

L3VPN Routing Working Group  
Internet-Draft  
Intended status: Standards Track  
Expires: March 19, 2014

H. Jeng  
AT&T  
R. Bonica  
Y. Rekhter  
Juniper Networks  
September 15, 2013

Covering Prefixes Outbound Route Filter for BGP-4  
draft-bonica-l3vpn-orf-covering-prefixes-00

Abstract

This document defines a new ORF-type, called the "Covering Prefixes ORF (CP-ORF)". The CP-ORF is applicable in the context of a Virtual Hub-and-Spoke VPN. It may also be applicable in other BGP/MPLS VPN environments.

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 March 19, 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. Problem Statement . . . . .	2
1.1. Terminology . . . . .	2
2. CP-ORF Encoding . . . . .	3
3. Processing Rules . . . . .	4
4. Applicability In Virtual Hub-and-Spoke VPNs . . . . .	5
4.1. CP-ORF Clean-up . . . . .	7
5. IANA Considerations . . . . .	8
6. Security Considerations . . . . .	8
7. Acknowledgements . . . . .	9
8. Normative References . . . . .	9
Authors' Addresses . . . . .	9

## 1. Problem Statement

[RFC5291] provides a mechanism through which a BGP [RFC4271] speaker can send a set of Outbound Route Filters (ORFs) to a peer. The peer uses those ORFs to filter routing updates that it sends to the BGP speaker.

The ORF mechanism allows the speaker to realize the "route pull" paradigm with BGP, where the speaker, on demand, can pull certain routes from the peer.

This document defines a new ORF-type, called the "Covering Prefixes ORF (CP-ORF)". The CP-ORF is applicable in the context of a Virtual Hub-and-Spoke VPN [I-D.ietf-l3vpn-virtual-hub]. However, it may also be applicable in other BGP/MPLS VPN [RFC4364] environments.

### 1.1. Terminology

This document uses the terms "Address Family Identifier (AFI)" and "Subsequent Address Family Identifier (SAFI)". In the context of this document, the meaning of these terms is the same as in [RFC4760].

This document also uses the terms "VPN IP default route", "V-hub" and "V-spoke". In the context of this document, the meaning of these terms is the same as in [I-D.ietf-l3vpn-virtual-hub].

## 2. CP-ORF Encoding

[RFC5291] augments the BGP ROUTE-REFRESH message so that it can carry ORF entries. When the ROUTE-REFRESH message carries ORF entries, it includes the following fields:

- o AFI
- o SAFI
- o When-to-refresh (IMMEDIATE or DEFERRED)
- o ORF Type
- o Length (of ORF entries)

The ROUTE-REFRESH message also contains a list of ORF entries. Each ORF entry contains the following fields:

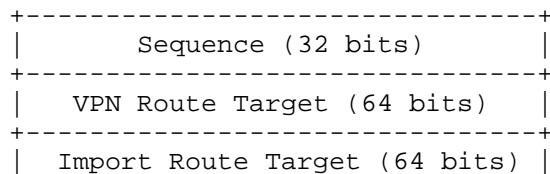
- o Action (ADD, REMOVE, or REMOVE-ALL)
- o Match (PERMIT or DENY)

The ORF entry may also contain Type-specific information. Type-specific information is present only when the Action is equal to ADD or REMOVE. It is not present when the Action is equal to REMOVE-ALL.

When the BGP ROUTE-REFRESH message carries CP-ORF entries, the following conditions must be true:

- o ORF Type MUST be equal to CP-ORF. (The value of CP-ORF is TBD. See Section 5 for details.)
- o AFI MUST be equal to either IPv4 or IPv6
- o SAFI MUST be equal to "MPLS-labeled VPN address" [IANA.SAFI]
- o Match field MUST be equal to PERMIT

Figure 1 depicts the encoding of the CP-ORF type-specific information.



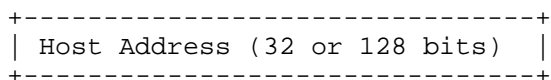


Figure 1: CP-ORF Type-specific Encoding

The Sequence field specifies the relative ordering of the entry among all CP-ORF entries.

The VPN Route Target field is used by the recipient of CP-ORF to identify the set of routes to which CP-ORF applies. See Section 3 for details.

The Import Route Target also is used by the recipient of CP-ORF. The CP-ORF recipient marks routes selected by CP-ORF with the value of the Route Target extended community before advertising them to the originator of the CP-ORF. See Section 3 for details.

If the AFI field in the ROUTE-REFRESH message is equal to IPv4, the Host Address field MUST contain exactly 32 bits and MUST represent an IPv4 host address. If the AFI field in the ROUTE-REFRESH message is equal to IPv6, the Host Address field MUST contain exactly 128 bits and MUST represent an IPv6 host address.

### 3. Processing Rules

When a BGP speaker receives a ROUTE-REFRESH message that contains a CP-ORF, and that ROUTE-REFRESH message that violates any of the encoding rules specified in Section 2, the BGP speaker MUST log the event and ignore the entire ROUTE-REFRESH message.

Otherwise, the BGP speaker processes each CP-ORF entry as indicated by the Action field. If the Action is equal to ADD, the BGP speaker adds a CP-ORF entry in the position specified by the Sequence field. If the Action is equal to REMOVE, the BGP speaker removes a CP-ORF entry. If the Action is equal to REMOVE-ALL, the BGP speaker removes all CP-ORF entries.

Whenever the BGP speaker advertises routes to a peer, it evaluates each route in terms of each CP-ORF entry received from that peer. A route matches the selection criteria of a CP-ORF if the following statements are true:

- o the route is more specific than a /64 (i.e., the route more specific than an IP VPN default route)
- o the route carries RT whose value is the same as the CP-ORF VPN Route Target

- o the route covers the CP-ORF Host Address

When evaluating whether the route covers the CP-ORF Host Address, the BGP speaker ignores Route Distinguishers. For example, assume that the CP-ORF Host Address is equal to 192.0.2.1. Assume also that the RIB contains routes for the following IPv4 VPN prefixes, and that all of these routes carry an RT whose value is the same as the CP-ORF VPN Route Target:

- o 1:0.0.0.0/64.
- o 2:192.0.2.0/88
- o 3:192.0.2.0/89

For the purposes of this search, 2:192.0.2.0/88 and 3:192.0.2.0/89 cover 192.0.2.1. This is because the search algorithm ignores Route Distinguishers. However, 1:0.0.0.0/64 does not cover 192.0.2.1, because the search algorithm requires a prefix length greater than /64.

If a route matches the selection criteria of a CP-ORF, the BGP speaker places the route into the Adj-RIB-Out associated with the peer from which CP-ORF was received. When placing the route into the Adj-RIB-Out, the speaker applies the following rules:

- o all BGP attributes except for Route Targets are unchanged
- o The Route Target specified by the CP-ORF Import Route Target is added to the list of Route Targets that the route already carries

As a result of placing the route into the Adj-RIB-Out, the route is advertised to the peer.

#### 4. Applicability In Virtual Hub-and-Spoke VPNs

In a Virtual Hub-and-Spoke environment, VPN sites are attached to Provider Edge (PE) routers, V-hubs and V-spokes. PE routers, V-hubs and V-spokes can exchange VPN routes through an iBGP mesh. Alternatively, they can exchange customer routes using Route Reflectors (RR).

For the purposes of this document, assume that RED-VPN sites are attached to PE routers, V-hub1 and V-spoke1. All of these devices advertise RED-VPN routes to a RR. They mark these routes with a route target, which we will call RT-RED.

V-hub1 serves the RED-VPN. Therefore, V-hub1 advertises a VPN IP default route for the RED-VPN to the RR, carrying the route target RT-RED-FROM-HUB1.

V-spokel establishes a BGP session with the RR, negotiating the CP-ORF capability, as well as the Multiprotocol Extensions Capability [RFC2858]. Upon establishment of the BGP session, the RR does not advertise any routes to V-spokel. The RR will not advertise any routes until it receives either a ROUTE-REFRESH message or a BGP UPDATE message containing a Route Target Membership NLRI [RFC4684].

Immediately after the BGP session is established, V-spokel sends the RR a BGP UPDATE message containing a Route Target Membership NLRI. The Route Target Membership NLRI specifies RT-RED-FROM-HUB1 as its route target. In response to the BGP-UPDATE message, the RR advertises the VPN IP default route for the RED-VPN to V-spokel. This route still carries the route target RT-RED-FROM-HUB1. V-spokel subjects this route to its import policy and accepts it because it carries the route target RT-RED-FROM-HUB1.

Now, V-spokel begins normal operation, sending all of its traffic through V-hub1. At some point, V-spokel determines that it might benefit from a more direct route to a destination. (Criteria by which V-spokel determines that it needs a more direct route are beyond the scope of this document.)

In order to discover a more direct route, V-spokel assigns a unique numeric identifier to the flow. V-spokel then sends a ROUTE-REFRESH message to the RR, containing the following information:

- o AFI is equal to IPv4 or IPv6, as appropriate
- o SAFI is equal to "MPLS-labeled VPN address"
- o When-to-refresh is equal IMMEDIATE
- o Action is equal to ADD
- o Match is equal to PERMIT
- o ORF Type is equal to CP-ORF
- o CP-ORF Sequence is equal to the identifier associated with the flow
- o CP-ORF VPN Route Target is equal to RT-RED
- o CP-ORF Import Route Target is equal to RT-RED-FROM-HUB1

- o CP-ORF Host Address is equal the destination address associated with the flow

Upon receipt of the ROUTE-REFRESH message, the RR must ensure that it carries all routes belonging to the RED-VPN. In at least one special case, where all of the RR clients are V-spokes and none of the RR clients are V-hubs, the RR will lack some or all of the required RED-VPN routes. So, the RR sends a BGP UPDATE message containing a Route Target Membership NLRI for VPN-RED to all of its peers. This causes the peers to advertise VPN-RED routes to the RR, if they have not done so already.

Next, the RR installs the CP-ORF and refreshes routes for V-spoke1. If the RR maintains any routes matching the CP-ORF selection criteria, it advertises those routes. As it advertises those routes, it adds the CP-ORF Import Route Target to the list of route targets that they carry. The advertised routes may specify either V-hub1 or any other node as the NEXT-HOP.

V-spoke1 subjects the advertised routes to its import policy and accepts them because they carry the route target RT-RED-FROM-HUB1.

V-spoke1 may repeat this process whenever it discovers another flow that might benefit from a more direct route to its destination.

#### 4.1. CP-ORF Clean-up

Each CP-ORF consumes memory and compute resources on the device that supports it. Therefore, in order to obtain optimal performance, the V-spoke periodically evaluates all CP-ORFs that it has originated and removes unneeded CP-ORFs. The V-spoke determines that a CP-ORF is unneeded if its forwarding table includes no route satisfying the following criteria:

- o Covers the CP-ORF Host Address
- o Carries the same route target as the CP-ORF VPN Route Target
- o Has prefix length greater than 64 (i.e., is not a default route)
- o Has NEXT-HOP different from that of any VPN IP default route (i.e., different from any V-hub with which the V-Spoke is associated)

When the V-spoke finds an unneeded CP-ORF, it removes the CP-ORF, as described below, and adds CP-ORF Host Address to a list of addresses known to be reachable only through the V-hub. The Host Address remains on that list for a configurable period of time. While the

Host Address is on that list, flows directed toward it will not be considered as candidates for a more direct route.

Also, the V-spoke removes all CP-ORFs when a configurable period of time has elapsed since their installation. When it does this, it does not add CP-ORF Host Address to the list of addresses known to be reachable only through a V-hub. If the V-spoke once again determines that a flow directed towards the Host Address might benefit from a more direct route, it will send another CP-ORF.

In order to removed unneeded CP-ORFs, the V-spoke sends a single ROUTE Refresh message containing the following information:

- o AFI is equal to IPv4 or IPv6, as appropriate
- o SAFI is equal to "MPLS-labeled VPN address"
- o When-to-refresh is equal IMMEDIATE
- o Action is equal to REMOVE
- o Match is equal to PERMIT
- o ORF Type is equal to CP-ORF
- o A list of CP-ORFs, with one element representing each unneeded CP-ORF

The recipient of this message responds to it as described in [RFC5291].

## 5. IANA Considerations

IANA is requested to assign a All Covering Prefixes ORF Type from the BGP Outbound Route Filtering (ORF) Types Registry.

## 6. Security Considerations

Each CP-ORF consumes finite memory and compute resources on the control plane of the V-hub. Therefore, the V-hub MUST take the following steps to protect itself from oversubscription:

- o When negotiating the ORF capability, advertise willingness to receive the CP-ORF only to known, trusted iBGP peers
- o Enforce a per-peer limit on the number of CP-ORFs that can be installed at any given time. Ignore all requests to add CP-ORFs beyond that limit



## 7. Acknowledgements

The authors wish to acknowledge Han Nguyen and James Uttaro for their comments and contributions.

## 8. 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.
- [IANA.SAFI]  
IANA, "abbrev="Subsequent Address Family Identifiers (SAFI) Parameters"", , <<http://www.iana.org/assignments/safi-namespace/safi-namespace.xhtml#safi-namespace-2>>.
- [RFC2119] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, March 1997.
- [RFC2858] Bates, T., Rekhter, Y., Chandra, R., and D. Katz, "Multiprotocol Extensions for BGP-4", RFC 2858, June 2000.
- [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.
- [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.
- [RFC4760] Bates, T., Chandra, R., Katz, D., and Y. Rekhter, "Multiprotocol Extensions for BGP-4", RFC 4760, January 2007.
- [RFC5291] Chen, E. and Y. Rekhter, "Outbound Route Filtering Capability for BGP-4", RFC 5291, August 2008.
- [RFC5292] Chen, E. and S. Sangli, "Address-Prefix-Based Outbound Route Filter for BGP-4", RFC 5292, August 2008.

## Authors' Addresses

Huajin Jeng  
AT&T

Email: [hj2387@att.com](mailto:hj2387@att.com)

Ron Bonica  
Juniper Networks  
2251 Corporate Park Drive  
Herndon, Virginia 20170  
USA

Email: [rbonica@juniper.net](mailto:rbonica@juniper.net)

Yakov Rekhter  
Juniper Networks  
1194 North Mathilda Ave.  
Sunnyvale, California 94089  
USA

Email: [yakov@juniper.net](mailto:yakov@juniper.net)

INTERNET-DRAFT  
Intended Status  
Expires: April 18, 2014

Luyuan Fang  
Rex Fernando  
Dhananjaya Rao  
Sami Boutros  
Cisco

October 18, 2013

BGP/MPLS IP VPN Data Center Interconnect  
draft-fang-l3vpn-data-center-interconnect-02

## Abstract

This document discusses two categories of inter-connections of BGP/MPLS IP VPN [RFC4364] and Data Center (DC) overlay networks. In the first category, DC overlay virtual network is built with BGP/MPLS IP VPN (IP VPN) technologies, the inter-connection of IP VPN in the DC to IP VPN in the WAN enables end-to-end IP VPN connectivity. In the second category, DC overlay network uses non IP VPN overlay technologies, the inter-connection of any overlay virtual network in the DC to IP VPN in the WAN provides end user connectivity through stitching of different overlay technologies.

## 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	2
1.1	Terminology	3
2.	Use Cases	3
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	4
3.1	BGP/MPLS IP VPN Inter-AS model	4
3.2	BGP/MPLS IP VPN Gateway PE to DC vCE Model	5
3.3	Hybrid inter-connection model	6
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	8
4.1.4	Use of RTC	9
5.	Inter-connect IP VPN and non-IP VPN overlay networks	9
6.	Security Considerations	10
7.	IANA Considerations	10
8.	References	11
8.1	Normative References	11
8.2	Informative References	11
	Authors' Addresses	12

## 1 Introduction

With the growth of cloud services, the need of inter-connecting DC overlay networks and Enterprise BGP/MPLS IP VPNs in the Wide Area Network (WAN) arises.

Two categories of inter-connections of BGP/MPLS IP VPN [RFC4364] and Data Center (DC) overlay networks are discussed in this document. In

the first category, DC overlay virtual network is built with BGP/MPLS IP VPN (IP VPN) technologies, the inter-connection of IP VPN in the DC to IP VPN in the WAN enables end-to-end IP VPN connectivity for Virtual Private Cloud (VPC) services. In the second category, DC overlay network uses non IP VPN overlay technologies, the inter-connection of any overlay virtual network in the DC to IP VPN in the WAN provides end user connectivity through stitching of different overlay technologies.

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
GRE	Generic Routing Encapsulation
Hypervisor	Virtual Machine Manager
I2RS	Interface to Routing System
MP-BGP	Multi-Protocol Border Gateway Protocol
NVGRE	Network Virtualization using GRE
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
vCE	virtual Customer Edge Router
VM	Virtual Machine
VN	Virtual Network
VPC	Virtual Private Cloud
vPE	virtual Provider Edge Router
VPN	Virtual Private Network
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

SPs have large deployments of BGP/MPLS IP VPNs. Many SPs are interested to extend the IP VPN capabilities into their DCs to provide end-to-end native BGP IP VPN services to their enterprise customers.

BGP IP VPN provides routing isolation, rich policy control, and QoS support. The technologies developed to extend BGP IP VPN into DC servers or ToR are work in progress in IETF, [I-D.fang-l3vpn-virtual-pe], and [I-D.ietf-l3vpn-end-system].

The WAN and DC may be managed by the same or different administrative domains.

## 2.2 Case 2: Hybrid cloud inter-connection

Inter-connecting network SPs Enterprise IP VPNs to Cloud/Content providers DCs for expanded services. The inter-connection between the SP BGP/MPLS IP VPNs and the cloud provider networks is needed to provide the service end-to-end. The inter-connection of different types of providers can be BGP/MPLS IP VPN to other VPN or overlay technologies which may be virtualized or non-virtualized.

## 3. Architecture reference models

The architecture reference models described below focus on the inter-connection aspects. Although the intra-DC implementation is not within the scope of this discussion, the intra-DC technology has a 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 DC. The IP VPN implementation is using vPE technology in DC. The two segments of the VPNs are inter-connected through ASBRs facing each other in the respective networks.

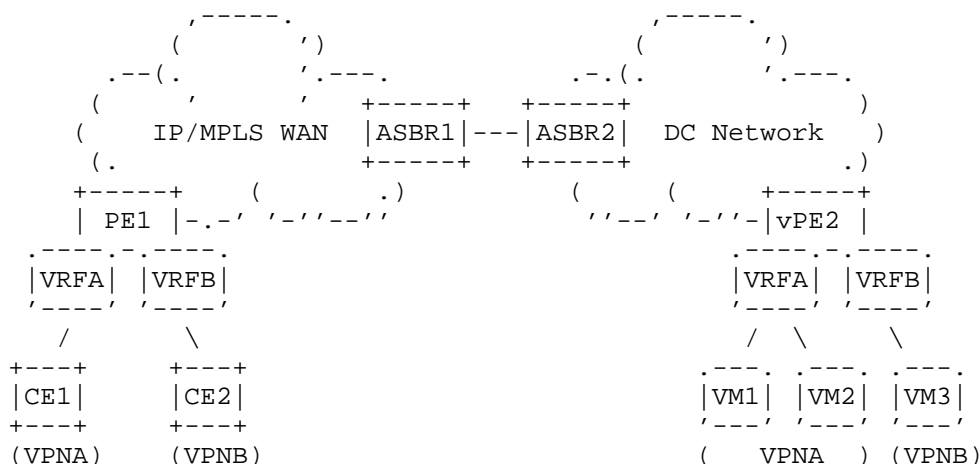


Figure 1. BGP/MPLS IP VPN Inter-Connection  
with ASBR in each network

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 illustrates this shared ASBR model.

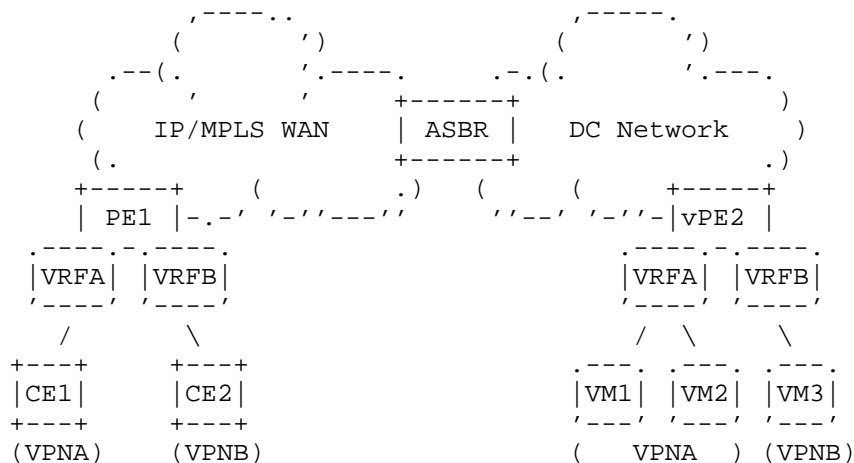


Figure 2. BGP/MPLS IP VPN Inter-Connection  
with share ASBR

### 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 is used by SPs to provide managed services, when scale can meet the service requirement.

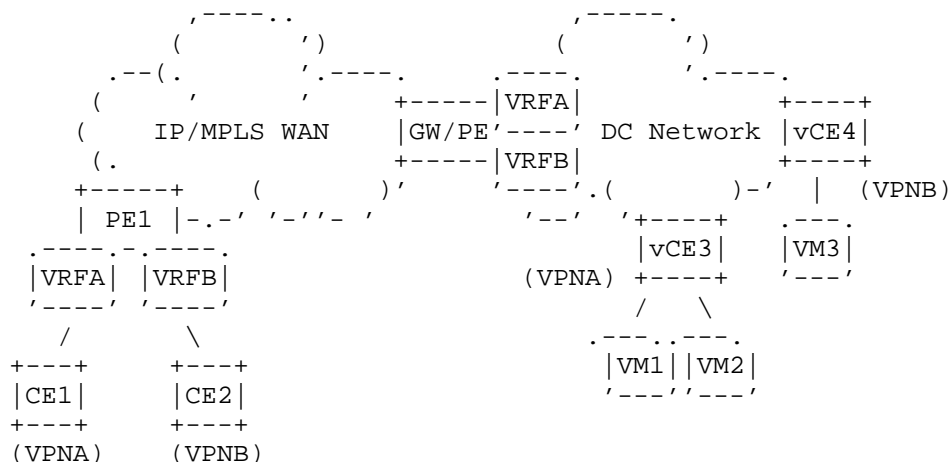


Figure 3. BGP/MPLS IP VPN GW/PE to vCEs  
without BGP/MPLS IP VPN in the DC

### 3.3 Hybrid inter-connection model

The BGP/MPLS IP VPNs are implemented in the WAN network, and non BGP/MPLS IP VPN Overlay are implemented in the DC. The connection of the two networks is outside of the technologies for Inter-AS connections for BGP IP VPNs. This model includes 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.



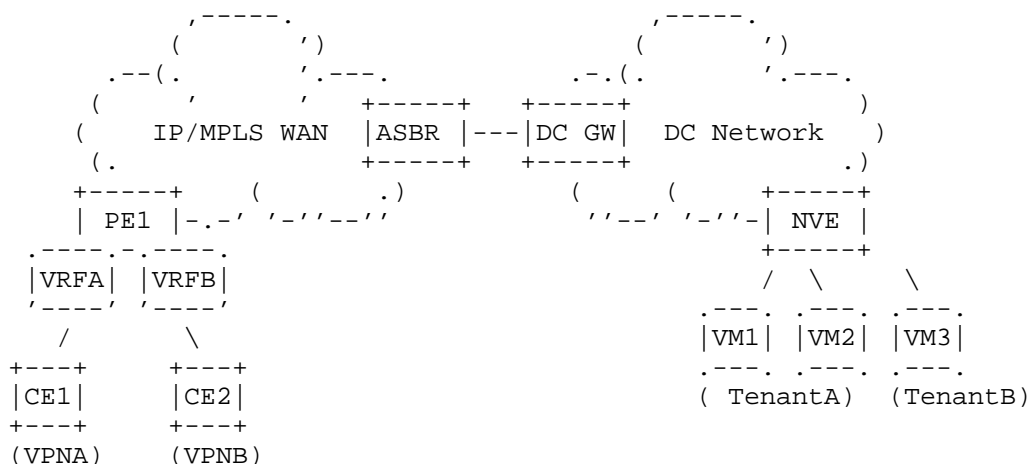


Figure 4. BGP/MPLS IP VPN Inter-Connection with  
non BGP/MPLS IP VPN Overlay in DC

#### 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 de-encapsulation and transmits packets to the peer PE-ASBR in the adjacent AS. The peer PE-ASBR performs MPLS or IP VPN encapsulation on the customer IPv4/IPv6 packets received, and transmits the packet through the IP backbone of the AS. 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: This 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: This option suffers from scaling limitations, because per Inter-AS VPN VRF and interface are needed on the PE-ASBR.

Option A has been commonly used in BGP/MPLS VPN Inter-Provider inter-

connections because of the security considerations and its clear operational demarcation.

DCI considerations: This is a simple way to connect DC and WAN if both sides are of 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 a large number) from the WAN side, and Vice Versa.

#### 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 ASes, 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 ASes exchange VPN routing information, such as, the destination network, the next hop field associated with the distributing router, a local MPLS label, and an RD. 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 improved scalability when compared with option A, since 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 the dynamic routing information exchange that is involved. The ASBR scaling may still be an issue because ASBR must maintain all VPN routes.

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

DCI considerations: Option B is one viable option to be used in DC inter-connection. However, it has the same scale concerns as other options because of the potentially large number of routes exchanged between the WAN and the DC.

#### 4.1.3 Option C pros and cons

In option C: Multihop eBGp redistribution of labeled VPN-IPv4/IPv6 routes between source and destination ASs, with eBGp redistribution of labeled IPv4/IPv6 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's perspective, the ASBRs act as P routers, with no knowledge about the VPNs concerned. Between the two inter-connecting ASBRs, the VPN traffic is treated just as between two P routers, each VPN 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 three. ASBR is no longer a bottle neck for VPN routes scaling as in Option B.

Cons: Major security issues as IGP reachability must be exchanged between the inter-connecting 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 for security reasons.

In this option, though ASBR is not longer the scaling bottleneck, the scaling issues still call for careful design, as the total numbers of VRFs, VPN interfaces, and the VPN routes being exchanged between the two ASes can be very large.

#### 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 scalable solution.

However, all IP VPN routes are exchanged between the two ASes (e.g. WAN and DC) as long as they have to support the same VPNs. The potential IP VPN routes distribution can still be very substantial in large WAN and DC deployment. Additional aggregation and abstraction mechanisms can be used to avoid large numbers of VPN routes being exchanges across the border between the interconnecting WAN and the DC in either directions.

### 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 Gateway router and traffic directed into a BGP/MPLS IP VPN, the DC Gateway router performs MPLS encapsulation towards the WAN and IP overlay based forwarding within the DC.

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 performs an IP lookup, appropriate MPLS or IP encapsulation, and forward traffic.

The DC Gateway router peers 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

BGP/MPLS Inter-AS security threats and defense techniques described in RFC 4111 [RFC4111] are applicable for the WAN/DC inter-connections.

In addition, the protocols between the Gateway routers and the controller/orchestrator MUST be mutually authenticated. Given the potentially very large scale and the dynamic nature in the cloud/DC environment, the choice of key management mechanisms need to be further studied.

### 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.
- [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.

### 8.2 Informative References

- [RFC4111] Fang, L., Ed., "Security Framework for Provider-Provisioned Virtual Private Networks (PPVPNs)", RFC 4111, July 2005.
- [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, work in progress.
- [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, work in progress.
- [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, work in progress.
- [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, work in progress.
- [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, work in progress.

[I-D.sridharan-virtualization-nvgre]: SridharanNetwork, M., et al.,  
"Virtualization using Generic Routing Encapsulation",  
draft-sridharan-virtualization-nvgre, work in progress.

## Authors' Addresses

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

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

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

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

INTERNET-DRAFT  
Intended Status: Informational  
Expires: April 18, 2014

Maria Napierala  
AT&T  
Luyuan Fang  
Cisco

October 18, 2013

Requirements for Extending BGP/MPLS VPNs to End-Systems  
draft-fang-l3vpn-end-system-requirements-02.txt

Abstract

The proven scalability and extensibility beyond the original design purposes of the BGP/MPLS IP VPNs (IP VPN) technology [RFC4364] has made it an attractive candidate for Data Center (DC)/Cloud virtualization. This document provides the requirements for extending IP VPN (in original or modified versions) into the end-systems/end-devices, such as a server in a DCs/Cloud. Physical separation of the control and the forwarding planes; virtualizing the network functions of the IP VPN network elements, such as a PE, are the key differences compared with the classic IP VPN solutions.

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.	Application of MPLS/BGP VPNs to End-Systems	4
2.1.	End-System CE and PE Functions	4
2.2.	PE Control Plane Function	5
3.	VPN Communication Requirements	5
3.1.	Unicast IPv4 and IPv6	5
3.2.	Multicast/VPN Broadcast IPv4 and IPv6	5
3.3.	IP Subnet Support	5
4.	Multi-Tenancy Requirements	6
5.	Decoupling of Virtualized Networking from Physical	7
6.	Decoupling of Layer 3 Virtualization from Layer 2 Topology	7
7.	Requirements for Encapsulation of Virtual Payloads	8
7.1.	Encapsulation Methods	9
7.2.	Routing of Virtual Payloads	9
8.	Optimal Forwarding of Traffic	9
9.	IP Mobility	10
9.1.	IP Addressing of Virtual Hosts	10
9.2.	Network Layer-Based Mobility	10
9.3.	Routing Convergence Requirements	10
10.	Inter-operability with Existing MPLS/BGP VPNs	11
11.	BGP Requirements in a Virtualized Environment	12
11.1.	BGP Convergence and Routing Consistency	12
11.1.1.	BGP IP Mobility Requirements	12
11.2.	Optimization of Route Distribution	12
11.3.	Service chaining	13
12.	Security Considerations	13
13.	IANA Considerations	13
14.	References	13
14.1.	Normative References	13
14.2.	Informative References	14
15.	Acknowledgements	14
	Authors' Addresses	14



## 1 Introduction

Enterprise networks are increasingly being consolidated and outsourced in an effort to improve the deployment time of services as well as reduce operational costs. This coincides with an increasing demand for compute, storage, and network resources from applications. Logical abstraction of these resources is needed to for improved scalability and cost efficiency. This is referred as server, storage, and network virtualization. It can be implemented in all layers of the computer systems or networks. The virtualized loads are executed or transferred over a common physical infrastructure. Compute nodes running guest operating systems are often executed as Virtual Machines (or VMs). Network virtualization is the next step after compute virtualization.

This document defines requirements for a network virtualization solution that provides BGP/MPLS IP VPN style connectivity to virtual resources on end-systems/end-device, such as a server, operating in a multi-tenant shared physical infrastructure. The requirements addresses the needs of virtual resources, applications reside on VMs, and focus on the appliances that require only IP connectivity. Non-IP communication is addressed by other documents and is not in scope of this document.

