Network Working Group
Internet-Draft
Intended status: Standards Track
Expires: September 8, 2011

Y. Gu
Huawei

March 7, 2011

Policies and dynamic data migration in DC
draft-gu-dc-management-problem-statement-00

Abstract

Virtualization provides Data Center with feasibility and improves the utilization of limited physical resource, e.g. switches/routers, servers and links. Virtual machines (VM) are allowed to migrate to any place in the Data Center. A variety of policies (e.g. ACL, firewalls, load balancers, IPS and QoS) are deployed in Data Center to guarantee the SLA the provider signed with their clients. Dynamic information, such as TCP Connection Table, dynamic ACLs and cumulated data, is generated on network devices. In order to keep running services uninterrupted while VM migrating, relevant policies and dynamic information, also need to migrate with VM.

This document describes some examples of the policies and dynamic information that need to migrate with VM, the influence if they are not migrated with VM, the problems that need to consider when migrate polices and dynamic information. It also describes some existing network management protocols standardized by IETF and the advantages and disadvantages of them for operating policies and dynamic information migration respectively. The goal is to justify that it is necessary for IETF to make effort on policy and dynamic information migration for large virtualized Data Center.

Copyright Notice

Table of Contents

1. Introduction

Data centers can host tens or even thousands of different applications. Some are simple applications such as web servers providing static content, while some may be very complex, e.g. e-commerce, that requiring all around privacy protection and data security. Clients of Data Center, unlike server hosting clients, raise more strict QoS and Security requirements. Clients may sign Service Level Agreement (SLA) with Data Center Provider to make sure their requirement can be guaranteed. To satisfy different level of security requirements and to manage and improve the performance of these applications, data centers typically deploy a large variety of middleboxes, including firewalls, load balancers, SSL offloaders, web caches, and intrusion prevention boxes.

To satisfy QoS requirements, Data Center also implement QoS mechanism as ISP network. For example, to deploy polices on Switches to execute traffic classification and marking. IEEE 802.1 DCB working group defines a series of standards to guarantee quality of service.

802.1Qau - Congestion Notification

802.1Qaz - Enhanced Transmission Selection

802.1Qbb - Priority-based Flow Control

Without regard to mobile network, the existing DC network management has a pre-assumption that the end hosts will not move. If an end host moves, because the physical link has to break down and the service also has to break down, the network can treat it as two separated parts: one host leave the network and another host join the network.

Server Virtualization and Virtual Machine Migration changes the situation and break the preassumption. Server Virtualization is not a new technology. But, because Cloud services become popular, which requires flexible resource assignment and effectively resource integration, server virtualization revitalizes again. Using server virtualization technologies, network adminitrator can reduce networking cost. To support the same volume of services, fewer network devices, servers and links are required than before. Multiple Virtual Machines (VMs) are established within a single physical server and the VMs are allowed to relocate to a different servers within the same subnet of Data Center, or even among different sites of a Data Center. This is so called VM Migration. VM Migration brings flexibility to Data Center, meanwhile it makes network management more complex and challenging.

While VM migrates, a very important requirement is that running services on the VM mustn't been interrupted. Though a 'zero delay' on running services is not realistic, but the services should be able to continue after a very short delay.

In order to avoid service interruption and minimize delay on running services, polices and dynamic information on network devices must be migrated timely and accurately. The policies and dynamic information includes those on switches, routers and middleboxes.

In the following section, we describe the policies and dynamic information migration on several example network devices. The influence to running services if they are not migrated accurately and timely. Then we will introduce the limitations of existing network management protocol.

2. Terminologies and concepts

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].

Source Network Device, Source switch, or Source device: the network device/switch/device from where the VM migrates. I.E. VM is originally located under the source network device/switch/device.

Destination Network Device, Destination switch, or Destination device: the network device/switch/device to where the VM migrates. I.E. VM is relocated to the destination network device/switch/device.

TCP connection table: A table containing TCP connection-specific information.

