Network Working Group
Internet Draft
Category: Standards Track
Expiration Date: March 2012

R. Aggarwal
Arktan Inc

Y. Rekhter
Juniper Networks

W. Henderickx
Alcatel-Lucent

R. Shekhar
Juniper Networks

September 6, 2011

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

draft-raggarwa-data-center-mobility-01.txt

Status of this Memo

Copyright and License Notice

Abstract

This document describes a set of solutions for seamless mobility in the data center. These solutions provide a tool-kit which is based on IP routing, BGP/MPLS MAC-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 solutions for seamless mobility in the data center. Mobility in the data center is defined as the ability to move a virtual machine (VM) from one server in the data center to another server in the same or different data center while retaining the IP address and the MAC address of the VM. The latter is necessary to provide seamless application experience. The term mobility or the reference to moving a VM in this document, should be considered to imply seamless mobility, unless otherwise stated. It is also to be noted that VM mobility doesn't change the VLAN/subnet associated with the VM. Infact VM mobility requires the VLAN to be "extended" to the new location of the VM.

Data center providers have expressed a desire to provide the ability to move VMs across data centers, where the data centers may be in different geographical locations. 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 provides a set of solutions for VM mobility. The practical applicability of these solutions will depend on these constraints. However the solutions described here provide a framework that enables VMs to be moved across both small and large geographical distances. In other words if these constraints relax 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 Data Center Switch (DCS) is used to refer to a switch in the data center that is connected to the servers that host VMs. A data center may have multiple DCSes. Each data center also has one or more Data Center Border Routers (DCB) that connect to other data centers and to the Wide Area Network (WAN). A DCS may act as a DCB.

This document also uses the terms MAC-VPN and Ethernet-VPN (E-VPN) inter-changeably.

3. Problem Statement

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

3.1. Layer 2 Extension

The first problem is to extend the VLAN of a VM across DCSes where the DCSes may be located in the same or different data centers. This is required to enable the VM to move between the DCSes. We will refer to this as the "layer 2 extension problem".

3.2. Optimal Intra-VLAN Forwarding

The second issue has to do with optimal forwarding in a VLAN in the presence of VM mobility, where VM mobility may involve multiple data centers.

Optimal forwarding in a VLAN by definition implies that traffic between VMs that are in the same VLAN, should not traverse DCBs in data centers that contain neither of these VMs, except if:

The DCBs in these data centers are on the layer 2 path between the DCBs in the data centers that contain the VM.

Optimal forwarding in a VLAN also implies that traffic between a client and a VM that are in the same VLAN, should not traverse DCBs in the data centers that do not contain the VM, except if:

The DCBs in these data centers are on the layer 2 path between the client site border router and the DCBs in the data centers that contain the VM.

3.3. Optimal Routing

Optimal routing, in the presence of intra-data center VM mobility, implies that traffic between VMs that are on different VLANs/subnets should not traverse a DCS or DCB in that data center that does not host these VMs, except if:

The DCS or DCBs are on an IP path between the DCSes that host the VMs.

Optimal routing, in the presence of inter-data center VM mobility, implies that traffic between VMs that are on different VLANs/subnets should not traverse DCBs in data centers that contain neither of these VMs, except if:

The DCBs in these data centers are on an IP path between the DCBs in the data centers that contain the VMs.

Optimal routing also implies that traffic between a VM and a client that are on different VLANs/subnets should not traverse any of the DCBs in data centers that do not contain the VM, except if:

The DCBs in these data centers are on an IP path between the client's site border roueter and the DCB of the data center that contains the VM.

Specifically optimal routing requires a mechanism that ensures that the default gateway of a VM can be in the geographical proximity of the VM as the VM moves. Consider a VM that moves from data center 1 (DC1) to data center 2 (DC2). Further consider that the default gateway of the VM is located in DC1. Once the VM moves it is desirable to avoid carrying traffic originating from the VM, destined to other subnets, back to the default gateway in DC1, as this may not be optimal. We will refer to this as the "VM default gateway problem".