The technical solutions to support these requirements are work in progress in IETF. [I-D.ietf-l3vpn-end-system], [I-D.fang-l3vpn-virtual-pe]. The solutions may referred as End-System solutions or virtual PE (vPE) solutions in different documents.

### 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
CE	Customer Edge router
End-System	A device where Guest OS, Host OS/Hypervisor reside
GRE	Generic Routing Encapsulation
Hypervisor	Virtual Machine Manager
PE	Provider Edge router
RT	Route Target
RTC	RT Constraint
SDN	Software Defined Network
ToR	Top-of-Rack switch

VM	Virtual Machine
vPE	virtual Provider Edge Router
VPN	Virtual Private Network

## 2. Application of MPLS/BGP VPNs to End-Systems

MPLS/BGP VPN technology [RFC4364] have proven to be able to very scale to a large number of VPNs (tens of thousands) and customer routes (millions) while providing for aggregated management capability. In traditional WAN deployments of BGP IP VPNs a Customer Edge (CE) is a physical device, residing a customer's location, connected to a Provider Edge (PE), residing in a Service Provider's location. CE devices are logically part of a customer's VPN while PE routers are logically part of the SP's network. In a traditional MPLS/BGP VPN deployment, a CE device is a router and it is a routing peer of a PE to which it is attached via an attachment circuit.

In addition, the forwarding function and control function of a Provider Edge (PE) device co-exist within a single physical router.

MPLS/BGP VPN technology can be evolved and adapted to new virtualized environments by implementing the VPN edge functionality of the PE line-cards on the end-system hosts and thereby extending VPN service directly to end-systems.

### 2.1. End-System CE and PE Functions

When end-system/end-device attaches to MPLS/BGP VPN, CE corresponds to a non-routing host that can reside in a VM or be an application residing on the end-system itself.

As in traditional MPLS/BGP VPN deployments, it is undesirable for the end-system VPN forwarding knowledge to extend to the core transport network infrastructure. Hence, optimally, with regard to forwarding, the end-system should become both the CE and the PE simultaneously.

The network virtualization solution should also support deployments where it is not possible or not desirable to co-locate the PE and CE functionality. In such deployments PE may be implemented on an external device with remote CE attachments. This external PE device should be as close as possible to the end-system where the CE resides. The external PE devices that attach to a particular VPN, need to know, for each attachment circuit leading to that VPN, the host address that is reachable over that attachment circuit. The end-system MPLS/BGP VPN solution must specify a method to convey this information from the end-system to the PE.

The same network virtualization solution should support deployments

with mixed, internal (co-located with CE) and external PE (i.e., remote CE) implementations.

## 2.2. PE Control Plane Function

It is a current practice to implement MPLS/BGP VPN PE forwarding and control functions in different processors of the same device and to use internal (proprietary) communication between those processors. Typically, the PE control functionality is implemented in one (or very few) components of a device and the PE forwarding functionality is implemented in multiple components of the same device (a.k.a., "line cards").

In end-system environment, a single end-system, effectively, corresponds to a line card in a traditional PE router. For scalable and cost effective deployment of end-system MPLS/BGP VPNs the PE forwarding function should be decoupled from PE control function such that the former can be implemented on multiple standalone devices. This separation of functionality will allow for implementing the end-system PE forwarding on multiple end-system devices, for example, in operating systems of application servers or network appliances. Moreover, the separation of PE forwarding and control plane functions allows for the PE control plane function to be itself virtualized and run as an application in end-system.

## 3. VPN Communication Requirements

### 3.1. Unicast IPv4 and IPv6

A network virtualization solution should be able to provide IPv4 and IPv6 unicast connectivity between hosts in the same and different subnets without any assumptions regarding the underlying media layer.

### 3.2. Multicast/VPN Broadcast IPv4 and IPv6

Furthermore, the multicast transmission, i.e., allowing IP applications to send packets to a group of IPv4 or IPv6 addresses should be supported. The multicast service should also support a delivery of traffic to all endpoints of a given VPN even if those endpoints have not sent any control messages indicating the need to receive that traffic. In other words, the multicast service should be capable of delivering the IP broadcast traffic in a virtual topology. A solution for supporting VPN multicast and VPN broadcast must not require that the underlying transport network supports IP multicast transmission service.

### 3.3. IP Subnet Support

In some deployments, Virtual Machines or applications are configured to belong to an IP subnet. A network virtualization solution should support grouping of virtual resources into IP subnets regardless of whether the underlying implementation uses a multi-access network or not. While some applications may expect to find other peers in a particular user defined IP subnet, this does not imply the need to provide a layer 2 service that preserves MAC addresses. End-system network virtualization solution should be able to provide IP (unicast, multicast, VPN broadcast) connectivity between hosts in the same and different subnets without any assumptions regarding the underlying media layer.

#### 4. Multi-Tenancy Requirements

One of the main goals of network virtualization is to provide traffic and routing isolation between different virtual components that share a common physical infrastructure. Networks use various VPN technologies to isolate disjoint groups of virtual resources. Some use VLANs [IEEE.802-1Q] as a VPN technology, others use layer 3 based solutions, often with proprietary control planes. Service Providers are interested in interoperability and in openly documented protocols rather than in proprietary solutions. Further more, it is more favorable if the solution can provide Open Source codes in public forums, this will give the most flexibility and agility for SPs to create new services.

A collection of virtual resources might provide external or internal services. Such collection may serve an external "customer" or internal "tenant" to whom a Service Provider provides service(s). In MPLS/BGP VPN terminology a collection of virtual resources dedicated to a process or application corresponds to a VPN.

A network virtualization multi-tenancy solution should support the following:

- Tenant or application isolation, in data plane and control plane, while sharing the same underlying physical network. Tenants should be able to independently select and deploy their choice of IP address space: public or private IPv4 and/or IPv6.
- Multiple distinct VPNs per tenant. Tenant's inter-VPN traffic should be allowed to cross VPN boundaries, subject to access controls and/or routing policies.
- Inter-VPN communication, subject to access policies. Typically VPNs that belong to different external tenants do not communicate with each other directly but they should be allowed to access shared services or shared network resources. It is often the case

that SP infrastructure services are provided to multiple tenants, for example voice-over-IP gateway services or video-conferencing services for branch offices.

- VM or application end-point should be able to directly access multiple VPNs without a need to traverse a gateway.
- End-system network virtualization solution should support both, isolated VPNs as well as overlapping VPNs (often referred to as "extranets"). It should also support any-to-any and hub-and-spoke topologies.

## 5. Decoupling of Virtualized Networking from Physical Infrastructure

One of the main goals in designing a large scale transport network is to minimize the cost and complexity of its "fabric" by delegating the virtual resource communication processing to the network edge. It has been proven (in Internet and in large MPLS/BGP VPN deployments) that moving complexity to network edge while keeping network core simple has very good scaling properties.

The transport network infrastructure should not maintain any information that pertains to the virtual resources in end-systems. Decoupling of virtualized networking from the physical infrastructure has the following advantages: 1) provides better scalability; 2) simplifies the design and operation; 3) reduces network cost.

Decoupling of virtualized networking from underlying physical network consists in the following:

- Separation between the virtualized segments (i.e., interface associated with virtual resources) and the physical network (i.e., physical interfaces associated with network infrastructure).
- Separation of the virtual network IP address space from the physical infrastructure network IP address space. In the case of a transport other than IP, for example MPLS or Ethernet, the infrastructure address refers to the Subnetwork Point of Attachment (SNPA) address in a given multi-access network.
- The physical infrastructure addresses should be routable (or switchable) in the underlying transport network, while the virtual network addresses should be routable only in the virtual network.
- The virtual network control plane should be decoupled from the underlying transport network.

## 6. Decoupling of Layer 3 Virtualization from Layer 2 Topology

The layer 3 approach to network virtualization dictates that the virtualized communication should be routed, not bridged. The layer 3 virtualization solution should be decoupled from the layer 2 topology. Thus, there should be no dependency on VLANs and layer 2 broadcast.

In solutions that depend on layer 2 broadcast domains, host-to-host communication is established based on flooding and data plane MAC learning. Layer 2 MAC information has to be maintained on every switch where a given VLAN is present. Even if some solutions are able to minimize data plane MAC learning and/or unicast flooding, they still rely on MAC learning at the network edge and on maintaining the MAC addresses on every switch where the layer 2 VPN is present.

The MAC addresses known to guest OS in end-system are not relevant to IP services and introduce unnecessary overhead. Hence, the MAC addresses associated with virtual resources should not be used in the virtual layer 3 networks. Rather, only what is significant to IP communication, namely the IP addresses of the virtual machines and application endpoints should be maintained by the virtual networks.

## 7. Requirements for Encapsulation of Virtual Payloads

In order to scale the transport networks, the virtual network payloads must be encapsulated with headers that are routable (or switchable) in the physical network infrastructure. The IP addresses of the virtual resources are not to be advertised within the physical infrastructure address space.

The encapsulation (and de-encapsulation) function should be implemented on a device as close to virtualized resources as possible. Since the hypervisors in the end-systems are the devices at the network edge they are the most optimal location for the encap/decap functionality.

The network virtualization solution should also support deployments where it is not possible or not desirable to implement the virtual payload encapsulation in the hypervisor/Host OS. In such deployments encap/decap functionality may be implemented in an external device. The external device implementing encap/decap functionality should be as close as possible to the end-system itself. The same network virtualization solution should support deployments with both, internal (in a hypervisor) and external(outside of a hypervisor) encap/decap devices.

Whenever the virtual forwarding functionality is implemented in an external device, the virtual service itself must be delivered to an end-system such that switching elements connecting the end-system to the encap/decap device are not aware of the virtual topology.

### 7.1. Encapsulation Methods

MPLS/VPN technology based on [RFC4364] specifies that different encapsulation methods could be for connecting PE routers, namely Label Switched Paths (LSPs), IP tunneling, and GRE tunneling.

If LSPs are used in the transport network they could be signaled with LDP, in which case host (/32) routes to all PE routers must be propagated throughout the network, or with RSVP-TE, in which case a full mesh of RSVP-TE tunnels is required.

If the transport network is only IP-capable then MPLS in IP or MPLS in GRE [RFC4023] encapsulation could be used. Due to route aggregation property of IP protocols, with IP/GRE encapsulation the PE host routes do not have to be present in the transport network.

Multi-access technologies, especially Ethernet, may also need to be supported as transport networks, for example, 802.1ah.

### 7.2. Routing of Virtual Payloads

A device implementing the encap/decap functionality acts as the first-hop router in the virtual topology.

In a layer 3 end-system virtual network, IP packets should reach the first-hop router in one IP-hop, regardless of whether the first-hop router is an end-system itself (i.e., a hypervisor/Host OS) or it is an external (to end-system) device. The first-hop router should always perform an IP lookup on every packet it receives from a virtual machine or an application. The first-hop router should encapsulate the packets and route them towards the destination end-system.

## 8. Optimal Forwarding of Traffic

The network virtualization solutions that optimize for the maximum utilization of compute and storage resources require that those resources may be located anywhere in the network. The physical and logical spreading of appliances and workloads implies a very significant increase in the infrastructure bandwidth consumption. In order to be efficient in terms of traffic forwarding, the virtualized networking solutions must assure that packets traverse the transport network only once.

It must be also possible to send the traffic directly from one end-system to another end-system without traversing through a midpoint router.

## 9. IP Mobility

Another reason for a network virtualization is the need to support IP mobility. IP mobility consists in IP addresses used for communication within or between applications being located anywhere across the virtual network. Using a virtual topology, i.e., abstracting the externally visible network address from the underlying infrastructure address is an effective way to solve IP mobility problem.

IP mobility consists in a device physically moving (e.g., a roaming wireless device) or a workload being transferred from one physical server/appliance to another. IP mobility requires preserving device's active network connections (e.g., TCP and higher-level sessions). Such mobility is also referred to as "live" migration with respect to a Virtual Machine. IP mobility is highly desirable for many reasons such as efficient and flexible resource sharing, data center migration, disaster recovery, server redundancy, or service bursting.

### 9.1. IP Addressing of Virtual Hosts

To accommodate live mobility of a virtual machine (or a device), it is desirable to assign to it a semi-permanent IP address that remains with the VM/device as it moves. The semi-permanent IP address can be configured through VM configuration process or by means of DHCP.

### 9.2. Network Layer-Based Mobility

When dealing with IP-only applications it is not only sufficient but optimal to forward the traffic based on layer 3 (network layer) rather than on layer 2 (data-link layer) information. The MAC addresses of devices or applications are irrelevant to IP services and introduce unnecessary overhead and complications when devices or VMs move. For example, when a VM moves between physical servers, the MAC learning tables in the switches must be updated. Moreover, it is possible that VM's MAC address might need to change in its new location. In IP-based network virtualization solution a device or a workload move is handled by an IP route advertisement.

### 9.3. Routing Convergence Requirements

IP mobility has to be transparent to applications and any external entity interacting with the applications. This implies that the network connectivity restoration time is critical. The transport sessions can typically survive over several seconds of disruption, however, applications may have sub-second latency requirement for their correct operation.

To minimize the disruption to established communication during



workload or device mobility, the control plane of a network virtualization solution should be able to differentiate between the activation of a workload in a new location from advertising its route to the network. This will enable the remote end-points to update their routing tables prior to workload's migration as well as allowing the traffic to be tunneled via the workload's old location.

#### 10. Inter-operability with Existing MPLS/BGP VPNs

Service Providers want to tie their server-based offerings to their MPLS/BGP VPN services. MPLS/BGP VPNs provide secure and latency-optimized remote connectivity to the virtualized resources in SP's data center. The Service Provider-based VPN access can provide additional capabilities compared with public internet access, such as QoS, OAM, multicast service, VoIP service, video conferencing, wireless connectivity.

MPLS/BGP VPN customers may require simultaneous access to resources in both SP and their own data centers.

Service Providers want to "spin up" the L3VPN access to data center VPNs as dynamically as the spin up of compute and other virtualized resources.

The network virtualization solution should be fully inter-operable with MPLS/BGP VPNs, including:

- Inter-AS MPLS/BGP VPN Options A, B, and C [RFC4364].
- BGP/MPLS VPN-capable network devices (such as routers and network appliances) should be able to participate directly in a virtual network that spans end-systems.
- The network devices should be able to participate in isolated collections of end-systems, i.e., in isolated VPNs, as well as in overlapping VPNs (called "extranets" in BGP/MPLS VPN terminology).
- The network devices should be able to participate in any-to-any and hub-and-spoke end-systems topologies.

When connecting an end-system VPN with other services/networks, it should not be necessary to advertize the specific host routes but rather the aggregated routing information. A BGP/MPLS VPN-capable router or appliance can be used to aggregate VPN's IP routing information and advertize the aggregated prefixes. The aggregated prefixes should be advertized with the router/appliance IP address as BGP next-hop and with locally assigned aggregate 20-bit label. The aggregate label should trigger a destination IP lookup in its

corresponding VRF on all the packets entering the virtual network.

The inter-connection of end-system VPNs with traditional VPNs requires an integrated control plane and unified orchestration of network and end-system resources.

## 11. BGP Requirements in a Virtualized Environment

### 11.1. BGP Convergence and Routing Consistency

BGP was designed to carry very large amount of routing information but it is not a very fast converging protocol. In addition, the routing protocols, including BGP, have traditionally favored convergence (i.e., responsiveness to route change due to failure or policy change) over routing consistency. Routing consistency means that a router forwards a packet strictly along the path adopted by the upstream routers. When responsiveness is favored, a router applies a received update immediately to its forwarding table before propagating the update to other routers, including those that potentially depend upon the outcome of the update. The route change responsiveness comes at the cost of routing blackholes and loops.

Routing consistency in virtualized environments is important because multiple workloads can be simultaneously moved between different physical servers due to maintenance activities, for example. If packets sent by the applications that are being moved are dropped (because they do not follow a live path), the active network connections will be dropped. To minimize the disruption to the established communications during VM migration or device mobility, the live path continuity is required.

#### 11.1.1. BGP IP Mobility Requirements

In IP mobility, the network connectivity restoration time is critical. In fact, Service Provider networks already use routing and forwarding plane techniques that support fast failure restoration by pre-installing a backup path to a given destination. These techniques allow to forward traffic almost continuously using an indirect forwarding path or a tunnel to a given destination, and hence, are referred to as "local repair". The traffic path is restored locally at the destination's old location while the network converges to a backup path. Eventually, the network converges to an optimal path and bypasses the local repair. BGP assists in the local repair techniques by advertizing multiple and not only the best path to a given destination.

### 11.2. Optimization of Route Distribution

When virtual networks are triggered based on the IP communication, the Route Target Constraint extension [RFC4684] of BGP should be used to optimize the route distribution for sparse virtual network events. This technique ensures that only those VPN forwarders that have local participants in a particular data plane event receive its routing information. This also decreases the total load on the upstream BGP speakers.

### 11.3. Service chaining

It is important to provide service chaining ability without major impact to the existing protocols deployed. One solution currently work in progress in IETF is [I-D.rfernando-l3vpn-service-chaining].

## 12. Security Considerations

The document presents the requirements for end-systems MPLS/BGP VPNs. The security considerations for traditional MPLS/BGP VPN deployments are described in [RFC4364] in Section 13. Security issues associated with deployments using MPLS-in-GRE or MPLS-in-IP encapsulations are described in [RFC4023] in Section 8. And [RFC4111] provides general IP VPN security guidelines.

The additional security requirements specific to end-system MPLS/BGP VPNs are as follows:

- End-systems MPLS/BGP VPNs solution should guarantee that packets originating from a specific end-system virtual interface are accepted only if the corresponding VPN IP host is present on that end-system.
- Virtual network must ensure that traffic arriving at the egress end-system is being sent from the correct ingress end-system.
- One virtual host or VM should not be able to impersonate another, during steady-state operation and during live migration.

The security considerations for specific solutions will be documented in the relevant documents.

## 13. IANA Considerations

This document contains no new IANA considerations.

## 14. References

### 14.1. Normative References

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

- [RFC4023] Worster, T., Rekhter, Y., and E. Rosen, Ed.,  
"Encapsulating MPLS in IP or Generic Routing Encapsulation  
(GRE)", RFC 4023, March 2005.
- [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.
- [IEEE.802-1Q] Institute of Electrical and Electronics Engineers,  
"Local and Metropolitan Area Networks: Virtual Bridged  
Local Area Networks", IEEE Std 802.1Q-2005, May 2006.

#### 14.2. Informative References

- [RFC4111] Fang, L., Ed., "Security Framework for Provider-  
Provisioned Virtual Private Networks (PPVPNs)", RFC 4111,  
July 2005.
- [I-D.ietf-l3vpn-end-system] Marques, P., Fang, L., Pan, P., Shukla,  
A., Napierala, M., "BGP-sigaled end-system IP/VPNs",  
draft-ietf-l3vpn-end-system, work in progress.
- [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, work  
in progress.
- [I-D.rfernando-l3vpn-service-chaining] Fernando, R., Rao, D., Fang,  
L., Napierala, M., So, N., draft-rfernando-l3vpn-service-  
chaining, work in progress.

#### 15. Acknowledgements

The authors would like to thank Pedro Marques and Han Nguyen for the  
comments and suggestions.

#### Authors' Addresses

Maria Napierala  
AT&T  
200 Laurel Avenue  
Middletown, NJ 07748

Email: mnapierala@att.com

Luyuan Fang

Cisco

111 Wood Avenue South

Iselin, NJ 08830, USA

Email: luyuanf@gmail.com

INTERNET-DRAFT  
Intended Status: Standards track  
Expires: April 18, 2014

Luyuan Fang  
John Evans  
David Ward  
Rex Fernando  
John Mullooly  
Cisco  
Ning So  
Tata Communications  
Nabil Bitar  
Verizon  
Maria Napierala  
AT&T

October 18, 2013

BGP/MPLS IP VPN Virtual CE  
draft-fang-l3vpn-virtual-ce-02

## Abstract

This document describes the architecture and solutions of using virtual Customer Edge (vCE) of BGP IP MPLS VPN. The solution is aimed at providing efficient service delivery capability through CE virtualization, and is especially beneficial in virtual Private Cloud (vPC) environments for extending BGP/MPLS IP VPN into tenant virtual Data Center containers.

## 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 Problem statement . . . . .	5
1.3 Scope of the document . . . . .	6
2. Virtual CE Architecture and Reference Model . . . . .	6
2.1 Virtual CE . . . . .	6
2.2 Architecture . . . . .	7
3. Control Plane . . . . .	10
3.1 vCE Control Plane . . . . .	10
4. Forwarding Plane . . . . .	10
4.1 Forwarding between vCE and PE/vPE . . . . .	11
4.2 Forwarding between vCE and VM . . . . .	11
5. Addressing and QoS . . . . .	11
5.1 Addressing . . . . .	11
5.2 QoS . . . . .	12
6. Management plane . . . . .	12
6.1 Network abstraction and management . . . . .	12
6.2 Service VM Management . . . . .	12
7. Orchestration and IP VPN inter-provisioning . . . . .	12
7.1 DC Instance to WAN BGP/MPLS IP VPN instance "binding" Requirements . . . . .	12
7.2. Provisioning/Orchestration . . . . .	13
7.2.1 vCE Push model . . . . .	13
7.2.1.1 Inter-domain provisioning vCE Push Model . . . . .	14
7.2.1.2 Cross-domain provisioning vCE Push Model . . . . .	14
7.1.1 vCE Pull model . . . . .	15
8. Security Considerations . . . . .	16
9. IANA Considerations . . . . .	16
10. References . . . . .	16
10.1 Normative References . . . . .	16
10.2 Informative References . . . . .	17

11. Acknowledgement . . . . .	17
Authors' Addresses . . . . .	18



## 1. Introduction

In the typical enterprise BGP/MPLS IP VPN [RFC4364] deployment, the Provider Edge (PE) and Customer Edge (CE) are physical routers which support the PE and CE functions. With the recent development of cloud services, using virtual instances of PE or CE functions, which reside in a compute device such as a server, can be beneficial to emulate the same logical functions as the physical deployment model but now achieved via cloud based network virtualization principles. This would be considered as part of the Network functions Virtualization (NFV) effort.

This document describes BGP/MPLS IP VPN virtual CE (vCE) solutions, while Virtual PE (vPE) concept and implementation options are discussed in [I-D.fang-l3vpn-virtual-pe], [I-D.ietf-l3vpn-end-system]. vPE and vCE solutions provide two avenues to realize network virtualization.

### 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
-----	-----
AAA	Authentication, Authorization, and Accounting
ACL	Access Control List
AS	Autonomous Systems
ASBR	Autonomous Systems Border Router
BGP	Border Gateway Protocol
CE	Customer Edge
DB	Data Base
DMZ	Demilitarized Zone, a.k.a. perimeter networking
FE	Front End
FTP	File Transfer Protocol
GRE	Generic Routing Encapsulation
HTTP	Hypertext Transfer Protocol
Hypervisor	Virtual Machine Manager
I2RS	Interface to Routing System
LDAP	Lightweight Directory Access Protocol
MP-BGP	Multi-Protocol Border Gateway Protocol
NAT	Network Address Translation
NVGRE	Network Virtualization using GRE
PE	Provider Edge
QinQ	Provider Bridging, stacked VLANs
RR	Route Reflector
SDN	Software Defined Network

SLA	Service Level Agreement
SMTP	Simple Mail Transfer Protocol
ToR	Top of the Rack switch
vCE	virtual Customer Edge Router
vLB	virtual Load Balancer
VM	Virtual Machine
VLAN	Virtual Local Area Network
vPE	virtual Provider Edge Router
VPN	Virtual Private Network
vSG	virtual Security Gateway
VXLAN	Virtual eXtensible Local Area Network
WAN	Wide Area Network

Virtual CE (vCE): A virtual instance of the Customer Edge (CE) routing function which resides in one or more network or compute devices. For example, the vCE data plane may reside in an end device, such as a server, and as co-resident with application Virtual Machines (VMs) on the server; the vCE control plane may reside in the same device or in a separate entity such as a controller.

End device: A device where Guest OS, Host OS/Hypervisor, applications, VMs, and virtual router may reside.

Network Container/Tenant Container: An abstraction of a set of network and compute resources which can be physical and virtual, providing the cloud services for a tenant. One tenant can have more than one Tenant Containers.

Zone: A logical grouping of VMs and service assets within a tenant container. Different security policies may be applied within and between zones.

DMZ: Demilitarized zone, a.k.a. perimeter networking. It is often a machine or a small subnet that sits between a trusted internal network, such as a corporate private LAN, and an un-trusted external network, such as the public Internet. Typically, the DMZ contains devices accessible to Internet traffic, such as Web (HTTP) servers, FTP servers, SMTP (e-mail) servers and DNS servers.

## 1.2 Problem statement

With the growth of cloud services and the increase in the number of CE devices, routers/switches, and appliances, such as Firewalls (FWs) and Load Balancers (LBs), that need to be supported, it is beneficial to virtualize the Data Center tenant container. The virtualized container can increase resource sharing, optimize routing and forwarding of inter-segment and inter-service traffic, and allow simplified design, provisioning, and management.

The following two aspects of the virtualized Data Center tenant container for the IP VPN CE solution are discussed in this document.

#### 1. Architecture re-design for virtualized DC.

The optimal architecture of the virtualized container includes virtual CE, virtual appliances, and application VMs. All these functions are co-residents on virtualized servers. CEs and appliances can be created and removed easily on demand, and the virtual CE can interconnect the virtual appliances (e.g., FW, LB, NAT), applications (e.g., Web, App., and DB) in a co-located fashion for simplicity, routing/forwarding optimization, and easier service chaining. Virtualizing these functions on a per-tenant basis provides simplicity for the network operator in regards to managing per tenant service orchestration, tenant container moves, capacity planning across tenants and per-tenant policies.

#### 2. Provisioning/orchestration. Two issues need to be addressed:

- a) The provisioning/orchestration system of the virtualized data center need to support VM life cycle and VM migration.
- b) The provisioning/orchestration systems of the DC and the WAN networks need to be coordinated to support end-to-end BGP/MPLS IP VPN from DC to DC or from DC to enterprise remote offices in the same VPN. The DC and the WAN network are often operated by separate departments, even if they belong to the same provider. Today, the process of inter-connecting is often slow and painful, and automation is highly desirable.

#### 1.3 Scope of the document

As the majority (all in some networks) of applications are IP, this vCE solution is focusing on IP VPN solutions to cover the most common cases and keep matters as simple as possible.

### 2. Virtual CE Architecture and Reference Model

#### 2.1 Virtual CE

As described in [RFC4364], IP uses a "peer model" - the customers' edge routers (CE routers) exchange routes with the Service Provider's edge routers (PE routers); the CEs do not peer with each other. MP-BGP [RFC4271, RFC4760] is used between the PEs (often with RRs) which have a particular VPN attached to them to exchange the VPN routes. A CE sends IP packets to the PE; no VPN labels for packets forwarded between CE and PE.

A virtual CE (vCE) is a software instance of BGP/MPLS IP VPN CE function which can reside in ANY network or compute devices. For example, a vCE MAY reside in an end device, such as a server in a Data Center, where the application VMs reside.

Using the virtual CE model, the CE functions CAN easily co-located with the VM/applications, e.g., in the same server. This allows tenant inter-segment and inter-service routing to be optimized. Likewise the vCE can be in a separate server (in the same DC rack or across racks) than the application VMs, in which case VMs would typically use standard L2 technologies to access the vCE via the DC network.

Similar to the virtual PE solution, the control and forwarding of a virtual CE can be on the same device, or decoupled and reside on different physical devices. The provisioning of a virtual CE, associated applications, and the tenant network container can be supported through DC orchestration systems.

Unlike a physical or virtual PE which can support multi-tenants, a physical or virtual CE supports a single tenant only. A single tenant CAN use multiple physical or virtual CEs. An end device, such as a server, CAN support one or more vCE(s). While the vCE is defined as a single tenant device, each tenant can have multiple logical departments which are under the tenant administrative control, requiring logical separation, this is the same model as today's physical CE deployments.

vCE and vPE are complimentary approaches for extending IP VPN into tenant containers. In the vCE solution, there is no BGP/MPLS IP VPN within the data center or other type of service network, the vCE can connect to the PE which is a centralized BGP/MPLS IP VPN PE/ASBR/DC Gateway, or connect to distributed vPE on a server or on the Top of the Rack switch (ToR). vCE can be used to extend the existing SP managed CE solution to create new cloud enabled services and provide the same topological model and features that are consistent with the physical CE systems.

## 2.2 Architecture

Figure 1 illustrates the topology where vCE is resident in the servers where the applications are hosted.

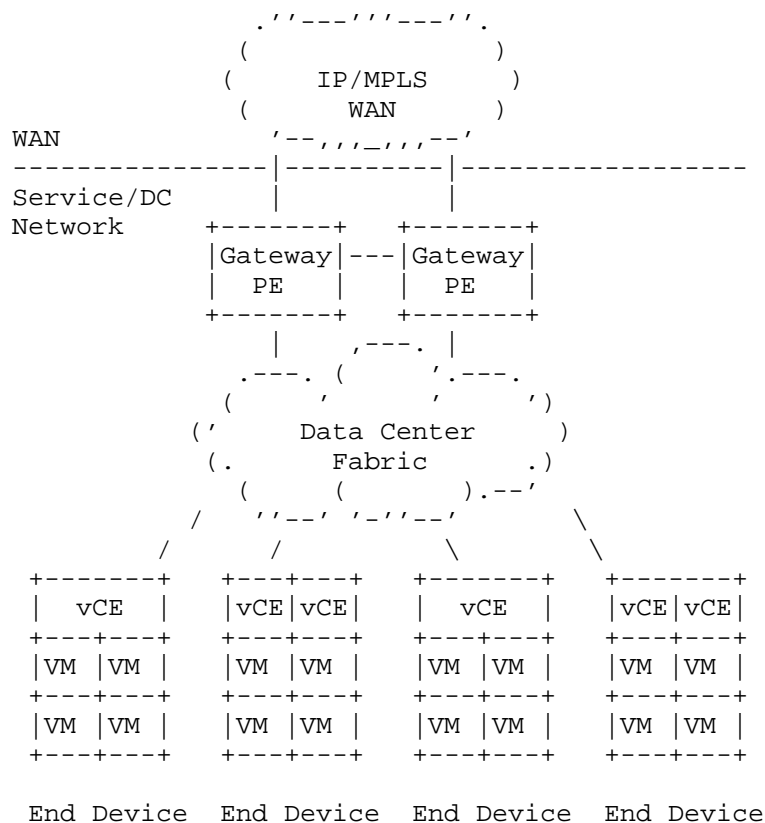


Figure 1. Virtualized Data Center with vCE

Figure 1 above illustrate a vCE solution in a virtualized Data Center with application VMs on the servers. One or more vCEs MAY be used on each server.

The vCEs logically connect to the PEs/Gateway to join the particular BGP/MPLS IP VPN which the tenant belongs to. Gateway PEs connect to the BGP/MPLS IP VPN in the WAN network for inter-DC and DC to enterprise VPN sites connection. The server physically connects to the DC Fabric for packet forwarding.

