INTERNET-DRAFT
Intended Status: Informational
Expires: January 4, 2015

Maria Napierala
AT&T
Luyuan Fang
Microsoft

July 4, 2014

Requirements for Extending BGP/MPLS VPNs to End-Systems
draft-fang-l3vpn-end-system-requirements-03.txt

Abstract

The proven scalability and extensibility of the BGP/MPLS IP VPNs (IP VPN) technology has made it an attractive candidate for data center/cloud virtualization. Virtualized end-system environment imposes additional requirements to MPLS/BGP VPN technology. This document provides the requirements for extending IP VPN technology (in original or modified versions) into the end-systems/hosts, such as a server in a data center.

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

Requirements Language

Although this document is not a protocol specification, 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].

1 Introduction

Enterprise networks are increasingly being consolidated and outsourced in an effort to improve the deployment time of services as well as reduce operational costs. This coincides with an increasing demand for compute, storage, and network resources from applications. Logical abstraction of these resources is needed to for improved scalability and cost efficiency. This is referred as server, storage, and network virtualization. It can be implemented in all layers of the computer systems or networks. The virtualized loads are executed or transferred over a common physical infrastructure. Compute nodes running guest operating systems are often executed as Virtual Machines (or VMs).

This document defines requirements for a network virtualization solution that provides secure IP VPN connectivity to virtual resources on end-systems operating in a multi-tenant shared physical infrastructure. The requirements address the needs of virtual resources, defined as Virtual Machines, applications, and appliances that require only IP connectivity. Non-IP communication is addressed by other solutions and is not in scope of this document.

The technical solutions to support these requirements are work in progress in IETF [I-D.ietf-l3vpn-end-system], [I-D.fang-l3vpn-virtual-pe]. The solutions may referred as End-System solutions or virtual PE (vPE) solutions in different documents.

1.1 Terminology

The key words "MUST", "MUST NOT", "REQUIRED", "SHALL", "SHALL NOT", "SHOULD", "SHOULD NOT", "RECOMMENDED", "MAY", and "OPTIONAL" in this document are to be interpreted as described in RFC 2119 [RFC2119].

Term	Definition
AS	Autonomous System

CE	Customer Edge router
End-System	A device where Guest OS, Host OS/Hypervisor reside
GRE	Generic Routing Encapsulation
Hypervisor	Virtual Machine Manager
Iaa	Infrastructure as a Service
PE	Provider Edge router
RT	Route Target
RTC	RT Constraint
SDN	Software Defined Network
ToR	Top-of-Rack switch
VM	Virtual Machine
vPE	virtual Provider Edge Router
VPN	Virtual Private Network

2. Application of MPLS/BGP VPNs to End-Systems

MPLS/BGP VPN technology [RFC4364] have proven to be able to scale to a large number of VPNs (tens of thousands) and customer routes (millions) while providing for aggregated management capability. In traditional WAN deployments of BGP IP VPNs a Customer Edge (CE) is a physical device, residing a customer’s location, connected to a Provider Edge (PE), residing in a Service Provider’s location. CE devices are logically part of a customer’s VPN while PE routers are logically part of the SP’s network. In a traditional MPLS/BGP VPN deployment, a CE device is a router and it is a routing peer of a PE to which it is attached via an attachment circuit. In addition, the forwarding function and control function of a Provider Edge (PE) device co-exist within a single physical router.

MPLS/BGP VPN technology can be evolved and adapted to new virtualized environments by implementing the VPN forwarding edge functionality on the end-system hosts and thereby extending VPN service directly to end-systems.

2.1. End-System CE and PE Functions

When end-system attaches to MPLS/BGP VPN, CE corresponds to a non-routing host that can reside in a Virtual Machine or be an application residing on the end-system itself.

As in traditional MPLS/BGP VPN deployments, it is undesirable for the end-system VPN forwarding knowledge to extend to the transport network infrastructure. Hence, optimally, with regard to forwarding, the end-system should become both the CE and the PE simultaneously.

The network virtualization solution should also support deployments where it is not possible or not desirable to co-locate the PE and CE functionality. In such deployments PE may be implemented on an external

device with remote CE attachments. This external PE device should be as close as possible to the end-system where the CE resides. The external PE devices that attach to a particular VPN, need to know, for each attachment circuit leading to that VPN, the host address that is reachable over that attachment circuit. The end-system MPLS/BGP VPN solution must specify a method to convey this information from the end-system to the PE.

The same network virtualization solution should support deployments with mixed, internal (co-located with CE) and external PE (i.e., remote CE) implementations.

2.2. PE Control Plane Function

It is a current practice to implement MPLS/BGP VPN PE forwarding and control functions in different processors of the same device and to use internal (proprietary) communication between those processors. Typically, the PE control functionality is implemented in one (or very few) components of a device and the PE forwarding functionality is implemented in multiple components of the same device (a.k.a., "line cards").

In end-system environment, a single end-system, effectively, corresponds to a line card in a traditional PE router. For scalable and cost effective deployment of end-system MPLS/BGP VPNs the PE forwarding function should be decoupled from PE control function such that the former can be implemented on multiple standalone devices. This separation of functionality will allow for implementing the end-system PE forwarding on multiple end-system devices, for example, in operating systems of application servers or network appliances. Moreover, the separation of PE forwarding and control plane functions allows for the PE control plane function to be itself virtualized and run as an application in end-system.

3. VPN Communication Requirements

3.1. Unicast IPv4 and IPv6

A network virtualization solution should be able to provide IPv4 and IPv6 unicast connectivity between hosts in the same and different subnets without any assumptions regarding the underlying media layer.

3.2. Multicast/VPN Broadcast IPv4 and IPv6

Furthermore, the multicast transmission, i.e., allowing IP applications to send packets to a group of IPv4 or IPv6 addresses should be supported. The multicast service should also support a delivery of traffic to all endpoints of a given VPN even if those endpoints have not

sent any control messages indicating the need to receive that traffic. In other words, the multicast service should be capable of delivering the IP broadcast traffic in a virtual topology. A solution for supporting VPN multicast and VPN broadcast must not require that the underlying transport network supports IP multicast transmission service.

3.3. IP Subnet Support

In some deployments, Virtual Machines or applications are configured to belong to an IP subnet. A network virtualization solution should support grouping of virtual resources into IP subnets regardless of whether the underlying implementation uses a multi-access network or not. While some applications may expect to find other peers in a particular user defined IP subnet, this does not imply the need to provide a layer 2 service that preserves MAC addresses. End-system network virtualization solution should be able to provide IP (unicast, multicast, VPN broadcast) connectivity between hosts in the same and different subnets without any assumptions regarding the underlying media layer.

4. Multi-Tenancy Requirements

One of the main goals of network virtualization is to provide traffic and routing isolation between different virtual components that share a common physical infrastructure. Networks use various VPN technologies to isolate disjoint groups of virtual resources. Some use VLANs [IEEE.802-1Q] as a VPN technology, others use layer 3 based solutions, often with proprietary control planes. Service Providers are interested in interoperability and in openly documented protocols rather than in proprietary solutions.

A collection of virtual resources might provide external or internal services. Such collection may serve an external "customer" or internal "tenant" to whom a Service Provider provides service(s). In MPLS/BGP VPN terminology a collection of virtual resources dedicated to a process or application corresponds to a VPN.

A network virtualization multi-tenancy solution should support the following:

- Tenant or application isolation, in data plane and control plane, while sharing the same underlying physical network. Tenants should be able to independently select and deploy their choice of IP address space: public or private IPv4 and/or IPv6.

- Multiple distinct VPNs per tenant. Tenant’s inter-VPN traffic should be allowed to cross VPN boundaries, subject to access controls and/or routing policies.

- Inter-VPN communication, subject to access policies. Typically VPNs that belong to different external tenants do not communicate with each other directly but they should be allowed to access shared services or shared network resources. It is often the case that SP infrastructure services are provided to multiple tenants, for example voice-over-IP gateway services or video-conferencing services for branch offices.

- VM or application end-point should be able to directly access multiple VPNs without a need to traverse a gateway.

End-system network virtualization solution should support both, isolated VPNs as well as overlapping VPNs (often referred to as "extranets"). It should also support any-to-any and hub-and-spoke topologies.

5. Decoupling of Virtualized Networking from Physical Infrastructure

One of the main goals in designing a large scale transport network is to minimize the cost and complexity of its "fabric" by delegating the virtual resource communication processing to the network edge. It has been proven (in Internet and in large MPLS/BGP VPN deployments) that moving complexity to network edge while keeping network core simple has very good scaling properties.

The transport network infrastructure should not maintain any information that pertains to the virtual resources in end-systems. Decoupling of virtualized networking from the physical infrastructure has the following advantages: 1) provides better scalability; 2) simplifies the design and operation; 3) reduces network cost.

Decoupling of virtualized networking from underlying physical network consists in the following:

- Separation between the virtualized segments (i.e., interface associated with virtual resources) and the physical network (i.e., physical interfaces associated with network infrastructure).

- Separation of the virtual network IP address space from the physical infrastructure network IP address space. In the case of a transport other than IP, for example MPLS or Ethernet, the infrastructure address refers to the Subnetwork Point of Attachment (SNPA) address in a given multi-access network.

- The physical infrastructure addresses should be routable (or switchable) in the underlying transport network, while the virtual network addresses should be routable only in the virtual network.

- The virtual network control plane should be decoupled from the

underlying transport network.

6. Decoupling of Layer 3 Virtualization from Layer 2 Topology

The layer 3 approach to network virtualization dictates that the virtualized communication should be routed, not bridged. The layer 3 virtualization solution should be decoupled from the layer 2 topology. Thus, there should be no dependency on VLANs and layer 2 broadcast.

In solutions that depend on layer 2 broadcast domains, host-to-host communication is established based on flooding and data plane MAC learning. Layer 2 MAC information has to be maintained on every switch where a given VLAN is present. Even if some solutions are able to minimize data plane MAC learning and/or unicast flooding, they still rely on MAC learning at the network edge and on maintaining the MAC addresses on every switch where the layer 2 VPN is present.

The MAC addresses known to guest OS in end-system are not relevant to IP services and introduce unnecessary overhead. Hence, the MAC addresses associated with virtual resources should not be used in the virtual layer 3 networks. Rather, only what is significant to IP communication, namely the IP addresses of the virtual machines and application endpoints should be maintained by the virtual networks.

7. Requirements for Encapsulation of Virtual Payloads

In order to scale the transport networks, the virtual network payloads must be encapsulated with headers that are routable (or switchable) in the physical network infrastructure. The IP addresses of the virtual resources are not to be advertized within the physical infrastructure address space.

The encapsulation (and de-capsulation) function should be implemented on a device as close to virtualized resources as possible. Since the hypervisors in the end-systems are the devices at the network edge they are the most optimal location for the encap/decap functionality.

The network virtualization solution should also support deployments where it is not possible or not desirable to implement the virtual payload encapsulation in the hypervisor/Host OS. In such deployments encap/decap functionality may be implemented in an external device. The external device implementing encap/decap functionality should be a close as possible to the end-system itself. The same network virtualization solution should support deployments with both, internal (in a hypervisor) and external (outside of a hypervisor) encap/decap devices.

Whenever the virtual forwarding functionality is implemented in an external device, the virtual service itself must be delivered to an end-system such that switching elements connecting the end-system to the encap/decap device are not aware of the virtual topology.

7.1. Encapsulation Methods

MPLS/VPN technology based on [RFC4364] specifies that different encapsulation methods could be for connecting PE routers, namely Label Switched Paths (LSPs), IP tunneling, and GRE tunneling.

If LSPs are used in the transport network they could be signaled with LDP, in which case host (/32) routes to all PE routers must be propagated throughout the network, or with RSVP-TE, in which case a full mesh of RSVP-TE tunnels is required. The label forwarding tables can also be constructed using SDN controllers without the need of distributed signaling protocols.

If the transport network is only IP-capable then MPLS in IP or MPLS in GRE [RFC4023] encapsulation could be used. Due to route aggregation property of IP protocols, with IP/GRE encapsulation the PE host routes do not have to be present in the transport network.

7.2. Routing of Virtual Payloads

A device implementing the encap/decap functionality acts as the first-hop router in the virtual topology.

In a layer 3 end-system virtual network, IP packets should reach the first-hop router in one IP-hop, regardless of whether the first-hop router is an end-system itself (i.e., a hypervisor/Host OS) or it is an external (to end-system) device. The first-hop router should always perform an IP lookup on every packet it receives from a virtual machine or an application. The first-hop router should encapsulate the packets and route them towards the destination end-system.

8. Optimal Forwarding of Traffic

The network virtualization solutions that optimize for the maximum utilization of compute and storage resources require that those resources may be located anywhere in the network. The physical and logical spreading of appliances and workloads implies a very significant increase in the infrastructure bandwidth consumption. In order to be efficient in terms of traffic forwarding, the virtualized networking solutions must assure that packets traverse the transport network only once.

It must be also possible to send the traffic directly from one end-system to another end-system without traversing through a midpoint router.

9. IP Mobility

Another reason for a network virtualization is the need to support IP mobility. IP mobility means that IP addresses used for communication within or between applications can be located anywhere across the virtual network. Using a virtual topology, i.e., abstracting the externally visible network address from the underlying infrastructure address is an effective way to solve IP mobility problem.

IP mobility consists in a device physically moving (e.g., a roaming wireless device) or a workload being transferred from one physical server/appliance to another. IP mobility requires preserving device’s active network connections (e.g., TCP and higher-level sessions). Such mobility is also referred to as "live" migration with respect to a Virtual Machine. IP mobility is highly desirable for many reasons such as efficient and flexible resource sharing, data center migration, disaster recovery, server redundancy, or service bursting.

9.1. IP Addressing of Virtual Hosts

To accommodate live mobility of a virtual machine (or a device), it is desirable to assign to it a semi-permanent IP address that remains with the VM/device as it moves. The semi-permanent IP address can be configured through VM or device configuration process or by means of DHCP.

9.2. Network Layer-Based Mobility

When dealing with IP-only applications it is not only sufficient but optimal to forward the traffic based on layer 3 (network layer) rather than on layer 2 (data-link layer) information. The MAC addresses of devices or applications are irrelevant to IP services and introduce unnecessary overhead and complications when devices or VMs move. For example, when a VM moves between physical servers, the MAC learning tables in the switches must be updated. Moreover, it is possible that VM’s MAC address might need to change in its new location. In IP-based network virtualization solution a device or a workload move is handled by an IP route advertisement.

9.3. Routing Convergence Requirements

IP mobility has to be transparent to applications and any external entity interacting with the applications. This implies that the network connectivity restoration time is critical. The transport

sessions can typically survive over several seconds of disruption, however, applications may have sub-second latency requirement for their correct operation.

To minimize the disruption to established communication during workload or device mobility, the control plane of a network virtualization solution should be able to differentiate between the activation of a workload in a new location from advertizing its route to the network. This will enable the remote end-points to update their routing tables prior to workload’s migration as well as allowing the traffic to be tunneled via the workload’s old location.

10. Inter-operability with Existing MPLS/BGP VPNs

Service Providers want to tie their server-based offerings to their MPLS/BGP VPN services. MPLS/BGP VPNs provide secure and latency-optimized remote connectivity to the virtualized resources in SP’s data center. The Service Provider-based VPN access can provide additional capabilities compared with public internet access, such as QoS, OAM, multicast service, VoIP service, video conferencing, wireless connectivity.

MPLS/BGP VPN customers may require simultaneous access to resources in both SP and their own data centers.

Service Providers want to "spin up" the L3VPN access to data center VPNs as dynamically as the spin up of compute and other virtualized resources.

The network virtualization solution should be fully inter-operable with MPLS/BGP VPNs, including:

- Inter-AS MPLS/BGP VPN Options A, B, and C [RFC4364].

- BGP/MPLS VPN-capable network devices (such as routers and network appliances) should be able to participate directly in a virtual network that spans end-systems.

- The network devices should be able to participate in isolated collections of end-systems, i.e., in isolated VPNs, as well as in overlapping VPNs (called "extranets" in BGP/MPLS VPN terminology).

- The network devices should be able to participate in any-to-any and hub-and-spoke end-systems topologies.

  When connecting an end-system VPN to other networks, it should not be necessary to advertize the specific host routes but rather the aggregated routing information. A BGP/MPLS VPN-capable router or

appliance can be used to aggregate VPN’s IP routing information and advertize the aggregated prefixes. The aggregated prefixes should be advertized with the router/appliance IP address as BGP next-hop and with locally assigned aggregate 20-bit label. The aggregate label should trigger a destination IP lookup in its corresponding VRF on all the packets entering the virtual network.

The inter-connection of end-system VPNs with traditional VPNs requires an integrated control plane and unified orchestration of network and end-system resources.

11. BGP Requirements in a Virtualized Environment

11.1. BGP Convergence and Routing Consistency

BGP was designed to carry very large amount of routing information but it is not a very fast converging protocol. In addition, the routing protocols, including BGP, have traditionally favored convergence (i.e., responsiveness to route change due to failure or policy change) over routing consistency. Routing consistency means that a router forwards a packet strictly along the path adopted by the upstream routers. When responsiveness is favored, a router applies a received update immediately to its forwarding table before propagating the update to other routers, including those that potentially depend upon the outcome of the update. The route change responsiveness comes at the cost of routing blackholes and loops.

Routing consistency in virtualized environments is important because multiple workloads can be simultaneously moved between different physical servers due to maintenance activities, for example. If packets sent by the applications that are being moved are dropped (because they do not follow a live path), the active network connections will be dropped. To minimize the disruption to the established communications during VM migration or device mobility, the live path continuity is required.

11.1.1. BGP IP Mobility Requirements

In IP mobility, the network connectivity restoration time is critical. In fact, Service Provider networks already use routing and forwarding plane techniques that support fast failure restoration by pre-installing a backup path to a given destination. These techniques allow to forward traffic almost continuously using an indirect forwarding path or a tunnel to a given destination, and hence, are referred to as "local repair". The traffic forwarding path is restored locally at the destination’s old location while the network converges to a backup path. Eventually, the network converges to an optimal path and bypasses the local repair. BGP assists in the local

repair techniques by advertizing multiple paths and not only the best path to a given destination.

