INTERNET-DRAFT
Intended Status: Informational
Expires: April 25, 2013

R. Fernando
D. Rao
L. Fang
Cisco

October 22, 2012

Virtual Service Topologies in BGP VPNs
draft-rfernando-virt-topo-bgp-vpn-01

Abstract

This document presents techniques that build on MPLS/VPN control plane mechanisms to construct virtual service topologies in data centers. These virtual service topologies interconnect network zones and help to constrain the flow of traffic that go between zones so that interesting services can be applied to them.

The techniques suggested are required to create a rich overlay network that mimics topology and routing functions of physical networks. Steps to create a virtual service topology and the ability to constrain routing and traffic to flow in this topology are outlined in this document.

Status of this Memo

Table of Contents

1 Introduction

Network topologies and routing in the enterprise, data center and campus networks reflect the needs of the organization in terms of performance, scale, security and availability. For scale and security reasons, networks are composed of multiple small domains or zones each serving one or more logical functions of the organization.

Hosts within a zone can freely communicate with one another but traffic between hosts in different zones is subjected to additional services that help in scaling and securing the end applications. Traditional networks achieve this using a combination of physical topology constraints and routing.

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.

Furthermore, data centers might need multiple virtual topologies per tenant to handle different types of application 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 manner specific to their organizational objectives. Therefore, the mechanisms devised should be flexible to accommodate the custom needs of a tenant and their applications at the same time 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 Intra-Zone Routing and Traffic Forwarding

This section provides a brief overview of how L3VPNs [RFC4364] can be used in data center networks to create a single zone to host customer

applications. The subsequent sections in the document builds on this base model to create richer topologies by interconnecting these zones and enforcing services for inter-zone traffic.

In a DC, servers host virtual machines where end applications reside. A collection of VMs that can communicate freely form a zone.

The notions of L3VPN when applied to the virtual data center works in the following manner.

The VM that runs the applications that is the CE. A CE/VM belongs to a zone. As in traditional L3VPN, 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).

The PE helps create the zone that the CE belongs to by placing the CE-PE link in a VRF corresponding to that zone. Intra-zone connectivity is achieved by designating an RT per zone (zone-RT) that is applied on all PE VRFs that terminate the CE/VMs that belong to the zone.

It is further assumed that the CE/VM’s are associated with network policies that get 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 achieve this step, a description of which is outside the scope of this proposal.

When the CE/VM gets activated the attached PE starts exporting its IP address with the corresponding zone-RT. This creates a full mesh connectivity between the newly active VM and the rest of the VMs in the zone.

Note that the IP address mask of the CE/VM need not necessarily be a /32. 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 and use L3VPN style forwarding to carry traffic between the VMs.

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.

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 are CE's attached to a service-PE.

A service-PE is a normal L3VPN 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-PE's.

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 service PE closest to the source zone is the source service-PE and the service PE closest to the destination zone is called the destination service-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.

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 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) 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 toplogy. 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 with the set of service-topology-RT's that they belong to at the point of origination.

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.

o 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.

o 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: Paul Quinn, David Ward, Ashok Ganesan, Peter Bosch.

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

Email: dhrao@cisco.com

Rex Fernando
Cisco
170 W Tasman Dr
San Jose, CA

Email: rex@cisco.com

Luyuan Fang
Cisco
170 W Tasman Dr
San Jose, CA

Email: lufang@cisco.com