

NV03 Working Group  
Internet Draft  
Intended status: Standards track  
Expires: April 2015

Y. Rekhter  
Juniper Networks  
L. Dunbar  
Huawei  
R. Aggarwal  
Arktan Inc  
R. Shekhar  
Juniper Networks  
W. Henderickx  
Alcatel-Lucent  
L. Fang  
Microsoft  
A. Sajassi  
Cisco

October 3, 2014

NV03 VM Mobility Scheme  
draft-merged-nvo3-vm-mobility-scheme-00.txt

Status of this Memo

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

This Internet-Draft is submitted in full conformance with the provisions of [BCP 78](#) and [BCP 79](#). This document may not be modified, and derivative works of it may not be created, except to publish it as an RFC and to translate it into languages other than English.

Internet-Drafts are working documents of the Internet Engineering Task Force (IETF), its areas, and its working groups. Note that other groups may also distribute working documents as Internet-Drafts.

Internet-Drafts are draft documents valid for a maximum of six months and may be updated, replaced, or obsoleted by other documents at any time. It is inappropriate to use Internet-Drafts as reference material or to cite them other than as "work in progress."

The list of current Internet-Drafts can be accessed at <http://www.ietf.org/ietf/lid-abstracts.txt>

The list of Internet-Draft Shadow Directories can be accessed at <http://www.ietf.org/shadow.html>

This Internet-Draft will expire on April 3, 2009.

## Copyright Notice

Copyright (c) 2014 IETF Trust and the persons identified as the document authors. All rights reserved.

This document is subject to [BCP 78](#) and the IETF Trust's Legal Provisions Relating to IETF Documents (<http://trustee.ietf.org/license-info>) in effect on the date of publication of this document. Please review these documents carefully, as they describe your rights and restrictions with respect to this document. Code Components extracted from this document must include Simplified BSD License text as described in Section 4.e of the [Trust Legal Provisions](#) and are provided without warranty as described in the Simplified BSD License.

## Abstract

This document describes the schemes to overcome the network-related issues to achieve seamless Virtual Machine mobility in the data center and between data centers.

## Table of Contents

<a href="#">1.</a>	<a href="#">Introduction.....</a>	<a href="#">3</a>
<a href="#">2.</a>	<a href="#">Conventions used in this document.....</a>	<a href="#">4</a>
<a href="#">3.</a>	<a href="#">Terminology.....</a>	<a href="#">4</a>
<a href="#">4.</a>	<a href="#">Scheme to resolve VLAN-IDs usage in L2 access domains.....</a>	<a href="#">8</a>
<a href="#">5.</a>	<a href="#">Layer 2 Extension.....</a>	<a href="#">10</a>
<a href="#">5.1.</a>	<a href="#">Layer 2 Extension Problem.....</a>	<a href="#">10</a>
<a href="#">5.2.</a>	<a href="#">NVA based Layer 2 Extension Solution.....</a>	<a href="#">10</a>
<a href="#">5.3.</a>	<a href="#">E-VPN based Layer 2 Extension Solution.....</a>	<a href="#">10</a>
<a href="#">6.</a>	<a href="#">Optimal IP Routing.....</a>	<a href="#">14</a>
<a href="#">6.1.</a>	<a href="#">Preserving Policies.....</a>	<a href="#">15</a>
<a href="#">6.2.</a>	<a href="#">VM Default Gateway solutions.....</a>	<a href="#">16</a>
<a href="#">6.2.1.</a>	<a href="#">E-VPN based VM Default Gateway Solutions.....</a>	<a href="#">16</a>
<a href="#">6.2.1.1.</a>	<a href="#">E-VPN based VM Default Gateway Solution 1.....</a>	<a href="#">17</a>

6.2.1.2. E-VPN based VM Default Gateway Solution 2.....	18
6.2.2. Distributed Proxy Default Gateway Solution.....	18
6.3. Triangular Routing.....	19
6.3.1. NVA based Intra Data Center Triangular Routing Solution .....	19
6.3.2. E-VPN based Intra Data Center Triangular Routing Solution.....	20
7. Manageability Considerations.....	21
8. Security Considerations.....	21
9. IANA Considerations.....	22
10. Acknowledgements.....	22
11. References.....	22
11.1. Normative References.....	22
11.2. Informative References.....	22

## 1. Introduction

An important feature of data centers identified in [[nvo3-problem](#)] is the support of Virtual Machine (VM) mobility within the data center and between data centers. This document describes the schemes to overcome the network-related issues to achieve seamless Virtual Machine mobility in the data center and between data centers, where seamless mobility is defined as the ability to move a VM from one server in a 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.