11.2. Optimization of Route Distribution

When virtual networks are triggered based on the IP communication, the Route Target Constraint extension [RFC4684] of BGP should be used to optimize the route distribution for sparse virtual network events. This technique ensures that only those VPN forwarders that have local participants in a particular data plane event receive its routing information. This also decreases the total load on the upstream BGP speakers.

12. Service chaining

A service chain is a deployment where a sequence of appliances intermediate traffic between networks. In fact, traffic from one virtual network may go through an arbitrary graph of service nodes before reaching another virtual network. Service chains can contain a mixture of virtual services (implemented as VMs on compute nodes) and physical services (hosted on service nodes). Network appliances tend to be designed to operate on an "inside/outside" interface model. This type of applications do not terminate traffic and are transparent to packets. In an SDN approach, the service chain is configured and managed in software that adds and removes services from the chain in an automated way. It is a requirement that service chaining is supported on devices using MPLS/BGP VPN technology for virtual networking.

Connecting appliances in a sequence has been done for many years using VLANs. However, "service-chaining" cannot be implemented without solving the problem of how to bring in traffic from a routed network into the set of appliances. The issue is always how to attract the traffic in and forward it out of the service-chain, i.e., how to integrate the service-chain with routing. By using the same mechanism to route traffic in and out of a service chain as well as through its intermediate hops, the implementation of service chains is significantly simplified.

One solution currently work in progress in IETF is [I-D.rfernando-l3vpn-service-chaining].

12.1. Load Balancing

One of the main requirements of service-chaining is horizontal scaling of a service in a service-chain to tens or hundreds of instances. When using MPLS/BGP VPN routing instance (or VRF)

construct to implement service chaining, the load balancing is built-in. The load balancing corresponds to BGP multipath where multiple routes for a single prefix are installed in a routing instance. The multiple BGP routes in the routing table translate to Equal Cost Multi-Path in the forwarding plane. The hash used in the load balancing algorithm can be per packet, per flow or per prefix. The forwarding plane should support load balancing over several hundreds next-hops.

Load balancing should support deployments where both, virtual and physical service appliances are present. It should support deployments where virtual service instances are spread across the same and different end-systems/hosts.

12.2. Symmetric Service Chain Support

If a service function is stateful, it is required that forward flows and reverse flows always pass through the same service instance. ECMP does not provide this capability, since the hash calculation will see different input data for the same flow in the forward and reverse directions. Additionally, if the number of service instances changes, either to expand/decrease capacity or due to an instance failure, the hash table in ECMP is recalculated, and most flows will be re-directed to a different service instance, causing user session disruption.

It is a requirement that service chaining solution satisfies the requirements of symmetric forward/reverse paths for flows and a minimal traffic disruption when service instances are added to or removed from a set of instances.

12.3. Packet Header Transforming Services

A service in a service chain might perform an action that changes the packet header information, e.g., the packet’s source address (such as performed by NAT service). In order to support the reverse traffic flow traffic in this case, the routing and forwarding information has to be modified such that the traffic can be directed via the instances of the transforming service. For example, the original routes with a source prefix (Network-A) are replaced with a route that has a prefix that includes all the possible addresses that the source address could be mapped to. In the case of network address translation, this would correspond to the NAT pool.

It is a requirement that service chaining solution supports services that manipulate packet headers.

13. Security Considerations

The document presents the requirements for end-systems MPLS/BGP VPNs. The security considerations for traditional MPLS/BGP VPN deployments are described in [RFC4364] in Section 13. The additional security issues associated with deployments using MPLS-in-GRE or MPLS-in-IP encapsulations are described in [RFC4023] in Section 8. In addition, [RFC4111] provides general IP VPN security guidelines.

The additional security requirements specific to end-system MPLS/BGP VPNs are as follows:

- End-systems MPLS/BGP VPNs solution should guarantee that packets originating from a specific end-system virtual interface are accepted only if the corresponding VPN IP host is present on that end-system.

- Virtual network must ensure that traffic arriving at the egress end-system is being sent from the correct ingress end-system.

- One virtual host or VM should not be able to impersonate another, during steady-state operation and during live migration.

The security considerations for specific solutions will be documented in the relevant documents.

13. IANA Considerations

This document contains no new IANA considerations.

14. References

14.1. Normative References

[RFC4023] Worster, T., Rekhter, Y., and E. Rosen, Ed., "Encapsulating MPLS in IP or Generic Routing Encapsulation (GRE)", RFC 4023, March 2005.

[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.

[IEEE.802-1Q] Institute of Electrical and Electronics Engineers, "Local and Metropolitan Area Networks: Virtual Bridged Local Area Networks", IEEE Std 802.1Q-2005, May 2006.

14.2. Informative References

[RFC2119] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, March 1997.

[RFC4111] Fang, L., Ed., "Security Framework for Provider-Provisioned Virtual Private Networks (PPVPNs)", RFC 4111, July 2005.

[I-D.ietf-l3vpn-end-system] Marques, P., Fang, L., Pan, P., Shukla, A., Napierala, M., "BGP-signaled end-system IP/VPNs", draft-ietf-l3vpn-end-system, work in progress.

[I-D.fang-l3vpn-virtual-pe] Fang, L., Ward, D., Fernando, R., Napierala, M., Bitar, N., Rao, D., Rijsman, B., So, N., "BGP IP VPN Virtual PE", draft-fang-l3vpn-virtual-pe, work in progress.

[I-D.rfernando-l3vpn-service-chaining] Fernando, R., Rao, D., Fang, L., Napierala, M., So, N., draft-rfernando-l3vpn-service-chaining, work in progress.

Acknowledgements

The authors would like to thank Pedro Marques and Han Nguyen for the comments and suggestions.

Authors' Addresses

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

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

INTERNET-DRAFT
Intended Status: Standards track
Expires: January 4, 2015

Luyuan Fang, Ed.
Microsoft
David Ward
Rex Fernando
Cisco
Maria Napierala
AT&T
Nabil Bitar
Verizon
Dhananjaya Rao
Cisco
Bruno Rijsman
Juniper

Ning So
Vinci Systems
Jim Guichard
Cisco
Wen Wang
CenturyLink
Manuel Paul
Deutsche Telekom
Wim Henderichx
Alcatel-Lucent

July 4, 2014

BGP/MPLS VPN Virtual PE
draft-fang-l3vpn-virtual-pe-05

Abstract

This document describes the architecture solutions for BGP/MPLS L3 and L2 Virtual Private Networks (VPNs) with virtual Provider Edge (vPE) routers. It provides a functional description of the vPE control, forwarding, and management. The proposed vPE solutions support both the Software Defined Networks (SDN) approach which allows physical decoupling of the control and the forwarding, and the traditional distributed routing approach. A vPE can reside in any network or compute devices, such as a server as co-resident with the application virtual machines (VMs), or a Top-of-Rack (ToR) switch in a Data Center (DC) network.

Status of this Memo

Table of Contents

1 Introduction

Network virtualization enables multiple isolated individual networks over a shared common network infrastructure. BGP/MPLS IP Virtual Private Networks (IP VPNs) [RFC4364] have been widely deployed to provide network based Layer 3 VPNs solutions. [RFC4364] provides routing isolation among different customer VPNs and allow address overlap among these VPNs through the implementation of per VPN Virtual Routing and Forwarding instances (VRFs) at a Service Provider Edge (PE) routers, while forwarding customer traffic over a common IP/MPLS network. For L2 VPN, a similar technology is being defined in [I-D.ietf-l2vpn-evpn] on the basis of BGP/MPLS, to provide switching isolation and allow MAC address overlap.

With the advent of compute capabilities and the proliferation of virtualization in Data Center servers, multi-tenant Data Centers are becoming the norm. As applications and appliances are increasingly being virtualized, support for virtual edge devices, such as virtual L3/L2 VPN PE routers, becomes feasible and desirable for Service Providers who want to extend their existing MPLS VPN deployments into Data Centers to provide end-to-end Virtual Private Cloud (VPC) services. Virtual PE work is also one of early effort for Network Functions Virtualization (NFV). In general, scalability, agility, and cost efficiency are primary motivations for vPE solutions.

The virtual Provider Edge (vPE) solution described in this document allows for the extension of the PE functionality of L3/L2 VPN to an end device, such as a server where the applications reside, or to a first hop routing/switching device, such as a Top of the Rack (ToR) switch in a DC.

The vPE solutions support both the Software Defined Networks (SDN) approach, which allows physical decoupling of the control and the forwarding, and the traditional distributed routing approach.

1.1 Terminology

The key words "MUST", "MUST NOT", "REQUIRED", "SHALL", "SHALL NOT", "SHOULD", "SHOULD NOT", "RECOMMENDED", "MAY", and "OPTIONAL" in this document are to be interpreted as described in RFC 2119 [RFC2119].

Term	Definition
ASBR	Autonomous System Border Router
BGP	Border Gateway Protocol
CE	Customer Edge
Forwarder	IP VPN forwarding function

GRE	Generic Routing Encapsulation
Hypervisor	Virtual Machine Manager
I2RS	Interface to Routing Systems
LDP	Label Distribution Protocol
MP-BGP	Multi-Protocol Border Gateway Protocol
MPLS	Multi-Protocol Label Switching
PCEF	Policy Charging and Enforcement Function
QoS	Quality of Service
RR	Route Reflector
RT	Route Target
RTC	RT Constraint
SDN	Software Defined Networks
ToR	Top-of-Rack switch
VI	Virtual Interface
vCE	virtual Customer Edge Router
VM	Virtual Machine
vPC	virtual Private Cloud
vPE	virtual Provider Edge Router
vPE-C	virtual Provider Edge Control plane
vPE-F	virtual Provider Edge Forwarder
VPN	Virtual Private Network
vRR	virtual Route Reflector
WAN	Wide Area Network

End device: where Guest OS, Host OS/Hypervisor, applications, VMs, and virtual router may reside.

1.2 Requirements

The following are key requirements for vPE solutions.

1) MUST support end device multi-tenancy, per tenant routing isolation and traffic separation.

2) MUST support large scale MPLS VPNs in the Data Center, upto tens of thousands of end devices and millions of VMs in the single Data Center.

3) MUST support end-to-end MPLS VPN connectivity, e.g. MPLS VPN can start from a DC end device, connect to a corresponding MPLS VPN in the WAN, and terminate in another Data Center end device.

4) MUST allow physical decoupling of MPLS VPN PE control and forwarding for network virtualization and abstraction.

5) MUST support the control plane with both SDN controller approach, and the traditional distributed control plane approach with MP-BGP protocol.

6) MUST support VM mobility.

7) MUST support orchestration/auto-provisioning deployment model.

8) SHOULD be capable to support service chaining as part of the solution [I-D.rfernando-l3vpn-service-chaining], [I-D.bitar-i2rs-service-chaining].

The architecture and protocols defined in BGP/MPLS IP VPN [RFC4364] and BGP/MPLS EVPN [I-D.ietf-l2vpn-evpn] provide the foundation for vPE extension. Certain protocol extensions may be needed to support the virtual PE solutions.

2. Virtual PE Architecture

2.1 Virtual PE definitions

As defined in [RFC4364] and [I-D.ietf-l2vpn-evpn], an MPLS VPN is created by applying policies to form a subset of sites among all sites connected to the backbone networks. It is a collection of "sites". A site can be considered as a set of IP/ETH systems maintaining IP/ETH inter-connectivity without direct connecting through the backbone. The typical use of L3/L2 VPN has been to inter-connect different sites of an Enterprise networks through a Service Provider’s BGP MPLS VPNs in the WAN.

A virtual PE (vPE) is a BGP/MPLS L3/L2 VPN PE software instance which may reside in any network or computing devices. The control and forwarding components of the vPE can be decoupled, they may reside in the same physical device, or in different physical devices.

A virtualized Provider Edge Forwarder (vPE-F) is the forwarding element of a vPE. vPE-F can reside in an end device, such as a server in a Data Center where multiple application Virtual Machines (VMs) are supported, or a Top-of-Rack switch (ToR) which is the first hop switch from the Data Center edge. When a vPE-F is residing in a server, its connection to a co-resident VM can be viewed as similar to the PE- CE relationship in the regular BGP L3/L2 VPNs, but without routing protocols or static routing between the virtual PE and end-host because the connection is internal to the device.

The vPE Control plane (vPE-C) is the control element of a vPE. When using the approach where control plane is decoupled from the physical topology, the vPE-F may be in a server and co-resident with application VMs, while one vPE-C can be in a separate device, such as an SDN Controller where control plane elements and orchestration functions are located. Alternatively, the vPE-C can reside in the same physical device as the vPE-F. In this case, it is similar to the

traditional implementation of VPN PEs where, distributed MP-BGP is used for L3/L2 VPN information exchange, though the vPE is not a dedicated physical entity as it is in a physical PE implementation.

2.2 vPE Architecture and Design options

2.2.1 vPE-F host location

Option 1a. vPE-F is on an end device as co-resident with application VMs. For example, the vPE-F is on a server in a Data Center.

Option 1b. vPE-F forwarder is on a ToR or other first hop devices in a DC, not as co-resident with the application VMs.

Option 1c. vPE-F is on any network or compute devices in any types of networks.

2.2.2 vPE control plane topology

Option 2a. vPE control plane is physically decoupled from the vPE-F. The control plane may be located in a controller in a separate device (a stand alone device or can be in the gateway as well) from the vPE forwarding plane.

Option 2b. vPE control plane is supported through dynamic routing protocols and located in the same physical device as the vPE-F.

2.2.3 Data Center orchestration models

Option 3a. Push model: It is a top down approach, push IP VPN provisioning state from a network management system or other centrally controlled provisioning system to the IP VPN network elements.

Option 3b. Pull model: It is a bottom-up approach, pull state information from network elements to network management/AAA based upon data plane or control plane activity.

2.3 vPE Architecture reference models

2.3.1 vPE-F in an end-device and vPE-C in the controller

Figure 1 illustrates the reference model for a vPE solution with the vPE-F in the end device co-resident with applications VMs, while the vPE-C is physically decoupled and residing on a controller.

The Data Center is connected to the IP/MPLS core via the Gatways/ASBRs. The MPLS VPN , e.g. VPN RED, has a single termination point within the DC at one of the VPE-F, and is inter-connected in the WAN to other member sites which belong to the same client, and the remote ends of VPN RED can be a PE which has VPN RED attached to it, or another vPE in a different Data Center.

Note that the DC fabrics/intermediate underlay devices in the DC do not participate IP VPNs, their function is the same as provider backbone routers in the IP/MPLS back bone and they do not maintain the VPN states, nor they are VPN aware.

```
                          ,-----.
                         (       ')
                       .--(.       '.---.
                      (     '      '     )
                     (     IP/MPLS WAN    )
                      (.                .)
                       (     (        .)
            WAN         ''--' '-''---'
            ----------------|----------|------------------------
            Service/DC      |          |
            Network     +-------+   +-------+
                        |Gateway|---|Gateway|  *
                        | /ASBR |   | /ASBR |     *
                        +-------+   +-------+       *
                            |          |         +-------------+
                            |    ,---. |         |Controller   |
                          .---. (     '.---.     |(vPE-C and   |
                         (     '      '     ')   |orchestrator)|
                        (     Data Center     )  +-------------+
                         (.      Fabric      )           *
                          (     (       ).--'            *
                       /   ''--' '-''--'       \        *
                      /     /            \      \     *
             +-------+   +-------+   +-------+   +-------+
             | vPE-F |   | vPE-F |   | vPE-F |   | vPE-F |
             +---+---+   +---+---+   +---+---+   +---+---+
             |VM |VM |   |VM |VM |   |VM |VM |   |VM |VM |
             +---+---+   +---+---+   +---+---+   +---+---+
             |VM |VM |   |VM |VM |   |VM |VM |   |VM |VM |
             +---+---+   +---+---+   +---+---+   +---+---+

             End Device  End Device  End Device  End Device
```

Figure 1. Virtualized Data Center with vPE at the end device and vPE-C and vPE-F physically decoupled

Note:

a) *** represents Controller logical connections to the all Gateway/ASBRs and to all vPE-F.

b) ToR is assumed included in the Data Center cloud.

2.3.2 vPE-F and vPE-C on the same end-device

In this option, vPE-F and vPE-C functionality are both resident in the end-device. The vPE functions the same as it is in a physical PE. MP-BGP is used for the VPN control plane. Virtual or physical Route Reflectors (RR) (not shown in the diagram) can be used to assist scaling.

```
                          ,-----.
                         (       ')
                      .--(.       '.---.
                     (     ,      '     )
                    (     IP/MPLS WAN    )
                     (.                 .)
                      (     (        .)
          WAN          ''--' '-''---'
          ----------------|----------|----------------------
          Service/DC      |          |
          Network     +-------+   +-------+
                      |Gateway|---|Gateway|
                      | /ASBR |   | /ASBR | *
                      +-------+   +-------+   *
                          |          |          * MP-BGP
                          |    ,---. |            *
                      .---. (     '.---.           *
                     (     ,      '     ')          *
                    (     Data Center     )          *
                     (.      Fabric      )           *
                      (     (       ).--'            *
                   /   ''--' '-''--'       \        *
                  /     /            \      \     *
           +-------+   +-------+   +-------+   +-------+
           |  vPE  |   |  vPE  |   |  vPE  |   |  vPE  |
           +---+---+   +---+---+   +---+---+   +---+---+
           |VM |VM |   |VM |VM |   |VM |VM |   |VM |VM |
           +---+---+   +---+---+   +---+---+   +---+---+
           |VM |VM |   |VM |VM |   |VM |VM |   |VM |VM |
           +---+---+   +---+---+   +---+---+   +---+---+

           End Device  End Device  End Device  End Device
```

Figure 2. Virtualized Data Center with vPE at the end device, VPN control signal uses MP-BGP

Note:

a) *** represents the logical connections using MP-BGP among the Gateway/ASBRs and to the vPEs on the end devices.

b) ToR is assumed included in the Data Center cloud.