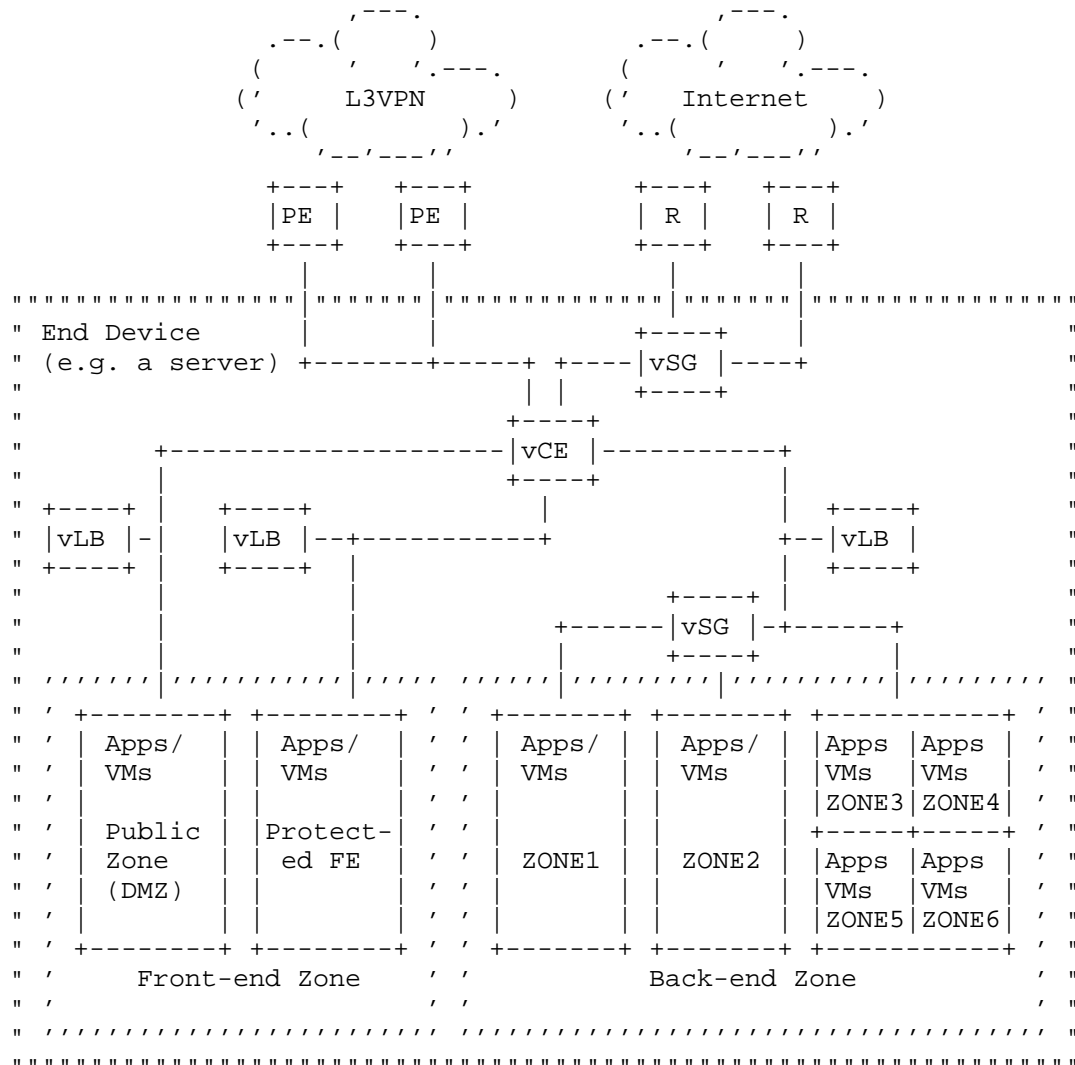


Figure 2. A Virtualized Container with vCE in an End Device

An end device shown in Figure 2 is a physical server supporting multiple virtualized appliances and applications, and hosts multiple client VMs.

In the traditional deployment, the topology often involves multiple physical CEs, physical Security Gateways and Load Balancers residing in the same Data Center.

The virtualized approach provides the benefit of reduced number of physical devices, simplified management, optimal routing due to the co-location of vCE, services, and client VMs.

While the above diagram represents a simplified view of all of the tenant service and application VMs residing in the same physical server, the above model can also be represented with the VMs spread across many physical servers and the DC network would provide the physical inter-connectivity while the vCE and the VMs connected to the vCE form the logical connections.

### 3. Control Plane

#### 3.1 vCE Control Plane

The vCE control plane can be distributed or centralized.

##### 1) Distributed control plane

vCE CAN exchange BGP routes with PE or vPE for the particular BGP/MPLS IP VPN as described in [RFC4364]. The vCE must support BGP if this approach is used.

The advantage of using distributed protocols is to avoid single point of failure and bottleneck. Service chaining can be easily and efficiently supported in this approach.

BGP as PE-CE protocol is used in majority deployment in typical Enterprise BGP/MPLS IP VPN PE-CE connections. BGP supports rich policy compared to other alternatives.

2) Static routing. It is also used in Enterprise BGP/MPLS IP VPN PE-CE connections based on past observation. It MAY be used if the operator prefers.

##### 2. Using controller approach

Controller can be used as part of the Software Defined Network (SDN) approach. A controller can be distributed or centralized, or physically distributed and logically centralized. The controller performs the control plane functions, and sends instructions to the vCE on the end devices to configure the data plane.

This requires standard interface to routing system (I2RS). The Interface to Routing System (I2RS) is work in progress in IETF [I-D.ietf-i2rs-architecture], [I-D.ietf-i2rs-problem-statement].

### 4. Forwarding Plane

#### 4.1 Forwarding between vCE and PE/vPE

No MPLS forwarding is required between PE and CE in typical PE-CE connection scenarios, though MPLS label forwarding is required for implementing Carriers' Carrier (CSC) model.

IPv4 and IPv6 packet forwarding MUST be supported.

Native fabric CAN be used to support isolation between vCEs to PE connections.

Examples of native fabric include:

- VLANs [IEEE 802.1Q], Virtual Local Area Network
- IEEE 802.1ad [IEEE 802.1ad]/QinQ, Provider Bridge

Or overlay segmentation with better scalability:

- VXLANs, Virtual Extensible LAN, work in progress in IETF, [I-D.mahalingam-dutt-dcops-vxlan].
- NVGRE, Network Virtualization using GRE, work in progress in IETF [I-D.sridharan-virtualization-nvgre].

#### 4.2 Forwarding between vCE and VM

If the vCE and the VM that the vCE is connecting are co-located in the same server, the connection is internal to the server, no external protocol involved.

If the vCE and the VM that the vCE is connecting are located in different devices, standard external protocols are needed. The forwarding can be native or overlay techniques as listed in the above sub-section.

### 5. Addressing and QoS

#### 5.1 Addressing

IPv4 and IPv6 addressing MUST be supported.

IP address allocation for vCEs and applications/client:

- 1) IP address MAY be assigned by central management/provisioning with predetermined blocks through planning process.
- 2) IP address MAY be obtained through DHCP server.



Address space separation: The IP addresses used for clients in the BGP/MPLS IP VPNs in the DC SHOULD be in separate address blocks outside the blocks used for the underlay infrastructure of the DC. The purpose is to protect the DC fabric from being attacked if the attacker gain access of the tenant VPNs.

## 5.2 QoS

Differentiated Services [RFC2475] Quality of Service (QoS) is standard functionality for physical CEs and MUST be supported on vCE. This is important to ensure seamless end-to-end SLA from BGP/MPLS IP VPN in the WAN into service network/Data center. The use of MPLS Diffserv tunnel model Pipe Mode (RFC3270) with explicit null LSP must be supported.

## 6. Management plane

### 6.1 Network abstraction and management

The use of vCE with single tenant virtual service instances can simplify management requirements as there is no need to discover device capabilities, track tenant dependencies and manage service resources.

vCE North bound interface SHOULD be standards based.

The programmatic interface are currently under definition in the IETF's Interface to Routing Systems (I2RS) initiative, [I-D.ietf-i2rs-architecture], [I-D.ietf-i2rs-problem-statement].

vCE element management MUST be supported, it can be in the similar fashion as for physical CE, without the hardware aspects.

### 6.2 Service VM Management

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

The management tools SHOULD be open standards.

## 7. Orchestration and IP VPN inter-provisioning

### 7.1 DC Instance to WAN BGP/MPLS IP VPN instance "binding" Requirements

- MUST support service activation in the physical and virtual environment.

For example, assign VLAN to correct VRF.

- MUST support per VLAN Authentication, Authorization, and Accounting (AAA).

The PE function is an OAM boundary.

- MUST be able to apply other policies to VLAN.

For example, per VLAN QOS, ACLs.

- MUST ensure that WAN BGP/MPLS IP VPN state and DC state are dynamically synchronized.

Ensure that there is no possibility of customer being connected to the wrong VRF. For example, remove all tenant state when service an instance is terminated.

- MUST integrate with existing WAN BGP/MPLS IP VPN provisioning processes.
- MUST scale to 10,000 or higher tenant service instances.
- MUST cope with rapid (sub minute) tenant mobility.
- SHOULD support automated cross provisioning accounting correlation between WAN BGP/MPLS IP VPN and Cloud/DC for the same tenant.
- MAY support Automated cross provisioning state correlation between WAN BGP/MPLS IP VPN and Cloud/DC for the same tenant.

## 7.2. Provisioning/Orchestration

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

### 7.2.1 vCE Push model

Push model: It is a top down approach - push IP VPN provisioning from network management system or other central control provisioning systems to the IP VPN network elements.

This approach supports service activation and it is commonly used in the existing BGP/MPLS IP VPN enterprise deployment. When extending BGP/MPLS IP VPN solution into the Cloud/DC, it MUST support off-line accounting correlation between the WAN BGP/MPLS 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, bind interface state to tenant.

#### 7.2.1.1 Inter-domain provisioning vCE Push Model

Provisioning process:

- 1) Cloud/DC orchestrator configures vCE.
- 2) Orchestrator initiates WAN IP VPN provisioning; passes connection IDs (e.g., of VLAN/VXLAN/NVGRE) and tenant context to WAN IP VPN provisioning systems.
- 3) WAN IP VPN provisioning system provisions PE VRF and other policies per normal enterprise IP VPN provisioning processes.

This model requires the following:

- The DC orchestration system or the WAN IP VPN provisioning system know the topology inter-connecting the DC and WAN VPN. For example, which interface on the WAN core device connects to which interface on the DC PE.
- Offline state correlation.
- Offline accounting correlation.
- Per SP integration.

Dynamic BGP session 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. Other protocols can be used for this purpose as well, for example, use Enhanced Interior Gateway Routing Protocol (EIGRP) for dynamic neighbour relationship establishment.

The dynamic routing prevents the needs to configure the PEs in PE-vCE model.

Caution: This is only under the assumption that the DC provisioning system is trusted and could support dynamic establishment of PE-vCE BGP neighbor relationships, for example, the WAN network and the cloud/DC belongs to the same SP.

#### 7.2.1.2 Cross-domain provisioning vCE Push Model

Provisioning Process:

- 1) Cross-domain orchestration system initiates DC orchestration.
- 2) DC orchestration system configures vCE.

- 3) DC orchestration system passes back VLAN/VXLAN/NVGRE and tenant context.  
to cross-domain orchestration system
- 4) Cross-domain orchestration system initiates WAN IP VPN provisioning.
- 5) WAN IP VPN provisioning system provisions PE VRF and other policies as per normal enterprise IP VPN provisioning processes.

This model requires the following:

- Cross-domain orchestration system knows the topology connecting the DC and WAN IP VPN, for example, which interface on core device connects to which interface on DC PE.
- Offline state correlation.
- Offline accounting correlation.
- Per SP integration.

#### 7.1.1 vCE Pull model

Pull model: It is a bottom-up approach - 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 deployment. Dynamic accounting correlation and dynamic state correlation are supported. For example, session based accounting is implicitly includes tenant context state correlation, as well as session based state which implicitly includes tenant context.

Inter-domain Provisioning:

Process:

- 1) Cloud/DC orchestration system configures vCE.
- 2) Cloud/DC orchestration system primes WAN IP VPN provisioning/AAA for new service, passes connection IDs (e.g., VLAN/VXLAN/NVGRE) and tenant context WAN IP VPN provisioning systems.
- 3) Cloud/DC PE detects new VLAN, send Radius Access-Request.
- 4) Radius Access-Accept with VRF and other policies.

This model requires VLAN/VXLAN/NVGRE information and tenant context

to be passed on a per transaction basis. In practice, it may simplify to use DC orchestration updating LDAP directory.

Auto accounting correlation and auto state correlation are supported in this model.

## 8. Security Considerations

When vCE is created on a network or compute device, such as a server, the operator MUST evaluate the following conditions: Is server owned by the the operator? Is it using a managed CE model? How to authenticate? The ownership of the device where the vCE resides has major implication on the design, it determines where the boundary is between the trusted and un-trusted zones.

When a vCE in DC connecting BGP MPLS IP VPN in the WAN, the amount of information can be exchanged across the two domains through auto-provisioning will be different depending on if the DC and WAN are under same administrative domain. Only limited and/or abstracted information should be exchanged if the two domains are owned by different SPs. Additional authentication, and other security mechanism need to be deployed to prevent accidental or malicious attach from the other domain.

In addition, the connection authentication is very important for the pull models.

And the virtual FW placement needs to be carefully designed to protect against attacks.

## 9. IANA Considerations

None.

## 10. References

### 10.1 Normative References

- [RFC2119] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, March 1997.
- [RFC4271] Rekhter, Y., 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.

- [RFC4760] Bates, T., Chandra, R., Katz, D., and Y. Rekhter, "Multiprotocol Extensions for BGP-4", RFC 4760, January 2007.
- [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, work in progress.
- [I-D.fang-l3vpn-virtual-pe] Fang, L., et al., "BGP IP VPN Virtual PE", draft-fang-l3vpn-virtual-pe, work in progress.
- [IEEE 802.1ad] IEEE, "Provider Bridges", 2005.
- [IEEE 802.1q] IEEE, "802.1Q - Virtual LANs", 2006.
- [IEEE 802.1ag] IEEE "802.1ag - Connectivity Fault Management", 2007.

## 10.2 Informative References

- [RFC2475] Blake, S., Black, D., Carlson, M., Davies, E., Wang, Z., and W. Weiss, "An Architecture for Differentiated Services", RFC 2475, December 1998.
- [I-D.ietf-i2rs-architecture] Atlas, A., Halpern, J., Hares, S., Ward, D., and T Nadeau, "An Architecture for the Interface to the Routing System", draft-ietf-i2rs-architecture, work in progress.
- [I-D.ietf-i2rs-problem-statement] Atlas, A., Nadeau, T., and Ward D., "Interface to the Routing System Problem Statement", draft-ietf-i2rs-problem-statement, work in progress.
- [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, work in progress.
- [I-D.sridharan-virtualization-nvgre]: SridharanNetwork, M., et al., "Virtualization using Generic Routing Encapsulation", draft-sridharan-virtualization-nvgre, work in progress.

## 11. Acknowledgement

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

Authors' Addresses

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

John Evans  
Cisco  
16-18 Finsbury Circus  
London, EC2M 7EB, UK  
Email: joevans@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

John Mullooly  
Cisco  
111 Wood Ave. South  
Iselin, NJ 08830  
Email: jmullool@cisco.com

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

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

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

INTERNET-DRAFT  
Intended Status: Standards track  
Expires: April 18, 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

October 18, 2013

BGP/MPLS IP VPN Virtual PE  
draft-fang-l3vpn-virtual-pe-04

## 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, forwarding, and management. The proposed vPE solutions support both the Software Defined Networks (SDN) approach which allows physical decoupling of the control and the forwarding, and the traditional distributed routing approach. A vPE can reside in any network or compute devices, such as a server as co-resident with the application virtual machines (VMs), or a Top-of-Rack (ToR) switch in a Data Center (DC) network.

## 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	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	7
2.3	vPE Architecture reference models	7
2.3.1	vPE-F in an end-device and vPE-C in the controller	7
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	11
2.3.5	The server view of a vPE	11
3.	Control Plane	12
3.1	vPE Control Plane (vPE-C)	12
3.1.1	The SDN approach	12
3.1.2	Distributed control plane	13
3.3	Use of router reflector	13
3.4	Use of Constrained Route Distribution [RFC4684]	13
4.	Forwarding Plane	13
4.1	Virtual Interface	13
4.2	Virtual Provider Edge Forwarder (vPE-F)	14
4.3	Encapsulation	14

4.4 Optimal forwarding . . . . .	15
4.5 Routing and Bridging Services . . . . .	15
5. Addressing . . . . .	16
5.1 IPv4 and IPv6 support . . . . .	16
5.2 Address space separation . . . . .	16
6.0 Inter-connection considerations . . . . .	16
7. Management, Control, and Orchestration . . . . .	17
7.1 Assumptions . . . . .	17
7.2 Management/Orchestration system interfaces . . . . .	18
7.3 Service VM Management . . . . .	18
7.4 Orchestration and IP VPN inter-provisioning . . . . .	18
7.4.1 vPE Push model . . . . .	19
7.4.2 vPE Pull model . . . . .	20
8. Security Considerations . . . . .	21
9. IANA Considerations . . . . .	21
10. References . . . . .	21
10.1 Normative References . . . . .	21
10.2 Informative References . . . . .	22
Authors' Addresses . . . . .	23

## 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 Layer 3 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.

With the advent of compute capabilities and the proliferation of virtualization in Data Center servers, multi-tenant Data Centers are becoming the norm. As applications and appliances are increasingly being virtualized, support for virtual edge devices, such as virtual IP VPN PE routers, becomes feasible and desirable for Service Providers who want to extend their existing IP VPN deployments into Data Centers to provide end-to-end Virtual Private Cloud (VPC) services. Virtual PE work is also one of early effort for Network Functions Virtualization (NFV). In general, scalability, agility, and cost efficiency are primary motivations for vPE solutions.

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

The vPE solutions support both the Software Defined Networks (SDN) approach, which allows physical decoupling of the control and the forwarding, and the traditional distributed routing approach.

### 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
ASBR	Autonomous System Border Router
BGP	Border Gateway Protocol
CE	Customer Edge
Forwarder	IP VPN forwarding function
GRE	Generic Routing Encapsulation
Hypervisor	Virtual Machine Manager

I2RS	Interface to Routing Systems
LDP	Label Distribution Protocol
MP-BGP	Multi-Protocol Border Gateway Protocol
PCEF	Policy Charging and Enforcement Function
QoS	Quality of Service
RR	Route Reflector
RT	Route Target
RTC	RT Constraint
SDN	Software Defined Networks
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

End device: where Guest OS, Host OS/Hypervisor, applications, VMs, and virtual router may reside.

## 1.2 Requirements

The following are 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 DC 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 and forwarding for network virtualization and abstraction.
- 5) MUST support the control plane with both SDN controller approach, and the traditional distributed control plane approach with MP-BGP protocol.
- 6) MUST support VM mobility.

7) MUST support orchestration/auto-provisioning 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 vPE 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 networks. 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/MPLS 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 DC, not as co-resident with the application VMs.

Option 1c. vPE-F is on any network or compute devices in any types 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 DC at one of the VPE-F, and is inter-connected 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 DC fabrics/intermediate underlay devices in the DC do not participate IP VPNs, their function is the same as provider backbone routers in the IP/MPLS back bone and they do not maintain the IP VPN states, nor IP VPN aware.

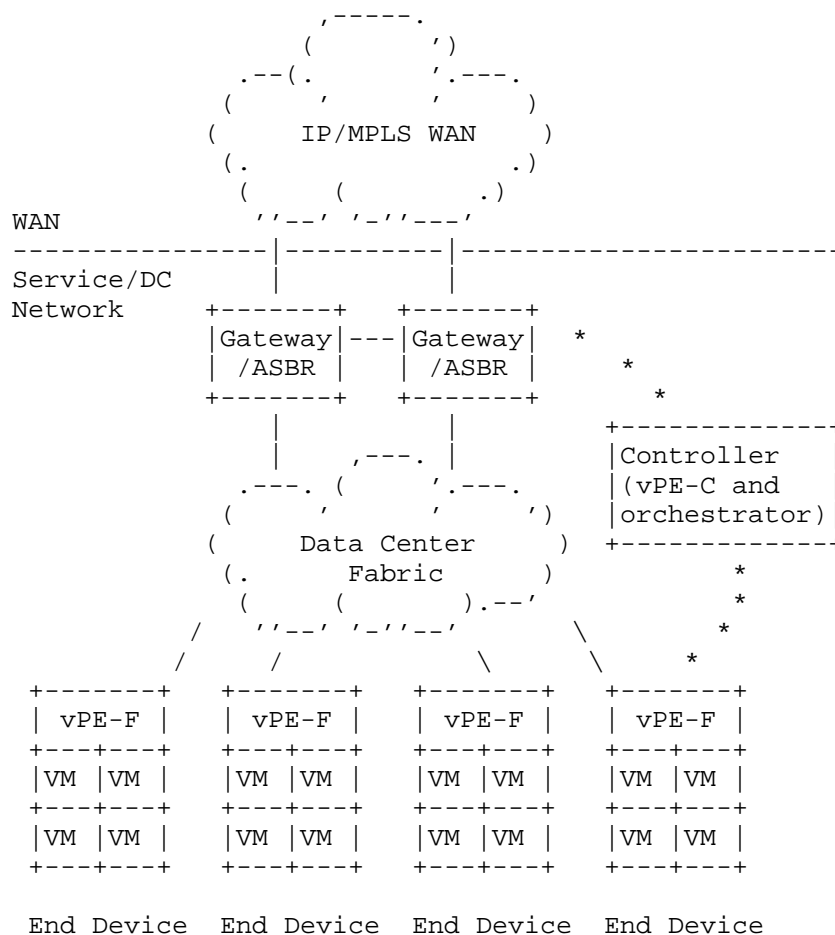


Figure 1. Virtualized Data Center with vPE at the end device and vPE-C and vPE-F physically decoupled

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.

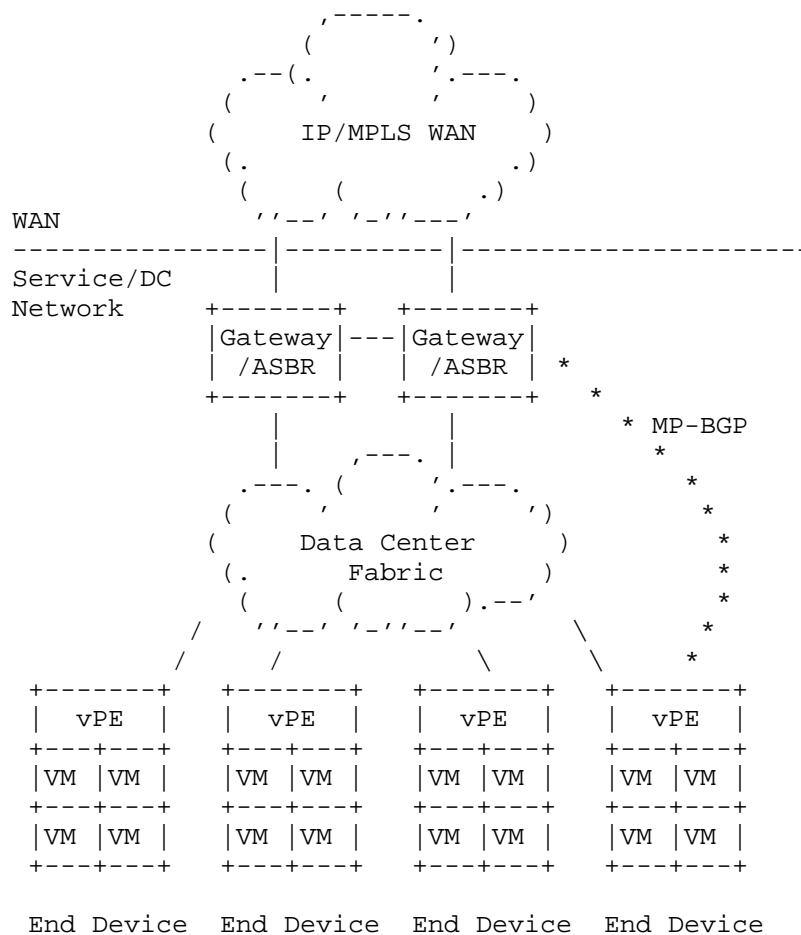


Figure 2. Virtualized Data Center with vPE at the end device, VPN control signal uses MP-BGP

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.

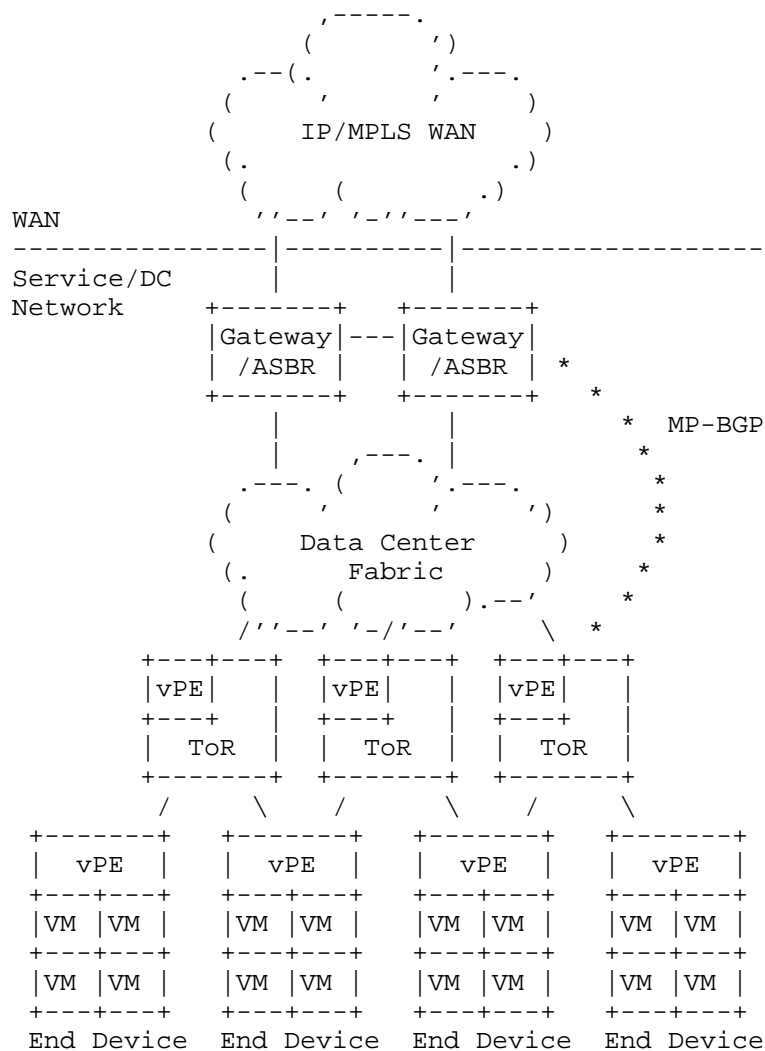


Figure 3. Virtualized Data Center with vPE at the ToR, VPN control signal uses MP-BGP

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 The server view of a 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 with application VMs. 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.

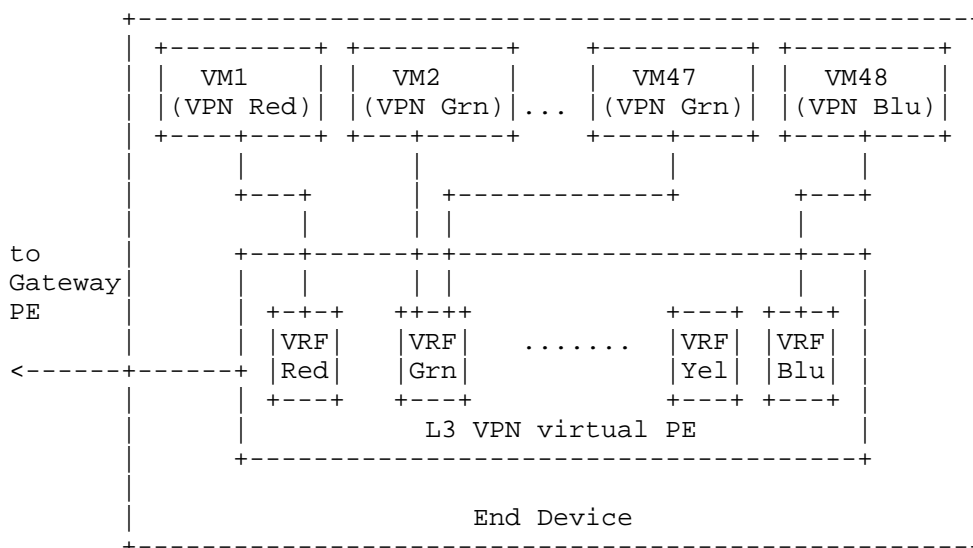


Figure 4. Server View of vPE to VM relationship

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.

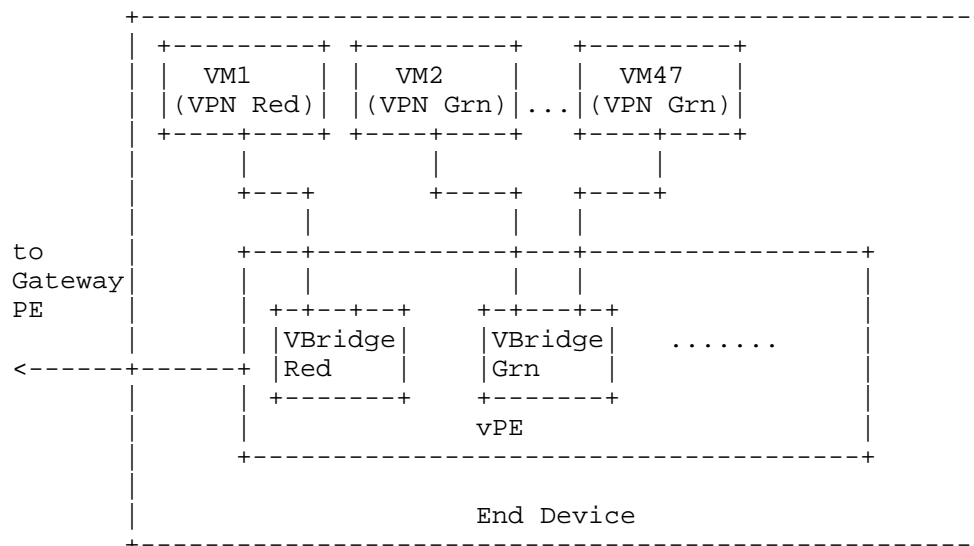


Figure 4. Bridging Service at vPE

### 3. Control Plane

#### 3.1 vPE Control Plane (vPE-C)

##### 3.1.1 The 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, e.g. in 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.ietf-i2rs-architecture], [I-D.ietf-i2rs-problem-statement].

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 DC can surpass the scale of those in a Service Provider backbone VPN networks. There may be tens of thousands of end devices in a single DC.

