Network Working Group
Internet Draft
Category: Standards Track
Expiration Date: December 2014

R. Aggarwal
Arktan Inc

Y. Rekhter
Juniper Networks

W. Henderickx
Alcatel-Lucent

R. Shekhar
Juniper Networks

Luyuan Fang
Cisco Systems

Ali Sajassi
Cisco Systems

June 2 2014

Data Center Mobility based on E-VPN, BGP/MPLS IP VPN, IP Routing and NHRP

draft-raggarwa-data-center-mobility-07.txt

Status of this Memo

Abstract

This document describes a set of network-based solutions for seamless Virtual Machine mobility in the data center. These solutions provide a toolkit which is based on IP routing, E-VPNs, BGP/MPLS IP VPNs, and NHRP.

Table of Contents

1. Specification of requirements

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 [RFC2119].

2. Introduction

This document describes network-based solutions for seamless Virtual Machine (VM) mobility, where seamless mobility is defined as the ability to move a VM from one server in the data center to another server in the same or different data center, while retaining the IP and MAC address of the VM. In the context of this document the term mobility, or a reference to moving a VM should be considered to imply seamless mobility, unless otherwise stated.

The solutions described in this document provide a network-based toolkit which is based on IP routing, E-VPN [E-VPN], BGP/MPLS IP VPNs [RFC4364], and NHRP [RFC2332].

Note that in the scenario where a VM is moved between servers located in different data centers, there are certain constraints to how far such data centers may be located geographically. This distance is limited by the current state of the art of the Virtual Machine technology, by the bandwidth that may be available between the data centers, the ability to manage and operate such VM mobility etc. This document describes a set of solutions for VM mobility. These solutions form a toolkit that enables VMs to move across both small and large geographical distances. However, the practical applicability of these solutions will depend on these constraints. If these constraints are relaxed over time, allowing VMs to move across larger geographical boundaries, the solutions described here will continue to be applicable.

2.1. Terminology

In this document the term "Top of Rack Switch (ToR)" is used to refer to a switch in a data center that is connected to the servers that host VMs. A data center may have multiple ToRs.

Several data centers could be connected by a network. In addition to providing interconnect among the data centers, such a network could provide connectivity between the VMs hosted in these data centers and the sites that contain hosts communicating with such VMs. Each data center has one or more Data Center Border Router (DCBR) that connects the data center to the network, and provides (a) connectivity between

VMs hosted in the data center and VMs in other data centers, and (b) connectivity between VMs hosted in the data center and hosts communicating with these VMs.

The data centers and the network that interconnects them may be either (a) under the same administrative control, or (b) controlled by different administrations.

Consider a set of VMs that (as a matter of policy) are allowed to communicate with each other, and a collection of devices that interconnect these VMs. If communication among any VMs in that set could be accomplished in such a way as to preserve MAC source and destination addresses in the Ethernet header of the packets exchanged among these VMs (as these packets traverse from their sources to their destinations), we will refer to such set of VMs as an Layer 2 based Closed User Group (L2-based CUG).

A given VM may be a member of more than one L2-based CUG.

In terms of IP address assignment this document assumes that all VMs of a given L2-based CUG have their IP addresses assigned out of a single IP prefix. Thus, in the context of this document a single IP subnet corresponds to a single L2-based CUG.

A VM that is a member of a given L2-based CUG may (as a matter of policy) be allowed to communicate with VMs that belong to other L2-based CUGs, or with other hosts. Such communication involves IP forwarding, and thus would result in changing MAC source and destination addresses in the Ethernet header of the packets being exchanged.

In this document the term "L2 site" refers to a collection of interconnected devices that perform forwarding based on the information carried in the Ethernet header. Forwarding within an L2 site could be provided by such layer 2 technologies as Spanning Tree Protocol (STP), etc... Note that any multi-chassis LAG is treated as a single L2 site.

Servers connected to a given L2 site may host VMs that belong to different L2-based CUGs. Enforcing L2-based CUGs boundaries among these VMs within a single L2 site is accomplished by relying on Layer 2 mechanisms (e.g., VLANs).

We say that an L2 site contains a given VM (or that a given VM is in a given L2 site), if the server presently hosting this VM is connected to a ToR that is part of that site.

We say that a given L2-based CUG is present within a given data

center if one or more VMs that are part of that CUG are presently hosted by the servers located in that data center.

This document assumes that VMs that belong to the same L2-based CUG, and are in the same L2 site MUST use the same VLAN-ID. This document assumes that VMs that belong to the same L2-based CUG, and are in different L2 sites MAY use either the same or different VLAN-IDs.

This document assumes that VMs that belong to different L2-based CUGs, and are in the same L2 site MUST use different VLAN-IDs. This document assumes that VMs that belong to different L2-based CUGs, and are in different L2 sites MAY use either the same, or different VLAN-IDs.

3. Problem Statement

This section describes the specific problems that need to be addressed to enable seamless VM mobility.

3.1. Maintaining Connectivity in the Presence of VM Mobility

In the context of this document the ability to maintain connectivity in the presence of VM mobility means the ability to exchange traffic between a VM and its peer(s), as the VM moves from one server to another, where the peer(s) may be either other VM(s) or hosts.

