Network Working Group
Internet-Draft
Intended status: Standards Track
Expires: September 6, 2012

Y. Gu
Huawei
C. Li
China Mobile
K. Li
China Telecom
Z. Zhuo
Ruijie Network
D. Zhang
Huawei

Mar 5, 2012

# State Migration
# draft-gu-opsawg-policies-migration-02

Abstract

In accompany with the migration of a Virtual Machine (VM), state associated with the VM located on the Hypervisors and the network side devices (e.g., Firewalls) need to be updated in order to guarantee that the services executed on the migrated VM will not be disrupted. VM vendors have their own ways to migrate VM's state on Hypervisors, and so this is out the scope of this draft.This draft introduces the background of state migration on network devices using several application scenarios and tries to specify a clear scope for the future standardization work on state migration on network devices.

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). Note that other groups may also distribute working documents as Internet-Drafts. The list of current Internet-Drafts is at http://datatracker.ietf.org/drafts/current/.

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."

This Internet-Draft will expire on September 2, 2012.

Table of Contents

## 1. Introduction

Under the assistance of a VM live migration mechanism, a VM can move across physical servers, racks, subnets, or even DCs (Data Centers). During the migration, the services executed on the VM should not be significantly interrupted. In order to achieve objective, not only the hypervisor that the VM moves to but also any related devices on the network side must update their state cooperatively. For instance, assume that a VM executed on a server is under the protection of a Firewall A. The VM creates several connections with external clients. The state information associated with the connections is generated and maintained in the session table of A. Any inbound packets to the VM will be checked against the session table, and the packets that doesn't match any recorded connection will be discarded. Now, the VM migrates to another rack, which is under the management of Firewall B. Because B has no knowledge about the connections created by the VM, any packets belonging to the connections will be discarded by B. As a result, the connections will finally be broken. A more detailed description can be found in Section 4.

There are various middleboxes embedded in the network (e.g., Firewalls, IPSes, IDSes, NATs and etc) which may need state migration. To benefit the discussion, Firewalls are used as an example to analyze the state updating issues caused by VM live migration.

## 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].

Virtual Machine (VM), A completely isolated operation system which is installed by software on a normal operation system. An normal operation system can be virtualized into several VM.

Firewall (FW), A policy based security device, typically used for restricting access to/from specific devices and applications.

## 3. States On Firewalls

Basically, there are In this draft, we consider two complimentary physical Firewall deployment solutions in DCs.

- Centralized Deployment. In this case, typically, a pair of centralized powerful Firewalls are deployed at WAN connect point. Any traffic, even the traffic between VMs within the same LAN, need to pass though the Firewall. Centralized deployed Firewalls need to be reliable and capable to deal with large amounts of traffic.

- Distributed deployment. In this case, Firewalls are deployed at aggregation switches or lower access switches in a distributed way. The goal of this kind of distributed deployed Firewall is to off load the huge workload on centralized Firewall. This case is especially reasonable for large layer 2 network with tens even hundreds of thousands of Virtual Machines.Note that both the centralized and distributed solutions can co-exist in DCs.

In either solution, the Firewalls need to generate and maintain the state (e.g., security policies and connection information) to process packets. In the remainder of this section, three types of state information supported by most stateful Firewalls are introduced respectively.

## 3.1. Session Table

A stateful Firewall needs to establish a record for every connection associated a host (which could a physical server or a VM) under its protection. Typically, a connection record should contain following parameters.

| Item | Interpretation |
|---|---|
| Src-IP | Source IP Address of the connection |
| Dst-IP | Destination IP Address of the connection |
| Src-Port | Source Port Number used to establish the session |
| Dst-Port | Destination Port Number used to establish the session |
| Protocol | Protocol type, e.g. TCP |
| Status | The status of a connection, e.g. Established or SYN_ACK. |
| Interface | The inbound interface of the connection |
| Creation-Time | The time that the session is created |
| Last-heard | The last time that a packet belong to this connection is received by Firewall |

Of course, not all of the information in the session table on a source Firewall is meaningful for the destination Firewall. For instance, the Interface parameter is meaningless to destination

Firewall. Thus, not all of the information need to be migrated to destination Firewall in state migration,.

