

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 24, 2014

Overlay Network Tenant System Address Migration
draft-merged-nvo3-ts-address-migration-01.txt

Status of this Memo

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 24, 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 data centers.

Table of Contents

1.	Introduction.....	3
2.	Conventions used in this document.....	3
3.	Terminology.....	4
4.	Scheme to resolve VLAN-IDs usage in L2 access domains.....	7
5.	Layer 2 Extension.....	9
	5.1. Layer 2 Extension Problem.....	9
	5.2. NVA based Layer 2 Extension Solution.....	10
6.	Optimal IP Routing.....	11
	6.1. Preserving Policies.....	13
	6.2. TS Default Gateway solutions.....	13
	6.2.1. Solution with Anycast for TS Default Gateways.....	13
	6.2.2. Distributed Proxy Default Gateway Solution.....	15
	6.3. Triangular Routing.....	16
7.	L3 Address Migration.....	16
8.	Managing duplicated addresses.....	18
9.	Manageability Considerations.....	18
10.	Security Considerations.....	18
11.	IANA Considerations.....	19

12.	Acknowledgements.....	19
13.	References.....	19
13.1.	Normative References.....	19
13.2.	Informative References.....	19

1. Introduction

An important feature of data centers identified in [[nvo3-problem](#)] is the support of Virtual Machine (TS) 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 TS 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 TS. In the context of this document the term mobility or a reference to moving a TS should be considered to imply seamless mobility, unless otherwise stated.

Note that in the scenario where a TS 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 TS mobility, storage-related issues (the moved TS 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

TS: Tenant System (used interchangeably with VM on servers supporting Virtual Machines)

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

VN: Virtual Network

3. Terminology

In this document "Mobility" refers to "address migration", meaning TSs move to different locations without changing their addresses (IP/MAC).

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 TSs. A data center may have multiple ToRs. Some servers may have embedded blade switches, some servers may have virtual switches to interconnect the TSs, 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 TSs hosted in these data centers and the sites that contain hosts communicating with such TSs. 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 TSs hosted in the data center and TSs hosted in other data centers, and (b) connectivity between TSs hosted in the data center and hosts communicating with these TSs.

The following figure illustrates the above:

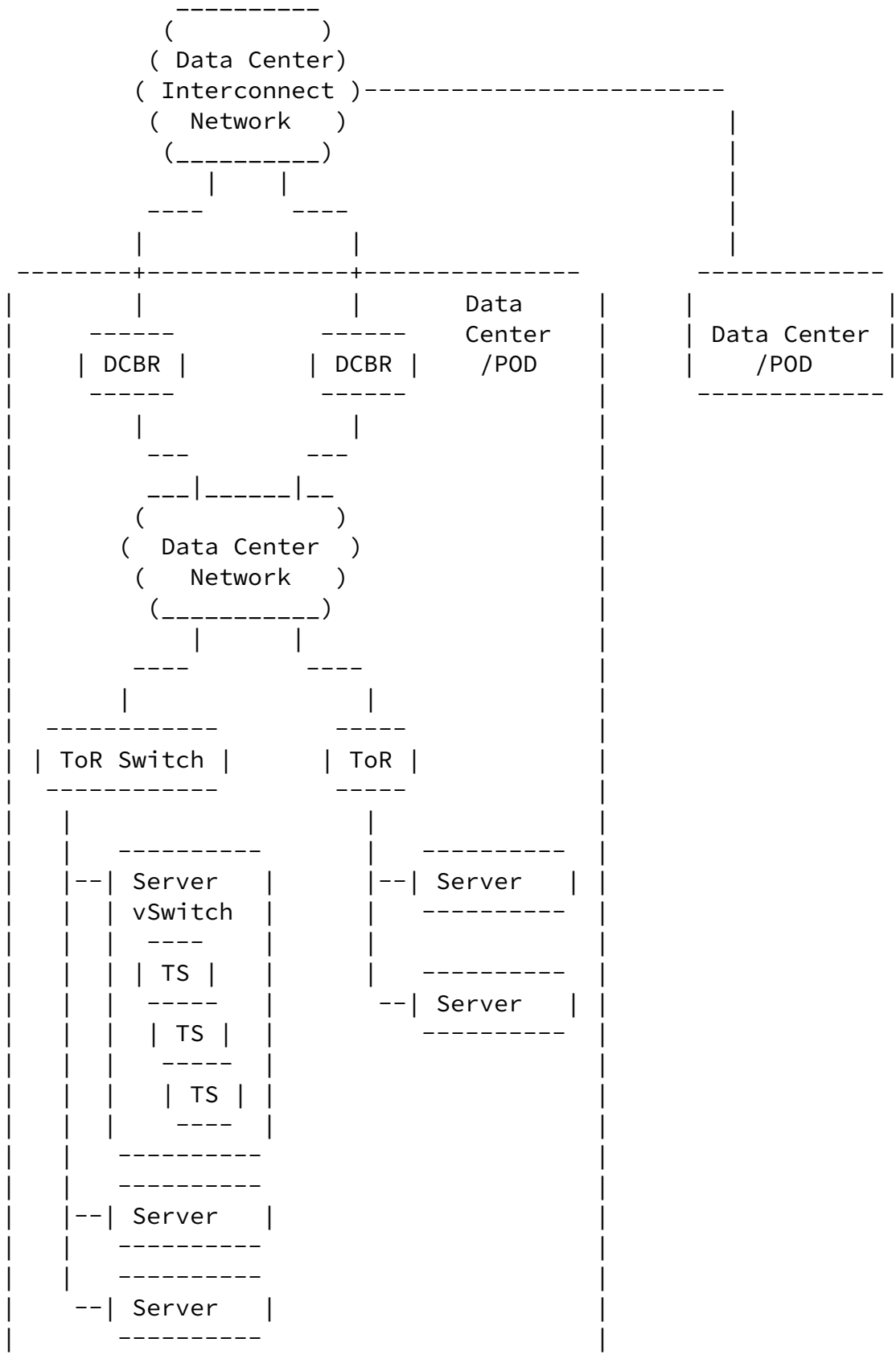


Figure 1: A Typical Data Center Network

merged, et al.

Expires April 24, 2015

[Page 5]

Internet-Draft

NV03 Mobility Scheme

October 2014

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 TSs that (as a matter of policy) are allowed to communicate with each other, and a collection of devices that interconnect these TSs. If communication among any TSs 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 TSs (as these packets traverse from their sources to their destinations), we will refer to such set of TSs 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 TS may be a member of more than one VN or L2-based VN.

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

A TS that is a member of a given L2-based VN may (as a matter of policy) be allowed to communicate with TSs 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 attachment" refers to a collection of interconnected devices attached to an NVE that perform forwarding based on the information carried in the Ethernet header. A trivial L2 physical attachment consists of just one non-

virtualized server. In a non-trivial L2 physical attachment (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 attachment. This document assumes that a layer 2 access domain is an L2 physical attachment.

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

We say that an L2 physical attachment contains a given TS (or that a given TS is in a given L2 physical attachment), if the server presently hosting this TS 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 TSs 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 TS, we refer to the VLAN-ID carried by the traffic that is within the same L2 physical attachment as the TS, and that is either originated or destined to that TS - e.g., VLAN-ID only has local significance within the L2 physical attachment, unless it is stated otherwise.

Some of the TS-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

This document assumes that within a given non-trivial L2 physical attachment traffic from/to TSs belonging to different L2-based VNs

MUST have different VLAN-IDs.

To support tens of thousands of virtual networks, the local VLAN-ID 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 TS moves from one non-trivial L2 physical attachment to another, the VLAN-ID of the traffic from/to TS in the former may be different than in the latter, and thus cannot assume to stay the same.

