Problem statement on address resolution in virtual machine migration
draft-liyz-armd-vm-migration-ps-00.txt

Status of this Memo

This Internet-Draft is submitted 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

This Internet-Draft will expire on April 18, 2009.

Copyright Notice

Copyright (c) 2010 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.
VM migration is one of the key features provided by larger scale virtualized data center. Various optimizations for address resolution in such network are expected to be provided by ARMD. This draft describes the problems that may be introduced by VM migration. It is expected that solutions provided by ARMD would address those problems.

Table of Contents

1. Introduction
2. Conventions used in this document
3. Problems in address resolution in VM migration
4. Security Considerations
5. IANA Considerations
6. Conclusions
7. References
   7.1. Normative References
   7.2. Informative References
8. Acknowledgments

1. Introduction

When virtualization is used in data center, it makes the server management more flexible and consequently more complex. One of the reasons is it would be much easier to move a VM (virtual machine) without the service interruption among physical servers. It is called VM migration. VM migration may occur due to server pool re-arrangement for maintenance, relocation, energy saving, load balancing, utilization optimization and other management purposes.

Figure 1 shows a typical VM migration scenario within a data center. VM1 moves from server 1 to server 2. VM migration is under control of the virtual machine management tools. It is known in advance by VM manager that where the VM would be moved to. Movement could occur between different servers of the same rack or across different racks or even across data centers.

The assumptions of VM migration include

- VM does not change its MAC and IP address after migration
- Service provided by VM should not be interrupted. Some packet loss may be observed at the moment of migration; however it should be recoverable by upper layer protocol and should not cause connection termination.
VM itself has no knowledge about its movement and therefore it should not be expected that VM would do anything special to accommodate the migration. On the other hand, hypervisor in a server participates in the whole migration process. Hypervisor in the destination server knows when the migration finishes and usually it will send certain data or control packet to signal the network entities that VM migration completes and it is ready to receive packets from the new location. Such signaling packet may be gratuitous ARP request, gratuitous ARP reply or reverse ARP depending on different implementation.

It has been shown in [I-D. dunbar-arp-for-large-dc-problem-statement] that there are basically two types of approaches used in virtualized larger layer 2 data center to solve the scaling issue,

1. Address translation: map raw flat MAC address to some hierarchical or manageable MAC address.

2. Address encapsulation: use additional header to encapsulate the frame/packet.

Either address translation or encapsulation could be performed by address registration or source address learning. In any case, VM live migration is a fundamental scenario to handle. The following sections talk about the problems caused by VM migration.
VM1 moves to server2

Figure 1 VM migration scenario
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].

3. Problems in address resolution in VM migration

Take figure 1 as example. During the process of VM1 movement, other hosts may still keep sending data packet to VM1. The switches including ToR1 have no knowledge that VM1 is going to move. All the packets still go to server 1 as normal. At the moment VM1 stops receiving packet from server 1, the incoming packet could be lost as the destination becomes a black hole to other hosts. After a short while, VM1 should be able to receive the packet from its new location server 2. It is very common that hypervisor at server 2 will flood a gratuitous ARP request/reply for VM1 to inform the whole broadcast domain about VM1’s new location.

There are a few problems may rise during the procedures described above:

3.1. No signaling message to indicate VM1 having left server.

Gratuitous ARP is a message to inform others a new node coming up for free. It is used for IP/MAC correspondence announcement. At same time, switches perform source MAC address learning to know the MAC/port/vlan correspondence. However there is no gratuitous ARP "leave" message to make others forget the previous learned source address and location information. Aging is a normal way to delete the cached information. Black hole may last existing as long as aging timout period. In virtualized system architecture, virtual machine management tool like vCenter knows a VM is going to move at management level. Therefore it is also possible to delete the stale cache through management plane and it needs collaboration between virtual machine manager and network manager. Another way might be to have some lightweight keepalive mechanism.

