Network Working Group   Y. Rekhter
Internet Draft   Juniper Networks
Category: Standards Track
Expiration Date: February 2013

                                                    W. Henderickx
                                                    Alcatel-Lucent

                                                       R. Shekhar
                                                  Juniper Networks

                                                      Luyuan Fang
                                                    Cisco Systems

                                                     Linda Dunbar
                                                           Huawei

                                                      Ali Sajassi
                                                    Cisco Systems

                                                  August 20 2012

Network-related VM Mobility Issues

draft-rekhter-nvo3-vm-mobility-issues-00.txt

Status of this Memo

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".

Table of Contents

1. Specification of requirements

The key words "MUST", "MUST NOT", "REQUIRED", "SHALL", "SHALL NOT",
"SHOULD", "SHOULD NOT", "RECOMMENDED", "MAY", and "OPTIONAL" in this
document are to be interpreted as described in [RFC2119].

2. Introduction

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 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.

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)
                   ( Interconnect )-------------------------
                    (   Network    )                        |
                     (_____)                           |
                       |    |                               |
                     ----      ----                         |
                     |            |                         |
          --------+-------------+--------------     -------------
          |       |             |     Data     |    |           |
          |    ------        ------   Center    |    | Data Center |
          |   | DBCR |      | DBCR |            |    |           |
          |    ------        ------             |    -------------
          |       |             |               |
          |      ---           ---              |
          |      ___|_____|__                 |
          |      (            )                 |
          |    (   Data Center  )               |
          |     (    Network    )               |
          |       (_____)                  |
```