To describe the solution more clearly, here are the terminologies used:

- Customer administered VLAN-IDs (usually hard coded in a TS's Guest OS and can't be changed when the TS move from one NVE to another). Some TSs may not have VLAN-ID attached.
- Provider administered VLAN-IDs of local significance, and
- Provider administered VN-IDs of global significance.

In the scenario where there are provider administered VLAN-IDs of local significance (e.g. NVE in a TOR), the value is selected by NVA from the pool of unused VIDs when the first local TS of a VN is being added, and returned by NVA to the unused pool of VLAN-IDs when the last TS leaves. For TSs with hard coded VLAN-ID, it is necessary for an entity, most likely the first switch (virtual or physical) to which the TS is attached, to change the locally administered VLAN-IDs to the TSs' hard coded VLAN-IDs. For un-tagged TSs, the first switch has to remove the locally administered VLAN-IDs before sending packets to TSs.

The section is intended to describe:

- . NVA manages unused VLAN-IDs pool in each access L2 domain
- . NVE reports to NVA when first local TS of a VN is reachable, or none of TS in a VN is reachable by the NVE
- . NVA can push the global VN ID <-> locally administered VID mapping to NVE, or NVE can pull upon detecting a newly attached VN.
- . NVA manages the first switch to which TS is attached on mapping between TS's own VLAN-ID and "locally administered VID".

Here is the detailed procedure:

Internet-Draft

NV03 Mobility Scheme

October 2014

- . 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, here are the possible ways for ingress NVE to assign VLAN-ID in the overlay header for data frames destined to other NVEs:

- a) carry what comes in at ingress Virtual Access point. Preserving vlan-id can be used to provide bundled service/PVLAN. In this case many vlan-ids in ingress could map to one logical VN (n to 1 mapping).
- b) not carrying any vlan-id and using logical VN identifier. The egress NVE gets the vlan-id from NVA to put on the packet before sending to attached TSs. This is 1-to-1 mapping between vlan-id and logical-VN.
- . If the data frame is already tagged before reaching the NVE's Virtual Access Point, the NVA should 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 a TS are already tagged, the first port facing the TS 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 TSs (i.e. inbound traffic towards TSs), the first switching port facing TSs have to convert the VLAN-IDs encoded in the data frames to the VLAN-IDs used by TSs.

5. Layer 2 Extension

5.1. Layer 2 Extension Problem

Consider a scenario where a TS 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 attachments, where these domains may be located in the same or different data centers (or PODs). In order to enable communication between this TS and other TSs of that L2-based

VN, the new L2 physical attachment must become interconnected with the other L2 physical attachment(s) that presently contain the rest

of the TSs 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 TS and other members of that VN.

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

Note that supporting TS mobility implies that the set of L2 physical attachments that contain TSs 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 TS mobility, as the former restricts communicating TSs 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:

- . 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 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 TSs of the given VN are attached.

When the last TS of a VN is moved out of a NVE, NVE can either confirm with the NVA or the NVA notifies the NVE for it to remove its connectivity to the VN. When an NVE needs to support connectivity to a VN not currently supported (as a result of TS turn up, or TS migration), the NVA will push the necessary VN information into the NVE.

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

- . the inner-outer address mapping information for all the TSs 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.

6. Optimal IP Routing

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

- Optimal routing of a TS's outbound traffic. This means that as a given TS moves from one server to another, the TS's default gateway should be in a close topological proximity to the ToR that connects the server presently hosting that TS. Note that when we talk about optimal routing of the TS's outbound traffic, we mean

traffic from that TS to the destinations that are outside of the TS's L2-based VN. This document refers to this problem as the TS default gateway problem.

