Network working group
Internet Draft
Category: Informational

X. Xu
Huawei Technologies

S. Hares

Y. Fan
China Telecom

C. Jacquenet
France Telecom

Expires: January 2014

July 15, 2013

Virtual Subnet: A L3VPN-based Subnet Extension Solution

draft-xu-l3vpn-virtual-subnet-00

Abstract

This document describes a Layer3 Virtual Private Network (L3VPN)-based subnet extension solution referred to as Virtual Subnet, which can be used as a kind of Layer3 network virtualization overlay approach for data center interconnect.

Status of this Memo

This Internet-Draft will expire on January 15, 2014.

Copyright Notice

Conventions used in this document

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

Table of Contents

1. Introduction

For business continuity purposes, Virtual Machine (VM) migration across data centers is commonly used in those 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.

In Infrastructure-as-a-Service (IaaS) cloud data center environments, to achieve subnet extension across multiple data centers in a scalable way, the following requirements SHOULD be considered for any data center interconnect solution:

1) 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. Hence, the data center interconnect solution SHOULD be capable of providing a large enough Virtual Private Network (VPN) instance space for tenant isolation.

2) Forwarding Table Scalability

With the development of server virtualization technologies, a single cloud data center containing millions of VMs is not uncommon. This number already implies a big challenge for data center switches, especially for core/aggregation switches, from the perspective of forwarding table scalability. Provided that multiple data centers of such scale were interconnected at layer2, this challenge would be even worse. Hence an ideal data center interconnect solution SHOULD prevent the forwarding table size of data center switches from growing by folds as the number of data centers to be interconnected increases. Furthermore, if any kind of L2VPN or L3VPN technologies is used for interconnecting data centers, the scale of forwarding tables on PE routers SHOULD be taken into consideration as well.

3) ARP/ND Cache Table Scalability on Default Gateways

[RFC6820] notes that the Address Resolution Protocol (ARP)/Neighbor Discovery (ND) cache tables maintained by data center default gateways in cloud data centers can raise both scalability and security issues. Therefore, an ideal data center interconnect solution SHOULD prevent the ARP/ND cache table size from growing by multiples as the number of data centers to be connected increases.

4) ARP/ND and Unknown Unicast Flood Suppression or Avoidance

It's well-known that the flooding of Address Resolution Protocol (ARP)/Neighbor Discovery (ND) broadcast/multicast and unknown unicast traffic within a large Layer2 network are likely to affect performances of networks and hosts. As multiple data centers each containing millions of VMs are interconnected together across the Wide Area Network (WAN) at layer2, the impact of flooding as mentioned above will become even worse. As such, it becomes increasingly desirable for data center operators to suppress or even avoid the flooding of ARP/ND broadcast/multicast and unknown unicast traffic across data centers.

5) 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 from a cloud user (i.e., a VPN user) which is destined for a given server located at one data center location of such extended subnet may arrive at another data center location firstly according to the subnet route, and then be forwarded to the location where the service is actually located. This suboptimal routing would obviously result in the unnecessary consumption of the bandwidth resources which are intended for data center interconnection. Furthermore, in the case where the traditional VPLS technology [RFC4761, RFC4762] is used for data center interconnect and default gateways of different data center locations are configured within the same virtual router redundancy group, the returning traffic from that server to the cloud user may be forwarded at layer2 to a default gateway located at one of the remote data center premises, rather than the one placed at the local data center location. This suboptimal routing would also unnecessarily consume the bandwidth resources which are intended for data center interconnect.

This document describes a L3VPN-based subnet extension solution referred to as Virtual Subnet (VS), which can meet all of the

requirements of cloud data center interconnect as described above. Since VS mainly reuses existing technologies including BGP/MPLS IP VPN [RFC4364] and ARP/ND proxy [RFC925][RFC1027][RFC4389], it allows those service providers offering IaaS public cloud services to interconnect their geographically dispersed data centers in a much scalable way, and more importantly, data center interconnection design can rely upon their existing MPLS/BGP IP VPN infrastructures and their experiences in the delivery and the operation of MPLS/BGP IP VPN services.

Although Virtual Subnet is described as a data center interconnection solution in this document, there is no reason to assume that this technology couldn't be used within data centers.

