

Network Working Group  
Internet-Draft  
Intended status: Informational  
Expires: May 14, 2016

X. Xu  
Huawei  
R. Raszuk  
Bloomberg LP  
C. Jacquenet  
Orange  
T. Boyes  
Bloomberg LP  
B. Fee  
Extreme Networks  
November 11, 2015

**Virtual Subnet: A BGP/MPLS IP VPN-based Subnet Extension Solution**  
**draft-ietf-bess-virtual-subnet-05**

Abstract

This document describes a BGP/MPLS IP VPN-based subnet extension solution referred to as Virtual Subnet, which can be used for building Layer 3 network virtualization overlays within and/or between data centers.

Status of This Memo

This Internet-Draft is submitted in full conformance with the provisions of [BCP 78](#) and [BCP 79](#).

Internet-Drafts are working documents of the Internet Engineering Task Force (IETF). Note that other groups may also distribute working documents as Internet-Drafts. The list of current Internet-Drafts is at <http://datatracker.ietf.org/drafts/current/>.

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."

This Internet-Draft will expire on May 14, 2016.

Copyright Notice

Copyright (c) 2015 IETF Trust and the persons identified as the document authors. All rights reserved.

This document is subject to [BCP 78](#) and the IETF Trust's Legal Provisions Relating to IETF Documents (<http://trustee.ietf.org/license-info>) in effect on the date of

publication of this document. Please review these documents carefully, as they describe your rights and restrictions with respect to this document. Code Components extracted from this document must include Simplified BSD License text as described in Section 4.e of the Trust Legal Provisions and are provided without warranty as described in the Simplified BSD License.

## Table of Contents

<a href="#">1.</a>	<a href="#">Introduction . . . . .</a>	<a href="#">2</a>
<a href="#">2.</a>	<a href="#">Terminology . . . . .</a>	<a href="#">4</a>
<a href="#">3.</a>	<a href="#">Solution Description . . . . .</a>	<a href="#">4</a>
<a href="#">3.1.</a>	<a href="#">Unicast . . . . .</a>	<a href="#">4</a>
<a href="#">3.1.1.</a>	<a href="#">Intra-subnet Unicast . . . . .</a>	<a href="#">4</a>
<a href="#">3.1.2.</a>	<a href="#">Inter-subnet Unicast . . . . .</a>	<a href="#">5</a>
<a href="#">3.2.</a>	<a href="#">Multicast . . . . .</a>	<a href="#">8</a>
<a href="#">3.3.</a>	<a href="#">Host Discovery . . . . .</a>	<a href="#">9</a>
<a href="#">3.4.</a>	<a href="#">ARP/ND Proxy . . . . .</a>	<a href="#">9</a>
<a href="#">3.5.</a>	<a href="#">Host Mobility . . . . .</a>	<a href="#">9</a>
<a href="#">3.6.</a>	<a href="#">Forwarding Table Scalability on Data Center Switches . .</a>	<a href="#">10</a>
<a href="#">3.7.</a>	<a href="#">ARP/ND Cache Table Scalability on Default Gateways . .</a>	<a href="#">10</a>
<a href="#">3.8.</a>	<a href="#">ARP/ND and Unknown Uncast Flood Avoidance . . . . .</a>	<a href="#">10</a>
<a href="#">3.9.</a>	<a href="#">Path Optimization . . . . .</a>	<a href="#">10</a>
<a href="#">4.</a>	<a href="#">Limitations . . . . .</a>	<a href="#">11</a>
<a href="#">4.1.</a>	<a href="#">Non-support of Non-IP Traffic . . . . .</a>	<a href="#">11</a>
<a href="#">4.2.</a>	<a href="#">Non-support of IP Broadcast and Link-local Multicast . .</a>	<a href="#">11</a>
<a href="#">4.3.</a>	<a href="#">TTL and Traceroute . . . . .</a>	<a href="#">11</a>
<a href="#">5.</a>	<a href="#">Acknowledgements . . . . .</a>	<a href="#">12</a>
<a href="#">6.</a>	<a href="#">IANA Considerations . . . . .</a>	<a href="#">12</a>
<a href="#">7.</a>	<a href="#">Security Considerations . . . . .</a>	<a href="#">12</a>
<a href="#">8.</a>	<a href="#">References . . . . .</a>	<a href="#">12</a>
<a href="#">8.1.</a>	<a href="#">Normative References . . . . .</a>	<a href="#">12</a>
<a href="#">8.2.</a>	<a href="#">Informative References . . . . .</a>	<a href="#">13</a>
	<a href="#">Authors' Addresses . . . . .</a>	<a href="#">13</a>

## [1. Introduction](#)

For business continuity purpose, Virtual Machine (VM) migration across data centers is commonly used in situations such as data center maintenance, data center migration, data center consolidation, data center expansion, and data center disaster avoidance. It's generally admitted that IP renumbering of servers (i.e., VMs) after the migration is usually complex and costly at the risk of extending the business downtime during the process of migration. To allow the migration of a VM from one data center to another without IP renumbering, the subnet on which the VM resides needs to be extended across these data centers.