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

When a RR software instance is residing in a physical device, e.g., a server, which is partitioned to support multi-functions and application VMs, the RR becomes a virtualized RR (vRR). Since RR performs control functions only, a dedicated 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 in a 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 IP VPN route distribution. With RTC, only the BGP receivers (e.g, PE/vPE/RR/vRR/ASBRs, etc.) with the particular IP VPNs attached will receive the route update for the corresponding VPNs. It is critical to use constrained route distribution to support large-scale IP VPN 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 IP VPN'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 DC 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], and NVGRE [I-D.sridharan-virtualization-nvgre] are work in progress in IETF.

The most common BGP/MPLS IP VPN deployments in SP 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 DC environment is different than the SP VPN networks or large Enterprise backbones. MPLS deployment 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 a more practical approach for DC and computing environment. Note that when IP encapsulation is used, the associated security properties must be analyzed carefully.

In addition, there are new encapsulation proposals for DCs currently as work in progress within IETF, including several UDP based encapsulation proposals and some TCP based proposal. These overlay encapsulations can be suitable alternatives for the vPE solutions.

#### 4.4 Optimal forwarding

Many large cloud service providers have reported the DC traffic is now 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 newer DC 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 capable to provide direct communication among the VMs/application senders/receivers without the need of going through Gateway devices. If the senders and the receivers are on the same end device, the traffic SHOULD NOT need to leave the device. If they are on different end devices, optimal routing SHOULD be applied.

Extranet IP VPN techniques can be used for multiple VPNs access without the need of Gateway facilitation. This is done through the use of IP VPN policy control mechanisms.

In addition, ECMP is a built in IP mechanism 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, VXLAN is such an example.

#### 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 entries provisioned by the vPE controller.

When the vPE is providing Layer 3 service to the attached CEs, the VPN forwarder has 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 Layer 2 service to the attached end-hosts, the VPN forwarder has 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 performs the necessary provisioning functions 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 this normally is not an issue for the newer generation of forwarding devices and servers. Note that 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 DC, SHOULD be taken from separate address blocks outside the ones used for the underlay infrastructure of the DC. This practice is to protect the DC infrastructure from being attacked if the attacker gains access to the tenant VPNs.

Similarity, the addresses used for the DC 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 BGP/MPLS IP VPN may not be supported in every segment of the networks to provide end-to-end IP VPN connectivity. A 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 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.

The IP VPN vPE should ideally be located as close to the "customer" edge devices as possible. When 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.

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

One can consider using a combined vPE and vCE [I-D.fang-l3vpn-virtual-ce] solution to solved the problem. Use 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 IP VPN 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 DC 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.ietf-i2rs-architecture], and [I-D.ietf-i2rs-problem-statement].

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.ietf-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 DC.
- 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.

## 8. Security Considerations

As vPE is an extended BGP/MPLS IP VPN solution, security threats and defense techniques described in RFC 4111 [RFC4111] for IP VPN generally apply.

When the SDN approach is used, the protocols between the vPE agent and the vPE-C in the controller MUST be mutually authenticated. Given the potentially very large scale and the dynamic nature in the cloud/DC environment, the choice of key management mechanisms need to be further studied.

## 9. IANA Considerations

None.

## 10. References

### 10.1 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.
- [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.

## 10.2 Informative References

- [RFC4111] Fang, L., Ed., "Security Framework for Provider-Provisioned Virtual Private Networks (PPVPNs)", RFC 4111, July 2005.
- [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.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, work in progress.
- [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, work in progress.
- [I-D.rfernando-l3vpn-service-chaining] Fernando, R., Rao, D., Fang, L., Napierala, M., So, N., draft-rfernando-l3vpn-service-chaining, work in progress.
- [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, work in progress.
- [I-D.ietf-i2rs-architecture] Atlas, A., Halpern, J., Hares, S., Ward,

D., and Nadeau, T., "An Architecture for the Interface to the Routing System", draft-ietf-i2rs-architecture, work in progress.

[I-D.ietf-i2rs-problem-statement] Atlas, A., Nadeau, T., and Ward, D., "Interface to the Routing System Problem Statement", draft-ietf-i2rs-problem-statement, work in progress.

[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, work in progress.

[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, work in progress.

[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, work in progress.

[I-D.sridharan-virtualization-nvgre] SridharanNetwork, M., et al., "Virtualization using Generic Routing Encapsulation", draft-sridharan-virtualization-nvgre, work in progress.

#### Authors' Addresses

Luyuan Fang  
Cisco  
111 Wood Ave. South  
Iselin, NJ 08830  
Email: luyuanf@gmail.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

Network Working Group  
Internet-Draft  
Intended status: Informational  
Expires: April 23, 2014

Z. Li  
Y. Yin  
Huawei Technologies  
October 20, 2013

An Architecture of Instant VPN  
draft-li-l3vpn-instant-vpn-arch-00

Abstract

With the wide application of cloud computing technology, more and more enterprises will hire public cloud data center resources, reduce their own costs. Providers need to enterprise data center rental network and enterprise own network connected together, provide enterprise lease line services. L3VPN is most providers provide this service selection. New VPN line business needs to rapidly deploy, but the current technology cannot meet this requirement. This document introduces the architecture to deploy L3VPN rapidly which can satisfy the requirement for service provision.

Requirements Language

The key words "MUST", "MUST NOT", "REQUIRED", "SHALL", "SHALL NOT", "SHOULD", "SHOULD NOT", "RE

COMMENDED", "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 April 23, 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. Requirement . . . . .	4
4. Architecture . . . . .	4
4.1. VPN Controller . . . . .	5
4.1.1. VPN User Interface . . . . .	6
4.1.2. VPN Authentication . . . . .	6
4.1.3. VPN User Management . . . . .	6
4.1.4. VPN Route Management . . . . .	7
4.2. PE . . . . .	7
4.2.1. VPN User Management . . . . .	7
4.2.2. VRF Instance Automatically Create . . . . .	7
4.2.3. Access Side Configure Automatically . . . . .	7
4.2.4. VPN Tunnel Automatically Setup . . . . .	8
4.3. CE . . . . .	8
5. Procedures . . . . .	8
5.1. VPN User Information Input . . . . .	8
5.2. VPN User Authentication Process . . . . .	9
5.3. VRF Instance Configuration . . . . .	9
5.4. Route Protocol Configuration Between PE and CE . . . . .	9
5.5. VRF Route Distribution . . . . .	10
5.6. VPN Tunnel Establishment . . . . .	10
6. Protocol Extension Requirements . . . . .	11
6.1. Authentication Between CE and PE . . . . .	11
6.2. Authentication Between PE and VPN Controller . . . . .	11
6.3. Configuration Distributed from VPN Controller to PE . . . . .	11
6.4. Tunnel Information Distributed from VPN Controller to PE . . . . .	11
7. IANA Considerations . . . . .	12
8. Security Considerations . . . . .	12
9. Normative References . . . . .	12
Authors' Addresses . . . . .	12

## 1. Introduction

Enterprise leased line is one of the provider's principal business. Many providers are currently using L3VPN technology to provide this type of service, and with the development of cloud computing, the public cloud data centers will provide network service for a large number of enterprises and the provider's network need deploy massive L3VPNs in a short time. But the current L3VPN deployment has following problems to be unable to meet fast deployment requirements.

1. The deployment of an L3VPN need to configure the VPN instance and tunnel on PE nodes. If an L3VPN has a lot of access points, many PE nodes need to be configured and creation of tunnels will become complicated. For data centers to support massive tenants, the PE nodes need to configure massive VPNs. The configuration work will be huge and error-prone.

2. The L3VPN deployment need to configure RD, RT, etc. which need complex unified planning and design.

3. For enterprise lease line services , providers and enterprises need to plan ahead between CE and PE IP address and routing protocols, and enterprise CE position changes, need to reconfigure the corresponding PE.

This document defines an architecture of instant L3VPN to deploy L3VPN fast which is achieved by central control.

## 2. Terminology

VPN: Virtual Private Network

CE: Customer Edge Router

PE: Provider Edge Router

MP-BGP: Multiprotocol BGP

RT: Route Target

RD: Route Distinguisher

EAP: Extensible Authentication Protocol

RADIUS: Remote Authentication Dial In User Service

I2RS: Interface to The Internet Routing System

### 3. Requirement

This chapter describe requirements of rapid deployment of L3VPN for cloud computing scenarios from the point of view of the enterprise and the provider:

For enterprises:

Requirement 1: Enterprise users can apply for VPN services from providers through web pages.

Requirement 2: Enterprise users can have access to the VPN network at any position.

Requirement 3: When the position of the enterprise's CE can be changed, configuration after migration need not to be changed and CE can still access L3VPN network real time.

Requirement 4: Enterprises can define access strategy between any two CEs.

Requirement 5: Enterprises can define the bandwidth and other constraints between any two CEs.

Requirement 6: Enterprises can apply for virtual machine resources from provider data center, and the virtual machine can quickly access the enterprise L3VPN networks.

For providers:

Requirement 1: Providers can deploy VPN service rapidly according to business requirements of the enterprise.

Requirement 2: Providers can do for enterprise user authentication, authorization, and accounting functions.

Requirement 3: Providers can manage all VPN, including viewing each VPN's CE ID, PE address which CE access to, the tunnel state information between PEs, etc., and can provide enterprises with a VPN SLA reports.

### 4. Architecture

The architecture of Instant VPN is shown in the following figure. There is a VPN controller in the network to directly control all PE devices. This document focuses on the VPN deployment in one AS. The central control architecture for the inter-AS VPN will be described in the future version.

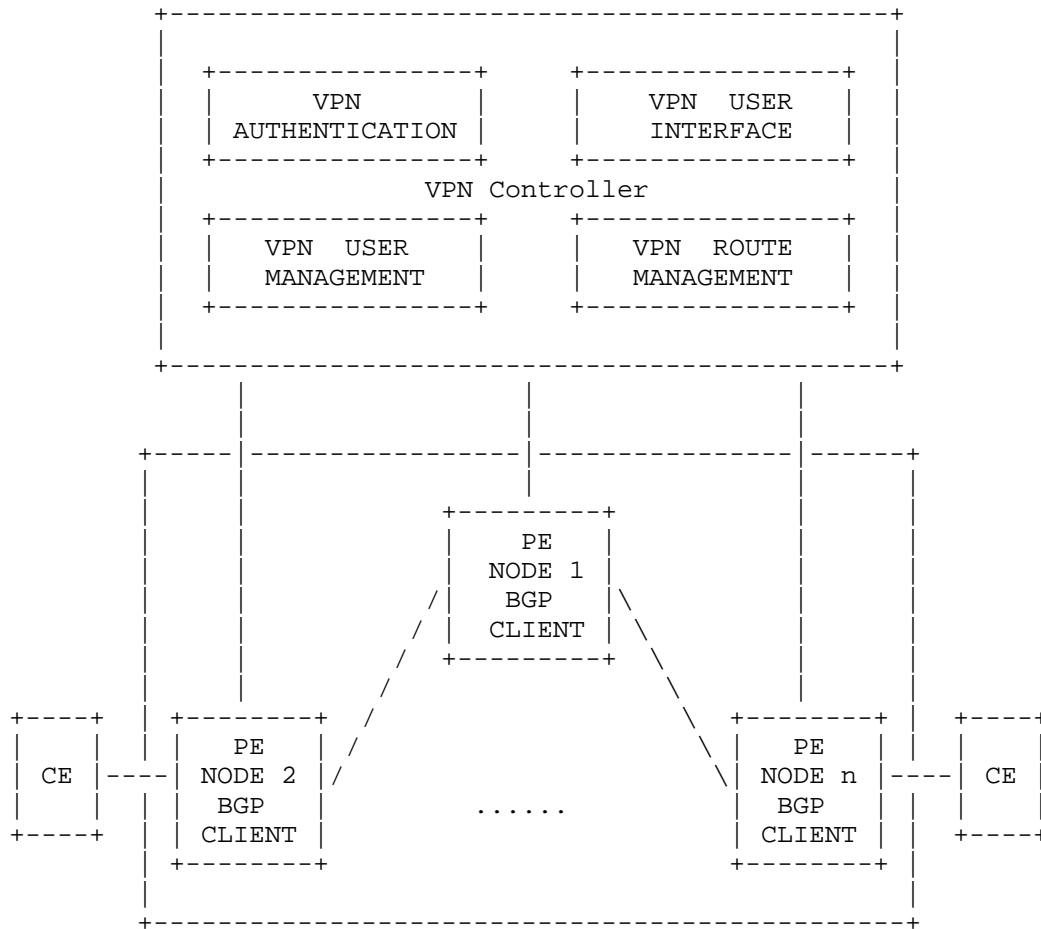


Figure 1 An Architecture of Instant VPN

#### 4.1. VPN Controller

There are four main function modules in the VPN Controller which controls all PE devices:

- VPN User Interface
- VPN Authentication
- VPN User Management
- VPN Route Management

#### 4.1.1. VPN User Interface

1. Provides the interface for the enterprise VPN user to input VPN information which will be saved to the VPN User Management module.
2. Provide VPN information query interface to view the VPN information through the interface, such as the current on-line of CE ID, PE address which CE accessed to, VPN tunnel information, VPN SLA etc.

#### 4.1.2. VPN Authentication

1. Receive the authentication request message of CE from PE.
2. Being responsible for the authentication of VPN users on-line by examining the VPN ID+CE ID and CE password.
3. Reply message with VPN information for the users that pass the authentication successfully .

#### 4.1.3. VPN User Management

1. Being responsible for the management of all providers of VPN user information, including:

VPN ID: unique ID of enterprise VPN users.

CE ID: unique ID of CE, which should be guaranteed to be unique within an enterprise VPN user.

CE password: CE access authentication password.

Access strategy between CEs: definition of strategy which defines how two different CEs in the same VPN access to each other.

Routing protocol between CE and PE: choice of routing protocol which runs between CE and PE. If BGP is chosen, the private network AS number should be provided.

IP address and mask connected between CE and PE: IP address and subnet mask of the PE's interface which accessed to CE, the IP address and subnet mask of the CE's interface which accessed to PE.

All above Information should be provided by the enterprises when they apply for VPN service.

2. Automatic generating of configuration for each access CE, including VRF name, RD, RT information.

3. Management of all on-line VPN users, including all CE IDs, PE addresses which CE accessed to, etc.

#### 4.1.4. VPN Route Management

1. Receives all VPN routes from all PEs. Calculate routes for each VRF based on each VRF's IRT.
2. Send the VRF routes to PE based on the strategy of center control.
3. Apply the route policy centrally to control the route distribution.

#### 4.2. PE

There are four main function modules in PE devices: VPN user management function, VRF instance automatically create function, Access side configuration automatically function, VPN tunnel automatically set up function.

##### 4.2.1. VPN User Management

1. Receive authentication request message from CE, including VPN ID, CE ID, CE password, and send authentication request to VPN controller.
2. Receive authentication response message from VPN controller, including authentication result, VPN ID, CE ID, the routing protocol running with CE, the CE interface IP address and mask which access to PE, the PE interface IP address and mask which access to CE, the CE AS number and PE AS number if route protocol is BGP, and send the authentication response message to CE.
3. Manage all CE users, recording VPN ID, CE ID, the CE access interface, accounting for VPN user.

##### 4.2.2. VRF Instance Automatically Create

1. Create VRF instance based on the configuration which VPN controller sent, configure the RD and RT automatically
2. Bind the CE access interface with VRF automatically.

##### 4.2.3. Access Side Configure Automatically

1. Configure the IP address and subnet mask of the interface which access to CE based on the authentication response message.

2. Configure the route protocol based on the message which VPN controller sent.

#### 4.2.4. VPN Tunnel Automatically Setup

1. For each VRF, receive the following information which VPN controller sent:

---The PE IP address list which can access to the VPN.

---The tunnel information including tunnel type, tunnel bandwidth and other constraints if MPLS TE tunnel is used.

2. Set up the tunnel automatically with each PE.

#### 4.3. CE

1. Automatically initiate VPN user authentication request to PE.
2. Receive VPN user authentication response message including the CE IP address and mask, the route protocol between CE and PE, the PE IP address and AS number if protocol is BGP.
3. Configure the interface IP address, mask and route protocol automatically.

### 5. Procedures

#### 5.1. VPN User Information Input

Step 1: Enterprise users input following information based on the VPN User interface which VPN controller provided:

-- VPN ID

-- CE ID

-- CE password

-- access strategy between CEs

-- routing protocol between CE and PE

-- IP address and mask connected between the CE and the PE.

Step 2: VPN Controller saves the VPN information to the VPN User Management.

## 5.2. VPN User Authentication Process

Step 1: CE initiates VPN service request to PE automatically , carrying the VPN ID, CE ID, CE password.

Step 2: PE receives the authentication request and sends the authentication request to VPN Controller.

Step 3: VPN Controller receives and processes the authentication request through the VPN Authentication module based on the information input by the enterprise users.

Step 4: If the authentication passes, VPN Controller sends authentication success message to PE, carrying information: VPN ID, CE ID, the routing protocol between CE and PE, the CE IP address and mask, the PE IP address and mask, the CE AS number and PE AS number if routing protocol is BGP.

Step 5: PE decapsulates authentication success message. If the routing protocol between CE and PE is BGP, PE's own AS number and IP address will be encapsulated in an authentication success message and sent to CE.

## 5.3. VRF Instance Configuration

Step 1: After the CE authentication passes, VPN Authentication module informs VPN User Management module.

Step 2: VPN User Management module creates an VRF instance for the CE user, automatically generating VRF RD, import RT and export RT information.

Step 3: VPN User Management module sends VRF configuration information to the corresponding PE.

Step 4: PE receive the VRF configuration from VPN controller, and create VRF automatically, binding VRF with the interface which CE access to.

## 5.4. Route Protocol Configuration Between PE and CE

Step 1: After the CE user authentication passes, VPN User Management gets the routing protocol information between CE and PE and automatically generates the routing protocol configuration of the PE. If the BGP protocol runs between the CE and the PE, the CE's IP address and AS number are also required. If the enterprise define the access strategy, also generates the route policy configuration.



Step 2: VPN User Management sends the configuration to PE to complete the configuration automatically.

Step 3: After the CE receives the authentication success message from PE, it can get the route protocol type between CE and PE. If the protocol is BGP, it can also get the PE's IP address and AS number from the message. Then the CE can complete the routing configuration automatically.

#### 5.5. VRF Route Distribution

Step 1: BGP peers establish automatically between the PE and the VPN Controller.

Step 2: CE advertises its routes to PE.

Step 3: PE advertises the routes which are received from CE to the VPN controller. VPN controller processes the routes through VPN Route Management module.

Step 4: VPN Route Management module receives VPN routes of all on-line CE and calculate routes for each VRF according to the VRF RT value.

Step 5: In the controller the enterprise users can configure specific policy for each CE or PE to control the route distribution, such as the removal of specific routing prefixes.

Step 6: VPN Controller advertises the VRF routes to the corresponding PE.

Step 7: PE receives and installs the VRF routes. Then the PE sends routes to CE.

#### 5.6. VPN Tunnel Establishment

If the enterprise user does not define the bandwidth and other constraints between CEs, the tunnel for the VPN can use LDP or GRE, VXLAN and other types of tunnels. If the enterprise users define the bandwidth or other constraints between CEs, MPLS TE tunnel can be used.

Based on VRF RT information of PEs, VPN controller determine the tunnels which should be set up among these PEs of the VPN.

When a new CE is online, VPN Controller will send the PE lists to a specific PE. The PE lists determines the tunnels which the specific PE need to set up to these PEs. In addition, the tunnel type can

also be advertised to the PE. If MPLS TE tunnel is used, the MPLS TE constraint for the tunnel can also be advertised. When the PE received the PE list, it should establish the tunnel with the other PE members specified by the PE list.

## 6. Protocol Extension Requirements

### 6.1. Authentication Between CE and PE

Needs to be extended as follows:

1. The authentication request message need carry the VPN ID information, CE ID, CE password information.
2. The authentication response message need carry the route protocol between CE and PE, the IP address and mask of CE, if route protocol is BGP, also need carry the IP address PE, the CE AS number and the PE AS number.

### 6.2. Authentication Between PE and VPN Controller

Needs to be extended as follows:

1. The authentication request message need carry the VPN ID information, CE ID, CE password information.
2. The authentication response message need carry the route protocol between CE and PE, the IP address and mask of CE, the IP address and mask of PE, if route protocol is BGP, also need carry the CE AS number and the PE AS number.

### 6.3. Configuration Distributed from VPN Controller to PE

Needs to be extended as follows:

1. VRF configuration, including VRF name, VRF RD, VRF RT.
2. The route protocol configuration, including protocol process number, route policy etc.

### 6.4. Tunnel Information Distributed from VPN Controller to PE

Needs to be extended as follows:

1. PE members list information which VRF accessed to.
2. The tunnel type, tunnel constraint information if MPLS TE is used.

## 7. IANA Considerations

This document makes no request of IANA.

## 8. Security Considerations

In this solution, the main need to consider the security between CE and PE communication, PE and VPN Controller, in the existing protocol already has good mechanism, this paper does not introduce in detail.

## 9. Normative References

- [RFC2119] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, March 1997.
- [RFC2865] Rigney, C., Willens, S., Rubens, A., and W. Simpson, "Remote Authentication Dial In User Service (RADIUS)", RFC 2865, June 2000.
- [RFC2866] Rigney, C., "RADIUS Accounting", RFC 2866, June 2000.
- [RFC3579] Aboba, B. and P. Calhoun, "RADIUS (Remote Authentication Dial In User Service) Support For Extensible Authentication Protocol (EAP)", RFC 3579, September 2003.
- [RFC3580] Congdon, P., Aboba, B., Smith, A., Zorn, G., and J. Roese, "IEEE 802.1X Remote Authentication Dial In User Service (RADIUS) Usage Guidelines", RFC 3580, September 2003.
- [RFC3748] Aboba, B., Blunk, L., Vollbrecht, J., Carlson, J., and H. Levkowetz, "Extensible Authentication Protocol (EAP)", RFC 3748, June 2004.
- [RFC4364] Rosen, E. and Y. Rekhter, "BGP/MPLS IP Virtual Private Networks (VPNs)", RFC 4364, February 2006.
- [RFC4761] Kompella, K. and Y. Rekhter, "Virtual Private LAN Service (VPLS) Using BGP for Auto-Discovery and Signaling", RFC 4761, January 2007.

## Authors' Addresses

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

Email: lizhenbin@huawei.com

Yuanbin Yin  
Huawei Technologies  
Huawei Bld., No.156 Beiqing Rd.  
Beijing 100095  
China

Email: yinyuanbin@huawei.com

INTERNET-DRAFT  
Intended Status: Standards track  
Expires: April 18, 2014

R. Fernando  
D. Rao  
L. Fang  
Cisco  
M. Napierala  
AT&T  
N. So  
Tata Communications  
A. Farrel  
Juniper Networks

October 18, 2013

## Virtual Topologies for Service Chaining in BGP/IP MPLS VPNs

draft-rfernando-l3vpn-service-chaining-03

### Abstract

This document presents techniques built upon BGP/IP MPLS VPN control plane mechanisms to construct virtual topologies for service chaining. These virtual service topologies interconnect network zones and constrain the flow of traffic between these zones via a sequence of service nodes so that service functions can be applied to the traffic.

This document also describes approaches enabled by both the routing control plane and by network orchestration to realize these virtual service topologies.

### 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 . . . . .	4
2. Intra-Zone Routing and Traffic Forwarding. . . . .	5
3. Inter-Zone Routing and Traffic Forwarding. . . . .	7
3.1 Traffic Forwarding Operational Flow . . . . .	8
4. Inter-Zone Model . . . . .	9
4.1 Constructing the Virtual Service Topology . . . . .	9
4.2 Per-VM Service Chains. . . . .	12
5. Routing Considerations . . . . .	12
5.1 Multiple Service Topologies . . . . .	12
5.2 Multipath . . . . .	12
5.3 Supporting Redundancy . . . . .	12
5.4 Route Aggregation . . . . .	13
6. Orchestration Driven Approach . . . . .	13
7. Security Considerations. . . . .	13
8. Management Considerations. . . . .	13
9. IANA Considerations. . . . .	13
10. Acknowledgements. . . . .	14
11. References. . . . .	14
11.1 Normative References . . . . .	14
11.2 Informative References . . . . .	14
Authors' Addresses . . . . .	15

## 1. Introduction

Network topologies and routing design in enterprise, data center, and campus networks typically reflect the needs of the organization in terms of performance, scale, security, and availability. For scale and security reasons, these networks may be composed of multiple small domains or zones each serving one or more functions of the organization.

A network zone is a logical grouping of physical assets that supports certain applications. Hosts can communicate freely within a zone. That is, a datagram traveling between two hosts in the same zone is not routed through any servers that examine the datagram payload and apply services (such as security or load balancing) to the traffic. But a datagram traveling between hosts in different zones may be subject to additional services to meet the needs of scaling, performance, and security for the applications or the networks themselves.

Networks have achieved division into zones and the imposition of services through a combination of physical topology constraints and routing. For example, one can force datagrams to go through a firewall (FW) by putting the FW in the physical data path from a source to the destination, or by causing the routed path from source to destination to go via an W that would not normally be on the path. Similarly, the datagrams may need to go through a security gateway for security services, or a Load Balancer (LB) for load balancing services.

In virtualized data centers, appliances, applications, and network functions, including IP VPN provider edge (PE) and customer edge (CE) functions are all commonly virtualized. That is, they exist as software instances residing in servers or appliances instead of individual (dedicated) physical devices.

Migrating a network with all its functions and infrastructure elements to realization in a virtualized data center requires network overlay mechanisms that provide the ability to create virtual network topologies that mimic physical networks, and that provide the ability to constrain the flow of routing and traffic over these virtual network topologies.

A data center uses a virtual topology in which the servers are in the "virtual" data path, rather than in the physical data path. For example, a traffic flow might previously have had the source PE-1 and destination at an Autonomous System Border Router (ASBR), ASBR-1, and the flow might have needed to be serviced by FW-1 and LB-1. In this case its path would have been PE-1, FW-1, LB-1, ASBR-1. However, in a

virtualized data center, the functions of all four nodes could be provided by virtual nodes that could be placed at arbitrary locations across the data center. Thus the "virtual service chain" vPE-1, vFW-1, vLB-1, vASBR-1, that is the sequence of virtual service nodes that packet must traverse, could be realized by a path between arbitrary physical locations in the data center.

A data center will likely support multiple tenants. A tenant is a customer who uses the virtualized data center services. Each tenant might require different connectedness (i.e., a different virtual topology) between their zones and applications, and might need the ability to apply different network policies such that the services for inter-zone traffic are applied in a specific order according to the organization objectives of the tenant. Furthermore, a data center might need multiple virtual topologies per tenant to handle different types of application traffic.

Additionally, a data center operator may choose to provide services for multiple tenants on the same virtualized end device, for example, a server. Such multi-tenant devices must utilize techniques such as routing isolation to retain separation between tenants' traffic.

To address all of these requirements, the mechanisms devised for use in a data center need to be flexible enough to accommodate the custom needs of the tenants and their applications, and at the same time must be robust enough to satisfy the scale, performance, and high availability needs that are demanded by the operator of the virtual network infrastructure that has a very large number of tenants each with different application types, large networks, multiple services, and high-volume traffic.

Toward this end, this document introduces the concept of virtual service topologies and extends IP MPLS VPN control plane mechanisms to constrain routing and traffic flow over virtual service topologies.

The creation of these topologies and the setting up of the forwarding tables to steer traffic over them may be carried out either by extensions to IP MPLS VPN procedures and functionality at the PEs, or via a "software defined networking" (SDN) approach. This document specifies the use of both approaches, but uses the IP MPLS VPN option to illustrate the various steps involved.



## 1.1 Terminology

This document uses the following acronyms and terms.

Terms	Meaning
AS	Autonomous System
ASBR	Autonomous System Border Router
CE	Customer Edge
FW	Firewall
I2RS	Interface to the Routing System
L3VPN	Layer Three VPN
LB	Load Balancer
NLRI	Network Layer Reachability Information [RFC4271]
P	Provider backbone router
proxy-arp	proxy-Address Resolution Protocol
RR	Route Reflector
RT	Route Target
SDN	Software Defined Network
vCE	virtual Customer Edge router [I-D.fang-l3vpn-virtual-ce]
vFW	virtual Firewall
vLB	virtual Load Balancer
VM	Virtual Machine
vPC	virtual Private Cloud
vPE	virtual Provider Edge router [I-D.fang-l3vpn-virtual-pe]
VPN	Virtual Private Network
VRF	VPN Routing and Forwarding table [RFC4364]
vRR	virtual Route Reflector

This document also uses the following general terms:

### Service-PE:

A BGP/IP MPLS VPN PE to which a service node in a virtual service topology is attached. The PE directs incoming traffic from other PEs or from attached hosts to the service node via an MPLS VPN label or IP lookup. The PE also forwards traffic from the service node to the next node in the chain. A Service-PE is a logical entity and a given PE may be attached to both a service node and an application host VM.

### Service node:

A physical or virtual service appliance/application which inspects and/or redirects the flow of inter-zone traffic. Examples of service nodes include FWs, LBs, and deep packet inspectors. The service node acts as a CE in the VPN network.

Service chain: A sequence of service nodes that interconnect the zones containing the source and destination hosts. The service chain is unidirectional and creates a one way traffic flow between source zone and destination zone.

Virtual service topology:

A virtual service topology consists of a sequence of service-PEs and their attached service nodes created in a specific order. A service topology is constructed via one or more routes that direct the traffic flow among the PEs that form the service chain.

Service-topology-RT:

A BGP route attribute that identifies the specific service topology.

Tenant:

A tenant is a higher-level management construct. In the control/forwarding plane it is the collection of various virtual networks that get instantiated. A tenant may have more than one virtual network or VPN.

Zone:

A logical grouping of physical assets that supports certain applications or a subset thereof. VMs or hosts can communicate freely within a zone.

## 2. Intra-Zone Routing and Traffic Forwarding

This section provides a brief overview of how the BGP/IP MPLS VPN [RFC4364] control plane can be used in a DC network to used to divide the network into a number of zones. The subsequent sections in the document build on this base model to create inter-zone service topologies by interconnecting these zones and forcing inter-zone traffic to travel through a sequence of servers where the sequence of servers depends on the tuple <source zone, destination zone, application>.

The notion of a BGP/IP VPN when applied to the virtual data center works in the following manner.

The VM that runs the applications in the server is treated as a CE attached to the VPN. A CE/VM belongs to a zone. The PE is the first hop router from the CE/VM and the PE-CE link is single hop from a layer-3 perspective. Any of the available physical, logical or tunneling technologies can be used to create this "direct" link between the CE/VM and its attached PE(s).

If a PE attaches to one or more CEs of a certain zone, the PE must

have exactly one VRF for that zone, and the PE-CE links to those CEs must all be associated with that VRF. Intra-zone connectivity between CE/VMs that attach to different PEs is achieved by designating an RT per zone (zone-RT) that is both an import RT and an export RT of all

PE VRFs that terminate the CE/VMs that belong to the zone. A VM may have multiple virtual interfaces that attach to different zones.

It is further assumed that the CE/VMs are associated with network policies that are activated on an attached PE when a CE/VM is instantiated. These policies dictate how the network is set up for the CE/VM including the properties of the CE-PE link, the IP address of the CE/VM, the zones to which it belongs, QoS policies, etc. There are many ways to accomplish this step, but a description of such mechanisms is outside the scope of this document.

