

Network Working Group
Internet Draft
Category: Standards Track
Expiration Date: April 2013

Y. Rekhter
Juniper Networks

W. Henderickx
Alcatel-Lucent

R. Shekhar
Juniper Networks

Luyuan Fang
Cisco Systems

Linda Dunbar
Huawei

Ali Sajassi
Cisco Systems

October 7 2012

Network-related VM Mobility Issues

[draft-rekhter-nvo3-vm-mobility-issues-03.txt](#)

Status of this Memo

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

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/1id-abstracts.txt>.

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

Internet Draftdraft-rekhter-nvo3-vm-mobility-issues-03.txt October 2012

Copyright and License Notice

Copyright (c) 2011 IETF Trust and the persons identified as the document authors. All rights reserved.

This document is subject to [BCP 78](http://trustee.ietf.org/license-info) 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.

This document may contain material from IETF Documents or IETF Contributions published or made publicly available before November 10, 2008. The person(s) controlling the copyright in some of this material may not have granted the IETF Trust the right to allow modifications of such material outside the IETF Standards Process. Without obtaining an adequate license from the person(s) controlling the copyright in such materials, this document may not be modified outside the IETF Standards Process, and derivative works of it may not be created outside the IETF Standards Process, except to format it for publication as an RFC or to translate it into languages other than English.

Abstract

This document describes a set of network-related issues presented by the desire to support seamless Virtual Machine mobility in the data center and between data centers. In particular, it looks at the implications of meeting the requirements for "seamless mobility".

Internet Draftdraft-rekhter-nvo3-vm-mobility-issues-03.txt October 2012

Table of Contents

1	Specification of requirements	3
2	Introduction	3
2.1	Terminology	4
3	Problem Statement	7
3.1	Usage of VLAN-IDs	7
3.2	Maintaining Connectivity in the Presence of VM Mobility ...	8
3.3	Layer 2 Extension	8
3.4	Optimal IP Routing	9
3.5	Preserving Policies	10
4	IANA Considerations	10
5	Security Considerations	10
6	Acknowledgements	10
7	References	10
8	Author's Address	11

[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

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 a set of network-related issues presented by the desire to support seamless Virtual Machine mobility in the data center, where seamless mobility is

defined as the ability to move a VM from one server in the data center to another server in the same or different data center, while retaining the IP and MAC address of the VM. In the context of this document the term mobility, or a reference to moving a VM should be considered to imply seamless mobility, unless otherwise stated.

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.