2.3.3 vPE-F and vPE-C are on the ToR

In this option, vPE functionality is the same as a physical PE. MP-BGP is used for the VPN control plane. Virtual or physical Route Reflector (RR) (not shown in the diagram) can be used to assist scaling.

```
                          ,-----.
                         (       )
                      .--(.       ).---.
                     (     ,      ,     )
                    (     IP/MPLS WAN    )
                     (.                 .)
                      (     (        .)
           WAN         ''--' '-''---'
           ----------------|----------|----------------------
           Service/DC      |          |
           Network     +-------+   +-------+
                       |Gateway|---|Gateway|
                       | /ASBR |   | /ASBR | *
                       +-------+   +-------+   *
                           |          |          *  MP-BGP
                           |    ,---. |           *
                         .---. (     ).---.        *
                        (     ,      ,     )       *
                       (     Data Center     )     *
                        (.      Fabric      )     *
                         (     (       ).--'     *
                         /''--' '-/'--'     \  *
                   +---+---+  +---+---+  +---+---+
                   |vPE|   |  |vPE|   |  |vPE|   |
                   +---+   |  +---+   |  +---+   |
                   |  ToR  |  |  ToR  |  |  ToR  |
                   +-------+  +-------+  +-------+
                    /     \    /      \    /     \
            +-------+   +-------+   +-------+   +-------+
            |  vPE  |   |  vPE  |   |  vPE  |   |  vPE  |
            +---+---+   +---+---+   +---+---+   +---+---+
            |VM |VM |   |VM |VM |   |VM |VM |   |VM |VM |
            +---+---+   +---+---+   +---+---+   +---+---+
            |VM |VM |   |VM |VM |   |VM |VM |   |VM |VM |
            +---+---+   +---+---+   +---+---+   +---+---+
            End Device  End Device  End Device  End Device
```

Figure 3. Virtualized Data Center with vPE at the ToP, VPN control signal uses MP-BGP

Note: *** represents the logical connections using MP-BGP among the Gateway/ASBRs and to the vPEs on the ToRs.

2.3.4 vPE-F on the ToR and vPE-C on the controller

In this option, the L3/L2 VPN termination is at the ToR, but the control plane decoupled from the data plane and resided in a controller, which can be on a stand alone device, or can be placed at the Gateway/ASBR.

2.3.5 The server view of a vPE

An end device shown in Figure 4 is a virtualized server that hosts multiple VMs. The virtual PE is co-resident in the server with application VMs. The vPE supports multiple VRFs, VRF Red, VRF Grn, VRF Yel, VRF Blu, etc. Each application VM is associated to a particular VRF as a member of the particular VPN. For example, VM1 is associated to VRF Red, VM2 and VM47 are associated to VRF Grn, etc. Routing/switching isolation applies between VPNs for multi-tenancy support. For example, VM1 and VM2 cannot communicate directly in a simple intranet VPN topology as shown in the configuration.

The vPE connectivity relationship between vPE and the application VM is similar to the PE-to-CE relationship in regular BGP VPNs. However, as the vPE and end-host functions are co-resident in the same server, the connection between them is an internal implementation of the server.

```
             +-------------------------------------------------------+
             | +---------+ +---------+    +---------+ +---------+ |
             | |  VM1    | |  VM2    |    |  VM47   | |  VM48   | |
             | |(VPN Red)| |(VPN Grn)|... |(VPN Grn)| |(VPN Blu)| |
             | +----+----+ +---+-----+    +----+----+ +----+----+ |
             |      |          |               |           |      |
             |      +---+      | +-------------+       +---+      |
             |          |      | |                     |          |
        to   |      +---+------+-+---------------------+---+      |
        Gateway|     |   |      | |                     |   |      |
        PE   |      | +-+-+   ++-++            +---+ +-+-+ |      |
             |      | |VRF|   |VRF|   .......  |VRF| |VRF| |      |
        <----+------+ |Red|   |Grn|            |Yel| |Blu| |      |
             |      | +---+   +---+            +---+ +---+ |      |
             |      |           L3 VPN virtual PE          |      |
             |      +--------------------------------------+      |
             |                                                    |
             |                     End Device                     |
             +-------------------------------------------------------+
```

Figure 4. Server View of vPE to VM relationship

An application VM may send packets to a vPE forwarder that need to be bridged, either locally to another VM, or to a remote destination. In this case, the vPE contains a virtual bridge instance to which the application VMs (CEs) are attached.

```
                +-----------------------------------------------------+
                | +---------+ +---------+   +---------+               |
                | |  VM1    | |  VM2    |   |  VM47   |               |
                | |(VPN Red)| |(VPN Grn)|...|(VPN Grn)|               |
                | +----+----+ +----+----+   +----+----+               |
                |      |           |             |                    |
                |      +---+       +----+   +----+                    |
                |          |            |   |                         |
          to    |      +---+------------+---+-----------------+       |
          Gateway|      |   |            |   |                 |       |
          PE    |      | +-+--+--+    +-+---+-+               |       |
                |      | |VBridge|    |VBridge|   .......     |       |
          <-----+------+ |Red    |    |Grn    |               |       |
                |      | +-------+    +-------+               |       |
                |      |              vPE                     |       |
                |      +--------------------------------------+       |
                |                                                     |
                |                      End Device                     |
                +-----------------------------------------------------+
```

Figure 4. Bridging Service at vPE

3. Control Plane

3.1 vPE Control Plane (vPE-C)

3.1.1 The SDN approach

This approach is appropriate when the vPE control and data planes are physically decoupled. The control plane directing the data flow may reside elsewhere, e.g. in a SDN controller. This approach requires a standard interface to the routing system. The Interface to Routing System (I2RS) is work in progress in IETF as described in [I-D.ietf-i2rs-architecture], [I-D.ietf-i2rs-problem-statement].

Although MP-BGP is often the de facto preferred choice between vPE and gateway-PE/ASBR, the use of extensible signaling messaging protocols MAY often be more practical in a Data Center environment. One such proposal that uses this approach is detailed in [I-D.ietf-l3vpn-end-system].

3.1.2 Distributed control plane

In the distributed control plane approach, the vPE participates in the overlay L3/L2 VPN control protocol: MP-BGP [RFC4364].

When the vPE function is on a ToR, it participates the underlay routing through IGP protocols (ISIS or OSPF) or BGP.

When the vPE function is on a server, it functions as a host attached to a server.

3.3 Use of router reflector

Modern Data Centers can be very large in scale. For example, the number of VPNs routes in a very large DC can surpass the scale of those in a Service Provider backbone VPN networks. There may be tens of thousands of end devices in a single DC.

Use of Router Reflector (RR) is necessary in large-scale IP VPN networks to avoid a full iBGP mesh among all vPEs and PEs. The VPN routes can be partitioned to a set of RRs, the partitioning techniques are detailed in [RFC4364] and [I-D.ietf-l2vpn-evpn].

When a RR software instance is residing in a physical device, e.g., a server, which is partitioned to support multi-functions and application VMs, the RR becomes a virtualized RR (vRR). Since RR performs control functions only, a dedicated or virtualized server with large scale of computing power and memory can be a good candidate as host of vRRs. The vRR can also reside in a Gateway PE/ASBR, or in an end device.

3.4 Use of Constrained Route Distribution [RFC4684]

The Constrained Route Distribution [RFC4684] is a powerful tool for selective VPN route distribution. With RTC, only the BGP receivers (e.g, PE/vPE/RR/vRR/ASBRs, etc.) with the particular IP VPNs attached will receive the route update for the corresponding VPNs. It is critical to use constrained route distribution to support large-scale IP VPN developments.

4. Forwarding Plane

4.1 Virtual Interface

A Virtual Interface (VI) is an interface within an end device that is used for connection of the vPE to the application VMs in the same end device. Such application VMs are treated as CEs in the regular VPN’s view.

4.2 Virtual Provider Edge Forwarder (vPE-F)

The Virtual Provider Edge Forwarder (vPE-F) is the forwarding component of a vPE where the tenant identifiers (for example, MPLS VPN labels) are pushed/popped.

The vPE-F location options include:

1) Within the end device where the virtual interface and application VMs are located.

2) In an external device such as a Top of the Rack switch (ToR) in a DC into which the end device connects.

Multiple factors should be considered for the location of the vPE-F, including device capabilities, overall solution economics, QoS/firewall/NAT placement, optimal forwarding, latency and performance, operational impact, etc. There are design tradeoffs, it is worth the effort to study the traffic pattern and forwarding looking trend in your own unique Data Center as part of the exercise.

4.3 Encapsulation

BGP/MPLS VPNs can be tunneled through the network as overlays using MPLS-based or IP-based encapsulation.

In the case of MPLS-based encapsulation, most existing core deployments use distributed protocols such as Label Distribution Protocol (LDP), [RFC3032][RFC5036], or RSVP-TE [RFC3209].

Due to its maturity, scalability, and header efficiency, MPLS Label Stacking is gaining traction by service providers, and large-scale cloud providers in particular, as the unified forwarding mechanism of choice.

With the emergence of the SDN paradigm, label distribution may be achieved through SDN controllers, or via a combination of centralized control and distributed protocols.

In the case of IP-based encapsulation, MPLS VPN packets are encapsulated in IP or Generic Routing Encapsulation (GRE), [RFC4023], [RFC4797]. IP-based encapsulation has not been extensively deployed for BGP/MPLS VPN in the core; however it is considered as one of the tunneling options for carrying MPLS VPN overlays in the data center. Note that when IP encapsulation is used, the associated security properties must be analyzed carefully.

4.4 Optimal forwarding

Many large cloud service providers have reported the DC traffic is now dominated by East-West across subnet traffic (between the end device hosting different applications in different subnets) rather than North-South traffic (going in/out of the Data Center and to/from the WAN) or switched traffic within subnets. This is the primary reason that newer DC design has moved away from traditional Layer-2 design to Layer-3, especially for the overlay networks.

When forwarding the traffic within the same VPN, the vPE SHOULD be capable to provide direct communication among the VMs/application senders/receivers without the need of going through Gateway devices. If the senders and the receivers are on the same end device, the traffic SHOULD NOT need to leave the device. If they are on different end devices, optimal routing SHOULD be applied.

Extranet MPLS VPN techniques can be used for multiple VPNs access without the need of Gateway facilitation. This is done through the use of VPN policy control mechanisms.

In addition, ECMP is a built in IP mechanism for load sharing. Optimal use of available bandwidth can be achieved by virtue of using ECMP in the underlay, as long as the encapsulation includes certain entropy in the header, VXLAN is such an example.

4.5 Routing and Bridging Services

A VPN forwarder (vPE-F) may support both IP forwarding as well as Layer 2 bridging for traffic from attached end hosts. This traffic may be between end hosts attached to the same VPN forwarder or to different VPN forwarders.

In both cases, forwarding at a VPN forwarder takes place based on the IP or MAC entries provisioned by the vPE controller.

When the vPE is providing Layer 3 service to the attached CEs, the VPN forwarder has a VPN VRF instance with IP routes installed for both locally attached end-hosts and ones reachable via other VPN forwarders. The vPE may perform IP routing for all IP packets in this mode.

When the vPE provides Layer 2 service to the attached end-hosts, the VPN forwarder has an E-VPN instance with appropriate MAC entries.

The vPE may support an Integrated Routing and Bridging service, in which case the relevant VPN forwarders will have both MAC and IP table entries installed, and will appropriately route or switch incoming packets.

The vPE controller performs the necessary provisioning functions to support various services, as defined by an user.

5. Addressing

5.1 IPv4 and IPv6 support

IPv4 and IPv6 MUST be supported in the vPE solution.

This may present a challenge for older devices, but this normally is not an issue for the newer generation of forwarding devices and servers. Note that a server is replaced much more frequently than a network router/switch, and newer equipment SHOULD be capable of IPv6 support.

5.2 Address space separation

The addresses used for the IP VPN overlay in a DC, SHOULD be taken from separate address blocks outside the ones used for the underlay infrastructure of the DC. This practice is to protect the DC infrastructure from being attacked if the attacker gains access to the tenant VPNs.

Similarity, the addresses used for the DC SHOULD be separated from the WAN backbone addresses space.

6.0 Inter-connection considerations

The inter-connection considerations in this section are focused on intra-DC inter-connections.

There are deployment scenarios where BGP/MPLS IP VPN may not be supported in every segment of the networks to provide end-to-end IP VPN connectivity. A vPE may be reachable only via an intermediate inter-connecting network; interconnection may be needed in these cases.

When multiple technologies are employed in the solution, a clear demarcation should be preserved at the inter-connecting points. The problems encountered in one domain SHOULD NOT impact other domains.

From an IP VPN point of view: An IP VPN vPE that implements [RFC4364] is a component of the IP VPN network only. An IP VPN VRF on a physical PE or vPE contains IP routes only, including routes learnt over the locally attached network.

The IP VPN vPE should ideally be located as close to the "customer" edge devices as possible. When this is not possible, simple existing

"IP VPN CE connectivity" mechanisms should be used, such as static, or direct VM attachments such as described in the vCE [I-D.fang-l3vpn-virtual-ce] option below.

Consider the following scenarios when BGP MPLS VPN technology is considered as whole or partial deployment:

Scenario 1: All VPN sites (CEs/VMs) support IP connectivity. The most suited BGP solution is to use IP VPNs [RFC4364] for all sites with PE and/or vPE solutions.

Scenario 2: Legacy Layer 2 connectivity must be supported in certain sites/CEs/VMs, and the rest of the sites/CEs/VMs need only Layer 3 connectivity.

One can consider using a combined vPE and vCE [I-D.fang-l3vpn-virtual-ce] solution to solve the problem. Use IP VPN for all sites with IP connectivity, and a physical or virtual CE (vCE, may reside on the end device) to aggregate the Layer 2 sites which for example, are in a single container in a Data Center. The CE/vCE can be considered as inter-connecting points, where the Layer 2 network is terminated and the corresponding routes for connectivity of the L2 network are inserted into IP VPN VRFs. The Layer 2 aspect is transparent to the L3VPN in this case.

Reducing operation complicity and maintaining the robustness of the solution are the primary reasons for the recommendations.

The interconnection of MPLS VPN in the data center and the MPLS core through ASBR using existing inter-AS options is discussed in detail in [I-D.fang-l3vpn-data-center-interconnect].

7. Management, Control, and Orchestration

7.1 Assumptions

The discussion in this section is based on the following set of assumptions:

- The WAN and the inter-connecting Data Center, MAY be under control of separate administrative domains

- WAN Gateways/ASBRs/PEs are provisioned by existing WAN provisioning systems

- If a single Gateway/ASBR/PE connecting to the WAN on one side, and connecting to the Data Center network on the other side, then this Gateway/ASBR/PE is the demarcation point between the two networks.

- vPEs and VMs are provisioned by Data Center Orchestration systems.

- Managing IP VPNs in the WAN is not within the scope of this document except the inter-connection points.

7.2 Management/Orchestration system interfaces

The Management/Orchestration system CAN be used to communicate with both the DC Gateway/ASBR, and the end devices.

The Management/Orchestration system MUST support standard, programmatic interface for full-duplex, streaming state transfer in and out of the routing system at the Gateway.

The programmatic interface is currently under definition in IETF Interface to Routing Systems (I2RS)) initiative. [I-D.ietf-i2rs-architecture], and [I-D.ietf-i2rs-problem-statement].

Standard data modeling languages will be defined/identified in I2RS. YANG - A Data Modeling Language for the Network Configuration Protocol (NETCONF) [RFC6020] is a promising candidate currently under investigation.

To support remote access between applications running on an end device (e.g., a server) and routers in the network (e.g. the DC Gateway), a standard mechanism is expected to be identified and defined in I2RS to provide the transfer syntax, as defined by a protocol, for communication between the application and the network/routing systems. The protocol(s) SHOULD be lightweight and familiar by the computing communities. Candidate examples include ReSTful web services, JSON [RFC7159], NETCONF [RFC6241], XMPP [RFC6120], and XML. [I-D.ietf-i2rs-architecture].

7.3 Service VM Management

Service VM Management SHOULD be hypervisor agnostic, e.g. On demand service VMs turning-up SHOULD be supported.

7.4 Orchestration and MPLS VPN inter-provisioning

The orchestration system

1) MUST support MPLS VPN service activation in virtualized DC.

2) MUST support automated cross-provisioning accounting correlation between the WAN MPLS VPN and Data Center for the same tenant.

3) MUST support automated cross provisioning state correlation

between WAN MPLS VPN and Data Center for the same tenant

There are two primary approaches for IP VPN provisioning - push and pull, both CAN be used for provisioning/orchestration.

7.4.1 vPE Push model

Push model: push IP VPN provisioning from management/orchestration systems to the IP VPN network elements.

This approach supports service activation and it is commonly used in existing MPLS VPN Enterprise deployments. When extending existing WAN IP VPN solutions into the a Data Center, it MUST support off-line accounting correlation between the WAN MPLS VPN and the cloud/DC MPLS VPN for the tenant. The systems SHOULD be able to bind interface accounting to particular tenant. It MAY requires offline state correlation as well, for example, binding of interface state to tenant.

Provisioning the vPE solution:

1) Provisioning process

   a. The WAN provisioning system periodically provides to the DC orchestration system the VPN tenant and RT context.
   b. DC orchestration system configures vPE on a per request basis

2) Auto state correlation

3) Inter-connection options:

   Inter-AS options defined in [RFC4364] may or may not be sufficient for a given inter-connection scenario. BGP IP VPN inter-connection with the Data Center is discussed in [I-D.fang-l3vpn-data-center-interconnect].

   This model requires offline accounting correlation

   1) Cloud/DC orchestration configures vPE

   2) Orchestration initiates WAN IP VPN provisioning; passes connection IDs (e.g., of VLAN/VXLAN) and tenant context to WAN IP VPN provisioning systems.

   3) WAN MPLS VPN provisioning system provisions PE VRF and policies as in typical Enterprise IP VPN provisioning processes.

   4) Cloud/DC Orchestration system or WAN IP VPN provisioning system

MUST have the knowledge of the connection topology between the DC and WAN, including the particular interfaces on core router and connecting interfaces on the DC PE and/or vPE.

In short, this approach requires off-line accounting correlation and state correlation, and requires per WAN Service Provider integration.

Dynamic BGP sessions between PE/vPE and vCE MAY be used to automate the PE provisioning in the PE-vCE model, that will remove the needs for PE configuration. Caution: This is only under the assumption that the DC provisioning system is trusted and can support dynamic establishment of PE-vCE BGP neighbor relationships, for example, the WAN network and the cloud/DC belong to the same Service Provider.