When the CE/VM is activated, the attached PE starts to export the CEs IP address with the corresponding zone-RT. This allows unrestricted any-to-any communication between the newly active VM and the rest of the VMs in the zone.

The classification of VMs into a zone is driven by the communication and security policy and is independent of the addressing scheme for the VMs. The VMs in a zone may be in the same or different IP subnets with user-defined mask-lengths. The PE advertises /32 routes to advertise reachability to locally attached VMs. If two VMs are in the same IP subnet, the PE may employ proxy-ARP to assist the VM to resolve ARP for other VMs in the IP subnet, and may use IP forwarding to carry traffic between the VMs. When a VM is attached to a remote PE, IP VPN forwarding is used to tunnel packets to the remote PE.

### 3. Inter-Zone Routing and Traffic Forwarding

A simple form of inter-zone traffic forwarding can be achieved using extranets or hub-and-spoke L3VPN configurations [RFC7024]. However, the ability to enforce constrained traffic flow through a set of services is non-existent in extranets and is limited in hub-and-spoke setups.

Note that the inter-zone services cannot always be assumed to reside and be in-lined on a PE. There is a need to virtualize the services themselves so that they can be implemented on commodity hardware and scaled out 'elastically' when traffic demands increase. This creates a situation where services for traffic between zones may be applied not only at the source-zone PE or the destination-zone PE. Mechanisms are required that make it easy to direct inter-zone traffic through the appropriate set of service nodes that might be remote or virtualized.

### 3.1 Traffic Forwarding Operational Flow

Traffic from a source endpoint (a VM/CE) in a source zone reaches an ingress zone-PE and is associated with a VRF in that zone as described above. The zone-PE will forward the traffic and direct it toward the first service-node. If the service-node is attached to the zone-PE, the zone-PE will forward the packets out of one of its access interfaces. If the service-node is attached to a different service-PE, the zone-PE will encapsulate the packets and send them toward the service-PE. The zone-PE and service PE may be connected via an intermediate network of devices and the encapsulation causes the packets to be tunneled across this intermediate network.

The service-PE will receive these encapsulated packets from the source zone-PE, decapsulate them, and forward them to its attached service-node. The traffic that comes back to the service-PE from the service-node must now be forwarded to the next service-node in the chain. As above, the next service-node may be locally attached or at a remote service-PE.

At the last service-PE in the chain, the traffic that comes back from a service-node must be forwarded to the destination in the target zone. Just as with the service-nodes, the destination may be attached to the service-PE or reachable via another PE.

As can be seen from this description, a given packet flow needs to be forwarded differently at each PE depending on whether it is arriving from a node attached to the PE or from a remote PE, and depending on whether the traffic is to be routed toward a node attached to the PE or attached to a remote PE. The next-hop for a flow changes depending on the relative position within the service chain.

Figure 1 illustrates a virtual service topology, where hosts in Zone 1 are interconnected with hosts in Zone 2 via two service nodes (Serv-A and Serv-B) attached to two service-PEs (S-PE-A and S-PE-B respectively).

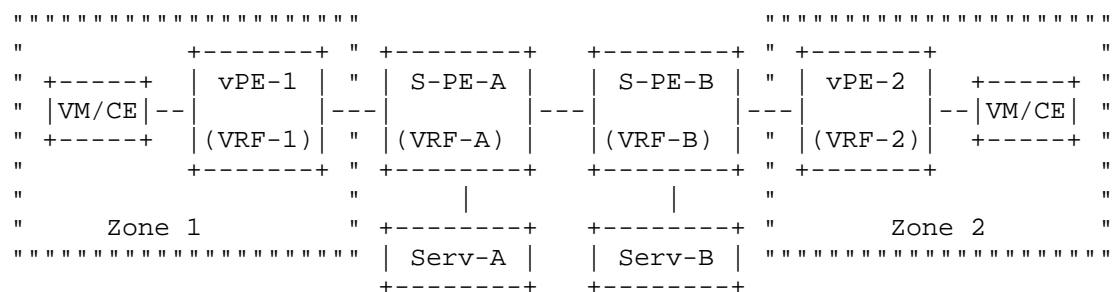


Figure 1. Virtual Service Topology Illustration

The different forwarding paths can be achieved at any PE as follows.

- o Each service node is associated with two VRFs at the service PE to which it is attached: an in-VRF for traffic toward the service node, and an out-VRF for traffic from the service node.
- o Traffic for the in-VRF arrives from the previous node in the service chain, and traffic for the out-VRF is destined toward the next node in the service chain, or toward the destination zone.
- o The in-VRF has one or more routes with a next-hop of a local access interface where the service node is attached. The out-VRF has routes with a next-hop of the next service node, which may be situated locally on the service-PE or at a remote PE.

The installation of the forwarding entries to implement the flow described above may be achieved either via IP VPN mechanisms described in Sections 4 and 5, or using an SDN approach, as described in Section 6.

#### 4. Inter-Zone Model

The inter-zone model has the following steps.

##### 4.1 Constructing the Virtual Service Topology

The virtual service topology described in the previous section is constructed via one or more routes that direct the traffic flow among the PEs forming the service chain. There should be a route per service node. The service topologies, and hence the service routes, are constructed on a per-VPN basis. This service topology is independent of the routes for the actual destination for a flow, i.e., the addresses of the VMs present in the various zones. There can be multiple service topologies for a given VPN.

###### 4.1.1 Reachability to the Service Nodes

Each service node is identified by an IP address that is scoped within the VPN. The service node is also associated with an in-VRF and out-VRF at the attached service node.

Reachability to the various service nodes in the service chain occurs via regular BGP/IP VPN route advertisements.

A service-PE will export a route for each service node attached to it. Each route will contain the Route-Target configured for the VPN, and a forwarding label that is associated with the logical in-VRF for a service node on the service-PE. This label enables the service-PE

to directly forward incoming traffic from the other PEs to the service node.

The routes to reach the various service nodes are imported into and installed in each out-VRF at a service-PE, as well as in the zone VRF on the ingress zone-PE.

#### 4.1.2 Provisioning the Service Chain

At each PE supporting a given VPN, the sequence of service nodes in a service chain can be specified in a VPN service route policy.

To create the service chain and give it a unique identity, each PE may be provisioned with the following tuple for every service chain that it belongs to:

{Service-topology-RT, Service-node-Sequence}

where Service-node-Sequence is simply an ordered list of the service node IP addresses that are in the chain.

Every service chain has a single unique service-topology-RT that is provisioned in all participating PEs.

A PE will also be provisioned with the tables and/or configuration that support the various zones and the in- and out- VRFs for the services.

#### 4.1.3 Zone Prefix Next-Hop Resolution

Routes representing hosts or VMs from a zone are called zone prefixes. A zone prefix will have its regular zone RTs attached when it is originated. This will be used by PEs in the same zone to import these prefixes to enable direct communication between VMs in the same zone.

In addition to the intra-zone RTs, zone prefixes are also tagged at the point of origination with the set of service-topology-RTs to which they belong.

Since they are tagged with the zone-RT, zone prefixes get imported into the VRFs of the service-PEs that form the service chain associated to that topology RT. Note that the zone-RT was added to the relevant VRF's import RT list during the virtual topology construction phase. These routes may be installed in the in-VRF and out-VRF at the service-PEs as well as in the ingress zone VRF.

Note that this approach introduces a change in the behavior of the

service-PEs compared to normal BGP VPN behavior, but does not require protocol changes to BGP. This modification to PE behavior allows the automatic and constrained flow of traffic via the service chain.

The PE, based on the presence of the configured Service-topology-RT in the received zone routes, will perform the following actions:

1. It will ignore the next-hop and VPN label that were advertised in the NLRI.
2. Instead, it will select the appropriate Service next-hop from the Service-node sequence associated with the Service-topology-RT.
3. It will further resolve this Service next-hop IP address locally in the associated VRF, instead of in the global routing table. It will use the next-hop and label associated with this IP address to encapsulate traffic toward the next service node.
4. If the importing service-PE is the last service-PE, it uses the next hop that came with the zone prefix for route resolution. It also uses the VPN label that came with the prefix.

In this way the zone prefixes in the intermediate service-PE hops recurse over the service chain forcing the traffic destined to them to flow through the virtual service topology.

Traffic for the zone prefix goes through the service hops created by the service topology. At each service hop, the service-PE directs the traffic to the service node. Once the service node is done processing the traffic, it sends it back to the service-PE which forwards the traffic to the next service-PE, and so on.

A significant benefit of this next-hop indirection is to avoid redundant advertisement of zone prefixes from the end-zone or service-PEs. Also, when the virtual service topology is changed (due to addition or removal of service-PEs), there should be no change to the zone prefix's import/export RT configuration.

There should be one service topology RT per virtual service topology. There can be multiple virtual service topologies and hence service topology RTs for a given VPN.

Virtual service topologies are constructed unidirectionally. Traffic in opposite directions between the same pair of zones will be supported by two different service topologies and hence two service topology routes. These two service topologies might or might not be symmetrical, i.e.m they might or might not traverse the same sequence of service-PEs/service-nodes in opposite directions.

As noted above, a service node route can be advertised with a label that directs incoming traffic to the attached service node. Alternatively, an aggregate label may be used for the service route and an IP route lookup done at the service-PE to send traffic to the service node.

Note that a new service node could be inserted into the service chain seamlessly by just configuring the service policy appropriately.

#### 4.2 Per-VM Service Chains

While the service-topology-RT allows an efficient inheritance of the service chain for all VMs in a zone, there may be a need to create a distinct service chain for an individual VM. This may be done by provisioning a separate service-topology RT and service node sequence. The VM route carries the service-topology RT, and the destination service-zone is provisioned with this RT as its Service-Import RT.

### 5. Routing Considerations

#### 5.1 Multiple Service Topologies

A service-PE can support multiple distinct service topologies for a VPN.

#### 5.2 Multipath

One could use all tools available in BGP to constrain the propagation and resolution of state created by the service topology [RFC4684].

Additional service nodes can be introduced to scale out a particular service. Each such service would be represented by a virtual IP address, and multiple service nodes associated with it. Multiple service-PEs may advertise a route to this address based on the presence of an attached service node instance, thereby creating multiple equal cost paths. This technique could be used to elastically scale out the service nodes with traffic demand.

#### 5.3 Supporting Redundancy

For stateful services an active-standby mechanism could be used at the service level. In this case, the inter-zone traffic should prefer the active service node over the standby service node.

At a routing level, this is achieved by setting up two paths for the same service node route: one path goes through the active service



node and the other through the standby service node. The active service path can then be made to win over the standby service path by appropriately setting the BGP path attributes of the service topology route such that the active path succeeds in path selection. This forces all inter-zone traffic through the active service node.

#### 5.4 Route Aggregation

Instead of the actual zone prefixes being imported and used at various points along the chain, the zone prefixes may be aggregated at the destination service-PE and the aggregate zone prefix used in the service chain between zones. In such a case, it is the aggregate zone prefix that carries the service-topology-RT and gets imported in the service-PEs that comprise the service chain.

#### 6. Orchestration Driven Approach

In an orchestration driven approach, there is no need for the zone or service PEs to determine the appropriate next-hops based on the specified service node sequence. All the necessary policy computations are carried out, and the forwarding tables for the various VRFs at the PEs determined, by the central orchestrator.

The orchestrator communicates with the various PEs (typically virtual PEs on the end-servers) to populate the forwarding tables.

The protocol used to communicate between the controller/orchestration and the PE/vPE must be a standard, programmatic interface. The programmatic interface are currently under definition in the IETF's Interface to Routing Systems (I2RS) initiative, [I-D.ietf-i2rs-architecture], [I-D.ietf-i2rs-problem-statement].

#### 7. Security Considerations

To be added.

#### 8. Management Considerations

To be added.

#### 9. IANA Considerations

This proposal does not have any IANA implications.

## 10. Acknowledgements

The authors would like to thank the following individuals for their review and feedback on the proposal: Eric Rosen, Jim Guichard, Paul Quinn, Peter, Bosch, David Ward, Ashok Ganesan. The option of configuring an ordered sequence of service nodes via policy is derived from a suggestion from Eric Rosen.

## 11. References

### 11.1 Normative References

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

### 11.2 Informative References

- [RFC4271] Rekhter, Y., Li, T., and S. Hares, "A Border Gateway Protocol 4 (BGP-4)", RFC 4271, January 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.
- [RFC7024] Jeng, H., Uttaro, J., Jalil, L., Decraene, B., Rekhter, Y., and R. Aggarwal, "Virtual Hub-and-Spoke in BGP/MPLS VPNs", RFC 7024, October 2013.
- [I-D.fang-l3vpn-virtual-ce]  
L. Fang, et al., "BGP/MPLS IP VPN Virtual CE",  
draft-fang-l3vpn-virtual-ce, work in progress.
- [I-D.fang-l3vpn-virtual-pe]  
L. Fang, et al., "BGP/MPLS IP VPN Virtual PE",  
draft-fang-l3vpn-virtual-pe, work in progress.
- [I-D.ietf-i2rs-architecture]  
Atlas, A., Halpern, J., Hares, S., Ward, D., and T Nadeau,  
"An Architecture for the Interface to the Routing System",  
draft-ietf-i2rs-architecture, work in progress.
- [I-D.ietf-i2rs-problem-statement]  
Atlas, A., Nadeau, T., and D. Ward, "Interface to the  
Routing System Problem Statement",  
draft-ietf-i2rs-problem-statement, work in progress.

## Authors' Addresses

Dhananjaya Rao  
Cisco  
170 W Tasman Dr  
San Jose, CA  
Email: [dhrao@cisco.com](mailto:dhrao@cisco.com)

Rex Fernando  
Cisco  
170 W Tasman Dr  
San Jose, CA  
Email: [rex@cisco.com](mailto:rex@cisco.com)

Luyuan Fang  
Cisco  
170 W Tasman Dr  
San Jose, CA  
Email: [luyuanf@gmail.com](mailto:luyuanf@gmail.com)

Maria Napierala  
AT&T  
200 Laurel Avenue  
Middletown, NJ 07748  
Email: [mnapierala@att.com](mailto:mnapierala@att.com)

Ning So  
Tata Communications  
Plano, TX 75082, USA  
Email: [ning.so@tatacommunications.com](mailto:ning.so@tatacommunications.com)

Adrian Farrel  
Juniper Networks  
Email: [adrian@olddog.co.uk](mailto:adrian@olddog.co.uk)



L3VPN Working Group  
Internet Draft  
Intended Status: Standards Track  
Updates: 6513,6514  
Expires: April 18, 2014

Eric C. Rosen  
Karthik Subramanian  
Cisco Systems, Inc.  
  
Jeffrey Zhang  
Juniper Networks, Inc.

October 18, 2013

## Ingress Replication Tunnels in Multicast VPN

draft-rosen-l3vpn-ir-00.txt

### Abstract

RFCs 6513, 6514, and other RFCs describe procedures by which a Service Provider may offer Multicast VPN service to its customers. These procedures create point-to-multipoint (P2MP) or multipoint-to-multipoint trees across the Service Provider's backbone. One type of P2MP tree that may be used is known as an "Ingress Replication (IR) tunnel". In an IR tunnel, a parent node need not be "directly connected" to its child nodes. When a parent node has to send a multicast data packet to its child nodes, it does not use layer 2 multicast, IP multicast, or MPLS multicast to do so. Rather, it makes n individual copies, and then unicasts each copy, through an IP or MPLS unicast tunnel, to exactly one child node. While the prior MVPN specifications allow the use of IR tunnels, those specifications are not always very clear or explicit about how the MVPN protocol elements and procedures are applied to IR tunnels. This document updates RFCs 6513 and 6514 by adding additional details that are specific to the use of IR tunnels.

### 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/ietf/lid-abstracts.txt>.

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
2	What is an IR P-tunnel? .....	6
3	How are IR P-tunnels Identified? .....	8
4	How to Join an IR P-tunnel .....	10
4.1	Advertised P-tunnels .....	10
4.1.1	If the 'Leaf Info Required Bit' is Set .....	10
4.1.2	If the 'Leaf Info Required Bit' is Not Set .....	11
4.2	Unadvertised P-tunnels .....	12
5	The PTA's 'Tunnel Identifier' Field .....	12
6	The PTA's 'MPLS Label' Field .....	13
6.1	Leaf A-D Route Originated by an Egress PE .....	13
6.2	Leaf A-D Route Originated by an Intermediate Node .....	15
6.2.1	Upstream and Downstream Segments are IR Segments .....	15
6.2.2	Only One Segment is IR .....	16
6.3	Intra-AS I-PMSI A-D Route .....	16
7	How A Child Node Prunes Itself from an IR P-tunnel .....	17
8	Parent Node Actions Upon Receiving Leaf A-D Route .....	17
9	Use of Timers when Switching UMH .....	18
10	IANA Considerations .....	19
11	Acknowledgments .....	19
12	Security Considerations .....	19
13	Authors' Addresses .....	19
14	Normative References .....	20
15	Informational References .....	20

## 1. Introduction

RFCs 6513, 6514, and others describe procedures by which a Service Provider (SP) may offer Multicast VPN (MVPN) service to its customers. These procedures create point-to-multipoint (P2MP) or multipoint-to-multipoint (MP2MP) tunnels, called "P-tunnels" (Provider-tunnels), across the SP's backbone network. Customer multicast traffic is carried through the P-tunnels.

A number of different P-tunnel technologies are supported. One of the supported P-tunnel technologies is known as "ingress replication" or "unicast replication". We will use the acronym "IR" to refer to this P-tunnel technology.

An IR P-tunnel is a P2MP tree, but a given node on the tree is not necessarily "directly attached" to its parent node or to its child nodes. To send a multicast data packet from a parent node to one of its child nodes, the parent node encapsulates the packet and then unicasts it (through a P2P or MP2P MPLS LSP or a unicast IP tunnel) to the child node. If a node on an IR tree has *n* child nodes, and has a multicast data packet that must be sent along the tree, the parent node makes *n* individual copies of the data packet, and then sends each copy, through a unicast tunnel, to exactly one child node. No lower layer multicast technology is used when sending traffic from a parent node to a child node; multiple copies of the packet may therefore be sent out a single interface.

With the single exception of IR, the P-tunnel technologies supported by the MVPN specifications are pre-existing IP multicast or MPLS multicast technologies. Each such technology has its own set of specifications, its own setup and maintenance protocols, its own syntax for identifying specific multicast trees, and its own procedures for enabling a router to be added to or removed from a particular multicast tree. For IR P-tunnels, on the other hand, there is no prior specification for setting up and maintaining the P2MP trees; the procedures and protocol elements used for setting up and maintaining the P2MP trees are specified in the MVPN specifications themselves, and all the signaling/setup is done by using the BGP A-D (Auto-Discovery) routes that are defined in [MVPN-BGP]. (The unicast tunnels used to transmit multicast data from one node to another in an IR P-tunnel may of course have their own setup and maintenance protocols, e.g., [LDP], [RSVP-TE].)

Since the transmission of a multicast data packet along an IR P-tunnel is done by transmitting the packet through a unicast tunnel, previous RFCs sometimes speak of an IR P-tunnel as "consisting of" a set of unicast tunnels. However, that way of speaking is not quite accurate. For one thing, it obscures the fact that an IR P-tunnel is



really a P2MP tree, whose nodes must maintain multicast state in both the control and data planes. For another, it obscures the fact the unicast tunnels used by a particular IR P-tunnel need not be specific to that P-tunnel; a single unicast tunnel can carry the multicast traffic of many different IR P-tunnels (and can also carry unicast traffic as well).

In this document, we provide a clearer and more explicit conceptual model for IR P-tunnels, clarifying the relationship between an IR P-tunnel and the unicast tunnels that are used for data transmission along the IR P-tunnel.

RFC 6514 defines a protocol element called a "tunnel identifier", which for most P-tunnel technologies is used to identify a P-tunnel (i.e., to identify a P2MP or MP2MP tree). However, when IR P-tunnels are used, this protocol element does not identify an IR P-tunnel. In some cases it identifies one of the P-tunnel's constituent unicast tunnels, and in other cases it is not used to identify a tunnel at all. In this document, we provide an explicit specification for how IR P-tunnels are actually identified.

Some of the MVPN specifications use phrases like "join the identified P-tunnel", even though there has up to now not been an explicit specification of how to identify an IR P-tunnel, of how a route joins such a P-tunnel, or of how a router prunes itself from such a P-tunnel. In this document, we make these procedures more explicit.

RFC 6514 does provide a method for binding an MPLS label to a P-tunnel, but does not discuss the label allocation policies that are needed for correct operation when the P-tunnel is an IR P-tunnel. Those policies are discussed in this document.

This document does not provide any new protocol elements or procedures; rather it makes explicit just how a router is to use the protocol elements and procedures of [MVPN] and [MVPN-BGP] to identify an IR P-tunnel, to join an IR P-tunnel, and to prune itself from an IR P-tunnel. This document also discusses the MPLS label allocation policies that need to be supported when binding MPLS labels to IR P-tunnels, and the timer policies that need to be supported when switching a customer multicast flow from one P-tunnel to another. As the material in this document must be understood in order to properly implement IR P-tunnels, this document is considered to update [MVPN] and [MVPN-BGP]. This document also discusses the application of "seamless multicast" [SMLS-MC] and "extranet" [MVPN-XNET] procedures to IR P-tunnels.

This draft does not discuss the use of IR P-tunnels to support a VPN customer's use of BIDIR-PIM. [C-BIDIR-IR] explains how to adapt the

procedures of [MVPN], [MVPN-BGP], and [MVPN-BIDIR] so that a customer's use of BIDIR-PIM can be supported by IR P-tunnels.

The key words "MUST", "MUST NOT", "REQUIRED", "SHALL", "SHALL NOT", "SHOULD", "SHOULD NOT", "RECOMMENDED", "NOT RECOMMENDED", "MAY", and "OPTIONAL", when and only when appearing in all capital letters, are to be interpreted as described in [RFC2119].

## 2. What is an IR P-tunnel?

An IR P-tunnel is a P2MP tree. Its nodes are BGP speakers that support the MVPN procedures of [MVPN-BGP] and related RFCs. In general, the nodes of an IR P-tunnel are either PE routers, ASBRs, or (if [SMLS-MC] is supported) ABRs. (MVPN procedures are sometimes used to support non-MVPN, or "global table" multicast; one way of doing this is defined in [SMLS-MC]. In such a case, IR P-tunnels can be used outside the context of MVPN.)

MVPN P-tunnels may be either "segmented" or "non-segmented" (as these terms are defined in [MVPN] and [MVPN-BGP]).

A "non-segmented" IR P-tunnel is a two-level P2MP tree, consisting only of a root node and a set of nodes that are children of the root node. When used in an MVPN context, the root is an ingress PE, and the child nodes of the root are the egress PEs.

In a segmented P-tunnel, IR may be used for some or all of the segments. If a particular segment of a segmented P-tunnel uses IR, then the root of that segment may have child nodes that are ABRs or ASBRs, rather than egress PEs.

As with any type of P2MP tree, each node of an IR P-tunnel holds "multicast state" for the P-tunnel. That is, each node knows the identity of its parent node on the tree, and each node knows the identities of its child nodes on the tree. In the MVPN specs, the "parent" node is also known as the "Upstream Multicast Hop" or "UMH".

What distinguishes an IR P-tunnel from any other kind of P2MP tree is the method by which a data packet is transmitted from a parent node to a child node. To transmit a multicast data packet from a parent node to a child node along a particular IR P-tunnel, the parent node does the following:

- It labels the packet with a label (call it a "P-tunnel label") that the child node has assigned to that P-tunnel,

- It then places the packet in a unicast encapsulation and unicasts the packet to the child node. That is, the parent node sends the packet through a "unicast tunnel" to a particular child node. This unicast tunnel need not be specially created to be part of the IR P-tunnel; it can be any P2P or MP2P unicast tunnel that will get the packets from the parent node to the child node. A single such unicast tunnel may be carrying multicast data packets of several different P2MP trees, and may also be carrying unicast data packets.

The parent node repeats this process for each child node, creating one copy for each child node, and sending each copy through a unicast tunnel to corresponding child node. It does not use layer 2 multicast, IP multicast, or MPLS multicast to transmit packets to its child nodes. As a result, multiple copies of each packet may be sent out a single interface; this may happen, e.g., if that interface is the next hop interface, according to unicast routing, from the parent node to several of the child nodes.

Since data traveling along an IR P-tunnel is always unicast from parent node to child node, it can be convenient to think of an IR P-tunnel as a P2MP tree whose arcs are unicast tunnels. However, it is important to understand that the unicast tunnels need not be specific to any particular IR P-tunnel. If R1 is the parent node of R2 on two different IR P-tunnels, a single unicast tunnel from R1 to R2 may be used to carry data along both IR P-tunnels. All that is required is that when the data packets arrive at R2, R2 will see the "P-tunnel label" at the top of the packets' label stack; R2's further processing of the packets will depend upon that label. Note that the same unicast tunnel between R1 and R2 may also be carrying unicast data packets.

Typically the unicast tunnels are the Label Switched Paths (LSPs) that already exist to carry unicast traffic; either MP2P LSPs created by LDP [LDP] or P2P LSPs created by RSVP-TE [RSVP-TE]. However, any other kind of unicast tunnel may be used. A unicast tunnel may have an arbitrary number of intermediate routers; those routers do not maintain any multicast state for the IR P-tunnel, and in general are not even aware of its existence.

As with all other P-tunnel types, IR P-tunnels may be used as Inclusive P-tunnels or as Selective P-tunnels.

### 3. How are IR P-tunnels Identified?

There are four MVPN BGP route types in which P-tunnels can be identified: Intra-AS I-PMSI A-D routes, Inter-AS I-PMSI A-D routes, S-PMSI A-D routes, and Leaf A-D routes. (These route types are all defined in [MVPN-BGP]).

Whenever it is necessary to identify a P-tunnel in a route of one of these types, a "PMSI Tunnel Attribute" (PTA) is added to the route. As defined in [MVPN-BGP] section 5, the PTA contains four fields: "Tunnel Type", "MPLS Label", "Tunnel Identifier", and "Flags". [MVPN-BGP] defines only one bit in the "Flags" field, the "Leaf Information Required" bit.

If a route identifies an IR P-tunnel, the "Tunnel Type" field of its PTA is set to the value 6, meaning "Ingress Replication".

Most types of P-tunnel are associated with specific protocols that are used to set up and maintain tunnels of that type. For example, if the "Tunnel Type" field is set to 2, meaning "mLDP P2MP LSP", the associated setup protocol is mLDP [mLDP]. The associated setup protocol always has a method of identifying the tunnels that it sets up. For example, mLDP uses a "FEC element" to identify a tree. If the "Tunnel type" field is set to 3, meaning "PIM SSM Tree", the associated setup protocol is PIM, and "(S,G)" is used to identify the tree. In these cases, the "Tunnel Identifier" field of the PTA carries a tree identifier as defined by the setup protocol used for the particular tunnel type.

IR P-tunnels, on the other hand, are entirely setup and maintained by the use of BGP A-D routes, and are not associated with any other setup protocol. (The unicast tunnels used to transmit multicast data along an IR P-tunnel may have their own setup and maintenance protocols, of course.) Further, the identifier of an IR P-tunnel does not appear in the PTA at all. Rather, the P-tunnel identifier is in the "Network Layer Reachability Information" (NLRI) field of the A-D routes that are used to advertise and to setup the P-tunnel.

When an IR P-tunnel is identified in an S-PMSI A-D route, an Intra-AS I-PMSI A-D route, or an Inter-AS I-PMSI A-D route (we will refer to these three route types as "advertising A-D routes"), its identifier is hereby defined to be the NLRI of that route. See sections 4.1, 4.2, and 4.3 of [MVPN-BGP] for the specification of these NLRIs. Note that the P-tunnel identifier includes the "route type" and "length" octets of the NLRI.

An advertising A-D route is considered to identify an IR P-tunnel only if it carries a PTA whose "Tunnel Type" field is set to "IR".

When an IR P-tunnel is identified in an S-PMSI A-D route or in an Inter-AS I-PMSI A-D route, the "Leaf Info Required" bit of the Flags field of the PTA MUST be set.

In an advertising A-D route:

- If the "Leaf Info Required" bit of the Flags field of the PTA is set, then the "Tunnel Identifier" field of the PTA has no significance whatsoever, and MUST be ignored upon reception.

Note that, per RFC6514, the length of the "Tunnel Identifier" field is variable, and is inferred from the length of the PTA. Even when this field is of no significance, its length MUST be the length of an IP address in the address space of the SP's backbone, as specified in section 4.2 of [P-ADDR]. In this case, it is RECOMMENDED that it be set to a routable address of the router that constructed the PTA. (While it might make more sense to allow or even require the field to be omitted entirely, that might raise issues of backwards compatibility with implementations that were designed prior to the publication of this document.)

- If the "Leaf Info Required" bit is not set, the "Tunnel Identifier" field of the PTA does have significance, but it does not identify the IR P-tunnel. The use of the PTA's "Tunnel Identifier" field in this case is discussed in section 5 of this document.

Note that according to the above definition, there is no way for two different advertising A-D routes (i.e., two advertising A-D routes with different NLRIs) to advertise the same IR P-tunnel. In the terminology of [MVPN], an IR P-tunnel can instantiate only a single PMSI. If an ingress PE, for example, wants to bind two customer multicast flows to a single IR P-tunnel, it must advertise that tunnel in an I-PMSI A-D route or in an S-PMSI A-D route whose NLRI contains wildcards [MVPN-WC].

When an IR P-tunnel is identified in a Leaf A-D route, its identifier is the "route key" field of the route's NLRI. See section 4.4 of [MVPN-BGP].

A Leaf A-D route is considered to identify an IR P-tunnel only if it carries a PTA whose "Tunnel Type" field is set to "IR". In this type of route, the "Tunnel Identifier" field of the PTA does have significance, but it does not identify the IR P-tunnel. The use of the PTA's "Tunnel Identifier" field in this case is discussed in section 5.

#### 4. How to Join an IR P-tunnel

The procedures for joining an IR P-tunnel depend upon whether the P-tunnel has been previously advertised, and if so, upon how the P-tunnel was advertised. Note that joining an unadvertised P-tunnel is only possible when using the "Global Table Multicast" procedures of [SMLS-MC].

##### 4.1. Advertised P-tunnels

The procedures in this section apply when the P-tunnel to be joined has been advertised in an S-PMSI A-D route, an Inter-AS I-PMSI A-D route, or an Intra-AS I-PMSI A-D route.

The procedures for joining an advertised IR P-tunnel depend upon whether the A-D route that advertises the P-tunnel has the "Leaf Info Required" bit set in its PTA.

###### 4.1.1. If the 'Leaf Info Required Bit' is Set

The procedures in this section apply when the P-tunnel to be joined has been advertised in a route whose PTA has the "Leaf Info Required Bit" set.

The router joining a particular IR P-tunnel must determine its UMH for that P-tunnel. If the route that advertised the P-tunnel contains a P2MP Segmented Next Hop Extended Community, the UMH is determined from the value of this community (see [SMLS-MC]). Otherwise the UMH is determined from the route's next hop (see [MVPN-BGP]).

