ARMD
Internet Draft
Intended status: Informational
Expires: September 2011

Y. Li
Huawei Technologies

March 11, 2011

Problem statement on address resolution in virtual machine migration
draft-liyz-armd-vm-migration-ps-01.txt

Status of this Memo

Copyright Notice

Abstract

VM migration is one of the key features provided by larger scaled virtualized data center. Various optimizations for address resolution in such network are expected to be provided by ARMD. This draft describes the problems that are introduced by VM migration. It is expected that solutions provided by ARMD would address these problems.

Table of Contents

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

o VM does not change its MAC and IP address after migration

- o 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 at 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.

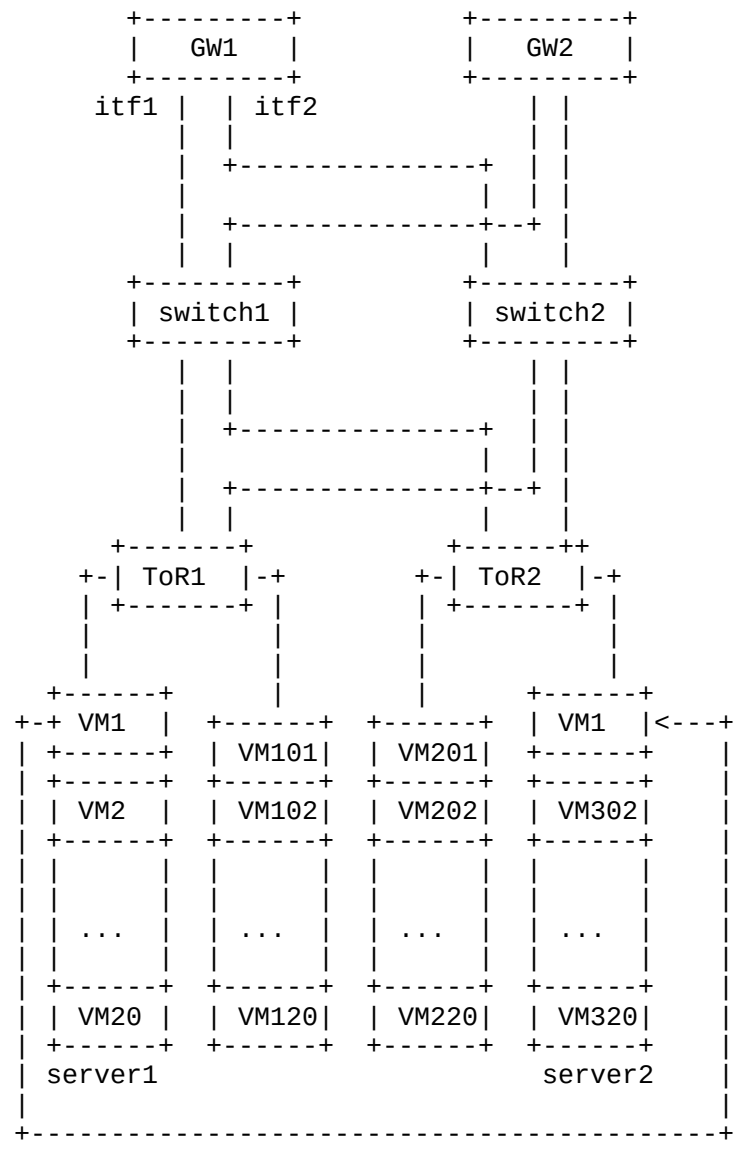

Figure 1 VM migration scenario

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. Some dimensions to consider in supporting VM migration

When we investigate the impact on ARP traffic by VM migration in data center, there are several dimensions to examine.

- Network topology. VM can be moved within a single layer 2 domain in current practice. The range of the domain restricts the movement. Therefore position of default gateways normally determines the size of the layer 2 network as they terminates the layer 2 traffic and handles the layer 3 traffic. If the default gateway is aligned with ToR, VM can only migrate within the same rack. If the default gateway is aligned with core switches, VM can be moved within the whole network. Therefore larger sized layer 2 network is more preferred considering VM migration.

- Protocol used at layer 2. Traditionally STP is used. In order to enjoy more efficient use of all links and faster convergence and support multipathing for fat tree structure based data center, routing based layer 2 protocol like TRILL or SPB are expected to be used in data center. They both provide additional encapsulation at the edge switches and make the core nodes simpler at the forwarding plane. Different operational recommendation may be needed for each.

4. ARP 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.

In traditional switches, there is no ARP table. Only routers/gateways keep the ARP table. In some of the approaches, switches have the ARP

cache for local host and/or remote host. We will study the impact for both.

