Network working group
Internet Draft
Category: Standard Track
Expires: July 2011

X. Xu
Huawei Technologies

January 24, 2011

Virtual Subnet: A Scalable Data Center Network Architecture

draft-xu-virtual-subnet-04

Status of this Memo

Copyright Notice

Abstract

This document proposes a new IP-only L2VPN solution which uses BGP/MPLS IP VPN technology [RFC4364] with some extensions, together with some other proven technologies including ARP proxy [RFC925][RFC1027] to provide a more scalable IP-only L2VPN services across a MPLS/IP backbone. This solution is intended to be a scalable data center network architecture which can be deployed today as an alternative to the spanning tree protocol bridge technology.

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. Problem Statement

With the popularity of cloud services, the scale of today's data centers expands larger and larger. In addition, virtual machine migration technology, which allows a virtual machine to be able to migrate to any physical server while keeping the same IP address, is becoming more and more prevalent for achieving service agility in data centers. As a result, large Layer 2 networks are needed for server-to-server connectivity. Meanwhile, due to the huge-volume traffic exchanged between servers, the Layer 2 networks SHOULD provide enough capacity for server-to-server interconnections.

Unfortunately, today's data center network using the Spanning-Tree Protocol (STP) bridge technology, can not address the above challenges facing today's large-scale data centers in several ways. First, STP can calculate out only one single forwarding tree for all connected servers of a particular Virtual Local Area Network (VLAN) and it can not support multi-path routing, e.g., Equal Cost Multi-Path (ECMP), hence the available network capacity in data center networks can't be highly utilized so as to provide enough bandwidth between servers; Second, since the bridge forwarding is based on the flat MAC addresses, the scalability of the bridge forwarding table would become a big issue, especially when the existing large Layer 2 network scales even larger; Third, broadcast storm impacts imposed by some protocols, e.g., Address Resolution Protocol (ARP) and the flooding of unknown destination unicast frames become much more serious and unpredictable in the continually growing large-scale STP bridge networks.

2. Terminology

This memo makes use of the terms defined in [RFC4364], [MVPN], [RFC2236] and [RFC2131]. Below are provided terms specific to this document:

- Service Domain: A group of servers which are dedicated for a given service and are usually located on a separate IP subnet.

3. Design Goals

To overcome the limitations of the STP bridge networks as mentioned above, this document describes Virtual Subnet (VS), a new IP-only L2VPN solution which is intended to be a practical and scalable data center network architecture meeting the following objectives:

- Bandwidth Utilization Maximization

To provide enough bandwidth between servers, the server-to-server traffic SHOULD always be delivered along the shortest paths while multi-path routing is used for load-balancing purpose.

- Layer 2 Connectivity

To be backwards compatible with existing applications and protocols running in today's data centers (e.g., virtual machine migration), servers of a given service domain SHOULD be connected as if they were on a Local Area Network (LAN) or an IP subnet.

- Domain Isolation

To achieve performance and security isolation, servers belonging to different service domains SHOULD be isolated just as if they were located on separate Virtual LANs (VLAN) or IP subnets.

- Forwarding Table Scalability

To accommodate tens to hundreds of thousands of servers in a single data center network, the forwarding tables of those forwarding devices (e.g., routers or bridges) SHOULD be scalable enough.

- Broadcast Storm Suppression

To alleviate the serious impacts on network performance which are imposed by broadcast storms, broadcast domains SHOULD be limited to their smallest scopes.

4. Architecture Description

VS uses BGP/MPLS IP VPN technology [RFC4364] with some extensions, together with other proven technologies including ARP proxy [RFC925][RFC1027] to provide scalable IP-only L2VPN services across a MPLS/IP backbone.

Since VS constructs large-scale IP subnets, rather than real LANs, across the MPLS/IP backbone, the non-IP traffic would not be supported in VS anymore. However, given that IP traffic is the predominant type of traffic in today's data center networks and the non-IP traffic will disappear from the data center networks with the elapse of time, we believe that VS can be used as a practical data center network solution in most cases.

The following sections describe VS in detail.

