INTERNET-DRAFT
Intended Status: Informational
Expires: April 22, 2013

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

October 22, 2012

BGP L3VPN Virtual PE Framework
draft-fang-l3vpn-virtual-pe-framework-01

Abstract

This document describes a framework for BGP/MPLS L3VPN with virtual PE solutions. It provides functional description of the control plane and data plane of the virtual PE solutions. It also describes interactions among the vPE solutions and other network elements. The virtual PE solutions support further control plane and forwarding plane separation when compared with traditional L3VPN PE solutions. It allows the L3VPN functions to be extended to application end devices for large scale and efficient IP application support.

Status of this Memo

Copyright and License Notice

Table of Contents

1 Introduction

Network virtualization is to provide multiple individual network services through shared common network resources. Network virtualization is not a new concept. For example, BGP/MPLS layer 3 Virtual Private Networks (L3VPNs) [RFC4364] have been widely deployed to provide network based virtual private network services. It provides routing isolation and forwarding separation for individual VPNs, allow IP address overlapping among different VPNs while forwarding traffic over common network infrastructure.

Network virtualization enables the support of multiple isolated individual networks over a common network infrastructure. Network virtualization is not a new concept. For example, BGP/MPLS IP Virtual Private Network (IP VPNs) [RFC4364] have been widely deployed to provide network based, service provider provisioned IP VPNs for multiple customers with overlapping IP address spaces over a common service provider IP/MPLS network. BGP/MPLS IPVPNs provide routing isolation among customers and allow address overlapping among different VPNs by having per-customer Virtual Routing and Forwarding Instance (VRF) at a service provided Edge (PE), while forwarding customer traffic over a common IP/MPLS network infrastructure.

With the advent of compute capabilities and the proliferation of virtualization in end devices for systems and applications, PE functionality virtualization on such end device is becoming feasible, and in some cases attractive for scale and efficiency. Scale and efficiency are crutial factors in the cloud computing environment supporting various applications and services, and in traditional service provider space.

The virtual Provider Edge (vPE) solution described in this document is to extend the functionality of BGP/MPLS L3VPN to the application end device.

1.1 Terminology

The key words "MUST", "MUST NOT", "REQUIRED", "SHALL", "SHALL NOT", "SHOULD", "SHOULD NOT", "RECOMMENDED", "MAY", and "OPTIONAL" in this document are to be interpreted as described in RFC 2119 [RFC2119].

Term	Definition
3GPP	3rd Generation Partnership Project (3GPP)
AS	Autonomous Systems
ASBR	Autonomous Systems Border Router

BGP	Border Gateway Protocol
ED	End device: where Guest OS, Host OS/Hypervisor, applications, VMs, and virtual router may reside
Forwarder	L3VPN forwarding function
GRE	Generic Routing Encapsulation
IaaS	Infrastructure as a Service
IRS	Interface to Routing System
LTE	Long Term Evolution
MP-BGP	Multi-Protocol Border Gateway Protocol
PCEF	Policy Charging and Enforcement Function
P	Provider backbone router
RR	Route Reflector
RT	Route Target
RTC	RT Constraint
ToR	Top-of-Rack switch
VM	Virtual Machine
Hypervisor	Virtual Machine Manager
VM	Virtual Machine
SDN	Software Defined Network
VI	Virtual Interface
vCE	virtual Customer Router
vPC	virtual Private Cloud
vPE	virtual Provider Edge
VPN	Virtual Private Network
vRR	virtual Route Reflector1.2 Scope of the document
WAN	Wide Area Network

Virtual PE is a PE resides in an end device (e.g., a server) along with client/application VMs.

Through out this document, the term virtual PE (vPE) is used to denote BGP/MPLS L3VPN virtual Provider Edge router.

1.2 Motivation

The recent rapid adoption of Cloud Services by enterprises and the phenomenal growth of mobile IP applications accelerate the needs to extend the L3VPN capability to the end devices. For example, Enterprise customers requested Service Providers to extend and integrate their L3VPN services available in the WAN into the new Cloud services; large enterprise have existing L3VPN deployment are extending them into their data centers; mobile providers are adopting L3VPN into their 3GPP Mobile infrastructure are looking to extend the L3VPNs to their end device of their call processing center.

