INTERNET-DRAFT
Intended Status: Standards track
Expires: August 18, 2013

R. Fernando
D. Rao
L. Fang
Cisco
M. Napierala
AT&T
N. Bitar
Verizon
N. So
Tata Communications

February 18, 2013

Virtual Topologies for Service Chaining in BGP IP VPNs
draft-rfernando-l3vpn-service-chaining-00

Abstract

This document presents the techniques built upon BGP/MPLS IP VPN control plane mechanisms to construct virtual service topologies for service chaining. These virtual service topologies allow a sequence of service nodes to be visited across multiple zones in a data center to form a service chain. The method uses service topology specific Route Targets (RTs) in addition to general purpose RTs.

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

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 support certain applications or a subset thereof. 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, but a datagram traveling between hosts in different zones is subject to additional services to meet the needs of scaling, performance, and security for specific applications. Example of such services can be a security gateway or a load-balancer.

Traditional networks achieve this by using a combination of physical topology constraints and routing. For example, one can force datagrams going through a FireWall (FW) by putting the firewall in the data path from a source to a destination. In some other cases, the datagrams needs to go through a security gateway for security service, and a Load Balancer (LB) for load balancing service.

In the modern virtualized Data Centers, appliances, applications, and network functions are virtualized, they are software instances residing in servers or appliances instead of individual physical devices.

Porting a traditional network with all its functions and infrastructure elements to a virtualized DC requires network overlay mechanisms that provide the ability to create virtual network topologies that mimic physical networks and the ability to constrain the flow of routing and traffic over these virtual network topologies.

A Data Center needs 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 in the traditional network has the resource as Provider Edge (PE) 1, and destination as Autonomous System Border Router (ASBR) 1, the flow must be serviced by FW1 and LB2, its path would be PE1 -> FW1 -> LB1 -> ASBR1. In a virtualized DC, the virtual topology for this path may be vPE1 -> vFW1 -> vLB1 -> ASBR1, assume PE1, FW1 and LB1 are virtual nodes. This sequence represents an example of virtual service chain. The nodes in the chain may be placed at arbitrary physical locations.

Furthermore, data centers might need multiple virtual topologies per tenant to handle different types of application traffic. A tenant is a customer who uses the virtualized data center services. The term Multi-tenant means virtualized single end device, for example, a server, supports multiple tenants which requires routing isolation among the tenants' traffic. Each tenant might dictate a different topology of connectedness between their zones and applications and might need the ability to apply network policies and services for inter-zone traffic in specific order to the organization objectives of the tenant. Therefore, the mechanisms devised should be flexible to accommodate the custom needs of a tenant and their applications at the same time MUST be robust enough to satisfy the scale, performance and HA needs that they demand from the virtual network infrastructure.

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

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].

2. Suggested text under terminology in 1.1 (after the key word paragraph)

Terms description

AS	Autonomous System
ASBR	Autonomous System Border Router
BGP	Border Gateway Protocol
CE	Customer Edge
ED	End device: where Guest OS, Host OS/Hypervisor, applications, VMs, and virtual router may reside
Forwarder	L3VPN forwarding function
FW	FireWall
GRE	Generic Routing Encapsulation
Hypervisor	Virtual Machine Manager running on each end device
I2RS	Interface to Routing System
LB	Load Balancer
LTE	Long Term Evolution
MP-BGP	Multi-Protocol Border Gateway Protocol
PCEF	Policy Charging and Enforcement Function
P	Provider backbone router
proxy-arp	proxy-Address Resolution Protocol

QoS	Quality of Service
RR	Route Reflector
RT	Route Target
RTC	RT Constraint
SDN	Software Defined Network
ToR	Top-of-Rack switch
VI	Virtual Interface
vCE	virtual Customer Router
vFW	virtual FireWall
vLB	virtual Load Balancer
VM	Virtual Machine
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

General terminologies:

Service-PE: A BGP IP-VPN PE to which a service node in a virtual service topology is attached. The PE directs incoming traffic from other PEs to the service node via an MPLS VPN label or IP lookup; and

forwards traffic from the service node to the next node in the chain. A Service-PE is a logical entity, in that a given PE may be attached to both a service node and an application VM.

Service node: A physical or virtual service appliance/application which inspects and/or redirects the flow of inter-zone traffic. Examples of service CEs: Firewalls, load-balancers, deep packet inspectors. The Service node acts as a CE in the VPN network.

Service Chain: A sequence of service-PE's and the corresponding service nodes created in a specific order. The service chain is unidirectional and creates a one way traffic flow between source Service-PE in the source zone and Destination Service-PE in the destination zone.

Service topology route: A topology service route is a route that is used to direct the traffic flow along the service topology. There should be one such route per service topology. The service topology - and hence the service route - is constructed on a per-VPN basis. This service topology is independent of the routes for the actual addresses of the VMs present in the various zones. There can be multiple service topologies for a given VPN. Topologies are constructed unidirectonally. Between the same pair of zones, traffic in opposite directions will be supported by two service topologies.

Source Service-PE: The service-PE closest to the source zone.

