Routing Optimization for Live VM Migration between Datacenters

Yukio Nagafuchi*, Yasuhiro Teramoto*, Bo Hu*, Toshiharu Kishi*, Takaaki Koyama*and Hideo Kitazume*
*NTT Secure Platform Laboratories
Email:{nagafuchi.yukio, teramoto.yasuhiro, hu.bo, kishi.toshiharu, koyama.takaaki, kitazume.hideo}@lab.ntt.co.jp

***Abstract*—Virtual Machines (VMs) and network virtualization technologies have been widely used for cloud computing. Active VMs can move to other execution environments, an action called "live migration." VMs must have the same IP address before and after migration. Therefore, live migration must be in the same segment. For this reason, network virtualization technologies are used to make a wide area L2 network between data centers (DC). Such L2 networks have a redundant route problem resulting from the distance between a particular VM and the default gateway. To solve this problem, each DC has a default gateway and VMs use the nearest default gateway. However, if live migration is used, the VMs' nearest default gateway changes. In this paper, we explain these problems, and propose solutions to change to the proper default gateway.**

I. Introduction

With the recent development of virtualization technologies, virtual machines (VMs) and network virtualization technologies have become widely used for cloud computing[1].

One major advantage of VMs is live migration. Live migration is a technology to move a VM from the current Hypervisor to another. The network virtualization technologies([2][3][4]) support live migration between DCs(Fig. 1). Any communication sessions are held during live migration, so the VM IP address must not be different before and after migration.

Logical L2 network between DCs where a VM and a default gateway (DGW) are at different DCs as shown in Fig. 2-(i), if the VM communicates with a nearby user terminal, traffic goes and comes back again. This redundant route (Route 1 in Fig. 2), called a "trombone phenomenon ", increases the delay. This problem can be avoided by installing a DGW at each DC, so a VM uses a nearby DGW. If the VM moves to another DC by live migration, however, the same problem occurs again without no route control because the DGW before migration is used. To solve this problem, this paper proposes methods of optimizing routes after live migration while holding communication sessions.

This paper is organized as follows: Section II introduces usecases of live migration between DCs. Sec.III describes network topologies and route selection to avoid the "trombone phenomenon" under a live migration environment across DCs. Sec. IV describes the form of live migration between DCs and the necessity of making the route after live migration equivalent to the original route to prevent redundancy. Sec.V describes the switching of a route from a VM to a user terminal (uplink). Sec.VI describes the switching of a route from a user terminal to a VM (downlink). Sec.VII concludes this paper.

II. Usecases

This section gives live migration usecases (a) to (c).

(a) Disaster recovery: If a corporate server is operated in a VM and the DC is struck by a disaster, the VM is moved to a DC in an unaffected area.

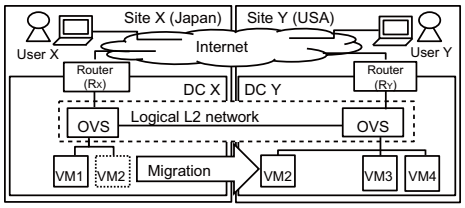


Fig. 1: Live migration between datacenters

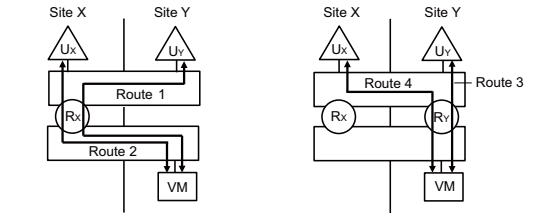


Fig. 2: Network topologies

(b) Desktop as a Service: In DaaS, if a DC is prepared in each country and a user travels overseas on a business trip or for other purposes, the VM is moved to a DC near the user to reduce delays.

(c) Cloud service maintenance and operation: For maintenance or operation, such as legal inspection or nighttime power saving, all VMs are moved to other areas.

III. Environment of Live Migration between DCs

This section describes network topologies to avoid the problem of a redundant route after migration between DCs (trombone phenomenon) and route selection to control routes according to network topologies.

A. Network Topologies

An IP network usually has one or more routers for external network connections. For the environment of live migration between DCs, it is preferable to install a router at each DC. This section explains the reason with reference to Fig. 2.

In Fig. 2-(i), a DGW is not installed at each DC but the DGW at Site X is shared by VMs at Site Y. In Fig. 2-(ii), a DGW is installed at each DC.