3.2. Layer 2 Extension

Consider a scenario where a VM that is a member of a given L2-based CUG moves from one server to another, and these two servers are in different L2 sites, where these sites may be located in the same or different data centers. In order to enable communication between this VM and other VMs of that L2-based CUG, the new L2 site must become interconnected with the other L2 site(s) that presently contain the rest of the VMs of that CUG, and the interconnect must not violate the L2-based CUG requirement to preserve source and destination MAC addresses in the Ethernet header of the packets exchange between this VM and other members of that CUG.

Moreover, if the previous site no longer contains any VMs of that CUG, the previous site no longer needs to be interconnected with the other L2 site(s) that contain the rest of the VMs of that CUG.

We will refer to this as the "layer 2 extension problem".

Note that the layer 2 extension problem is a special case of maintaining connectivity in the presence of VM mobility, as the former restricts communicating VMs to a single/common L2-based CUG, while the latter does not.

3.3. Optimal IP Routing

In the context of this document optimal IP routing, or just optimal routing, in the presence of VM mobility could be partitioned into two problems:

+ Optimal routing of a VM's outbound traffic. This means that as a given VM moves from one server to another, the VM's default gateway should be in a close topological proximity to the ToR that connects the server presently hosting that VM. Note that when we talk about optimal routing of the VM's outbound traffic, we mean traffic from that VM to the destinations that are outside of the VM's L2-based CUG. This document refers to this problem as the VM default gateway problem.

+ Optimal routing of VM's inbound traffic. This means that as a given VM moves from one server to another, the (inbound) traffic originated outside of the VM's L2-based CUG, and destined to that VM be routed via the router of the VM's L2-based CUG that is in a close topological proximity to the ToR that connects the server presently hosting that VM, without first traversing some other router of that L2-based CUG. This is also known as avoiding "triangular routing". This document refers to this problem as the triangular routing problem.

Note that optimal routing is a special case of maintaining connectivity in the presence of VM mobility, as the former assumes not only the ability to maintain connectivity, but also that this connectivity is maintained using optimal routing. On the other hand, maintaining connectivity does not make optimal routing a pre-requisite.

4. Layer 2 Extension Solution

This document assumes that the solution for the layer 2 extension problem, relies on [E-VPN]. That is, the L2 sites that contain VMs of a given L2-based CUG are interconnected together using E-VPN. Thus a given E-VPN corresponds/associated with one or more L2-based CUGs (e.g., VLANs). An L2-based CUG is associated with a single E-VPN Ethernet Tag Identifier.

This section provides a brief overview of how E-VPN is used as the solution for the "layer 2 extension problem". Details of E-VPN operations can be found in [E-VPN].

A single L2 site could be as large as the whole network within a single data center, in which case the DCBRs of that data center, in addition to acting as IP routers for the L2-based CUGs present in the data center, also act as PEs. In this scenario E-VPN is used to handle VM migration between servers in different data centers.

A single L2 site could be as small as a single ToR with the servers connected to it, in which case the ToR acts as a PE. In this scenario E-VPN is used to handle VM migration between servers that are either in the same or in different data centers. Note that even in this scenario this document assumes that DCBRs, in addition to acting as IP routers for the L2-based CUGs present in their data center, also participate in the E-VPN procedures, acting as BGP Route Reflectors for the E-VPN routes originated by the ToRs acting as PEs.

In the case where E-VPN is used to interconnect L2 sites in different data centers, the network that interconnects DCBRs of these data centers could provide either (a) only Ethernet or IP/MPLS connectivity service among these DCBRs, or (b) may offer the E-VPN service. In the former case DCBRs exchange E-VPN routes among themselves relying only on the Ethernet or IP/MPLS connectivity service provided by the network that interconnects these DCBRs. The network does not directly participate in the exchange of these E-VPN routes. In the latter case the routers at the edge of the network may be either co-located with DCBRs, or may establish E-VPN peering with DCBRs. Either way, in this case the network facilitates exchange of E-VPN routes among DCBRs (as in this case DCBRs would not need to exchange E-VPN routes directly with each other).

Please note that for the purpose of solving the layer 2 extension problem the propagation scope of E-VPN routes for a given L2-based CUG is constrained by the scope of the PEs connected to the L2 sites that presently contain VMs of that CUG. This scope is controlled by the Route Target of the E-VPN routes. Controlling propagation scope could be further facilitated by using Route Target Constrain [RFC4684].

Use of E-VPN ensures that traffic among members of the same L2-based CUG is optimally forwarded, irrespective of whether members of that CUG are within the same or in different data centers. This follows from the observation that E-VPN inherently enables (disaggregated) forwarding at the granularity of the MAC address of the VM.

Optimal forwarding among VMs of a given L2-based CUG that are within

the same data center requires propagating VM MAC addresses, and comes at the cost of disaggregated forwarding within a given data center. However such disaggregated forwarding is not necessary between data centers if a given L2-based CUG spans multiple data centers. For example when a given ToR acts as a PE, this ToR has to maintain MAC advertisement routes only to the VMs within its own data center (and furthermore, only to the VMs that belong to the L2-based CUGs whose site(s) are connected to that ToR), and then point a "default" MAC route to one of the DCBRs of that data center. In this scenario a DCBR of a given data center, when it receives MAC advertisement routes from DCBR(s) in other data centers, does not re-advertise these routes to the PEs within its own data center, but just advertises a single "default" MAC advertisement route to these PEs.