### 3.2. Cumulative Data

In order to deal with DOS (Deny of Service ) attacks, Firewalls need to cumulate certain types of data. For instance, assume there are both individual clients and enterprise servers in a DC. A malicious client tries to perform a SYN Flooding attack to degrade the performance of a server in the same DC. That is, the client keeps sending SYN message to the server with a un-reasonably high rate. To detect this attack, Firewalls in the DC need to cumulate the SYN message sent from the clients. If a Firewall finds that the frequency of SYN message sent by a client exceed a pre-defined rate, the IP address of this client will be drawn into a black list by the Firewall.

### 3.3. Access Control List

Static Access Control List (ACL) can be configured on Firewall to filter packets between internal and external network, or between different security zones. Usually 5-tuple, VLAN information and interface information are designated in ACL. Static ACL is not generated dynamically based on flow, so there could be other way to re-configure ACL except for state migration, e.g. manually configure the static ACL on destination Firewall before VM migration.

## 4. Scenarios for Migration of States on Firewall

This section demonstrates the necessity of migrating Firewall states when a VM is moved to a new place using several real-life scenarios.

### 4.1. VM Migration between different DCs

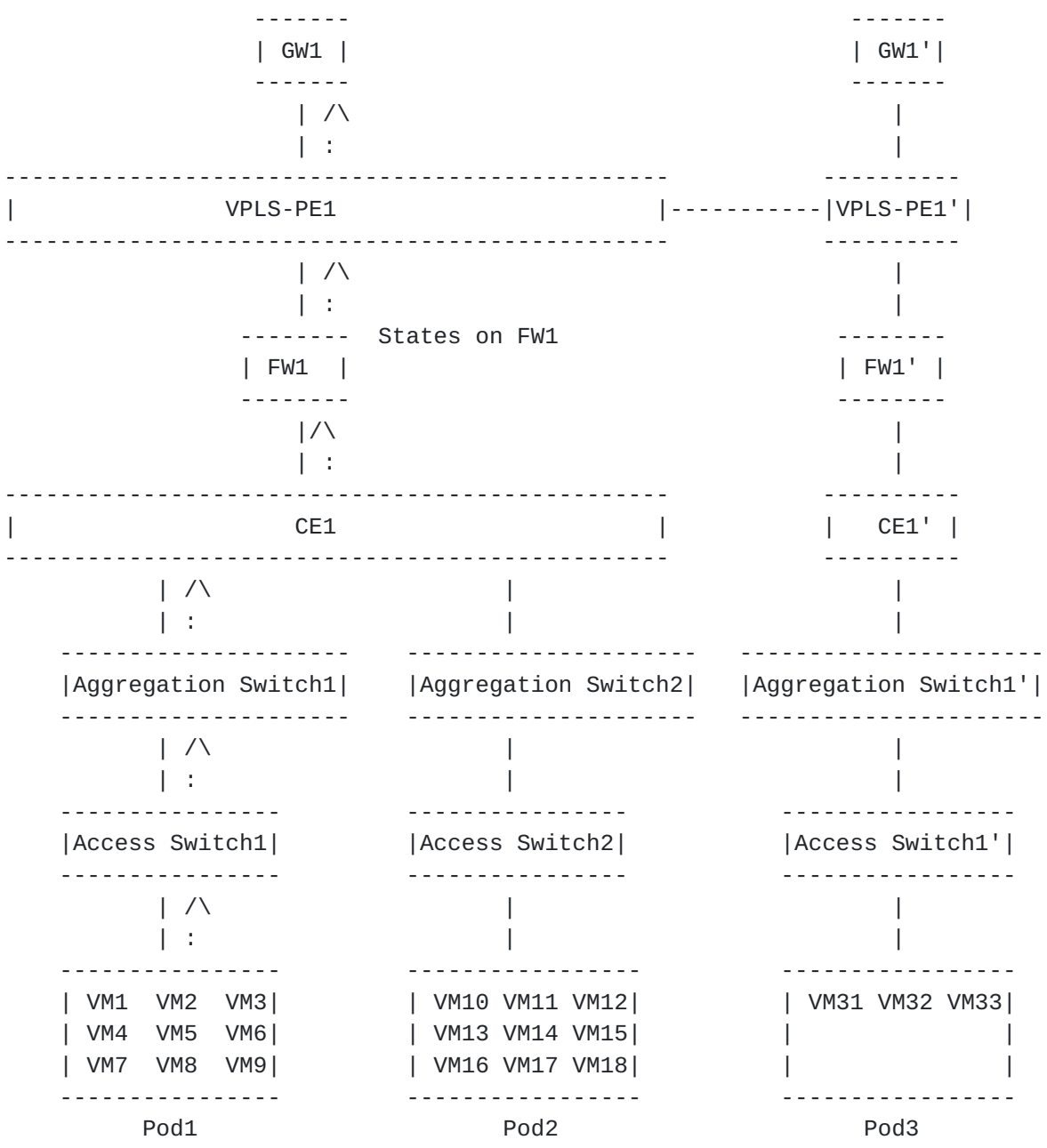
China Telecom has several geographically distributed DCs. These DCs were built several years ago and can only provide limited computing and storage services. Because the accelerating urbanization in China, the places where the DCs locate has become downtown areas, and it is very expensive to extend their scales, which cannot be afforded by China Telecom. As a result, sometimes, China Telecom has to use computing and storage resources of multiple DCs to support a single service (e.g., Instance messaging or search engine). Such combination of DCs should be seamless so that the service provider can work in the same way as they work in a single DC, even when its VMs need to migrate from one DC to another.

