INTERNET-DRAFT
Intended Status: Standards track
Expires: January 14, 2014

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

July 15, 2013

Virtual Topologies for Service Chaining in BGP IP VPNs
draft-rfernando-l3vpn-service-chaining-02

Abstract

This document presents techniques built upon BGP 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 interesting service functions can be applied to such trafic.

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

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 modern virtualized Data Centers, appliances, applications, and network functions, including IP VPN PE and CE functions are commonly virtualized, i.e, 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 data center 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.

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-VPN procedures and functionality at the PEs, or via an SDN based approach. This document specifies the use of both approaches, but uses the IP-VPN based option
to illustrate the various steps involved.

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
DPI	Deep Packet imspection
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

PBR	Policy Based Routing
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 or from attached hosts 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 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 CEs: Firewalls, load-balancers, deep packet inspectors. The Service node acts as a CE in the VPN network.

Service Chain: A sequence of service nodes that interconnect two end-host zones.
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-PE's 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
forming 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 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 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.

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 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. 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 inlined 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 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.

3.1 Traffic Forwarding operational flow

Traffic from an endpoint in a source zone lands on an ingress zone-PE in a VRF associated with the zone. The zone-PE will forward the traffic and direct it towards the first service-node.If the service-node is attached to the zone-

PE,
it will forward the packet out one of its access interfaces. If the service-node
is attached to a different service-PE, it will encapsulate the packets
appropriately and send them towards the service-PE. The PEs may be physically
connected via an intermediate network of devices.

The service-PE will receive these encapsulated packets from the source zone-PE
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 now be forwarded directly to the destination in the target zone. The
destination may be attached or reachable via another PE.

As can be determined by the above example, a given packet flow needs to be
forwarded differently at any PE depending on whether the traffic is destined
towards an attached node on the PE or is arriving from an attached node and
destined to a node at a remote PE. The next-hop for the flows changes depending
on the relative position within the logical service chain.

The following figure 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.

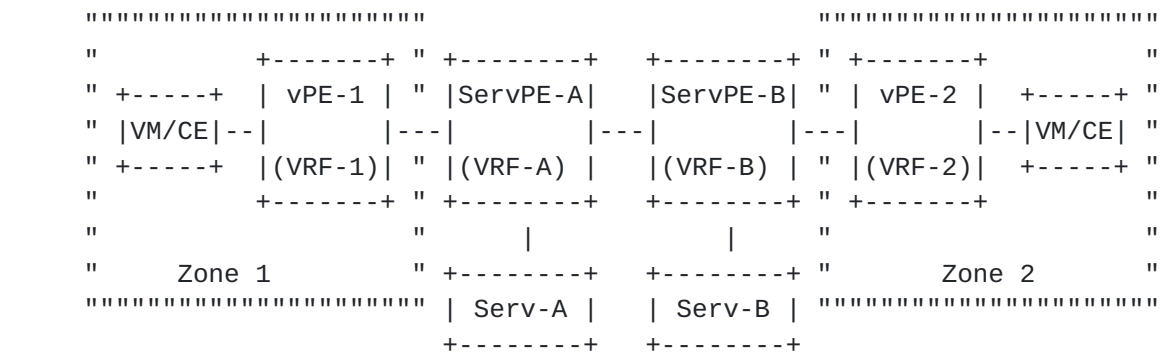

Figure 1. Virtual Service Topology illustration

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

Each service node is associated with two VRF tables at the service PE that it is

attached to - an in-VRF for traffic towards the service node, and an out-VRF for
traffic from the service node.

Traffic in the in-VRF arrives from the previous node in the service chain, and
traffic in the out-VRF is destined towards the next node in the service chain,
or towards the destination zone.

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 appropriate forwarding entries to implement the
forwarding flow described above may be achieved either via IP-VPN mechanisms
or via an SDN approach, as described further.

4. Proposed Inter-Zone Model

The proposed model has the following steps to it.

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, ie 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's provisioned in all participating PEs.

A PE will also be provisioned with the tables and/or configuration that support the various zone, service in- and out- VRFs.

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 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 zone-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 zone RT was added to the relevant service-VRF's import RT list during the virtual topology construction phase. These routes may be installed in the in-VRF, out-VRF tables at the service-PEs as well as in the ingress zone-VRF.

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

A modification is proposed to a standard PE behavior to allow 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 table. It will use the next-hop and label associated with this IP address to encapsulate traffic towards 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.

This way the zone prefixes in the intermediate service-PE hops recurse over the service chain 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 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 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 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. 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 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 seamlessly into the chain 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 state created by the service topology.

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-PE's 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 then uses a suitable means of communication with the various PEs, typically virtual PEs on the end-servers to populate the forwarding tables.

The controller/Orchestration system used to communicate between PE/vPE MUST support standard, programmatic interface. The programmatic interface are current under definition in IETF Interface to Routing Systems (I2RS)) initiative. [I-D.ward-irs-framework], [I-D.rfernando-irs-fw-req]. Standard data modeling languages will be defined/identified in I2RS. YANG - A Data Modeling Language for the Network Configuration Protocol (NETCONF) [RFC6020] is one of the candidates currently under investigation.

7. Security Considerations

To be added.

8. IANA Considerations

This proposal does not have any IANA implications.

9. 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.

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.

[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.

[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.

10.2 Informative References

[RFC4627] Crockford, D., "The application/json Media Type for JavaScript Object Notation (JSON)", RFC 4627, July 2006.

[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-01, Feb. 2013.

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

[I-D.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

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

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