When a given VM moves to a new L2 site, if in the new site this VM is the only VM from its L2-based CUG, then the PE(s) connected to the new site need to be provisioned with the E-VPN Instances (EVI) of the E-VPN associated with this L2-based CUG. Likewise, if after the move the old site no longer has any VMs that are in the same L2-based CUG as the VM that moved, the PE(s) connected to the old site need to be de-provisioned with the EVI of the E-VPN. Procedures to accomplish this are outside the scope of this document.

5. VM Default Gateway Solutions

Once VM moves to a new L2 site, solving the VM Default Gateway problem would require PE(s) connected to that L2 site to apply IP forwarding to the inter-CUG/inter-subnet traffic originated from that VM. That implies that (a) PE(s) should be capable of performing both MAC-based and IP-based forwarding (although IP-based forwarding functionality could be limited to just forwarding either based on IP host routes, or based on the IP default route), and (b) PE(s) should be able to distinguish between intra-CUG/intra-subnet and inter-CUG/inter-subnet traffic originated by that VM (in order to apply MAC-based forwarding to the former and IP-based forwarding to the latter).

As VM moves to a new L2 site, the default gateway IP address of the VM may not change. Further, the ARP cache of the VM may not time out. Thus the destination MAC address in the inter-CUG/inter-subnet traffic originated by that VM would not change as VM moves to the new site. Given that, how would PE(s) connected to the new L2 site be able to recognize inter-CUG/inter-subnet traffic originated by that VM ? The following describes two possible solutions.

Both of the solutions assume that for inter-CUG/inter-subnet traffic between VM and its peers outside of VM's own data center, one or more

DCBRs of that data center act as fully functional default gateways for that traffic.

Both of these solutions also assume that VLAN-aware VLAN bundling mode of E-VPN is used as the default mode such that different L2-CUGs (different subnets) for the same tenant can be accommodated in a single EVI. This facilitates provisioning since E-VPN related provisioning (such as RT configuration) could be done on a per-tenant basis as opposed to on a per-subnet (per L2-CUG) basis. In this default mode, VMs' MAC addresses are maintained on a per bridge domain basis (per subnet) within the EVI; however, VM's IP addresses are maintained across all the subnets of that tenant in that EVI. In the scenarios where communications among VMs of different subnets belonging to the same tenant is to be restricted based on some policies, then the VLAN mode of E-VPN should be used with each VLAN/subnet mapping to its own EVI and E-VPN RT filtering can be leveraged to enforce flexible policy-based communications among VMs of different subnets for that tenant.

5.1. VM Default Gateway Solution - Solution 1

The first solution relies on the use of an anycast default gateway IP address and an anycast default gateway MAC address.

If DCBRs act as PEs for an E-VPN corresponding to a given L2-based CUG, then these anycast addresses are configured on these DCBRs. Likewise, if ToRs act as PEs, then these anycast addresses are configured on these ToRs. All VMs of that L2-based CUG are (auto)configured with the (anycast) IP address of the default gateway.

DCBRs (or ToRs) acting as PEs use these anycast addresses as follows:

+ When a particular DCBR (or ToR) acting as a PE receives a packet with the (anycast) default gateway MAC address, the DCBR (or ToR) applies IP forwarding to the packet.

+ When a particular DCBR (or ToR) acting as a PE receives an ARP Request for the default gateway (anycast) IP address, the DCBR (or ToR) generates ARP Reply.

This ensures that a particular DCBR (or ToR), acting as a PE, can always apply IP forwarding to the packets sent by a VM to the (anycast) default gateway MAC address. It also ensures that such DCBR (or ToR) can respond to the ARP Request generated by a VM for the default gateway (anycast) IP address.

DCBRs (or ToRs) acting as PEs must never use the anycast default gateway MAC address as the source MAC address in the packets originated by these DCBRs (or ToRs).

Note that multiple L2-based CUGs may share the same MAC address for the purpose of using as the (anycast) MAC address of the default gateway for these CUGs.

If the default gatewat functionality is not in TORs, then the default gateway MAC/IP addresses need to be distributed using E-VPN procedures. Note that with this approach when originating E-VPN MAC advertisement routes for the MAC address of the default gateways of a given L2-based CUG, all these routes MUST indicate that this MAC address belongs to the same Ethernet Segment Identifier (ESI).

5.2. VM Default Gateway Solution - Solution 2

The second solution does not require to configure the anycast default gateway IP and MAC address on the PEs.

Each DCBR (or each ToR) that acts as a default gateway for a given L2-based CUG advertises in the E-VPN control plane its default gateway IP and MAC address using the MAC advertisement route, and indicates that such route is associated with the default gateway. The MAC advertisement route MUST be advertised as per procedures in [E-VPN]. The MAC address in such an advertisement MUST be set to the default gateway MAC address of the DCBR (or ToR). The IP address in such an advertisement MUST be set to the default gateway IP address of the DCBR (or ToR). To indicate that such a route is associated with a default gateway, the route MUST carry the Default Gateway extended community [Default-Gateway].