Note that in the scenario where a VM is moved between servers located in different data centers, there are certain issues related to the current state of the art of the Virtual Machine technology, the bandwidth that may be available between the data centers, the distance between the data centers, the ability to manage and operate such VM mobility, storage-related issues (the moved VM has to have access to the same virtual disk), etc. Discussion of these issues is outside the scope of this document.

## 2. 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](#)].

In this document, these words will appear with that interpretation only when in ALL CAPS. Lower case uses of these words are not to be interpreted as carrying [RFC-2119](#) significance.

DC: Data Center

DCBR: Data Center Bridge Router

LAG: Link Aggregation Group

POD: Modular Performance Optimized Data Center. POD and Data Center are used interchangeably in this document.

ToR: Top of Rack switch

VEPA: Virtual Ethernet Port Aggregator (IEEE802.1Qbg)

VN: Virtual Network

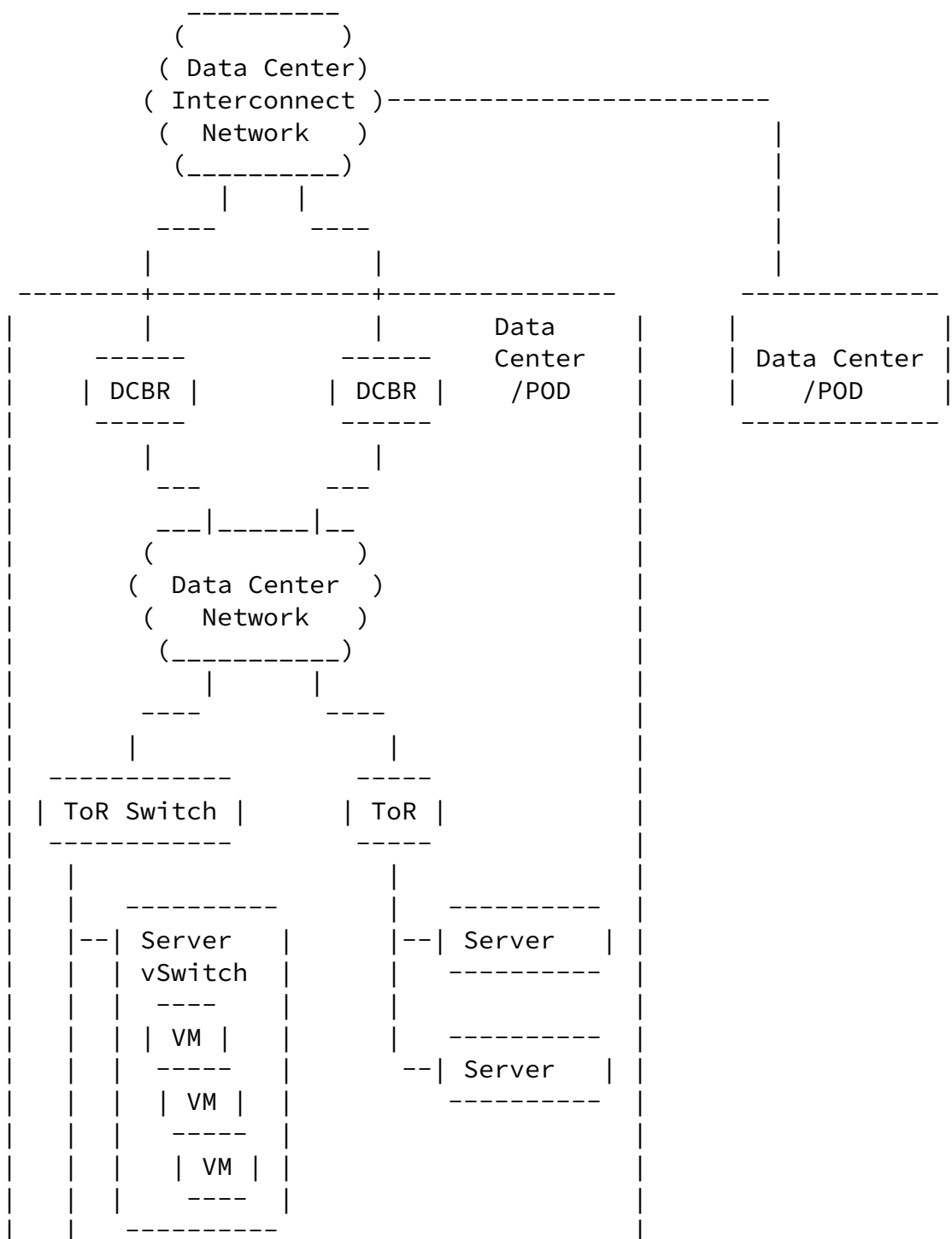
## 3. Terminology

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. Some servers may have embedded blade switches, some servers may have virtual switches to interconnect the VMs, and some servers may not have any embedded switches. When External Bridge Port Extenders (as defined by 802.1BR) are used to connect the servers to the data center network, the ToR switch is the Controlling Bridge.

Several data centers or PODs could be connected by a network. In addition to providing interconnect among the data centers/PODs, 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 hosted in other data centers, and (b) connectivity between VMs hosted in the data center and hosts communicating with these VMs.

The following figure illustrates the above:



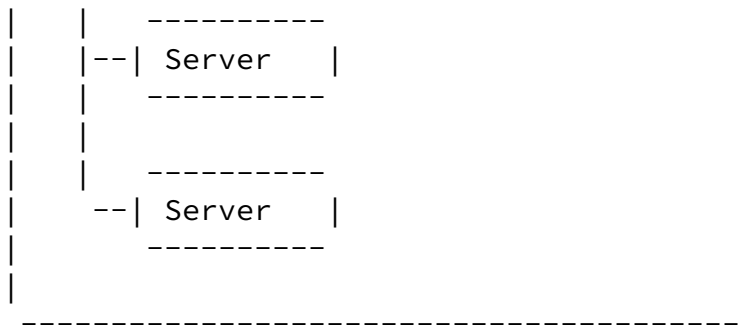


Figure 1: A Typical Data Center Network

The data centers/PODs 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 Virtual Network (VN) or Closed User Group (L2-based CUG). In this document, the Closed User Group and Virtual Network (VN) are used interchangeably.

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