Once the UMH is determined, the router joining the IR P-tunnel originates a Leaf A-D route. The NLRI of the Leaf A-D route MUST contain the tunnel identifier (as defined in section 3 above) as its "route key". The UMH MUST be identified by attaching an "IP Address Specific Route Target" (or an "IPv6 Address Specific Route Target") to the Leaf A-D route. The IP address of the UMH appears in the "global administrator" field of the Route Target (RT). Details can be found in [MVPN-BGP] and [SMLS-MC].

The Leaf A-D route MUST also contain a PTA whose fields are set as follows:

- The "Tunnel Type" field is set to "IR".
- The "Tunnel Identifier" field is set as described in section 5 of this document.
- The "MPLS Label" field is set to a non-zero value. This is the "P-tunnel label". The value must be chosen so as to satisfy various constraints, as discussed in section 6 of this document.

#### 4.1.2. If the 'Leaf Info Required Bit' is Not Set

The procedures in this section apply when the P-tunnel to be joined has been advertised in a route whose PTA does not have the "Leaf Info Required Bit" set. This can only be the case if the P-tunnel was advertised in an Intra-AS I-PMSI A-D route.

If an IR P-tunnel is advertised in the Intra-AS I-PMSI A-D routes originated by the PE routers of a given MVPN, the Intra-AS I-PMSI can be thought of as being instantiated by a set of IR P-tunnels. Each PE is the root of one such P-tunnel, and the other PEs are children of the root. A PE simultaneously joins all these P-tunnels by originating (if it hasn't already done so) an Intra-AS I-PMSI A-D route with a PTA whose fields are set as follows:

- The "Tunnel Type" field is set to "IR".
- The "Tunnel Identifier" field is set as described in section 5 of this document.
- The "MPLS Label" field MUST be set to a non-zero value. This label value will be used by the child node to associate a received packet with the I-PMSI of a particular MVPN. The MPLS label allocation policy must be such as to ensure that the binding from label to I-PMSI is one-to-one.

The NLRI and the RTs of the originated I-PMSI A-D route are set as specified in [MVPN-BGP].

Note that if a set of IR P-tunnels is joined in this manner, the "discard from the wrong PE" procedures of [MVPN] section 9.1.1 cannot be applied to that P-tunnel. Thus duplicate prevention on such IR P-tunnels requires the use of either Single Forwarder Selection ([MVPN] section 9.1.2) or native PIM procedures ([MVPN] section 9.1.3).

#### 4.2. Unadvertised P-tunnels

In [SMLS-MC], a procedure is defined for "Global Table Multicast", in which a P-tunnel can be joined even if the P-tunnel has not been previously advertised. See the sections of that document entitled "Leaf A-D Route for Global Table Multicast" and "Constructing the Rest of the Leaf A-D Route". The route key of the Leaf A-D route has the form of the "S-PMSI Route-Type Specific NLRI" in this case, and that should be considered to be the P-tunnel identifier. Note that the procedure for finding the UMH is different in this case; the UMH is the next hop of the best UMH-eligible route towards the "ingress PE". See the section of that document entitled "Determining the Upstream ABR/PE/ASBR (Upstream Node)".

#### 5. The PTA's 'Tunnel Identifier' Field

If the "Tunnel Type" field of a PTA is set to "IR", its "Tunnel Identifier" field is significant only when one of the following two conditions holds:

- The PTA is carried by a Leaf A-D route, or
- The "Leaf Information Required" bit of the "Flags" field of the PTA is not set.

If one of these conditions holds, then the "Tunnel Identifier" field must contain a routable IP address of the originator of the route. (See [MVPN-BGP] sections 9.2.3.2.1 and 9.2.3.4.1 for the detailed specification of the contents of this field.) This address is used by the UMH to determine the unicast tunnel that it will use in order to send data, along the IR P-tunnel identified by the route key, to the originator of the Leaf A-D route.

The means by which the unicast tunnel is determined from this IP address is outside the scope of this document. The means by which the unicast tunnel is set up and maintained is also outside the scope of this document.

Section 4 of [P-ADDR] MUST be applied when a PTA is carried in a Leaf A-D route, and describes how to determine whether the "Tunnel Identifier" field carries an IPv4 or an IPv6 address.

If neither of the above conditions hold, then the "Tunnel Identifier" field is of no significance, and MUST be ignored upon reception.



## 6. The PTA's 'MPLS Label' Field

When a PTA is carried by an S-PMSI A-D route or an Inter-AS I-PMSI A-D route, and the "Tunnel Type" field is set to "IR", the "MPLS Label" field is of no significance. In this case, it SHOULD be set to zero upon transmission and MUST be ignored upon reception.

The "MPLS Label" field is significant only when the PTA appears either in a Leaf A-D route or in an Intra-AS I-PMSI A-D route that does not have the "Leaf Information Required" bit set. In these cases, the MPLS label is the label that the originator of the route is assigning to the IR P-tunnel(s) identified by the route's NLRI. (That is, the MPLS label assigned in the PTA is what we have called the "P-tunnel label".)

### 6.1. Leaf A-D Route Originated by an Egress PE

As previously stated, when a Leaf A-D route is used to join an IR P-tunnel, the "route key" of the Leaf A-D route is the P-tunnel identifier.

We now define the notion of the "root of an IR P-tunnel".

- If the identifier of an IR P-tunnel is of the form of an S-PMSI NLRI, the "root" of the P-tunnel is the router identified in the "Originating Router's IP Address" field of that NLRI.
- If the identifier of an IR P-tunnel is of the form specified in Section "Leaf A-D Route for Global Table Multicast" of [SMLS-MC], the "root" of the P-tunnel is the router identified in the "Ingress PE's IP Address" field of that NLRI.
- If the identifier of an IR P-tunnel is of the form of an Intra-AS I-PMSI NLRI, the "root" of the P-tunnel is the router identified in the "Originating Router's IP Address" field of that NLRI.
- If the identifier of an IR P-tunnel is of the form of an Inter-AS I-PMSI NLRI, the "root" of the P-tunnel is same as the identifier of the P-tunnel, i.e., the combination of an RD and an AS.

Note that if a P-tunnel is segmented, the root of the P-tunnel, by this definition, is actually the root of the entire P-tunnel, not the root of the local segment.

In order to apply the procedures of RFC 6513 Section 9.1.1 ("Discarding Packets from Wrong PE"), the following condition MUST be met by the MPLS label allocation policy:.

Suppose an egress PE originates two Leaf A-D routes, each with a different route key in its NLRI, and each with a PTA specifying a "Tunnel Type" of "IR". Thus each of the Leaf A-D routes identifies a different IR P-tunnel. Suppose further that each of those IR P-tunnels has a different root. Then the egress PE MUST NOT specify the same MPLS label in both PMSI Tunnel attributes.

That is, to apply the "Discarding Packets from the Wrong PE" duplicate prevention procedures ([MVPN] section 9.1.1), the same MPLS label MUST NOT be assigned to two IR P-tunnels that have different roots.

If segmented P-tunnels are in use, the above rule is necessary but not sufficient to prevent a PE from forwarding duplicate data to the CEs. For various reasons, a given egress PE or egress ABR or egress ASBR may decide to change its parent node, on a given segmented P-tunnel, from one router to another. It does this by changing the RT of the Leaf A-D route that it originated in order to join that P-tunnel. Once the RT is changed, there may be a period of time during which the old parent node and the new parent node are both sending data of the same multicast flow. To ensure that the egress node not forward duplicate data, whenever the egress node changes the RT that it attaches to a Leaf A-D route, it MUST also change the "MPLS Label" specified in the Leaf A-D route's PTA. This allows the egress router to distinguish between packets arriving on a given P-tunnel from the old parent and packets arriving on that same P-tunnel from the new parent. At any given time, a router MUST consider itself to have only a single parent node on a given P-tunnel, and MUST discard traffic that arrives on that P-tunnel from a different parent node.

If extranet functionality [MVPN-XNET] is not implemented in a particular egress PE, or if an egress PE is provisioned with the knowledge that extranet functionality is not needed, the PE may adopt the policy of assigning a label that is unique for the ordered triple <root, parent node, egress VRF>. This will enable the egress PE to apply the duplicate prevention procedures discussed above, and to determine the VRF to which an arriving packet must be directed.

However, this policy is not sufficient to support the "Discard Packets from the Wrong P-tunnel" procedures that are specified in [MVPN-XNET]. To support those procedures, the labels specified in the PTA of Leaf A-D routes originated by a given egress PE MUST be unique for the ordered triple <root, root RD, parent node>, where the "root RD" is taken from the RD field of the IR P-tunnel identifier. (All forms of IR P-tunnel identifier contain an embedded "RD" field.) This policy is also sufficient for supporting non-extranet cases, but in some cases may result in the use of more labels than the policy of the previous paragraph.

## 6.2. Leaf A-D Route Originated by an Intermediate Node

When a P-tunnel is segmented, there will be "intermediate nodes" (nodes that have a parent and also have children on the P-tunnel). Each intermediate node is a leaf node of an "upstream segment" and a parent node of a "downstream segment". The intermediate node "splices" together the two segments, so that data it receives on the upstream segment gets transmitted on the downstream segment. If either the upstream or downstream segments (or both) are instantiated by IR, the need to do this splicing places certain constraints on the MPLS label allocation policy.

### 6.2.1. Upstream and Downstream Segments are IR Segments

An intermediate node N (i.e., a node that has a parent and also has children) on an IR P-tunnel may originate a Leaf A-D route with a particular route key as a result of receiving a Leaf A-D route with that same route key. This will happen only if the received Leaf A-D route carries an IP address specific RT whose Global Administrator field identifies node N.

Suppose intermediate node N originates two Leaf A-D routes, one whose route key is K1, and one whose route key is K2, where  $K1 \neq K2$ . In general, the respective PTAs of these Leaf A-D routes MUST specify distinct non-zero MPLS labels, such that it is possible to map uniquely from the specified label value to a single IR P-tunnel (call this the "uniqueness rule"). There is one exception to this rule; the exception is specified below.

Consider the set of Leaf A-D routes with route key K1 or route key K2 such that:

- N has received these Leaf A-D routes and has them currently installed.
- Each of these Leaf A-D routes carries an IP Address Specific Route Target that identifies N in its Global Administrator field.

Now suppose that all the Leaf A-D routes in this set have the same originating router, and that the PTAs of these Leaf A-D routes all specify the same MPLS label. Suppose further that N's UMH for K1 is the same as N's UMH for K2. In this particular case, N MAY specify the same MPLS label in the PTA of Leaf A-D route it originates for K1 as in the PTA of the route it originates for K2. However, if at any future time these conditions no longer hold, N must reoriginate at least one of the Leaf A-D routes with a different label so that the "uniqueness rule" holds.

### 6.2.2. Only One Segment is IR

To handle the case where an intermediate node, call it N, is splicing together two P-tunnel segments, only one of which is IR, it is necessary to generalize the rules of the preceding sub-section.

Suppose N is a leaf node of two (upstream) P-tunnel segments, call them U1 and U2. Suppose also that N is a parent node of two (downstream) P-tunnel segments, call them D1 and D2. And suppose that N needs to splice U1 to D1, and U2 to D2.

To follow the uniqueness rule of section 6.2.1 of this document, N must assign a different MPLS label to U1 than it assigns to U2. How this assignment is made depends, of course, on the control protocol used to set up U1 and U2.

There is one case in which the uniqueness rule need not be followed. Suppose that there is a node M such that (a) M is N's only child node on D1, and (b) M is N's only child node on D2. M will have advertised to N a label L1 bound to D1, and a label L2 bound to D2. If (and for as long as)  $L1=L2$ , then N MAY violate the uniqueness rule by advertising to its parent node for U1 the same MPLS label it advertises to its parent node for U2.

Section 6.2.1 of this document specifies in detail the way this requirement is applied when the upstream and downstream segments are all IR segments.

### 6.3. Intra-AS I-PMSI A-D Route

When a router joins a set of IR P-tunnels using the procedures of section 4.1.2 of this document, the procedures of section 9.1.1 of [MVPN] cannot be applied, no matter what the label allocation policy is. In this case, the ingress PE is the same as the UMH, but it is not possible to assign a label uniquely to a particular ingress PE or UMH. However, the label in the MPLS label field of the PTA MUST NOT appear in the MPLS label field of the PTA carried by any other route originated by the same router.

## 7. How A Child Node Prunes Itself from an IR P-tunnel

If a particular IR P-tunnel was joined via the procedures of section 4.1.2 of this document, a router can prune itself from the P-tunnel by withdrawing the Intra-AS I-PMSI A-D route it used to join the P-tunnel. This is not usually done unless the router is removing itself entirely from a particular MVPN.

The procedures in the remainder of this section apply when a router joined a particular IR P-tunnel by originating a Leaf A-D route (as described in section 4.1.1 or 4.2 of this document).

If a router no longer has a need to receive any multicast data from a given IR P-tunnel, it may prune itself from the P-tunnel by withdrawing the Leaf A-D route it used to join the tunnel. This is done, e.g., if the router no longer needs any of the flows traveling over the P-tunnel, or if all the flows the router does need are being received over other P-tunnels.

A router that is attached to a particular IR P-tunnel via a particular parent node may determine that it needs to stay joined to that P-tunnel, but via a different parent node. This can happen, for example, if there is a change in the Next Hop or the P2MP Segmented Next Hop Extended Community of the S-PMSI A-D route in which that P-tunnel was advertised. In this case, the router changes the Route Target of the Leaf A-D route it used to join the IR P-tunnel, so that the Route Target now identifies the new parent node.

A parent node must notice when a child node has been pruned from a particular tree, as this will affect the parent node's multicast data state. Note that the pruning of a child node may appear to the parent node as the explicit withdrawal of a Leaf A-D route, or it may appear as a change in the Route Target of a Leaf A-D route. If the Route Target of a particular Leaf A-D route previously identified a particular parent node, but changes so that it no longer does so, the effect on the multicast state of the parent node is the same as if the Leaf A-D route had been explicitly withdrawn.

## 8. Parent Node Actions Upon Receiving Leaf A-D Route

These actions are detailed in [MVPN-BGP] and [SMLS-MC]. Two points of clarification are made:

- If a router R1 receives and installs a Leaf A-D route originated by router R2, R1's multicast state is affected only if the Leaf A-D route carries an "IP Address Specific RT" (or "IPv6 Address Specific RT") whose "global administrator" field identifies R1.

(This is as specified in [MVPN-BGP] and [SMLS-MC].) If a Leaf A-D route's RT does not identify R1, but then changes so that it does identify R1, R1 must take the same actions it would take if the Leaf A-D route were newly received.

- It is possible that router R1 will receive and install a Leaf A-D route originated by router R2, where:

- \* the route's RT identifies R1,
- \* the route's NLRI contains a route key whose first octet indicates that it is identifying a P-tunnel advertised in an S-PMSI A-D route,
- \* R1 has neither originated nor installed any such S-PMSI A-D route.

If at some later time, R1 installs the corresponding S-PMSI A-D route, and the Leaf A-D route is still installed, and the Leaf A-D route's RT still identifies R1, then R1 MUST follow the same procedures it would have followed if the S-PMSI A-D route had been installed before the Leaf A-D route was installed. (I.e., implementers must not assume that events occur in the "usual" or "expected" order.)

## 9. Use of Timers when Switching UMH

Suppose a child node has joined a particular IR P-tunnel via a particular UMH, and it now determines (for whatever reason) that it needs to change its UMH on that P-tunnel. It does this by modifying the RT of a Leaf A-D route.

It is desirable for such a "switch of UMH" to be done using a "make before break" technique, so that the older UMH does not stop transmitting the packets on the given P-tunnel to the child until the newer UMH has a chance to start transmitting the packets on the given P-tunnel to the child. However, the control plane operation (modifying the RT of the Leaf A-D route) does not permit the child node to first join the P-tunnel at the new UMH, and then later prune itself from the old UMH; a single control plane operation has both effects. Therefore, to achieve "make before break", timers must be used as follows:

1. The old UMH must continue transmitting to the child node for a period of time after it sees the child's Leaf A-D route being withdrawn (or its RT changing to identify a different UMH).

2. The child node must continue to accept packets from the old UMH for a period of time before it starts to accept packets from the new UMH (and discard packets from the old).

Further, the timer in 1 should be longer than the timer in 2. This allows the child to switch from one UMH to another without any loss of data.

## 10. IANA Considerations

This document contains no actions for IANA.

## 11. Acknowledgments

The authors wish to thank Yakov Rekhter for his contributions to this work. We also wish to thank Huajin Jeng and Samir Saad for their contributions, and to thank Thomas Morin for pointing out some of the issues that needed further elaboration.

Section 6.1 discusses the importance of having an MPLS label allocation policy that, when ingress replication is used, allows an egress PE to infer the identity of a received packet's ingress PE. This issue was first raised in earlier work by Xu Xiaohu.

## 12. Security Considerations

No security considerations are raised by this document beyond those already discussed in [MVPN] and [MVPN-BGP].

## 13. Authors' Addresses

Eric C. Rosen  
Cisco Systems, Inc.  
1414 Massachusetts Avenue  
Boxborough, MA, 01719  
Email: [erosen@cisco.com](mailto:erosen@cisco.com)

Karthik Subramanian  
Cisco Systems, Inc.  
170 Tasman Drive  
San Jose, CA, 95134  
Email: kartsubr@cisco.com

Jeffrey Zhang  
Juniper Networks  
10 Technology Park Dr.  
Westford, MA 01886  
Email: zzhang@juniper.net

#### 14. Normative References

- [MVPN] "Multicast in MPLS/BGP IP VPNs", E. Rosen and R. Aggarwal, editors, RFC 6513, February 2012
- [MVPN-BGP] "BGP Encodings and Procedures for Multicast in MPLS/BGP IP VPNs", R. Aggarwal, E. Rosen, T. Morin, Y. Rekhter, RFC 6514, February 2012
- [P-ADDR] "IPv4 and IPv6 Infrastructure Addresses in Updates for Multicast VPN", R. Aggarwal and E. Rosen, RFC 6515, February 2012
- [RFC2119] "Key words for use in RFCs to Indicate Requirement Levels.", Bradner, March 1997
- [SMLS-MC] "Inter-Area P2MP Segmented LSPs", Y. Rekhter, R. Aggarwal, T. Morin, I. Grosclaude, N. Leymann, S. Saad, draft-ietf-mppls-seamless-mcast-07.txt, May 2013

#### 15. Informational References

- [C-BIDIR-IR] "Simulating "Partial Mesh of MP2MP P-Tunnels" with Ingress Replication", Zhang, Rekhter, Dolganow, draft-zzhang-l3vpn-mvpn-bidir-ingress-replication-00.txt, June 2013
- [LDP] "LDP Specification", L. Andersson, I. Minei, and B. Thomas, editors, RFC 5036, October 2007
- [MVPN-BIDIR] "MVPN: Using Bidirectional P-Tunnels", Rosen, Wijnands, Cai, Boers, draft-ietf-l3vpn-mvpn-bidir-06.txt, October 2013



[MVPN-WC] "Wildcards in Multicast VPN Auto-Discovery Routes", Rosen, Rekhter, Henderickx, Qiu, RFC 6625, May 2012

[MVPN-XNET] "Extranet Multicast in BGP/IP MPLS VPNs", Y. Rekhter and E. Rosen (editors), draft-ietf-l3vpn-mvpn-extranet-02.txt, August 2013

[RSVP-TE] "RSVP-TE: Extensions to RSVP for LSP Tunnels", D. Awduche, L. Berger, D. Gan, T. Li, V. Srinivasan, G. Swallow, RFC 3209, December 2001

[SMLS-MC] "Inter-Area P2MP Segmented LSPs", Y. Rekhter, R. Aggarwal, T. Morin, I. Grosclaude, N. Leymann, S. Saad, draft-ietf-mpls-seamless-mcast-07.txt, May 2013

Network working group  
Internet Draft  
Category: Informational

X. Xu  
Huawei

S. Hares

Y. Fan  
China Telecom

C. Jacquenet  
Orange

T. Boyes  
Bloomberg LP

B. Fee  
Enterasys

Expires: March 2014

September 9, 2013

## Virtual Subnet: A L3VPN-based Subnet Extension Solution

draft-xu-l3vpn-virtual-subnet-01

### 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 March 9, 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 .....	11
3.6.1. MAC Table Reduction on Data Center Switches .....	11
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 .....	16
8.1. Normative References .....	16
8.2. Informative References .....	16
Authors' Addresses .....	17

## 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.

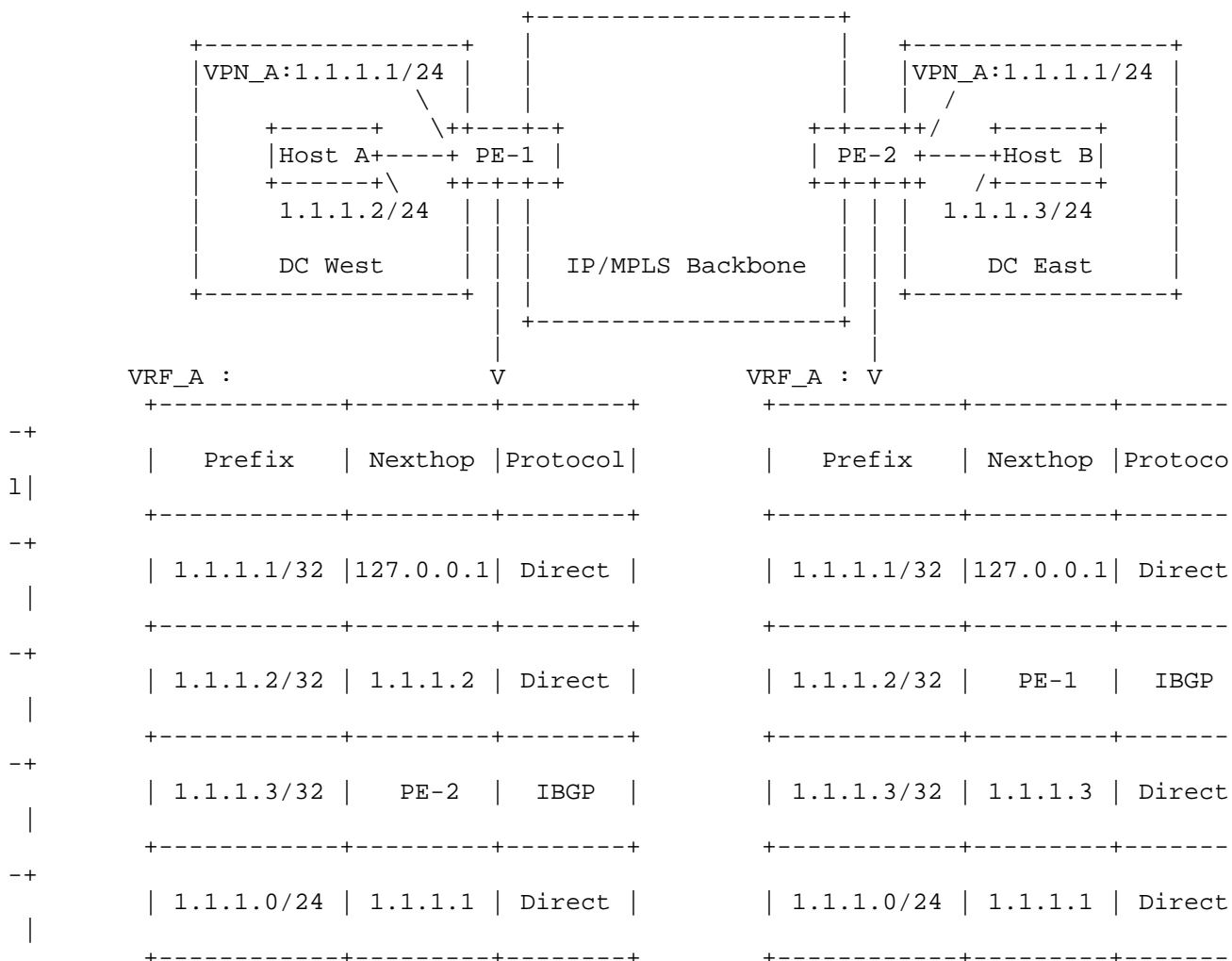
## 2. Terminology

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 March 9, 2014

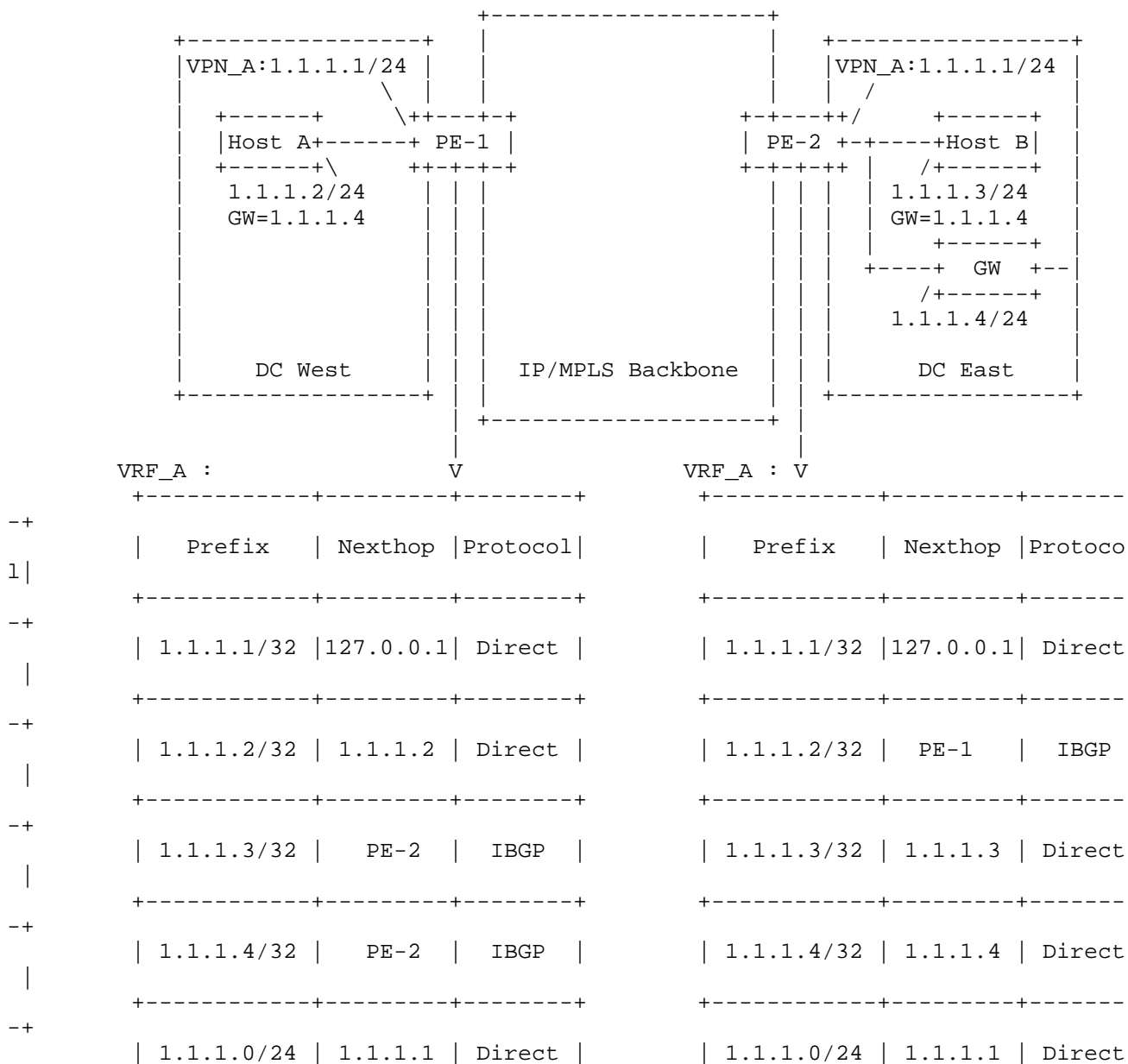
[Page 6]



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



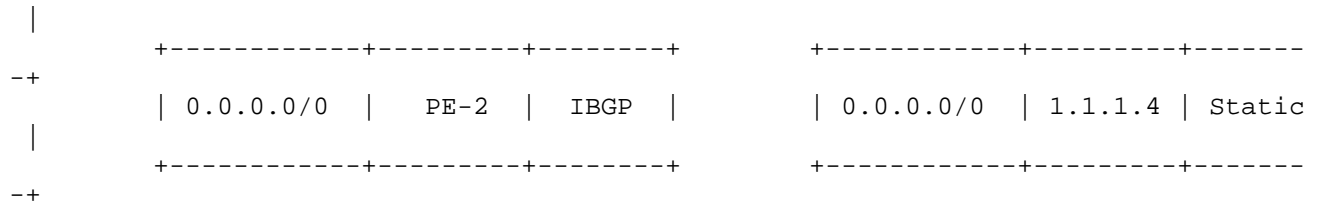
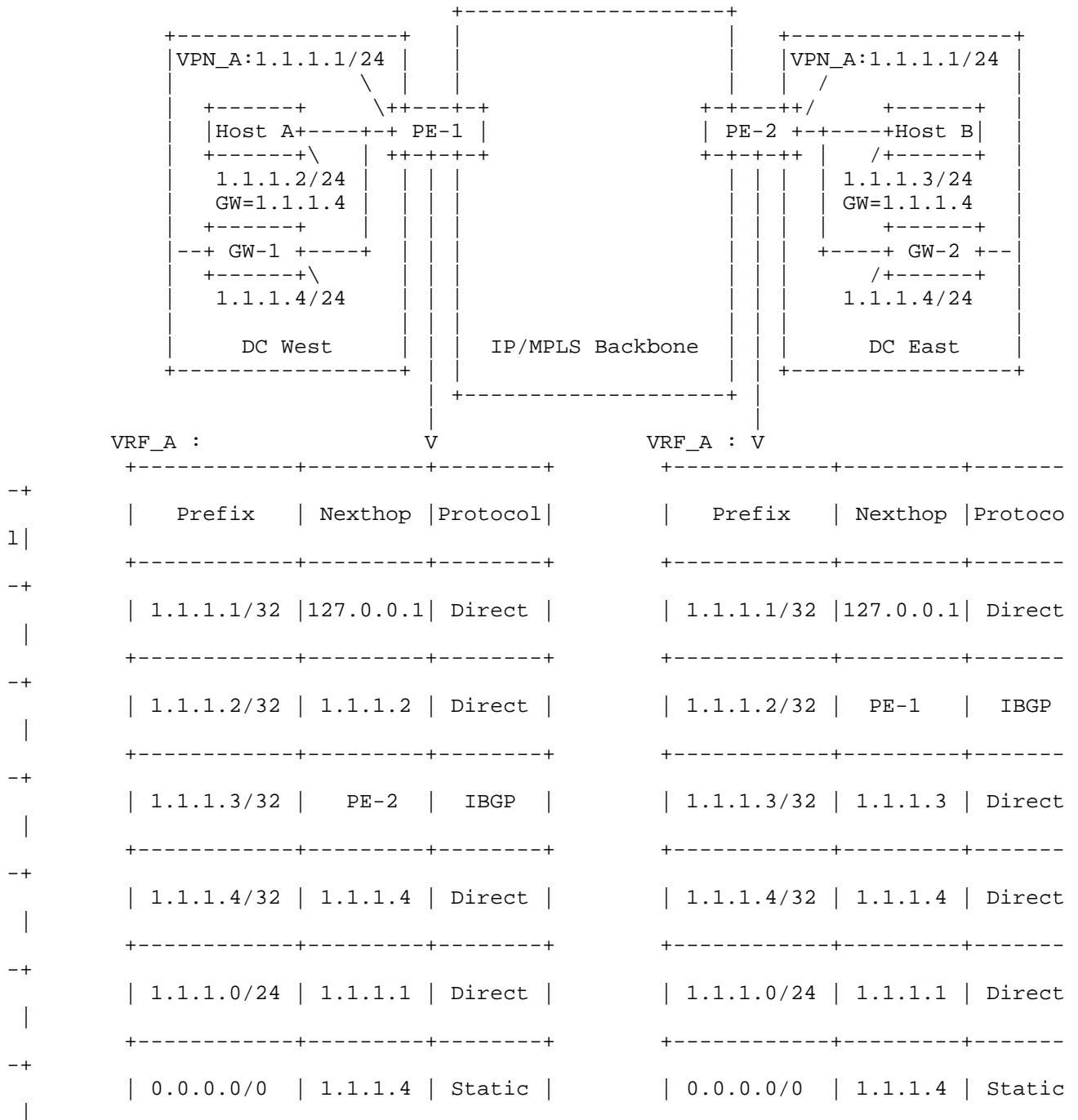


Figure 2: Inter-subnet Unicast Example (1)

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.