Each PE that receives this route and imports it as per procedures of [E-VPN] MUST create MAC forwarding state that enables it to apply IP forwarding to the packets destined to the MAC address carried in the route. The PE that receives this E-VPN route follows procedures in Section 12 of [E-VPN] when replying to ARP Requests that it receives if such Requests are for the IP address in the received E-VPN route.

6. Triangular Routing Solution

The triangular routing solution could be partitioned into two components: intra data center triangular routing solution, and inter data center triangular routing solution. The former handles the situation where communicating VMs are in the same data center. The latter handles all other cases.

Both of these solutions assume that as a PE originates MAC advertisement routes, such routes, in addition to MAC addresses of the VMs, also carry IP addresses of these VMs. Procedures by which a PE can learn the IP address associated with a given MAC address are specified in [E-VPN].

6.1. Intra Data Center Triangular Routing Solution

Consider a set of L2-based CUGs, such that VMs of these CUGs, as a matter of policy, are allowed to communicate with each other. To avoid triangular routing among such VMs that are in the same data center this document relies on the E-VPN procedures, as follows.

Procedures in this section assume that ToRs act as PEs, and also able to support IP forwarding functionality.

For a given set of L2-based CUGs whose VMs are allowed to communicate with each other, consider a set of E-VPN instances (EVIs) of the E-VPNs associated with these CUGs. We further restrict this set of EVIs to only the EVIs that are within the same data center. To avoid triangular routing among VMs within the same data center, E-VPN routes originated by one of the EVIs within such set should be imported by all other EVIs in that set, irrespective of whether these other EVIs belong to the same E-VPN as the EVI that originates the routes.

One possible way to accomplish this is (a) for each set of L2-based CUGs whose VMs are allowed to communicate with each other, and for each data center that contains such CUGs have a distinct RT (distinct RT per set, per data center), (b) provision each EVI of the E-VPNs associated with these CUGs to import routes that carry this RT, and (c) make the E-VPN routes originated by such EVIs to carry this RT. Note that these RTs are in addition to the RTs used to form individual E-VPNs. Note also, that what is described here is conceptually similar to the notion of "extranets" in BGP/MPLS VPNs [RFC4364].

When a PE imports an E-VPN route into a particular EVI, and this route is associated with a VM that is not part of the L2-based CUG

associated with the E-VPN of that EVI, the PE creates IP forwarding state to forward traffic to the IP address present in the NLRI of the route towards the Next Hop, as specified in the route.

To illustrate how the above procedures avoid triangular routing, consider the following example. Assume that a particular VM, VM-A, is currently hosted by a server connected to a particular ToR, ToR-1, and another VM, VM-B, is currently hosted by a server connected to ToR-2. Assume that VM-A and VM-B belong to different L2-based CUGs, and (as a matter of policy) VMs in these CUGs are allowed to communicate with each other. Now assume that VM-B moves to another server, and this server is connected to ToR-3. Assume that ToR-1, ToR-2, and ToR-3 are in the same data center. While initially ToR-1 would forward data originated by VM-A and destined to VM-B to ToR-2, after VM-B moves to the server connected to ToR-3, using the procedures described above, ToR-1 would forward the data to ToR-3 (and not to ToR-2), thus avoiding triangular routing.

Note that for the purpose of redistributing E-VPN routes among multiple L2-based CUGs, the above procedures limit the propagation scope of routes to individual VMs to a single data center, and furthermore, to only a subset of the PEs within that data center - the PEs that have EVIs of the E-VPNs associated with the L2-based CUGs whose VMs are allowed to communicate with each other. As a result, the control plane overhead needed to avoid triangular routing within a data center is localized to these PEs.

6.2. Inter Data Center Triangular Routing Solution

This section describes procedures to avoid triangular routing between VMs in different data centers, or between a VM located in a given data center and a host located in some other site that are based on propagating host routes.

There are two inter data center triangular routing solutions proposed in this document.

The first solution is based on propagating host routes to VMs IP addresses, with careful consideration given to constraining the propagation scope of these routes in order to be able to limit the scope of the devices that need to carry additional control plane load. In this solution a DCBR of a given data center originates host routes for the VMs that are hosted by the servers in that data center. Such routes could be either IP host routes or VPN-IP host routes.

The second solution relies on using Next Hop Resolution Protocol

(NHRP). In this solution NHRP is used to provide (on demand) mapping from a given VM's IP address into an IP address of the DCBR of the data center that contains the server presently hosting this VM.

6.2.1. Propagating IP host routes

The approach described in this section assumes that all the communicating VMs belong to the same routing/addressing realm.

Note that while the material in this section is presented in terms of avoiding triangular routing between VMs that are in different data centers, procedures described in this section are equally applicable to communication between a VM and a host.