4.1 No ARP message to indicate VM having left a 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 as long as aging out time.

There are several ways to make it up.

- Operationally if the VM sends out the gratuitous ARP or reverse ARP right after the migration, and the message is not lost, it will fresh the ARP table entry on gateways and switches. It is the most common way given that migration process, i.e. the time from VM stopping receiving frame at old location to VM starting receiving frame at new location, is very short and the frame lost is rare.

- In virtualized system architecture, virtual machine management tool like vCenter knows a VM is going to move at management level. Therefore it is possible to delete the stale cache through management plane and it needs collaboration between virtual machine manager and network manager.

- Use some lightweight keepalive mechanism to guarantee the freshness of the local ARP entry. It is called ARP detection in some implementations. It decreases the possibility of re-issuing gratuitous ARP for silent hosts. If an ARP entry becomes invalid, some specific message needs to be flooded to let remote switches "forget" the entry if switch also has the ARP cache for remote hosts.

4.2 Uncertainty of ARP message type 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 switches in market 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 in which case black hole problem would become more significant. The expected result should be GW1 updates its ARP table entry to correlate VM1 with interface 2 (itf2) as soon as possible when VM finishes migration.

#	packet sent aft VM1 migration	Is VM1's interface updated to itf2 on GW1?
1	std gratuitous ARP	Y
2	broadcast ARP reply	N
3	RARP	N
4	ARP request with GW1 as target IP	Y
5	ARP request with other host as target IP	N
6	unicast ARP reply with GW1 as destination	Y
7	unicast ARP reply with other host as destination	N

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

There are various implementations 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.

It is recommended that interface number for an ARP table entry on gateway should be updated for any ARP messages including ARP request/reply and reverse ARP no matter if the frame is destined for itself.

4.3 ARP message unreliable delivery

Gratuitous ARP from an end host is normally sent three times in order to survive from frame loss. However it is hard to 100% avoid ARP frame loss. Some analysis says a typical congestion is about 10-20 seconds which is longer than 3 retries of gratuitous ARP. In case the ARP frames are lost after VM migration, the gateway is not able to correctly update the corresponding interface number in ARP table entry. For inbound traffic from gateway, the gateway will keep sending it to the old location which is a black hole. It is noted that the ARP table will not be updated by data frames. Hence even the VM sends out data frame from new location, gateway will not update the relevant entry of ARP table.

For internal traffic within data center, if switches do not have any ARP cache, MAC/port correspondence will be updated accordingly along the path. As most of the data traffic should be bidirectional, MAC table should be correctly updated after a short while. Everything should be ok. On the other hand, if switches have ARP caching table, situation would be more completed depending on where the frame is lost, if switches cache remote ARP entry.

If ARP table is updated by data frames in addition ARP frame, it will solve most of the problems here. However, it may bring some performance and security issue.

4.4 Duplicate address detection

Gratuitous ARP is also used for duplicate address detection. For example, in Windows NT 4.0 with Service Pack 3 or higher installed, a statically addressed Windows NT computer will perform a gratuitous ARP up to 3 times: 1 time when the TCP/IP stack initializes, and 2 more times after .5 and 1 second intervals, if no response is received. Whenever a statically configured IP address is changed, Windows NT sends a single gratuitous ARP. If Windows NT receives a response to a gratuitous ARP, it disables the interface that issued the gratuitous ARP, generates an event (event ID 26), and generates a pop-up dialog box on the console warning the user that a duplicate IP address has been detected resulting in the shutdown of the affected interface. For DHCP leased address, Windows NT sends a single gratuitous ARP.

VM migration normally takes time in magnitude of second depending on the amount of memory to be copied over at the last stage. If another VM starts up and tries to use the same IP address of the migrated VM right within its migration process, there will be no duplicate address detected. Therefore the new VM can safely uses that IP address. Then after the migrated VM completes the movement, there will be duplicated IP address running at same time or migrated VM will block itself from using that IP address. Neither behavior is desired.

5. 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 mechanisms are not 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 does 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.

6. IANA Considerations

This document requires no IANA actions.

7. 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 consider the VM migration carefully. It is suggested to include the information from the draft in the problem statement of impact on address resolution for massive number of hosts in the data center.

8. References

8.1. Normative References

[ARP] D.C. Plummer, "An Ethernet address resolution protocol." RFC826, Nov 1982.

8.2. Informative References

[I-D. dunbar-arp-for-large-dc-problem-statement]Dunbar, L. and Hares, S., " Scalable Address Resolution for Large Data Center Problem Statements", draft-dunbar-arp-for-large-dc-problem-statement-00, July 2010.

9. 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