In terms of IP address assignment this document assumes that all VMs of a given L2-based VN 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 VN. If a given VM is a member of more than one L2-based VN, this VM would have multiple IP addresses and multiple logical interfaces, one IP address and one logical interface per each such VN.

A VM that is a member of a given L2-based VN may (as a matter of

policy) be allowed to communicate with VMs that belong to other L2-based VNs, 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 physical domain" refers to a collection of interconnected devices that perform forwarding based on the information carried in the Ethernet header. A trivial L2 physical domain consists of just one non-virtualized server. In a non-trivial L2 physical domain (domain that contains multiple forwarding entities) forwarding could be provided by such layer 2 technologies as Spanning Tree Protocol (STP), VEPA (IEEE802.1Qbg), etc. Note that any multi-chassis LAG cannot span more than one L2 physical domain. This document assumes that a layer 2 access domain is an L2 physical domain.

A physical server connected to a given L2 physical domain may host VMs that belong to different L2-based VNs (while each of these VNs may span multiple L2 physical domains). If an L2 physical domain contains servers that host VMs belonging to different L2-based VNs, then enforcing L2-based VNs boundaries among these VMs within that domain is accomplished by relying on Layer 2 mechanisms (e.g. VLANs).

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

We say that a given L2-based VN is present within a given data center if one or more VMs that are part of that VN are presently hosted by the servers located in that data center.

In the context of this document when we talk about VLAN-ID used by a given VM, we refer to the VLAN-ID carried by the traffic that is within the same L2 physical domain as the VM, and that is either originated or destined to that VM - e.g., VLAN-ID only has local significance within the L2 physical domain, unless it is stated otherwise.

Some of the VM-mobility solutions described in this document are E-VPN based. When using E-VPN in NV03 environment, the NVE function is on the PE node. NVE-PE is used to describe the E-VPN PE node that supports the NVE function.