Figure 1 illustrates an example in which a DC provider has two DCs on

different locations. One is at City A; the other is at City B, which is 30 kilometers away from City A. Assume that the physical distance and network bandwidth between City A and B satisfy the requirements of VM live migration. Two DCs are interconnected by, for example, VPLS (, [RFC4761][RFC4762]) or OTV ([OTV]). Each DC has a pair of centralized Firewalls. For simplicity, each DC only has only one Firewall as shown the figure.

In the figure, Firewalls are shown separately from CEs. But in real-life deployment, they could be integrated within CEs.

At the very beginning, VMs are evenly created on Pod1 and Pod2. As time elapses, the overload imposed on Pod1 and Pod2 increases. Now, for some reasons (e.g., hardware errors), Pod2 need to be switched off and Pod1 does not have sufficient capability to support the VMs on Pod2. In order to guarantee SLA, Pod3 is created and some of the VMs on Pod1 and Pod2 are migrated to Pod3, and the running service must be kept during the migration.




VM1: 10.1.1.1 VLAN 1

.... VM Traffic

Figure 1: Example architecture

One way to achieve this is to migrate VMs but let the flow still pass through FW1, VPLS-PE1 and GW1. But in that case, the traffic is not optimized, which means more packet delay and more bandwidth consumption. And another essential problem is that FW1' would drop the packets since FW1' cannot map the flow to any recorded session.

So another way is to migrate states on FW1 to FW1' and let the flow pass through FW1', VPLS-PE1' and GW1'.

The DCs could be in different subnets and we need to make sure the IP address of VM be kept unchanged during the VM live migration. Some existing work, e.g. in IETF LISP WG ([LISP]), can make VM migration between subnets feasible. In State Migration, we won't try to define technologies that can be used to keep VM's IP address unchanged while migrating between subnets.