- Optimal routing of TS's inbound traffic. This means that as a given TS moves from one server to another, the (inbound) traffic originated outside of the TS's L2-based VN, and destined to that TS be routed via the router of the TS's L2-based VN that is in a close topological proximity to the ToR that connects the server presently hosting that TS, without first traversing some other router of that L2-based VN (the router of the TS'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 TSs. When TSs in a single VN are spread across many different DCBRs, all individual TSs' 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 TSs IP addresses have to be exposed to the WAN. So overlay hides the TSs 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 TS from one L2 physical attachment to another means (among other things) that the NVE in the new domain that provides connectivity between this TS and TSs in other L2 physical attachments must be able to implement the policies that control connectivity between this TS and TSs in other L2 physical attachments. In other words, the policies that control connectivity between a given TS and its peers MUST NOT change as the TS moves from one L2 physical attachment to another. Moreover, policies, if any, within the L2 physical attachment that contains a given TS MUST NOT preclude realization of the policies that control connectivity between this TS and its peers. All of the above is irrespective of whether the L2 physical attachments are trivial or not.

There could be policies guarding TSs 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 TSs 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. TS Default Gateway solutions

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

Thus the destination MAC address in the inter-VN/inter-subnet traffic originated by that TS would not change as TS 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 TS? The following describes possible solutions.

6.2.1. Solution with Anycast for TS Default Gateways

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

If DCBRs act as default gateway to a given L2-based VN, then these anycast addresses are configured on these DCBRs. Likewise, if ToRs act as default gateways, then these anycast addresses are configured on these ToRs. All TSs of that L2-based VN are (auto) configured with the (anycast) IP address of the default gateway.

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

- When a particular NVE receives a packet from local L2 attachment with the (anycast) default gateway MAC address, the NVE 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 default gateway receives an ARP/ND Request from local L2 attachment 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 default gateway, can always apply IP forwarding to the packets sent by a TS to the (anycast) default gateway MAC address. It also ensures that such DCBR (or ToR) can respond to the ARP Request generated by a TS for the default gateway (anycast) IP address.

Except for gratuitous ARP/ND, DCBRs (or ToRs) acting as default gateway 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 to all NVEs.

6.2.2. Distributed Proxy Default Gateway Solution

This solution does not require configuring the anycast default gateway IP and MAC address for TSs.

In this solution, NVEs perform the function of the default gateway for all the TSs attached. Those NVEs are called "Proxy Default Gateway" in this document because those NVEs might not be the Default Gateways explicitly configured on TSs 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 TS would not change as TS 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 TSs for their default gateway IP addresses, the NVE suppresses the ARP/ND request from being forwarded and 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 TSs are in the same data center. The latter handles all other cases. This draft only describes the solution for intra data center triangular routing.

To avoid triangular routing, each NVE needs to have the egress NVEs for potential designations of packets originated from the attached TSs.

One approach is for each NVE to announce its directly attached TSs addresses to all other NVEs that participate in the VNs of the TSs'

Another approach is for NVA to distribute the VN scoped TS Address <-> NVE mappings to all the NVEs. See [Section 7](#) for the detailed mechanism.

7. L3 Address Migration

When the attachment to NVE is L3 based, TS migration can cause one subnetwork to be scatted among many NVEs, or fragmented addresses.

The outbound traffic of fragmented L3 addresses doesn't have the same issue as L2 address migration, but the inbound traffic has the same issues as L2 address migration ([Section 6](#)).

Optimal routing of TS's inbound traffic: This means that as a given TS moves from one server to another, the (inbound) traffic originated outside of the TS's directly attached NVE, and destined to that TS be routed optimally to the NVE to which the server presently hosting that TS, without first traversing some other NVEs. This is also known as avoiding "triangular routing".

In theory, host hosting by every NVE (including the NVEs attached to DCBR) can achieve the optimal inbound forwarding in very fragmented network. When TSs' IP addresses under all the NVEs can't be aggregated at all, a NVE needs to support the combined number of TSs

of all the VNs enabled on the NVE. Here is the math showing that

host routing on server based NVE or ToR based NVE can be relatively easy to be supported even under the worst case scenario:

- . Suppose a NVE has TSs belonging to X number of VNs and suppose each VN has 200 hosts (spread among many NVEs), then the worst case scenario (or the maximum routes that NVE needs to have) is $200 \times X$.
- . For Server based NVE, the number of VNs enabled on the NVE has to be less than number of VMs instantiated on the server. The industry state of art virtualization technology allows maximum 100 VMs on one server. So the worst case scenario (or the maximum routes that NVE needs to have) is $100 \times 200 = 20,000$
- . For ToR based NVE, the number of TSs can be number of TSs per server * the number of servers attached to ToR (typical ToR has 48 downstream ports to servers). So the worst case scenario is $40 \times 100 \times 200 = 800,000$.