Optimal routing also requires mechanisms to avoid "triangular routing" to ensure that the traffic destined to a given VM would not traverse through a DCB of a data center that does not contain the VM. For example packets from VM1 and VM2 that are both in data center 1 (DC1) but on different VLANs/subnets should not go to data center 2 (DC2) and back to DC1. This can be the case if VM2 moves from DC2 to DC1, unless additional mechanisms are built to prevent this.

4. Layer 2 Extension and Optimal Intra-VLAN Forwarding Solution

The solution for the "layer 2 extension problem", particularly when the DCSes are located in different data centers, relies on MAC-VPNs [MAC-VPN]. A DCS may be enabled with MAC-VPN, in which case it acts as an MPLS Edge Switch (MES). However this is not a requirement. It is required for the DCBs to be enabled with MAC-VPN to enable layer 2 extension across data centers. DCBs learns MAC routes within their own data center either via MAC-VPN state exchange with the DCSes, or via data plane learning, or other layer 2 protocols between the DCSes

and the DCBs. The DCBs MUST advertise these MAC routes as MAC-VPN routes. This way DCBs in one data center learns about MAC routes in other data centers. The specifics of such advertisement depends on the inter-connect between the DCBs as described below.

- IP, MPLS (e.g., or Layer 2 Interconnect between the DCBs; and between the client site border router and the DCBs. In this case the provider of the IP, MPLS or Layer 2 Interconnect does not participate in MAC-VPN. The DCBs MUST exchange MAC-VPN routes using either IBGP or (multi-hop) EBGP peering. In addition if DCSes support MAC-VPN the DCBs MUST act as BGP Route Reflectors (RRs). IBGP peering may utilize additional RRs in the data center infrastructure (RR hierarchy). Note that in this scenario the provider of the IP, MPLS or Layer 2 Interconnect is not involved in these IBGP or EBGP peerings/exchanges.

- MAC-VPN as a Data Center Interconnect (DCI) service. The DCI service may be offered by a Service Provider (SP). There are two variants to this model. In the first variant the WAN Border Router is the same as the DCB. In other words DCB is provided by the SP and may be used to provide DCI for DCSes belonging to multiple enterprises. The DCSes may connect to the DCBs using Layer 2 Protocols or even MAC-VPN peering. The DCBs MUST exchange MAC-VPN routes between themselves. The DCBs may utilize BGP RRs to exchange such routes. If there is MAC-VPN peering between the DCB and the DCSes within the DCB's own data center, then the DCB propagates the MAC-VPN routes that it learns from other DCBs to the DCSes within its own data center,

  In the second variant the WAN Border Router is not the same device as the DCB. In this variant the DCBs may connect to the WAN Border routers using layer 2 protocols. Or WAN Border Routers may establish MAC-VPN peering with the DCBs in which case the DCBs MUST advertise the MAC-VPN routes using either IBGP or (multi-hop) EBGP to the WAN Border routers. The WAN Border routers MUST exchange MAC-VPN routes between themselves. The WAN Border routers may utilize BGP RRs to exchange such routes. A WAN Border router propagates the MAC-VPN routes that it learns from other WAN Border routers to the DCBs that it is connected to if there is MAC-VPN peering between the DCBs and the WAN Border Routers.

Please note that the propagation scope of MAC-VPN routes for a given VLAN/subnet is constrained by the scope of data centers that span that VLAN/subnet and this is controlled by the Route Target of the

MAC-VPN routes.

The use of MAC-VPN ensures that traffic between VMs and clients, that are on the same VLAN, is optimally forwarded irrespective of the geographical extension of the VLAN. This follows from the observation that MAC-VPN inherently enables disaggregated forwarding at the granularity of the MAC address of the VM. MAC-VPN also allows aggregating MAC-VPN addresses into MAC prefixes. Optimal intra-VLAN forwarding requires propagating VM MAC addresses and comes at the cost of of disaggregated forwarding within a given data center. However such disaggregated forwarding is not necessary between data centers. For example for a MAC-VPN enabled DCS, this DCS has to maintain MAC routes only to the VMs within its own data center, and then point a "deafult MAC route" to the DCB of that data center. Another example would be advertisement of prefix-MAC routes by a DCS/DCB when its possible to assign a structure to the MAC addresses.