3.2. Uncertainty of signaling message after VM migration.

Currently there is no standard behavior defined for hypervisor in VM migration. Hypervisor may send gratuitous ARP request/reply and even reverse ARP after migration completes. The reason for sending the signaling message is to inform the switches and gateways about the new location of VM1 and make them have the correct entry for interface/port in the ARP/MAC table.
However, there are a large variety of ARP implementations. We have tested on one of Huawei’s switch on various ARP messages; the result is in figure 2.

The testing scenario is as follows. VM1 moves from server 1 to server 2 which connect to GW1 via interface 1 and interface 2 accordingly. Before migration, ARP table of GW1 has the entry to include IP/MAC of VM1 and its outgoing interface is itf1. After migration, hypervisor of server 2 may flood ARP or other signaling message; it is also possible that it keeps silent and does not send out any signaling packet and deals with ARP request/reply as normal. The expected result should be GW1 updates its ARP table entry to correlate VM1 with interface 2 (itf2).

<table>
<thead>
<tr>
<th>#</th>
<th>packet sent after VM1 migration</th>
<th>Is VM1’s interface updated to itf2 on GW1?</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>std gratuitous ARP</td>
<td>Y</td>
</tr>
<tr>
<td>2</td>
<td>broadcast ARP reply</td>
<td>N</td>
</tr>
<tr>
<td>3</td>
<td>RARP</td>
<td>N</td>
</tr>
<tr>
<td>4</td>
<td>ARP request with GW1 as target IP</td>
<td>Y</td>
</tr>
<tr>
<td>5</td>
<td>ARP request with other host as target IP</td>
<td>N</td>
</tr>
<tr>
<td>6</td>
<td>unicast ARP reply with GW1 as destination</td>
<td>Y</td>
</tr>
<tr>
<td>7</td>
<td>unicast ARP reply with other host as destination</td>
<td>N</td>
</tr>
</tbody>
</table>

Figure 2 Test result of GW ARP table update in VM migration

There are various implementation of switches and hypervisors. Figure 2 shows one example that depending on the type of ARP message sent by hypervisor and handling of switch, result may not be always as what we expect.

3.3. Difficulty of traffic redirection after migration.
Traffic redirection here has two meanings: first, as VM starts to operate in the new location, all switches in the network should be able to correctly send the frame destined to that VM. That is to say, the interface/port corresponding to VM has to be updated in relevant switches as soon as possible; second, if packets received at VM’s old location were cached by some network entity during VM migration, those packets could be redirected to VM in the new location. Such redirection may probably last for a while after VM completes migration to tolerate the unexpected delay caused by some switches keeping sending frames to VM’s old location.

Normally we want to minimize broadcast ARP message to alleviate the burden on switch and server. On the other hand, it would be necessary to do blind flooding of ARP in VM migration as it is hard to determine who are the entities that should be informed with new location and who are not. It is also not easy for a network entity to know if a VM is a newly starting one or a migrated one. Optimization for new VM and migrated VM may go through different procedures.

4. Security Considerations

It may not be easy to tell if an ARP sent from a new location is really for a migrated VM or it is a spoofed one. With VM migration, some security mechanism may not be applicable any more, like:

- MAC locking: locking a MAC address to a specific physical port of the switch.

- DHCP snooping: binding IP/MAC by snooping DHCP ACK to port of switch. VM may not send DHCP request again after migration. Some mechanism should be introduced to move the binding to the new port in migration case.

VM migration itself does not introduce more risk to ARP messages. However some existing solutions to solve ARP security issues may wrongly treat ARP after migration as illegal one.

5. IANA Considerations

This document requires no IANA actions.
6. Conclusions

VM migration brings extra problem to larger scale virtualized data center. Any solution in ARMD, like directory based address resolution, distributed caching, or specially designed control protocol, should take the problems into consideration.

7. References

7.1. Normative References


7.2. Informative References


8. Acknowledgments

This document was prepared using 2-Word-v2.0.template.dot.
Authors’ Addresses

Li Yizhou
Huawei Technologies
101 Software Avenue,
Nanjing 210012
China

Phone: +86-25-56622310
Email: liyizhou@huawei.com