Virtual Machine Mobility Solutions for L2 and L3 Overlay Networks
draft-ietf-nvo3-vmm-05

Abstract

This document describes virtual machine mobility solutions commonly used in data centers built with overlay-based network. This document is intended for describing the solutions and the impact of moving VMs (or applications) from one Rack to another connected by the Overlay networks.

For layer 2, it is based on using an NVA (Network Virtualization Authority) - NVE (Network Virtualization Edge) protocol to update ARP (Address Resolution Protocol) table or neighbor cache entries after VM (virtual machine) moves from Old NVE to the New NVE. For Layer 3, it is based on address and connection migration after the move.

Status of this Memo

This Internet-Draft is submitted in full conformance with the provisions of BCP 78 and BCP 79.

This Internet-Draft is submitted in full conformance with the provisions of BCP 78 and BCP 79. This document may not be modified, and derivative works of it may not be created, except to publish it as an RFC and to translate it into languages other than English.
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 February 22, 2009.

Copyright Notice

Copyright (c) 2019 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. Code Components extracted from this document must include Simplified BSD License text as described in Section 4.e of the Trust Legal Provisions and are provided without warranty as described in the Simplified BSD License.

Table of Contents

1. Introduction...................................................3
2. Conventions used in this document..............................4
3. Requirements...................................................5
4. Overview of the VM Mobility Solutions.........................6
   4.1. VM Migration in Layer 2 Network........................6
1. Introduction

This document describes the overlay-based data center networks solutions in supporting multitenancy and VM (Virtual Machine) mobility. Many large DCs, especially Cloud DCs, host tasks (or workloads) for multiple tenants, which can be multiple departments of one organization or multiple organizations. There is communication among tasks belonging to one tenant and communications among tasks belonging to different tenants or with external entities.

Server Virtualization, which is being used in almost all of today’s data centers, enables many VMs to run on a single physical computer or compute server sharing the processor/memory/storage. Network connectivity among VMs is provided by the network virtualization edge (NVE) [RFC8014]. It is highly desirable [RFC7364] to allow VMs to be moved dynamically (live, hot, or cold move) from one server to another for dynamic load balancing or optimized work distribution.

There are many challenges and requirements related to VM mobility in large data centers, including dynamic attaching/detaching VMs to/from Virtual Network Edges (VNEs). Retaining IP addresses after a move is a key requirement [RFC7364]. Such a requirement is needed in order to maintain existing transport connections. In traditional Layer-3 based networks, retaining IP addresses after a move is generally not recommended because the frequent move will cause non-aggregated IP addresses (a.k.a. fragmented IP addresses), which introduces complexity in IP address management.

In view of many VM mobility schemes that exist today, there is a desire to document comprehensive VM mobility solutions that cover...
both IPv4 and IPv6. The large Data Center networks can be organized as one large Layer-2 network geographically distributed in several buildings/cities or Layer-3 networks with large number of host routes that cannot be aggregated as the result of frequent move from one location to another without changing their IP addresses. The connectivity between Layer 2 boundaries can be achieved by the network virtualization edge (NVE) functioning as Layer 3 gateway routing across bridging domain such as in Warehouse Scale Computers (WSC).

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] and [RFC8014].

This document uses the terminology defined in [RFC7364]. In addition, we make the following definitions:

VM: Virtual Machine

Tasks: Task is a program instantiated or running on a virtual machine or container. Tasks in virtual machines or containers can be migrated from one server to another. We use task, workload and virtual machine interchangeably in this document.

Hot VM Mobility: A given VM could be moved from one server to another in running state.

Warm VM Mobility: In case of warm VM mobility, the VM states are mirrored to the secondary server (or domain) at a predefined (configurable) regular intervals. This reduces the overheads and complexity, but this may also lead to a situation when both servers may not contain the exact same data (state information)
Cold VM Mobility: A given VM could be moved from one server to another in stopped or suspended state.

Old NVE: refers to the old NVE where packets were forwarded to before migration.

New NVE: refers to the new NVE after migration.

Packets in flight: refers to the packets received by the Old NVE sent by the correspondents that have old ARP or neighbor cache entry before VM or task migration.

Users of VMs in diskless systems or systems not using configuration files are called end user clients.

Cloud DC: Third party data centers that host applications, tasks or workloads owned by different organizations or tenants.

3. Requirements

This section states requirements on data center network virtual machine mobility.