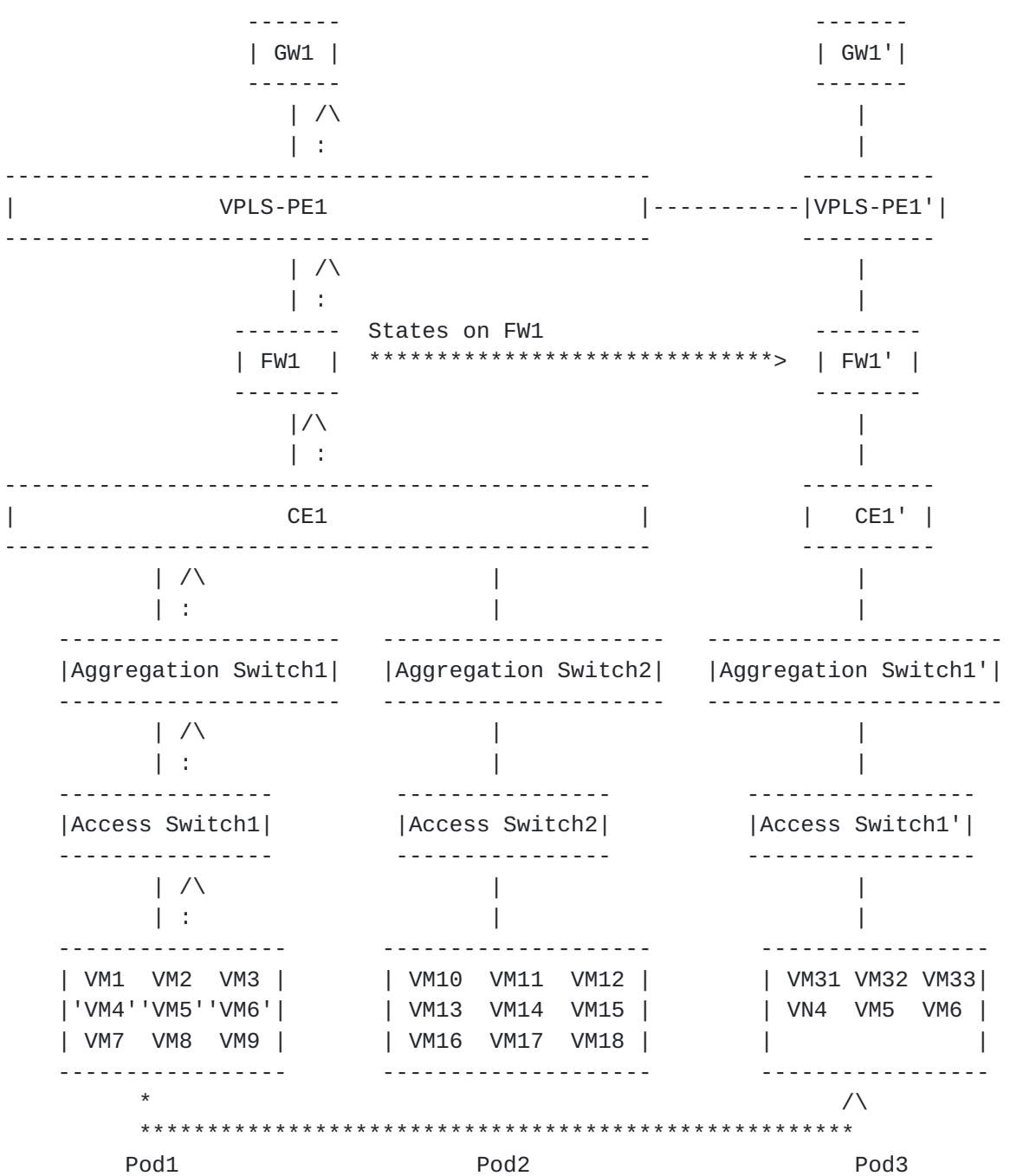



Figure 2: VM