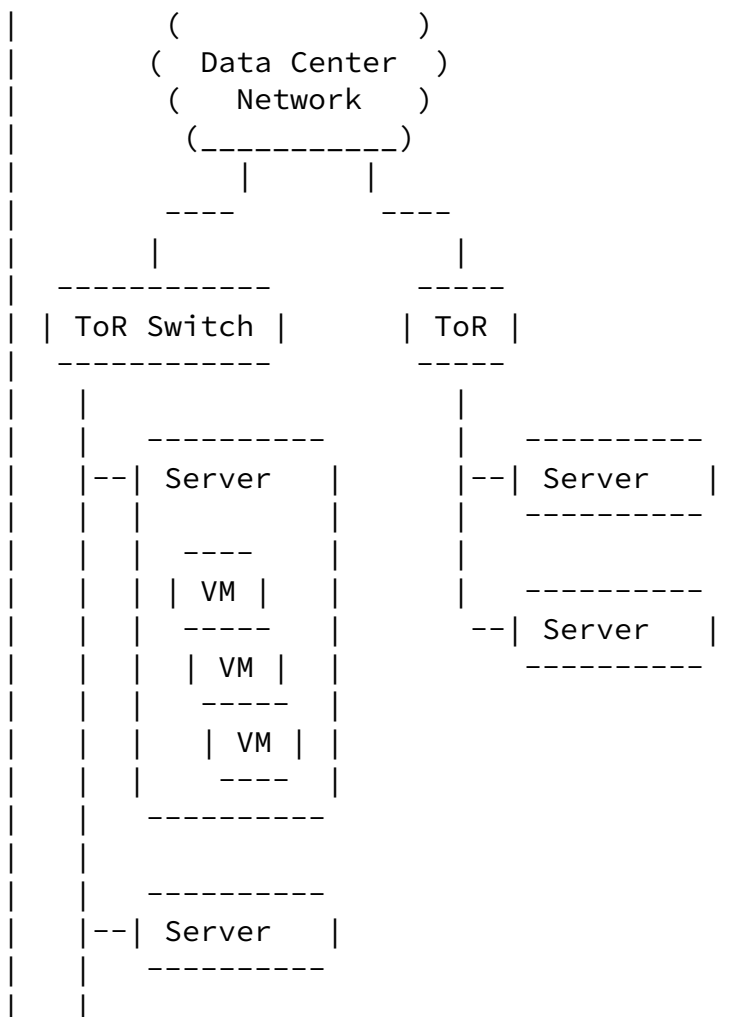
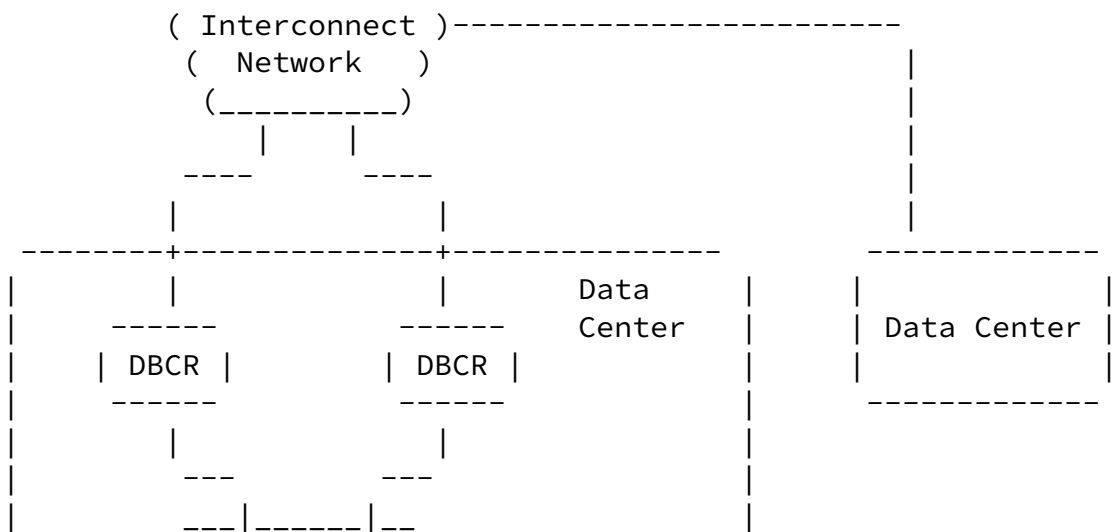
[2.1](#). Terminology

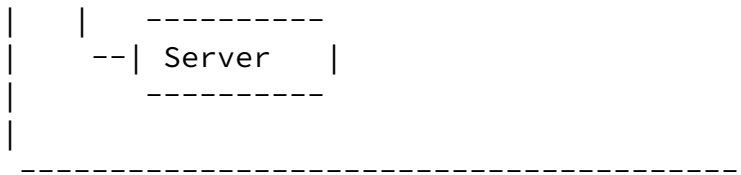
In this document the term "Top of Rack Switch (ToR)" is used to refer to a switch in a data center that is connected to the servers that host VMs. A data center may have multiple ToRs. When External Bridge Port Extenders (as defined by 802.1BR) are used to connect the servers to the data center network, the ToR switch is the Controlling Bridge.

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

()
(Data Center)





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

Consider a set of VMs that (as a matter of policy) are allowed to communicate with each other, and a collection of devices that interconnect these VMs. If communication among any VMs in that set could be accomplished in such a way as to preserve MAC source and destination addresses in the Ethernet header of the packets exchanged among these VMs (as these packets traverse from their sources to

their destinations), we will refer to such set of VMs as an Layer 2 based Closed User Group (L2-based CUG).

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

In terms of IP address assignment this document assumes that all VMs of a given L2-based CUG have their IP addresses assigned out of a single IP prefix. Thus, in the context of this document a single IP subnet corresponds to a single L2-based CUG. If a given VM is a member of more than one L2-based CUG, this VM would have multiple IP addresses and multiple logical interface, one IP address and one logical interface per each such CUG.