But host routing can be challenging on NVEs attached to Data Center Gateways. Those NVEs usually need to support all the VNs enabled in the data center. There could be hundreds of thousands of hosts/VMs, sometimes in millions, due to business demand and highly advanced server virtualization technologies.

For those data centers with millions of TSs, the following approach should be considered:

- . Some NVEs (e.g. ToR/Server based NVEs) support host route, and
- . Some NVEs (e.g. the NVEs attached to Data center gateways) that participate in large number of VNs (if not all VNs), support "non-host-route". Those NVEs are called "non-host-route" NVEs in the draft.

Those non-host-route NVEs have one or two egress NVEs as the designated forwarders for a VN (subnet) even if the VN (subnet) is spread across many NVEs. For example, if high percentage of TSs of one subnet is attached to NVE "X", the remaining small percentage of the subnet is spread around many NVEs. The non-host-route NVEs can have NVE "X" as the designated egress for the VN. By doing so, it can greatly reduce the "triangular routing" for the traffic destined

to TSs in this VN (subnet).

To avoid loops, the designated NVEs must support host route.

It worth noting that for the NVEs that have host route, they send traffic directly to the egress NVEs because they have the detailed information. Only for the NVEs (most likely the NVEs attached to the Gateway), they send traffic to the VN's (subnet) designated NVEs if they don't have host routes for the VN. The NVEs that prefer not to have host routes need to notify NVA that they only want designated NVEs, or can be configured in the NVA.

ECMP can be another approach that can be used by those non-host-route NVEs, when VNs are spread across many NVEs. The ECMP approach basically assigns all the NVEs that have the TSs of a VN attached as the "designated egress NVEs" for the VN. Again, to avoid loops, those designated egress NVEs have to support host route. ECMP approach may cause most packets from those non-host-route NVEs (it not all) to traverse two NVEs before reaching packets' destinations.

8. Managing duplicated addresses

This document assumes that during VM migration a given MAC address within a VN can only exist at one TS at a time. As TSs move around NVEs, it is possible that the network state may not be immediately synchronized. It is important for NVEs to report directly attached TSs to NVA on periodically bases so that NVA can generate alarms and fix duplicated address issues.

9. Manageability Considerations

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

10. Security Considerations

In addition to the security considerations described in [nvo3-problem], it is clear that allowing TSs migrating across Data Center will require more stringent security enforcement. The traditional placement of security functions, e.g. firewall, at data center gateways is no longer enough. TS mobility will require security

functions to enforce policies among east-west traffic among TSs.

When TSs move across Data Center, the associated policies have to be updated and enforced.

11. IANA Considerations

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

12. Acknowledgements

The authors would like to thank Adrian Farrel, David Black, Dave Allen, Tom Herbert and Larry Kreeger for their review and comments. The authors would like to thank Ivan Pepelnjak for his contributions to this document.

13. References

13.1. Normative References

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

13.2. Informative References

[nvo3-problem] Narten T.et al., "Overlays for Network Virtualization", [draft-ietf-nvo3-overlay-problem-statement-04](#), July 2013.

[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

Internet-Draft

NV03 Mobility Scheme

October 2014

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

[DC-mobility] R. Aggarwal, et al, "Data Center Mobility based on E-VPN, BGP/MPLS IP VPN, IP Routing and NHRP", [draft-raggarwa-data-center-mobility-07](#), June 2014

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 24, 2015

[Page 21]