2. Terminology

This memo makes use of the terms defined in [RFC4364], [RFC2338] [MVPN] and [VA-AUTO].

3. Solution Description

3.1. Unicast

3.1.1. Intra-subnet Unicast

```
                                  +--------------------+
            +-----------------+   |                    |   +-----------------+
            |VPN_A:1.1.1.1/24 |   |                    |   |VPN_A:1.1.1.1/24 |
            |              \  |   |                    |   |  /              |
            |    +------+   \++---+-+                +-+---++/   +------+    |
            |    |Host A+----+ PE-1 |                | PE-2 +----+Host B|    |
            |    +------+\   ++-+-+-+                +-+-+-++   /+------+    |
            |     1.1.1.2/24  | | |                    | | |  1.1.1.3/24     |
            |                 | | |                    | | |                 |
            |     DC West     | | |  IP/MPLS Backbone  | | |     DC East     |
            +-----------------+ | |                    | | +-----------------+
                                | +--------------------+ |
                                |                        |
        VRF_A :                 V                VRF_A : V
         +------------+---------+--------+        +------------+---------
+--------+
         |   Prefix   | Nexthop |Protocol|        |   Prefix   | Nexthop |
Protocol|
         +------------+---------+--------+        +------------+---------
+--------+
         | 1.1.1.1/32 |127.0.0.1| Direct |        | 1.1.1.1/32 |127.0.0.1|
Direct |
         +------------+---------+--------+        +------------+---------
+--------+
         | 1.1.1.2/32 | 1.1.1.2 | Direct |        | 1.1.1.2/32 |   PE-1  |
IBGP  |
         +------------+---------+--------+        +------------+---------
+--------+
         | 1.1.1.3/32 |   PE-2  |  IBGP  |        | 1.1.1.3/32 | 1.1.1.3 |
Direct |
         +------------+---------+--------+        +------------+---------
+--------+
         | 1.1.1.0/24 | 1.1.1.1 | Direct |        | 1.1.1.0/24 | 1.1.1.1 |
Direct |
         +------------+---------+--------+        +------------+---------
+--------+
```

Figure 1: Intra-subnet Unicast Example

As shown in Figure 1, two CE hosts (i.e., Hosts A and B) belonging to the same subnet (i.e., 1.1.1.0/24) 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 local CE hosts respectively and then advertise them via L3VPN signaling. Meanwhile, 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

```
                                  +--------------------+
            +-----------------+   |                    |   +-----------------+
            |VPN_A:1.1.1.1/24 |   |                    |   |VPN_A:1.1.1.1/24 |
            |              \  |   |                    |   |  /              |
            |  +------+     \++---+-+                +-+---++/     +------+  |
            |  |Host A+------+ PE-1 |                | PE-2 +-+----+Host B|  |
            |  +------+\     ++-+-+-+                +-+-+-++ |   /+------+  |
            |   1.1.1.2/24    | | |                    | | |  | 1.1.1.3/24   |
            |   GW=1.1.1.4    | | |                    | | |  | GW=1.1.1.4   |
            |                 | | |                    | | |  |    +------+  |
            |                 | | |                    | | |  +----+  GW  +--|
            |                 | | |                    | | |      /+------+  |
            |                 | | |                    | | |    1.1.1.4/24   |
            |                 | | |                    | | |                 |
            |     DC West     | | |  IP/MPLS Backbone  | | |      DC East    |
            +-----------------+ | |                    | | +-----------------+
                                | +--------------------+ |
                                |                        |
        VRF_A :                 V                VRF_A : V
         +------------+---------+--------+        +------------+---------
+--------+
         |   Prefix   | Nexthop |Protocol|        |   Prefix   | Nexthop |
Protocol|
         +------------+---------+--------+        +------------+---------
+--------+
         | 1.1.1.1/32 |127.0.0.1| Direct |        | 1.1.1.1/32 |127.0.0.1|
Direct |
         +------------+---------+--------+        +------------+---------
+--------+
         | 1.1.1.2/32 | 1.1.1.2 | Direct |        | 1.1.1.2/32 |  PE-1   |
IBGP  |
         +------------+---------+--------+        +------------+---------
```

```
+--------+
         | 1.1.1.3/32 |   PE-2  |  IBGP  |        | 1.1.1.3/32 | 1.1.1.3 |
Direct |
         +------------+---------+--------+        +------------+---------
+--------+
         | 1.1.1.4/32 |   PE-2  |  IBGP  |        | 1.1.1.4/32 | 1.1.1.4 |
Direct |
         +------------+---------+--------+        +------------+---------
+--------+
         | 1.1.1.0/24 | 1.1.1.1 | Direct |        | 1.1.1.0/24 | 1.1.1.1 |
Direct |
         +------------+---------+--------+        +------------+---------
+--------+
         | 0.0.0.0/0  |   PE-2  |  IBGP  |        | 0.0.0.0/0  | 1.1.1.4 |
Static |
         +------------+---------+--------+        +------------+---------
+--------+
```