Destination Service-PE: The service PE closest to the destination zone.

Service-import-RT: A RT allows the routes to be imported into a Service-VRF. The route which is imported through service-import-RT MUST be re-originated with the corresponding service-export-RT.

Service-export-RT: A RT allows the routes to be exported from a Service-VRF.

Service-topology-RT: identifies the specific service topology.

Tenant. A tenant is a higher-level management construct. In the control/forwarding plane, it is the 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 can communicate freely within a zone.

2 Intra-Zone Routing and Traffic Forwarding

This section provides a brief overview of how BGP/MPLS IP VPNs [RFC4364] control plane can be used in DC networks to used to divide a DC 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 <source zone, destination zone, application>.

The notion of 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 an L3 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/VM's are associated with network policies that become activated on an attached PE when a CE/VM becomes alive. These policies dictate how networking should be set up for the CE/VM including the properties of the CE-PE link, the IP address of the CE/VM, the zone(s) that it belongs to, QoS policies etc. There are many ways to accomplish this step, a description of which is outside the scope of this document.

When the CE/VM is activated, the attached PE starts exporting its 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.

Note that the IP address mask of the CE/VM that the PE advertises along with the CE's address need not necessarily be a /32 for IPv4, and /128 for IPv6. This is the case when the CE/VM's in a zone belong to a single IP subnet. The PE, in this case, would use proxy-arp to resolve ARP's for remote destinations in the IP subnet.

Alternatively, there may be a pool of dedicated service appliances which support multiple contexts.

The classification of VMs into a zone is driven by the communication and security policy and is independent of the addressing 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 a locally attached VM. If two VMs are in the same IP subnet, the PE employs proxy-ARP to assist the VM resolve ARP for other VMs in the IP subnet, and uses IP forwarding to carry traffic between the VMs. When a VM is remotely attached to another PE, BGP IP-VPN forwarding is used.

3 Inter-Zone Routing and Traffic Forwarding

A simple form of inter-zone traffic forwarding can be achieved using extranets BGP IP VPN configurations. However, extranet procedures do not by themselves provide the ability to force inter-zone traffic flows through a set of servers.

Note that the inter-zone services cannot always be assumed to reside 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. Alternatively, there may be a pool of dedicated service appliances which support multiple contexts and hence multiple tenants, that are attached to different PEs distributed across the DC.This creates a situation where services for traffic between zones may not be applied 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 and virtualized.

A service node for the purposes of this proposal is a physical or virtual service appliance that inspects and/or impacts the flow of inter-zone traffic. Firewalls, load-balancers, deep packet inspectors are examples of service nodes. Service nodes modeled as CEs, either they are reside on a PE, or they are then attached to a service-PE.

A service-PE is a normal BGP IP VPN PE that recognizes and directs the appropriate traffic flows to its attached service nodes through VPN label lookup. Service nodes may be integrated or attached to service-PEs.

A sequence of service-PE's and the corresponding service nodes create a service chain for inter-zone traffic. The service chain is unidirectional and creates a one way traffic flow between source zone and destination zone. The first service PE in the path is called as the source service-PE and the last service PE in the path is called the destination service-PE, there can be arbitrary numbers of service-PE between the source service-PE and the destination service-PE. An example of a service chain may look like this: ingress-PE --> source-service-PE --> other-service-PE --> destination-service-PE --> egress-PE.

4 Proposed Inter-Zone Model

The proposed model has two steps to it.

4.1 Constructing the Virtual Service Topology

The first step involves creating the virtual service topology that ties two or more zones through one or more service nodes.

This is done by originating a service topology route that creates the route resolution state for the zone prefixes in a set of service-PEs. The service topology route is originated in the destination service-PE. It then propagates through the series of service-PE's from the destination service-PE to the source service-PE.