To achieve subnet extension across multiple Infrastructure-as-a-Service (IaaS) cloud data centers in a scalable way, the following requirements and challenges must be considered:

- a. **VPN Instance Space Scalability:** In a modern cloud data center environment, thousands or even tens of thousands of tenants could be hosted over a shared network infrastructure. For security and performance isolation purposes, these tenants need to be isolated from one another.
- b. **Forwarding Table Scalability:** With the development of server virtualization technologies, it's not uncommon for a single cloud data center to contain millions of VMs. This number already implies a big challenge on the forwarding table scalability of data center switches. Provided multiple data centers of such scale were interconnected at Layer 2, this challenge would become even worse.
- c. **ARP/ND Cache Table Scalability:** [[RFC6820](#)] notes that the Address Resolution Protocol (ARP)/Neighbor Discovery (ND) cache tables maintained on default gateways within cloud data centers can raise scalability issues. Therefore, it's very useful if the ARP/ND cache table size could be prevented from growing by multiples as the number of data centers to be connected increases.
- d. **ARP/ND and Unknown Unicast Flooding:** It's well-known that the flooding of ARP/ND broadcast/multicast and unknown unicast traffic within large Layer 2 networks would affect the performance of networks and hosts. As multiple data centers with each containing millions of VMs are interconnected at Layer 2, the impact of flooding as mentioned above would become even worse. As such, it becomes increasingly important to avoid the flooding of ARP/ND broadcast/multicast and unknown unicast traffic across data centers.
- e. **Path Optimization:** A subnet usually indicates a location in the network. However, when a subnet has been extended across multiple geographically dispersed data center locations, the location semantics of such subnet is not retained any longer. As a result, the traffic between a specific user and server, in different data centers, may first be routed through a third data center. This suboptimal routing would obviously result in an unnecessary consumption of the bandwidth resource between data centers. Furthermore, in the case where traditional VPLS technology [[RFC4761](#)] [[RFC4762](#)] is used for data center interconnect, return traffic from a server may be forwarded to a default gateway located in a different data center due to the



configuration in a virtual router redundancy group. This suboptimal routing would also unnecessarily consume the bandwidth resource between data centers.

This document describes a BGP/MPLS IP VPN-based subnet extension solution referred to as Virtual Subnet, which can be used for data center interconnection while addressing all of the requirements and challenges as mentioned above. Here the BGP/MPLS IP VPN means both BGP/MPLS IPv4 VPN [[RFC4364](#)] and BGP/MPLS IPv6 VPN [[RFC4659](#)]. In addition, since Virtual Subnet is mainly built on proven technologies such as BGP/MPLS IP VPN and ARP/ND proxy [[RFC0925](#)][[RFC1027](#)][[RFC4389](#)], those service providers offering IaaS public cloud services could rely upon their existing BGP/MPLS IP VPN infrastructures and their corresponding experiences to realize data center interconnection.

Although Virtual Subnet is described in this document as an approach for data center interconnection, it actually could be used within data centers as well.

Note that the approach described in this document is not intended to achieve an exact emulation of Layer 2 connectivity and therefore it can only support a restricted Layer 2 connectivity service model with limitations declared in [Section 4](#). As for the discussion about in which environment this service model should be suitable, it's outside the scope of this document.

## **[2. Terminology](#)**

This memo makes use of the terms defined in [[RFC4364](#)].