The virtual PE solution described in this document is aimed to meet the following key requirement [I-D.fang-l3vpn-end-system-req].

1) Support end device multi-tenancy, per tenant routing isolation and traffic separation.

2) Support large scale L3VPNs in service network, upto tens of thousands of end devices and Millions of VMs in the single service network, e.g., a data center.

3) Support end-to-end L3VPN connectivity, e.g. L3VPN can start from a service network end device, connect to a corresponding L3VPN in the WAN, and terminate in another service network end device.

4) Decoupling control plane and forwarding for network virtualization and abstraction.

L3VPN is the proven technologies which is capable of providing routing and forwarding separation, and it is proven with large scale deployment (e.g. supporting 7-8 million L3VPN routes in single Service Provider networks today).

By extending L3VPN solution to the end device with vPE solution, application end-to-end (VM to VM, applications to the end user) L3VPN connectivity cab be achieved, and well as the true network virtualization and abstraction.

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

1.3 Scope of the document

It is assumed that the readers are familiar with BGP/MPLS IP VPN [RFC4364] terms and technologies, the base technology and its operation are not reviewed in details in this document.

The following network elements are discussed in this document: the concept of BGP L3VPN vPE; the interaction of vPE with other network elements, including BGP L3VPN physical PE, physical or virtual BGP Route Reflectors (RR, vRR), and Autonomous System Border Router (ASBR), Service Network gateway routers, external controllers, provisioning/orchestration systems, and the vPE inter-connections with other non L3VPN networks.

The definitions of protocols extensions are out of the scope of this document.

2. Virtual PE Architecture and Reference Model

2.1 Virtual PE

As defined in [RFC4364], a L3VPN is created by applying policies to form a subset of sites among all sites connected the backbone network. It is collection of "sites". A site can be considered as a set of IP systems maintain IP inter-connectivity without connecting through the backbone. The typical use of L3VPM has been to inter-connect different sites of an Enterprise networks through Service Provider’s L3VPNs in the WAN.

A virtual PE (vPE) is a PE instance which resides in one or more physical devices, it is commonly placed in a network service (e.g. a Data Center) end device (e.g., a Server) where the client/application VMs are hosted. The control and forwarding components of the vPE are decoupled, they may reside in the same physical device or in different physical devices.

In the case that a vPE is in a Data Center server along with client/application VMs, one can view the vPE to VM relationship as a typical PE-CE relationship. Unlike a regular physical PE, vPE allows L3VPN control plane and forwarding function residing on different physical devices. The full MP-BGP control plane may reside on the end device, or may be external to the end device, e.g., in a BGP L3VPN boarder router (ASBR)/DC gateway router, a Route Reflector (RR), or an external controller.

Virtual PE solution allows the placement of L3VPN termination point right inside the end device (e.g., a server). In this case, the vPE to CE (VM) connection is internal to the device. If both control and forwarding elements are placed on the end device, L3VPN routing and forwarding starts from the end device, the eliminate the needs for additional process in the next hop (e.g., layer2 and layer 3 integration). This approach helps to simplify the operation and improve the routing and forwarding efficiency in large scale deployment.

Another important benefit is that vPE solution allows full control and forwarding decoupling for scale and achieving true network virtualization to allow network abstraction, flexible and dynamic policy control, quick service turn up time and VM mobility support.