#### 4. Scheme to resolve VLAN-IDs usage in L2 access domains

To support tens of thousands of virtual networks, the local VID associated with client payload under each NVE has to be locally significant. Therefore, the same L2-based VN MAY have either the same or different VLAN-IDs under different NVEs. Thus when a given VM moves from one non-trivial L2 physical domain to another, the VLAN-ID of the traffic from/to VM in the former may be different than in the latter, and thus cannot assume to stay the same.

For data frames traverse through the NV03 underlay network, if ingress NVE simply encapsulates an outer header to data frames received from VMs and forward the encapsulated data frames to egress NVE via underlay network, the egress NVE can't simply decapsulate the outer header and send the decapsulated data frames to the attached VMs as done by TRILL.

It is possible that within a trivial L2 physical domain traffic from/to VMs that are in this domain may not have VLAN-IDs at all.

If a given VM's Guest OS sends packets that carry VLAN-ID, then the VLAN-ID used by the Guest OS may not change when the VM moves from one L2 physical domain to another (this is irrespective of whether L2 physical domains are trivial or non-trivial). In other words, the VLAN-IDs used by a tagged VM network interface are part of the VM's state and may not be changed when the VM moves from one L2 physical domain to another. Therefore, it is necessary for an entity, most likely the first switch (virtual or physical) to which the VM is attached, to change the VLAN-ID from the value used by NVE to the value expected by the VM (in contrast, a VLAN tag assigned by a hypervisor for use with an untagged VM network interface can change). If the L2 physical domain is extended to include VM tagged interfaces, the hypervisor virtual switch, and the DC bridged network, then special consideration described below is needed in



assignment of VLAN tags for the VMs, the L2 physical domain and other domains into which the VM may move.

This document assumes that within a given non-trivial L2 physical domain traffic from/to VMs that are in that domain, and belong to different L2-based VNs MUST have different VLAN-IDs.

The above assumptions about VLAN-IDs are driven by (a) the assumption that within a given L2 physical domain VLANs are used to identify individual L2-based VNs, and (b) the need to overcome the limitation on the number of different VLAN-IDs.

NVA can facilitate NVE for local VID assignment and dynamic mapping between local VID and global virtual network instances. NVE needs to free up VLAN-IDs when there is no VMs underneath under the VLAN-IDs. Here is the detailed procedure:

- . NVE should get the specific VNID from NVA for untagged data frames arriving at the each Virtual Access Point [VNo3-framework 3.1.1] of a NVE.

Since local VLAN-IDs under each NVE are locally significant, ingress NVE should remove the local VLAN-ID attached to the data frame. So that egress NVE can always assign its own local VLAN-ID to data frame before sending the decapsulated data frame to the attached VMs.

If, for whatever reasons, it is necessary to have local VLAN-ID in the data frames before encapsulating outer header (i.e. EgressNVE-DA, IngressNVE-SA, VNID), NVE should get the specific local VLAN-ID from the NVA for those untagged data frames coming to each Virtual Access Point.

- . If the data frame is already tagged before reaching the NVE's Virtual Access Point, the NVA can inform the first switch port that is responsible for adding VLAN-ID to the untagged data frames of the specific VLAN-ID to be inserted to data frames.
- . If data frames from VMs are already tagged, the first port facing the VMs has be informed by the NVA of the new local VLAN-ID to replace the VLAN-ID encoded in the data frames.

For data frames coming from network side towards VMs (i.e. inbound traffic towards VMs), the first switching port facing

VMs have to convert the VLAN-IDs encoded in the data frames to the VLAN-IDs used by VMs.

## 5. Layer 2 Extension

### 5.1. Layer 2 Extension Problem

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

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

Note that supporting VM mobility implies that the set of L2 physical domains that contain VMs that belong to a given L2-based VN may change over time (new domains added, old domains deleted).

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 VN, while the latter does not.

### 5.2. NVA based Layer 2 Extension Solution

Assume NV03's NVA has at least the following information for each TS (or VM):

- . Inner Address: TS (host) Address family (IPv4/IPv6, MAC, virtual network Identifier MPLS/VLAN, etc)
- . Outer Address: The list of locally attached edges (NVEs); normally one TS is attached to one edge, TS could also be

merged, et al.