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., 1.1.1.4) 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.

```
                                  +--------------------+
            +-----------------+   |                    |   +-----------------+
            |VPN_A:1.1.1.1/24 |   |                    |   |VPN_A:1.1.1.1/24 |
            |              \  |   |                    |   |  /              |
            |  +------+     \++---+-+                +-+---++/     +------+  |
            |  |Host A+----+-+ PE-1 |                | PE-2 +-+----+Host B|  |
            |  +------+\   | ++-+-+-+                +-+-+-++ |   /+------+  |
            |   1.1.1.2/24 |  | | |                    | | |  | 1.1.1.3/24   |
            |   GW=1.1.1.4 |  | | |                    | | |  | GW=1.1.1.4   |
            |  +------+    |  | | |                    | | |  |    +------+  |
            |--+ GW-1 +----+  | | |                    | | |  +----+ GW-2 +--|
            |  +------+\      | | |                    | | |      /+------+  |
            |   1.1.1.4/24    | | |                    | | |   1.1.1.4/24    |
            |                 | | |                    | | |                 |
            |     DC West     | | |  IP/MPLS Backbone  | | |      DC East    |
            +-----------------+ | |                    | | +-----------------+
                                | +--------------------+ |
                                |                        |
        VRF_A :                 V                VRF_A : V
         +------------+---------+--------+        +------------+---------
+--------+
         |   Prefix   | Nexthop |Protocol|        |   Prefix   | Nexthop |
Protocol|
         +------------+---------+--------+        +------------+---------
+--------+
         | 1.1.1.1/32 |127.0.0.1| Direct |        | 1.1.1.1/32 |127.0.0.1|
Direct |
         +------------+---------+--------+        +------------+---------
+--------+
         | 1.1.1.2/32 | 1.1.1.2 | Direct |        | 1.1.1.2/32 |  PE-1   |
IBGP  |
         +------------+---------+--------+        +------------+---------
+--------+
         | 1.1.1.3/32 |   PE-2  |  IBGP  |        | 1.1.1.3/32 | 1.1.1.3 |
Direct |
         +------------+---------+--------+        +------------+---------
+--------+
         | 1.1.1.4/32 | 1.1.1.4 | Direct |        | 1.1.1.4/32 | 1.1.1.4 |
Direct |
         +------------+---------+--------+        +------------+---------
+--------+
         | 1.1.1.0/24 | 1.1.1.1 | Direct |        | 1.1.1.0/24 | 1.1.1.1 |
Direct |
         +------------+---------+--------+        +------------+---------
+--------+
         | 0.0.0.0/0  | 1.1.1.4 | Static |        | 0.0.0.0/0  | 1.1.1.4 |
Static |
         +------------+---------+--------+        +------------+---------
+--------+
```