Procedures in this section assumes that DCBRs, in addition to acting as routers for the L2-based CUGs present in their data center, also participate in the E-VPN procedures either (a) acting as PEs, or (b) acting as BGP Route Reflectors for the E-VPN routes originated by the ToRs within their data center if these ToRs are acting as PEs. As a result, a DCBR that acts as a router for a given L2-based CUG can determine whether a particular VM that is a member of this CUG is in the same data center as the DCBR itself.

Procedures in this section rely on DCBRs performing what amounts to a redistribution of routes between E-VPN and OSPF/ISIS/BGP. In other words, DCBRs in one data center use the E-VPN functionality to obtain the information about IP addresses of the VMs currently being present in their data center, and then advertise into OSPF/ISIS/BGP host routes to these IP addresses.

DCBRs in other data centers receive these route, and use the information carried in these routes to avoid triangular routing. Note that even if ToRs within a given data center act as both PEs and also perform IP-based forwarding, DCBRs of that data center SHOULD NOT redistribute to these ToRs the host routes they receive from DCBRs in other data centers - DCBRs SHOULD advertise only the IP default route to these ToRs.

To illustrate how the above procedures avoid triangular routing consider the following example. Assume that a particular VM, VM-A, is currently being hosted by a server located in data center DC-1 with DCBR-1 as its DCBR, and another VM, VM-B, is currently being hosted by a server located in data center DC-2 with DCBR-2 as its DCBR. Assume that VM-A and VM-B belong to different L2-based CUGs. Using the E-VPN procedures DCBR-2 determines that VM-B is presently in its own data center, and thus originates an IP host route to VM-B's IP address. Using OSPF/ISIS/BGP this route ultimately gets propagated to

DCBR-1. Using this information DCBR-1 would forward data originated by VM-A and destined to VM-B to DCBR-2.

Now assume that VM-B moves to another server, and this server is located in data center DC-3 with DCBR-3 as its DCBR. Using the E-VPN procedures, DCBR-2 determines that VM-B is no longer present in DCBR-2's data center, and thus withdraws the previously originated IP host route to VM-B's IP address. Using the E-VPN procedures, DCBR-3 now determines that VM-B is now present in DCBR-3's data center, and thus originates an IP host route to VM-B's IP address. Using the OSPF/ISIS/BGP procedures, this route ultimately gets propagated to DCBR-1. Using this information DCBR-1 would now forward data originated by VM-A and destined to VM-B to DCBR-3, thus avoiding triangular routing.

As we mentioned above, essential to the scheme that relies on propagating (host) routes to individual VM's IP addresses is the ability to constrain the propagation scope of these routes. The following describes possible approaches to accomplish this.

6.2.1.1. Constraining propagation scope with OSPF/ISIS

When DCBRs use OSPF or ISIS to exchange routing information among themselves, OSPF/ISIS areas may be used as a boundary to constrain propagation scope of host routes. That is, a host route originated by a given DCBR is propagated only within the OSPF/ISIS area of that DCBR, and thus received only by the DCBRs that are in the same OSPF/ISIS area. ABRs connected to a particular OSPF/ISIS area advertise outside of this area only routes to the IP subnets associated with the L2-based CUGs present in the data centers whose DCBRs are in that area, but do not advertise any host routes.

Note that this approach avoids triangular routing when VM is moved between servers that are located in the data centers whose DCBRs belong to the same OSPF/ISIS area, but does not avoid triangular routing if these DCBRs belong to different OSPF/ISIS areas. However, when these DCBRs belong to the same OSPF/ISIS area this approach avoid triangular routing irrespective of whether the peer is in the same or different OSPF/ISIS area as the VM itself.

Since this approach avoids triangular routing avoidance only within a limited scope, to provide connectivity to the peers that are outside of that scope, DCBRs connected to a given L2-based CUG, in addition to advertising host routes, also advertise into OSPF/ISIS a route associated with the IP subnet of that CUG. Propagation of such route need not be limited to the OSPF/ISIS area(s) of these DCBRs.

6.2.1.2. Constraining propagation scope with BGP

When DCBRs use BGP to exchange routing information among themselves, one could use Route Targets (RTs) to constrain the propagation scope of host routes to a particular set of data centers, or to be more precise to a particular set of DCBRs of these data centers.

To accomplish that, DCBRs in a particular set of data centers may be configured with a particular import RT. DCBRs that originate host routes and wish to constrain the propagation scope of these routes to a particular set of data centers would advertise these routes with the import RT provisioned for the DCBRs of the data centers in that set. Route Target Constrain [RFC4684] MAY be used to facilitate constrained distribution of these host routes.

Note that this approach avoids triangular routing only if both communicating VMs are in the data centers whose DCBRs are provisioned with the same import RT, and moreover, when VM moves between servers that are located in the data centers whose DCBRs are configured with the same import RT.

Note that at least in principle RIPv2 by carefully using routing policies and tags in the routes can achieve similar results.

Since this approach avoids triangular routing avoidance only within a limited scope, to provide connectivity to the peers that are outside of that scope, DCBRs connected to a given L2-based CUG, in addition to advertising host routes, also advertise into BGP a route to the IP subnet associated with that CUG.