and State Migration stage

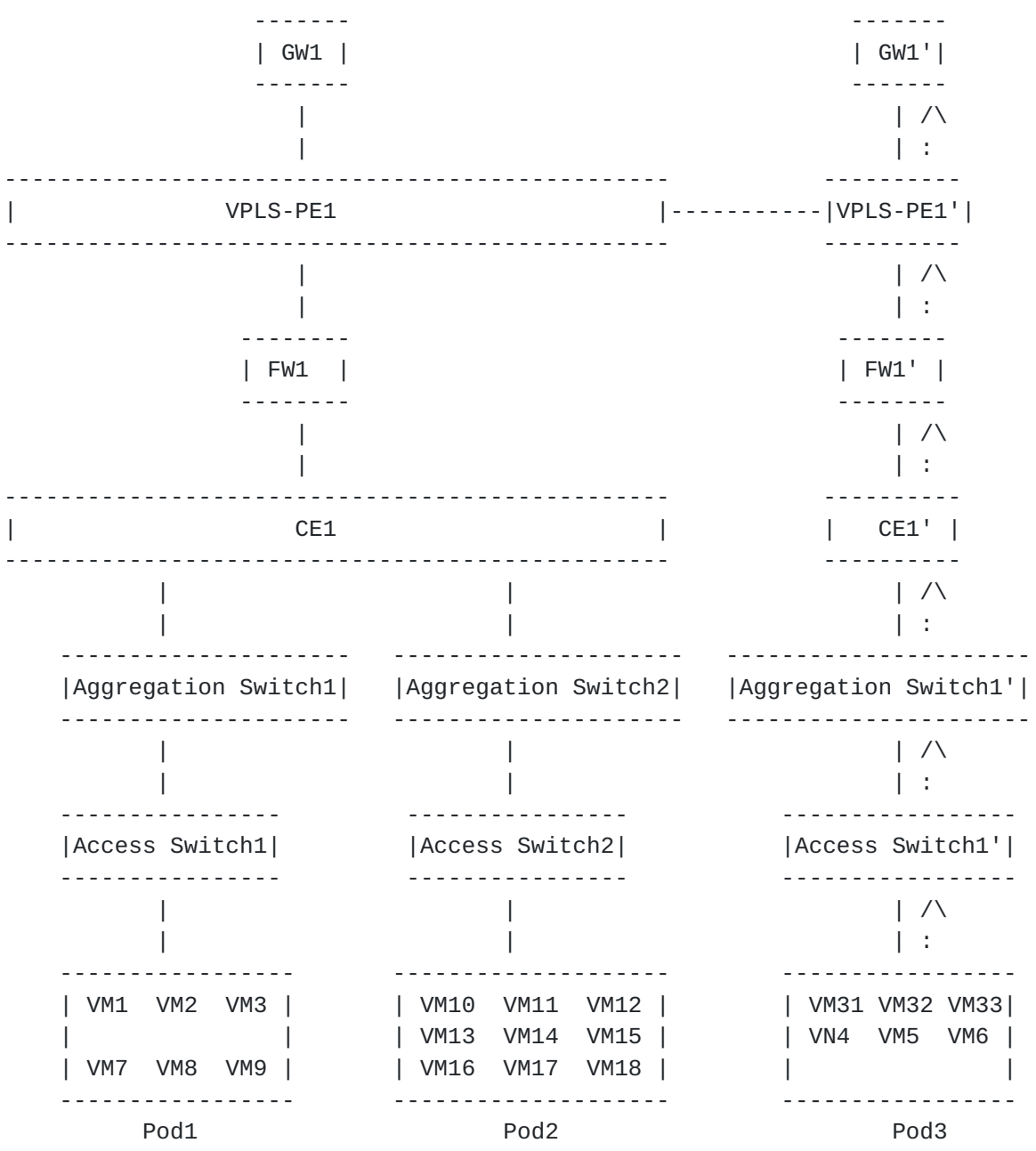