## **[3. Solution Description](#)**

### **[3.1. Unicast](#)**

#### **[3.1.1. Intra-subnet Unicast](#)**



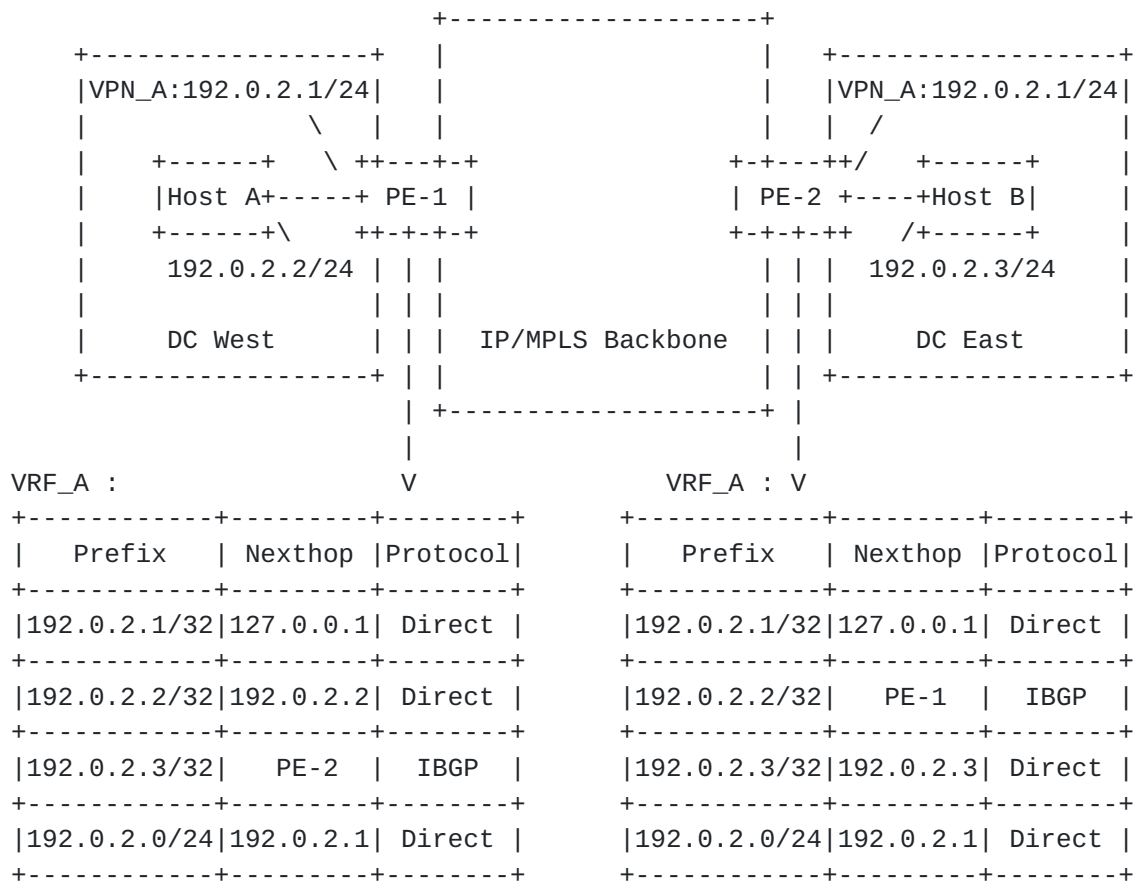


Figure 1: Intra-subnet Unicast Example

As shown in Figure 1, two hosts (i.e., Hosts A and B) belonging to the same subnet (i.e., 192.0.2.0/24) are located at different data centers (i.e., DC West and DC East) respectively. PE routers (i.e., PE-1 and PE-2) which are used for interconnecting these two data centers create host routes for their own local hosts respectively and then advertise them via the BGP/MPLS IP VPN signaling. Meanwhile, an ARP proxy is enabled on VRF attachment circuits of these PE routers.

Now assume host A sends an ARP request for host B before communicating with host B. Upon receiving the ARP request, PE-1 acting as an ARP proxy returns its own MAC address as a response. Host A then sends IP packets for host B to PE-1. PE-1 tunnels such packets towards PE-2 which in turn forwards them to host B. Thus, hosts A and B can communicate with each other as if they were located within the same subnet.