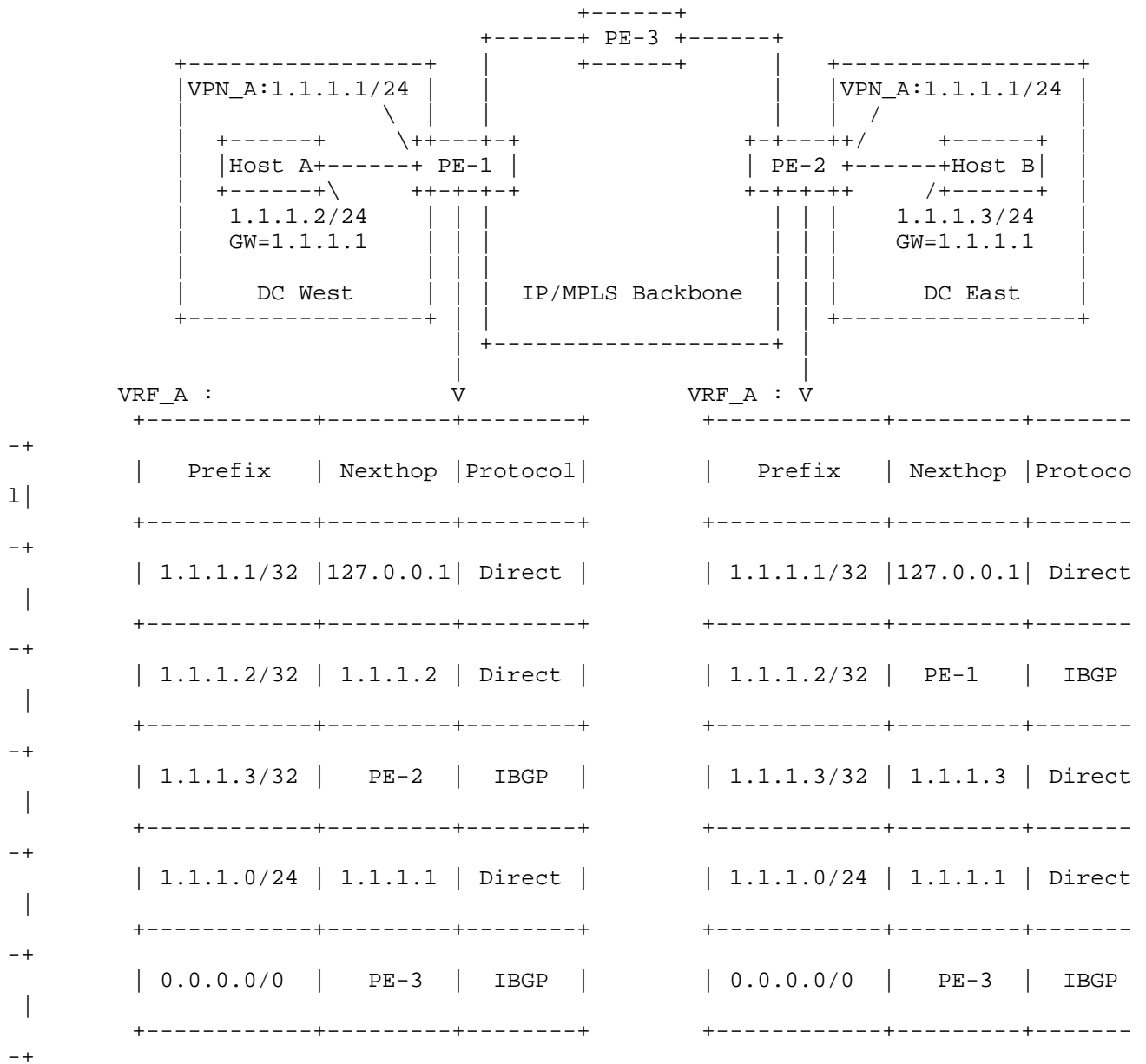


Figure 4: Inter-subnet Unicast Example (3)

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 e.g., a gratuitous ARP or unsolicited NA message. 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 (e.g., a VDP message about VM detachment). Meanwhile, the latter PE router could optionally broadcast a gratuitous ARP or send an unsolicited NA message on behalf of that CE host with source MAC address being one of its own. In this way, the ARP/ND entry of this CE host that moved and which has been cached on any local CE host would be updated accordingly. In the case where there is no explicit VM detachment notification mechanism, the PE router could also use the following trick to determine the VM detachment event: upon learning a route update for a local CE host from a remote PE router for the first time, the PE router could immediately check whether that local CE host is still attached to it by some means (e.g., ARP/ND PING and/or ICMP PING).

### 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

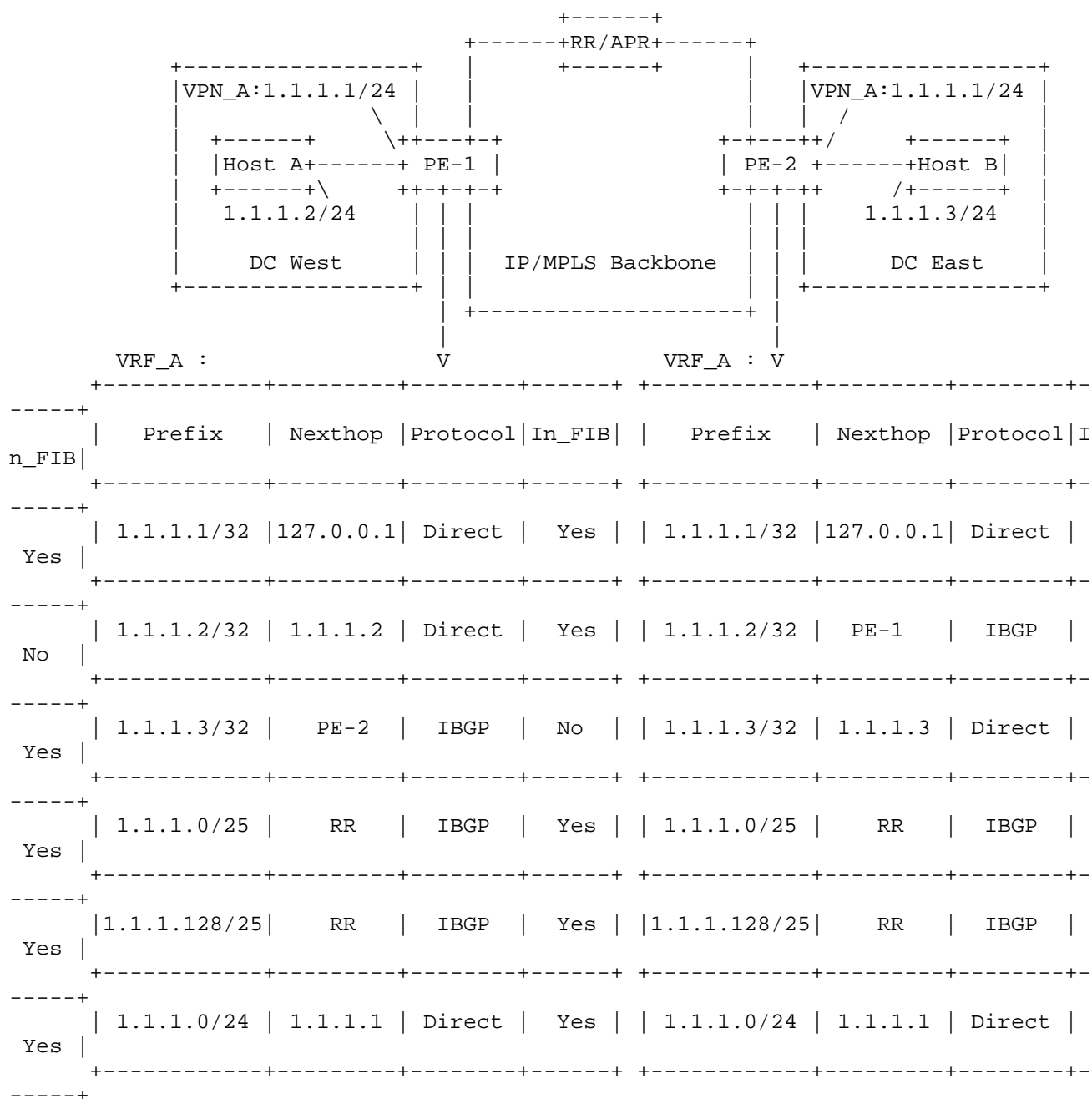


Figure 5: FIB Reduction Example



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, which in turn would install these VP routes into their FIBs.
- 3) Upon receiving a packet destined for a remote CE host from a local CE host, if there is no host route for that remote CE host in the FIB, the ingress PE router will forward the packet to the RR (i.e., APR) 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 corresponding host route for the destination CE host which is learnt from that egress PE router. In a word, the FIB 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 reasons (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 as well. Thus, the RR only needs to attach a "can-suppress" tag to the host routes learnt from one of 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 which has not yet been installed into the FIB. In this way, 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 host routes for remote CE hosts which have been attached with "can-suppress" tags would expire if they have not been used for forwarding for a certain period of time.

### 3.6.3. PE Router RIB Reduction

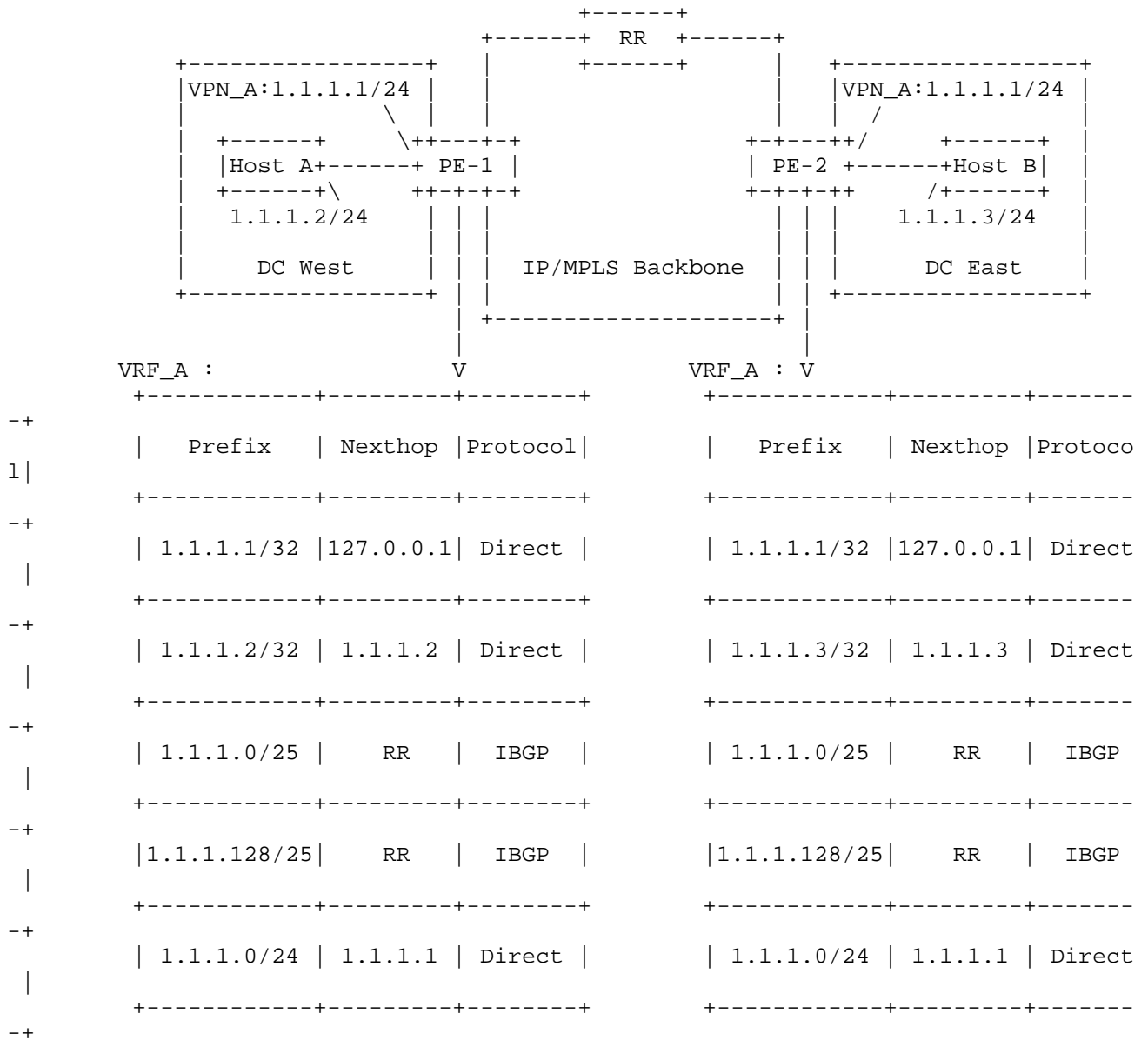


Figure 6: RIB Reduction Example

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 destined for a remote CE host from a local CE host, if there is no host route for that remote CE host in the FIB, 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 learnt 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 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

Yongbing Fan  
Guangzhou Institute, China Telecom  
Guangzhou, China.  
Phone: +86 20 38639121  
Email: fanyb@gsta.com

Christian Jacquenet  
Orange  
Rennes France  
Email: christian.jacquenet@orange.com

Truman Boyes  
Bloomberg LP  
Phone: +1 2126174826  
Email: tboyes@bloomberg.net

Brendan Fee  
Enterasys  
9 Northeastern Blvd.  
Salem, NH, 03079  
Email: bfee@enterasys.com



Network working group  
Internet Draft  
Category: Standard Track

L. Yong  
X. Xu  
Huawei

Expires: April 2014

October 17, 2013

NVGRE and VXLAN Encapsulation Extension for L3 Overlay  
draft-yong-l3vpn-nvgre-vxlan-encap-03

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/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 April 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.

## Abstract

Both NVGRE and VXLAN encapsulations were originally designed for L2 overlay only. This draft proposes the enhancement on both to support L3 overlay as well. The proposed method completely decouples the L3 overlay from the L2 overlay in terms of encoding schema and data processing.

## 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.....	3
2. NVGRE Encapsulation Extension for L3 Overlay.....	3
3. VXLAN Encapsulation Extension for L3 Overlay.....	3
4. Security Considerations.....	4
5. IANA Considerations.....	5
6. References.....	5
6.1. Normative References.....	5
6.2. Informative References.....	5

## 1. Introduction

Network Virtualization Overlay [NVO3FRWK] explicitly states that both L2 and L3 overlays are needed in practice. However both NVGRE encapsulation [NVGRE] and VXLAN encapsulation [VXLAN] were originally designed for L2 overlay only.

This document proposes enhancements to NVGRE and VXLAN encapsulations to allow the same data encapsulation semantics for both L2 overlay and L3 overlay. The benefits of this approach are generalizing the data encapsulation semantics for overlay technologies, maintaining L3 overlay natively, and decoupling it from L2 overlay completely.

## 2. NVGRE Encapsulation Extension for L3 Overlay

NVGER [NVGRE] leverages the GRE protocol [RFC2890] and specifies that the protocol type field in the GRE header MUST be filled with the value of 0x6558, which means for Transparent Ethernet.

This document proposes the protocol type field to be filled with the value of 0x6558, 0x0800(IPv4), or 0x86dd(IPv6). The value of 0x0800 and 0x86dd means that the payload is IP. The value 0x6558 MUST be used if the inner header is an Ethernet header. When NVGRE encapsulation is used for L3 overlay, it MUST use the value of 0x0800 or 0x86dd in the protocol type field and MUST encode an IPv4 or IPv6 header as the inner header. Other fields in the outer header and the GRE header remain the same.

To support backward compatibility, when the remote tunnel end point only support the NVGRE described in [NVGRE], the tunnel end point that supports NVGRE described in this document MUST only encapsulate L2 packets. This capability can be either manually configured or be dynamically informed. How tunnel end points inform each other the encapsulation capabilities is beyond the scope of this document. Note that a tunnel may have more than two end points.

## 3. VXLAN Encapsulation Extension for L3 Overlay

This document proposes adding a protocol type field in the VXLAN header as shown below. It takes 16 bits from the reserved 24 bits as the protocol type field. The remained 8 reserved bits MUST be filled with zero. For L2 overlay encapsulation, the protocol type field MUST be filled with the value of 0x6558 and inner header MUST be an Ethernet header. For L3 overlay encapsulation, the protocol type

field MUST be filled with the value of 0x0800(IPv4) or 0x86dd(IPv6), and inner header MUST be an IPv4 or IPv6 header. Other fields in the outer header and VXLAN header remain the same.

```

      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
Outer Ethernet Header:
      As described in VXLAN [VXLAN]
Outer IP Header:
      As described in VXLAN [VXLAN]
Outer UDP Header:
      As described in VXLAN [VXLAN]
VXLAN Header:
+++++
|R|R|R|R|I|R|R|R|   Reserved   |Prot. Type=0x6558/0x0800/0x86dd|
+++++
|               VXLAN Network Identifier (VNI) |   Reserved   |
+++++
Inner Header:
+++++
|               Ethernet header or IP Header               ~
+++++

```

To be backward compatible with the existing VXLAN encapsulation [VXLAN], the value 0x0000 in the Protocol Type field MUST be treated as Ethernet payload too. When the end points of a tunnel support different VXLAN formats, i.e. one, say A, supports old VXLAN format and another, say B, supports the new format described in this document, B MUST only encapsulate L2 packets and set value 0x0000 in the protocol type field. This capability can be either manually configured at B or be dynamically informed. How tunnel end points inform each other the encapsulation capabilities is beyond the scope of this document. Note that a tunnel may have more than two end points.

Having protocol type field in the VXLAN header enables other overlay payload type beside L2 and L3 overlays. The application for other payload type is for future study.

#### 4. Security Considerations

The mechanism proposed in this document does not add any additional security concern beside what has been described in the NVGRE [NVGRE] and VXLAN [VXLAN].

## 5. IANA Considerations

The document does not require any IANA action.

## 6. References

### 6.1. Normative References

- [RFC2119] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC2119, March 1997.
- [RFC2890] Dommety, G., "Key and Sequence Number Extension to GRE", RFC2890, September 2000

### 6.2. Informative References

- [NVO3FRWK] Lasserre, M., et al, "Framework for DC Network Virtualization", draft-ietf-nvo3-framework-03.txt, work in progress.
- [NVGRE] Sridharan, M., et al, "NVGRE: Network Virtualization using Generic Routing Encapsulation", draft-sridharan-virtualization-nvgre-03, work in progress
- [VXLAN] Mahalingam, M., Dutt, D., etc, "VXLAN: A Framework for Overlaying Virtualized Layer 2 Networks over Layer 3 Networks", draft-mahalingam-dutt-dcops-vxlan-05.txt, work in progress

## Authors' Addresses

Lucy Yong  
Huawei Technologies, USA  
  
Phone: 918-808-1918  
Email: lucy.yong@huawei.com

Xiaohu Xu  
Huawei Technologies,  
Beijing, China  
  
Phone: +86-10-60610041  
Email: xuxiaohu@huawei.com



INTERNET-DRAFT  
Intended Status: Proposed Standard  
Expires: April 24, 2014

Mingui Zhang  
Peng Zhou  
Huawei  
Russ White  
Verisign  
October 21, 2013

Label Sharing for Fast PE Protection  
draft-zhang-l3vpn-label-sharing-01.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, so that the failure affected traffic can be smoothly switched to the backup PE for delivery without changing its VPN route label.

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
2.2. Link Costs Set Up for IGP FRR . . . . .	5
2.3 Label Assignment and Processing . . . . .	6
2.3.1. The VPN Route Label . . . . .	6
2.3.2. The Tunnel Label . . . . .	6
3. Security Considerations . . . . .	7
4. IANA Considerations . . . . .	7
5. References . . . . .	7
5.1. Normative References . . . . .	7
5.2. Informative References . . . . .	8
Appendix A: Generating OSPF LSAs . . . . .	8
Appendix B: Generating ISIS LSPs . . . . .	10
Author's Addresses . . . . .	13



## 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 a 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 (Point of Local Repair) 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 document supposes BGP/MPLS IP VPN is deployed on the backbone and Label Distribution Protocol (LDP) is used to distribute MPLS labels. Through generating virtual LSAs/LSPs in OSPF/ISIS, egress PEs in an RG create a virtual router (the vNH) in the domain of 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 when the PLR detects the primary egress PE fails.

### 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

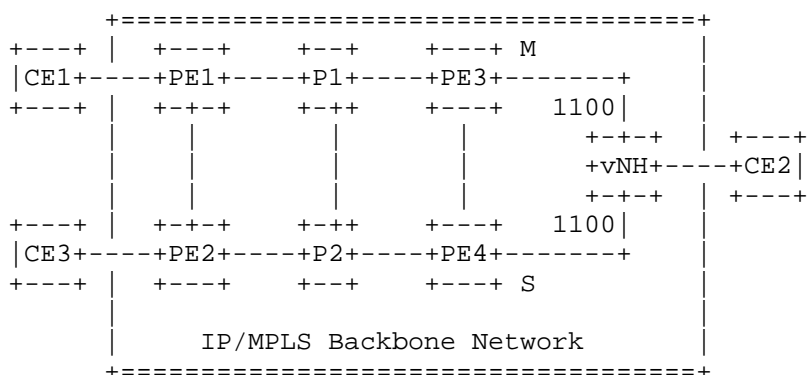


Figure 2.1: Egress PE routers share the same VPN route label 1100.

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. 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]. 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.

### 2.1. The Virtual Next Hop

A vNH router is created in IGP to represent the set of CEs which are dual-homed to the same egress PEs in the Service Provider's backbone. A master PE is elected in the same way the DR is selected (see section 7.3 of [RFC2328]), or the DIS is selected [ISIS]. This master PE determines the loopback IP address for the vNH. This loopback IP address can be configured manually or assigned automatically. The SystemID of the vNH under ISIS is composed based on this loopback IP address. The master PE generates the router link state information (LSA/LSP) on behalf of the vNH. Links to each PE and each CE in the group are included in router link state information PDUs of the PE and CE.

Multiple vNHs may be created 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 mode can be set up through setting the link weights from the vNH to

egress PEs to the maximum link weight which is 0xFFFF. In ISIS, this overload mode is realized as setting the overload bit in the LSP of the vNH.

See Appendix A and B for the detail setting up of LSAs/LSPs.

## 2.2. Link Costs Set Up for IGP FRR

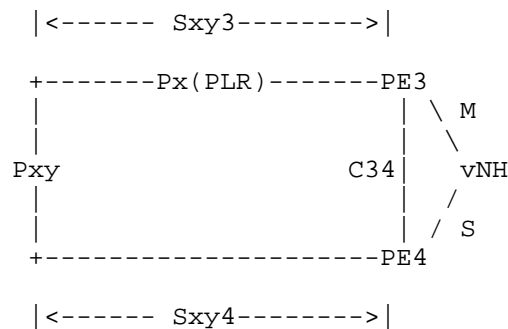


Figure 2.2: The illustration of equations.

LSP tunnels 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 the other PE appears on the backup path, the PLR can make use of the multiple egress PEs to achieve fast failure protection. Link weights can be set up according to the following rule in order to leverage the well supported [LFA] as the IGP FRR mechanism.

1. This document supposes bidirectional link weights are being used. As illustrated in Figure 2.2, 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.

$$\text{eq1: } Sxy4 + S < Sxy3 + M$$

$$\text{eq2: } C34 + S > M$$

The eq1 guarantees that Pxy is safe, i.e., no loop occurs, to be used as the next hop by the PLR for bypass. The eq2 is designed to insure that the primary path does not go through the primary egress PE and backup egress PE in series.

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 learn from CEs [RFC4364]. When egress PEs in an RG 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.

The shared label may be manually configured or negotiated through signaling between egress PEs. [LS-ICCP] extends [ICCP] and defines application TLVs to achieve such kind of signaling. If global label is supported, egress PEs in the RG may use the global label as the shared label. For global label, see Section 3.2.2/[G-use] and [G-frame] for more information.

Suppose PE3 fails and the packet with VPN route label 1100 is redirected to PE4. PE4 can recognize this shared label. It simply looks up the packet's destination address in the VRF identified by this label. 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. In this way, the failed egress PE is smoothly protected.

The handling of PE-CE link failure is out the scope of this document. When the PE-CE link on the primary path fails, the primary PE may resort to existing PE-CE protection mechanisms. It might require that the backup PE had advertised the route to CE using its own IP address as the BGP next hop. When this route is advertised, its preference should be turned down so that the route 'advertised' by the vNH always precedes.

### 2.3.2. The Tunnel Label

In this document, all operations on the tunnel label are widely supported on existing PEs. Suppose Label Distribution Protocol is being used to distribute MPLS labels. The LSP tunnel follows an IGP route from ingress PEs to the vNH. The backup path to vNH can be calculated according to existing 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 pops 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

This document requires no IANA actions. RFC Editor: please remove this section before publication.

### 5. References

#### 5.1. Normative References

- [LFA]      Filssils, 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-03.txt, work in progress.
- [G-use]    Z. Li and Q. Zhao, "Use Cases of MPLS Global Label", draft-li-mpls-global-label-usecases-00.txt, work in progress.
- [G-frame] Z. Li and Q. Zhao, and T. Yang, "A Framework of MPLS Global Label", draft-li-mpls-global-label-framework-00.txt, work in progress.

## Appendix A: Generating OSPF LSAs

The following Type 1 Router-LSA is flooded by the egress PE with the highest priority. As defined in [RFC2328], this LSA can only be flooded throughout a single area.

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								
LS age										Options										LS type																			
Link State ID																																							
Advertising Router																																							
LS sequence number																																							
LS checksum										length																													
0		V		E		B		0		# links																													
Link ID																																							
Link Data																																							
Type										# TOS										metric																			
...																																							
TOS										0										TOS metric																			

```

+-----+
|                                     |
|                               Link ID |
|-----+-----+-----+-----+
|                                     |
|                               Link Data |
|-----+-----+-----+-----+
|                                     |
|                               ...      |
|                                     |
+-----+

```

LS age  
     The time in seconds since the LSA was originated. (Set to 0x708 by default.)

Options  
     As defined in [RFC2328], options = (E-bit).

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.

### Appendix B: Generating ISIS LSPs

The primary egress PE generates the following level 1 LSP to describe the vNH node.

	No. of octets
+-----+	
Intradomain Routeing	1
Protocol Discriminator	
+-----+	
Length Indicator	1
+-----+	
Version/Protocol ID	1
Extension	
+-----+	
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.

Author's Addresses

Mingui Zhang  
Huawei Technologies  
No.156 Beiqing Rd. Haidian District,  
Beijing 100095 P.R. China

Email: zhangmingui@huawei.com

Peng Zhou  
Huawei Technologies  
No.156 Beiqing Rd. Haidian District,  
Beijing 100095 P.R. China

Email: Jewpon.zhou@huawei.com

Russ White  
Verisign  
12061 Bluemont Way  
Reston, VA 20190  
USA

Email: riwhite@verisign.com

L3VPN Working Group  
Internet Draft  
Intended Status: Standards Track  
Expires: June 16, 2014

Jeffrey Zhang  
Lenny Giuliano  
Juniper Networks, Inc.

Eric C. Rosen  
Karthik Subramanian  
Cisco Systems, Inc.

Dante J. Pacella  
Verizon

Jason Schiller  
Google

December 16, 2013

## Global Table Multicast with BGP-MVPN Procedures

draft-zzhang-l3vpn-mvpn-global-table-mcast-02.txt

### Abstract

RFC6513, RFC6514, and other RFCs describe protocols and procedures which a Service Provider (SP) may deploy in order offer Multicast Virtual Private Network (Multicast VPN or MVPN) service to its customers. Some of these procedures use BGP to distribute VPN-specific multicast routing information across a backbone network. With a small number of relatively minor modifications, the very same BGP procedures can also be used to distribute multicast routing information that is not specific to any VPN. Multicast that is outside the context of a VPN is known as "Global Table Multicast", or sometimes simply as "Internet multicast". In this document, we describe the modifications that are needed to use the MVPN BGP procedures for Global Table Multicast.

### 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/ietf/lid-abstracts.txt>.

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
2	Adapting MVPN Procedures to GTM .....	6
2.1	Use of Route Distinguishers .....	7
2.2	Use of Route Targets .....	7
2.3	UMH-eligible Routes .....	9
2.3.1	Routes of SAFI 1, 2 or 4 with MVPN ECs .....	10
2.3.2	MVPN ECs on the Route to the Next Hop .....	11
2.3.3	Non-BGP Routes as the UMH-eligible Routes .....	12
2.3.4	Why SFS Does Not Apply to GTM .....	13
2.4	Inclusive and Selective Tunnels .....	14
2.5	I-PMSI A-D Routes .....	14
2.5.1	Intra-AS I-PMSI A-D Routes .....	14
2.5.2	Inter-AS I-PMSI A-D Routes .....	15
2.6	S-PMSI A-D Routes .....	15
2.7	Leaf A-D Routes .....	15
2.8	Source Active A-D Routes .....	15
2.9	C-multicast Source/Shared Tree Joins .....	16
3	Differences from other MVPN-like GTM Procedures .....	17
4	IANA Considerations .....	18
5	Security Considerations .....	18
6	Acknowledgments .....	18
7	Authors' Addresses .....	19
8	References .....	20
8.1	Normative References .....	20
8.2	Informative References .....	20

## 1. Introduction

[RFC4364] specifies architecture, protocols, and procedures that a Service Provider (SP) can use to provide Virtual Private Network (VPN) service to its customers. In that architecture, one or more Customer Edge (CE) routers attach to a Provider Edge (PE) router. Each CE router belongs to a single VPN, but CE routers from several VPNs may attach to the same PE router. In addition, CEs from the same VPN may attach to different PEs. BGP is used to carry VPN-specific information among the PEs. Each PE router maintains a separate Virtual Routing and Forwarding table (VRF) for each VPN to which it is attached.