7.4.2 vPE Pull model

Pull model: pull from network elements to network management/AAA based upon data plane or control plane activity. It supports service activation. This approach is often used in broadband deployments. Dynamic accounting correlation and dynamic state correlation are supported. For example, session based accounting is implicitly includes tenant context state correlation, as well as session-based state that implicitly includes tenant context. Note that the pull model is less common for vPE deployment solutions.

Provisioning process:

1) Cloud/DC orchestration configures vPE

2) Orchestration primes WAN MPLS VPN provisioning/AAA for new service, passes connection IDs (e.g., VLAN/VXLAN) and tenant context.

3) Cloud/DC ASBR detects new VLAN and sends Radius Access-Request (or Diameter Base Protocol request message [RFC6733]).

4) Radius Access-Accept (or Diameter Answer) with VRF and other policies

Auto accounting correlation and auto state correlation is supported.

8. Security Considerations

As vPE is an extended BGP/MPLS VPN solution, security threats and defense techniques described in RFC 4111 [RFC4111] generally apply.

When the SDN approach is used, the protocols between the vPE agent and the vPE-C in the controller MUST be mutually authenticated. Given the potentially very large scale and the dynamic nature in the cloud/DC environment, the choice of key management mechanisms need to be further studied.

VMs in the servers can belong to different tenants with different characteristics depending on the application. Classification of the VMs must be done through the orchestration system and appropriate security policies must be applied based on such classification before turning on the services.

9. IANA Considerations

None.

10. Acknowledgments

The authors would like to thank Daniel Voyer for his review and comments.

11. References

11.1 Normative References

[RFC2119] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, March 1997.

[RFC3032] Rosen, E., Tappan, D., Fedorkow, G., Rekhter, Y., Farinacci, D., Li, T., and A. Conta, "MPLS Label Stack Encoding", RFC 3032, January 2001.

[RFC3209] Awduche, D., et al., "RSVP-TE: Extension to RSVP for LSP Tunnels", RFC 3209, December 2001.

[RFC4023] Worster, T., Rekhter, Y., and E. Rosen, Ed., "Encapsulating MPLS in IP or Generic Routing Encapsulation (GRE)", RFC 4023, March 2005.

[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.

[RFC5036] Andersson, L., Ed., Minei, I., Ed., and B. Thomas, Ed., "LDP Specification", RFC 5036, October 2007.

[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.

[RFC6241] Enns, R., Ed., Bjorklund, M., Ed., Schoenwaelder, J., Ed., and A. Bierman, Ed., "Network Configuration Protocol (NETCONF)", RFC 6241, June 2011.

[RFC6733] Fajardo, V., Ed., Arkko, J., Loughney, J., and G. Zorn, Ed., "Diameter Base Protocol", RFC 6733, October 2012.

11.2 Informative References

[RFC4111] Fang, L., Ed., "Security Framework for Provider-Provisioned Virtual Private Networks (PPVPNs)", RFC 4111, July 2005.

[RFC7159] Bray, T., "The JavaScript Object Notation (JSON) Data Interchange Format", RFC 7159, March 2014.

[RFC4797] Rekhter, Y., Bonica, R., and E. Rosen, "Use of Provider Edge to Provider Edge (PE-PE) Generic Routing Encapsulation (GRE) or IP in BGP/MPLS IP Virtual Private Networks", RFC 4797, January 2007.

[I-D.ietf-l3vpn-end-system] Marques, P., Fang, L., Pan, P., Shukla, A., Napierala, M., Bitar, N., "BGP-signaled end-system IP/VPNs", draft-ietf-l3vpn-end-system, work in progress.

[I-D.rfernando-l3vpn-service-chaining] Fernando, R., Rao, D., Fang, L., Napierala, M., So, N., draft-rfernando-l3vpn-service-chaining, work in progress.

[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, work in progress.

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

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

[I-D.bitar-i2rs-service-chaining] Bitar, N., Geron, G., Fang, L., Krishnan, R., Leymann, N., Shah, H., Chakrabarti, S., Haddad, W., draft-bitar-i2rs-service-chaining, work in progress.

[I-D.fang-l3vpn-data-center-interconnect] Fang, L., Fernando, R., Rao, D., Boutros, S., "BGP IP VPN Data Center Interconnect", draft-fang-l3vpn-data-center-interconnect, work in progress.

[I-D.ietf-l2vpn-evpn] Sajassi, A., et al., "BGP MPLS Based Ethernet VPN", draft-ietf-l2vpn-evpn, work in progress.

Authors' Addresses

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

David Ward
Cisco
170 W Tasman Dr
San Jose, CA 95134
Email: wardd@cisco.com

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

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

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

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

Bruno Rijsman
Juniper Networks
10 Technology Park Drive
Westford, MA 01886
Email: brijsman@juniper.net

Ning So
Vinci Systems
Plano, TX 75082, USA
Email: ning.so@vinci-systems.com

Jim Guichard
Cisco
Boxborough, MA 01719
Email: jguichar@cisco.com

Wen Wang
CenturyLink
2355 Dulles Corner Blvd.
Herndon, VA 20171
Email:Wen.Wang@CenturyLink.com

Manuel Paul
Deutsche Telekom
Winterfeldtstr. 21-27
10781 Berlin, Germany
Email: manuel.paul@telekom.de

Wim Henderichx
Alcatel-Lucent

Email: wim.henderichx@alcatel-lucent.com

L3VPN
Internet Draft
Intended status: Informational
Expires: January 2015

Weiguo Hao
Lucy Yong
Huawei
S. Hares
Hickory Hill Consulting

July 1, 2014

Inter-AS Option B between NVO3 and BGP/MPLS IP VPN network
draft-hao-l3vpn-inter-nvo3-vpn-00.txt

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/ietf/1id-abstracts.txt.

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

This Internet-Draft will expire on January 1, 2015.

Copyright Notice

Abstract

This draft describes option-B inter-as connection between NVO3 network and MPLS/IP VPN network. Comparing to traditional Option-B inter-as connection defined in [RFC 4364], this draft provides enhancement for heterogeneous network multi-as connection, the control plane and data plane procedures in NVO3 network are newly designed.

Table of Contents

1. Introduction

In cloud computing era, multi-tenancy has become a core requirement for data centers. Since NVO3 can satisfy multi-tenancy key requirements, this technology is being deployed in an increasing number of cloud data center network. NVO3 focuses on the construction of overlay networks that operate over an IP (L3)

underlay transport network. It can provide layer 2 bridging and layer 3 IP service for each tenant. VXLAN and NVGRE are two typical NVO3 technologies. NVO3 overlay network can be controlled through centralized NVE-NVA architecture or through distributed BGP VPN protocol.

NVO3 has good scaling properties from relatively small networks to networks with several million tenant systems (TSs) and hundreds of thousands of virtual networks within a single administrative domain. In NVO3 network, 24-bit VN ID is used to identify different virtual networks, theoretically 16M virtual networks can be supported in a data center. In a data center network, each tenant may include one or more layer 2 virtual network and in normal cases each tenant corresponds to one routing domain (RD). Normally each layer 2 virtual network corresponds to one or more subnets.

To provide cloud service to external data center client, data center networks should be connected with WAN networks. BGP MPLS/IP VPN has already been widely deployed at WAN networks. Normally internal data center and external MPLS/IP VPN network belongs to different autonomous system(AS). This requires the setting up of inter-as connections at Autonomous System Border Routers(ASBRs) between NVO3 network and external MPLS/IP network.

Currently, a typical connection mechanism between a data center network and an MPLS/IP VPN network is similar to Inter-AS Option-A of RFC4364, but it has scalability issue if there is huge number of tenants in data center networks. To overcome the issue, inter-as Option-B between NVO3 network and BGP MPLS/IP VPN network is proposed in this draft.

2. Conventions used in this document

Network Virtualization Edge (NVE) -An NVE is the network entity that sits at the edge of an underlay network and implements network virtualization functions.

Tenant System - A physical or virtual system that can play the role of a host, or a forwarding element such as a router, switch, firewall, etc. It belongs to a single tenant and connects to one or more VNs of that tenant.

VN - A VN is a logical abstraction of a physical network that provides L2 network services to a set of Tenant Systems.

RD - Route Distinguisher. RDs are used to maintain uniqueness among identical routes in different VRFs, The route distinguisher is an 8-

octet field prefixed to the customer’s IP address. The resulting 12-octet field is a unique "VPN-IPv4" address.

RT - Route targets. It is used to control the import and export of routes between different VRFs.

3. Reference model

```
|-----------------------------------------------------|
|  ------           AS1                               |
|  | TS1| -                                           |
|  ------  -                                          |
|            - ------    ------                       |
|            - |NVE1| -- |TOR1|----------------      |
|  ------  -   ------    ------                |     |
|  | TS2|-                                      |     |
|  ------                                       |     |
|                                            --------  |
|                            |------------ | ASBR1|  |
|  ------                    |             --------  |
|  | TS3| -                  |                 |     |
|  ------  -                 |                 |     |
|            - ------    ------               |     |
|            - |NVE2| -- |TOR2|               |     |
|  ------  -   ------    ------               |     |
|  | TS4|-                                      |     |
|  ------                                       |     |
|-------------------------------------------------------
|                                               |     |
|  ------                                       |     |
|  | CE1| -                                     |     |
|  ------  -                                    |     |
|            - ------                       --------  |
|            - | PE1| --------------------| ASBR2|  |
|   ------  -  ------                       --------  |
|   | CE2|-                                           |
|   ------          AS2                               |
|-----------------------------------------------------|
```

Figure 1 Reference model

Figure 1 shows an arbitrary Multi-AS VPN interconnectivity scenario between NVO3 network and BGP MPLS/IP VPN network. NVE1, NVE2, and ASBR1 forms NVO3 overlay network in internal DC. TS1 and TS2 connect to NVE1, TS3 and TS4 connect to NVE2. PE1 and ASBR2 forms MPLS

IP/VPN network in external DC. CE1 and CE2 connect to PE1. The NVO3 network belongs to AS 1, the MPLS/IP VPN network belongs to AS 2.

There are two tenants in NVO3 network, TSs in tenant 1 can freely communicate with CEs in VPN-Red, TSs in tenant 2 can freely communicate with CEs in VPN-Green. TS1 and TS3 belong to tenant 1, TS2 and TS4 belong to tenant 2. CE1 belongs to VPN-Red , CE2 belongs to VPN-Green.

4. Option-A inter-as solution overview

In Option-A inter-as solution, peering ASBRs are connected by multiple sub-interfaces, each ASBR acts as a PE, and thinks that the other ASBR is a CE. Virtual routing and forwarding (VRF)data bases (RIB/FIB) are configured at AS border routers (ASBR1 and ASBR2) so that each ASBRs associate each such sub-interface with a VRF and use EBGP to distribute unlabeled IPv4 addresses to each other. In the data-plane, VLANs are used for tenant traffic separation. In normal case ASBR1 also acts as NVO3 layer 3 gateway, it can terminate NVO3 encapsulation for inter-subnet traffic between TS in internal DC and CE in external DC.

For the traffic from internal DC to external DC, the forwarding process of ASBR1 at internal DC is as follows:

1. Terminates NVO3 encapsulation and gets VN ID.

2. Finds corresponding VRF relying on the VN ID.

3. Looks up IP forwarding table in the VRF, and then forwards the traffic to a sub-interface connecting to peer ASBR.

The forwarding process of ASBR2 at MPLS/VPN network is as follows:

1. Finds corresponding VRF based on sub-interface.

2. Looks up IP forwarding table in the VRF, encapsulates the traffic with MPLS VPN label, and then sends the traffic to MPLS VPN network.

For the traffic from external DC to internal DC, the traffic forwarding process is similar to the above process.

Option-A inter-as solution has following issues:

1. Up to 16 million (16M) gateway interfaces (virtual/physical) and 16M EBGP session need to exist between the ASBRs.

2. UP to 16M VRFs need to be supported on border routers.

3. Several million routing entries need to be supported on border routers.

Inter-as option B between NVO3 network and MPLS IP/VPN network can be used to address these issues. Due to it is for multi-as interconnection between heterogeneous networks, so there are some differences from traditional Inter-AS Option-B of RFC4364.

5. Option-B inter-as solution overview

Similar to the solution described in section 10, part (b) of [RFC4364] (commonly referred to as Option-B) peering ASBRs are connected by one or more sub-interfaces that are enabled to receive MPLS traffic. An MP-BGP session is used to distribute the labeled VPN prefixes between the ASBRs. In data plane, the traffic that flows between the ASBRs is placed upon MPLS tunnels, traffic separation among different VPNs between the ASBRs relies on MPLS VPN Label.

In this solution, the procedures in MPLS/IP VPN network are same as defined in [RFC4364], but the procedures in NVO3 network need to be newly designed to support inter-as Option-B.

The advantage of this option is that it’s more scalable, as there is no need to have one sub-interface and BGP session per VPN/Tenant.

6. Inter-As Option-B procedures

The TS and CE information in above figure 1 are as follows:

TS	Tenant	IP Address	VN ID
TS1	1	10.1.1.2	10
TS2	2	20.1.1.2	20
TS3	1	10.1.1.3	10
TS4	2	20.1.1.3	20

Table 1 TS information in NVO3 network

CE	Route Distinguisher	Route Target	IP Address	Mask
CE1	VPN-Red1	1:1	30.1.1.1	24
CE2	VPN-Green1	2:2	40.1.1.1	24

Table 2 CE information in MPLS/IP VPN network

Section 6.1 below describes the route distribution process for this option, and section 6.2 describes the data forwarding process.

NVO3 network can pass routing data for the NVEs (IP Address, VN ID) through either: a) RFC 4364 running between the NVEs and the ASBR1, or b) NVE-NVA architecture. Therefore, the routing distribution process is different for these two options. Section 6.1.1 describes the routing distribution procedures using RFC 4364 on NVO3 network, and section 6.1.2 describes the procedures using NVE-NVA architecture.

The Data plane process is same in these two cases.

6.1. Routing distribution procedures

6.1.1. Using RFC 4364

The route distribution in NVO3 network makes use of the BGP multiple tunnel identifiers [BGP Remote-Next-Hop] to create an RFC4364 Option-B solution. This section provides a step by step explanation of the process.

In internal DC network, VRF1 and VRF2 are created on NVE1 and NVE2 to isolate IP forwarding process between tenant 1 and tenant 2. Route distinguishers (RD) and RT are specified for each VRF on these

NVEs. BGP MPLS/IP VPN protocol extension is running between NVEs and ASBR1 utilizing the [BGP Remote-Next-Hop] which describes the BGP MPLS/IP VPN protocol extension detail to specify a set of remote tunnels (1 to N) that occur between two BGP speakers. The VRF configuration information on each NVE are as follows:

NVE	Tenant	Route Distinguisher	Route Target
NVE1	1	VPN-Red2	1:1
NVE1	2	VPN-Green2	2:2
NVE2	1	VPN-Red3	1:1
NVE2	2	VPN-Green3	2:2

6.1.1.1. Internal DC to external DC direction

1. NVE1 and NVE2 operate as a layer 3 gateway for local connecting TS. NVE1 and NVE2 learn the local TS's IP Address via ARP, and advertise this information to the ASBR1. The routing information from NVE1 and NVE2 are as follows:

NVE	RD:IP Prefix	Route Target	VN ID
NVE1	VPN-Red2:10.1.1.2/32	1:1	10
NVE1	VPN-Green2:20.1.1.2/32	2:2	20
NVE2	VPN-Red3:10.1.1.3/32	1:1	10
NVE2	VPN-Green3:20.1.1.3/32	2:2	20

Routing information sent form NVE1 and NVE2

2. ASBR1 allocates MPLS VPN Label per tenant (VN ID) per NVE and the RD and RT remain the same. Then the ASBR1 advertises the VPN route with new allocated MPLS VPN Label to ASBR2. The allocated MPLS VPN label and its corresponding NVE+VN ID forms incoming forwarding table which is used to forward MPLS traffic from external DC to internal DC. The incoming forwarding table on ASBR1 is as follows:

MPLS VPN Label	NVE + VN ID
1000	NVE1 + 10
2000	NVE1 + 20
1001	NVE2 + 10
2001	NVE2 + 20

Incoming forwarding table

3. ASBR2 allocates new local VPN label for each receiving VPN label from ASBR1 firstly, then ASBR2 advertises the VPN route with new allocated MPLS VPN Label to PE1. As for data plane forwarding process, the new local VPN label are in VPN label, the receiving VPN label from ASBR1 are out VPN label. The VPN label switch table on ASBR2 is as follows:

In VPN Labels	Out VPN Labels
1000	1000
2000	2000
1001	1001
2001	2001

VPN label switch table

4. PE1 matches the Route Target Attribute in BGP MPLS/IP VPN protocol with local VRF’s import RT configuration. Then it populates local VRF with these matched VPN routes. The routing tables of VPN-Red and VPN-Green are as follows:

IP Prefix	VPN Label	BGP Peer
10.1.1.2/32	1000	ASBR2
10.1.1.3/32	1001	ASBR2

Routing table in VPN-Red

IP Prefix	VPN Label	BGP Peer
20.1.1.2/32	2000	ASBR2
20.1.1.3/32	2001	ASBR2

Routing table in VPN-Green

6.1.1.2. External DC to internal DC direction

1. PE1 learns IP prefix from CE1 and CE2 and populates these IP prefix to local VRF. Then the PE allocates VPN Label for these IP prefix and announces VPN routing information with allocated VPN Label to ASBR1. The VPN routing information are as follows:

Route Target	IP Prefix	VPN Label
1:1	VPN-Red1:30.1.1.1/24	3000
2:2	VPN-Green1:40.1.1.1/24	4000

2. ASBR2 allocates new local VPN label for each receiving VPN label from PE1, then ASBR2 advertises the VPN route with new allocated MPLS VPN Label to ASBR1. As for data plane forwarding process, the new local VPN label are in VPN label, the receiving VPN label from ASBR1 are out VPN label. The VPN label switch table on ASBR2 is as follows:

In VPN Label	Out VPN Label
3000	3000
4000	4000

3. ASBR1 allocates VN ID for each VPN Label receiving from ASBR2, and then ASBR2 advertises the VPN route with new allocated VN ID to each NVE (NVE1 and NVE2). The role of the VN ID is similar to the role of In VPN Label in ASBR1, it has local significance on ASBR1, each VN ID corresponds to per MPLS VPN Label on peer ASBR2; The VN ID space should be assigned in beforehand and should be orthogonal to the VN ID space for tenant identification(for example, assuming ASBR1 has local connecting TSs of tenant 1 to tenant 100, VN ID 1 to 100 are allocated for these tenants, other VN ID other than 1 to 100 can be allocated for outgoing forwarding table purpose). The allocated VN ID and its corresponding out VPN Label forms an outgoing forwarding table which is used to forward NVO3 traffic from internal DC to external DC. The outgoing forwarding table on ASBR1 is as follows:

VN ID	Out VPN Label
10000	3000
10001	4000

Outgoing forwarding table

4. NVE1 matches the Route Target Attribute in BGP MPLS/IP VPN protocol with local VRF’s import RT configuration. Then it populates local VRF with these matched VPN routes. The routing tables of tenant 1 and tenant 2 are as follows:

```
+------------------+--------------------+--------------------+
|   IP Prefix      |        VN ID       |     Dest Outer IP  |
+------------------+--------------------+--------------------+
|   30.1.1.1/24    |        10000       |        ASBR1       |
+------------------+--------------------+--------------------+
       Tenant 1 routing table on NVE1 and NVE2
+------------------+--------------------+--------------------+
|   IP Prefix      |        VN ID       |     Dest Outer IP  |
+------------------+--------------------+--------------------+
|   40.1.1.1/24    |        10001       |        ASBR1       |
+------------------+--------------------+--------------------+
       Tenant 2 routing table on NVE1 and NVE2
```

6.1.2. NVE-NVA architecture

No distributed BGP VPN protocol (RFC4364) is running on all NVEs and ASBR1 in NVO3 network, NVEs and ASBR1 are controlled by centralized NVA. The NVA runs EBGP VPN protocol with peer ASBR2 and exchanges VPN routing information between NVO3 network and MPLS/IP VPN network.