### [3.1.2. Inter-subnet Unicast](#)





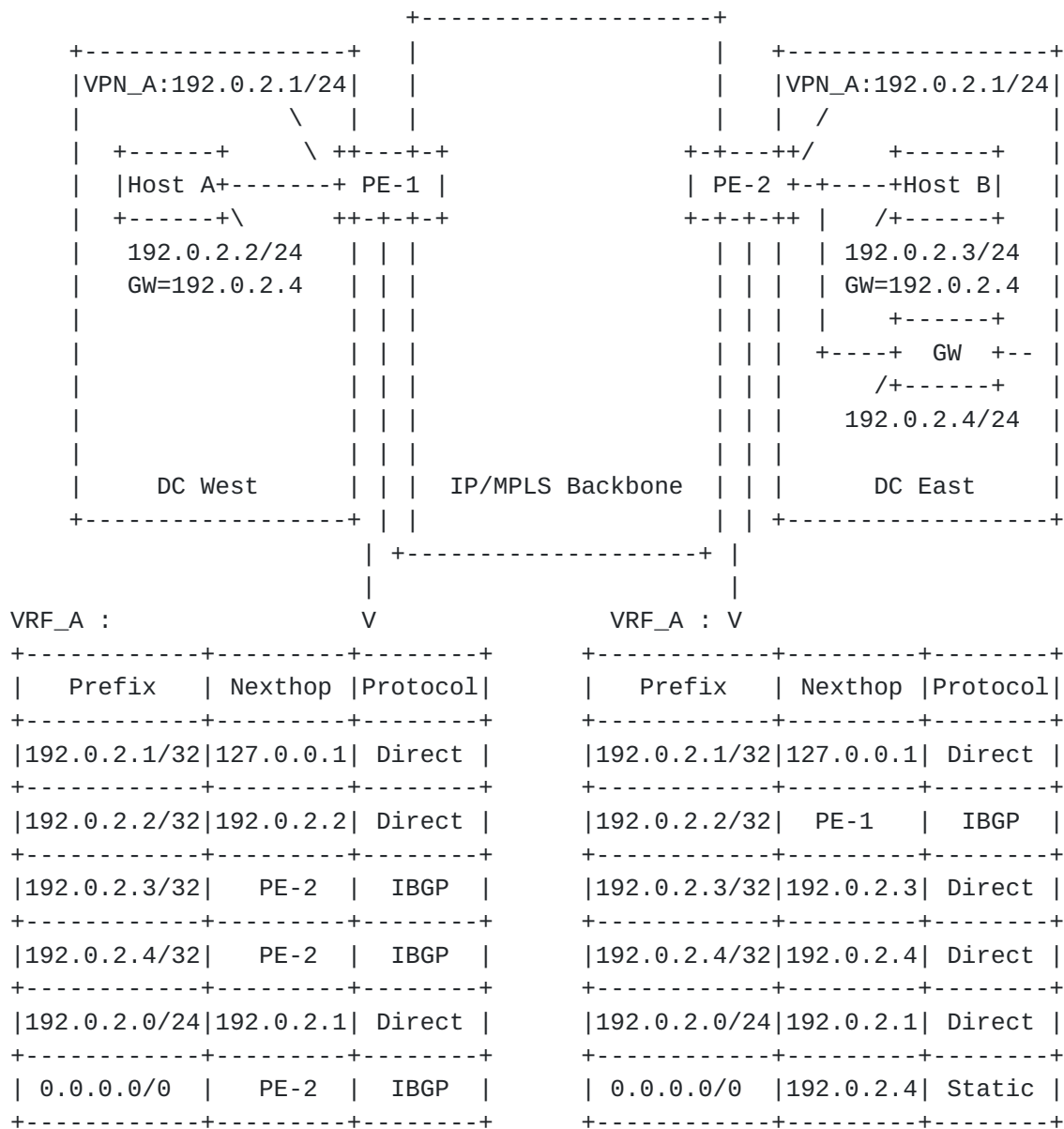


Figure 2: Inter-subnet Unicast Example (1)

As shown in Figure 2, only one data center (i.e., DC East) is deployed with a default gateway (i.e., GW). PE-2 which is connected to GW would either be configured with or learn from GW a default route with next-hop being pointed to GW. Meanwhile, this route is distributed to other PE routers (i.e., PE-1) as per normal [\[RFC4364\]](#) operation. Assume host A sends an ARP request for its default gateway (i.e., 192.0.2.4) prior to communicating with a destination host outside of its subnet. Upon receiving this ARP request, PE-1 acting as an ARP proxy returns its own MAC address as a response. Host A then sends a packet for Host B to PE-1. PE-1 tunnels such packet towards PE-2 according to the default route learnt from PE-2, which in turn forwards that packet to GW.