For VM communication with a user at Site Y, Route 1 is used in Fig. 2-(i) and Route 3 is used in Fig. 2-(ii). The route of going far once and coming back, like Route 1, is called the "trombone phenomenon". When Route 1 is used, the delay is equal to one round trip to R_x within the network. The delay by Route 3 is obviously shorter than that by Route 1. The delay of one round trip within a network increases as the inter-DC distance becomes longer. For VM communication with a user at Site X, there is no difference of delay by distance between Routes 2 and 3 in Fig. 2.

For these reasons, a network topology where a router exists at each DC is preferable for the environment of live migration between DCs. Therefore, the topology is used in this paper.

B. Route Selection

This section describes the uplink and downlink route selection methods based on the network topology introduced in Section III-A. Communication from a VM to a user terminal is called uplink and that from a user terminal to a VM is called downlink.

1) Selection of uplink route: An uplink route is selected by setting the IP address of via routers ($\mathrm{R_X}$ and $\mathrm{R_Y}$) for DGW by the VM OS.

2) Selection of downlink route: A downlink route should be selected by the NAT or routing method according to the network type, such as Internet or VPN. The methods are distinguished as explained below. The NAT method is used for a network where a user cannot control the route of advertising from a site in real time (e.g. IP-VPN where a user must apply to the network operator for an Internet or advertising route address). In Fig. 3a, different IP address bands are assigned to DCs at Sites X and Y. For example, the address band of aaa.bbb.ccc.0/24 is assigned to the DC at Site X and that of xxx.yyy.zzz.0/24 to the DC at Site Y. From these IP address bands, an IP address is selected for VM access to each site. In Fig. 3a, $\mathrm{IP\_R_X}$ is selected as the address of Site X and $\mathrm{IP\_R_Y}$ as that of Site Y. A VM at a DC has a private IP address and $\mathrm{IP_{VM}}$ is set. Router $\mathrm{IP\_R_X}$ at Site X has a NAT table (Table 1 in Fig. 3a) for mapping $\mathrm{IP\_R_X}$ and $\mathrm{IP_{VM}}$. $\mathrm{R_Y}$ at Site Y has a NAT table (Table 2 in Fig. 3a) for mapping $\mathrm{IP\_R_Y}$ and $\mathrm{IP_{VM}}$. In this status, when a user terminal sets $\mathrm{IP\_R_X}$ as the destination IP address, Router X is used. When a user terminal sets $\mathrm{IP\_R_Y}$ as the destination IP address, Route Y is used. (To describe router selection, NAT tables were set to $\mathrm{R_X}$ and $\mathrm{R_Y}$ here. As described in the previous section, the optimum route in Fig. 3a is Route X and Route Y is not used. Therefore, there is no need to set a NAT table for $\mathrm{R_Y}$. This is also true for the routing method described next.)

The routing method is used for a network where a user can control the advertising route from a site in real time (e.g. IP-VPN service not limiting address information for an advertising route). The mechanism of the routing method is route control usually used on a network. In Fig. 3b, when Routers $\mathrm{R_U}$, $\mathrm{R_X}$, and $\mathrm{R_Y}$ are in the same segment because of tunneling or VPN, the $\mathrm{R_U}$ routing table is rewritten statically for this method. When not connecting these routers in the same segment but relaying them through external network routers, either $\mathrm{R_X}$ or $\mathrm{R_Y}$ can be selected as a route for advertising the IP address $\mathrm{IP_{VM}}$ of a VM. For independent route control in units of VM, however, route control by "IP address/32", called the host route, is necessary. Since frequent use increases the number of entries in the routing table of the relay router and causes a heavy load, the NAT method is considered better if the same segment cannot be constructed.

IV. Live Migration between DCs

To implement live migration between DCs, this paper assumes the system configuration shown in Fig. 3c. The usecases introduced in Section II were for live migration by a user and that by a maintenance engineer. Live migration is controlled by a cloud controller(CLC) at a DC. A CLC is a device to manage the cloud environment. An operator or maintenance engineer who creates VMs, constructs a virtual network, or migrates VMs can directly operate a CLC. However, CLCs cannot be accessed directly by users and are generally operated through web servers that are extended to the Internet or other external networks. For security, only functions to be released to users are made public on web servers. In this paper, a web server makes the live migration function public to users. The web server and CLC are connected by using IPsec and other encryption technologies and the route tunneling technologies. Similarly, a CLC has functions to connect VMs, routers, web servers, and DNS servers and to manage IP addresses assigned to VMs.