This document assumes that the VM's VLAN and policy, e.g., firewalls, associated with a VM are present on the DCS to which the VM moves. If this is not the case then in addition to MAC-VPNs layer 2 extension requires the ability to move policies dynamically. The procedures for doing so are for further study.

5. Optimal VM Default Gateway Solution

The solution for the "VM default gateway problem" relies on requiring the ability to perform routing at each DCB. This is in addition to requiring layer 2 forwarding and MAC-VPN functionality on a DCB. In addition it is desirable to be able to perform routing on the DCSes.

Please note that when a VM moves the default gateway IP address of the VM may not change. Further the ARP cache of the VM may not time out. Rest of this section is written with this in mind.

First consider the case where each DCB acts as a router but the DCSes do not act as routers. In this case the default gateway of a VM, that moves in the geographical proximity of a new DCB, may be the new DCB as long as there is a mechanism for the new DCB to be able to route packets that the VM sent to the "original" default gateway's MAC address.

Now consider the case where one or more DCSes act as a router. In this case the default gateway of a VM, that moves to a particular DCS, may be the new DCS as long as there is a mechanism for the new DCS to be able to route packets sent by the VM to the "original" default gateway's MAC address.

There are two mechanisms to address the above cases.

The first mechanism relies on the use of an anycast default gateway IP address and an anycast default gateway MAC address. These anycast addresses are configured on each DCB that is part of the layer 2 domain. This requires co-ordination to ensure that the same anycast addresses are configured on DCBs, which may or may not be in the same data center, that are part of the same layer 2 domain. The anycast addresses are also configured on the DCSes that act as routers. This ensures that a particular DCB or DCS, when the DCS acts as a router, can always route packets sent by a VM to the anycast default gateway MAC address. It also ensures that such DCB or DCS can respond to the ARP request for the anycast IPaddress, generated by a VM. This mechanism

The second mechanism lifts the restriction to configure the anycast default gateway addresses on each DCB or DCSes. This is accomplished by each DCB and the DCSes that act as routers, propagating, in the BGP MAC-VPN control plane, its default gateway IP and MAC address using the MAC advertisement route. To accomplish this the MAC advertisement route MUST be advertised as per the procedures in [MAC-VPN]. The MAC address in such an advertisement MUST be set to the default gateway MAC address of the DCB or DCS. The IP address in such an advertisement MUST be set to the default gateway IP address of the DCB or DCS. A new BGP community called the "Default Gateway Community" MUST be included with the route. Each DCB or DCS that receives this route and imports it as per the procedures o f [MAC-VPN] SHOULD:

- Create forwarding state that enables it to route packets destined to the default gateway MAC address of the advertising DCB or DCS.

- As an optimization, optionally, reply to ARP requests, that it receives, destined to the default gateway IP address of the advertising DCB or DCS. The MAC address in the ARP response should be the MAC address associated with the IP address to which the ARP was sent.

6. Triangular Routing Solution

There are two Triangular Routing solutions proposed in this document.

The first Triangular Routing Solution is based on propagating routes to VM host IP addresses (/32 IPv4 or /128 IPv6) using IP routing or BGP/MPLS VPNs [RFC 4364] with careful consideration given to constraining the propagation of these addresses.

The second solution relies on using Next Hop Resolution Protocol (NHRP).

The section "Triangular Routing Solution based on Host Routes" describes the details of the first solution. The section "Triangular Routing Solution based on NHRP" describes details of the second solution.

7. Triangular Routing Solution Based on Host Routes

The solution to the triangular routing problem based on MAC-VPN, IP routing or BGP/MPLS VPNs [RFC 4364] relies on the propagation of the host IP address of the VM. Further the solution provides a toolkit to constrain the scope of the distribution of the host IP address of the VM. In other words the solution relies on disaggregated routing with the ability to control which nodes in the network have the disaggregated information and also the ability to aggregate this information as it propagates in the network.