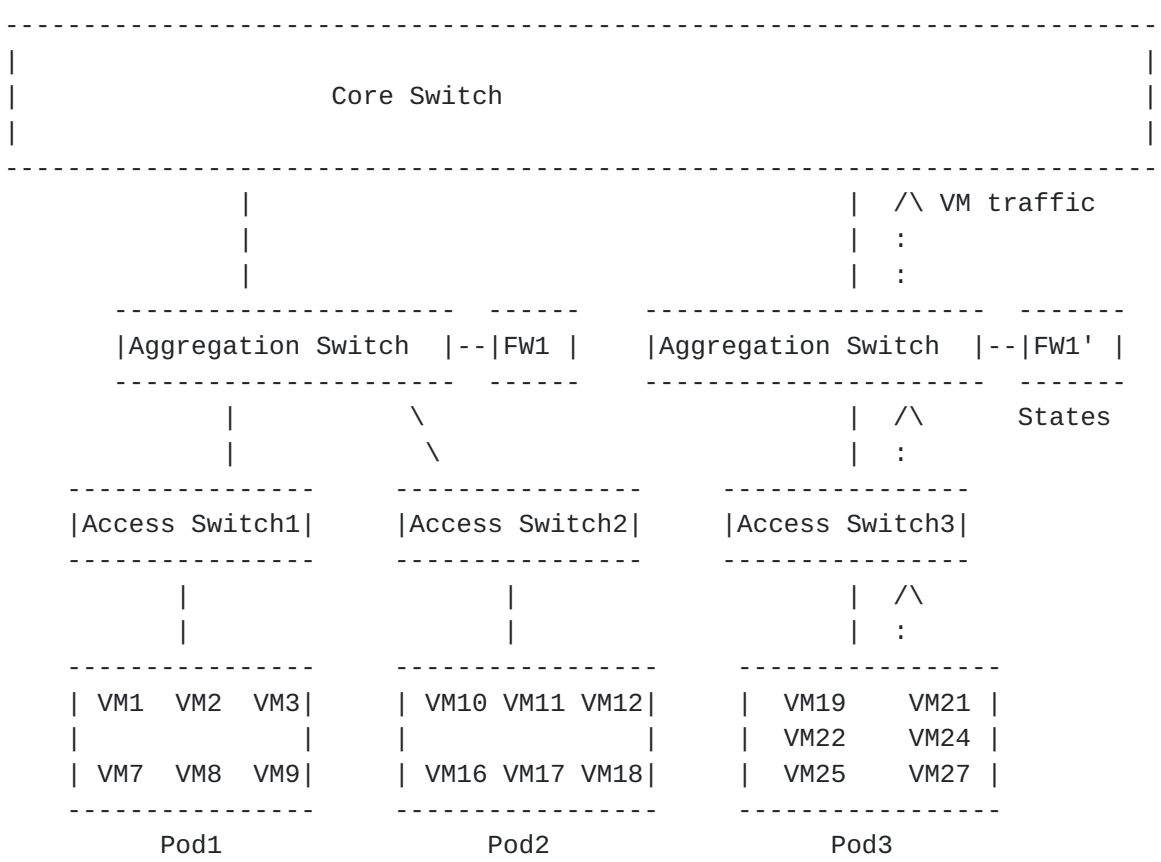