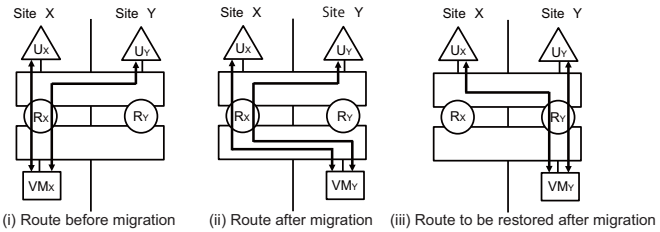


Fig. 4: Scope of live migration between DCs

A. System Configuration and Implementation Form for Live Migration between Datacenters

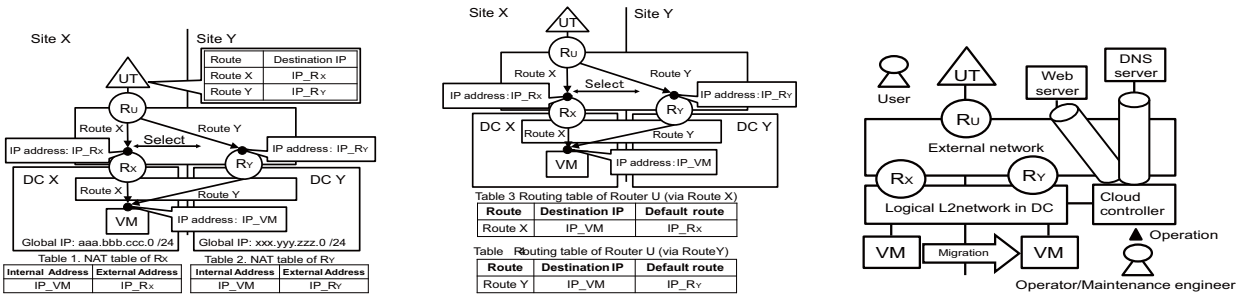
To implement live migration between DCs, this paper assumes the system configuration shown in Fig. 3c. Live migration is controlled by a cloud controller(CLC) at a DC. A CLC is a device to manage the cloud environment. An operator who creates VMs, constructs a virtual network, or migrates VMs can directly operate a CLC. Similarly, a CLC has functions to connect VMs, routers, web servers, and DNS servers and to manage IP addresses assigned to VMs.

B. Route after Live Migration

This section clarifies that live migration makes a route redundant with reference to Fig. 4. Fig. 4-(i) shows the route before live migration. The via router after live migration between DCs is that shown in Fig. 4-(ii) because the one set after live migration is used. As mentioned in the description of network topologies in, if no DGW exists at each DC, the route becomes redundant as equally as that when there is no DGW at the DC (Fig. 2-(ii)). After live migration, therefore, it is necessary to switch the route before live migration to the equivalent route shown in Fig. 4-(iii) where the via router and post-migration VM are at the same DC. Route switching is necessary each for uplink and downlink because different setting methods are used for uplink and downlink as described in Section III-B.

V. Uplink Route Switching

The simplest method of switching the uplink route is that a user or operator changes the DGW setting of the VM after migration. However, this method is not convenient as a communication service because the user or operator needs to know the IP address of the switch-to DGW in advance and some work is necessary. One solution to this problem is that a CLC rewrites the DGW setting of the VM automatically after migration. If a DGW is switched by these methods, however, the live migration requirement for keeping a communication session cannot be satisfied. Therefore, this paper proposes the construction of a network where the DGW at each DC uses



(a) Downlink route selection (NAT) (b) Downlink route selection (Routing) (c) System configuration

Fig. 3: Downlink route switching

the same IP address so that a route can be switched without changing the IP address of a VM DGW before and after migration while keeping a communication session.

A. Route Switching Using the Same IP for DGW

This section describes the proposed method of making the DGW IP address at each DC identical. To implement this method, however, functions need to be added to routers. Routers require the following additional functions.