2.2 Architecture reference model

Figure 1 illustrate the topology that vPE is reside in the end device where the applications are hosted.

```
                   +---------------------------------------------+
                   |                                             |
                   |                            +-----------+    |
                   |                            |vPE| End   |    |
                   |                            |---+ Device|    |
                   |              +---------+   +-----------+    |
      .------.     |              |Transport|   +-----------+    |
     (        )    | +-------+    | Device  |   |vPE| End   |    |
    (          )   | |Gateway|    +---------+   |---+ Device|    |
   :            :--+-|  PE   |                  +-----------+    |
   :            :  | +-------+    +---------+   +-----------+    |
   :   IP/MPLS  :  |              |Transport|   |vPE| End   |    |
   :    WAN     :  | +-------+    | Device  |   |---+ Device|    |
   :            :--+-|Gateway|    +---------+   +-----------+    |
    (          )   | |  PE   |                  +-----------+    |
     (        )    | +-------+    +---------+   |vPE| End   |    |
      ’------’     |              |Transport|   |---+ Device|    |
                   |              | Device  |   +-----------+    |
                   |              +---------+   +-----------+    |
                   |                            |vPE| End   |    |
                   |                            |---+ Device|    |
                   |                            +-----------+    |
                   | Virtualized Service Network                 |
                   +---------------------------------------------+
```

Figure 1. Virtualized Service network with vPE

The Virtualized Service Network in Figure 1 consists of WAN gateway PE devices, transport devices, and end devices. In some networks, it is feasible the VPN Gateways may be implemented as vPEs as well.

Examples of service network may be a network that supports cloud computing services, mobile call centers, and SP or enterprise data centers.

Note that the transport devices in the service network in the diagram do not participate L3VPNs, they function similar as P routers in MPLS back bone, they do not maintain the L3VPN states, and are not L3VPN aware.

```
             +----------------------------------------------------+
             | +---------+ +---------+    +---------+ +---------+ |
             | |  VM1    | |  VM2    |    |  VM47   | |  VM48   | |
             | |(VPN Red)| |(VPN Grn)|... |(VPN Grn)| |(VPN Blu)| |
             | +----+----+ +---+-----+    +----+----+ +----+----+ |
             |      |          |               |           |      |
             |      +---+      | +-------------+       +---+      |
             |          |      | |                     |          |
      to     |      +---+------+-+---------------------+---+      |
      Gateway|      |   |      | |                     |   |      |
      PE     |      | +-+-+   ++-++            +---+ +-+-+ |      |
             |      | |VRF|   |VRF|   .......  |VRF| |VRF| |      |
      <------+------+ |Red|   |Grn|            |Yel| |Blu| |      |
             |      | +---+   +---+            +---+ +---+ |      |
             |      |            L3 VPN virtual PE         |      |
             |      +--------------------------------------+      |
             |                                                    |
             |                     End Device                     |
             +----------------------------------------------------+
```

Figure 2. VM in end device to VRF in vPE mapping

An end device shown in Figure 2 is a virtualized server or system which hosts multiple VMs, the virtual PE resides in the end device. The vPE supports multiple VRFs, VRF Red, VRF Grn, VRF Yel, VRF Blu, etc. Each client or 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 RFC Grn, etc. Routing isolation applies between VPNs for multi-tenancy support. For example, VM1 and VM2 can not communicate with each other 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 L3VPNs.

```
               +----------------------+   +----------------------------+
               |                      |   |              +-----------+ |
   +------+--+ | +-----+      +-----+ |   | +-----+      |      +---+| |
   |VPN   +--+ | |     |      |     | |   | |+---+|      +---+  |VM1|| |
   |Red   |CE|-+-|+---+|      |+---+| |   | ||VRF||      |vPE|  +---+| |
   |Site A+--+ | ||VRF||      ||VRF|| |   | ||Red||      +---+  |VM2|| |
   +------+--+ | ||Red||      ||Red|| |   | |+---+|      |Server+---+| |
               | |+---+|      |+---+|-+---+-|+---+|      +-----------+ |
               | |+---+|      |+---+| |   | ||VRF||      +-----------+ |
   +------+--+ | ||VRF||      ||VRF|| |   | ||Grn||      |      +---+| |
   |VPN   +--+-+-||Grn||      ||Grn|| |   | |+---+|      +---+  |VM1|| |
   |Grn   |CE| | |+---+|      |+---+| |   | |GWay |      |vPE|  +---+| |
   |Site A+--+ | | PE  |      | PE  | |   | | PE  |      +---+  |VM2|| |
   +------+--+ | +-----+      +-----+ |   | +-----+      |Server+---+| |
               |                      |   |              +-----------+ |
               |      IP/MPLS WAN     |   |Virtualized Data Center     |
               +----------------------+   +----------------------------+
```

Figure 3. Connecting Enterprise CE to DC VM over WAN

The example of connection from an Enterprise site to application VMs through vPE on the end device of a SP provisioned virtualized data center.