3. Policies on several network devices

In this draft, our discussion using the following figure as an example networking. The links between AGG1/AGG2 and Gateway2, AGG3/ AGG4 and Gateway1 are omitted for simplicity. the new VM1 under Server4 represents the VM1 after migration. VM1 and new VM1 don't exist at the same time. In the real world, the networking of DC could be different.

```
                      ------------------               ------------------
                      |    Gateway1    |-------------|    Gateway2    |
                      ------------------               ------------------
                         /     \                          /       \
                        /       \                        /         \
 ------------      --------     --------              --------
--------       ------------
 |Firewall-A|------| AGG1  |----| AGG2  |              | AGG3  |----| AGG4
|----- |Firewall-B|
 ------------       --------     --------              --------
--------       ------------
                    |      |     |     |                 |      |    |     |
                    |      \     /     |                 |      \    /     |
                    |       \   /      |                 |       \  /      |
                    |        \/        |                 |        \/       |
                    |        /\        |                 |        /\       |
                    |       /  \       |                 |       /  \      |
                    |      /    \      |                 |      /    \     |
                    |     |      |     |                 |     |      |    |
                   ----------   ----------              ----------   ----------
                   |Switch1 |   |Switch2 |              |Switch3 |   |Switch4 |
                   ----------   ----------              ----------   ----------
                    |       \   /      |                 |       \  /      |
                    |        \/        |                 |        \/       |
                    |        /\        |                 |        /\       |
                    |       /  \       |                 |       /  \      |
                  ----------   ----------              ----------   ----------
                  |Server1 |   |Server2 |              |Server3 |   |Server4 |
                  ----------   ----------              ----------   ----------
                   |     |                                           |     |
                 VM1    VM2                                        VM3    new
VM1
```

Fig1. Basic networking for discussion in this draft.

3.1. Policies and configurations

SLA is parsed into a collection of polices, which can be described by natural languages or mathematics fomula. Then policies are represented by specific configurations on different network elements, e.g. physical ports on routers and switches. In this draft, we discuss the migration of policies, but the policies migration also implies the migration of configurations on network devices, because configurations are embodiment of policies.

Policies that need to migrate with VM are those can influence VM's running services. The policies could be different on different network devices. For example, it can be static Access Control Lists on Access switches, QoS on switches and routers, security rules on Firewalls, etc.

Take Access Control List as an example. Figure 1 shows the influence of lack of ACLs on destination switch. There is an ACL 100 on source switch (Switch1) deny all packets from IP subnet 10.138.3.0 to Internet. And another ACL 101 allows IP Address 10.138.3.1, VM1's IP Address, to send packets to Internet. VM1 has a running service on it. During service provisioning, VM1 is migrated to Server4 under Switch4, Where there is no ACL 100 and ACL 101. VM1's IP Address falls into a default ACL which deny all unmatching packets. As a result, packets belonging to the running services are dropped, hence the running service is interrupted.

```
                      ------------------             ------------------
                      |    Gateway1    |-------------|    Gateway2    |
                      ------------------             ------------------
                         /     \                           /       \
                        /       \                         /         \
 ------------      --------     --------               --------
--------       ------------
 |Firewall-A|------| AGG1  |----| AGG2  |              | AGG3  |----| AGG4
|----- |Firewall-B|
 ------------       --------     --------              --------
--------       ------------
                    |      |     |     |                 |      |    |     |
                    |      \     /     |                 |      \    /     |
                    |       \   /      |                 |       \  /      |
                    |        \/        |                 |        \/       |
                    |        /\        |                 |        /\       |
                    |       /  \       |                 |       /  \      |
                    |      /    \      |                 |      /    \     |
                    |     |     |      |                 |     |     |     |
    ACL 101        ----------   ----------              ----------   ----------
    ACL 100        |Switch1 |   |Switch2 |              |Switch3 |   |Switch4
|   Default ACL:
                   ----------   ----------              ----------
----------   deny all
                    |       \   /      |                 |       \  /      |
                    |        \/        |                 |        \/       |
                    |        /\        |                 |        /\       |
                    |       /  \       |                 |       /  \      |
                  ----------   ----------              ----------   ----------
                  |Server1 |   |Server2 |              |Server3 |   |Server4 |
                  ----------   ----------              ----------   ----------
                   |     |                                            |     |
                 VM1    VM2                                       VM3    new
VM1
 Fig.2 VM migration without ACL migration
```

