-D route withdrawl). To speed up convergency (so that PEy starts

receiving traffic from its new UMH immediately instead of waiting until the new Leaf A-D route corresponding to the new UMH is received by sending PEs), PEy MAY advertise a Leaf A-D route even if does not choose PEx as its UMH wrt the C-RPA. With that, it will receive traffic from all PEs, but some will arrive with the label corresponding to its choice of UMH while some will arrive with a different label, and the traffic in the latter case will be discarded.

Similar to the (C-*,C-BIDIR) case, if PEy receives and imports into one of its VRFs such a (C-*,C-G-BIDIR) S-PMSI A-D route, and PEy chooses PEx as its UMH wrt the C-RPA, and it has corresponding local (C-*,C-G-BIDIR) join state that it has received from its CEs in the VRF, PEy MUST advertise a Leaf A-D route in response. Or, if PEy has received and imported into one of its VRFs a (C-*,C-G-BIDIR) S-PMSI A-D route before, then upon receiving its local (C-*,C-G-BIDIR) join state from its CEs in the VRF, it MUST advertise a Leaf A-D route.

The encoding of the Leaf A-D route is as specified in RFC 6514, except that the Route Targets are set to the same as in the corresponding S-PMSI A-D route so that the Leaf A-D route will be imported by all VRFs that import the corresponding S-PMSI A-D route. This is irrespective of whether from the receiving PE, PEz's perspective PEx (originator of the S-PMSI A-D route) is the UMH PE or not. The label in the PTA of the Leaf A-D route originated by PEy MUST be allocated specifically for PEx, so that when traffic arrives with that label, the traffic can associated with the partition (represented by the PEx).

Whenever the (C-*,C-BIDIR) or (C-*,C-G-BIDIR) S-PMSI A-D route is withdrawn, or if PEy no longer chooses the originator PEx as its UMH wrt C-RPA and PEy only advertises Leaf A-D routes in response to its UMH's S-PMSI A-D route, or if relevant local join state is pruned, PEy MUST withdraw the corresponding Leaf A-D route.

3.2. Forwarding State

The following specification regarding forwarding state matches the "When an S-PMSI is a 'Match for Transmission'" and "When an S-PMSI is a 'Match for Reception'" rules for "Flat Partitioning" method in [draft-ietf-l3vpn-mvpn-bidir-05], except that the rules about (C-*,C-*) are not applicable, because this document requires that (C-*,C-BIDIR) S-PMSI A-D routes are always originated for a VPN that supports C-Bidir flows.

For the (C-*,C-G-BIDIR) S-PMSI A-D route that a PEy receives and imports into one of its VRFs from its UMH wrt the C-RPA, or if PEy itself advertises the S-PMSI A-D route in the VRF, PEy maintains a

(C-*,C-G-BIDIR) forwarding state in the VRF, with the Ingress Replication provider tunnel leaves being the originators of the S-PMSI A-D route and all relevant Leaf-A-D routes. The relevant Leaf A-D routes are the routes whose Route Key field contains the same information as the MCAST-VPN NLRI of the (C-*, C-G-BIDIR) S-PMSI A-D route advertised by the UMH.

For the (C-*,C-BIDIR) S-PMSI A-D route that a PEy receives and imports into one of its VRFs from its UMH wrt a C-RPA, or if PEy itself advertises the S-PMSI A-D route in the VRF, it maintains appropriate forwarding states in the VRF for the ranges of bidirectional groups for which the C-RPA is responsible. The provider tunnel leaves are the originators of the S-PMSI A-D route and all relevant Leaf-A-D routes. The relevant Leaf A-D routes are the routes whose Route Key field contains the same information as the MCAST-VPN NLRI of the (C-*, C-BIDIR) S-PMSI A-D route advertised by the UMH. This is for the so-called "Sender Only Branches" where a router only has data to send upstream towards C-RPA but no explicit join state for a particular bidirectional group. Note that the traffic must be sent to all PEs (not just the UMH) in the partition, because they may have specific (C-*,C-G-BIDIR) join states that this PEy is not aware of, while there is no corresponding (C-*,C-G-BIDIR) S-PMSI A-D and Leaf A-D routes.

For a (C-*,C-G-BIDIR) join state that a PEy has received from its CEs in a VRF, if there is no corresponding (C-*,C-G-BIDIR) S-PMSI A-D route from its UMH in the VRF, PEy maintains a corresponding forwarding state in the VRF, with the provider tunnel leaves being the originators of the (C-*,C-BIDIR) S-PMSI A-D route and all relevant Leaf-A-D routes (same as the above Sender Only Branch case). The relevant Leaf A-D routes are the routes whose Route Key field contains the same information as the MCAST-VPN NLRI of the (C-*, C-BIDIR) S-PMSI A-D route originated by the UMH. If there is no (C-*,C-BIDIR) S-PMSI A-D route from its UMH either, then the provider tunnel has an empty set of leaves and PEy does not forward relevant traffic across the provider network.

4. Security Considerations

This document raises no new security issues. Security considerations for the base protocol are covered in [RFC6514].

5. IANA Considerations

This document has no IANA considerations.

This section should be removed by the RFC Editor prior to final publication.

6. Acknowledgements

7. Normative References

- [RFC2119] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, March 1997.
- [RFC6513] Rosen, E. and R. Aggarwal, "Multicast in MPLS/BGP IP VPNs", RFC 6513, February 2012.
- [RFC6514] Aggarwal, R., Rosen, E., Morin, T., and Y. Rekhter, "BGP Encodings and Procedures for Multicast in MPLS/BGP IP VPNs", RFC 6514, February 2012.
- [RFC5015] Handley, M., Kouvelas, I., Speakman, T., and L. Vicisano, "Bidirectional Protocol Independent Multicast (BIDIR-PIM)", RFC 5015, October 2007.
- [I-D.ietf-l3vpn-mvpn-bidir]
Rosen, E., Wijnands, I., Cai, Y., and A. Boers, "MVPN: Using Bidirectional P-Tunnels",
draft-ietf-l3vpn-mvpn-bidir-05 (work in progress),
April 2013.

Authors' Addresses

Jeffrey Zhang
Juniper Networks
10 Technology Park Dr.
Westford, MA 01886
US

Email: zzhang@juniper.net

Yakov Rekhter
Juniper Networks
1194 North Mathilda Ave.
Sunnyvale, CA 94089
US

Email: yakov@juniper.net

Andrew Dolganow
Alcatel-Lucent
600 March Rd.
Ottawa, ON K2K 2E6
CANADA

Email: andrew.dolganow@alcatel-lucent.com