Data center network should support both IPv4 and IPv6 VM mobility.

Virtual machine mobility should not require changing their IP addresses after the move.

There is "Hot Migration" with transport service continuing, and there is a "Cold Migration" with transport service restarted, i.e. stop the task running on the Old NVE and move to the New NVE before restart as described in the Task Migration.

VM mobility solutions/procedures should minimize triangular routing except for handling packets in flight.

VM mobility solutions/procedures should not need to use tunneling except for handling packets in flight.
4. Overview of the VM Mobility Solutions

Layer 2 and Layer 3 mobility solutions are described respectively in the following sections.

4.1. VM Migration in Layer 2 Network

Being able to move VMs dynamically, from one server to another, makes it possible for dynamic load balancing or work distribution. Therefore, it is highly desirable for large scale multi-tenants data centers.

In a Layer-2 based approach, VM moving to another server does not change its IP address, but this VM is now under a new NVE, previously communicating NVEs will continue to send their packets to the Old NVE. To solve this problem, Address Resolution Protocol (ARP) cache in IPv4 [RFC0826] or neighbor cache in IPv6 [RFC4861] in the NVEs need to be updated. NVEs need to change their caches associating the VM Layer-2 or Medium Access Control (MAC) address with the NVE’s IP address. Such a change enables NVEs to encapsulate the outgoing MAC frames with the current target NVE address. It may take some time to refresh ARP/ND cache when a VM is moved to a New NVE. During this period, a tunnel is needed so that Old NVE can forwards packets destined to the VM to the New NVE.

In IPv4, the VM immediately after the move should send a gratuitous ARP request message containing its IPv4 and Layer 2 MAC address in its new NVE. This message’s destination address is the broadcast address. Old NVE receives this message. Both Old and New NVEs should update VM’s ARP entry in the central directory at the NVA, to update its mappings to record the IPv4 address & MAC address of the moving VM along with the new NVE IPv4 address. An NVE-to-NVA protocol is used for this purpose [RFC8014].

Reverse ARP (RARP) which enables the host to discover its IPv4 address when it boots from a local server [RFC0903], is not used by VMs because the VM already knows its IPv4 address. Next, we describe a case where RARP is used.

There are some vendor deployments (diskless systems or systems without configuration files) wherein VM users, i.e. end-user clients ask for the same MAC address upon migration. This can be achieved by the clients sending RARP request message which carries the old MAC address looking for an IP address allocation. The server, in this case the new NVE needs to communicate with NVA, just like in the gratuitous ARP case to ensure that the same IPv4
address is assigned to the VM. NVA uses the MAC address as the key in the search of ARP cache to find the IP address and informs this to the new NVE which in turn sends RARP reply message. This completes IP address assignment to the migrating VM.

Other NVEs communicating with this VM could have the old ARP entry. If any VMs in those NVEs need to communicate with the VM attached to the New NVE, old ARP entries might be used. Thus, the packets are delivered to the Old NVE. The Old NVE MUST tunnel these in-flight packets to the New NVE.

When an ARP entry for those VMs times out, their corresponding NVEs should access the NVA for an update.

IPv6 operation is slightly different:

In IPv6, after the move, the VM immediately sends an unsolicited neighbor advertisement message containing its IPv6 address and Layer-2 MAC address to its new NVE. This message is sent to the IPv6 Solicited Node Multicast Address corresponding to the target address which is the VM's IPv6 address. The NVE receiving this message should send request to update VM’s neighbor cache entry in the central directory of the NVA. The NVA’s neighbor cache entry should include IPv6 address of the VM, MAC address of the VM and the NVE IPv6 address. An NVE-to-NVA protocol is used for this purpose [RFC8014].

Other NVEs communicating with this VM might still use the old neighbor cache entry. If any VM in those NVEs need to communicate with the VM attached to the New NVE, it could use the old neighbor cache entry. Thus, the packets are delivered to the Old NVE. The Old NVE MUST tunnel these in-flight packets to the New NVE.

When a neighbor cache entry in those VMs times out, their corresponding NVEs should access the NVA for an update.

4.2. Task Migration in Layer-3 Network

Layer-2 based data center networks become quickly prohibitive because ARP/neighbour caches don’t scale. Scaling can be accomplished seamlessly Layer-3 data center networks by just giving each virtual network an IP subnet and a default route that points to NVE. This means no explosion of ARP/neighbour cache in VMs and NVEs (just one ARP/neighbour cache entry for default
route) and there is no need to have Ethernet header in encapsulation [RFC7348] which saves at least 16 bytes.