4. Dynamic Information and the influence if lack or unaccurate

Network Manager (NM) can configure static configuration on network devices. Except for the static configuration, some dynamic information could also be recorded and processed by network devices. TCP connection table is an obvious example. Normally, TCP Connection Table is not configured by NM, but is generated by network devices, e.g. Firewalls, by looking into the packets passing them. TCP

Connection Table can be used for forwarding and security reasons. Another example is cumulated data, e.g. how many packets/TCP connecition requests an end host has sent. This information can only be generated by network devices according to real traffic. Configurations could be generated dynamically by network devices themselves according to the dynamic informaiton, e.g. Dynamic ACLs.

4.1. TCP connection tables

A typical TCP Connection Table includes the following data:

tcpConnState: The state of this TCP connection.

tcpConnLocalAddress: The local IP address for this TCP connection.

tcpConnLocalPort: The local port number for this TCP connection.

tcpConnRemAddress: The remote IP address for this TCP connection.

tcpConnRemPort: The remote port number for this TCP connection.

A TCP Connectin Table could also includes the following information:

Sequence Number: the sequence number in the packet header the sender is going to send.

Acknowledgement Number: the sequence number in the packet header the receiver is hoping to receive.

Idle time: the time that the tcp connection table hasn't been updated.

4.1.1. If TCP Connection Table isn't migrated

Assuming TCP Connection Table item is generated for VM1 on Firewall-A, the information is as follows:

tcpConnState == Established

tcpConnLocalAddress == 10.138.3.1

tcpConnLocalPort == 1234

tcpConnRemAddress: == 192.167.22.3

tcpConnRemPort == 4321

Assuming VM1 is migrated to Server4 under Switch4, without TCP

Connection Table migration. In order to keep the running service uninterrupted, the IP Address of VM1 will keep unchanged. The packets belonging to this TCP Connection will continue coming, which will pass Firewall-B, instead of Firewall-A. Because there is no TCP Connection Table for VM1 on Firewall-B, the following packets belonging to the TCP Connection will be dropped, hence the running service is broken down.

4.1.2. If TCP Connection Table is not accurately migrated

VM migration needs a period to finish memory and register copy. Fig.3 shows the VMware VMotion process. There are three points we should pay attention to.

Pre-copy period: VM begins to prepare for migration. In this period, VM pre-copy memory state to the new VM on destination device. The original VM is still power on and service is still running, which means the memory and register could keep changing. The new VM is power-on.

VM not running period: The end phase of memory copy. In this period, original VM stop running service, the memory will not change. Original VM finish copying the rest changed memory and register to new VM. New VM is still power-on.

VM power-off point: After original VM receives the OK message from new VM, it turns off the power, and meanwhile the new memory starts to run.

We can see that it's unrealistic to make a 'zero delay' VM migration, because there is at least about 1 second period (VM not running period) when neither VM is running.

Assuming there is a NM can GET and SET dynamic information. The NM GET dynamic information at Time A, and finish SET at Time B. At Time A, the Sequence Number of VM1's TCP Connection is 99. After NM GET dynamic information, VM1's TCP connection keeps transferring packets and Sequence Number increase to 110, until VM not running period begins. During VM not running period, no TCP packet is acknowledged by VM1, so the Sequence Number is 110. At Time B, the destination Firewall is SET by Sequence Number 99. When new VM1 starts, the packets belonging to VM1's TCP connection comes to Firewall-B with Sequence Number 111. Since the receiving Sequence Number doesn't equal to the Acknowleadge Sequence Number of Firewall-B, this packet will be dropped and the running service is broken down.

```
----------------------------------------------------------------------------
   ^     ||          |   Power-on destination VM                          ^
   |     ||          |                                                    |
   |     ||--------->| ---------------------------
VMotion
  VM     ||          |    Pre-copy memory state
abortable
 running ||--------->|
|------------>Time A
on source||          |                                                    | ^
   |     ||--------->|                                                    |
Dynamic Information
   V     ||          |                                                    | V
keep changing
--------------------------------------------------------------------------|----
   ^     |--------->| Checkpoint-save src VM and transfer state           |
   |     |          |-------------------------------------------          |
  VM not |--------->| transfer remaining changed memory and
|------------->Time B
 running |          | checkpoint-restore dst VM                           V
   |     |          |-----------------------------------------------------
 ~1sec   |<---------| send restore OK
   V     |          |
----------------------------------------------------------------
         |          ||  Power-off source VM
         |          ||  VM running on destination
```