NVA maintains tenant information collected from all tenants. This information includes VN ID to identify each tenant and the corresponding RD and RT. This information can be statically configured by operators or dynamically notified by cloud management systems.

NVA also maintains all TS’s MAC/IP address and its attached NVE information for each tenant.

```
                 ------     EBGP      --------
                 |NVA | ------------- |ASBR2 |
                 ------               --------
                   .
                   . Southbound interface(Openflow,OVSDB,etc)
      .......................
      .          .           .
      .          .           .
      .          .           .
   ------     ------       -------
   |NVE1|     |NVE2|       |ASBR1|
   ------     ------       -------
```

Figure 2 NVE-NVA Archetecture

6.1.2.1. Internal DC to external DC direction

1. NVA allocates MPLS VPN Label per tenant per NVE.

2. NVA advertises all internal data center VPN routing information to peer ASBR2, which includes RD, IP prefix, RT, and MPLS VPN Label.

3. NVA downloads incoming forwarding table to ASBR1.

6.1.2.2. External DC to internal DC direction

1. NVA receives VPN routing information from peer ASBR2.

2. NVA allocates VN ID for each MPLS VPN Label receiving from ASBR2.

3. NVA downloads outgoing forwarding table to ASBR1.

4. NVA matches local Route Target configuration, imports VPN route to each tenant, and downloads routing table to corresponding NVE.

6.2. Data plane procedures

This section describes the step by step procedures of data forward for either: a) internal DC to external DC IP data flows, or b) the external DC to internal DC IP data flows.

6.2.1. Internal DC to external DC direction

1. TS1 sends traffic to NVE1, the destination IP is CE1’s IP address of 30.1.1.1.

2. NVE1 looks up VRF1’s IP forwarding table, then it gets NVO3 tunnel encapsulation information. The destination outer address is ASBR1’s IP address, VN ID is 10000. NVE1 performs NVO3 encapsulation and sends the traffic to ASBR1.

3. ASBR1 decapsulates NVO3 encapsulation and gets VN ID 10000. Then it looks up outgoing forwarding table based on the VN ID and gets MPLS VPN label 3000. Finally it pushes MPLS VPN label for the IP traffic and sends it to ASBR2.

4. ASBR2 swaps VPN Label, then sends the traffic to PE1 through IGP tunnel.

5. PE1 terminates IGP tunnel, pops MPLS VPN label 3000, looks up local IP forwarding table in VRF1, and then forwards the traffic to CE1.

6.2.2. External DC to internal DC direction

1. CE1 sends traffic to PE1, destination IP is TS1's IP address of 10.1.1.2.

2. PE1 looks up local IP forwarding table in VPN-RED, pushes MPLS VPN label 1000, then searches IGP tunnel, then the PE sends the traffic to ASBR2 through IGP tunnel.

3. ASBR2 terminates IGP tunnel, swaps MPLS VPN label, then sends the traffic to ASBR1.

4. ASBR1 looks up incoming forwarding table and gets NVO3 encapsulation, then performs NVO3 encapsulation and sends the traffic to NVE1. The destination outer IP is NVE1's IP, VN ID is 10.

5. NVE1 decapsulates NVO3 encapsulation, gets local VRF1 relying on VN ID 10, looks up local IP forwarding table in VRF1, then sends the traffic to TS1.

7. Security Considerations

Similar to the security considerations for inter-as Option-B in [RFC4364] the appropriate trust relationship must exist between NVO3 network and MPLS/IP VPN network. VPN-IPv4 routes in NVO3 network should neither be distributed to nor accepted from the public Internet, or from any BGP peers that are not trusted. For other general VPN Security Considerations, see [RFC4364].

8. IANA Considerations

This document requires no IANA actions. RFC Editor: Please remove this section before publication.

9. References

9.1. Normative References

[1] [RFC2119] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, March 1997.

9.2. Informative References

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

[2] [NVA] D.Black, etc, "An Architecture for Overlay Networks (NVO3)", draft-ietf-nvo3-arch-01, February 14, 2014

[3] [BGP Remote-Next-Hop] G. Van de Velde,etc, "BGP Remote-Next-Hop", draft-vandevelde-idr-remote-next-hop-05, January, 2014

[4] [RFC7047] B. Pfaff, B. Davie,"The Open vSwitch Database Management Protocol", RFC 7047, December 2013

[5] [OpenFlow1.3]OpenFlow Switch Specification Version 1.3.0 (Wire Protocol 0x04). June 25, 2012. (https://www.opennetworking.org/images/stories/downloads/sdn-resources/onf-specifications/openflow/openflow-spec-v1.3.0.pdf)

10. Acknowledgments

Authors like to thank Xiaohu Xu, Liang Xia, Shunwan Zhang for his valuable inputs.

Authors' Addresses

Weiguo Hao
Huawei Technologies
101 Software Avenue,
Nanjing 210012
China
Phone: +86-25-56623144
Email: haoweiguo@huawei.com

Lucy Yong
Huawei Technologies
Phone: +1-918-808-1918
Email: lucy.yong@huawei.com

Susan Hares
Huawei Technologies
Phone: +1-734-604-0323
Email: shares@ndzh.com.

This Internet-Draft, draft-morin-l3vpn-mvpn-fast-failover-05.txt, has expired, and has been deleted from the Internet-Drafts directory. An Internet-Draft expires 185 days from the date that it is posted unless it is replaced by an updated version, or the Secretariat has been notified that the document is under official review by the IESG or has been passed to the RFC Editor for review and/or publication as an RFC. This Internet-Draft was not published as an RFC.

Internet-Drafts are not archival documents, and copies of Internet-Drafts that have been deleted from the directory are not available. The Secretariat does not have any information regarding the future plans of the authors or working group, if applicable, with respect to this deleted Internet-Draft. For more information, or to request a copy of the document, please contact the authors directly.

Draft Authors:
Thomas Morin<thomas.morin@orange-ftgroup.com>
Yakov Rekhter<yakov@juniper.net>
Rahul Aggarwal<rahul@juniper.net>
Wim Henderickx<wim.henderickx@alcatel-lucent.com>
Praveen Muley<Praveen.muley@alcatel-lucent.com>
Ray Qiu<rayq@huawei.com>

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.

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).

```
"""""""""""""""""""""""                         """""""""""""""""""""""
"          +-------+ " +--------+   +--------+ " +-------+          "
" +-----+  | vPE-1 | " | S-PE-A |   | S-PE-B | " | vPE-2 |  +-----+ "
" |VM/CE|--|       |---|        |---|        |---|       |--|VM/CE| "
" +-----+  |(VRF-1)| " |(VRF-A) |   |(VRF-B) | " |(VRF-2)|  +-----+ "
"          +-------+ " +--------+   +--------+ " +-------+          "
"                    "      |            |     "                    "
"     Zone 1         " +--------+   +--------+ "       Zone 2       "
""""""""""""""""""""""" | Serv-A |   | Serv-B | """""""""""""""""""""""
                        +--------+   +--------+
```

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

- o 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.

- o 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.

- o 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

L3VPN Working Group
Internet Draft
Intended Status: Standards Track
Updates: 6513,6514
Expires: October 11, 2014

Eric C. Rosen
Karthik Subramanian
Cisco Systems, Inc.

Jeffrey Zhang
Juniper Networks, Inc.

April 11, 2014

Ingress Replication Tunnels in Multicast VPN

draft-rosen-l3vpn-ir-01.txt

Abstract

RFCs 6513, 6514, and other RFCs describe procedures by which a Service Provider may offer Multicast VPN service to its customers. These procedures create point-to-multipoint (P2MP) or multipoint-to-multipoint trees across the Service Provider’s backbone. One type of P2MP tree that may be used is known as an "Ingress Replication (IR) tunnel". In an IR tunnel, a parent node need not be "directly connected" to its child nodes. When a parent node has to send a multicast data packet to its child nodes, it does not use layer 2 multicast, IP multicast, or MPLS multicast to do so. Rather, it makes n individual copies, and then unicasts each copy, through an IP or MPLS unicast tunnel, to exactly one child node. While the prior MVPN specifications allow the use of IR tunnels, those specifications are not always very clear or explicit about how the MVPN protocol elements and procedures are applied to IR tunnels. This document updates RFCs 6513 and 6514 by adding additional details that are specific to the use of IR tunnels.

Status of this Memo

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/ietf/1id-abstracts.txt.

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

Copyright and License Notice

Table of Contents

1. Introduction

RFCs 6513, 6514, and others describe procedures by which a Service Provider (SP) may offer Multicast VPN (MVPN) service to its customers. These procedures create point-to-multipoint (P2MP) or multipoint-to-multipoint (MP2MP) tunnels, called "P-tunnels" (Provider-tunnels), across the SP’s backbone network. Customer multicast traffic is carried through the P-tunnels.

A number of different P-tunnel technologies are supported. One of the supported P-tunnel technologies is known as "ingress replication" or "unicast replication". We will use the acronym "IR" to refer to this P-tunnel technology.

An IR P-tunnel is a P2MP tree, but a given node on the tree is not necessarily "directly attached" to its parent node or to its child nodes. To send a multicast data packet from a parent node to one of its child nodes, the parent node encapsulates the packet and then unicasts it (through a P2P or MP2P MPLS LSP or a unicast IP tunnel) to the child node. If a node on an IR tree has n child nodes, and has a multicast data packet that must be sent along the tree, the parent node makes n individual copies of the data packet, and then sends each copy, through a unicast tunnel, to exactly one child node. No lower layer multicast technology is used when sending traffic from a parent node to a child node; multiple copies of the packet may therefore be sent out a single interface.

With the single exception of IR, the P-tunnel technologies supported by the MVPN specifications are pre-existing IP multicast or MPLS multicast technologies. Each such technology has its own set of specifications, its own setup and maintenance protocols, its own syntax for identifying specific multicast trees, and its own procedures for enabling a router to be added to or removed from a particular multicast tree. For IR P-tunnels, on the other hand, there is no prior specification for setting up and maintaining the P2MP trees; the procedures and protocol elements used for setting up and maintaining the P2MP trees are specified in the MVPN specifications themselves, and all the signaling/setup is done by using the BGP A-D (Auto-Discovery) routes that are defined in [MVPN-BGP]. (The unicast tunnels used to transmit multicast data from one node to another in an IR P-tunnel may of course have their own setup and maintenance protocols, e.g., [LDP], [RSVP-TE].)

Since the transmission of a multicast data packet along an IR P-tunnel is done by transmitting the packet through a unicast tunnel, previous RFCs sometimes speak of an IR P-tunnel as "consisting of" a set of unicast tunnels. However, that way of speaking is not quite accurate. For one thing, it obscures the fact that an IR P-tunnel is

really a P2MP tree, whose nodes must maintain multicast state in both the control and data planes. For another, it obscures the fact the unicast tunnels used by a particular IR P-tunnel need not be specific to that P-tunnel; a single unicast tunnel can carry the multicast traffic of many different IR P-tunnels (and can also carry unicast traffic as well).

In this document, we provide a clearer and more explicit conceptual model for IR P-tunnels, clarifying the relationship between an IR P-tunnel and the unicast tunnels that are used for data transmission along the IR P-tunnel.

RFC 6514 defines a protocol element called a "tunnel identifier", which for most P-tunnel technologies is used to identify a P-tunnel (i.e., to identify a P2MP or MP2MP tree). However, when IR P-tunnels are used, this protocol element does not identify an IR P-tunnel. In some cases it identifies one of the P-tunnel’s constituent unicast tunnels, and in other cases it is not used to identify a tunnel at all. In this document, we provide an explicit specification for how IR P-tunnels are actually identified.

Some of the MVPN specifications use phrases like "join the identified P-tunnel", even though there has up to now not been an explicit specification of how to identify an IR P-tunnel, of how a route joins such a P-tunnel, or of how a router prunes itself from such a P-tunnel. In this document, we make these procedures more explicit.

RFC 6514 does provide a method for binding an MPLS label to a P-tunnel, but does not discuss the label allocation policies that are needed for correct operation when the P-tunnel is an IR P-tunnel. Those policies are discussed in this document.

This document does not provide any new protocol elements or procedures; rather it makes explicit just how a router is to use the protocol elements and procedures of [MVPN] and [MVPN-BGP] to identify an IR P-tunnel, to join an IR P-tunnel, and to prune itself from an IR P-tunnel. This document also discusses the MPLS label allocation policies that need to be supported when binding MPLS labels to IR P-tunnels, and the timer policies that need to be supported when switching a customer multicast flow from one P-tunnel to another. As the material in this document must be understood in order to properly implement IR P-tunnels, this document is considered to update [MVPN] and [MVPN-BGP]. This document also discusses the application of "seamless multicast" [SMLS-MC] and "extranet" [MVPN-XNET] procedures to IR P-tunnels.

This draft does not discuss the use of IR P-tunnels to support a VPN customer’s use of BIDIR-PIM. [C-BIDIR-IR] explains how to adapt the

procedures of [MVPN], [MVPN-BGP], and [MVPN-BIDIR] so that a customer’s use of BIDIR-PIM can be supported by IR P-tunnels.

The key words "MUST", "MUST NOT", "REQUIRED", "SHALL", "SHALL NOT", "SHOULD", "SHOULD NOT", "RECOMMENDED", "NOT RECOMMENDED", "MAY", and "OPTIONAL", when and only when appearing in all capital letters, are to be interpreted as described in [RFC2119].

2. What is an IR P-tunnel?

An IR P-tunnel is a P2MP tree. Its nodes are BGP speakers that support the MVPN procedures of [MVPN-BGP] and related RFCs. In general, the nodes of an IR P-tunnel are either PE routers, ASBRs, or (if [SMLS-MC] is supported) ABRs. (MVPN procedures are sometimes used to support non-MVPN, or "global table" multicast; one way of doing this is defined in [SMLS-MC]. In such a case, IR P-tunnels can be used outside the context of MVPN.)

MVPN P-tunnels may be either "segmented" or "non-segmented" (as these terms are defined in [MVPN] and [MVPN-BGP]).

A "non-segmented" IR P-tunnel is a two-level P2MP tree, consisting only of a root node and a set of nodes that are children of the root node. When used in an MVPN context, the root is an ingress PE, and the child nodes of the root are the egress PEs.

In a segmented P-tunnel, IR may be used for some or all of the segments. If a particular segment of a segmented P-tunnel uses IR, then the root of that segment may have child nodes that are ABRs or ASBRs, rather than egress PEs.

As with any type of P2MP tree, each node of an IR P-tunnel holds "multicast state" for the P-tunnel. That is, each node knows the identity of its parent node on the tree, and each node knows the identities of its child nodes on the tree. In the MVPN specs, the "parent" node is also known as the "Upstream Multicast Hop" or "UMH".

What distinguishes an IR P-tunnel from any other kind of P2MP tree is the method by which a data packet is transmitted from a parent node to a child node. To transmit a multicast data packet from a parent node to a child node along a particular IR P-tunnel, the parent node does the following:

- It labels the packet with a label (call it a "P-tunnel label") that the child node has assigned to that P-tunnel,

- It then places the packet in a unicast encapsulation and unicasts the packet to the child node. That is, the parent node sends the packet through a "unicast tunnel" to a particular child node. This unicast tunnel need not be specially created to be part of the IR P-tunnel; it can be any P2P or MP2P unicast tunnel that will get the packets from the parent node to the child node. A single such unicast tunnel may be carrying multicast data packets of several different P2MP trees, and may also be carrying unicast data packets.

The parent node repeats this process for each child node, creating one copy for each child node, and sending each copy through a unicast tunnel to corresponding child node. It does not use layer 2 multicast, IP multicast, or MPLS multicast to transmit packets to its child nodes. As a result, multiple copies of each packet may be sent out a single interface; this may happen, e.g., if that interface is the next hop interface, according to unicast routing, from the parent node to several of the child nodes.

Since data traveling along an IR P-tunnel is always unicast from parent node to child node, it can be convenient to think of an IR P-tunnel as a P2MP tree whose arcs are unicast tunnels. However, it is important to understand that the unicast tunnels need not be specific to any particular IR P-tunnel. If R1 is the parent node of R2 on two different IR P-tunnels, a single unicast tunnel from R1 to R2 may be used to carry data along both IR P-tunnels. All that is required is that when the data packets arrive at R2, R2 will see the "P-tunnel label" at the top of the packets’ label stack; R2’s further processing of the packets will depend upon that label. Note that the same unicast tunnel between R1 and R2 may also be carrying unicast data packets.