The solution places the following requirements on DCSes and DCBs:

- A given DCB MUST implement IP routing using OSPF/IS-IS or/and BGP. A given DCB MAY implement BGP/MPLS VPNs. A DCB MUST implement MAC-VPN.

- A given DCS MAY implement IP routing using OSFP/IS-IS. A DCS MAY implement IP routing using BGP. A DCS MAY implement BGP/MPLS VPNs. A DCS MAY implement MAC-VPN.

To accomplish this each DCS/DCB SHOULD advertise the IP addresses of the VMs, in MAC-VPN, IP routing or using VPN IPv4 or VPN IPv6 address family, as per IP VPN [RFC 4364] procedures. The IP address of a VM maybe learned by an DCS either from data plane packets generated by the VM or from the control/management plane, if there is a control/management plane integration between the server hosting the VM and the DCS.

The propagation of the VM host IP addresses advertised by an DCS/DCB is constrained to a set of DCSes/DCBs. Such constrained distribution needs to address three main scenarios:

- Scenario 1. Traffic between VMs that are on different VLANs/subnets in the same data center. This scenario assumes that VM can move only among DCSes that are in the same data center.

- Scenario 2. Traffic between VMs (or between a VM and a client) that are on different VLANs/subnets in different DCs, but the DCs are in close geographical proximity. An example of this is multiple DCs in San Francisco or DCs in San Francisco and Los Angeles. This scenario assumes that VM can move only among DCs that are in close geographical proximity.

- Scenario 3. Traffic among VMs (or between a VM and a client) that are on different VLANs/subnets, in different DCs, and these DCs are not in close geographical proximity. An example of this is DCs in San Francisco and Denver. In this scenario VM may move among DCs that are not in close geographical proximity

7.1. Scenario 1

A DCS may originate /32 or /128 routes for all VMs connected to it. These routes may be propagated using MAC advertisement routes in MAC-VPN, along with the MAC address of the VM. Or they may be propagated using OSPF or IS-IS or BGP or even using BGP VPN IPv4/IPv6 routes [RFC 4364]. In either case the distribution scope of such routes is constrained to only the DCSes and the DCBs in the data center to which the DCS belongs. If BGP is the distribution protocol then this can be achieved by treating DCBs as the Route Reflectors. If OSPF/IS-IS is the routing protocol then this can be achieved by treating the data center as an IGP area.

When MAC-VPN is used for distributing VM host IP routes by DCSes, within the data center, then the Route Target of such routes must be such that the routes can be imported by all the DCSes and DCBs in the data center, even if they do not have members in the VLAN associated with the MAC address in the route. When a DCS or DCB imports such a route, then it should create IP forwarding state to route the IP address present in the advertisement with the next-hop as the DCS/DCB from which the advertisement was received.

Consider a VM in a VLAN connected to DCS1 that sends a packet to a VM, in another VLAN, connected to DCS2. Further consider that DCS1 and DCS2 are in the same data center. Then DCS1 will be able to route

the packet optimally to DCS2. For instance this packet may be sent directly from DCS1 to DCS2 without having to go through a DCB, if there is physical connectivity between DCS and DCS2. This is because DCS1 would have received and imported the host IP route to reach the destination VM.

7.2. Scenario 2: BGP as the Routing Protocol between DCBs

A DCS MAY advertise /32 or /128 routes for all VMs connected to it using the procedures described in "Scenario 1". Note that the DCSes may use OSPF or IS-IS or BGP as the routing protocol.

If a DCS advertises host routes as described above then the DCBs in the data center MUST learn the VM host routes within their data center from the routes advertised by the DCSes. If the DCSes do not advertise host routes but implement MAC-VPN then the DCSes SHOULD advertise the IP address of a VM along with the MAC advertisement for that VM. In this case the DCBs MUST learn the VM host IP addresses from the MAC advertisement routes. If the DCSes neither advertise VM host routes nor implement MAC-VPN then DCBs must rely on data plane snooping to learn the MAC addresses of the VMs.