Expires April 3, 2015

[Page 10]

---

Internet-Draft

NV03 Mobility Scheme

October 2014

attached to 2 edges for redundancy (dual homing). One TS is rarely attached to more than 2 edges, though it could be possible;

- . VN Context (VN ID and/or VN Name)
- . Timer for NVEs to keep the entry when pushed down to or pulled from NVEs.
- . Optionally the list of interested remote edges (NVEs). This information is for NVA to promptly update relevant edges (NVEs) when there is any change to this TS' attachment to edges (NVEs). However, this information doesn't have to be kept per TS. It can be kept per VN.

NVA can offer services in a Push, Pull mode, or the combination of the two.

In this solution, the NVEs are connected via underlay IP network. For each VN, the NVA informs all the NVEs to which the VMs of the given VN are attached.

When the last VM of a VN is moved out of a NVE, the NVA notifies the NVE for it to remove its connectivity to the VN. When a VM of a given VN is moved into a NVE for the first time (i.e. the NVE didn't have any VMs belonging to this VN yet), the NVA will notify the NVE for it to be connected to VN.

The term "NVE being connected to a VN" means that the NVE at least has:

- . the inner-outer address mapping information for all the VMs in the VN or being able to pull the mapping from the NVA,
- . the mapping of local VLAN-ID to the VNID used by overlay header, and
- . has the VN's default gateway IP/MAC address.

### 5.3. E-VPN based Layer 2 Extension Solution

This section describes a [[E-VPN](#)] based solution for the layer 2 extension problem, i.e. the L2 sites that contain VMs of a given L2-based VN are interconnected together using E-VPN. Thus a given E-VPN corresponds/associated with one or more L2-based VNs (e.g., VLANs). An L2-based VN 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 POD or a data center, in which case the DCBRs of that POD/data center, in addition to acting as IP routers for the L2-based VNs present in the POD/data center, also act as PEs. In this scenario E-VPN is used to handle VM migration between servers in different POD/data centers and the PE nodes support the NVE function.

A single L2 site could be as small as a single ToR with the servers connected to it or virtual switch with VMs attached, in which case the ToR or the virtual switch acts as a PE-NVE. 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 VNs 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 PE-NVEs.

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 VN is constrained by the scope of the PEs connected to the L2 sites that presently contain VMs of that VN. 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 VN is optimally forwarded, irrespective of whether members of that VN are within the same or in different data centers/PODs. 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 VN 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 VN spans multiple data centers. For example when a given ToR acts as a PE-NVE, 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 VNs 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 PE-NVEs within its own data center, but just advertises a single "default" MAC advertisement route to these PE-NVEs.

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 VN, then the PE-NVE(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 VN. Likewise, if after the move the old site no longer has any VMs that are in the same L2-based VN as the VM that moved, the PE-NVE(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.

## 6. 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 VN. 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 VN, and destined to that VM be routed via the router of the VM's L2-based VN 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 VN (the router of the VM's L2-based VN may be either DCBR or ToR itself). This is also known as avoiding "triangular routing". This document refers to this problem as the triangular routing problem.

In order to avoid the "triangular routing", routers in the Wide Area Network have to be aware which DCBRs can reach the designated VMs. When VMs in a single VN are spread across many different DCBRs, all individual VMs' addresses have to be visible to those routers, which can dramatically increase the number of routes in those routers.

If a VN is spread across multiple DCBRs and all those DCBRs announce the same IP prefix for the VN, there could be many issues,

including:

- Traffic could go to DCBR A where target is in DCBR B. and DCBR "A" is connected to DCBR "B" via WAN

- If majority of one VN members are under DCBR "A" and rest are spread across X number of DCBRs. Will DCBR "A" have same weight as DCBR "B", "C", etc?

If all those DCBRs announce individual IPs that are directly attached and those IPs are not segmented well, then all the VMs IP addresses have to be exposed to the WAN. So overlay hides the VMs IP from the core switches in one DC or one POD, but exposes them to the WAN. There are more routers in the WAN than the number of core switches in one DC/POD.

The ability to deliver optimal routing (as defined above) in the presence of stateful devices is outside the scope of this document.