6.2.1.3. Policy based origination of VM Host IP Address Routes

When a DCBR (using E-VPN procedures) learns that a particular VM is now moved to the DCBR's data center, the DCBR may not originate a corresponding VM host route by default. Instead, it may optionally do so based on a dynamic policy. For example, the policy may be to originate such a route only when the traffic to the VM flowing through that DCBR exceeds a certain threshold. Note that delaying origination of the host route, while impacting routing optimality, does not impact the ability to maintain connectivity between this VM and its peers.

6.2.1.4. Policy based instantiation of VM Host IP Address Forwarding State

When a ToR/DCBR learns (from another ToR or DCBR) a host route of a VM, it may not immediately install this route in its forwarding table. Instead, it may optionally do so based on a dynamic policy. For example, the policy may be to install such forwarding state only when the ToR/DCBR needs to forward the first packet to that particular VM. Note that delaying installation of the host route, while impacting routing optimality, does not impact the ability to maintain connectivity between this VM and its peers.

6.2.2. Propagating VPN-IP host routes

In the scenario where one wants to restrict communication between VMs in different L2-based CUGs to a particular set of L2-based CUGs, and/or when one need to support multiple routing/addressing realms (e.g., IP VPNs) this document proposes to use mechanisms of BGP/MPLS VPN [RFC4364] as follows.

The set of L2-based CUGs whose VMs are allowed to communicate with each other is considered as a single Layer 3 VPN.

A DCBR, in addition to implementing the E-VPN functionality, also implements functionality of a Provider Edge (PE) router, as specified in [RFC4364]. Specifically, this PE router would maintain multiple VRFs, oner per each Layer 3 VPN whose L2-based CUGs are present in the DCBR's data center. Such VRF would be populated from two sources: (1) VPN-IP routes received from other VRFs that belong to the same Layer 3 VPN, and (2) MAC advertisement routes received from the EVIs that are in the same data center as the DCBR hosting the VRF, and that belong to the E-VPNs associated with the L2-based CUGs that form the Layer 3 VPN associated with the VRF.

Procedures of [RFC4364] constrain the propagation scope of the VPN-IP host routes originated from a given VRF on a given DCBR to only the other VRFs who belong to the same VPN as the VRF that originated the routes (or even to a subset of such VRFs).

Using the extranet procedures such VPN-IP host routes could be propagated to other VPNs. Alternatively, one or more VRFs of a given Layer 3 VPN, in addition to originating the VPN-IP host routes, MAY also originate a VPN-IP route to the IP subnet associated with the L2-based CUG that belongs to the Layer 3 VPN associated with that VRF. Such route could be distributed to other Layer 3 VPNs using the extranet procedures.

Note that applicability of the approach described in this section is not limited to the environment where one need to support multiple routing/addressing realms (e.g., IP VPN environment) - this approach is also well suitable for the environment that consists of a single routing/addressing realm.

6.2.3. Triangular Routing Solution Based on NHRP

Triangular routing solution based on NHRP utilizes a subset of the functionality provided by the Next Hop Resolution Protocol [RFC2332] as follows.

Note that while most of the material in this section is presented in terms of avoiding triangular routing between a VM located in a given data center and a host located in some other site, procedures described in this section are equally applicable to communication between VMs in different data centers.

Consider a scenario where a host within a given site communicates with a VM, and the VM could move among servers located in different data centers. The following example illustrates how NHRP allows to avoid triangular routing.

Assume that a given L2-based CUG spans two data centers, one in San Francisco (SF) and another in Los Angelos (LA). DCBR-SF is the DCBR for the SF data center. DCBR-LA is the DCBR for the LA data center. Since this CUG spans both the SF data center and the LA data center, at least one of DCBR-SF or DCBR-LA advertises a route to the IP prefix of the IP subnet associated with the CUG (this is a route to a prefix, and not a host route). Let's denote this IP prefix as X. Advertising a route to this prefix is essential to avoid transient disruptions in maintaining connectivity in the presence of VM mobility.

DCBR-LA and DCBR-SF can determine whether a particular VM of that L2-based CUG is in LA or SF by using the E-VPN procedures.

There is a site in Denver, and that site contains a host B that wants to communicate with a particular VM, VM-A, that belong to that L2-based CUG.

Assume that there is an IP infrastructure that connects the border router of the site in Denver, DCBR-SF, and DCBR-LA. This infrastructure could be provided by either 2547 VPNs, or IPSec tunnels over the Internet, or by L2 circuits. [Note that this infrastructure does not assume that the border router in Denver is 1 IP hop away from either DCBR-SF or DCBR-LA].

To avoid triangular routing, if VM-A is in LA, then the border route in Denver should send traffic for VM-A via DCBR-LA without going first through DCBR-SF. If VM-A is in SF, then the border route in Denver should send traffic for VM-A via DCBR-SF without going first through DCBR-LA. This should be true except for some transients during the move of VM-A between SF and LA.

To accomplish this we would require the border router in Denver, DCBR-SF, and DCBR-LA to support a subset of the NHRP functionality, as follows. In NHRP terminology DCBR-SF and DCBR-LA are NHRP Servers (NHSs), while the border router in Denver is an NHRP Client (NHC).