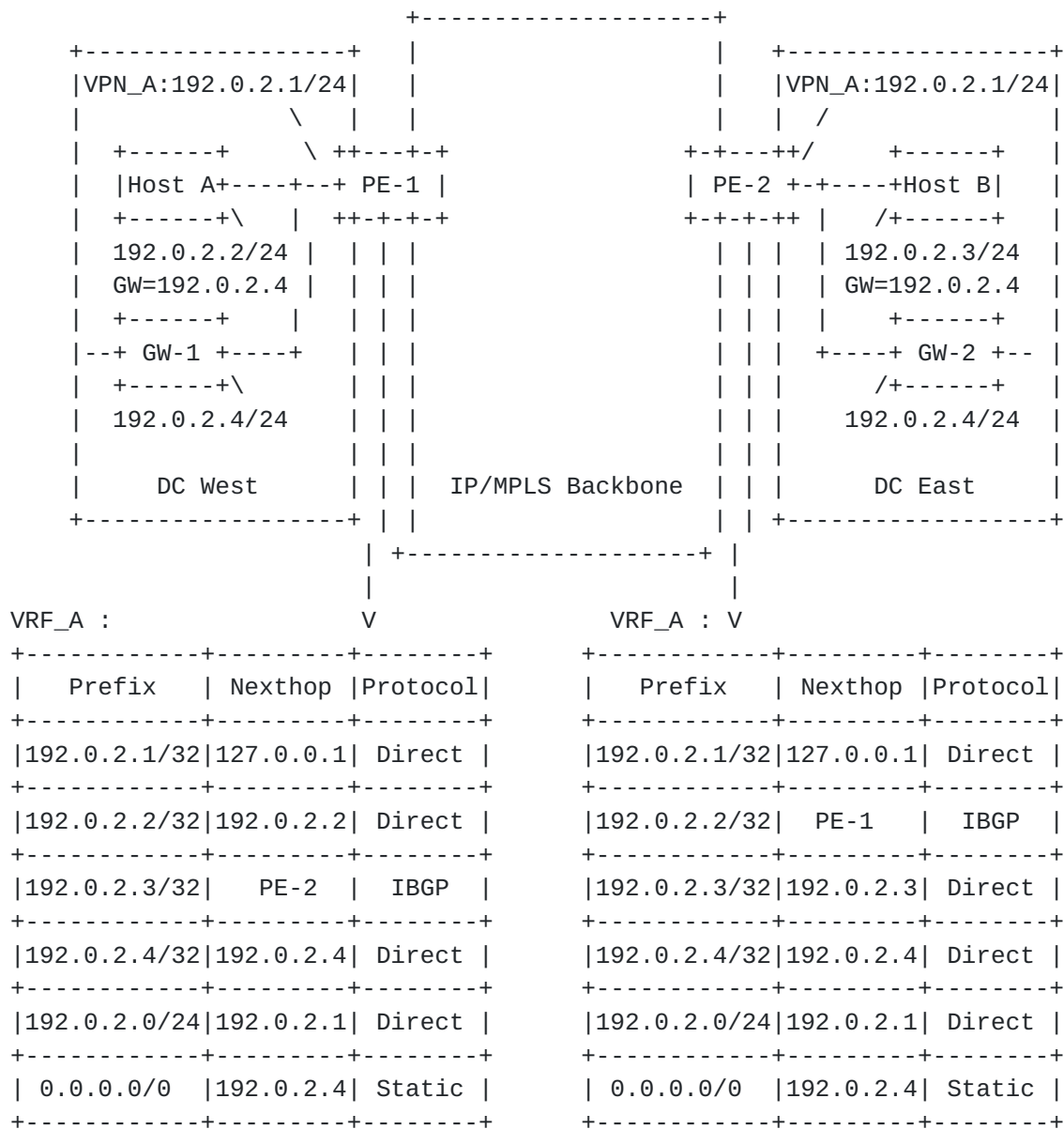


Figure 3: Inter-subnet Unicast Example (2)

As shown in Figure 3, in the case where each data center is deployed with a default gateway, hosts will get ARP responses directly from their local default gateways, rather than from their local PE routers when sending ARP requests for their default gateways.