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, CE 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.

```
                                         +------+
                                  +------+ PE-3 +------+
            +-----------------+   |      +------+      |   +-----------------+
            |VPN_A:1.1.1.1/24 |   |                    |   |VPN_A:1.1.1.1/24 |
            |              \  |   |                    |   |  /              |
            |  +------+     \++---+-+                +-+---++/     +------+  |
            |  |Host A+------+ PE-1 |                | PE-2 +------+Host B|  |
            |  +------+\     ++-+-+-+                +-+-+-++     /+------+  |
            |   1.1.1.2/24    | | |                    | | |    1.1.1.3/24   |
            |   GW=1.1.1.1    | | |                    | | |    GW=1.1.1.1   |
            |                 | | |                    | | |                 |
            |     DC West     | | |  IP/MPLS Backbone  | | |      DC East    |
            +-----------------+ | |                    | | +-----------------+
                                | +--------------------+ |
                                |                        |
        VRF_A :                 V                VRF_A : V
         +------------+---------+--------+        +------------+---------
+--------+
         |   Prefix   | Nexthop |Protocol|        |   Prefix   | Nexthop |
Protocol|
         +------------+---------+--------+        +------------+---------
+--------+
         | 1.1.1.1/32 |127.0.0.1| Direct |        | 1.1.1.1/32 |127.0.0.1|
Direct |
         +------------+---------+--------+        +------------+---------
+--------+
         | 1.1.1.2/32 | 1.1.1.2 | Direct |        | 1.1.1.2/32 |  PE-1   |
IBGP  |
         +------------+---------+--------+        +------------+---------
+--------+
         | 1.1.1.3/32 |   PE-2  |  IBGP  |        | 1.1.1.3/32 | 1.1.1.3 |
Direct |
         +------------+---------+--------+        +------------+---------
+--------+
         | 1.1.1.0/24 | 1.1.1.1 | Direct |        | 1.1.1.0/24 | 1.1.1.1 |
Direct |
         +------------+---------+--------+        +------------+---------
+--------+
         | 0.0.0.0/0  |   PE-3  |  IBGP  |        | 0.0.0.0/0  |   PE-3  |
IBGP  |
         +------------+---------+--------+        +------------+---------
+--------+
```

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 CE hosts as long as these PE routers have routes for outside networks.

3.2. Multicast

To support IP multicast between CE hosts of the same virtual subnet, MVPN technology [MVPN] could be directly reused. 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 CE hosts, forward them towards remote PE routers through the corresponding default provider multicast distribution tree.

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

3.3. CE Host Discovery

PE routers SHOULD be able to discover their local CE 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 CE host discovery by some traditional host discovery mechanisms using ARP or ND protocols. Furthermore, Link Layer Discovery Protocol (LLDP) described in [802.1AB] or VSI Discovery and Configuration Protocol (VDP) described in [802.1Qbg], or even interaction with the data center orchestration system could also be considered as a means to dynamically discover local CE hosts.

3.4. ARP/ND Proxy

Acting as ARP or ND proxies, PE routers SHOULD only respond to an ARP request or Neighbor Solicitation (NS) message for the target host when there is a corresponding host route in the associated VRF and the outgoing interface of that route is different from the one over which the ARP request or the NS message arrived.

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. CE 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 CE host upon receiving a notification message of VM attachment while 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. Meanwhile, the latter PE router could optionally broadcast a gratuitous ARP/ND message on behalf of that CE host with source MAC address being one of its own. In the way, the ARP/ND entry of that moved CE host which has been cached on any local CE host would be updated accordingly.

3.6. Forwarding Table Scalability

3.6.1. MAC Table Reduction on Data Center Switches

In a VS 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.

3.6.2. PE Router FIB Reduction

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

Figure 5: FIB Reduction Example

To reduce the FIB size of PE routers, Virtual Aggregation (VA) [VA-

AUTO] technology can be used. Take the VPN instance A shown in Figure 5 as an example, the procedures of FIB reduction are as follows:

1) Multiple more specific prefixes (e.g., 1.1.1.0/25 and 1.1.1.128/25) corresponding to the prefix of virtual subnet (i.e., 1.1.1.0/24) are configured as Virtual Prefixes (VPs) and a Route-Reflector (RR) is configured as an Aggregation Point Router (APR) for these VPs. PE routers as RR clients advertise host routes for their own local CE hosts to the RR which in turn, as an APR, installs those host routes into its FIB and then attach the "can-suppress" tag to those host routes before reflecting them to its clients.

2) Those host routes which have been attached with the "can suppress" tag would not be installed into FIBs by clients who are VA-aware since they are not APRs for those host routes. In addition, the RR as an APR would advertise the corresponding VP routes to all of its

clients, and those of which who are VA-aware in turn would install these VP routes into their FIBs.