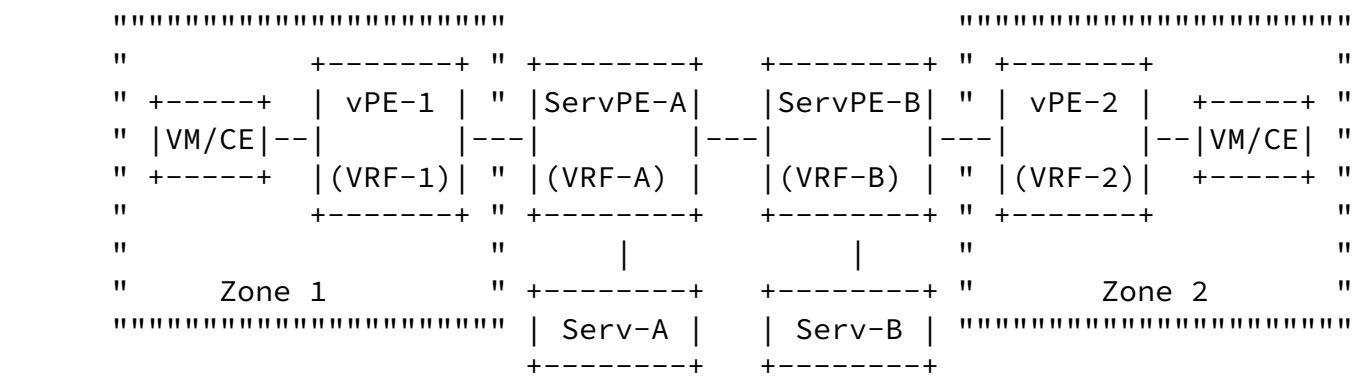

Figure 1. Construct of Service Chain Topology

A modification is proposed to the service-PE behavior to allow the automatic and constrained propagation of service topology routes through the service-PE's that form the service chain. A service-PE in a given service chain is provisioned to accept the service topology route and re-originate it such that the upstream service-PE imports it and so on. The sequential import and export of the service topology route along the service chain is controlled by RTs provisioned appropriately at each service-PE.

To create the service chain and give it a unique identity, each service-PE is provisioned with three service RT's per VRF for every service chain that it belongs to: {service-import-RT, service-export-RT, service-topology-RT}.

A service-import-RT acts exactly as a regular import RT importing any route that carries that RT into the service-VRF. Additionally, any route that was imported using the service-import-RT MUST be automatically re-originated with the corresponding service-export-RT.

The next-hop of the re-originated route points to the service node attached to the service-PE. The VPN label carried in the re-originated route directs all traffic received by the service-PE to the service node.

The service-export-RT of a downstream service-PE MUST be equal to the service-import-RT of the immediate upstream service-PE. The service topology route MUST be originated in the destination service-PE carrying its service-export-RT. The flow of the service topology route creates both the service chain as well as the route resolution state for the zone prefixes.

Finally, the presence of the service topology route in a service-PE triggers the addition of the service-topology-RT to the regular

import RT's of the service-VRF. Every service chain has a single unique service-topology-RT that's provisioned in all participating

service-PE's.

The three service RT's (import, export and topology) are special RTs which should not be reused for other purposes within the network. The service RT's that establish the chain and give it its identity can be pre-provisioned or activated due to the appearance of a attached virtual service node. The provisioning system is assumed to have the intelligence to create loop-free virtual service topologies.

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

Virtual service topologies are constructed unidirectionally. Between the same pair of zones, traffic in opposite directions will be supported by two service topologies and hence two service topology routes. These two service topologies might or might not be symmetrical, i.e. they might or might not traverse the same service-PE's/service-nodes in opposite directions.

As noted above, a service topology route can be advertised with a per-next-hop 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 seamlessly by just configuring the three service RT's in the attached service-PE. This technique could be used to elastically scale out the service nodes with traffic demand.

The distribution of the service topology route itself can be controlled by RT constrains [RFC4684] to only the interesting service-PE's.

Finally, note that the service topology route is independent of the zone prefixes which are the actual addresses of the VMs present in the various zones. The zone prefixes use the service topology route to resolve their next-hop.

4.2 Inter-zone Routing and Service Chaining

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 VM's of the same zone.

In addition to the intra-zone RT's, 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 service-topology-RT, zone prefixes get imported into the appropriate service-VRF's of particular service-PE's that form the service chain associated to that topology RT. Note that the topology RT was added to the relevant service-VRF's import RT list during the virtual topology construction phase.

Once the zone prefixes are imported into the service-PE, their next-hops are resolved as follows.

- If the importing service-PE is the destination 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.

- If the importing service-PE is not the destination service-PE, it rewrites the received next-hop of the zone prefix with the service topology route.

In an MPLS VPN, the zone prefixes come with VPN labels. The labels also must be ignored when in the intermediate service-PEs. Instead, the zone prefix gets resolved via the service topology route and uses the associated service route's VPN label.

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

Traffic for the zone prefix goes through the service hops created by the service topology route. At each service hop, the service-PE

directs the traffic to the service node. Once the service node is done processing the traffic, it then 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 service-PE's. 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.

Note that this proposal introduces a change in the behavior of the service-PE's but does not require protocol changes to BGP.

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 state created by the service topology route. A service topology route can have multiple equal cost paths, for inter-zone traffic to get load-balanced over.

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 topology 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-PE's that comprise the service chain.

6 Security Considerations

This proposal does not change the security model of MPLS/VPN BGP.

7 IANA Considerations

This proposal does not have any IANA implications.

8 Acknowledgements

The authors would like to thank the following individuals for their review and feedback on the proposal: Eric Rosen, Jim Guichard, Paul Quinn, David Ward, Ashok Ganesan.

9 References

9.1 Normative References

[RFC4364] Rosen, E., "BGP/MPLS IP Virtual Private Networks (VPNs)", RFC4364.

[RFC4684] Marques, P., "Constrained Route Distribution for Border Gateway Protocol/Multiprotocol Label Switching (BGP/MPLS) Internet Protocol (IP) Virtual Private Networks (VPNs)

Authors' Addresses

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

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

Luyuan Fang
Cisco
170 W Tasman Dr
San Jose, CA
US
Email: lufang@cisco.com

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

US
Email: mnapierala@att.com

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

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