The DCBs in the data center originate /32 or /128 routes for all the VMs within their own data center as BGP IPv4/IPv6 routes or as BGP VPN IPv4/IPv6 routes. These routes are propagated to other DCBs that are in data centers in close geographical proximity of the data center originating the routes. To achieve this the routes carry one or more Route Targets (RT). These route targets control which of the other DCBs or Route Reflectors import the route.

One mechanism to constrain the distribution of such routes is to assign a RT per DCB or per set of DCBs. This set of DCBs may be chosen based on geographical proximity. Note that when BGP/MPLS VPNs are used this RT is actually per {VPN, DCB} tuple or {VPN, set of DCBs} tuple. The rest of this section will refer to this as "DCB Set RT" for simplicity.

Each DCB in a particular set of data centers is then configured with this RT. A DCB may belong to multiple data center sets and hence may be configured with multiple DCB Set RTs. If a DCB that is in one or more Data Center Sets advertises a VM host IP address route, it MUST include all the DCB Set RTs it is configured with along with the route. This results in each DCB that is part of one or more of these Data Center Sets to import the route.

A DCB MAY advertise a default IP route to the DCSes in its own data center employing a "vitual hub-and-spoke" methodology. Or a DCB MAY

advertise the IP routes received from other DCBs to the DCSes in its own data center.

Consider a VM or a client in a VLAN in an IP VPN in particular data center that sends a packet to a VM, in another VLAN. Further consider that the destination VM is in a data center which is in the same data center set as the sender VM or client. Then the DCS that the sender VM or client is connected to will able to route the packet optimally. This is because the DCB in this DCS's data center would have received and imported the host IP route to reach the destination VM. Note that the DCS may have imported only a default route advertised by the DCB in the DCS's own data center.

Now consider that the sender VM's or client's data center and the destination VM's data center are not in the same Data Center Set. In this case the packet sent by the sender VM or client will first be routed as per the best IP prefix route to reach the destination VM. The next-hop DCB of this route may be in the same Data Center Set as the destination VM's data center, in which case this next-hop DCB will be able to route the packet optimally. If this is not the case then the packet will be forwarded by the next-hop DCB as per its best route.

Constraining the VM host IP address route using the DCB Set RT provides a mechanism for optimal routing within the set of data centers that are configured with the DCB Set RT.

For example consider data centers in San Francisco and Los Angeles. All the DCBs in these data centers may be assigned a particular Data Center Set import RT, RT1. Further each DCB advertises VM host IP addresses with RT1. As a result it is possible to perform optimal routing of packets destined to a VM in one of these data centers if the packet is originated by a VM or client in one of these data centers. It is also possible to perform this optimal routing for a packet that is originated outside these data centers, once the packet reaches a DCB in these data centers. However if there are multiple entry points i.e., DCBs in these data centers then this mechanism is not sufficient for WAN routers to optimally route the packet to the DCB, that the VM is closest to. Please see the section on "Sceanrio 3: Using BGP as the Routing Protocol" for procedures on how to achieve this.

7.3. Scenario 2: OSPF/IS-IS as the Routing Protocol between DCBs

A DCS MAY advertise /32 or /128 routes for all VMs connected to it using the procedures described in "Scenario 1". Note that the DCSes may use OSPF or IS-IS or BGP as the routing protocol.

If a DCS advertises host routes as described above then the DCBs in the data center MUST learn the VM host routes within their data center from the routes advertised by the DCSes. If the DCSes do not advertise host routes but implement MAC-VPN then the DCSes SHOULD advertise the IP address of a VM along with the MAC advertisement for that VM. In this case the DCBs MUST learn the VM host IP addresses from the MAC advertisement routes. If the DCSes neither advertise VM host routes nor implement MAC-VPN then DCBs must rely on data plane snooping to learn the MAC addresses of the VMs.

DCBs must follow IGP procedures to propagate the host routes within the non-backbone IGP area to which they belong.

