Network Working Group
Internet Draft
Intended status: Informational
Expires: September 2012

Youval Nachum
Tal Mizrahi
Ilan Yerushalmi
Marvell

March 4, 2012

Scaling the Address Resolution Protocol for Large Data Centers (SARP)
draft-nachum-sarp-00.txt

Abstract

This document provides a recommended architecture and network operation named SARP. SARP is based on fast proxies that significantly reduce broadcast domains and ARP/ND broadcast transmissions. SARP supports smooth and fast virtual machine (VM) mobility without any modification to the VM, while keeping the connection up and running efficiently. SARP is targeted for massive scaling data centers with a significant number of VMs using ARP and ND protocols.

Status of this Memo

Copyright Notice

Table of Contents

1. Introduction

1.1. SARP Motivation

SARP provides operational recommendations that mitigate performance derogation due to the data center architecture. SARP can be used in large data centers with large amount of VMs where VMs migrate from one system to another while keeping their network connections up and running. Data center operators are required to allow the VMs to keep their IP and MAC identity while migrating between systems. The direct outcome of having VMs keep their respective IP and MAC identities is that Layer 2 broadcast domains are scaling up and protocols such as [ARP] and [ND] cause network performance derogation. SARP addresses a scaling problem that is also discussed in [ARMD].

1.2. SARP Overview

SARP uses FAST proxies that break down the large Layer 2 broadcast domains into small segments. The SARP proxies are located at the boundaries where the local Layer 2 infrastructure connects to its Layer 2 cloud. Figure 1 depicts an example of two remote data centers that are managed as a single flat Layer 2 domain. SARP proxies are implemented at the edge devices connecting the data center to the transport network. The direct outcome is significant reduction of broadcast domains and ARP/ND transmissions. The large L2 broadcast domains are bounded by the SARP proxies. ARP/ND transmissions are reduced due to the limited broadcast domains and the use of ARP/ND proxies and caching.

SARP proxies enable fast migration of a VM between clouds and data centers, keeping their connections up and running while the mobile VMs retain their IP and MAC addresses.

```
                        *-------------------*
                        |                   |
                +-------|     TRANSPORT     |-------+
                |       |                   |       |
                |       *-------------------*       |
                |                                   |
        *-----------------*                 *----------------*
        |  SARP Proxies   |                 |  SARP Proxies  |
        *-----------------*                 *----------------*
           |           |                       |           |
       *-------*   *-------*               *-------*   *-------*
       |  Agg  |   |  Agg  |               |  Agg  |   |  Agg  |
       *-------*   *-------*               *-------*   *-------*
           |
      *----------*
      |Hypervisor|
      *----------*
           |
       *--------*
       |Virtual |
       |Machine |
       *--------*

          (West Site)                          (East Site)
```

Figure 1 SARP Networking Architecture Example.

SARP distributes the Layer 2 Forwarding Information Base (FIB) from the edge devices (functioning as SARP proxies) to the VMs. By doing so, it significantly reduces table sizes on the edge devices. The source VM maintains the mapping of its destination VMs to the destination site/cloud in the ARP table. The destination VM IP is translated to the destination MAC address of the SARP proxy at the destination site. The SARP proxies only maintain Layer 2 FIB of local VMs and remote edge devices.

SARP proxies can support FAST VM migration and provide minimum transition phase. When SARP proxy indicates or is informed of VM migration, it can update all its peers and triggers a fast update.

SARP seamlessly supports Layer 2 network virtualization services over the overlay network and significantly reduces their complexity in terms of table size and performance. The overlay networks are only required to map MAC addresses of the SARP proxies to the correct tunnel.

1.3. SARP Deployment Options

SARP deployment is tightly coupled with the data center architecture. SARP proxies are located at the point where the Layer 2 infrastructure connects to its Layer 2 cloud using overlay networks. SARP proxies can be located at the data center edge (As Figure 1 depicts), data center core, or data center aggregation. SARP can also be implemented by the hypervisor (As Figure 2 depicts).