3) Upon receiving a packet from a local CE host, if no matching host route found, the ingress PE router will forward the packet to the RR according to one of the VP routes learnt from the RR, which in turn forwards the packet to the relevant egress PE router according to the host route learnt from that egress PE router. In a word, the FIB table size of PE routers can be greatly reduced at the cost of path stretch. Note that in the case where the RR is not available for transferring L3VPN traffic between PE routers for some reason (e.g., the RR is implemented on a server, rather than a router), the APR function could actually be performed by a given PE router other than the RR as long as that PE router has installed all host routes belonging to the virtual subnet into its FIB. Thus, the RR only needs to attach a "can-suppress" tag to the host routes learnt from its clients before reflecting them to the other clients. Furthermore, PE routers themselves could directly attach the "can-suppress" tag to those host routes for their local CE hosts before distributing them to remote peers as well.

4) Provided a given local CE host sends an ARP request for a remote CE host, the PE router that receives such request will install the host route for that remote CE host into its FIB, in case there is a host route for that CE host in its RIB and has not yet been installed into the FIB. Therefore, the subsequent packets destined for that remote CE host will be forwarded directly to the egress PE router. To save the FIB space, FIB entries corresponding to remote host routes which have been attached with "can-suppress" tags would expire if they have not been used for forwarding packets for a certain period of time.

3.6.3. PE Router RIB Reduction

```
                                         +------+
                                  +------+  RR  +------+
            +-----------------+   |      +------+      |   +-----------------+
            |VPN_A:1.1.1.1/24 |   |                    |   |VPN_A:1.1.1.1/24 |
            |              \  |   |                    |   |  /              |
            |  +------+     \++---+-+                +-+---++/     +------+  |
            |  |Host A+------+ PE-1 |                | PE-2 +------+Host B|  |
            |  +------+\     ++-+-+-+                +-+-+-++     /+------+  |
            |   1.1.1.2/24    | | |                    | | |    1.1.1.3/24   |
            |                 | | |                    | | |                 |
            |     DC West     | | |  IP/MPLS Backbone  | | |      DC East    |
            +-----------------+ | |                    | | +-----------------+
                                | +--------------------+ |
                                |                        |
```

```
       VRF_A :                V                VRF_A : V
         +------------+---------+--------+        +------------+---------
+--------+
         |   Prefix   | Nexthop |Protocol|        |   Prefix   | Nexthop |
Protocol|
         +------------+---------+--------+        +------------+---------
+--------+
         | 1.1.1.1/32 |127.0.0.1| Direct |        | 1.1.1.1/32 |127.0.0.1|
Direct |
         +------------+---------+--------+        +------------+---------
+--------+
         | 1.1.1.2/32 | 1.1.1.2 | Direct |        | 1.1.1.3/32 | 1.1.1.3 |
Direct |
         +------------+---------+--------+        +------------+---------
+--------+
         | 1.1.1.0/25 |    RR   |  IBGP  |        | 1.1.1.0/25 |    RR   |
IBGP  |
         +------------+---------+--------+        +------------+---------
+--------+
         |1.1.1.128/25|    RR   |  IBGP  |        |1.1.1.128/25|    RR   |
IBGP  |
         +------------+---------+--------+        +------------+---------
+--------+
         | 1.1.1.0/24 | 1.1.1.1 | Direct |        | 1.1.1.0/24 | 1.1.1.1 |
Direct |
         +------------+---------+--------+        +------------+---------
+--------+
```

Figure 6: RIB Reduction Example

To reduce the RIB size of PE routers, BGP Outbound Route Filtering (ORF) mechanism is used to realize on-demand route announcement. Take the VPN instance A shown in Figure 6 as an example, the procedures of RIB reduction are as follows:

1) PE routers as RR clients advertise host routes for their local CE hosts to a RR which however doesn't reflect these host routes by default unless it receives explicit ORF requests for them from its clients. The RR is configured with routes for more specific subnets (e.g., 1.1.1.0/25 and 1.1.1.128/25) corresponding to the virtual subnet (i.e., 1.1.1.0/24) with next-hop being pointed to Null0 and then advertises these routes to its clients via BGP.