"Geographical proximity" is defined by an IGP area. The /32 /128 routes are only propagated in the non-backbone IGP area to which the DCSes and DCB belong. This assumes that geographically proximate data centers are in their non-backbone IGP area. This solution is a natural fit with the OSPF/IS-IS model of operations. It avoids triangular routing when the sender VM/client and destination VM/client are in the same IGP area using principles that are very similar to those described in the section "Scenario 2: BGP as the Routing Protocol".

7.4. Scenario 3: Using BGP as the Routing Protocol

The mechanisms to address Scenario 2 does not address Scenario 3. Specifically they do not address the distribution of VM host IP routes between DCBs that are not in close geographical proximity. This distribution may be necessary if it is desirable to ensure that a packet from a data center, outside the set of data centers desribed above, is to be routed to the optimal entry point in the set. For example if a VM in VLAN1 moves from San Francisco to Los Angeles, then it may be desirable to route packets from New York to Los Angeles without going through San Francisco, if such a path exists from New York to Los Angeles.

The section "Base Solution" describes the base solution to address Scenario 3 based on BGP as the routing protocol. The section "Refinements" describes the modifications to these base procedures to improve the scale of the solution.

7.4.1. Base Solution

A given DCB MUST advertise in IP routing routes for the IP subnets configured on the DCB. These are NOT host (/32 /128) routes. Instead these are prefix/aggregated routes. Further DCB of a given data center MUST originate into BGP IPv4/IPv6 or VPN IPv4/IPv6 host routes for all the VMs currently being present within its own DC. These routes are propagated to all DCBs in all data centers. This requires all host routes to be maintained by all DCBs at least in the control plane.

This base solution may impose significant control plane overhead depending on the number of VM host IP addresses across all data centers. However it may be applicable as is in certain environments.

Please see the next section "Refinements" for procedures that may be employed to improve the scale of this solution.

7.4.2. Refinements: SP Unaware of DC Routes

We first consider the case where the SP does not participate in data center routing. Instead the SP just provides layer 2 or IP connectivity between the DCBs.

In this case the VM host routes are propagated by the DCBs to the Route Reflectors (RRs) where the RRs are part of the data center infrastructure. Distribution of these routes to the RRs is constrained using Route Target that is configured on all RRs. In addition such VM host routes also carry the DCB Set RTs as described in "Section 2: BGP as the Routing Protocol". The RRs propagate such routes to all the DCBs that belong to the DCB Set RTs present in the route.

In addition the propagation of these routes from RRs to other DCBs and/or client site border routers is done on demand. A given DCB, that needs to send traffic to a particular VM in some other data center would dynamically/on-demand request the host route to that VM from its RR using "prefix-based Outbound Route Filter (ORF)" A DCB can determine whether it requires a VM host IP address based on policy. For example the policy may be based on high volume of traffic to the destination IP address of the VM. This mechanism reduces the number of host routes that a DCB needs to maintain. Likewise, a given client site border router that needs to send traffic to a particular VM would dynamically/on-demand request the host route to that VM using prefix-based ORF. This reduces the number of host routes that client site border router needs to maintain.

7.4.3. Refinements: SP Participates in DC Routing

This section considers the case where the SP offers Inter-DC routing as a service. To enable this the IPv4/IPv6 or VPN IPv4/IPv6 host VM routes need to be propagated by the SP.

The first variant of this is the case where the DCBs are managed by the SP and the WAN Border Router is the same device as the DCB. The procedures of this variant are the same as those in "Refinements: WAN Unaware of DC Routes" except that the DCBs and the RR infrastructure is managed by the SP. In this variant it is desirable that the inter-DCB routing protocol is based on BGP/MPLS IP VPNs.

The second variant of this is the case where the WAN Border Router and DCBs are separate devices and DCBs are not managed by the SP. In this variant the DCBs first need to propagate the routes to the WAN border routers. This can be done by configuring the WAN border routers with the Data Center Set RTs of all the data centers that the WAN border routers are connected to. WAN border routers would then need to import BGP IPv4/IPv6 or VPN IPv4/IPv6 routes that carry one of these RTs.

Next the WAN border routers maybe configured to propagate such routes. As they propagate such routes, they MUST include a RT that controls which other routers in the WAN import such routes.