Typically the unicast tunnels are the Label Switched Paths (LSPs) that already exist to carry unicast traffic; either MP2P LSPs created by LDP [LDP] or P2P LSPs created by RSVP-TE [RSVP-TE]. However, any other kind of unicast tunnel may be used. A unicast tunnel may have an arbitrary number of intermediate routers; those routers do not maintain any multicast state for the IR P-tunnel, and in general are not even aware of its existence.

As with all other P-tunnel types, IR P-tunnels may be used as Inclusive P-tunnels or as Selective P-tunnels.

3. How are IR P-tunnels Identified?

There are four MVPN BGP route types in which P-tunnels can be identified: Intra-AS I-PMSI A-D routes, Inter-AS I-PMSI A-D routes, S-PMSI A-D routes, and Leaf A-D routes. (These route types are all defined in [MVPN-BGP]).

Whenever it is necessary to identify a P-tunnel in a route of one of these types, a "PMSI Tunnel Attribute" (PTA) is added to the route. As defined in [MVPN-BGP] section 5, the PTA contains four fields: "Tunnel Type", "MPLS Label", "Tunnel Identifier", and "Flags". [MVPN-BGP] defines only one bit in the "Flags" field, the "Leaf Information Required" bit.

If a route identifies an IR P-tunnel, the "Tunnel Type" field of its PTA is set to the value 6, meaning "Ingress Replication".

Most types of P-tunnel are associated with specific protocols that are used to set up and maintain tunnels of that type. For example, if the "Tunnel Type" field is set to 2, meaning "mLDP P2MP LSP", the associated setup protocol is mLDP [mLDP]. The associated setup protocol always has a method of identifying the tunnels that it sets up. For example, mLDP uses a "FEC element" to identify a tree. If the "Tunnel type" field is set to 3, meaning "PIM SSM Tree", the associated setup protocol is PIM, and "(S,G)" is used to identify the tree. In these cases, the "Tunnel Identifier" field of the PTA carries a tree identifier as defined by the setup protocol used for the particular tunnel type.

IR P-tunnels, on the other hand, are entirely setup and maintained by the use of BGP A-D routes, and are not associated with any other setup protocol. (The unicast tunnels used to transmit multicast data along an IR P-tunnel may have their own setup and maintenance protocols, of course.) Further, the identifier of an IR P-tunnel does not appear in the PTA at all. Rather, the P-tunnel identifier is in the "Network Layer Reachability Information" (NLRI) field of the A-D routes that are used to advertise and to setup the P-tunnel.

When an IR P-tunnel is identified in an S-PMSI A-D route, an Intra-AS I-PMSI A-D route, or an Inter-AS I-PMSI A-D route (we will refer to these three route types as "advertising A-D routes"), its identifier is hereby defined to be the NLRI of that route. See sections 4.1, 4.2, and 4.3 of [MVPN-BGP] for the specification of these NLRIs. Note that the P-tunnel identifier includes the "route type" and "length" octets of the NLRI.

An advertising A-D route is considered to identify an IR P-tunnel only if it carries a PTA whose "Tunnel Type" field is set to "IR".

When an IR P-tunnel is identified in an S-PMSI A-D route or in an Inter-AS I-PMSI A-D route, the "Leaf Info Required" bit of the Flags field of the PTA MUST be set.

In an advertising A-D route:

- If the "Leaf Info Required" bit of the Flags field of the PTA is set, then the "Tunnel Identifier" field of the PTA has no significance whatsoever, and MUST be ignored upon reception.

  Note that, per RFC6514, the length of the "Tunnel Identifier" field is variable, and is inferred from the length of the PTA. Even when this field is of no significance, its length MUST be the length of an IP address in the address space of the SP's backbone, as specified in section 4.2 of [P-ADDR]. In this case, it is RECOMMENDED that it be set to a routable address of the router that constructed the PTA. (While it might make more sense to allow or even require the field to be omitted entirely, that might raise issues of backwards compatibility with implementations that were designed prior to the publication of this document.)

- If the "Leaf Info Required" bit is not set, the "Tunnel Identifier" field of the PTA does have significance, but it does not identify the IR P-tunnel. The use of the PTA's "Tunnel Identifier" field in this case is discussed in section 5 of this document.

Note that according to the above definition, there is no way for two different advertising A-D routes (i.e., two advertising A-D routes with different NLRIs) to advertise the same IR P-tunnel. In the terminology of [MVPN], an IR P-tunnel can instantiate only a single PMSI. If an ingress PE, for example, wants to bind two customer multicast flows to a single IR P-tunnel, it must advertise that tunnel in an I-PMSI A-D route or in an S-PMSI A-D route whose NLRI contains wildcards [MVPN-WC].

When an IR P-tunnel is identified in a Leaf A-D route, its identifier is the "route key" field of the route's NLRI. See section 4.4 of [MVPN-BGP].

A Leaf A-D route is considered to identify an IR P-tunnel only if it carries a PTA whose "Tunnel Type" field is set to "IR". In this type of route, the "Tunnel Identifier" field of the PTA does have significance, but it does not identify the IR P-tunnel. The use of the PTA's "Tunnel Identifier" field in this case is discussed in section 5.

4. How to Join an IR P-tunnel

The procedures for joining an IR P-tunnel depend upon whether the P-tunnel has been previously advertised, and if so, upon how the P-tunnel was advertised. Note that joining an unadvertised P-tunnel is only possible when using the "Global Table Multicast" procedures of [SMLS-MC].

4.1. Advertised P-tunnels

The procedures in this section apply when the P-tunnel to be joined has been advertised in an S-PMSI A-D route, an Inter-AS I-PMSI A-D route, or an Intra-AS I-PMSI A-D route.

The procedures for joining an advertised IR P-tunnel depend upon whether the A-D route that advertises the P-tunnel has the "Leaf Info Required" bit set in its PTA.

4.1.1. If the ’Leaf Info Required Bit’ is Set

The procedures in this section apply when the P-tunnel to be joined has been advertised in a route whose PTA has the "Leaf Info Required Bit" set.

The router joining a particular IR P-tunnel must determine its UMH for that P-tunnel. If the route that advertised the P-tunnel contains a P2MP Segmented Next Hop Extended Community, the UMH is determined from the value of this community (see [SMLS-MC]). Otherwise the UMH is determined from the route’s next hop (see [MVPN-BGP]).

Once the UMH is determined, the router joining the IR P-tunnel originates a Leaf A-D route. The NLRI of the Leaf A-D route MUST contain the tunnel identifier (as defined in section 3 above) as its "route key". The UMH MUST be identified by attaching an "IP Address Specific Route Target" (or an "IPv6 Address Specific Route Target") to the Leaf A-D route. The IP address of the UMH appears in the "global administrator" field of the Route Target (RT). Details can be found in [MVPN-BGP] and [SMLS-MC].

The Leaf A-D route MUST also contain a PTA whose fields are set as follows:

- The "Tunnel Type" field is set to "IR".

- The "Tunnel Identifier" field is set as described in section 5 of this document.

- The "MPLS Label" field is set to a non-zero value. This is the "P-tunnel label". The value must be chosen so as to satisfy various constraints, as discussed in section 6 of this document.

4.1.2. If the ’Leaf Info Required Bit’ is Not Set

The procedures in this section apply when the P-tunnel to be joined has been advertised in a route whose PTA does not have the "Leaf Info Required Bit" set. This can only be the case if the P-tunnel was advertised in an Intra-AS I-PMSI A-D route.

If an IR P-tunnel is advertised in the Intra-AS I-PMSI A-D routes originated by the PE routers of a given MVPN, the Intra-AS I-PMSI can be thought of as being instantiated by a set of IR P-tunnels. Each PE is the root of one such P-tunnel, and the other PEs are children of the root. A PE simultaneously joins all these P-tunnels by originating (if it hasn’t already done so) an Intra-AS I-PMSI A-D route with a PTA whose fields are set as follows:

- The "Tunnel Type" field is set to "IR".

- The "Tunnel Identifier" field is set as described in section 5 of this document.

- The "MPLS Label" field MUST be set to a non-zero value. This label value will be used by the child node to associate a received packet with the I-PMSI of a particular MVPN. The MPLS label allocation policy must be such as to ensure that the binding from label to I-PMSI is one-to-one.

The NLRI and the RTs of the originated I-PMSI A-D route are set as specified in [MVPN-BGP].

Note that if a set of IR P-tunnels is joined in this manner, the "discard from the wrong PE" procedures of [MVPN] section 9.1.1 cannot be applied to that P-tunnel. Thus duplicate prevention on such IR P-tunnels requires the use of either Single Forwarder Selection ([MVPN] section 9.1.2) or native PIM procedures ([MVPN] section 9.1.3).

4.2. Unadvertised P-tunnels

In [SMLS-MC], a procedure is defined for "Global Table Multicast", in which a P-tunnel can be joined even if the P-tunnel has not been previously advertised. See the sections of that document entitled "Leaf A-D Route for Global Table Multicast" and "Constructing the Rest of the Leaf A-D Route". The route key of the Leaf A-D route has the form of the "S-PMSI Route-Type Specific NLRI" in this case, and that should be considered to be the P-tunnel identifier. Note that the procedure for finding the UMH is different in this case; the UMH is the next hop of the best UMH-eligible route towards the "ingress PE". See the section of that document entitled "Determining the Upstream ABR/PE/ASBR (Upstream Node)".

5. The PTA's 'Tunnel Identifier' Field

If the "Tunnel Type" field of a PTA is set to "IR", its "Tunnel Identifier" field is significant only when one of the following two conditions holds:

- The PTA is carried by a Leaf A-D route, or

- The "Leaf Information Required" bit of the "Flags" field of the PTA is not set.

If one of these conditions holds, then the "Tunnel Identifier" field must contain a routable IP address of the originator of the route. (See [MVPN-BGP] sections 9.2.3.2.1 and 9.2.3.4.1 for the detailed specification of the contents of this field.) This address is used by the UMH to determine the unicast tunnel that it will use in order to send data, along the IR P-tunnel identified by the route key, to the originator of the Leaf A-D route.

The means by which the unicast tunnel is determined from this IP address is outside the scope of this document. The means by which the unicast tunnel is set up and maintained is also outside the scope of this document.

Section 4 of [P-ADDR] MUST be applied when a PTA is carried in a Leaf A-D route, and describes how to determine whether the "Tunnel Identifier" field carries an IPv4 or an IPv6 address.

If neither of the above conditions hold, then the "Tunnel Identifier" field is of no significance, and MUST be ignored upon reception.

6. The PTA's 'MPLS Label' Field

When a PTA is carried by an S-PMSI A-D route or an Inter-AS I-PMSI A-D route, and the "Tunnel Type" field is set to "IR", the "MPLS Label" field is of no significance. In this case, it SHOULD be set to zero upon transmission and MUST be ignored upon reception.

The "MPLS Label" field is significant only when the PTA appears either in a Leaf A-D route or in an Intra-AS I-PMSI A-D route that does not have the "Leaf Information Required" bit set. In these cases, the MPLS label is the label that the originator of the route is assigning to the IR P-tunnel(s) identified by the route's NLRI. (That is, the MPLS label assigned in the PTA is what we have called the "P-tunnel label".)

6.1. Leaf A-D Route Originated by an Egress PE

As previously stated, when a Leaf A-D route is used to join an IR P-tunnel, the "route key" of the Leaf A-D route is the P-tunnel identifier.

We now define the notion of the "root of an IR P-tunnel".

- If the identifier of an IR P-tunnel is of the form of an S-PMSI NLRI, the "root" of the P-tunnel is the router identified in the "Originating Router's IP Address" field of that NLRI.

- If the identifier of an IR P-tunnel is of the form specified in Section "Leaf A-D Route for Global Table Multicast" of [SMLS-MC], the "root" of the P-tunnel is the router identified in the "Ingress PE's IP Address" field of that NLRI.

- If the identifier of an IR P-tunnel is of the form of an Intra-AS I-PMSI NLRI, the "root" of the P-tunnel is the router identified in the "Originating Router's IP Address" field of that NLRI.

- If the identifier of an IR P-tunnel is of the form of an Inter-AS I-PMSI NLRI, the "root" of the P-tunnel is same as the identifier of the P-tunnel, i.e., the combination of an RD and an AS.

Note that if a P-tunnel is segmented, the root of the P-tunnel, by this definition, is actually the root of the entire P-tunnel, not the root of the local segment.

In order to apply the procedures of RFC 6513 Section 9.1.1 ("Discarding Packets from Wrong PE"), the following condition MUST be met by the MPLS label allocation policy:.

> Suppose an egress PE originates two Leaf A-D routes, each with a different route key in its NLRI, and each with a PTA specifying a "Tunnel Type" of "IR". Thus each of the Leaf A-D routes identifies a different IR P-tunnel. Suppose further that each of those IR P-tunnels has a different root. Then the egress PE MUST NOT specify the same MPLS label in both PMSI Tunnel attributes.

That is, to apply the "Discarding Packets from the Wrong PE" duplicate prevention procedures ([MVPN] section 9.1.1), the same MPLS label MUST NOT be assigned to two IR P-tunnels that have different roots.

If segmented P-tunnels are in use, the above rule is necessary but not sufficient to prevent a PE from forwarding duplicate data to the CEs. For various reasons, a given egress PE or egress ABR or egress ASBR may decide to change its parent node, on a given segmented P-tunnel, from one router to another. It does this by changing the RT of the Leaf A-D route that it originated in order to join that P-tunnel. Once the RT is changed, there may be a period of time during which the old parent node and the new parent node are both sending data of the same multicast flow. To ensure that the egress node not forward duplicate data, whenever the egress node changes the RT that it attaches to a Leaf A-D route, it MUST also change the "MPLS Label" specified in the Leaf A-D route’s PTA. This allows the egress router to distinguish between packets arriving on a given P-tunnel from the old parent and packets arriving on that same P-tunnel from the new parent. At any given time, a router MUST consider itself to have only a single parent node on a given P-tunnel, and MUST discard traffic that arrives on that P-tunnel from a different parent node.

If extranet functionality [MVPN-XNET] is not implemented in a particular egress PE, or if an egress PE is provisioned with the knowledge that extranet functionality is not needed, the PE may adopt the policy of assigning a label that is unique for the ordered triple <root, parent node, egress VRF>. This will enable the egress PE to apply the duplicate prevention procedures discussed above, and to determine the VRF to which an arriving packet must be directed.

However, this policy is not sufficient to support the "Discard Packets from the Wrong P-tunnel" procedures that are specified in [MVPN-XNET]. To support those procedures, the labels specified in the PTA of Leaf A-D routes originated by a given egress PE MUST be unique for the ordered triple <root, root RD, parent node>, where the "root RD" is taken from the RD field of the IR P-tunnel identifier. (All forms of IR P-tunnel identifier contain an embedded "RD" field.) This policy is also sufficient for supporting non-extranet cases, but in some cases may result in the use of more labels than the policy of the previous paragraph.

6.2. Leaf A-D Route Originated by an Intermediate Node

When a P-tunnel is segmented, there will be "intermediate nodes" (nodes that have a parent and also have children on the P-tunnel). Each intermediate node is a leaf node of an "upstream segment" and a parent node of a "downstream segment". The intermediate node "splices" together the two segments, so that data it receives on the upstream segment gets transmitted on the downstream segment. If either the upstream or downstream segments (or both) are instantiated by IR, the need to do this splicing places certain constraints on the MPLS label allocation policy.

6.2.1. Upstream and Downstream Segments are IR Segments

An intermediate node N (i.e., a node that has a parent and also has children) on an IR P-tunnel may originate a Leaf A-D route with a particular route key as a result of receiving a Leaf A-D route with that same route key. This will happen only if the received Leaf A-D route carries an IP address specific RT whose Global Administrator field identifies node N.

Suppose intermediate node N originates two Leaf A-D routes, one whose route key is K1, and one whose route key is K2, where K1 != K2. In general, the respective PTAs of these Leaf A-D routes MUST specify distinct non-zero MPLS labels, such that it is possible to map uniquely from the specified label value to a single IR P-tunnel (call this the "uniqueness rule"). There is one exception to this rule; the exception is specified below.

Consider the set of Leaf A-D routes with route key K1 or route key K2 such that:

- N has received these Leaf A-D routes and has them currently installed.

- Each of these Leaf A-D routes carries an IP Address Specific Route Target that identifies N in its Global Administrator field.

Now suppose that all the Leaf A-D routes in this set have the same originating router, and that the PTAs of these Leaf A-D routes all specify the same MPLS label. Suppose further that N’s UMH for K1 is the same as N’s UMH for K2. In this particular case, N MAY specify the same MPLS label in the PTA of Leaf A-D route it originates for K1 as in the PTA of he route it originates for K2. However, if at any future time these conditions no longer hold, N must reoriginate at least one of the Leaf A-D routes with a different label so that the "uniqueness rule" holds.

6.2.2. Only One Segment is IR

To handle the case where an intermediate node, call it N, is splicing together two P-tunnel segments, only one of which is IR, it is necessary to generalize the rules of the preceding sub-section.

Suppose N is a leaf node of two (upstream) P-tunnel segments, call them U1 and U2. Suppose also that N is a parent node of two (downstream) P-tunnel segments, call them D1 and D2. And suppose that N needs to splice U1 to D1, and U2 to D2.

To follow the uniqueness rule of section 6.2.1 of this document, N must assign a different MPLS label to U1 than it assigns to U2. How this assignment is made depends, of course, on the control protocol used to set up U1 and U2.

There is one case in which the uniqueness rule need not be followed. Suppose that there is a node M such that (a) M is N’s only child node on D1, and (b) M is N’s only child node on D2. M will have advertised to N a label L1 bound to D1, and a label L2 bound to D2. If (and for as long as) L1==L2, then N MAY violate the uniqueness rule by advertising to its parent node for U1 the same MPLS label it advertises to its parent node for U2.

Section 6.2.1 of this document specifies in detail the way this requirement is applied when the upstream and downstream segments are all IR segments.

6.3. Intra-AS I-PMSI A-D Route

When a router joins a set of IR P-tunnels using the procedures of section 4.1.2 of this document, the procedures of section 9.1.1 of [MVPN] cannot be applied, no matter what the label allocation policy is. In this case, the ingress PE is the same as the UMH, but it is not possible to assign a label uniquely to a particular ingress PE or UMH. However, the label in the MPLS label field of the PTA MUST NOT appear in the MPLS label field of the PTA carried by any other route originated by the same router.