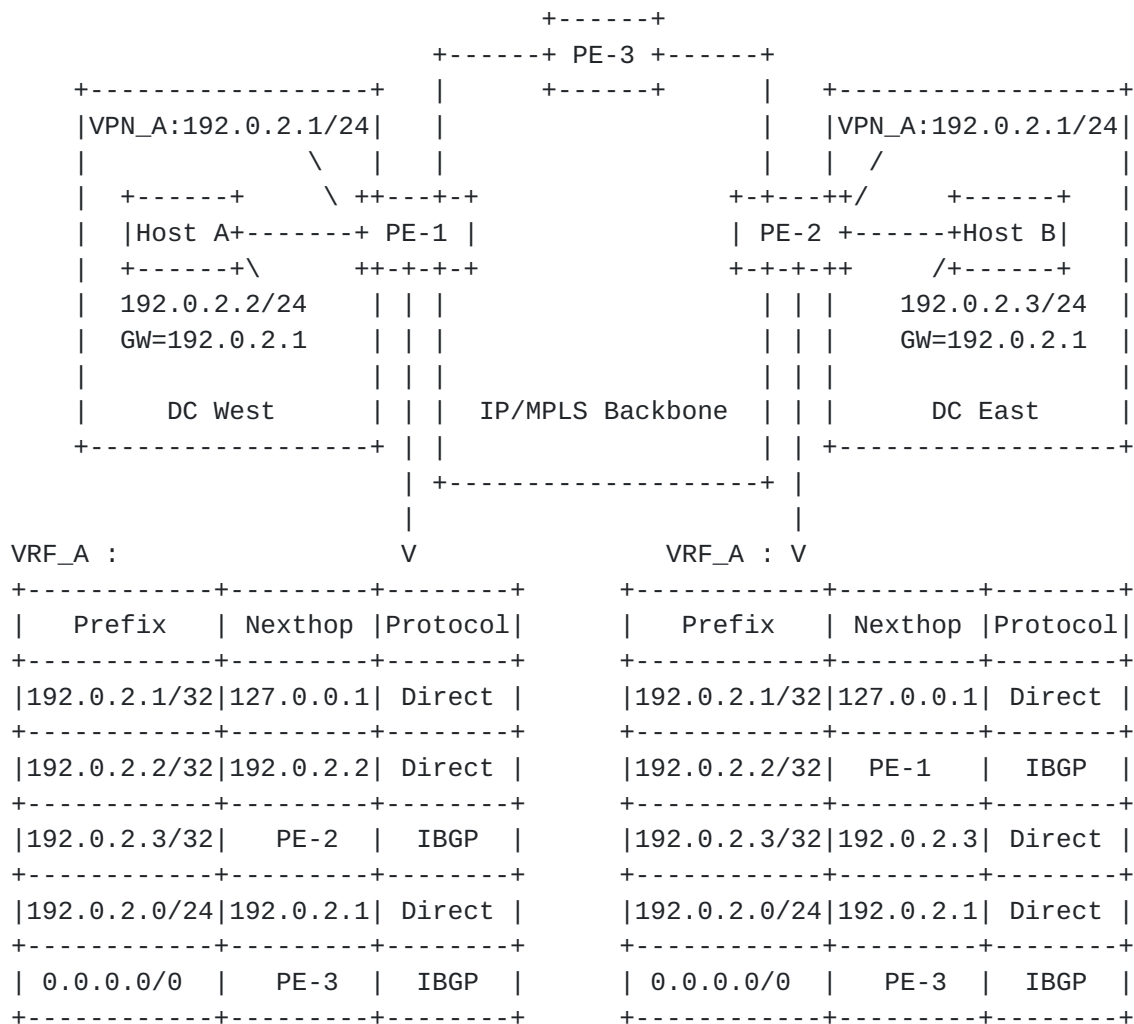


Figure 4: Inter-subnet Unicast Example (3)

Alternatively, as shown in Figure 4, PE routers themselves could be directly configured as default gateways of their locally connected hosts as long as these PE routers have routes for outside networks.

### 3.2. Multicast

To support IP multicast between hosts of the same Virtual Subnet, MVPN technologies [RFC6513] could be directly used without any change. For example, PE routers attached to a given VPN join a default provider multicast distribution tree which is dedicated for that VPN. Ingress PE routers, upon receiving multicast packets from their local hosts, forward them towards remote PE routers through the corresponding default provider multicast distribution tree. Note that here the IP multicast doesn't include link-local multicast.



### **3.3. Host Discovery**

PE routers should be able to discover their local hosts and keep the list of these hosts up to date in a timely manner so as to ensure the availability and accuracy of the corresponding host routes originated from them. PE routers could accomplish local host discovery by some traditional host discovery mechanisms using ARP or ND protocols.

### **3.4. ARP/ND Proxy**

Acting as an ARP or ND proxies, a PE routers should only respond to an ARP request or Neighbor Solicitation (NS) message for a target host when it has a best route for that target host in the associated VRF and the outgoing interface of that best route is different from the one over which the ARP request or NS message is received. In the scenario where a given VPN site (i.e., a data center) is multi-homed to more than one PE router via an Ethernet switch or an Ethernet network, Virtual Router Redundancy Protocol (VRRP) [[RFC5798](#)] is usually enabled on these PE routers. In this case, only the PE router being elected as the VRRP Master is allowed to perform the ARP/ND proxy function.