One possible mechanism is to propagate such routes only to Route Reflectors (RRs) in the WAN. This can be accomplished by configuring the RRs with a particular import RT and by propagating the routes at the WAN border routers along with this RT. Now DCBs or border routers or PEs in the WAN can d dynamically request routes using prefix-based ORF for one or more host VM addresses.

For instance the policy maybe to request such routes for a particular host address if the traffic to that host address exceeds a certain threshold. This does require data plane statistics to be maintained for flows. This policy may be implemented on a WAN border router or PE which can then dynamically request host routes from a RR using BGP Outbound Route Filtering (ORF).

7.5. VM Motion

The procedures described in this document require that a DCS that originates a VM host IP route MUST be able to detect when that VM moves to another DCS. If DCSes support MAC-VPN then the procedures in MAC-VPN MUST be used to detect VM motion. If DCSes do not support MAC-VPN then the DCSes must rely on layer 2 mechanisms or control plane/management plane interaction between the DCS and the VM to detect VM motion.

When the DCS detects such VM motion it MUST withdraw the host VM route, that it advertised, from IGP or BGP.

7.6. Policy based origination of VM Host IP Address Routes

When a DCS/DCB learns the host IP address of a VM it may not originate a corresponding VM host IP address route by default. Instead it may optionally do so based on a dynamic policy. For example the policy maybe to originate such a route only when the traffic to the VM exceeds a certain threshold.

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

When a DCS/DCB learns the host IP address of a VM, from another DCS or DCB, it may not immediately install this route in the forwarding table. Instead it may optionally do so based on a dynamic policy. For example the policy maybe to install such forwarding state only when the first packet to that particular VM is received.

8. Triangular Routing Solution Based on NHRP

8.1. Overview

The following describes a scenario where a client within a given customer site communicates with a VM, and the VM could move among several data centers (DCs).

Assume that a given VLAN/subnet, subnet X, spans two DCs, one in SF and another in LA. DCB-SF is the DCB for the SF DC. DCB-LA is the DCB for the LA DC. Since X spans both the SF DC and the LA DC, both DCB-SF and DCB-LA advertise a route to X (this is a route to a prefix, and not a /32 route).

DCB-LA and DCB-SF can determine whether a particular VM on that VLAN/subnet is in LA or SF by running MAC-VPN (and exchanging MAC-VPN

routes among themselves).

There is a site in Denver, and that site contains a host B that wants to communicate with a particular VM, VM-A, on the subnet X.

Assume that there is an IP infrastructure that connects the border router of the site in Denver, DCB-SF, and DCB-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 DCB-SF or DCB-LA].

Goal: If VM-A is in LA, then the border route in Denver sends traffic for VM-A via DCB-LA without going first through DCB-SF. If VM-A is in SF, then the border route in Denver send traffic for VM-A via DCB-SF without going first through DCB-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, DCB-SF, and DCB-LA to support NHRP, and support GRE encapsulation. In NHRP terminology DCB-SF and DCB-LA are NHSs, while the border router in Denver is an NHC.

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

DCB-SF will be an authoritative NHS for all the /32s from X that are presently in the SF DC. Likewise, DCB-LA will be an authoritative NHS for all the /32s from X that are presently in the LA DC. Note that as a VM moves from SF to LA, the authoritative NHS for the IP address of that VM moves from DCB-SF to DCB-LA.

We assume that the border router in Denver can determine the subset of the destination for which it has to apply NHRP. One way to do this would be for DCB-SF and DCB-LA to use OSPF tag to mark a route for X, and then make the border router in Denver to apply NHRP to any destination that matches any route that carries that particular tag. Another way to do this wouldbe for DCB-SF and DCB-LA to use a particular BGP community to mark a route for X, and then make the border router in Denver to apply NHRP to any destination that matches any route that carries that particular BGP community.

8.2. 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. [Note that the trigger for the originating an NHRP Request may be either the first packet destined to a particular /32, or a particular rate threshold for the traffic to that /32.] 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].