4.1. Unicast

4.1.1. Unicast inside a Service Domain

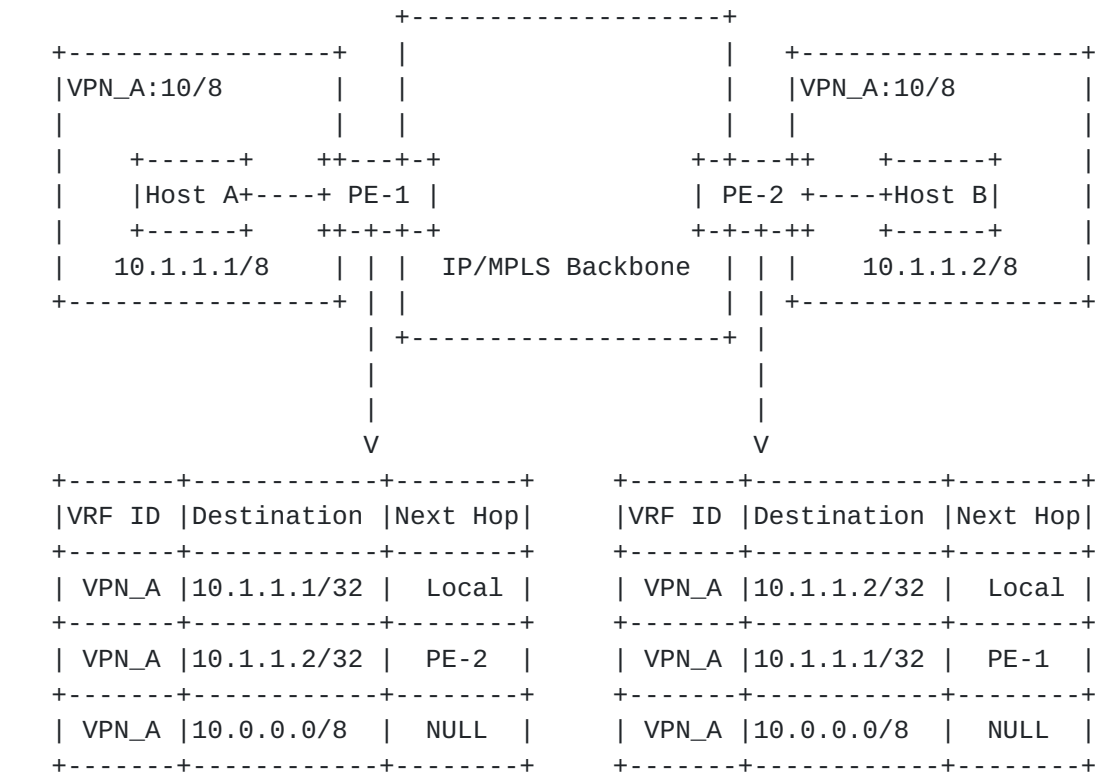

Figure 1: Unicast inside a Service Domain

As shown in Figure 1, BGP/MPLS IP VPN technology with some extensions is deployed in a data center network. To maintain proper isolation of one service domain from another, each service domain is mapped to a distinct VPN and servers of a given service domain, as Customer Edge (CE) hosts, are attached to Provider Edge (PE) routers directly or through one or more Ethernet bridges. In addition, to build large IP subnets across the MPLS/IP backbone, different sites of a particular VPN are associated with an identical IP subnet. PE routers create host routes for their local CE hosts automatically according to the corresponding ARP entries. Instead of distributing the routes for the configured VPN subnets, PE routers distribute host routes for their local CE hosts to each other. In addition, each PE router automatically creates a route for the configured VPN subnet whose next-hop is pointed to a null interface. With such special route, packets destined for the nonexistent hosts of that subnet will be discarded directly by the ingress PE routers. APR proxy is implemented on PE routers for every attached VPN, thus, upon receiving from a local CE host an ARP request for a remote CE

host, the PE as an ARP proxy returns its own MAC address as a response.