```
       |           |        |                        |
       |         ----      ----                       |
       |          |          |                        |
       |  ------------       -----                    |
       | | ToR Switch |     | ToR |                   |
       |  ------------       -----                    |
       |    |                  |                       |
       |    |   ----------     |   ----------          |
       |    |--| Server   |    |--| Server   | |
       |    |  |          |    |  |  ----------        |
       |    |  |  ----     |    |  |                    |
       |    |  | | VM |    |    |  |  ----------        |
       |    |  |  -----     |    |  --| Server   | |
       |    |  |  | VM |    |       ----------         |
       |    |  |   -----    |                           |
       |    |  |   | VM | |                             |
       |    |  |    ----   |                            |
       |    |   ----------                              |
       |    |                                           |
       |    |   ----------                              |
       |    |--| Server   |                             |
       |    |   ----------                              |
       |    |                                           |
       |    |   ----------                              |
       |    --| Server   |                             |
       |       ----------                              |
       |                                                |
         ------------------------------------------
```
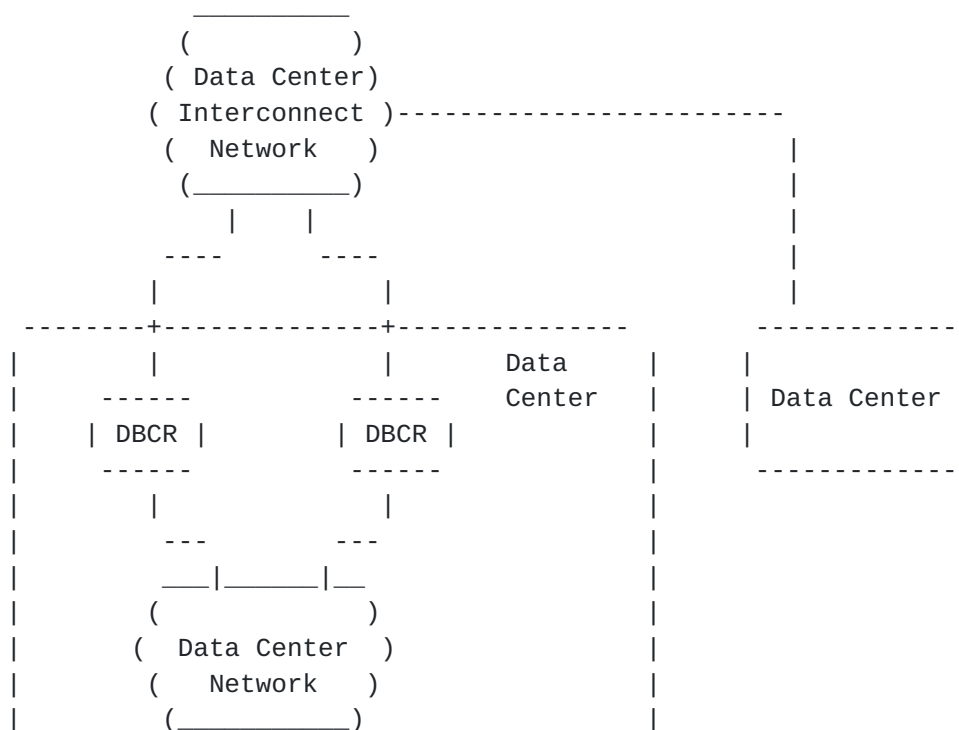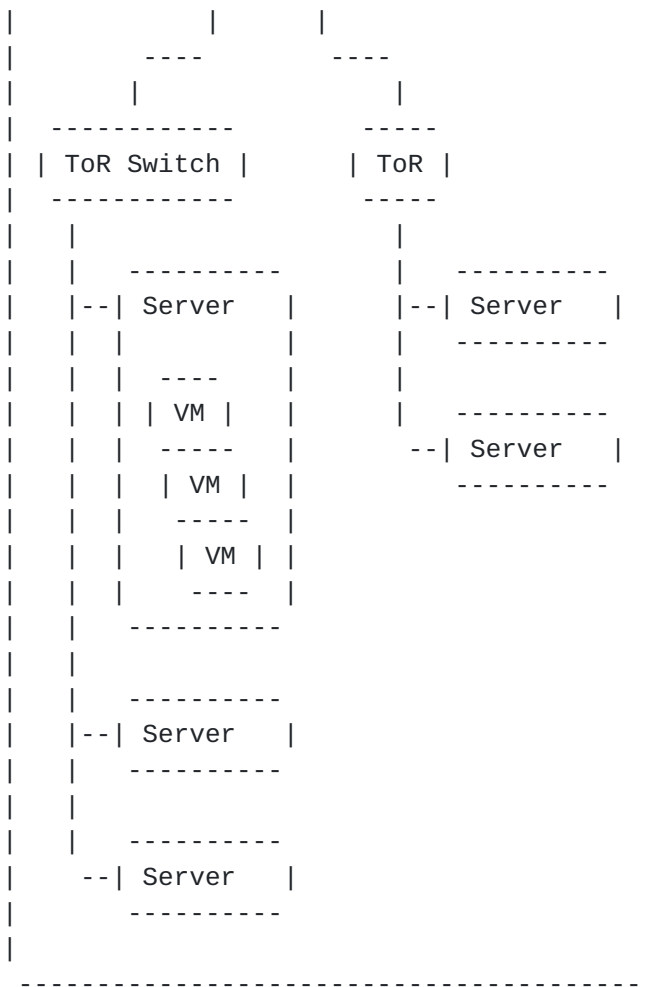
   The data centers and the network that interconnects them may be
   either (a) under the same administrative control, or (b) controlled
   by different administrations.

   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, one 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 site" refers to a collection of
interconnected devices that perform forwarding based on the
information carried in the Ethernet header. In a non-trivial L2 site
(site 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 is treated as a single
L2 site - a given multi-chassis LAG has to be contained within a
single L2 site. This document assumes that a layer 2 access domain
is an L2 site.

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

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

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

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 site as the VM, and that is either originated or
destined to that VM.

[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 VMs that belong to the same L2-based CUG,
   and are in the same L2 site MUST use the same VLAN-ID. This document
   assumes that VMs that belong to the same L2-based CUG, and are in
   different L2 sites MAY use either the same or different VLAN-IDs.
   Thus when a given VM moves from one L2 site to another, traffic the
   VLAN-ID used by the VM may change, and thus can not assume to stay
   the same. However, if a given VM's Guest OS sends packets that carry
   VLAN-ID, then when the VM moves from one L2 site to another the VLAN-
   ID used by the Guest OS can not change.

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

   The above assumptions about VLAN-IDs are driven by (a) the assumption
   that within a given L2 site 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.

[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 sites, where these sites may be located in the same or
   different data centers. In order to enable communication between this
   VM and other VMs of that L2-based CUG, the new L2 site must become
   interconnected with the other L2 site(s) that presently contain the

rest of the VMs of that CUG, and the interconnect must not violate
the L2-based CUG requirement to preserve source and destination MAC
addresses in the Ethernet header of the packets exchange between this
VM and other members of that CUG.

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

Note that supporting VM mobility implies that the set of L2 sites
that contain VMs that belong to a given L2-based CUG may change over
time (new sites added, old sites 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. This is also known as avoiding
    "triangular routing". This document refers to this problem as the
    triangular routing problem.

Note that optimal routing is a special case of maintaining
connectivity in the presence of VM mobility, as the former assumes

not only the ability to maintain connectivity, but also that this
connectivity is maintained using optimal routing. On the other hand,
maintaining connectivity does not make optimal routing a pre-
requisite.

4. 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.

7. References

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

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