Figure 3: VM Migration Completion

## 4.2. VM Migration under Distributed Deployed Firewalls

In a DC with distributed deployed Firewalls on Aggregation Switches, assuming an enterprise customer lease hundreds of physical servers, and each physical server carries 10 plus Virtual Machines (VM). The VMs provide VDI service to employees in remote branch. At day time, the VMs are evenly deployed on each Pod.




.... VM Traffic

Figure 4: VDI service in DC

During nighttime, most of the VMs are shut down. In order to save energy, the VMs still active are migrated to a few physical servers and other physical servers are switched off. In this case, the states on FW1 need to be updated under the assistance of FW1, otherwise the running service on migrating VM will be disrupted.

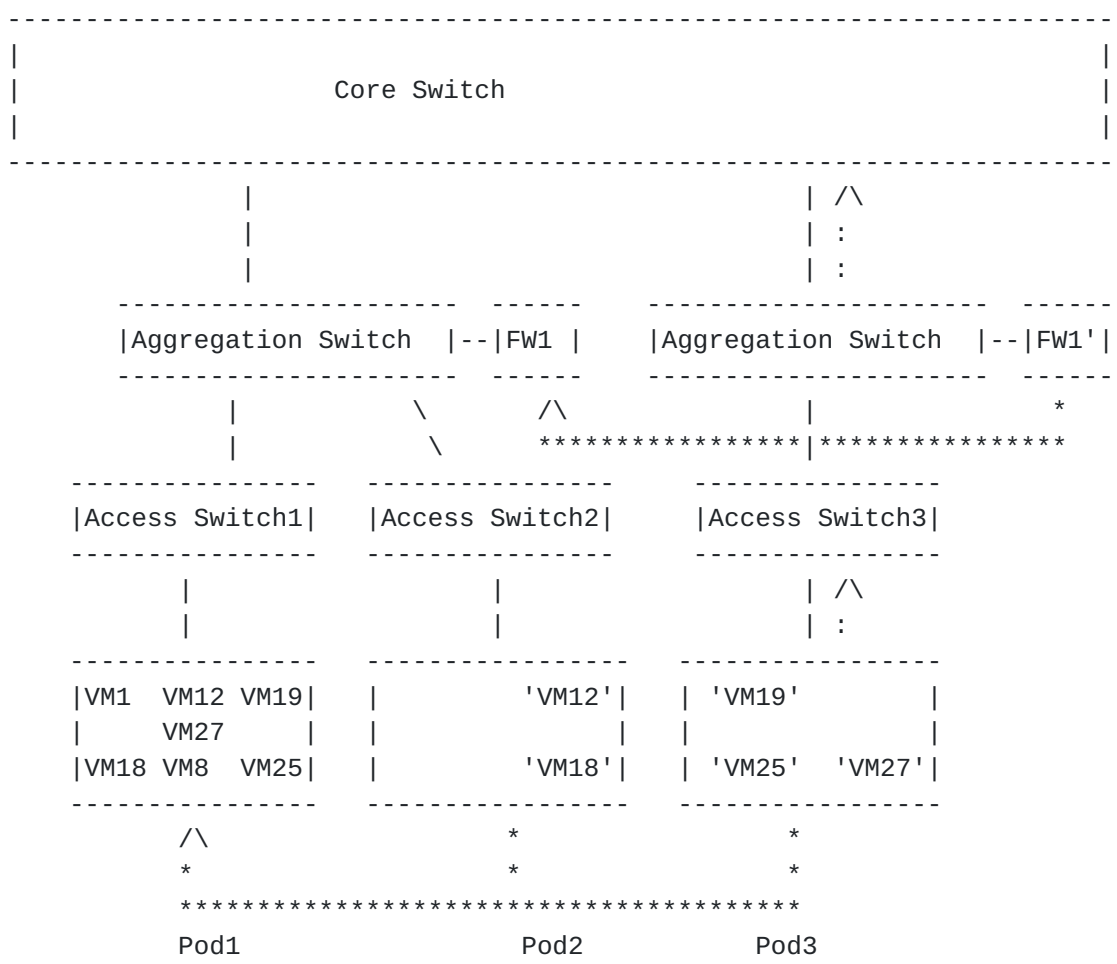



Figure 5: VM and State migration

## 5. Active-Active or VM Migration

In previous WG discussion on State Migration, there is suggestions to use alternative mechanism to migrate user's service instead of VM Migration, and hence no necessity to do State Migration. The suggested mechanism is to run active-active VMs on both old and new locations. The new services related to the old VM is directed to the new VM and shut down the old VM while all the existing services are finished. In fact, this is not within the scope of State Migration intention. But since State Migration proposal is based on the precondition that VM is migrating, the authors would like to list some reasons to clarify why VM Migration is necessary.

- Long lived connections: Some applications may establish connections with virtual machines and the connections are kept for a long time until the applications disconnect. An example is the HTTP persistent connection technique,. The idea is to generate a pool of TCP connections. Instead of opening a new one for every single request/response pair, a TCP connection may be re-used to send and receive multiple HTTP requests/ responses.Unused TCP connections are then stored in the pool. This solution is also widely used for improving performance of network systems (e.g., distributed database systems). The above examples make it very clear that it's unexpected how long the existing services will be alive, that is it's unexpected how long the old VM need to be kept active.