Assume host A broadcasts an ARP request for host B before communicating with B, upon the receipt of this ARP request, PE-1 lookups the associated VRF to find the host route for B. If found and the route is learnt from a remote PE, PE-1 acting as an ARP proxy returns its own MAC address in the response to that ARP request. Otherwise, no ARP reply SHOULD be sent. After obtaining the ARP reply from PE-1, A sends an IP packet to B with destination MAC address of PE-1's MAC address. Upon receiving this packet, PE-1 acting as an ingress PE, tunnels the packet towards PE-2 which in turn, as an egress PE, forwards the packet to B.

4.1.2. Unicast outside a Service Domain

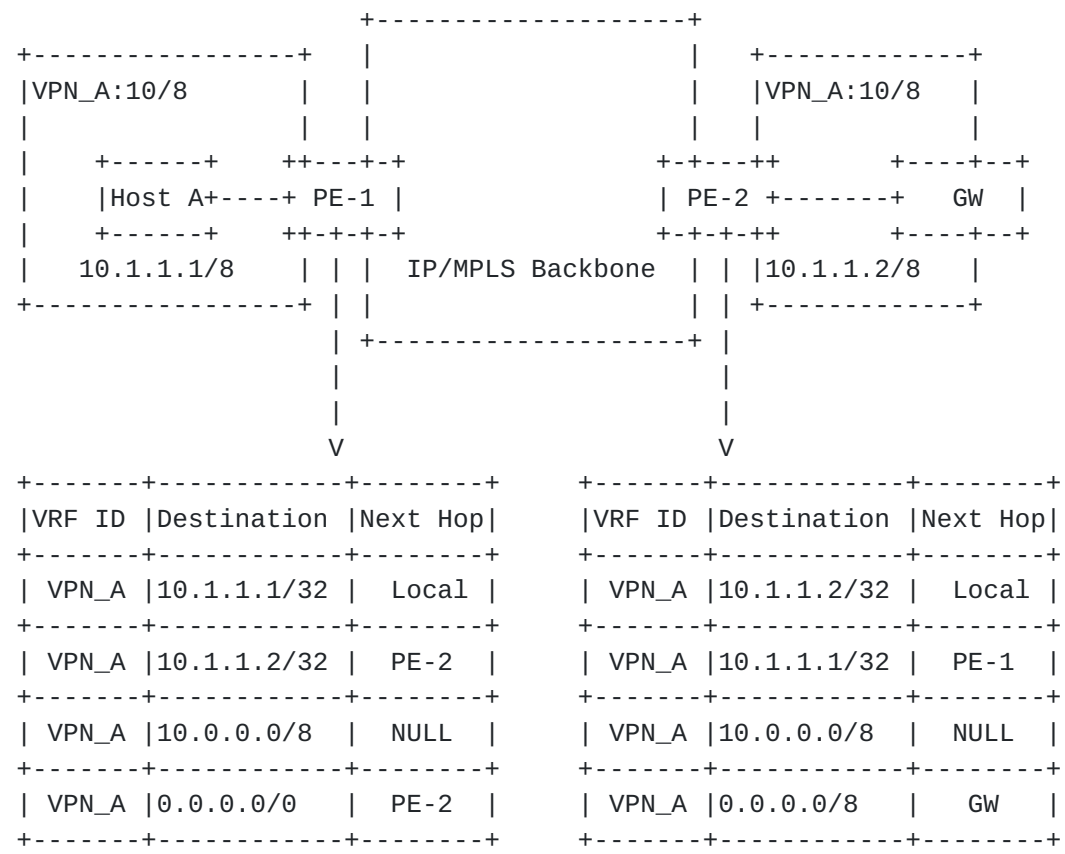

Figure 2: Unicast between Service Domains

As shown in Figure 2, for a CE host (e.g., host A) to communicate with other hosts outside its subnet, the PE router (e.g., PE-2) which is connected to the default gateway router (e.g., GW) for that VPN SHOULD be configured with a default route with the next-hop of