Following the route to X, the packet will eventually get to either DCB-SF or DCB-LA. Let's assume that it is DCB-SF that receives the packet. [None of the routers, if any, between the site border router in Denver and DCB-SF or DCB-LA would be required to support NHRP.] However, since both DCB-SF and DCB-LA assume to support NHRP, they would be required to process the NHRP Request carried in the packet.

If DCB-SF determines that VM-A is in LA (DCB-SF determines this from the information provided by MAC-VPN), then DCB-SF will forward the packet to DCB-LA, as DCB-SF is not an authoritative NHS for VM-A, while DCB-LA is. [A way for DCB-SF to forward the packet to DCB-LA would be for DCB-SF to change to DCB-LA the destination address in the IP header of the packet. Alternatively, DCB-SF could keep the original destination address in the IP header, but set the destination MAC address to the MAC address of DCF-LA.]

When the NHRP Request will reach DCB-LA, and DCB-LA determines that VM-A is in LA (DCB-LA determines this from the information provided by MAC-VPN), and thus DCB-LA is an authoritative NHS for VM-A, DCB-LA sends back to the border router in Denver an NHRP Reply indicating that DCB-LA should be used for forwarding traffic to VM-A (When sending the NHRP Reply, DCB-LA determines the address of the border router in Denver from the NHRP Request). 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 DCB-LA as the IP destination address. [In effect that means that the border router in Denver will install in its FIB a /32 route for VM-A indicating GRE encapsulation with DCB-LA as the destination IP address in the outer header.]

Now assume that VM-A moves from LA to SF. Once DCB-LA finds this out (DCB-LA finds this out from the information provided by MAC-VPN),

DCB-LA sends an NHRP Purge to the border router in Denver. [Note that DCB-LA can defer sending the Purge message until it receives GRE-encapsulated data destined to VM-A. Note also, that in this case DCB-LA does not have to keep track of all the requestors for VM-A to whom DCB-LA subsequently sent NHRP Replies, as DCB-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 previousely received information that VM-A is reachable via DCB-LA. In effect that means that the border router in Denver will remove /32 route for VM-A from its FIB (but will still retain a route for X).

>From that moment the border router in Denver will start forwarding packets destined to VM-A using the route to the subnet X (relying on plain IP routing). That means that these packets will get to DCB-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 DCB-SF, DCB-SF will send back to the border router in Denver an NHRP Reply (as at this point DCB-SF determines that VM-A is in SF, and therefore DCB-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 DCB-SF as the IP destination address. In effect that means that the border router in Denver will install in its FIB a /32 route for VM-A indicating GRE encapsulation with DCB-SF as the destination IP address in the outer header.

8.3. Failure scenarios

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

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

Now assume that DCB-SF1 crashes. At that point the border router in Denver should stop tunnelling the traffic through DCB-SF1, and should switch to DCB-SF2. A way to accomplish this is to make each DCB to originate /32 route for its own IP address that it would advertise in

the NHRP Replies. This way when DCB-SF1 crashes, the route to DCB-SF1 IP address goes away, providing indication to the border router in Denver that it no longer can use DCB-SF1. At that point the border router in Denver removes /32 route for VM-A from its FIB (but will still retain a route for X). From that moment the border router in Denver will start forwarding packets destined to VM-A using the route to the subnet X. Since DCB-SF1 crashes, these packets will be routed to DCB-SF2, as DCB-SF2 advertises a route to X.

However, once the border router in Denver detects that DCB-SF1 is down, the border router will issue another NHRP Request. This time, NHRP Request reaches DCB-SF2, and DCB-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 DCB-SF2 as the IP destination address. In effect that means that the border router in Denver will install in its FIB a /32 route for VM-A indicating GRE encapsulation with DCB-SF2 as the destination IP address in the outer header.

9. Acknowledgements

We would like to thank Dave Katz for reviewing the NHRP procedures.

10. References

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

[MAC-VPN] "BGP/MPLS Based Ethernet VPN", draft-raggarwa-sajassi-l2vpn-evpn-01.txt, R. Aggarwal et al.

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

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
e-mail: wim.henderickx@alcatel-lucent.com

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