Requirements for Extending BGP/MPLS VPNs to End-Systems
draft-fang-l3vpn-end-system-requirements-02.txt

Abstract

   The proven scalability and extensibility beyond the original design
   purposes of the BGP/MPLS IP VPNs (IP VPN) technology [RFC4364] has
   made it an attractive candidate for Data Center (DC)/Cloud
   virtualization. This document provides the requirements for extending
   IP VPN (in original or modified versions) into the end-systems/end-
   devices, such as a server in a DCs/Cloud. Physical separation of the
   control and the forwarding planes; virtualizing the network functions
   of the IP VPN network elements, such as a PE, are the key differences
   compared with the classic IP VPN solutions.

Status of this Memo

Copyright and License Notice

Table of Contents

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). Network virtualization is the next step after
compute virtualization.

This document defines requirements for a network virtualization
solution that provides BGP/MPLS IP VPN style connectivity to virtual
resources on end-systems/end-device, such as a server, operating in a
multi-tenant shared physical infrastructure. The requirements
addresses the needs of virtual resources, applications reside on VMs,
and focus on the appliances that require only IP connectivity. Non-IP
communication is addressed by other documents 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
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 very
   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 edge functionality of the PE
   line-cards 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/end-device attaches to MPLS/BGP VPN, CE corresponds
   to a non-routing host that can reside in a VM 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 core 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. Further more, it is more
favorable if the solution can provide Open Source codes in public
forums, this will give the most flexibility and agility for SPs to
create new services.

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.

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.

Multi-access technologies, especially Ethernet, may also need to be
supported as transport networks, for example, 802.1ah.

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 consists in IP addresses used for communication
   within or between applications being 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 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 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.

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 with other services/networks, it
      should not be necessary to advertise 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 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. 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.

11.3. Service chaining

It is important to provide service chaining ability without major
impact to the existing protocols deployed. One solution currently
work in progress in IETF is [I-D.rfernando-l3vpn-service-chaining].

12. 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. Security issues associated
with deployments using MPLS-in-GRE or MPLS-in-IP encapsulations are
described in [RFC4023] in Section 8. And [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

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

[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

[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.

15.   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
    Cisco
    111 Wood Avenue South
    Iselin, NJ 08830, USA
    Email: luyuanf@gmail.com