the default gateway router and then advertise such default route to other PE routers of the same VPN. Now host A sends an ARP request for its default gateway (i.e., GW) before sending a packet to a host outside its subnet. Upon receiving this ARP request, PE-1 lookups the associated VRF to find the host route for GW. If found and that found host route is learnt from a remote PE, PE-1 as an ARP proxy, returns its own MAC address in the ARP reply. Host A then sends an IP packet for the destination host with destination MAC address of PE-1's MAC. Upon receiving this packet, PE-1 as an ingress PE, tunnels it towards PE-2 according to the best route (i.e., the default route learnt from PE-2) for the destination host. PE-2 as an egress PE, in turn, forwards the received packet towards the default gateway router (i.e., GW). Due to the null route for the subnet, packets destined for those nonexistent CE hosts of that subnet would not be mistakenly forwarded to the default gateway router of that subnet.

In the scenario where more than one default gateway router running Virtual Router Redundancy Protocol (VRRP) [RFC2338] is connected to a given VPN for redundancy purpose, only the PE router which is connected to the Virtual Router Master SHOULD be allowed to announce a default route into that VPN. To achieve that goal, a default route with the next-hop of the corresponding Virtual Router IP address is configured for that VPN instance on each of the PE routers which are connected to the VRRP routers. In addition, the default route SHOULD not be deemed as valid until there is an active host route for its next-hop address. Since only the Virtual Router Master is allowed to respond to ARP requests for the Virtual Router IP address and broadcast gratuitous ARP requests containing the Virtual Router IP address, only the PE router which is connected to the Virtual Router Master could have an active ARP entry for the Virtual Router IP address and therefore could have an active host route for the Virtual Router IP address (i.e., the next-hop address of the configured default route). In this way, all packets destined for the outside would be sent to the corresponding Virtual Router Master.

4.2. Multicast/Broadcast

The MVPN technology [MVPN], in particular, the Protocol-Independent-Multicast (PIM) tree option with some extensions, is partially reused here to support IP multicast and broadcast between CE hosts of the same VPN. For example, PE routers attached to a given VPN join a default provider multicast distribution tree which is dedicated for that VPN. PE routers receiving customer multicast or broadcast traffic from local CE hosts forward such traffic to other remote PE routers over the corresponding default provider multicast distribution tree. When customer multicast or broadcast traffic is

received from a provider multicast distribution tree, PE routers forward such traffic to the associated VRF attachment circuits.

For the customer multicast group of a particular VPN which carries high-volume traffic and not all sites of that VPN need the traffic of that customer multicast group, a dedicated provider multicast distribution tree other than the default provider multicast distribution tree for that VPN can be assigned optionally. As a result, those PE routers of that VPN that have no local CE hosts which are interested in that customer multicast group will not receive such traffic from remote PE routers anymore.

More details about how to support multicast and broadcast traffic in VS will be explored in a later version of this document.

4.3. CE Host Discovery

To discover all local CE hosts including gateway routers, PE routers SHOULD perform at least once ARP scan on the attached VPN subnet after rebooting. For example, a PE broadcasts an ARP request for each IP address within the subnet of each attached VPN. Alternatively, this PE could also broadcast an ARP request for a directed broadcast address (i.e., 255.255.255.255) or an ALL-Systems multicast group address (i.e., 224.0.0.1), that is to say, the target protocol address field is filled with 2555.255.255.255 or 224.0.0.1. Any CE host receiving this ARP request SHOULD respond with an ARP reply containing its IP and MAC addresses. After a round of such ARP scan, the PE will discover all local CE hosts and cache their ARP entries in its ARP table. After that, the PE could send ARP requests in unicast to each already-learnt local CE host periodically so as to check whether the CE host is still present on the subnet. Using unicast ARP requests has the advantage that it is quieter than using the broadcast because it won't be received by all CE hosts on the subnet. When receiving a gratuitous ARP from a local CE host, the PE SHOULD cache the ARP entry of that CE host in its ARP table immediately if no ARP entry for that CE host exists yet. Otherwise, the PE SHOULD just update the corresponding ARP entry of that CE host. Most operating systems generate a gratuitous ARP request when the host boots up, the host's network interface or links comes up, or an address assigned to the interface changes. In the scarce scenarios where a host does not generate a gratuitous ARP, the PE would have to perform ARP scan periodically.