### **3.5. Host Mobility**

During the VM migration process, the PE router to which the moving VM is now attached would create a host route for that host upon receiving a notification message of VM attachment (e.g., a gratuitous ARP or unsolicited NA message). The PE router to which the moving VM was previously attached would withdraw the corresponding host route when receiving a notification message of VM detachment (e.g., a VDP message about VM detachment). Meanwhile, the latter PE router could optionally broadcast a gratuitous ARP or send an unsolicited NA message on behalf of that host with source MAC address being one of its own. In this way, the ARP/ND entry of this host that moved and which has been cached on any local host would be updated accordingly. In the case where there is no explicit VM detachment notification mechanism, the PE router could also use the following trick to determine the VM detachment event: upon learning a route update for a local host from a remote PE router for the first time, the PE router could immediately check whether that local host is still attached to it by some means (e.g., ARP/ND PING and/or ICMP PING). It is important to ensure that the same MAC and IP are associated to the default gateway active in each data center, as the VM would most likely continue to send packets to the same default gateway address after migrated from one data center to another. One possible way to achieve this goal is to configure the same VRRP group on each location so as to ensure the default gateway active in each data center share the same virtual MAC and virtual IP addresses.





### **3.6. Forwarding Table Scalability on Data Center Switches**

In a Virtual Subnet environment, the MAC learning domain associated with a given Virtual Subnet which has been extended across multiple data centers is partitioned into segments and each segment is confined within a single data center. Therefore data center switches only need to learn local MAC addresses, rather than learning both local and remote MAC addresses.

### **3.7. ARP/ND Cache Table Scalability on Default Gateways**

When default gateway functions are implemented on PE routers as shown in Figure 4, the ARP/ND cache table on each PE router only needs to contain ARP/ND entries of local hosts. As a result, the ARP/ND cache table size would not grow as the number of data centers to be connected increases.

### **3.8. ARP/ND and Unknown Uncast Flood Avoidance**

In a Virtual Subnet environment, the flooding domain associated with a given Virtual Subnet that has been extended across multiple data centers, is partitioned into segments and each segment is confined within a single data center. Therefore, the performance impact on networks and servers imposed by the flooding of ARP/ND broadcast/multicast and unknown unicast traffic is alleviated.

### **3.9. Path Optimization**

Take the scenario shown in Figure 4 as an example, to optimize the forwarding path for the traffic between cloud users and cloud data centers, PE routers located at cloud data centers (i.e., PE-1 and PE-2), which are also acting as default gateways, propagate host routes for their own local hosts respectively to remote PE routers which are attached to cloud user sites (i.e., PE-3). As such, the traffic from cloud user sites to a given server on the Virtual Subnet which has been extended across data centers would be forwarded directly to the data center location where that server resides, since the traffic is now forwarded according to the host route for that server, rather than the subnet route. Furthermore, for the traffic coming from cloud data centers and forwarded to cloud user sites, each PE router acting as a default gateway would forward the traffic according to the best-match route in the corresponding VRF. As a result, the traffic from data centers to cloud user sites is forwarded along an optimal path as well.



## **4. Limitations**

### **4.1. Non-support of Non-IP Traffic**

Although most traffic within and across data centers is IP traffic, there may still be a few legacy clustering applications which rely on non-IP communications (e.g., heartbeat messages between cluster nodes). Since Virtual Subnet is strictly based on L3 forwarding, those non-IP communications cannot be supported in the Virtual Subnet solution. In order to support those few non-IP traffic (if present) in the environment where the Virtual Subnet solution has been deployed, the approach following the idea of "route all IP traffic, bridge non-IP traffic" could be considered. That's to say, all IP traffic including both intra-subnet and inter-subnet would be processed by the Virtual Subnet process, while the non-IP traffic would be resorted to a particular Layer 2 VPN approach. Such unified L2/L3 VPN approach requires ingress PE routers to classify the traffic received from hosts before distributing them to the corresponding L2 or L3 VPN forwarding processes. Note that more and more cluster vendors are offering clustering applications based on Layer 3 interconnection.

### **4.2. Non-support of IP Broadcast and Link-local Multicast**

As illustrated before, intra-subnet traffic is forwarded at Layer 3 in the Virtual Subnet solution. Therefore, IP broadcast and link-local multicast traffic cannot be supported by the Virtual Subnet solution. In order to support the IP broadcast and link-local multicast traffic in the environment where the Virtual Subnet solution has been deployed, the unified L2/L3 overlay approach as described in [Section 4.1](#) could be considered as well. That's to say, the IP broadcast and link-local multicast would be resorted to the L2VPN forwarding process while the routable IP traffic would be processed by the Virtual Subnet process.