- Suppression of contention detection: If a router or other IP equipment finds its own IP address within the same segment, it usually notifies the detection of contention to the network administrator. Since two or more routers use the same IP address in the same segment for this method, the detection function should be stopped. Therefore, Routers $\mathrm{R_X}$ and $\mathrm{R_Y}$ require an additional function to make the interface of the logical L2 network stop detecting contention.
- Grouping: A group is created to synchronize information between routers. Router names and interfaces are registered to the group. Routers in the registered group are suppressed from detecting contention. If the DC-side interfaces of $\mathrm{R_X}$ and $\mathrm{R_Y}$ are registered in a group, the detection of DC-side interface contention between $\mathrm{R_X}$ and $\mathrm{R_Y}$ is suppressed.
- MAC address registration: To each router, the MAC addresses of VMs under the router are registered. This registered information is synchronized between groups ($\mathrm{R_X}$ and $\mathrm{R_Y}$).
- Response to ARP REQUEST: When ARP REQUESTs are received, each router responds only to requests having the MAC addresses of VMs registered in the router. This allows two or more routers in the same segment to have the same IP address as DGW.

After VM migration, the cloud controller(CLC) registers the VM to $\mathrm{R_Y}$ and sends Gratuitous ARP (GARP) for the VM to $\mathrm{R_Y}$ to make the router re-learn the MAC table of the VM and router and to switch the route. Without having to change the DGW setting, the VM can use the DGW at the destination DC.

VI. Downlink Route Switching

A. Route Switching by NAT

This section describes route switching by NAT with reference to Fig. 3a. A route can be switched by selecting a route by the method described in Section III-B. Unlike route selection, migration is executed by a CLC. If a CLC rewrites the NAT tables of $\mathrm{R_X}$ and $\mathrm{R_Y}$, through processing can be executed. Here is an example of processing by a CLC, including IP address management.

Before live migration: When creating a VM, the CLC selects a private address for use within the logical L2 network of the DC and an address to assign ($\mathrm{IP\_R_X}$ in Fig. 3a) from the pool of global IP addresses (aaa.bbb.ccc.0/24) assigned to Site X. The CLC accesses a router at the same site ($\mathrm{R_X}$ here) and creates a NAT table (Table 1 in Fig. 3a). Then the CLC makes public a global IP address to access a web server or DNS server and for a user terminal to access the VM.

After live migration: From the pool of global IP addresses assigned to DC1 (xxx.yyy.zzz.0/24), the CLC selects an address to assign to the VM ($\mathrm{IP\_R_Y}$ in Fig. 3a). Based on this information, the CLC accesses $\mathrm{R_Y}$ and creates a NAT table (Table 2 in Fig. 3a). Through a web server or DNS server, the CLC accesses, the CLC provides a global IP address to access the migrated VM. By using this information, a user can access the VM through the post-migration route. A NAT table written in the CLC is for static NAT and does not disappear unless erased from outside. Therefore, the CLC needs to delete the NAT table of a VM moved from the pre-switching router after a specified time (e.g. 1 or 24 hours).

B. Route Switching by Routing

This section explains route switching by routing with reference to Fig. 3b. After live migration, the CLC accesses $\mathrm{R_U}$ and rewrites the entry of destination IP address to the IP address of the VM in the routing table information from $\mathrm{R_X}$ to $\mathrm{R_Y}$ for route switching. In other words, the routing table of $\mathrm{R_U}$ before live migration is Table 3 in Fig. 3b for communication through $\mathrm{R_X}$. For communication through $\mathrm{R_Y}$ after VM live migration, the table is rewritten to Table 4 in Fig. 3b for route switching.

VII. Conclusion

This paper described the problem of a redundant route and a large delay after live migration between DCs because via routers remain unchanged from those before live migration. To optimize redundant routes while keeping communication sessions, this paper proposed different route switching methods for uplink and downlink. By using the proposed methods, we believe the problem of redundant routes after live migration can be solved. In future, the proposed methods will be implemented to confirm route switching without clearing TCP sessions.

References

[1] P. Mell and T. Grance, "The NIST definition of cloud computing (Draft)," *Recommendations of the National Institute of Standards and Technology*, NIST Special Publication 800-145(Draft), 2011

[2] VMware NSX Website, https://www.vmware.com/products/nsx/

[3] Open vSwitch Website, http://www.openvswitch.org/

[4] H. Kitazume, T. Koyama, T. Kishi, T. Inoue, "Network Virtualization Technology to Support Cloud Services," *IEICE TRANSACTIONS on Commu.*, Vol.E95-B No.8, pp.2530–2537, 2012.