### 6.1. Preserving Policies

Moving VM from one L2 physical domain to another means (among other things) that the NVE in the new domain that provides connectivity between this VM and VMs in other L2 physical domains must be able to implement the policies that control connectivity between this VM and VMs in other L2 physical domains. In other words, the policies that control connectivity between a given VM and its peers MUST NOT change as the VM moves from one L2 physical domain to another. Moreover, policies, if any, within the L2 physical domain that contains a given VM MUST NOT preclude realization of the policies that control connectivity between this VM and its peers. All of the above is irrespective of whether the L2 physical domains are trivial or not.

There could be policies guarding VMs across different VNs, with some being enforced by Firewall, some enforced by NAT/AntiDDOS/IPS/IDS, etc. It is less about NVE policies to be maintained when VMs move, it is more along the line of dynamically changing policies associated with the "middleware" boxes attached to NVEs (if those middle boxes are distributed).

## 6.2. VM Default Gateway solutions

As VM moves to a new L2 site, the default gateway IP address of the VM may not change. Further, while with cold VM mobility one may assume that VM's ARP/ND cache gets flushed once VM moves to another server, one cannot make such an assumption with hot VM mobility.

Thus the destination MAC address in the inter-VN/inter-subnet traffic originated by that VM would not change as VM moves to the new site. Given that, how would NVE(s) connected to the new L2 site be able to recognize inter-VN/inter-subnet traffic originated by that VM? The following describes possible solutions.

### 6.2.1. E-VPN based VM Default Gateway Solutions

The E-VPN based solutions assume that for inter-VN/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-VNs (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-VN) 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.



#### [6.2.1.1](#). E-VPN based VM Default Gateway 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 PE-NVEs for an E-VPN corresponding to a given L2-based VN, then these anycast addresses are configured on these DCBRs. Likewise, if ToRs act as PE-NVEs, then these anycast addresses are configured on these ToRs. All VMs of that L2-based VN are (auto) configured with the (anycast) IP address of the default gateway.

DCBRs (or ToRs) acting as PE-NVEs use these anycast addresses as follows:

- When a particular DCBR (or ToR) acting as a PE-NVE receives a packet with the (anycast) default gateway MAC address, the DCBR (or ToR) applies IP forwarding to the packet, and perform NVE function if the destination of the packet is attached to another NVE.
- When a particular DCBR (or ToR) acting as a PE-NVE receives an ARP/ND Request for the default gateway (anycast) IP address, the DCBR (or ToR) generates ARP/ND Reply.

This ensures that a particular DCBR (or ToR), acting as a PE-NVE, 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 PE-NVEs must never use the anycast default gateway MAC address as the source MAC address in the packets originated by these DCBRs (or ToRs), cannot use the anycast default gateway IP address as the source IP address in the overlay header.

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

If the default gateway functionality is not in NVEs (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 VN, all these routes MUST indicate that this MAC address belongs to the same Ethernet Segment Identifier (ESI).

#### [6.2.1.2](#). E-VPN based VM Default Gateway Solution 2

The second solution does not require configuring the anycast default gateway IP and MAC address on the PE-NVEs.

Each DCBR (or each ToR) that acts as a default gateway for a given L2-based VN 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-NVE 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-NVE that receives this E-VPN route follows procedures in Section 12 of [\[E-VPN\]](#) when replying to ARP/ND Requests that it receives if such Requests are for the IP address in the received E-VPN route.

#### 6.2.2. Distributed Proxy Default Gateway Solution

In this solution, NVEs perform the function of the default gateway for all the VMs attached. Those NVEs are called "Proxy Default Gateway" in this document because those NVEs might not be the Default Gateways explicitly configured on VMs attaches. Some of those proxy default gateway NVEs might not have the complete inter-subnet communications policies for the attached VNs.

In order to ensure that the destination MAC address in the inter-VN/inter-subnet traffic originated by that VM would not change as VM moves to a different NVE, a pseudo MAC address is assigned to all NVE-based Proxy Default Gateways.

When a particular NVE acting as Proxy Default Gateway receives an ARP/ND Request from the attached VMs for their default gateway IP addresses, the NVE generates ARP/ND Reply with the pseudo MAC address.