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

In this document the term "L2 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 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), etc... Note that any multi-chassis LAG can not 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 CUGs (while each of these CUGs may span multiple L2 physical domains). If an L2 physical domain contains servers that host VMs belonging to different L2-based CUGs, then enforcing L2-based CUGs 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 CUG is present within a given data center if one or more VMs that are part of that CUG are presently hosted by the servers located in that data center.

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.

[3.](#) Problem Statement

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

[3.1.](#) Usage of VLAN-IDs

This document assumes that within a given non-trivial L2 physical domain traffic from/to VMs that are in that domain, and belong to the same L2-based CUG MUST have the same VLAN-ID. This document assumes that in different non-trivial L2 physical domains traffic from/to VMs that are in these domains and belong to the same L2-based CUG MAY have either the same or different VLAN-IDs. 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 can not assume to stay the same.

This document assumes 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 when the VM moves from one L2 physical domain to another the VLAN-ID used by the Guest OS can not change (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 cannot be changed when the VM moves from one L2 physical domain to another, even though it is possible for an entity, such as hypervisor virtual switch, 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 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 CUG 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 CUGs, and (b) the need to overcome the limitation on the number of different VLAN-IDs.

[3.2.](#) Maintaining Connectivity in the Presence of VM Mobility

In the context of this document the ability to maintain connectivity in the presence of VM mobility means the ability to exchange traffic between a VM and its peer(s), as the VM moves from one server to another, where the peer(s) may be either other VM(s) or hosts. Furthermore, the peer(s) need not be within the same data center as the VM itself.

A given VM could be moved from one server to another in stopped or suspended state ("cold" VM mobility), or the hypervisors might move a running VM ("hot" VM mobility). IP address preservation is sometimes highly desired for cold VM mobility; it's mandatory to preserve transport connections when a running VM is moved.

VM mobility may result in transient loss of IP connectivity between VM and its peers. In the case of hot VM mobility the upper bound on the duration of such transients is (much) lower than in the case of cold VM mobility (due to the requirement of preserving transport connections and potential additional application requirements).

Furthermore, while with cold VM mobility one may assume that VM's ARP cache gets flushed once VM moves to another server, one can not make such an assumption with hot VM mobility.

[3.3.](#) Layer 2 Extension

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

Moreover, if the previous L2 physical domain no longer contains any VMs of that CUG, 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 CUG.

Note that supporting VM mobility implies that the set of L2 physical domains that contain VMs that belong to a given L2-based CUG 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 CUG, while the latter does not.

[3.4.](#) Optimal IP Routing

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

- + Optimal routing of a VM's outbound traffic. This means that as a given VM moves from one server to another, the VM's default gateway should be in a close topological proximity to the ToR that connects the server presently hosting that VM. Note that when we talk about optimal routing of the VM's outbound traffic, we mean traffic from that VM to the destinations that are outside of the VM's L2-based CUG. This document refers to this problem as the VM default gateway problem.
- + Optimal routing of VM's inbound traffic. This means that as a given VM moves from one server to another, the (inbound) traffic originated outside of the VM's L2-based CUG, and destined to that VM be routed via the router of the VM's L2-based CUG that is in a close topological proximity to the ToR that connects the server presently hosting that VM, without first traversing some other router of that L2-based CUG (the router of the VM's L2-based CUG 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.

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

Internet Draftdraft-rekhter-nvo3-vm-mobility-issues-03.txt October 2012

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

3.5. 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 contain 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.

4. IANA Considerations

This document introduces no new IANA Considerations.

5. Security Considerations

TBD.

6. Acknowledgements

The authors would like to thank Adrian Farrel for his review and comments. The authors would also like to thank Ivan Pepelnjak and David Black for their contributions to this document.

7. References

[nvo3-problem] Narten T. et al., "Overlays for Network Virtualization", [draft-narten-nvo3-overlay-problem-statement](#), work in progress.

Internet Draftdraft-rekhter-nvo3-vm-mobility-issues-03.txt October 2012

[8.](#) Author's Address

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

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

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

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

Linda Dunbar
Huawei Technologies
5340 Legacy Drive, Suite 175
Plano, TX 75024, USA
Phone: (469) 277 5840
Email: ldunbar@huawei.com

Ali Sajassi
Cisco Systems
Email: sajassi@cisco.com

Rahul Aggarwal

Arktan, Inc
Email: raggarwa_1@yahoo.com

Rekhter

[Page 11]