INTERNET-DRAFT
Intended Status: Standards track
Expires: January 5, 2015

R. Fernando
D. Rao
Cisco
L. Fang
Microsoft
M. Napierala
AT&T
N. So
Vinci Systems
A. Farrel
Juniper Networks

July 4, 2014

Virtual Topologies for Service Chaining in BGP/IP MPLS VPNs

draft-rfernando-l3vpn-service-chaining-04

Abstract

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

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

Status of this Memo

This Internet-Draft is submitted to IETF in full conformance with the provisions of BCP 78 and BCP 79.

Internet-Drafts are working documents of the Internet Engineering Task Force (IETF), its areas, and its working groups. Note that other groups may also distribute working documents as Internet-Drafts.

Internet-Drafts are draft documents valid for a maximum of six months and may be updated, replaced, or obsoleted by other documents at any time. It is inappropriate to use Internet-Drafts as reference material or to cite them other than as "work in progress."

The list of current Internet-Drafts can be accessed at

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

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

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

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

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

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

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

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

To address all of these requirements, the mechanisms devised for use

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

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

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

1.1 Terminology

This document uses the following acronyms and terms.

Terms	Meaning
AS	Autonomous System
ASBR	Autonomous System Border Router
CE	Customer Edge
FW	Firewall
I2RS	Interface to the Routing System
L3VPN	Layer 3 VPN
LB	Load Balancer
NLRI	Network Layer Reachability Information [RFC4271]
P	Provider backbone router

proxy-arp	proxy-Address Resolution Protocol
RR	Route Reflector
RT	Route Target
SDN	Software Defined Network
vCE	virtual Customer Edge router [I-D.fang-l3vpn-virtual-ce]
vFW	virtual Firewall
vLB	virtual Load Balancer
VM	Virtual Machine
vPC	virtual Private Cloud
vPE	virtual Provider Edge router [I-D.fang-l3vpn-virtual-pe]
VPN	Virtual Private Network
VRF	VPN Routing and Forwarding table [RFC4364]
vRR	virtual Route Reflector

This document also uses the following general terms:

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

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

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

Virtual service topology:
A virtual service topology consists of a sequence of service-PEs and their attached service nodes created in a specific order. A

service topology is constructed via one or more routes that direct the traffic flow among the PEs that form the service chain.

Service-topology-RT:
A BGP route attribute that identifies the specific service topology.

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

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

2. Intra-Zone Routing and Traffic Forwarding

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

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

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

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

have exactly one VRF for that zone, and the PE-CE links to those CEs

must all be associated with that VRF. Intra-zone connectivity between CE/VMs that attach to different PEs is achieved by designating an RT per zone (zone-RT) that is both an import RT and an export RT of all PE VRFs that terminate the CE/VMs that belong to the zone. A VM may have multiple virtual interfaces that attach to different zones.

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

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

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

3. Inter-Zone Routing and Traffic Forwarding

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

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

3.1 Traffic Forwarding Operational Flow

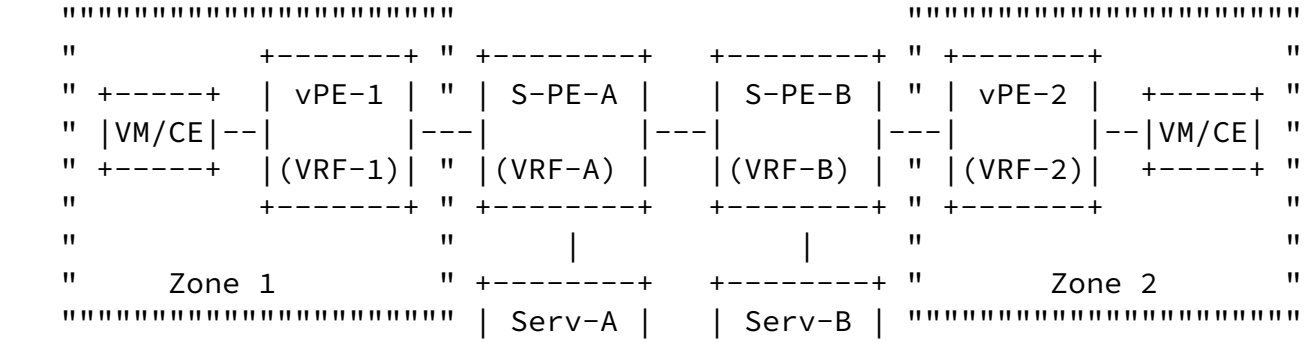
Traffic from a source endpoint (a VM/CE) in a source zone reaches an ingress zone-PE and is associated with a VRF in that zone as described above. The zone-PE will forward the traffic and direct it toward the first service-node. If the service-node is attached to the zone-PE, the zone-PE will forward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

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

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

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

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



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

- Each service node is associated with two VRFs at the service PE to which it is attached: an in-VRF for traffic toward the service node, and an out-VRF for traffic from the service node.