[RFC6513] and [RFC6514] extend the procedures of [RFC4364] to allow the SP to provide multicast service to its VPN customers. The customer's multicast routing protocol (e.g., PIM) is used to exchange multicast routing information between a CE and a PE. The PE stores a given customer's multicast routing information in the VRF for that customer's VPN. BGP is used to distribute certain multicast-related control information among the PEs that attach to a given VPN, and BGP may also be used to exchange the customer multicast routing information itself among the PEs.

While this multicast architecture was originally developed for VPNs, it can also be used (with a small number of modifications to the procedures) to distribute multicast routing information that is not specific to VPNs. The purpose of this document is to specify the way in which BGP MVPN procedures can be adapted to support non-VPN multicast.

Multicast routing information that is not specific to VPNs is stored in a router's "global table", rather than in a VRF; hence it is known as "Global Table Multicast" (GTM). GTM is sometimes more simply called "Internet multicast". However, we will avoid that term because it suggests that the multicast data streams are available on the "public" Internet. The procedures for GTM can certainly be used to support multicast on the public Internet, but they can also be used to support multicast streams that are not public, e.g., content distribution streams offered by content providers to paid subscribers. For the purposes of this document, all that matters is that the multicast routing information is maintained in a global table rather than in a VRF.

This architecture does assume that the network over which the multicast streams travel can be divided into a "core network" and one or more non-core parts of the network, which we shall call "attachment networks". The multicast routing protocol used in the attachment networks may not be the same as the one used in the core,

so we consider there to be a "protocol boundary" between the core network and the attachment networks. We will use the term "Protocol Boundary Router" (PBR) to refer to the core routers that are at the boundary. We will use the term "Attachment Router" (AR) to refer to the routers that are not in the core but that attach to the PBRs.

This document does not make any particular set of assumptions about the protocols that the ARs and the PBRs use to exchange unicast and multicast routing information with each other. For instance, multicast routing information could be exchanged between an AR and a PBR via PIM, IGMP, or even BGP. Multicast routing also depends on an exchange of routes that are used for looking up the path to the root of a multicast tree. This routing information could be exchanged between an AR and a PBR via IGP, via EBGp, or via IBGP ([RFC6368]). Note that if IBGP is used, the [RFC6368] "push/pop procedures" are not necessary.

The PBRs are not necessarily "edge" routers, in the sense of [RFC4364]. For example, they may be both be Autonomous System Border Routers (ASBR). As another example, an AR may be an "access router" attached to a PBR that is an OSPF Area Border Router (ABR). Many other deployment scenarios are possible. However, the PBRs are always considered to be delimiting a "backbone" or "core" network. A multicast data stream from an AR is tunneled over the core network from an Ingress PBR to one or more Egress PBRs. Multicast routing information that a PBR learns from the ARs attached to it is stored in the PBR's global table. The PBRs use BGP to distribute multicast routing and auto-discovery information among themselves. This is done following the procedures of [RFC6513], [RFC6514], and other MVPN specifications, as modified in this document.

In general, PBRs follow the same MVPN/BGP procedures that PE routers follow, except that these procedures are adapted to be applicable to the global table rather than to a VRF. Details are provided in subsequent sections of this document.

By supporting GTM using the BGP procedures designed for MVPN, one obtains a single control plane that governs the use of both VPN and non-VPN multicast. Most of the features and characteristics of MVPN carry over automatically to GTM. These include scaling, aggregation, flexible choice of tunnel technology in the SP network, support for both segmented and non-segmented tunnels, ability to use wildcards to identify sets of multicast flows, support for the Any Source Multicast (ASM), Single Source Multicast (SSM), and Bidirectional (bidir) multicast paradigms, support for both IPv4 and IPv6 multicast flows over either an IPv4 or IPv6 SP infrastructure, support for unsolicited flooded data (including support for BSR as RP-to-group mapping protocols), etc.



This document not only uses MVPN procedures for GTM, but also, insofar as possible, uses the same protocol elements, encodings, and formats. The BGP Updates for GTM thus use the same Subsequent Address Family Identifier (SAFI), and have the same Network Layer Reachability Information (NLRI) format, as the BGP Updates for MVPN.

Details for supporting MVPN (either IPv4 or IPv6 MVPN traffic) over an IPv6 backbone network can be found in [RFC6515]. The procedures and encodings described therein are also applicable to GTM.

The document [SEAMLESS-MCAST] extends [RFC6514] by providing procedures that allow tunnels through the core to be "segmented" at ABRs within the core. The ABR segmentation procedures are also applicable to GTM as defined in the current document. In general, the MVPN procedures of [SEAMLESS-MCAST], adapted as specified in the current document, are applicable to GTM.

The document [SEAMLESS-MCAST] also defines a set of procedures for GTM. Those procedures are different from the procedures defined in the current document, and the two sets of procedures are not interoperable with each other. The two sets of procedures can co-exist in the same network, as long as they are not applied to the same multicast flows or to the same multicast group addresses. See section 3 for more details.

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. Adapting MVPN Procedures to GTM

In general, PBRs support Global Table Multicast by using the procedures that PE routers use to support VPN multicast. For GTM, where [RFC6513] and [RFC6514] talk about the "PE-CE interface", one should interpret that to mean the interface between the AR and the PBR. For GTM, where [RFC6513] and [RFC6514] talk about the "backbone" network, one should interpret that to mean the part of the network that is delimited by the PBRs.

A few adaptations to the procedures of [RFC6513] and [RFC6514] need to be made. Those adaptations are described in the following subsections.

## 2.1. Use of Route Distinguishers

The MVPN procedures require the use of BGP routes, defined in [RFC6514], that have a SAFI value of 5 ("MCAST-VPN"). We refer to these simply as "MCAST-VPN routes". [RFC6514] defines the Network Layer Reachability Information (NLRI) format for MCAST-VPN routes. The NLRI field always begins with a "Route Type" octet, and, depending on the route type, may be followed by a "Route Distinguisher" (RD) field.

When a PBR originates an MCAST-VPN route in support of GTM, the RD field (for those routes types where it is defined) of that route's NLRI MUST be set to zero (i.e., to 64 bits of zero). Since no VRF may have an RD of zero, this allows "MCAST-VPN" routes that are "about" GTM to be distinguished from MCAST-VPN routes that are about VPNs.

## 2.2. Use of Route Targets

The MVPN procedures require all MCAST-VPN routes to carry Route Targets (RTs). When a PE router receives an MCAST-VPN route, it processes the route in the context of a particular VRF if and only if the route is carrying an RT that is configured as one of that VRF's "import RTs".

There are two different "kinds" of RT used in MVPN.

- One kind of RT is carried only by the following MCAST-VPN route types: C-multicast Shared Tree Joins, C-multicast Source Tree Joins, and Leaf A-D routes. This kind of RT identifies the PE router that has been selected by the route's originator as the "Upstream PE" or as the "Upstream Multicast Hop" (UMH) for a particular (set of) multicast flow(s). Per [RFC6514] and [RFC6515], this RT must be an IPv4-address-specific or IPv6-address-specific Extended Community (EC), whose "Global Administrator" field identifies the Upstream PE or the UMH. If the Global Administrator field identifies the Upstream PE, the "Local Administrator" field identifies a particular VRF in that PE.

The GTM procedures of this document require the use of this type of RT, in exactly the same situations where it is used in the MVPN specification. However, one adaptation is necessary: the "Local Administrator" field of this kind of RT MUST always be set to zero, thus implicitly identifying the global table, rather than identifying a VRF. We will refer to this kind of RT as a "PBR-identifying RT".

- The other kind of RT is the conventional RT first specified in [RFC4364]. It does not necessarily identify a particular router by address, but is used to constrain the distribution of VPN routes, and to ensure that a given VPN route is processed in the context of a given VRF if and only if the route is carrying an RT that has been configured as one of that VRF's "import RTs".

Whereas every VRF must be configured with at least one import RT, there is heretofore no requirement to configure any RTs for the global table of any router. As stated above, this document makes the use of PBR-identifying RTs mandatory for GTM. This document makes the use of non-PBR-identifying RTs OPTIONAL for GTM.

The procedures for the use of RTs in GTM are the following:

- If the global table of a particular PBR is NOT configured with any import RTs, then a received MCAST-VPN route is processed in the context of the global table only if it is carrying no RTs, or if it is carrying a PBR-identifying RT whose Global Administrator field identifies that PBR.
- The global table in each PBR MAY be configured with (a) a set of export RTs to be attached to MCAST-VPN routes that are originated to support GTM, and (b) with a set of import RTs for GTM.

If the global table of a given PBR has been so configured, the PBR will process a received MCAST-VPN route in the context of the global table if and only if the route carries an RT that is one of the global table's import RTs, or if the route carries a PBR-identifying RT whose global administrator field identifies the PBR.

If the global tables are configured with RTs, care must be taken to ensure that the RTs configured for the global table are distinct from any RTs used in support of MVPN (except in the case where it is actually intended to create an "extranet" [MVPN-extranet] in which some sources are reachable in global table context while others are reachable in VPN context.)

The "RT Constraint" procedures of [RFC4684] MAY be used to constrain the distribution of MCAST-VPN routes (or other routes) that carry RTs that have been configured as import RTs for GTM. (This includes the PBR-identifying RTs.)

In [RFC6513], the UMH-eligible routes (see section 5.1 of [RFC6513], "Eligible Routes for UMH Selection") are generally routes of SAFI 128 (Labeled VPN-IP routes) or 129 (VPN-IP multicast routes), and are required to carry RTs. These RTs determine which VRFs import which

such routes. However, for GTM, when the UMH-eligible routes may be routes of SAFI 1, 2, or 4, the routes are not required to carry RTs. This document does NOT specify any new rules for determine whether a SAFI 1, 2, or 4 route is to be imported into the global table of any PBR.

### 2.3. UMH-eligible Routes

[RFC6513] section 5.1 defines procedures by which a PE router determines the "C-root", the "Upstream Multicast Hop" (UMH), the "Upstream PE", and the "Upstream RD" of a given multicast flow. (In non-VPN multicast documents, the UMH of a multicast flow at a particular router is generally known as the "RPF neighbor" for that flow.) It also defines procedures for determining the "Source AS" of a particular flow. Note that in GTM, the "Upstream PE" is actually the "Upstream PBR".

The definition of the C-root of a flow is the same for GTM as for MVPN.

For MVPN, to determine the UMH, Upstream PE, Upstream RD, and Source AS of a flow, one looks up the C-root of the flow in a particular VRF, and finds the "UMH-eligible" routes (see section 5.1.1 of [RFC6513]) that "match" the C-root. From among these, one is chosen as the "selected UMH route".

For GTM, the C-root is of course looked up in the global table, rather than in a VRF. For MVPN, the UMH-eligible routes are routes of SAFI 128 or 129. For GTM, the UMH-eligible routes are routes of SAFI 1, SAFI 4, or SAFI 2. If the global table has imported routes of SAFI 2, then these are the UMH-eligible routes. Otherwise, routes of SAFI 1 or SAFI 4 are the UMH-eligible routes. For the purpose of UMH determination, if a SAFI 1 route and a SAFI 4 route contain the same IP prefix in their respective NLRI fields, then the two routes are considered by the BGP bestpath selection process to be comparable.

[RFC6513] defines procedures for determining which of the UMH-eligible routes that match a particular C-root is to become the "Selected UMH route". With one exception, these procedures are also applicable to GTM. The one exception is the following. Section 9.1.2 of [RFC6513] defines a particular method of choosing the Upstream PE, known as "Single Forwarder Selection" (SFS). This procedure MUST NOT be used for GTM (see section 2.3.4 for an explanation of why the SFS procedure cannot be applied to GTM).

In GTM, the "Upstream RD" of a multicast flow is always considered to

be zero, and is NOT determined from the Selected UMH route.

The MVPN specifications require that when BGP is used for distributing multicast routing information, the UMH-eligible routes MUST carry the VRF Route Import EC and the Source AS EC. To determine the Upstream PE and Source AS for a particular multicast flow, the Upstream PE and Source AS are determined, respectively, from the VRF Route Import EC and the Source AS EC of the Selected UMH route for that flow. These ECs are generally attached to the UMH-eligible routes by the PEs that originate the routes.

In GTM, there are certain situations in which it is allowable to omit the VRF Route Import EC and/or the Source AS EC from the UMH-eligible routes. The following sub-sections specify the various options for determining the Upstream PBR and the Source AS in GTM.

The procedures in sections 2.3.1 MUST be implemented. The procedures in sections 2.3.2 and 2.3.3 are OPTIONAL to implement. It should be noted that while the optional procedures may be useful in particular deployment scenarios, there is always the potential for interoperability problems when relying on OPTIONAL procedures.

#### 2.3.1. Routes of SAFI 1, 2 or 4 with MVPN ECs

If the UMH-eligible routes have a SAFI of 1, 2 or 4, then they MAY carry the VRF Route Import EC and/or the Source AS EC. If the selected UMH route is a route of SAFI 1, 2 or 4 that carries the VRF Route Import EC, then the Upstream PBR is determined from that EC. Similarly, if the selected UMH route is a route of SAFI 1, 2, or 4 route that carries the Source AS EC, the Source AS is determined from that EC.

When the procedure of this section is used, a PBR that distributes a UMH-eligible route to other PBRs is responsible for ensuring that the VRF Route Import and Source AS ECs are attached to it.

If the selected UMH-eligible route has a SAFI of 1, 2 or 4, but is not carrying a VRF Route Import EC, then the Upstream PBR is determined as specified in section 2.3.2 or 2.3.3 below.

If the selected UMH-eligible route has a SAFI of 1, 2 or 4, but is not carrying a Source AS EC, then the Source AS is considered to be the local AS.

### 2.3.2. MVPN ECs on the Route to the Next Hop

Some service providers may consider it to be undesirable to have the PBRs put the VRF Route Import EC on all the UMH-eligible routes. Or there may be deployment scenarios in which the UMH-eligible routes are not advertised by the PBRs at all. The procedures described in this section provide an alternative that can be used under certain circumstances.

The procedures of this section are OPTIONAL.

In this alternative procedure, each PBR MUST originate a BGP route of SAFI 1, 2 or 4 to itself. This route MUST carry a VRF Route Import EC that identifies the PBR. The address that appears in the Global Administrator field of that EC MUST be the same address that appears in the NLRI and in the Next Hop field of that route. This route MUST also carry a Source AS EC identifying the AS of the PBR.

Whenever the PBR distributes a UMH-eligible route for which it sets itself as next hop, it MUST use this same IP address as the Next Hop of the UMH-eligible route that it used in the route discussed in the prior paragraph.

When the procedure of this section is used, then when a PBR is determining the Selected UMH Route for a given multicast flow, it may find that the Selected UMH Route has no VRF Route Import EC. In this case, the PBR will look up (in the global table) the route to the Next Hop of the Selected UMH route. If the route to the Next Hop has a VRF Route Import EC, that EC will be used to determine the Upstream PBR, just as if the EC had been attached to the Selected UMH Route.

If recursive route resolution is required in order to resolve the next hop, the Upstream PBR will be determined from the first route with a VRF Route Import EC that is encountered during the recursive route resolution process. (The recursive route resolution process itself is not modified by this document.)

The same procedure can be applied to find the Source AS, except that the Source AS EC is used instead of the VRF Route Import EC.

Note that this procedure is only applicable in scenarios where it is known that the Next Hop of the UMH-eligible routes is not be changed by any router that participates in the distribution of those routes; this procedure MUST NOT be used in any scenario where the next hop may be changed between the time one PBR distributes the route and another PBR receives it. The PBRs have no way of determining dynamically whether the procedure is applicable in a particular deployment; this must be made known to the PBRs by provisioning.

Some scenarios in which this procedure can be used are:

- all PBRs are in the same AS, or
- the UMH-eligible routes are distributed among the PBRs by a Route Reflector (that does not change the next hop), or
- the UMH-eligible routes are distributed from one AS to another through ASBRs that do not change the next hop.

If the procedures of this section are used in scenarios where they are not applicable, GTM will not function correctly.

### 2.3.3. Non-BGP Routes as the UMH-eligible Routes

In particular deployment scenarios, there may be specific procedures that can be used, in those particular scenarios, to determine the Upstream PBR for a given multicast flow.

Suppose the PBRs neither put the VRF Route Import EC on the UMH-eligible routes, nor do they distribute BGP routes to themselves. It may still be possible to determine the Upstream PBR for a given multicast flow, using specific knowledge about the deployment.

For example, suppose it is known that all the PBRs are in the same OSPF area. It may be possible to determine the Upstream PBR for a given multicast flow by looking at the link state database to see which router is attached to the flow's C-root.

As another example, suppose it is known that the set of PBRs is fully meshed via Traffic Engineering (TE) tunnels. When a PBR looks up, in its global table, the C-root of a particular multicast flow, it may find that the next hop interface is a particular TE tunnel. If it can determine the identify of the router at the other end of that TE tunnel, it can deduce that that router is the Upstream PBR for that flow.

This is not an exhaustive set of examples. Any procedure that correctly determines the Upstream PBR in a given deployment scenario MAY be used in that scenario.

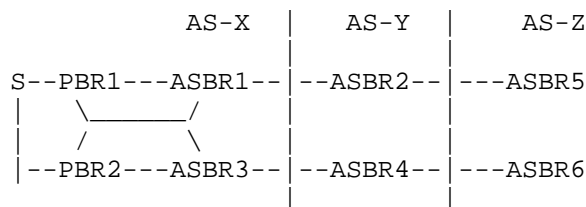
#### 2.3.4. Why SFS Does Not Apply to GTM

To see why the SFS procedure cannot be applied to GTM, consider the following example scenario. Suppose some multicast source S is homed to both PBR1 and PBR2, and suppose that both PBRs export a route (of SAFI 1, 2, or 4) whose NLRI is a prefix matching the address of S. These two routes will be considered comparable by the BGP decision process. A route reflector receiving both routes may thus choose to redistribute just one of the routes to S, the one chosen by the bestpath algorithm. Different route reflectors may even choose different routes to redistribute (i.e., one route reflector may choose the route to S via PBR1 as the bestpath, while another chooses the route to S via PBR2 as the bestpath). As a result, some PBRs may receive only the route to S via PBR1 and some may receive only the route to S via PBR2. In that case, it is impossible to ensure that all PBRs will choose the same route to S.

The SFS procedure works in VPN context as along the following assumption holds: if S is homed to VRF-x in PE1 and to VRF-y in PE2, then VRF-x and VRF-y have been configured with different RDs. In VPN context, the route to S is of SAFI 128 or 129, and thus has an RD in its NLRI. So the route to S via PE1 will not have the same NLRI as the route to S via PE2. As a result, all PEs will see both routes, and the PEs can implement a procedure that ensures that they all pick the same route to S.

That is, the SFS procedure of [RFC6513] relies on the UMH-eligible routes being of SAFI 128 or 129, and relies on certain VRFs being configured with distinct RDs. Thus the procedure cannot be applied to GTM.

One might think that the SFS procedure could be applied to GTM as long as the procedures defined in [ADD-PATH] are applied to the UMH-eligible routes. Using the [ADD-PATH] procedures, the BGP speakers could advertise more than one path to a given prefix. Typically [ADD-PATH] is used to report the n best paths, for some small value of n. However, this is not sufficient to support SFS, as can be seen by examining the following scenario.





In AS-X, PBR1 reports to both ASBR1 and ASBR3 that it has a route to S. Similarly, PBR2 reports to both ASBR1 and ASBR3 that it has a route to S. Using [ADD-PATH], ASBR1 reports both routes to ASBR2, and ASBR3 reports both routes to ASBR4. Now AS-Y sees 4 paths to S. The AS-Z ASBRs will each see eight paths (four via ASBR2 and four via ASBR4). To avoid this explosion in the number of paths, a BGP speaker that uses [ADD-PATH] is usually considered to report only the n best paths. However, there is then no guarantee that the reported set of paths will contain at least one path via PBR1 and at least one path via PBR2. Without such a guarantee, the SFS procedure will not work.

#### 2.4. Inclusive and Selective Tunnels

The MVPN specifications allow multicast flows to be carried on either Inclusive Tunnels or on Selective Tunnels. When a flow is sent on an Inclusive Tunnel of a particular VPN, it is sent to all PEs in that VPN. When sent on a Selective Tunnel of a particular VPN, it may be sent to only a subset of the PEs in that VPN.

This document allows the use of either Inclusive Tunnels or Selective Tunnels for GTM. However, any service provider electing to use Inclusive Tunnels for GTM should carefully consider whether sending a multicast flow to ALL its PBRs would result in problems of scale. There are potentially many more MBRs for GTM than PEs for a particular VPN. If the set of PBRs is large and growing, but most multicast flows do not need to go to all the PBRs, the exclusive use of Selective Tunnels may be a better option.

#### 2.5. I-PMSI A-D Routes

##### 2.5.1. Intra-AS I-PMSI A-D Routes

Per [MVPN-BGP}, there are certain conditions under which is it NOT required for a PE router implementing MVPN to originate one or more Intra-AS I-PMSI A-D routes. These conditions apply as well to PBRs implementing GTM.

In addition, a PBR implementing GTM is NOT required to originate an Intra-AS I-PMSI A-D route if both of the following conditions hold:

- The PBR is not using Inclusive Tunnels for GTM, and

- The distribution of the C-multicast Shared Tree Join and C-multicast Source Tree Join routes is done in such a manner that the next hop of those routes does not change.

Please see also the sections on RD and RT usage.

#### 2.5.2. Inter-AS I-PMSI A-D Routes

There are no GTM-specific procedures for the origination, distribution, and processing of these routes, other than those specified in the sections on RD and RT usage.

#### 2.6. S-PMSI A-D Routes

There are no GTM-specific procedures for the origination, distribution, and processing of these routes, other than those specified in the sections on RD and RT usage.

#### 2.7. Leaf A-D Routes

There are no GTM-specific procedures for the origination, distribution, and processing of these routes, other than those specified in the sections on RD and RT usage.

#### 2.8. Source Active A-D Routes

There are no MANDATORY GTM-specific procedures for the origination, distribution, and processing of these routes, other than those specified in the sections on RD and RT usage.

However, this document defines an OPTIONAL procedure to allow additional constraints on the distribution of the Source Active A-D routes for GTM. If some site has receivers for a particular ASM group G, then it is possible (by the procedures of [RFC6514]) that every PBR attached to site with a source for group G will originate a Source Active A-D route whose NLRI identifies that source and group. These Source Active A-D routes may be distributed to every PBR. If only a relatively small number of PBRs are actually interested in traffic from group G, but there are many sources for group G, this could result in a large number of (S,G) Source Active A-D routes being installed in a large number of PBRs that have no need of them.

For GTM, it is possible to constrain the distribution of (S,G) Source Active A-D routes to those PBRs that are interested in GTM traffic to

group G. This can be done using the following OPTIONAL procedures:

- If a PBR originates a C-multicast Shared Tree Join whose NLRI contains (RD=0,\*,G), then it dynamically creates an import RT for its global table, where the Global Administrator field of the RT contains the group address G, and the Local Administrator field contains zero. (Note that an IPv6-address-specific RT would need to be used if the group address is an IPv6 address.)
- When a PBR creates such an import RT, it uses "RT Constraint" [RFC4684] procedures to advertise its interest in routes that carry this RT.
- When a PBR originates a Source Active A-D route from its global table, it attaches the RT described above.
- When the C-multicast Shared Tree Join is withdrawn, so is the corresponding RT constrain route, and the corresponding RT is removed as an import RT of its global table.

These procedures enable a PBR to automatically filter all Source Active A-D routes that are about multicast groups in which the PBR has no interest.

This procedure does introduce the overhead of distributing additional "RT Constraint" routes, and therefore may not be cost-effective in all scenarios, especially if the number of sources per ASM group is small. This procedure may also result in increased join latency.

## 2.9. C-multicast Source/Shared Tree Joins

[RFC6514] section 11.1.3 has the following procedure for determining the IP-address-specific RT that is attached to a C-multicast route: (a) determine the upstream PE, RD, AS, (b) find the proper Inter-AS or Intra-AS I-PMSI A-D route based on (a), (c) find the next hop of that A-D route, (d) base the RT on that next hop.

However, for GTM, in environments where it is known a priori that that the next hop of the C-multicast Source/Shared Tree Joins does not change during the distribution of those routes, the proper procedure for creating the IP-address-specific RT is to just put the IP Address of the Upstream PBR in the Global Administrator field of the RT. In other scenarios, the procedure of the previous paragraph (as modified by this document's sections on "RD usage" and "RT usage") is applied by the PBRs.

### 3. Differences from other MVPN-like GTM Procedures

The document [SEAMLESS-MCAST] also defines a procedure for GTM that is based on the BGP procedures that were developed for MVPN.

However, the GTM procedures of [SEAMLESS-MCAST] are different than and are NOT interoperable with the procedures defined in this document.

The two sets of procedures can co-exist in the same network, as long as they are not applied to the same multicast flows or to the same ASM multicast group addresses.

Some of the major differences between the two sets of procedures are the following;

- The [SEAMLESS-MCAST] procedures for GTM do not use C-multicast Shared Tree Joins or C-multicast Source Tree Joins at all. The procedures of this document use these C-multicast routes for GTM, setting the RD field of the NLRI to zero.
- The [SEAMLESS-MCAST] procedures for GTM use Leaf A-D routes instead of C-multicast Shared/Source Tree Join routes. Leaf A-D routes used in that manner can be distinguished from Leaf A-D routes used as specified in [RFC6514] by means of the NLRI format; [SEAMLESS-MCAST] defines a new NLRI format for Leaf A-D routes. Whether a given Leaf A-D route is being used according to the [SEAMLESS-MCAST] procedures or not can be determined from its NLRI. (See [SEAMLESS-MCAST] section "Leaf A-D Route for Global Table Multicast".)
- The Leaf A-D routes used by the current document contain an NLRI that is in the format defined in [RFC6514], NOT in the format as defined in [SEAMLESS-MCAST]. The procedures assumed by this document for originating and processing Leaf A-D routes are as specified in [RFC6514], NOT as specified in [SEAMLESS-MCAST].
- The current document uses an RD value of zero in the NLRI in order to indicate that a particular route is "about" a Global Table Multicast, rather than a VPN multicast. No other semantics are inferred from the fact that RD is zero. [SEAMLESS-MCAST] uses two different RD values in its GTM procedures, with semantic differences that depend upon the RD values.
- In order for both sets of procedures to co-exist in the same network, the PBRs MUST be provisioned so that for any given IP group address in the global table, all egress PBRs use the same set of procedures for that group address (i.e., for group G,

either all egress PBRs use the GTM procedures of this document or all egress PBRs use the GTM procedures of [SEAMLESS-MCAST].

#### 4. IANA Considerations

This document has no IANA considerations.

#### 5. Security Considerations

The security considerations of this document are primarily the security considerations of the base protocols, as discussed in [RFC6514], [RFC4601], and [RFC5294].

This document makes use of a BGP SAFI (MCAST-VPN routes) that was originally designed for use in VPN contexts only. It also makes use of various BGP path attributes and extended communities (VRF Route Import Extended Community, Source AS Extended Community, Route Target Extended Community) that were originally intended for use in VPN contexts. If these routes and/or attributes leak out into "the wild", multicast data flows may be distributed in an unintended and/or unauthorized manner.

Internet providers often make extensive use of BGP communities (ie, adding, deleting, modifying communities throughout a network). As such, care should be taken to avoid deleting or modifying the VRF Route Import Extended Community and Source AS Extended Community. Incorrect manipulation of these ECs may result in multicast streams being lost or misrouted.

The procedures of this document require certain BGP routes to carry IP multicast group addresses. Generally such group addresses are only valid within a certain scope. If a BGP route containing a group address is distributed outside the boundaries where the group address is meaningful, unauthorized distribution of multicast data flows may occur.

#### 6. Acknowledgments

The authors would like to thank Rahul Aggarwal, Huajin Jeng, Yakov Rekhter, and Samir Saad for their contributions to this work.

The authors would also like to thank Cui Wang, Wei Meng, and Zhengbin Li for their review and critique of this work.

## 7. Authors' Addresses

Lenny Giuliano  
Juniper Networks  
2251 Corporate Park Drive  
Herndon, VA 20171  
US  
Email: lenny@juniper.net

Dante J. Pacella  
Verizon  
Verizon Communications  
22001 Loudoun County Parkway  
Ashburn, VA 20147  
US  
Email: dante.j.pacella@verizonbusiness.com

Eric C. Rosen  
Cisco Systems, Inc.  
1414 Massachusetts Avenue  
Boxborough, MA, 01719  
US  
Email: erosen@cisco.com

Karthik Subramanian  
Cisco Systems, Inc.  
170 Tasman Drive  
San Jose, CA, 95134  
US  
Email: kartsubr@cisco.com

Jeffrey Zhang  
Juniper Networks  
10 Technology Park Dr.  
Westford, MA 01886  
US  
Email: zzhang@juniper.net

Jason Schiller  
Google  
1818 Library Street  
Suite 400  
Reston, VA 20190  
US  
Email: jschiller@google.com

## 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.
- [RFC4364], Rosen, E. and Y. Rekhter, "BGP/MPLS IP Virtual Private Networks", RFC 4364, February 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.
- [RFC6515] Aggarwal, R., and E. Rosen, "IPv4 and IPv6 Infrastructure Addresses in BGP Updates for Multicast VPN", RFC 6515, February 2012.

### 8.2. Informative References

- [ADD-PATH] "Advertisement of Multiple Paths in BGP", D. Walton, A. Retana, E. Chen, J. Scudder, draft-ietf-idr-add-paths-09.txt, October 2013.
- [RFC6368] Marques, P., Raszuk, R., Patel, K., Kumaki, K., and T. Yamagata, "Internal BGP as the Provider/Customer Edge Protocol for BGP/MPLS IP Virtual Private Networks (VPNs)", RFC 6368, September 2011.
- [RFC4601] Fenner, B., Handley, M., Holbrook, H., and I. Kouvelas, "Protocol Independent Multicast - Sparse Mode (PIM-SM): Protocol Specification (Revised)", RFC 4601, August 2006.
- [RFC4684] P. Marques, et. al., "Constrained Route Distribution for Border Gateway Protocol/MultiProtocol Label Switching (BGP/MPLS)

Internet Protocol (IP) Virtual Private Networks (VPNs)", RFC 4684, November 2006.

[RFC5294] Savola, P. and J. Lingard, "Host Threats to Protocol Independent Multicast (PIM)", RFC 5294, August 2008.

[MVPN-extranet] Rekhter, Y. and E. Rosen (editors), "Extranet Multicast in BGP/IP MPLS VPNs", draft-ietf-l3vpn-mvpn-extranet-02.txt, August 2013

[SEAMLESS-MCAST] Rekhter, Y., Aggarwal, R., Morin, T., Grosclaude, I., Leymann, N., and S. Saad, "Inter-Area P2MP Segmented LSPs", draft-ietf-mpls-seamless-mcast-09.txt, December 2013