Even though the term VM and Task are used interchangeably in this document, the term Task is used in the context of Layer-3 migration mainly to have slight emphasis on the moving an entity (Task) that is instantiated on a VM or a container.

Traditional Layer-3 based data center networks require IP address of the task to change after moving because the prefixes of the IP address usually reflect the locations. It is necessary to have an IP based VM migration solution that can allow IP addresses staying the same after moving to different locations. The Identifier Locator Addressing or ILA [I-D.herbert-nvo3-ila] is one of such solutions.

Because broadcasting is not available in Layer-3 based networks, multicast of neighbor solicitations in IPv6 would need to be emulated.

Cold task migration, which is a common practice in many data centers, involves the following steps:

- Stop running the task.
- Package the runtime state of the job.
- Send the runtime state of the task to the New NVE where the task is to run.
- Instantiate the task’s state on the new machine.
- Start the tasks for the task continuing from the point at which it was stopped.

Address migration and connection migration in moving tasks or VMs are addressed next.

4.2.1. Address and Connection Migration in Task Migration

Address migration is achieved as follows:

- Configure IPv4/v6 address on the target Task.
- Suspend use of the address on the old Task. This includes handling established connections. A state may be established to drop packets or send ICMPv4 or ICMPv6 destination unreachable message when packets to the migrated address are received.
- Push the new mapping to VM. Communicating VMs will learn of the new mapping via a control plane either by participation in a protocol for mapping propagation or by getting the new mapping from a central database such as Domain Name System (DNS).

Connection migration involves reestablishing existing TCP connections of the task in the new place.

The simplest course of action is to drop TCP connections across a migration. It the migrations are relatively rare events, it is conceivable that TCP connections could be automatically closed in the network stack during a migration event. If the applications running are known to handle this gracefully (i.e. reopen dropped connections) then this may be viable.

More involved approach to connection migration entails pausing the connection, packaging connection state and sending to target, instantiating connection state in the peer stack, and restarting the connection. From the time the connection is paused to the time it is running again in the new stack, packets received for the connection could be silently dropped. For some period of time, the old stack will need to keep a record of the migrated connection. If it receives a packet, it can either silently drop the packet or forward it to the new location, similarly as in Section 5.

5. Handling Packets in Flight

The Old NVE may receive packets from the VM’s ongoing communications and these packets should not be lost, and they should be sent to the New NVE to be delivered to the VM. The steps involved in handling packets in flight are as follows:

Preparation Step: It takes some time, possibly a few seconds for a VM to move from its Old NVE to a New NVE. During this period, a tunnel needs to be established so that the Old NVE can forward packets to the New NVE. Old NVE gets New NVE address from NVA in the request to move the VM. The Old NVE can store the New NVE address for the VM with a timer. When the timer expired, the entry for the New NVE for the VM can be deleted.

Tunnel Establishment - IPv6: Inflight packets are tunneled to the New NVE using the encapsulation protocol such as VXLAN in IPv6.
Tunnel Establishment - IPv4: Inflight packets are tunneled to the New NVE using the encapsulation protocol such as VXLAN in IPv4.

Tunneling Packets - IPv6: IPv6 packets received for the migrating VM are encapsulated in an IPv6 header at the Old NVE. New NVE decapsulates the packet and sends IPv6 packet to the migrating VM.

Tunneling Packets - IPv4: IPv4 packets received for the migrating VM are encapsulated in an IPv4 header at the Old NVE. New NVE decapsulates the packet and sends IPv4 packet to the migrating VM.

Stop Tunneling Packets: When Old NVE stops receiving packets destined to the VM that has just moved to the New NVE. The Timer for storing the New NVE address for the VM should be long enough for all other NVEs that need to communicate with the VM to get their NVE-VM cache entries updated.

6. Moving Local State of VM
In addition to the VM mobility related signaling (VM Mobility Registration Request/Reply), the VM state needs to be transferred to the New NVE. The state includes its memory and file system if the VM cannot access the memory and the file system after moved to the New NVE. Old NVE opens a TCP connection with New NVE over which VM’s memory state is transferred.

