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/1id-abstracts.html

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

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.

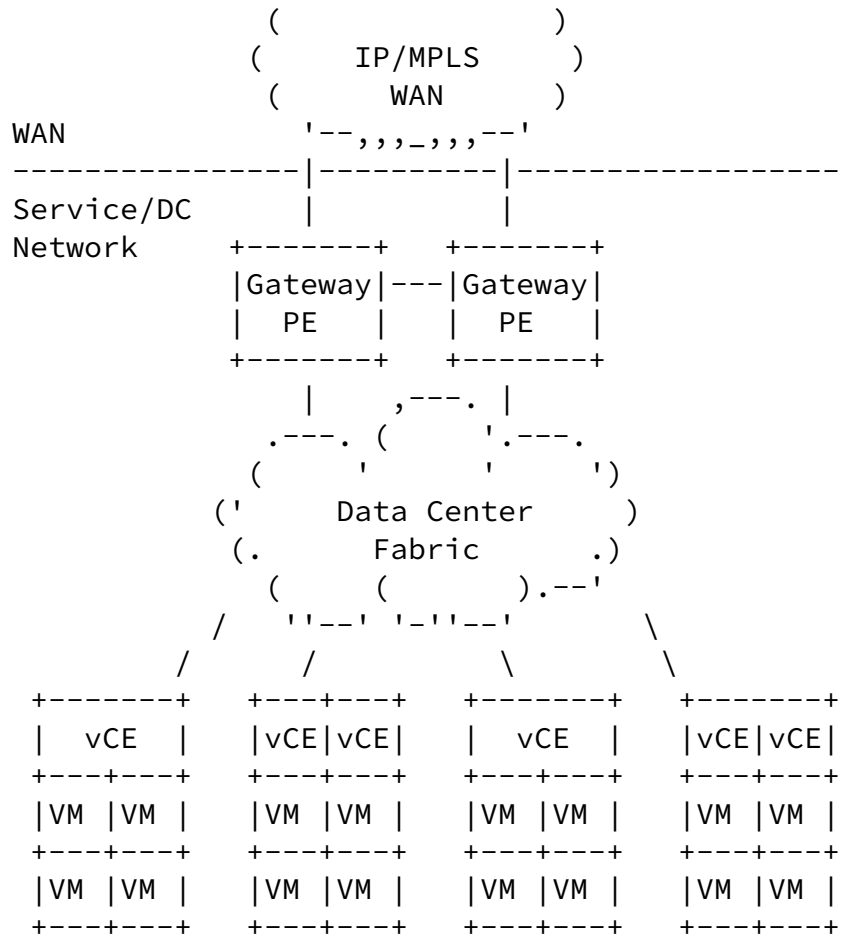

Figure 1. Virtualized Data Center with vCE

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.

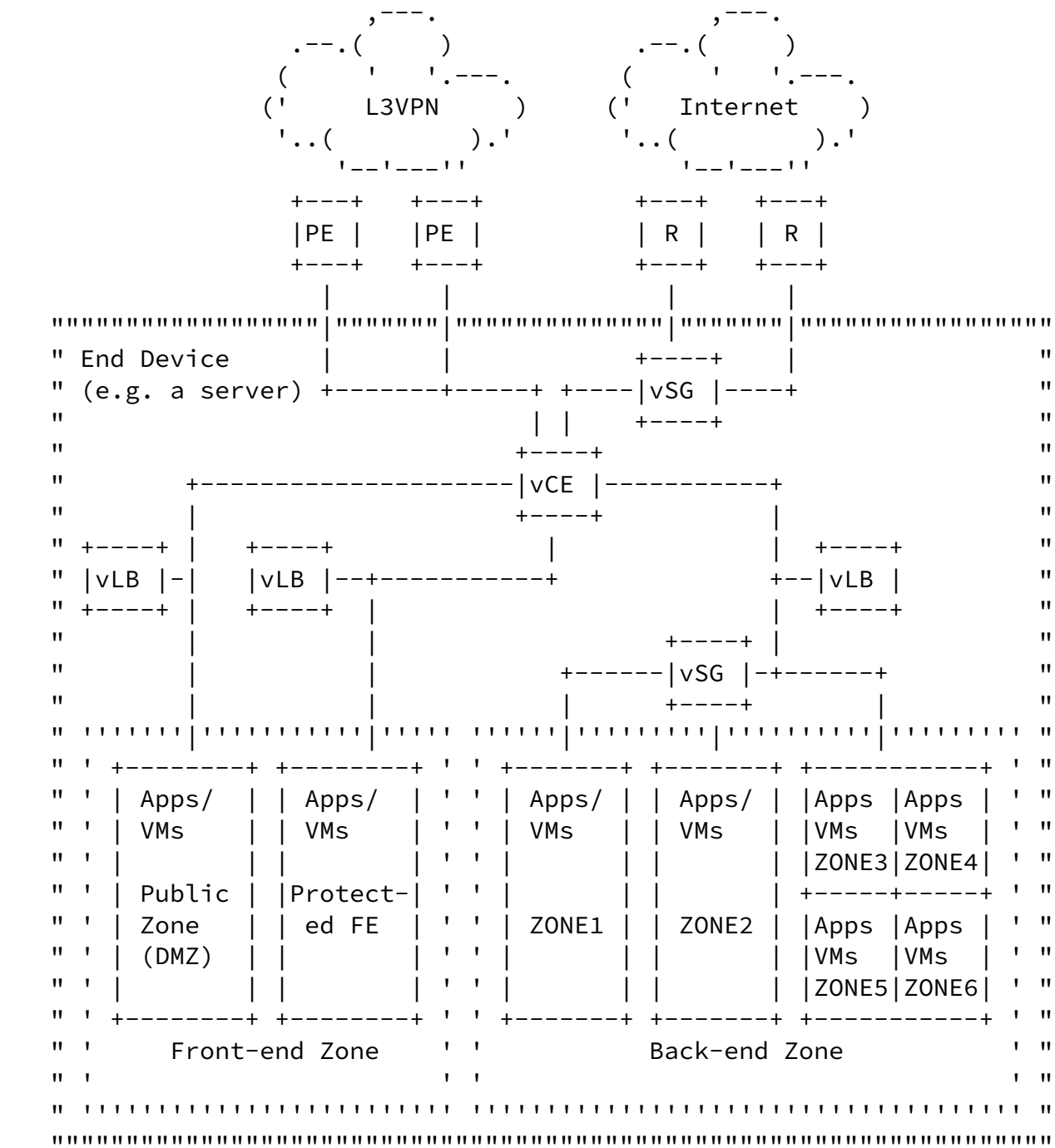

Figure 2. A Virtualized Container with vCE in an End Device

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