There are multiple options for VPN control plane signaling between the Gateway PE to vPE on the server within the data center. It can use MP-BGP as in regular L3VPN, or use other extensible IP messaging protocols defined in IETF, or use controller direct signaling as a SDN approach.

The inter-connection from DC Gateway PE to MPLS WAN may use one of the Inter-AS options if they are in different ASes. Option B may be more practical for the reasons it is more scalable than Option A, and more restricted than Option C. Consider route aggregation with Option B if both sides have large number of routes.

The connection between backbone VPN to VPN CE on the left hand side is regular L3VPN connection, e-BGP, or static, or other protocols can be used.

3. Control Plane

3.1 vPE Control Plane

The vPE control plane can be distributed or centralized.

1) Distributed control plane

vPE participates in underlay routing through IGP protocols: ISIS or OSPF.

vPE participates in overlay L3VPN control protocol: MP-BGP [RFC4364].

While MP-BGP is the de facto preferred choice between vPE and gateway-PE, using extensible signaling messaging protocols can be alternative, such technologies have been proposed for this segment of signaling [I-D.ietf-l3vpn-end-system].

2. Centralized routing controller

This is a SDN approach. In the virtual PE implementation, not only the service network infrastructure and the VPN overlay networks are decoupled, but also the vPE control plane and data plane are physically decoupled. The control plane directing the data flow may reside elsewhere, such a centralized controller. This requires standard interface to routing system (IRS). The Interface to Routing System (IRS) is work in progress in IETF [I-D.ward-irs-framework], [I-D.rfernando-irs-fw-req].

3.1 Route server of vPE

A virtual PE consist the control plane element and the forwarding plane element. Since the proposed solution decoupled the two element, they may or may not reside on the same physical device.

The Route Server of L3VPN vPE is a software application that implements the BGP/MPLS L3VPN PE control plane functionality.

In the case other control/signaling/messaging protocol are used, the route server is also the server of the particular protocol(s), it interacts with VPN forwarder.

3.3 Use of router reflector

Modern service networks can be very large in scale. For example, the number of VPNs routes in a very large data centers can pass the scale of those in SP backbone VPN networks. There are may be tens of thousands of end devices in a single service network.

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

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

3.4 Use of RT constraint

The Route Target Constraint (RT Constraint, RTC) [RFC4684] is a powerful tool for VPN selective L3VPN route distribution. With RT Constraint, only the BGP receiver (e.g, PE/vPE/RR/vRR/ASBRs, etc.) with the particular L3VPNs will receive the route update for the corresponding VPNs. It is critical to use RT constraint to support large scale L3VPN development.

4. Forwarding Plane

4.1 Virtual Interface

Virtual Interface (VI) is an interface in an end device which is used for connecting the vPE to the application VMs in the end device. The latter cab be treated as CEs in the regular L3VPN's view.

4.2 VPN forwarder

VPN Forwarder is the forwarding component of a vPE.

The VPN forwarder location options:

1) within the end device where the virtual interface and application VMs are.

2) in an external device which the end device connect to, for example, a Top of the Rack (ToR) in a data center.

Multiple factors should be considered for the location of the VPN forwarder, including device capability, overall solution economics, QoS/firewall/NAT placement, optimal forwarding, latency and performance, operation impact, etc. There are design trade offs, it is worth the effort to study the traffic pattern and forwarding looking trend in your own unique service network as part of the exercise.

4.3 Encapsulation

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

1. MPLS Encapsulation, [RFC3032].

2. Encapsulating MPLS in IP or Generic Routing Encapsulation (GRE), [RFC4023].

The most common BGP/MPLS L3VPNs deployment in SP networks are using MPLS forwarding. This requires MPLS, 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 network against attacks.