4.4. CE Host Multi-homing and Mobility

When a given PE receives a host route for one of its local CE hosts from a remote PE, it SHOULD immediately send an ARP request for that

CE host to the attached VPN subnet so as to determine whether that CE host is still connected locally. If an ARP reply is received in a short amount of time (imaging the CE host multi-homing scenario), the PE just needs to update the ARP entry for that CE host as normal. Otherwise (considering the virtual machine migration scenario), the PE SHOULD delete the ARP entry corresponding to that host from its APR table. Meanwhile, the PE SHOULD broadcast a gratuitous ARP on the attached VPN subnet on behalf of that CE host, with the sender hardware address field being filled with one of its own MAC addresses. As a result, the ARP entry for that CE host which has been cached on other local CE hosts is updated.

4.5. APR Proxy

The PE, acting as an ARP proxy, SHOULD only respond to the ARP requests for those CE hosts which have been learnt from other remote PE routers. Especially, the PE SHOULD not respond to ARP requests for local CE hosts. Otherwise, in case that the ARP reply from the PE covers that from the requested CE host, the packet for that local CE host which is sent from another local CE would be unnecessarily relayed by the PE.

When VRRP, together with ARP proxy is enabled on multiple PE routers which are attached to the same VPN site, only the PE acting as Virtual Router Master is delegated to perform ARP proxy function on the shared VPN subnet. In addition, it SHOULD use the Virtual Router MAC address in any ARP packet it sends.

4.6. DHCP Relay Agent

To avoid flooding Dynamic Host Configuration Protocol (DHCP) [RFC2131] broadcast messages through the data center network, DHCP Relay Agent can be implemented on PE routers for each attached VPN. Thus, DHCP broadcast messages received from DHCP clients on local CE hosts would be relayed by DHCP Relay Agents on PE routers to DHCP servers in unicast.

5. Comparison

5.1. VS vs VPLS

Virtual Private LAN Service (VPLS) [RFC4761, RFC4762] provides private LAN services for IP as well as other protocols. Since PE routers in VPLS work much similar as STP bridges, broadcast storm issues are intactly inherited from traditional STP bridge networks to VPLS.

At the cost of being lacking in support for non-IP traffic, VS alleviates the broadcast storm issues by using Layer 3 routing and ARP proxy technologies on PE routers.

In addition, if CE hosts of multiple VPNs are attached to a PE router through an intermediate Ethernet bridge, in VPLS, this intermediate bridge would have to learn the MAC addresses of both local CE hosts and remote CE hosts of these attached VPNs. However, in VS, such intermediate bridge only needs to learn MAC addresses of local CE hosts and local PE routers due to the ARP proxy implemented on PE routers.

5.2. VS vs IPLS

Both VS and IP LAN Service (IPLS) [IPLS] are IP only L2VPN technologies.

However, IPLS is different from VS in several aspects. First, in IPLS, ARP packets even including the unicast ARP reply packets are forwarded from attachment circuits to "multicast" PWs and the received APR packets from the "multicast" PWs will be flooded to all CE hosts (although broadcast ARP request packets can be suppressed by PE routers on which there are matching ARP entries for the ARP requests in their ARP caches). As a result, the broadcast storm imposed by ARP traffic is worsened to some extent, rather than being alleviated. In contrast, by using ARP Proxy on PE routers in VS, ARP traffic is limited within small network scopes. Second, as said in [IPLS], "An IP frame received over a unicast PW is prepended with a MAC header before transmitting it on the appropriate attachment circuits and the source MAC address is the PE router's own local MAC address or a MAC address which has been specially configured on the PE router for this use." However, the destination MAC address of the packet to a remote CE host which is sent from a local CE host is the MAC of the remote CE host, rather than the local PE router's MAC. Thus, flooding unknown destination unicast frames on the above Ethernet bridges would not be avoided anymore unless these intermediary bridges are configured to not age out the learned MAC entries (whether such configuration has any side-effects is uncertain). In contrast, such intermediate bridges in VS only need to learn MAC addresses of local CE hosts and local PE routers. Third, IPLS prohibits connection of a common LAN or VLAN to more than one PE router. In other words, IPLS can not allow CE hosts to be multi-homed to multiple PE routers for redundancy and load-balancing. In contrast, VS can support CE multi-homing easily.

