Network Working Group
Internet Draft
Intended status: Informational
Expires: April 15, 2013

Maria Napierala
AT&T
Luyuan Fang
Cisco Systems

October 15, 2012

Requirements for Extending BGP/MPLS VPNs to End-Systems
draft-fang-l3vpn-end-system-requirements-00.txt

Abstract

Service Providers commonly use BGP/MPLS VPNs [RFC 4364] as the control plane for wide-area virtual networks. This technology has proven to scale to a large number of VPNs and attachment points, and it is well suited to provide VPN service to end-systems. Virtualized environment imposes additional requirements to MPLS/BGP VPN technology when applied to end-system networking, which are defined in this document.

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 [RFC 2119].

1. Introduction

Networks are increasingly being consolidated and outsourced in an effort, both, 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.

In order to scale compute, storage, and network service functions, physical resources are being abstracted from their logical representation. This is referred as server, storage, and network virtualization. Virtualization can be implemented in various layers of computer systems or networks. The virtualized loads are executed 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 IP connectivity to virtual resources on end-systems. The requirements address the 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.

1.1. Terminology

AS	Autonomous Systems
End-System	A device where Guest OS and Host OS/Hypervisor reside
IaaS	Infrastructure as a Service
RT	Route Target
ToR	Top-of-Rack switch
VM	Virtual Machine
Hypervisor	Virtual Machine Manager
SDN	Software Defined Network
VPN	Virtual Private Network

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

MPLS/BGP VPN technology [RFC 4364] 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 connected to a Provider Edge (PE). 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 should to able to evolve and adapt to new virtualized environments by extending VPN service to end-systems.

When end-system attaches to MPLS/BGP VPN, CE becomes a Virtual Machine or 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. Moreover, it is a current practice to implement 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 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. The PE control plane function can itself be virtualized and run as an application in end-system.

3. Connectivity Requirements

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.

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.

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.

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. A collection of virtual resources might provide external or internal services. For example, such collection may serve an external "customer" or internal "tenant" to whom a Service Provider provides service(s). We will refer to collection of virtual resources dedicated to a process or application as a VPN, using the terminology of IP VPNs.

Any network virtualization solution has to assure the network isolation (in data plane and control plane) among tenants or applications sharing the same data center physical resources. 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 also common for tenants to require multiple distinct VPNs. In that scenario traffic might need to cross VPN boundaries, subject to access controls and/or routing policies.

A tenant should be able to create multiple VPNs. A network virtualization solution should allow a VM or application end-point to directly access multiple VPNs without a need to traverse a gateway. 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. A network virtualization solution should support both, isolated VPNs and overlapping VPNs (often referred to as "extranets"), as well as both, 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". It is often done by delegating the virtual resource communication processing to the network edge. Networks use various VPN technologies to isolate disjoint groups of virtual resources. Some use VLANs 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.

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. It has been proven (in Internet and in large BGP IP VPN deployments) that moving complexity to network edge while keeping network core simple has very good scaling properties.

There should be a total separation between the virtualized segments (i.e., interfaces associated with virtual resources) and the physical network (i.e., physical interfaces associated with network infrastructure). This separation should include the separation of the virtual network IP address space from the physical network IP address space. The physical infrastructure addresses should be routable in the underlying transport network, while the virtual network addresses should be routable only in the virtual network. Not only should the virtual network data plane be fully decoupled from the physical network, but its control plane should be decoupled as well.

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 (edge) 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. Encapsulation of Virtual Payloads

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.

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

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.

MPLS/VPN technology based on [RFC 4364] 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. If the transport network is only IP-capable then MPLS in IP or MPLS in GRE [RFC4023] encapsulation could be used. Other transport layers such 802.1ah might also need to be supported.

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. Hence, it is important that the virtualized networking solutions are efficient in terms of traffic forwarding and 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. 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 WAN connectivity to the virtualized resources in SP's data center. MPLS/BGP VPN customers may require simultaneous access to resources in both SP and their own data centers. The service provider-based VPN access can provide additional value compared with public internet access, such as security, QoS, OAM, multicast service, VoIP service, video conferencing, wireless connectivity. 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, or C [RFC 4364]. MPLS/BGP VPN technology is widely supported on routers and other appliances. BGP/MPLS VPN-capable network devices 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).

When connecting an end-system VPN with other services/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.

10. IP Mobility

Another reason for a network virtualization is the need to support IP mobility. IP mobility consists in IP addresses used for communication within or between applications being anywhere across the 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.

To accommodate live mobility of a virtual machine (or a device), it is desirable to assign to it a permanent IP address that remains with the VM/device after it moves. When dealing with IP-only applications it is not only sufficient but optimal to forward the traffic based on layer 3 rather than on layer 2 information. The MAC addresses of devices or applications should be irrelevant to IP services and introduce unnecessary overhead and complications when devices or VMs move (i.e., when a VM moves between physical servers, the MAC learning tables in the switches must be updated; also, 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 should be handled by an IP route advertisement.

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 advertising 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.

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 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 and not only the best path to a given destination.

11.2. Optimizing Route Distribution

When virtual networks are triggered based on the IP communication, the Route Target Constraint extension [RFC 4684] 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. Security Considerations

The document presents the requirements for end-systems MPLS/BGP VPNs. The security considerations for specific solutions will be documented in the relevant documents.

13. IANA Considerations

This document contains no new IANA considerations.

14. Normative References

[RFC 4363] Rosen, E. and Y. Rekhter, "BGP/MPLS IP Virtual Private Networks (VPNs)", RFC 4364, February 2006.

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

[RFC 4684] 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.

15. Informative References

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

16. Authors' Addresses

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

Luyuan Fang
Cisco Systems
111 Wood Avenue South
Iselin, NJ 08830, USA
Email: lufang@cisco.com

17. Acknowledgements

The authors would like to thank Pedro Marques for his comments and input.