This document does not rely on the use of NHRP Registration Request/Reply messages, as DCBRs/NHSs rely on the information provided by E-VPN.

DCBR-SF will be an authoritative NHS for all the IP addresses of the VMs that are presently in the SF data center. Likewise, DCBR-LA will be an authoritative NHS for all the IP addresses of the VMs that are presently in the LA data center. Note that as a VM moves from SF to LA, the authoritative NHS for the IP address of that VM moves from DCBR-SF to DCBR-LA.

We assume that the border router in Denver can determine the subset of the destination for which it has to apply NHRP. If DCBR-SF, DCBR-LA, and the border router in Denver use OSPF to exchange routing information, then a way to do this would be for DCBR-SF and DCBR-LA to use a particular OSPF tag to mark routes advertised by these DCBRs, and then make the border router in Denver to apply NHRP to any destination that matches any route that carries that particular tag. If DCBR-SF, DCBR-LA, and the border router in Denver use BGP to exchange routing information, then a way to do this would be for DCBR-SF and DCBR-LA to use a particular BGP community to mark routes advertised by these DCBRs, and then make the border router in Denver to apply NHRP to any destination that matches any route that carries that particular BGP community.

6.2.3.1. Detailed Procedures

The following describes details of NHRP operations.

When the border router in Denver first receives a packet from B destined to VM-A, the border router determines that VM-A falls into the subset of the destination for which the border router has to apply NHRP. Therefore, the border router originates an NHRP Request.

The mandatory part of the NHRP Request is constructed as follows.

The Source NBMA Address and the Source Protocol Address fields contain the IP address of the border router in Denver; the Destination Protocol Address field contains the IP address of VM-A. This Request is encapsulated into an IP packet, whose source IP address is the address of the border router, and whose destination IP address is the address of VM-A. The packet carries the Router Alert option. NHRP is carried directly over IP using IP Protocol Number 54 [RFC1700].

Note that the trigger for the originating an NHRP Request may be either the first packet destined to a particular host, or a particular rate threshold for the traffic to that host.

Following the route to the prefix X the packet that carries the NHRP Request will eventually get to either DCBR-SF or DCBR-LA. Let's assume that it is DCBR-SF that receives the packet. (Note that none of the routers, if any, between the site border router in Denver and DCBR-SF or DCBR-LA would be required to support NHRP.) Since both DCBR-SF and DCBR-LA assume to support NHRP, they would be required to process the NHRP Request carried in the packet.

If DCBR-SF determines that VM-A is in the LA data center (DCBR-SF determines this from the information provided by E-VPN), then DCBR-SF will forward the packet that contains the NHRP Request to DCBR-LA, as DCBR-SF is not an authoritative NHS for VM-A, while DCBR-LA is. DCBR-SF can accomplish this by setting the destination MAC address in the packet to the MAC address of DCBR-LA, in which case the packet will be forwarded to DCBR-LA using the E-VPN procedures. Alternatively, DCBR-SF could change to DCBR-LA the destination address in the IP header of the packet that carries the NHRP Request, in which case the packet will be forwarding to DCBR-LA using IP forwarding procedures.

When the NHRP Request will reach DCBR-LA, and DCBR-LA determines that VM-A is in the LA data center (DCBR-LA determines this from the information provided by E-VPN), and thus DCBR-LA is an authoritative NHS for VM-A, DCBR-LA sends back to the border router in Denver an NHRP Reply indicating that DCBR-LA should be used for forwarding traffic to VM-A.

The mandatory part of the NHRP Reply is constructed as follows. The Source NBMA Address, the Source Protocol Address, and the Destination Protocol Address fields in the mandatory part are copied from the corresponding fields in the NHRP Request. The Reply carries a Client Information Entry (CIE) with the Client NBMA Address field set to the IP address of DCBR-LA, and the Client Protocol Address field set to the IP address of VM-A. The Reply is encapsulated into an IP packet, whose source IP address is the address of DCBR-LA, and whose

destination IP address is the IP address of the border router in Denver (DCBR-LA determines this address from the information carried in the NHRP Request). The packet does not carries the Router Alert option.

Once the border router in Denver receives the Reply, the border router will encapsulate all the subsequent packets destined to VM-A into GRE with the outer header carrying DCBR-LA as the IP destination address. (In effect that means that the border router in Denver will install in its FIB a host route for VM-A indicating GRE encapsulation with DCBR-LA as the destination IP address in the outer header.)

Now assume that VM-A moves from the data center in LA to the data center in SF. Once DCBR-LA finds this out (from the information provided by E-VPN), DCBR-LA sends an NHRP Purge to the border router in Denver. Note that DCBR-LA can defer sending the Purge message until it receives GRE-encapsulated data destined to VM-A. Note also, that in this case DCBR-LA does not have to keep track of all the requestors for VM-A to whom DCBR-LA subsequently sent NHRP Replies, as DCBR-LA determines the address of these requestors from the outer IP header of the GRE tunnel.

When the border router in Denver receives the Purge message, it will purge the previously received information that VM-A is reachable via DCBR-LA. In effect that means that the border router in Denver will remove the host route for VM-A from its FIB (but will still retain a route for the prefix X).