- Hardware failure: While there is hardware problem, it may not leave enough time to wait until all the existing services are finished. For example, there may be a power shortage in a particular area, and the DC providers have to move their VMs, services and data to another location in limited time. In this case, active-active mechanism may cause services disruption, if the existing services on old VMs keep running.

Active-active and VM migration are both useful for particular scenarios. In State Migration concept, we only consider the scenarios where VM Migration is a better choice.

## 6. Scope

For the first stage, SAMI (StAte MIgration) only do research on and develop solution for Session Table migration on Firewall. But the solution should be extensible to enable migration of other states that we may find in the future which is necessary for VM live migration.

In SAMI, we require that the network, wherein VM live migration happens, must satisfy the basic network condition requirements raised by Virtualization Platform vendors. Examples of network condition requirements include reasonable geographic distance, higher than minimum bandwidth and acceptable packet delay. The requirements could vary from one Virtualization Platform vendor to another, and SAMI won't define the requirements. Another important requirement is that the IP address of VM must be kept unchanged during VM live migration, but SAMI won't work on these technologies or make any suggestions on these technologies. When we do research on SAMI, we assume there are technologies to guarantee IP address unchanged during VM live migration.

To be more specific on scope, we list some of scenarios that SAMI will consider. All of these scenarios, are in scope for now, and revision will be made when we get further achievements during research.

1) State migration within the same DC, same subnet and same administration domain;

2) State migration within the same DC and same administration domain, but between different subnets;

3) State migration between DCs, which is under different administration domains and different subnets;

Existing IETF work should be re-used to resolve the SAMI problem as much as possible. Only when there is no existing IETF work can use, to achieve State migration, a new mechanism needs to be developed. [GapAnalysis]makes a brief analysis on existing related IETF work, including MIDCOM, ForCES and PCP. And more will be introduced later.

## 7. Security Considerations

The states described above are all about security. Besides, we need to be careful to avoid poisoned states from untrusted source. That means no matter how the states are migrated, authentication and verification are required.

## 8. Acknowledgments

The authors 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.

## 9. Author List

Jingtao Yang

yangjingtao@huawei.com

Huiyang Xu

xuhuiyang@chinamobile.com

Yongbin Fan

fanyb@gsta.com

Ming Liu

lium@ruijie.com.cn

## 10. References

### 10.1. Normative Reference

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

[RFC3303] Srisuresh, P., Kuthan, J., Rosenberg, J., Molitor, A., and A. Rayhan, "Middlebox communication architecture and framework", August 2002.

[RFC4761] Kompella, K. and Y. Rekhter, "Virtual Private LAN Service (VPLS) Using BGP for Auto-Discovery and Signaling", Jan. 2007.

[RFC4762] Lasserre, M. and V. Kompella, "Virtual Private LAN Service (VPLS) Using Label Distribution Protocol (LDP) Signaling", Jan. 2007.

### 10.2. Informative Reference

[GapAnalysis] Wang, D. and Y. Gu, "I-D.wang-opsawg-policy-migration-gap-analysis", 2011.

[Vmotion_between_DCs] VMware, "VMotion between Data Centers--a VMware and Cisco Proof of Concept, (http://http://blogs.vmware.com/networking/2009/06/vmotion-between-data-centersa-vmware-and-cisco-proof-of-concept.html)", June 2009.

[OTV] Grover, H., Rao, D., and D. Farinacci, "Overlay Transport Virtualization", July 2011.

[LISP] "Location/ID separation protocol, http://tools.ietf.org/wg/lisp/".

Authors' Addresses

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

Email: guyingjie@huawei.com

Li Chen
China Mobile

Email: lichenyj@chinamobile.com

Li Kai
China Telecom

Email: leekai@ctbri.com.cn

Zhuo Zhiqiang
Ruijie Network

Email: zhuozq@ruijie.com.cn

Zhang Dacheng
Huawei

Phone: 86-01060610033
Fax:
Email: zhangdacheng@huawei.com