2) Upon receiving a packet from a local CE host, if no matching host route found, the ingress PE router will forward the packet to the RR according to one of the subnet routes learnt from the RR, which in turn forwards the packet to the relevant egress PE router according to the host route learnt from that egress PE router. In a word, the RIB table size of PE routers can be greatly reduced at the cost of path stretch.

3) Just as the approach mentioned in section 3.6.2, in the case where the RR is not available for transferring L3VPN traffic between PE routers for some reason, a PE router other than the RR could advertise the more specific subnet routes as long as that PE router has installed all host routes belonging to that virtual subnet into its FIB.

4) Provided a given local CE host sends an ARP request for a remote CE host, the ingress PE router that receives such request will request the corresponding host route from its RR by using the ORF mechanism (e.g., a group ORF containing Route-Target (RT) and prefix information) in case there is no host route for that CE host in its RIB yet. Once the host route for the remote CE host is

learnt from the RR, the subsequent packets destined for that CE host would be forwarded directly to the egress PE router. Note that the RIB entries of remote host routes could expire if they have not been used for forwarding packets for a certain period of time. Once the expiration time for a given RIB entry is approaching, the PE router would notify its RR not to pass the updates for corresponding host route by using the ORF mechanism.

3.7. ARP/ND Cache Table Scalability on Default Gateways

In case where data center default gateway functions are implemented on PE routers of the VS as shown in Figure 4, since the ARP/ND cache table on each PE router only needs to contain ARP/ND entries of local CE hosts, the ARP/ND cache table size will not grow as the number of data centers to be connected increases.

3.8. ARP/ND and Unknown Uncast Flood Avoidance

In VS, the flooding domain associated with a given virtual subnet that has been extended across multiple data centers, has been partitioned into segments and each segment is confined within a single data center. Therefore, the performance impact on networks and servers caused 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 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 data center default gateways, propagate host routes for their local CE hosts respectively to remote PE routers which are attached to cloud user sites (i.e., PE-3).

As such, 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 traffic is now forwarded according to the host route for that server, rather than the subnet route.

Furthermore, for traffic coming from cloud data centers and forwarded to cloud user sites, each PE router acting as a default gateway would forward the traffic received from its local CE hosts according to the best-match route in the corresponding VRF. As a result, traffic from data centers to cloud user sites is forwarded along the optimal path as well.

4. Considerations for 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). To support those few non-IP traffic (if present) in the Virtual Subnet solution, the approach following the idea of "route all IP traffic, bridge non-IP traffic" could be considered as an enhancement to the original Virtual Subnet solution.

Note that more and more cluster vendors are offering clustering applications based on Layer 3 interconnection.

5. Security Considerations

This document doesn't introduce additional security risk to BGP/MPLS L3VPN, nor does it provide any additional security feature for BGP/MPLS L3VPN.

6. IANA Considerations

There is no requirement for any IANA action.

7. Acknowledgements

Thanks to Dino Farinacci, Himanshu Shah, Nabil Bitar, Giles Heron, Ronald Bonica, Monique Morrow, Rajiv Asati and Eric Osborne for their valuable comments and suggestions on this document.

8. References

8.1. Normative References

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

8.2. 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.

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

[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.

[RFC4389] D. Thaler, M. Talwar, and C. Patel, "Neighbor Discovery Proxies (ND Proxy) ", RFC 4389, April 2006.

[RFC5798] S. Nadas., "Virtual Router Redundancy Protocol", RFC 5798, March 2010.

[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.

[802.1AB] IEEE Standard 802.1AB-2009, "Station and Media Access Control Connectivity Discovery", September 17, 2009.

[802.1Qbg] IEEE Draft Standard P802.1Qbg/D2.0, "Virtual Bridged Local Area Networks -Amendment XX: Edge Virtual Bridging", Work in Progress, December 1, 2011.

[RFC6820] Narten, T., Karir, M., and I. Foo, "Problem Statement for ARMD", RFC 6820, January 2013.

Authors' Addresses

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

Susan Hares
Email: shares@ndzh.com

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

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