7. How A Child Node Prunes Itself from an IR P-tunnel

If a particular IR P-tunnel was joined via the procedures of section 4.1.2 of this document, a router can prune itself from the P-tunnel by withdrawing the Intra-AS I-PMSI A-D route it used to join the P-tunnel. This is not usually done unless the router is removing itself entirely from a particular MVPN.

The procedures in the remainder of this section apply when a router joined a particular IR P-tunnel by originating a Leaf A-D route (as described in section 4.1.1 or 4.2 of this document).

If a router no longer has a need to receive any multicast data from a given IR P-tunnel, it may prune itself from the P-tunnel by withdrawing the Leaf A-D route it used to join the tunnel. This is done, e.g., if the router no longer needs any of the flows traveling over the P-tunnel, or if all the flows the router does need are being received over other P-tunnels.

A router that is attached to a particular IR P-tunnel via a particular parent node may determine that it needs to stay joined to that P-tunnel, but via a different parent node. This can happen, for example, if there is a change in the Next Hop or the P2MP Segmented Next Hop Extended Community of the S-PMSI A-D route in which that P-tunnel was advertised. In this case, the router changes the Route Target of the Leaf A-D route it used to join the IR P-tunnel, so that the Route Target now identifies the new parent node.

A parent node must notice when a child node has been pruned from a particular tree, as this will affect the parent node’s multicast data state. Note that the pruning of a child node may appear to the parent node as the explicit withdrawal of a Leaf A-D route, or it may appear as a change in the Route Target of a Leaf A-D route. If the Route Target of a particular Leaf A-D route previously identified a particular parent node, but changes so that it no longer does so, the effect on the multicast state of the parent node is the same as if the Leaf A-D route had been explicitly withdrawn.

8. Parent Node Actions Upon Receiving Leaf A-D Route

These actions are detailed in [MVPN-BGP] and [SMLS-MC]. Two points of clarification are made:

- If a router R1 receives and installs a Leaf A-D route originated by router R2, R1’s multicast state is affected only if the Leaf A-D route carries an "IP Address Specific RT" (or "IPv6 Address Specific RT") whose "global administrator" field identifies R1.

(This is as specified in [MVPN-BGP] and [SMLS-MC].) If a Leaf A-D route's RT does not identify R1, but then changes so that it does identify R1, R1 must take the same actions it would take if the Leaf A-D route were newly received.

- It is possible that router R1 will receive and install a Leaf A-D route originated by router R2, where:

  * the route's RT identifies R1,

  * the route's NLRI contains a route key whose first octet indicates that it is identifying a P-tunnel advertised in an S-PMSI A-D route,

  * R1 has neither originated nor installed any such S-PMSI A-D route.

If at some later time, R1 installs the corresponding S-PMSI A-D route, and the Leaf A-D route is still installed, and the Leaf A-D route's RT still identifies R1, then R1 MUST follow the same procedures it would have followed if the S-PMSI A-D route had been installed before the Leaf A-D route was installed. (I.e., implementers must not assume that events occur in the "usual" or "expected" order.)

9. Use of Timers when Switching UMH

Suppose a child node has joined a particular IR P-tunnel via a particular UMH, and it now determines (for whatever reason) that it needs to change its UMH on that P-tunnel. It does this by modifying the RT of a Leaf A-D route.

It is desirable for such a "switch of UMH" to be done using a "make before break" technique, so that the older UMH does not stop transmitting the packets on the given P-tunnel to the child until the newer UMH has a chance to start transmitting the packets on the given P-tunnel to the child. However, the control plane operation (modifying the RT of the Leaf A-D route) does not permit the child node to first join the P-tunnel at the new UMH, and then later prune itself from the old UMH; a single control plane operation has both effects. Therefore, to achieve "make before break", timers must be used as follows:

1. The old UMH must continue transmitting to the child node for a period of time after it sees the child's Leaf A-D route being withdrawn (or its RT changing to identify a different UMH).

2. The child node must continue to accept packets from the old UMH for a period of time before it starts to accept packets from the new UMH (and discard packets from the old).

Further, the timer in 1 should be longer than the timer in 2. This allows the child to switch from one UMH to another without any loss of data.

10. IANA Considerations

This document contains no actions for IANA.

11. Acknowledgments

The authors wish to thank Yakov Rekhter for his contributions to this work. We also wish to thank Huajin Jeng and Samir Saad for their contributions, and to thank Thomas Morin for pointing out some of the issues that needed further elaboration.

Section 6.1 discusses the importance of having an MPLS label allocation policy that, when ingress replication is used, allows an egress PE to infer the identity of a received packet's ingress PE. This issue was first raised in earlier work by Xu Xiaohu.

12. Security Considerations

No security considerations are raised by this document beyond those already discussed in [MVPN] and [MVPN-BGP].

13. Authors' Addresses

Eric C. Rosen
Cisco Systems, Inc.
1414 Massachusetts Avenue
Boxborough, MA, 01719
Email: erosen@cisco.com

Karthik Subramanian
Cisco Systems, Inc.
170 Tasman Drive
San Jose, CA, 95134
Email: kartsubr@cisco.com

Jeffrey Zhang
Juniper Networks
10 Technology Park Dr.
Westford, MA 01886
Email: zzhang@juniper.net

14. Normative References

[MVPN] "Multicast in MPLS/BGP IP VPNs", E. Rosen and R. Aggarwal, editors, RFC 6513, February 2012

[MVPN-BGP] "BGP Encodings and Procedures for Multicast in MPLS/BGP IP VPNs", R. Aggarwal, E. Rosen, T. Morin, Y. Rekhter, RFC 6514, February 2012

[P-ADDR] "IPv4 and IPv6 Infrastructure Addresses in Updates for Multicast VPN", R. Aggarwal and E. Rosen, RFC 6515, February 2012

[RFC2119] "Key words for use in RFCs to Indicate Requirement Levels.", Bradner, March 1997

15. Informational References

[C-BIDIR-IR] "Simulating "Partial Mesh of MP2MP P-Tunnels" with Ingress Replication", Zhang, Rekhter, Dolganow, draft-ietf-l3vpn-mvpn-bidir-ingress-replication-00.txt, February 2014

[LDP] "LDP Specification", L. Andersson, I. Minei, and B. Thomas, editors, RFC 5036, October 2007

[MVPN-BIDIR] "MVPN: Using Bidirectional P-Tunnels", Rosen, Wijnands, Cai, Boers, draft-ietf-l3vpn-mvpn-bidir-07.txt, January 2014

[MVPN-WC] "Wildcards in Multicast VPN Auto-Discovery Routes", Rosen, Rekhter, Henderickx, Qiu, RFC 6625, May 2012

[MVPN-XNET] "Extranet Multicast in BGP/IP MPLS VPNs", Y. Rekhter and E. Rosen (editors), draft-ietf-l3vpn-mvpn-extranet-05.txt, April 2014

[RSVP-TE] "RSVP-TE: Extensions to RSVP for LSP Tunnels", D. Awduche, L. Berger, D. Gan, T. Li, V. Srinivasan, G. Swallow, RFC 3209, December 2001

[SMLS-MC] "Inter-Area P2MP Segmented LSPs", Y. Rekhter, R. Aggarwal, T. Morin, I. Grosclaude, N. Leymann, S. Saad, draft-ietf-mpls-seamless-mcast-09.txt, December 2013

Network working group
Internet Draft
Category: Informational

X. Xu
Huawei

S. Hares

Y. Fan
China Telecom

C. Jacquenet
Orange

T. Boyes
Bloomberg LP

B Fee
Extreme Networks

Expires: August 2014 February 14, 2014

FIB Reduction in Virtual Subnet

draft-xu-l3vpn-virtual-subnet-fib-reduction-00

Abstract

Virtual Subnet is a L3VPN-based subnet extension solution which can be used to build Layer3 network virtualization overlays within and/or across data centers. This document describes a mechanism for reducing the FIB size of PE routers in the Virtual Subnet context.

Status of this Memo

Copyright Notice

Conventions used in this document

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].

Table of Contents

1. Introduction

Virtual Subnet [VS] is a L3VPN-based subnet extension solution which can be used to build Layer3 network virtualization overlays within and/or across data centers. Since host routes of a given VPN instance are usually exchanged among PE routers participating in that VPN instance in the context of Virtual Subnet, the forwarding table (a.k.a. FIB) size of PE routers may become a scaling concern once they need to install a huge amount of host routes into their forwarding tables, especially in the particular cloud data center interconnect scenario where millions of host routes are there.

To address the above FIB scaling concern, this document proposes a very simple mechanism for reducing the FIB size of PE routers. The basic idea of this mechanism is: Those host routes learnt from remote PE routers are installed into the FIB on demand, while the remaining routes including local host routes are installed into the FIB by default as before.

2. Solution Overview

```
                                        +--------+
                                   +-----+PE-3/APR+-----+
            +-----------------+    |     +--------+     |    +-----------------+
            |VPN_A:1.1.1.1/24 |    |                    |    |VPN_A:1.1.1.1/24 |
            |              \  |    |                    |    |  /              |
            |  +------+     \++---+-+                +-+---++/     +------+   |
            |  |Host A+------+ PE-1 |                | PE-2 +------+Host B|   |
            |  +------+\     ++-+-+-+                +-+-+-++     /+------+   |
            |   1.1.1.2/24    | | |                    | | |    1.1.1.3/24    |
            |                 | | |                    | | |                  |
            |     DC West     | | |  IP/MPLS Backbone  | | |      DC East     |
            +-----------------+ | |                    | | +-----------------+
                                | +--------------------+ |
                                |                        |
        VRF_A :                 V                VRF_A : V
      +------------+---------+--------+------+ +------------+---------+--------+
------+
      |   Prefix   | Nexthop |Protocol|In_FIB| |   Prefix   | Nexthop |Protocol|
In_FIB|
      +------------+---------+--------+------+ +------------+---------+--------+
------+
      | 1.1.1.1/32 |127.0.0.1| Direct |  Yes | | 1.1.1.1/32 |127.0.0.1| Direct |
  Yes |
      +------------+---------+--------+------+ +------------+---------+--------+
------+
      | 1.1.1.2/32 | 1.1.1.2 | Direct |  Yes | | 1.1.1.2/32 |  PE-1   |  IBGP  |
  No  |
      +------------+---------+--------+------+ +------------+---------+--------+
------+
      | 1.1.1.3/32 |   PE-2  |  IBGP  |  No  | | 1.1.1.3/32 | 1.1.1.3 | Direct |
  Yes |
      +------------+---------+--------+------+ +------------+---------+--------+
------+
      | 1.1.1.0/25 |   APR   |  IBGP  |  Yes | | 1.1.1.0/25 |   APR   |  IBGP  |
  Yes |
      +------------+---------+--------+------+ +------------+---------+--------+
------+
      |1.1.1.128/25|   APR   |  IBGP  |  Yes | |1.1.1.128/25|   APR   |  IBGP  |
  Yes |
      +------------+---------+--------+------+ +------------+---------+--------+
------+
      | 1.1.1.0/24 | 1.1.1.1 | Direct |  Yes | | 1.1.1.0/24 | 1.1.1.1 | Direct |
  Yes |
```

```
      +------------+---------+--------+------+ +------------+---------+--------+
------+
```

Figure 1: FIB Reduction Example

To reduce the FIB size of PE routers, the selective FIB installation concept as described in [VA] can be leveraged in the Virtual Subnet context. Take the VPN instance demonstrated in Figure 1 as an example, the FIB reduction procedures are described as follows:

1) Multiple more specific prefixes (e.g., 1.1.1.0/25 and 1.1.1.128/25) corresponding to an extended subnet (i.e., 1.1.1.0/24) are specified as Virtual Prefixes (VPs). Meanwhile, one or more PE routers are configured as Aggregation Point Routers (APR) for each VP. The APRs for a given VP would install a null route to that VP while propagating a route to that VP via the L3VPN signaling.

2) For a given host route in the routing table which is learnt from any remote PE router, PE routers which are non-APRs for any VP covering this host route would not install it into the FIB by default. In contrast, PE routers which are APRs for any VP covering that host route would install it into the FIB.

3) Upon receiving a packet destined for a given remote CE host, if no host route for that CE host is found in the FIB, the ingress PE router would forward the packet to a given APR according to the longest-matching VP route, which in turn forwards the packet to the final egress PE router. In this way, the FIB size of those non-APR PE routers can be greatly reduced at the potential cost of path stretch.

In order to forward packets destined for remote CE hosts directly to the final egress PE routers without the potential path stretch penalty, non-APR PE routers could perform on-demand FIB installation for remote host routes which are available in the routing table. For example, upon receiving an ARP request or Neighbor Solicitation (NS) message from a local CE host, the non-APR PE router would perform a lookup in the routing table. If a corresponding host route for the target host is found but not yet installed into the FIB, it would be installed into the FIB. Another possible way to trigger on-demand FIB installation is as follows: when receiving a packet whose longest-matching FIB entry is a particular VP route learnt from any APR, a copy of this packet would be sent to the control plane while this original packet is forwarded as normal. The above copy sent to the control plane would trigger a lookup in the routing table. If a corresponding host route is found but not yet installed into the FIB, it would be installed into the FIB. To provide robust protection against DoS attacks on the control plane, rate-limiting of the above packets sent to the control plane MUST be enabled. Those FIB entries

for remote CE host routes which are on-demand installed on non-APR PE routers would expire if not used for a certain period of time.

3. Security Considerations

This document doesn’t introduce additional security risk to BGP/MPLS IP VPN, nor does it provide any additional security feature for BGP/MPLS IP VPN.

4. IANA Considerations

There is no requirement for any IANA action.

5. Acknowledgements

Thanks Robert Raszuk for his review and suggestions on this document.

6. References

6.1. Normative References

[RFC2119] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, March 1997.

[VS] Xu, X et al., "Virtual Subnet: A L3VPN-based Subnet Extension Solution", draft-xu-l3vpn-virtual-subnet-03 (Work in Progress), January 2014.

6.2. Informative References

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

[VA] Francis, P., Xu, X., Ballani, H., Jen, D., Raszuk, R., and L. Zhang, "Virtual Aggregation", draft-ietf-grow-va-05.txt, (Work in Progress), December 2011.

Authors’ Addresses

Xiaohu Xu
Huawei Technologies,
Beijing, China.
Phone: +86 10 60610041
Email: xuxiaohu@huawei.com

Susan Hares
Email: shares@ndzh.com

Yongbing Fan
Guangzhou Institute, China Telecom
Guangzhou, China.
Phone: +86 20 38639121
Email: fanyb@gsta.com

Christian Jacquenet
Orange
Rennes France
Email: christian.jacquenet@orange.com

Truman Boyes
Bloomberg LP
Phone: +1 2126174826
Email: tboyes@bloomberg.net

Brendan Fee
Extreme Networks
9 Northeastern Blvd.
Salem, NH, 03079
Email: bfee@enterasys.com

INTERNET-DRAFT
Intended Status: Proposed Standard
Expires: December 18, 2014

Mingui Zhang
Peng Zhou
Huawei
Russ White
IETF
June 16, 2014

Label Sharing for Fast PE Protection
draft-zhang-l3vpn-label-sharing-02.txt

Abstract

This document describes a method to be used by VPN Service Providers to provide multi-homed CEs with fast protection of egress PEs. Egress PEs in a redundant group always share the same label in distribution of VPN routes of a VRF. A virtual Next Hop (vNH) in the IGP/MPLS backbone is created as the common end of LSP tunnels which would otherwise terminate at each egress PE. Primary and backup LSP tunnels ended at the vNH are set up by MPLS on basis of existing IGP FRR mechanisms. If the primary egress PE fails, the backup egress PE can recognize the "shared" VPN route label carried by the data packets. Therefore, the failure affected data packets can be smoothly rerouted to the backup PE for delivery without changing their VPN route label.

Status of this Memo

Copyright and License Notice

Table of Contents

1. Introduction

For the sake of reliability, ISPs often connect one CE to multiple PEs. When the primary egress PE fails, a backup egress PE continues to offer VPN connectivity to the CE. If local repair is performed by the upstream neighbor of the primary egress PE on the data path, it’s possible to achieve a 50msec switchover.

VPN routes learnt from CEs are distributed by egress PEs to ingress PEs that need to know these VPN routes. Egress PEs in a redundant group (RG) MUST advertise the same VPN route label for routes of the same VPN. When the primary egress PE fails, data packets are redirected to a backup egress PE by the PLR (Point of Local Repair) router, the backup PE can recognize the VPN route label in these data packets and deliver them correctly. The method developed in this document is so called "Label Sharing for Fast PE Protection".

1.1. Overview

```
                 +====================================+
           +---+ |  +---+    +--+    +---+ M          |
           |CE1+----+PE1+----+P1+----+PE3+-------+    |
           +---+ |  +-+-+    +-++    +---+   1100|    |
                 |    |        |       |       +-+-+  | +---+
                 |    |        |       |       +vNH+----+CE2|
                 |    |        |       |       +-+-+  | +---+
           +---+ |  +-+-+    +-++    +---+   1100|    |
           |CE3+----+PE2+----+P2+----+PE4+-------+    |
           +---+ |  +---+    +--+    +---+ S          |
                 |                                    |
                 |     IGP/MPLS Backbone Network      |
                 +====================================+
```

Figure 1.1: Egress PE routers share the same VPN route label 1100.

An example topology is shown in Figure 1.1. Let PE1 and PE2 be ingress routers, and let PE3 and PE4 be egress routers. CE2 is connected to both PE3 and PE4 so they form an Redundant Group (RG). Usually, egress PEs may be configured to be in the same RG or discover each other from the CE routes learning process which can be a dynamic routing algorithm or a static routing configuration [RFC4364]. Suppose PE3 is the primary while PE4 is the backup. For topologies with more than two egress PEs in an RG, one PE acts as the primary while other act as backups.

A vNH node is created in the backbone. The primary PE allocates a loopback IP address to vNH (say 2.2.2.2). Instead of the egress PEs, vNH acts as the common end node of LSP tunnels which otherwise end at

egress PEs. The metrics (’M’ and ’S’) for the links between egress PEs and vNH is set up in a way that the primary and backup LSP tunnels to vNH respectively use PE3 and PE4 as the penultimate hop.