However, the service network environment, such as a data center, is different than Service Provider VPN networks or large enterprise backbones. MPLS deployment may or may not be feasible. Two major challenges for MPLS deployment in this new environment: 1) the capabilities of the end devices and the transport/forwarding devices; 2) the workforce skill set.

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

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

4.4 Optimal forwarding

As reported by many large cloud service operators, the traffic pattern in their data centers were dominated by East-West across subnet traffic (between the end device hosting different applications in different subnets) than North-South traffic (going in and out the DC to the WAN) or switched traffic within subnets. This is a primary reason that many large scale new design has moved away from traditional L2 design to L3.

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

When multiple VPNs need to be accessed to accomplish the task the user requested (this is common too), the end device virtual interfaces should be able to directly access multiple VPNs via use of extranet VPN techniques without the need of Gateway facilitation. Use BGP L3VPN policy control mechanisms to support this function.

5. Addressing

5.1 IPv4 and IPv6 support

Both IPv4 and IPv6 should be supported in the virtual PE solution.

This may present challenging to older devices, but may not be issues to newer forwarding devices and servers. A server is replaced much more frequently than a network router/switch in the infrastructure network, newer equipment should be capable of IPv6 support.

5.2 Address space separation

The addresses used for L3VPNs in the service network should be in separate address blocks than the ones used the underlay infrastructure of the service network. This practice is to protect the service network infrastructure being attacked if the attacker gain access of the tenant VPNs.

Similarity, the addresses used for the service network, e.g., a cloud service center of a SP, should be separated from the WAN backbone addresses space, for security reasons.

6.0 Inter-connection considerations

There are also deployment scenarios that L3VPN may not be supported in every segment of the networks to provide end-to-end L3VPN connectivity, a L3VPN vPE may be reachable only via an intermediate inter-connecting network, interconnection may be needed in these cases.

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

From L3VPN point of view: A L3VPN vPE that implements [RFC4364] is a component of L3VPN network only. A L3VPN VRF on physical PE or vPE contains IP routes only, including routes learnt over the locally attached network.

As described earlier in this document, the L3VPN vPE should ideally be located as close to the "customer" edge devices. For cases, where this is not possible, simple existing "L3VPN CE connectivity" mechanisms should be used, such as static, or direct VM attachments such as described in the vCE 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 best suited BGP solution is to use L3 VPNs [RFC4364] for all sites with PE and/or vPE solutions. This is a straightforward case.

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

One can consider to use combined vPE and vCE solution to solved the problem. Use L3VPN for all sites with IP connectivity, and use a physical or virtual CE (vCE, may reside on the end device) to aggregate the L2 sites which, for example, are in a single container in a data center. The CE/vCE can be considered as inter-connecting point, where the L2 network are terminated and the corresponding routes for connectivity of the L2 network are inserted into L3VPN VRF. The L2 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. Security Considerations

vPE solution presented a virtualized L3VPN PE model. There are potential implications to L3VPN control plane, forwarding plane, and management plane. Security considerations are currently under study, will be included in the future revisions.

8. IANA Considerations

None.

9. References

9.1 Normative References

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

[RFC3032] Rosen, E., Tappan, D., Fedorkow, G., Rekhter, Y., Farinacci, D., Li, T., and A. Conta, "MPLS Label Stack Encoding", RFC 3032, January 2001.

[RFC4023] Worster, T., Rekhter, Y., and E. Rosen, Ed., "Encapsulating MPLS in IP or Generic Routing Encapsulation (GRE)", RFC 4023, March 2005.

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

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

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

[RFC5036] Andersson, L., Ed., Minei, I., Ed., and B. Thomas, Ed., "LDP Specification", RFC 5036, October 2007.

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

9.2 Informative References

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

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

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

Authors’ Addresses

Luyuan Fang
Cisco
111 Wood Ave. South
Iselin, NJ 08830

Email: lufang@cisco.com

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

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

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

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

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