Fig.3 VMware VMotion process

4.2. Cumulated data

One example for cumulated data is unfinished TCP Connection established by a specific VM. In order to avoid TCP SYN flood, a network device may control the unfinished TCP Connection established by a single end host by setting a threshold. For example, the NM set the threshold to 5, and VM1 has established 3 unfinished TCP connections. If the cumulated TCP connection number isn't migrated to destination devices, the destination device will allow VM1 to establish up to 5 unfinished TCP connections. For the single destination device, the unfinished TCP connections established by VM1 is under control, but for the whole DC, VM1 has established 8 unfinished TCP connections. So VM1 has consume more resoureces than allowed.

4.3. Dynamic ACLs

Assuming all traffic is denied unless the end host is authorized and authenticated. VM1 has been authenticated on source device and a dynamic ACL has been generated to allow VM1's traffic to pass. If VM1 migrates to destination device without the dynamic ACL, the destination device will drop VM1's traffic, because VM1 is an unathenticated end host for it. So in this case, the dynamic ACL needs to migrate with VM.

4.4. DHCP Snooping

Assuming source device is DHCP Snooping Enabled and a DHCP Snooping mapping item is created for VM1: (IP-VM1: MAC-VM1). This mapping is created dynamically by listening to DHCP Response message. If VM1 migrate to destination device, since the IP Address of VM1 doesn't change, there is no DHCP Request sent by VM1. So on destination device, there is (IP-VM1: MAC-VM1) mapping, all traffic from VM1 will be dropped. So DHCP Snooping mapping item need to migrate to destination device.

4.5. Multicast Membership

Multicast membership is similar to DHCP Snooping. Multicast membership is created on ports by listening to IGMP membership report messages. If VM1 migrates to destination, VM1 will not send IGMP membership report until next IGMP General Query. Before that, VM1 may not be able to recevie Multicast packets since network devices on and above destination devices don't knwo VM1's Multicast membership and don't forwarding the Multicast packets to VM1.

5. Existing network management protocol and the limitations

RFC3535 introduces many Network Management architectures and protocols. Basically, there are two kinds of architectures: network element oriented (SNMP and NETCONF) and Policy based (COPS-PR). In this section, we will introduce why these NMPs can not resolve the problem described in this draft.

5.1. Limitations

We analyze the problem described above into two aspects. One is Policies migration and the other is dynamic information migration.

5.1.1. For Policies migration

Existing NMP could be used to migrate Policies from source device to destination device. But we still need to face some questions:

Is device-oriented NMP suitable for policies migration?

Is C/S based NMP suitable for DC management?

How does NM know the source and destination device?

Do we need an automatic policies migration mechanism?

5.1.2. For Dynamic Information Migration

Currently, NMP is not used to configure dynamic information.

And, as Fig.3 shows, if we fail to begin migrating dynamic information at appropriate time (the time during VM not running period), the running services will be interrupted. For example, if dynamic information is migrated before 'VM not running period', dynamic information is inaccurate and the running services may be broken down when new VM restarts. In order to make accurate migration and keep running service uninterrupted, we need to know the exact timing for migration.

6. Security Considerations

The policies and dynamic information described above are all about security.

7. Acknowledgments

I would like to thank the following people for contributing to this draft: Ning Zong, David harrington, Linda dunbar, Susan Hares, Serge manning, Barry Leiba, Jiang xingfeng, Song Wei, Robert Sultan, and many others.

8. Normative Reference

[RFC2119] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", March 1997.

Author's Address

Gu Yingjie
Huawei
No. 101 Software Avenue
Nanjing, Jiangsu Province 210001
P.R.China

Phone: +86-25-56624760
Fax: +86-25-56624702
Email: guyingjie@huawei.com