File system or local storage is more complicated to transfer. The transfer should ensure consistency, i.e. the VM at the New NVE should find the same file system it had at the Old NVE. Pre-copying is a commonly used technique for transferring the file system. First the whole disk image is transferred while VM continues to run. After the VM is moved any changes in the file system are packaged together and sent to the New NVE Hypervisor which reflects these changes to the file system locally at the destination.

7. Handling of Hot, Warm and Cold VM Mobility
Both Cold and Warm VM mobility (or migration) refers to the VM being completely shut down at the Old NVE before restarted at the New NVE. Therefore, all transport services to the VM are restarted.

Upon starting at the New NVE, the VM should send an ARP or Neighbor Discovery message. Cold VM mobility also allows the Old NVE and all communicating NVEs to time out ARP/neighbor cache entries of the VM. It is necessary for the NVA to push the
updated ARP/neighbor cache entry to NVEs or for NVEs to pull the updated ARP/neighbor cache entry from NVA.

The Cold VM mobility can be facilitated by cold standby entity receiving scheduled backup information. The cold standby entity can be a VM or can be other form factors which is beyond the scope of this document. The cold mobility option can be used for non-critical applications and services that can tolerate interrupted TCP connections.

The Warm VM mobility refers the backup entities receive backup information at more frequent intervals. The duration of the interval determines the warmth of the option. The larger the duration, the less warm (and hence cold) the Warm VM mobility option becomes.

There is also a Hot Standby option in addition to the Hot Mobility, where there are VMs in both primary and secondary NVEs and they identical information and can provide services simultaneously as in load-share mode of operation. If the VMs in the primary NVE fails, there is no need to actively move the VMs to the secondary NVE because the VMs in the secondary NVE already contain identical information. The hot standby option is the most costly mechanism, and hence this option is utilized only for mission-critical applications and services. In hot standby option, regarding TCP connections, one option is to start with and maintain TCP connections to two different VMs at the same time. The least loaded VM responds first and pickup providing service while the sender (origin) still continues to receive Ack from the heavily loaded (secondary) VM and chooses not use the service of the secondary responding VM. If the situation (loading condition of the primary responding VM) changes the secondary responding VM may start providing service to the sender (origin).

8. VM Operation

Once VM moves to a New NVE, VM IP address does not change and VM should be able to continue to receive packets to its address(es).

VM needs to send a gratuitous Address Resolution message or unsolicited Neighbor Advertisement message upstream after each move.

The VM lifecycle management is a complicated task, which is beyond the scope of this document. Not only it involves monitoring server utilization, balanced distribution of workload, etc., but also
needs to manage seamlessly VM migration from one server to another.

9. Security Considerations
   Security threats for the data and control plane for overlay networks are discussed in [RFC8014]. There are several issues in a multi-tenant environment that create problems. In Layer-2 based overlay data center networks, lack of security in VXLAN, corruption of VNI can lead to delivery to wrong tenant. Also, ARP in IPv4 and ND in IPv6 are not secure, especially if we accept gratuitous versions. When these are done over a UDP encapsulation, like VXLAN, the problem is worse since it is trivial for a non-trusted entity to spoof UDP packets.

   In Layer-3 based overlay data center networks, the problem of address spoofing may arise. An NVE may have untrusted tasks attached. This usually happens in cases like the VMs (tasks) running third party applications. This requires the usage of stronger security mechanisms.

10. IANA Considerations
    This document makes no request to IANA.

11. Acknowledgments
    The authors are grateful to Bob Briscoe, David Black, Dave R. Worley, Qiang Zu, Andrew Malis for helpful comments.

12. Change Log
    - submitted version -00 as a working group draft after adoption
    - submitted version -01 with these changes: references are updated,
      - added packets in flight definition to Section 2
    - submitted version -02 with updated address.
    - submitted version -03 to fix the nits.
13. References

13.1. Normative References


13.2. Informative References

Authors’ Addresses

Linda Dunbar
Futurewei
Email: ldunbar@futurewei.com

Behcet Sarikaya
Denpel Informatique
Email: sarikaya@ieee.org

Bhumip Khasnabish
Independent
55 Madison Avenue, Suite 160
Morristown, NJ 07960
Email: vumip1@gmail.com

Tom Herbert
Intel
Email: tom@herbertland.com

Saumya Dikshit
Aruba-HPE
Bangalore, India
Email: saumya.dikshit@hpe.com