When a particular NVE acting as a Proxy Default Gateway receives a packet with the Pseudo default gateway MAC address:

- if the NVE has all the needed policies for the Source & Destination VNs, the NVE applies the IP forwarding, i.e. forward the packet from source VN to the destination VN, and apply the NVE encapsulation function with target NVE as destination address and destination VN identifier in the header,
- if the NVE doesn't have the needed policies from the source VN to the destination VN, the NVE applies the NVE encapsulation function with real host's default gateway as destination address and source VN identifier in the header

This solution assumes that the NVE-based proxy default gateways either get the mapping of hosts' default gateway IP <-> default gateway MAC from the corresponding NVA or via ARP/ND discovery.

### 6.3. Triangular Routing

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. This draft only describes the solution for intra data center triangular routing.

#### 6.3.1. NVA based Intra Data Center Triangular Routing Solution

To be added.

### 6.3.2. E-VPN based Intra Data Center Triangular Routing Solution

This solutions assumes that as a PE-NVE 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-NVE can learn the IP address associated with a given MAC address are specified in [[E-VPN](#)].

Consider a set of L2-based VNs, such that VMs of these VNs, 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 PE-NVEs, and also able to support IP forwarding functionality.

For a given set of L2-based VNs whose VMs are allowed to communicate with each other, consider a set of E-VPN instances (EVI) of the E-VPNs associated with these VNs. 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

- for each set of L2-based VNs whose VMs are allowed to communicate with each other, and for each data center that contains such VNs have a distinct RT (distinct RT per set, per data center),
- provision each EVI of the E-VPNs associated with these VNs to import routes that carry this RT, and
- 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 VNs [[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 VN

associated with the E-VPN of that EVI, the PE-NVE 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-NVE, ToR-1, and another VM, VM-B, is currently hosted by a server connected to ToR-2 (NVE). Assume that VM-A and VM-B belong to different L2-based VNs, and (as a matter of policy) VMs in these VNs are allowed to communicate with each other. Now assume that VM-B moves to another server, and this server is connected to ToR-3 (NVE). 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 VNs, 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 PE-NVEs within that data center – the PE-NVEs that have EVIs of the E-VPNs associated with the L2-based VNs 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 PE-NVEs.

## 7. Manageability Considerations

Several solutions described in this document depend on the presence of NVA in the data center.

## 8. Security Considerations

In addition to the security considerations described in [nvo3-problem], it is clear that allowing VMs migrating across Data Center will require more stringent security enforcement. The traditional placement of security functions, e.g. firewall, at data center

Internet-Draft

NV03 Mobility Scheme

October 2014

gateways is no longer enough. VM mobility will require security functions to enforce policies among east-west traffic among VMs. When VMs move across Data Center, the associated policies have to be updated and enforced.

## 9. IANA Considerations

This document requires no IANA actions. RFC Editor: Please remove this section before publication.

## 10. Acknowledgements

The authors would like to thank Adrian Farrel, David Black and Larry Kreeger their review and comments. The authors would also like to thank Ivan Pepelnj his contributions to this document.

## 11. References

### 11.1. Normative References

- [RFC2119] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", [BCP 14](#), [RFC 2119](#), March 1997.
- [RFC7297] Boucadair, M., "IP Connectivity Provisioning Profile", [RFC7297](#), April 2014.

### 11.2. Informative References

- [nvo3-problem] Narten T.et al., "Overlays for Network Virtualization", [draft-ietf-nvo3-overlay-problem-statement-04](#), July 2013.
- [RFC1700] Reynolds J., Postel J., "ASSIGNED NUMBERS", [RFC1700](#), October 1994
- [RFC2332] "NBMA Next Hop Resolution Protocol (NHRP)", [RFC 2332](#), J. Luciani et. al.

---

Internet-Draft

NV03 Mobility Scheme

October 2014

[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>

Internet-Draft

NV03 Mobility Scheme

October 2014

#### Authors' Addresses

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

Linda Dunbar  
Huawei Technologies  
5340 Legacy Drive, Suite 175  
Plano, TX 75024, USA  
Email: ldunbar@huawei.com

Rahul Aggarwal  
Arktan, Inc  
Email: raggarwa\_1@yahoo.com

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@microsoft.com

Ali Sajassi  
Cisco Systems  
Email: sajassi@cisco.com



merged, et al.

Expires April 3, 2015

[Page 24]