6. Conclusion

By using Layer 3 routing on the backbone of the data center network to replace the STP bridge forwarding, traffic between any two servers is forwarded along shortest paths between them and multi-path routing is easily achieved. Thus, the totally available bandwidth in data center networks is utilized to the maximum extent.

By reusing the BGP/MPLS IP VPN technology to build large IP subnets across the backbones of data center networks, servers of a given VPN are allowed to communicate with each other just as if they were on the same subnet.

Due to the BGP/MPLS IP VPN technology, forwarding tables of P routers are sized to the number of PE routers rather than the total number of CE hosts. Meanwhile, forwarding tables of PE routers can also scale well by distributing VPN instances and their corresponding routing tables among multiple PE routers. Especially, thanks to the Outbound Route Filtering (ORF) capability of BGP, PE routers only needs to maintain the routing tables of their attached VPN instances. Thus, the forwarding table scalability issues with today's data center networks are largely alleviated.

By enabling APR proxy function on PE routers, ARP broadcast messages from local CE hosts are blocked by local PE routers. Thus, APR broadcast messages will not flood the whole data center network. Besides, by enabling DHCP Relay Agent function on PE routers, DHCP broadcast messages from local CE hosts are intercepted by DHCP Relay Agents and forwarded to DHCP servers in unicast. Thus, the broadcast storms in data center networks are largely suppressed.

7. Future work

How to support IPv6 CE hosts in VS is for future study.

8. Security Considerations

TBD.

9. IANA Considerations

There is no requirement for IANA.

10. Acknowledgements

Thanks to Dino Farinacci for his valuable 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.

11.2. Informative References

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

[MVPN] Rosen. E and Aggarwal. R, "Multicast in MPLS/BGP IP VPNs", draft-ietf-l3vpn-2547bis-mcast-10.txt (work in progress), Janurary 2010.

[MVPN-BGP] R. Aggarwal, E. Rosen, T. Morin, Y. Rekhter, C. Kodeboniya, "BGP Encodings for Multicast in MPLS/BGP IP VPNs", draft-ietf-l3vpn-2547bis-mcast-bgp-08.txt (work in progress), September 2009.

[RFC826] Plummer, D., "An Ethernet Address Resolution Protocol or Converting Network Protocol Addresses to 48-bit Ethernet Addresses for Transmission on Ethernet Hardware", RFC-826, Symbolics, November 1982.

[RFC925] Postel, J., "Multi-LAN Address Resolution", RFC-925, USC Information Sciences Institute, October 1984.

[RFC1027] Smoot Carl-Mitchell, John S. Quarterman, "Using ARP to Implement Transparent Subnet Gateways", RFC 1027, October 1987.

[RFC2338] Knight, S., et. al., "Virtual Router Redundancy Protocol", RFC 2338, April 1998.

[RFC2131] Droms, R., "Dynamic Host Configuration Protocol", RFC 2131, March 1997.

[RFC2236] Fenner, W., "Internet Group Management Protocol, Version 2", RFC 2236, November 1997.

[RFC4761] Kompella, K. and Y. Rekhter, "Virtual Private LAN Service (VPLS) Using BGP for Auto-Discovery and Signaling", RFC 4761, January 2007.

[RFC4762] Lasserre, M. and V. Kompella, "Virtual Private LAN Service (VPLS) Using Label Distribution Protocol (LDP) Signaling", RFC 4762, January 2007.

[IPLS] H. Shah., et. al., "IP-Only LAN Service (IPLS)", draft-ietf-l2vpn-ipls-09.txt (work in progress), February 2010.

Authors' Addresses

Xiaohu Xu
Huawei Technologies,
Hai-Dian District, Beijing 100085, P.R. China
Phone: +86 10 82882573
Email: xuxh@huawei.com