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

- The in-VRF has one or more routes with a next-hop of a local access interface where the service node is attached. The out-VRF has routes with a next-hop of the next service node, which may be situated locally on the service-PE or at a remote PE.

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

4. Inter-Zone Model

The inter-zone model has the following steps.

4.1 Constructing the Virtual Service Topology

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

4.1.1 Reachability to the Service Nodes

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

and out-VRF at the attached service node.

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

A service-PE will export a route for each service node attached to it. Each route will contain the Route-Target configured for the VPN, and a forwarding label that is associated with the logical in-VRF for to directly forward incoming traffic from the other PEs to the

service node.

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

4.1.2 Provisioning the Service Chain

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

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

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

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

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

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

4.1.3 Zone Prefix Next-Hop Resolution

Routes representing hosts, VMs or other destinations associated with 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

that have VRFs for the same zone to import these prefixes to enable direct communication between VMs in the same zone.

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

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

Note that the approach being described introduces a change in the

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

The PE, based on the presence of the Service-topology-RT in the zone routes it receives, 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. In the out-VRF associated with a service node, it will select the next service node in the sequence.

3. It will further resolve this Service next-hop IP address locally in the associated VRF, instead of in the global routing table. It will use the next-hop (and label, if remote) associated with this IP address to encapsulate traffic toward the next service node.

4. If the importing service-PE is the last service-PE, it uses the next hop that came with the zone prefix for route resolution. It also uses the VPN label that came with the prefix.

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

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

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

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

Virtual service topologies are constructed unidirectionally. Traffic in opposite directions between the same pair of zones will be

supported by two different service topologies and hence two service topology routes. These two service topologies might or might not be symmetrical, i.e. they might or might not traverse the same sequence As noted above, a service node route is 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 in the in-VRF at the service-PE to send traffic to the service node.

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

4.2 Per-VM Service Chains

While the service-topology-RT allows an efficient inheritance of the service chain for all VMs or prefixes in a zone, there may be a need to create a distinct service chain for an individual VM or prefix. 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 and service PEs are provisioned with this RT as described above.

5. Routing Considerations

5.1 Multiple Service Topologies

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

5.2 Multipath

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

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

5.3 Supporting Redundancy

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

At a routing level, this is achieved by setting up two paths for the same service 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 a specific PE and the aggregate zone prefix used in the service chain between zones. In such a case, it is the aggregate zone prefix that carries the service-topology-RT and gets imported in the service-PEs that comprise the service chain.

6. Orchestration Driven Approach

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

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

The protocol used to communicate between the controller/orchestration and the PE/vPE must be a standard, programmatic interface. There are several possible options to this programmatic interface, some being under discussion in the IETF's Interface to Routing Systems (I2RS) initiative, [I-D.ietf-i2rs-architecture], [I-D.ietf-i2rs-problem-statement]. One specific option is defined in [IPSE].

7. Security Considerations

To be added.

8. Management Considerations

To be added.

9. IANA Considerations

This proposal does not have any IANA implications.

10. Acknowledgements

The authors would like to thank the following individuals for their review and feedback on the proposal: Eric Rosen, Jim Guichard, Paul

Quinn, Peter, Bosch, David Ward, Ashok Ganesan. The option of configuring an ordered sequence of service nodes via policy is derived from a suggestion from Eric Rosen.

11. References

11.1 Normative References

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

11.2 Informative References

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

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

[RFC7024] Jeng, H., Uttaro, J., Jalil, L., Decraene, B., Rekhter, Y., and R. Aggarwal, "Virtual Hub-and-Spoke in BGP/MPLS VPNs", RFC 7024, October 2013.

[I-D.fang-l3vpn-virtual-ce] L. Fang, et al.,"BGP/MPLS IP VPN Virtual CE", draft-fang-l3vpn-virtual-ce, work in progress.

[I-D.fang-l3vpn-virtual-pe] L. Fang, et al., "BGP/MPLS IP VPN Virtual PE", draft-fang-l3vpn-virtual-pe, work in progress.

[I-D.ietf-i2rs-architecture] Atlas, A., Halpern, J., Hares, S., Ward, D., and T Nadeau, "An Architecture for the Interface to the Routing System", draft-ietf-i2rs-architecture, work in progress.

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

[IPSE]
Fernando, R., Boutros, S., Rao, D., "Interface to a Packet Switching Element", draft-rfernando-ipse-00, work in progress.

Authors' Addresses

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

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

Luyuan Fang
Microsoft
5600 148th Ave NE
Redmond, WA 98052
Email: lufang@microsoft.com

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

Ning So
Vinci Systems, Inc.
Email: ningso@yahoo.com

Adrian Farrel
Juniper Networks
Email: adrian@olddog.co.uk