From that moment the border router in Denver will start forwarding packets destined to VM-A using the route to the prefix X (relying on plain IP routing). That means that these packets will get to DCBR-SF (which is the desirable outcome anyway).

However, once the border router in Denver receives NHRP Purge, the border router will issue another NHRP Request. This time, once this NHRP Request reaches DCBR-SF, DCBR-SF will send back to the border router in Denver an NHRP Reply, as at this point DCBR-SF determines that VM-A is in SF, and therefore DCBR-SF is an authoritative NHS for VM-A. Once the border router in Denver receives the Reply, the router will encapsulate all the subsequent packets destined to VM-A into GRE with the outer header carrying DCBR-SF as the IP destination address. In effect that means that the border router in Denver will install in its FIB a host route for VM-A indicating GRE encapsulation with DCBR-SF as the destination IP address in the outer header.

6.2.3.2. Failure scenarios

To illustrate operations during failures let's modify the original example by assuming that each data center has more than one DCBR. Specifically, the data center in SF has DCBR-SF1 and DCBR-SF2. Both of these are authoritative NHSs for all the VMs whose addresses are take from prefix X, and who are presently in the SF data center. Note also that both DCBR-SF1 and DCBR-SF2 advertise a route to the prefix X.

Assume that VM-A is presently in SF, so the border router in Denver tunnels the traffic to VM-A through DCBR-SF1.

Now assume that DCBR-SF1 crashes. At that point the border router in Denver should stop tunnelling the traffic through DCBR-SF1, and should switch to DCBR-SF2. The following sections describe two possible options to accomplish this.

6.2.3.2.1. DCBR Failure - Option 1

One option to handle DCBRs failures is to make each DCBR to originate a host route for its own IP address that it would advertise in the NHRP Replies. This way when DCBR-SF1 crashes, the route to DCBR-SF1 IP address goes away, providing indication to the border router in Denver that it no longer can use DCBR-SF1. At that point the border router in Denver removes the route for VM-A from its FIB (but will still retain a route for the prefix X). From that moment the border router in Denver will start forwarding packets destined to VM-A using the route to the prefix X. Since DCBR-SF1 crashes, these packets will be routed to DCBR-SF2, as DCBR-SF2 advertises a route to prefix X (and the route to prefix X that has been previously advertised by DCBR-SF1 will be withdrawn due to crash of DCBR-SF1).

However, once the border router in Denver detects that DCBR-SF1 is down, the border router will issue another NHRP Request. This time, NHRP Request reaches DCBR-SF2, and DCBR-SF2 will send back to the border router in Denver an NHRP Reply. Once the border router in Denver receives the Reply, the router will encapsulate all the subsequent packets destined to VM-A into GRE with the outer header carrying DCBR-SF2 as the IP destination address. In effect that means that the border router in Denver will install in its FIB a host route for VM-A indicating GRE encapsulation with DCBR-SF2 as the destination IP address in the outer header.

6.2.3.2.2. DCBR Failure - Option 2

Another option to handle DCBRs failures is to make both DCBRs to advertise the same (anycast) IP address in the NHRP Replies. This way when DCBR-SF1 crashes, the route to this address would not go away (as DCBR-SF2 will continue to advertise a route to this address into IP routing), and thus the traffic destined to that address will now go to DCBR-SF2. Since DCBR-SF2 is an authoritative NHS for all the VMs whose addresses are take from prefix X, and who are presently in the SF data center, DCBR-SF2 will forward this traffic to VM-A.

7. IANA Considerations

This document introduces no new IANA Considerations.

8. Security Considerations

TBD.

9. Acknowledgements

We would like to thank Dave Katz for reviewing NHRP procedures. We would also like to thank John Drake for his review and comments.

10. References

[RFC1700] Reynolds J., Postel J., "ASSIGNED NUMBERS", RFC1700, October 1994

[RFC2332] "NBMA Next Hop Resolution Protocol (NHRP)", RFC 2332, J. Luciani et. al.

[RFC4364] Rosen, Rekhter, et. al., "BGP/MPLS IP VPNs", RFC4364, February 2006

[RFC4684] Pedro Marques, et al., "Constrained Route Distribution for Border Gateway Protocol/MultiProtocol Label Switching (BGP/MPLS) Internet Protocol (IP) Virtual Private Networks (VPNs)", RFC4684, November 2006

[E-VPN] Aggarwal R., et al., "BGP MPLS Based Ethernet VPN", draft-ietf-l2vpn-evpn, work in progress

[Default-Gateway] http://www.iana.org/assignments/bgp-extended-communities

11. Author's Address

Rahul Aggarwal
Arktan, Inc
Email: raggarwa_1@yahoo.com

Yakov Rekhter
Juniper Networks
1194 North Mathilda Ave.
Sunnyvale, CA 94089
Email: yakov@juniper.net

Wim Henderickx
Alcatel-Lucent
Email: wim.henderickx@alcatel-lucent.com

Ravi Shekhar
Juniper Networks
1194 North Mathilda Ave.
Sunnyvale, CA 94089
Email: rshekhar@juniper.net

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

Ali Sajassi
Cisco Systems
Email: sajassi@cisco.com