Egress PEs in an RG MUST advertise the same VPN route label for each VPN connected to this RG. When a route is learn from CE2 (say 10.9.8/24), PE3 and PE4 will distribute this route to other PEs sharing the same label (say 1100). In this way, when the primary PE fails, the VPN route label carried with the rerouted data packets need not be changed. It can be recognized by the backup PE as well.

This document supposes BGP/MPLS IP VPN [RFC4364] is deployed in the backbone and Label Distribution Protocol (LDP) is used to distribute MPLS labels. The approach developed in this document confines changes to routers in an RG. P and PE routers out of this RG are totally oblivious to these changes.

1.2. Conventions used in this document

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].

1.3. Terminology

VRF: Virtual Routing and Forwarding table [RFC4364]

FRR: Fast ReRouting

PLR: Point of Local Repair

LFA: Loop-Free Alternate [LFA]

RG: Redundant Group. A Redundant Group of Provider Edge nodes (PEs) to which a set of CEs are multi-homed.

2. The Virtual Next Hop

A virtual router (the virtual Next Hop, vNH) is created in IGP to represent the RG in the Service Provider’s backbone. For other routers in the backbone, the vNH acts as the common egress PE connecting a set of CEs. Multiple vNHs may be created for one RG. Then multiple paths can be computed from ingress PEs to the vNHs. Ingress PEs can choose from these paths to achieve load balance for the CEs.

Service Providers may configure one PE to be the primary when an RG is created. The primary PE may also be automatically elected out of

the RG in the same way the DR is selected (see section 7.3 of [RFC2328]), or the DIS is selected [ISIS]. Other PEs in the RG will act as backup ones. This primary PE determines the loopback IP address for the vNH. This loopback IP address can be configured manually or assigned automatically. The SystemID of the vNH under ISIS is composed based on this loopback IP address. The primary PE generates the router link state information (LSA/LSP) on behalf of the vNH. Links to each PE and each CE in the group are included in router link state information PDUs of the PE and CE.

The overload mode MUST be set so that the rest routers in the network will not route transit traffic through the vNH. In OSPF, the overload mode can be set up through setting the link weights from the vNH to egress PEs to the maximum link weight which is 0xFFFF. In ISIS, this overload mode is realized as setting the overload bit in the LSP of the vNH. (See Appendix A and B for the detail set up of LSAs/LSPs.)

3. Link Costs Set Up for IGP FRR

```
                  |<------ Sxy3-------->|

                  +-------Px(PLR)-------PE3
                  |                     | \ M
                  |                     |  \
                 Pxy                  C34|   vNH
                  |                     |  /
                  |                     | / S
                  +---------------------PE4

                  |<------ Sxy4-------->|
```

Figure 2.2: The illustration of equations.

If the IGP costs for the links between egress PEs and the vNH can be set up in a way that one egress PE appears on the primary path while the other PE appears on the backup path, the PLR can make use of the multiple egress PEs to achieve fast failure protection. Link weights can be set up according to the following rule in order to leverage the well supported [LFA] as the IGP FRR mechanism.

1. This document supposes bidirectional link weights are being used. As illustrated in Figure 2.2, assume the weight for the link between PE3 and vNH is "M" and the weight for the link between PE4 and vNH is "S". The weight for the link between PE3 and PE4 is C34.

2. Px is a neighbor of PE3. This Px will act as the PLR. Suppose Pxy is Px’s neighbor with the shortest path to PE4, after PE3 is removed from the topology. The cost of this path is Sxy4.

3. Add PE3 back to the topology. The cost of the path from Pxy to PE3 is Sxy3.

4. "M" and "S" can be set up as long as the following two equations hold.

$$eq1: Sxy4+S < Sxy3+M$$

$$eq2: C34+S > M$$

The eq1 guarantees that Pxy is safe, i.e., no loop occurs, to be used as the next hop by the PLR for bypass. The eq2 is designed to insure that the primary path does not go through the primary egress PE and backup egress PE in series.

Although this document designs the method based on [LFA] which is widely deployed, other IGP FRR mechanisms can also be utilized to achieve the protection. For example, [MRT] can be applicable regardless of how the link weights are set up.

4. The LSP Tunnels

Egress PEs use the IP address of the vNH to identify the FEC. Its LSPs on basis of IGP routes with vNH as the last hop are set up using LDP:

- The primary LSP tunnel follows the IGP route from ingress PEs to the vNH;

- The backup LSP tunnel is set up according to existing IGP FRR calculation, such as [MRT] and [LFA].

Data packets are tunneled through the backbone using a "tunnel label" at the top of the label stack. Egress PE will not really transmit a packet to the tunnel end node vNH. Rather, they need locally deliver the packet. It can be interpreted that at the egress PE, the packet’s next hop is the egress PE itself (see Section 3.10 of [RFC3031]). The tunnel label will be popped at the egress PE. The indication for popping is got from the tunnel label at the top of the stack since this is a label assigned to the FEC identified by the PE’s loopback IP address. Next, there will be a pop of the VPN route label followed by an address lookup in the VRF. Section 5 will explain how to set the VPN route label in order to leverage these LSP tunnels to achieve the egress PE protection.

5. The VPN Route Label

5.1. Sharing the VPN Route Label

In [RFC4364], egress PEs separately allocate and distribute the label for the route to an address prefix they learn from CEs. In this document, it's REQUIRED that backup PE(s) in an RG always advertises the label already advertised by the primary PE for the address prefix in question. The primary PE RG SHOULD distribute the same label for any address prefix in an attached VPN. This is per VRF label sharing. Others granularities, such as per address family per VRF label sharing, are also feasible.

Egress PEs continue to locally allocate VPN route labels so that the proposal need not modify existing forwarding processes of L3VPN egress PEs. At the backup egress PE, the allocated label and the distributed label would be inconsistent. The following two options arise to address this issue.

5.1.1. Option A: Reserved Label Ranges per RG

PEs in an RG are physically connected to the same set of CEs. It's viable for them allocate the same VPN route label per VPN. For each VPN served by an RG, the backup egress PE always allocates the same label as the primary PE. It acts as a 'compromised' network entity which always listens to the label advertised by the primary then allocates and also distributed the same label. By doing this, they are intimating the VPN route label allocation of the virtual node, vNH.

For this option, PEs in an RG are REQUIRED to reserve the same label range(s) for allocation at the management plane. PEs with h/w disjoint label ranges are not qualified for this option. This option SHOULD only be used in well managed and highly monitored networks. It's not intended to be applicable when the RG spans more than one administrative domain. It ought not to be deployed on or over the public Internet.

Note that if one PE participates in multiple RGs, a label range reserved for one RG can't be used by another RG on this PE. It increases the consumption of labels on this PE. So this option should be deployed with care in this case.

The architecture of the label sharing method allows a 'higher-layer' entity to allocate labels for all PEs across all RGs. This document leaves this choice as for future study.

5.1.2. Option B: The Label Swapping Table

```
+----+----+
|1100| 30 |
|1101| 31 |
|1102| 32 |
   .    .
   .    .
   .    .
+----+----+
```

Figure 2.3: The label ’swapping’ table

In the inter-AS L3VPN Option B defined in Section 10 of [RFC4364], when an ASBR distributes a VPN route to an ASBR in another AS, it need perform a label swap for this route. Similarly, the backup PE in this proposal uses a label swapping table to record the mapping between advertised labels and locally assigned labels for VPN routes. Obviously, the backup PE need maintain one such table per RG. Whenever a data packet to a route in a VPN attached to the RG arrives at the backup PE, the locally assigned label (e.g., 30) got from the swapping will be used in the VPN route label lookup followed by an address lookup.

5.2. Binding to LSP Tunnels

When the VPN route with a shared label is distributed to other PEs by the primary PE and backup PEs, the BGP next hop is set to the IP address of the vNH. As defined in Section 4, LSP tunnels are set up for the FEC identified also by the IP address of the vNH. By doing this, the VPN route is bound to these LSP tunnels. When data packets to this VPN route are tunneled through the backbone, these LSP tunnels will offer the protection.

6. Examples To Walk Through

Two examples are included in this section. Figure 1.1 is referred. The first one describes how to distribute VPN route label to peers. It’s westbound in the control plane. The second one interprets how egress PE act in the case of the primary PE failure. It’s eastbound in the data plane.

6.1. Label Distribution Procedure

Assume PE3 is elected as the primary while PE4 is the backup. The loopback IP address of vNH is 2.2.2.2.

1) PE3 learns the VPN route to address prefix 10.9.8/24 from CE2. It allocates the VPN route label 1100 and distributes it in BGP with 2.2.2.2 as the BGP Next Hop. (prefix = 10.9.8/24|label = 1100|BGP

Next Hop = 2.2.2.2)

2) PE4 also learns the VPN route to address prefix 10.9.8/24 and allocate the VPN route label 30. It then waits for the primary PE3 to advertise the VPN route label for this prefix.

3) PE4 monitors the VPN route label 1100 from PE3 for the prefix 10.9.8/24. The mapping from 1100 to 30 is inserted to the swapping table.

4) PE4 distributes the VPN route using the monitored label 1100. (prefix = 10.9.8/24|label = 1100|BGP Next Hop = 2.2.2.2)

6.2. Protection Procedure

Suppose the label for the primary LSP tunnel to vNH is 2100 while the backup LSP tunnel to vNH is 3100. P1 is the PLR.

1) In normal case, P1 sends data packets with tunnel label 2100 to PE3. When PE3 fails, P1 redirects data packets to the backup LSP tunnel (say P2-PE4-vNH) using tunnel label 3100.

2) PE4 will receive a packet with two levels of labels. It pops the outer label 3100 and use this label to identify a swapping table.

3) PE4 pops the VPN route label and looks up the swapping table. The VPN route label 1100 is mapped to 30.

4) The VPN route label 30 is looked up in the VPN route label table followed by an address lookup in the VRF.

7. Operations

7.1. Label Space Management for Option A

A label range should be reserved before an RG comes to operate. Operators need set a large label sharing space for label ranges reservation. When an RG is created, the operator needs reserve a unused label range for it. The label range should be reserved in a manner of ’enough is enough’. If a label range of an RG is being used out, the operator can reserve a new range from the unused label sharing space. The newly reserved range is then appended to the one being used out.

If a backup PE is partitioned from the primary PE, it continues to work with those allocated labels for the RG. However, it MUST NOT allocate any more labels in the reserved ranges. A label in a reserved range can only be allocated by a backup PE when it monitors

that the primary PE has distributed this label.

When a primary PE resumes from a failure, its reserved label ranges come to work again. It SHOULD conserve the labels it allocated for each range.

7.2. Backup LSP Tunnel Exceptions

The label sharing method requires that the backup LSP tunnel is set up as specified in Section 4, following the IGP route. However, Service Providers are allowed to have exceptions. For instance, an operator may use BGP Local_Pref to give a higher degree of preference to the route advertised by the primary PE. For another instance, the operator may have the primary PE advertise a more specific prefix. Take Figure 1.1 for example, the backup tunnel will actually goes through PE4->PE3->CE2 for both instances. When the VPN route is bound to this tunnel, it does not protect the primary egress PE. An alarm should be generated to notify the operator that such kind of configuration will jeopardize the VPN route’s resilience to egress PE node failure.

8. Security Considerations

This document raises no new security issues.

9. IANA Considerations

This document requires no IANA actions. RFC Editor: please remove this section before publication.

Acknowledgements

Authors would like to thank the comments and suggestions from Bruno Decraene, Eric Rosen, Eric Gray, Jakob Heitz, James Uttaro, Jeff Tantsura, Loa Andersson, Nagendra Kumar, Robert Raszuk, Stewart Bryant, Shunwan Zhuang, Wim Henderickx and Zhenbin Li.

10. References

10.1. Normative References

[LFA] Filsfils, C., Ed., Francois, P., Ed., Shand, M., Decraene, B., Uttaro, J., Leymann, N., and M. Horneffer, "Loop-Free Alternate (LFA) Applicability in Service Provider (SP) Networks", RFC 6571, June 2012.

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

[ISIS] ISO, "Intermediate system to Intermediate system routeing information exchange protocol for use in conjunction with the Protocol for providing the Connectionless-mode Network Service (ISO 8473)," ISO/IEC 10589:2002.

[RFC2328] Moy, J., "OSPF Version 2", STD 54, RFC 2328, April 1998.

[RFC1213] McCloghrie, K. and M. Rose, "Management Information Base for Network Management of TCP/IP-based internets:MIB-II", STD 17, RFC 1213, March 1991.

10.2. Informative References

[MRT] A. Atlas, Ed., R. Kebler, et al, "An Architecture for IP/LDP Fast-Reroute Using Maximally Redundant Trees", draft-ietf-rtgwg-mrt-frr-architecture, work in progress.

Appendix A: Generating OSPF LSAs

The following Type 1 Router-LSA is flooded by the egress PE with the highest priority. As defined in [RFC2328], this LSA can only be flooded throughout a single area.

```
 0                   1                   2                   3
 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|            LS age             |     Options   |    LS type    |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|                        Link State ID                          |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|                     Advertising Router                        |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|                     LS sequence number                        |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|         LS checksum           |             length            |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|    0    |V|E|B|        0      |            # links            |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|                          Link ID                              |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|                         Link Data                             |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|     Type      |     # TOS     |            metric             |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|                              ...                              |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|      TOS      |        0      |          TOS  metric          |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
```

```
      |                           Link ID                             |
      +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
      |                          Link Data                            |
      +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
      |                              ...                              |
```

LS age
The time in seconds since the LSA was originated. (Set to 0x708 by default.)

Options
As defined in [RFC2328], options = (E-bit).

LS type
1

Link State ID
Same as the Advertising Router

Advertising Router
The Router ID of the vNH.

LS sequence number
As defined in [RFC2328].

LS checksum
As defined and computed in [RFC2328].

length
The length in bytes of the LSA. This includes the 20 byte LSA header. (As defined and computed in [RFC2328].)

VEB
As defined in [RFC2328], set its value to 000.

#links
The number of router links described in this LSA. It equals to the number of Egress PEs in the RG.

The following fields are used to describe each router link connected to an egress PE. Each router link is typed as Type 1 Point-to-point connection to another router.

Link ID
The Router ID of one of the egress PEs in the RG.

Link Data
It specifies the interface’s MIB-II [RFC1213] ifIndex value. It

ranges between 1 and the value of ifNumber. The ifNumber equals to the number of the PEs in the RG. The PE with the highest priority sorts the PEs according to their unsigned integer Router ID in the ascend order and assigns the ifIndex for each.

Type
Value 1 is used, indicating the router link is a point-to-point connection to another router.

TOS
This field is set to 0 for this version.

Metric
It is set to 0xFFFF.

The fields used here to describe the virtual router links are also included in the Router-LSA of each egress PEs. The Link ID is replaced with the Router ID of the vNH. The Link Data specifies the interface’s MIB-II [RFC1213] ifIndex value. The "Metric" field is set as defined in Section 3.

Appendix B: Generating ISIS LSPs

The primary egress PE generates the following level 1 LSP to describe the vNH node.

```
                                 No. of octets
+-------------------------+
| Intradomain Routeing    |     1
| Protocol Discriminator  |
+-------------------------+
| Length Indicator        |     1
+-------------------------+
| Version/Protocol ID     |     1
| Extension               |
+-------------------------+
| ID Length               |     1
+-------------------------+
|R|R|R| PDU Type          |     1
+-------------------------+
| Version                 |     1
+-------------------------+
| Reserved                |     1
+-------------------------+
| Maximum Area Address    |     1
+-------------------------+
| PDU Length              |     2
```

```
+-------------------------+
| Remaining Lifetime      |     2
+-------------------------+
| LSP ID                  |     ID Length + 2
+-------------------------+
| Sequence Number         |     4
+-------------------------+
| Checksum                |     2
+-------------------------+
|P|ATT|LSPDBOL|IS Type    |     1
+-------------------------+
: Variable Length Fields  :     Variable
+-------------------------+
```

Intradomain Routeing Protocol Discriminator - 0x83 (as defined in [ISIS])

Length Indicator - Length of the Fixed Header in octets

Version/Protocol ID Extension - 1

ID Length - As defined in [ISIS]

PDU Type (bits 1 through 5) - 18

Version - 1

Reserved - transmitted as zero, ignored on receipt

Maximum Area Address - same as the primary egress PE

PDU Length - Entire Length of this PDU, in octets, including the header.

Remaining Lifetime - Number of seconds before this LSP is considered expired. (Set to 0x384 by default.)

LSP ID - the system ID of the source of the LSP. It is structured as follows:

```
+-------------------------+
| Source ID               |     6
+-------------------------+
| Pseudonode ID           |     1
+-------------------------+
| LSP Number              |     1
+-------------------------+
```

Source ID - SystemID of the vNH

Pseudonode ID - Transmitted as zero

LSP Number - Fragment number

Sequence Number - sequence number of this LSP (as defined in [ISIS])

Checksum - As defined and computed in [ISIS]

P - Bit 8 - 0

ATT - Bit 7-4 - 0

LSDBOL - Bit 3 - 1

IS Type - Bit 1 and 2 - bit 1 set, indicating the vNH is a Level 1 Intermediate System

In the Variable Length Field, each link outgoing from the vNH to an egress PE is depicted by a Type #22 Extended Intermediate System Neighbors TLV [RFC5305]. The egress PE is identified by the 6 octets SystemID plus one octet of all-zero pseudonode number. The 3 octets metric is set as that in Section 3. None sub-TLVs is used by this version, therefore the value of the one octet length of sub-TLVs is 0. The Type #22 TLV requires 11 octets.

The Type #22 TLV is also included in the LSP of each egress PE to depict the incoming link of the vNH. Only the 6 octets SystemID is replaced with the SystemID of the vNH.

Author’s Addresses

Mingui Zhang
Huawei Technologies
No.156 Beiqing Rd. Haidian District,
Beijing 100095 P.R. China

Email: zhangmingui@huawei.com

Peng Zhou
Huawei Technologies
No.156 Beiqing Rd. Haidian District,
Beijing 100095 P.R. China

Email: Jewpon.zhou@huawei.com

Russ White
Verisign
12061 Bluemont Way
Reston, VA 20190
USA

Email: russw@riw.us