### **4.3. TTL and Traceroute**

As illustrated before, intra-subnet traffic is forwarded at Layer 3 in the Virtual Subnet context. Since it doesn't require any change to the TTL handling mechanism of the BGP/MPLS IP VPN, when doing a traceroute operation on one host for another host (assuming that these two hosts are within the same subnet but are attached to different sites), the traceroute output would reflect the fact that these two hosts within the same subnet are actually connected via an Virtual Subnet, rather than a Layer 2 connection since the PE routers to which those two host are connected respectively would be displayed in the traceroute output. In addition, for any other applications which generate intra-subnet traffic with TTL set to 1, these



applications may not be workable in the Virtual Subnet context, unless special TTL processing for such case has been implemented (e.g., if the source and destination addresses of a packet whose TTL is set to 1 belong to the same extended subnet, neither ingress nor egress PE routers should decrement the TTL of such packet. Furthermore, the TTL of such packet should not be copied into the TTL of the transport tunnel and vice versa).

## **5. Acknowledgements**

Thanks to Susan Hares, Yongbing Fan, Dino Farinacci, Himanshu Shah, Nabil Bitar, Giles Heron, Ronald Bonica, Monique Morrow, Rajiv Asati, Eric Osborne, Thomas Morin, Martin Vigoureux, Pedro Roque Marque, Joe Touch and Wim Henderickx for their valuable comments and suggestions on this document. Thanks to Loa Andersson for his WG LC review on this document. Thanks to Alvaro Retana for his AD review on this document. Thanks to Ronald Bonica for his RtgDir review.

## **6. IANA Considerations**

There is no requirement for any IANA action.

## **7. 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.

## **8. References**

### **8.1. Normative References**

- [RFC0925] Postel, J., "Multi-LAN address resolution", [RFC 925](#), DOI 10.17487/RFC0925, October 1984, <<http://www.rfc-editor.org/info/rfc925>>.
- [RFC1027] Carl-Mitchell, S. and J. Quarterman, "Using ARP to implement transparent subnet gateways", [RFC 1027](#), DOI 10.17487/RFC1027, October 1987, <<http://www.rfc-editor.org/info/rfc1027>>.
- [RFC4364] Rosen, E. and Y. Rekhter, "BGP/MPLS IP Virtual Private Networks (VPNs)", [RFC 4364](#), DOI 10.17487/RFC4364, February 2006, <<http://www.rfc-editor.org/info/rfc4364>>.
- [RFC4389] Thaler, D., Talwar, M., and C. Patel, "Neighbor Discovery Proxies (ND Proxy)", [RFC 4389](#), DOI 10.17487/RFC4389, April 2006, <<http://www.rfc-editor.org/info/rfc4389>>.



## 8.2. Informative References

- [RFC4659] De Clercq, J., Ooms, D., Carugi, M., and F. Le Faucheur, "BGP-MPLS IP Virtual Private Network (VPN) Extension for IPv6 VPN", [RFC 4659](#), DOI 10.17487/RFC4659, September 2006, <<http://www.rfc-editor.org/info/rfc4659>>.
- [RFC4761] Kompella, K., Ed. and Y. Rekhter, Ed., "Virtual Private LAN Service (VPLS) Using BGP for Auto-Discovery and Signaling", [RFC 4761](#), DOI 10.17487/RFC4761, January 2007, <<http://www.rfc-editor.org/info/rfc4761>>.
- [RFC4762] Lasserre, M., Ed. and V. Kompella, Ed., "Virtual Private LAN Service (VPLS) Using Label Distribution Protocol (LDP) Signaling", [RFC 4762](#), DOI 10.17487/RFC4762, January 2007, <<http://www.rfc-editor.org/info/rfc4762>>.
- [RFC5798] Nadas, S., Ed., "Virtual Router Redundancy Protocol (VRRP) Version 3 for IPv4 and IPv6", [RFC 5798](#), DOI 10.17487/RFC5798, March 2010, <<http://www.rfc-editor.org/info/rfc5798>>.
- [RFC6513] Rosen, E., Ed. and R. Aggarwal, Ed., "Multicast in MPLS/BGP IP VPNs", [RFC 6513](#), DOI 10.17487/RFC6513, February 2012, <<http://www.rfc-editor.org/info/rfc6513>>.
- [RFC6820] Narten, T., Karir, M., and I. Foo, "Address Resolution Problems in Large Data Center Networks", [RFC 6820](#), DOI 10.17487/RFC6820, January 2013, <<http://www.rfc-editor.org/info/rfc6820>>.

### Authors' Addresses

Xiaohu Xu  
Huawei

Email: [xuxiaohu@huawei.com](mailto:xuxiaohu@huawei.com)

Robert Raszuk  
Bloomberg LP

Email: [robert@raszuk.net](mailto:robert@raszuk.net)





Christian Jacquenet  
Orange

Email: christian.jacquenet@orange.com

Truman Boyes  
Bloomberg LP

Email: tboyes@bloomberg.net

Brendan Fee  
Extreme Networks

Email: bfee@extremenetworks.com