To simplify the description, we will focus on data centers that are managed as a single flat Layer 2 network, where SARP proxies are located at the boundary where the data center connects to the transport network (as Figure 1 depicts).

```
                      *-------------------*
                      |                   |
              +-------|     TRANSPORT     |-------+
              |       |                   |       |
              |       *-------------------*       |
              |                                   |
      *-----------------*                 *----------------*
      |   Edge Device   |                 |  Edge Device   |
      *-----------------*                 *----------------*
              |                                   |
      *-----------------*                 *----------------*
      |       Core      |                 |      Core      |
      *-----------------*                 *----------------*
         |           |                       |           |
     *-------*   *-------*               *-------*   *-------*
     |  Agg  |   |  Agg  |               |  Agg  |   |  Agg  |
     *-------*   *-------*               *-------*   *-------*
         |
    *----------*
    |Hypervisor|
    *----------*

        (West Site)                         (East Site)
```

Figure 2 SARP deployment options.

2. Abbreviations Used in this Document

ARP: Address Resolution Protocol

FIB: Forwarding Information Base

IP-D: IP address of the destination virtual machine

IP-S: IP address of the source virtual machine

MAC-D: MAC address of the destination virtual machine

MAC-E: MAC address of the East Proxy SARP Device

MAC-S: MAC address of the source virtual machine

ND: Neighbor Discovery

SARP Proxy: The components that participate at SARP protocol.

VM: Virtual Machine

3. SARP Description

3.1. Control Plane: ARP/ND

This section describes the ARP/ND procedure scenarios. In the first scenario, VMs share the same site. In the second scenario, the source VM is local and the destination VM is located at the remote site.

In all scenarios, the VMs (source and destination) share the same L2 broadcast domain.

3.1.1. ARP/ND Request for a Local VM

When source and destination VMs are located at the same site, the Address Resolution process is as described in [ARP]. When the VM sends an ARP request to learn the IP to MAC mapping of another local VM, it receives a reply from the other local VM with the IP-D to MAC-D mapping.

3.1.2. ARP/ND Request for a Remote VM

When the source and destination VMs are located at different sites, the Address Resolution process is as follows.

In our example, the source VM is located at the west site and the destination VM is located at the east site.

When the source VM sends an ARP/ND request to find out the IP to MAC mapping of a remote VM, the ARP request is propagated to the Layer 2 broadcast domain in all sites, including the east site.

The destination VM responds to the ARP/ND request and transmits an ARP/ND reply having the IP-D to MAC-D mapping.

The east SARP proxy functions as the proxy ARP of its Local VMs. The east SARP proxy modifies the ARP reply message to be IP-D to MAC-E and forwards the modified ARP reply message to all the SARP proxies.

The West SARP Proxy forwards the modified ARP reply message to the source VM.

The west SARP proxy can also functions as an ARP cache of the Remote VMs. By doing so, it significantly reduces the volume of the ARP/ND transmission over the network.

3.2. Data Plane: Packet Transmission

3.2.1. Local Packet Transmission

When a VM transmits packets to a destination VM that is located at the same site, there is no change in the data plane. The packets are sent from (IP-S, MAC-S) to (IP-D, MAC-D).

3.2.2. Packet Transmission Between Sites

Packets that are sent between sites traverse the SARP proxy of both sites. In our example, all packets sent from the VM located at the west site to the destination VM located at the east site traverse the west SARP proxy and the east SARP proxy.

The source VM follows its ARP table and sends packets to (IP-D, MAC-E) destination addresses and with (IP-s, MAC-S) as the source addresses.

The west SARP proxy replaces the packet source address to its own source address (MAC-W), keeps the destination address to be (MAC-E), and forwards the packet to the east proxy SARP.

When the east proxy SARP receives the packet, it replaces the destination MAC address to be (MAC-D) based on the packet destination IP (i.e., IP-D), but it does not change the source MAC addresses.

3.3. VM Local Migration

When a VM migrates locally within its site, the SARP protocol is not required to perform any action. VM migration is resolved entirely by the Layer 2 mechanisms.

3.4. VM Migration from One Site to Another

VMs migration from one site to another is done seamlessly, without any changes to the VMs addressing at any level while keeping VMs connections up and running.

In our example, the VM migrates from the west site to the east site.

VM migration differently affects VMs and networking elements based on their respective location:

- Origin site (west site)

- Destination site (east site)

- Other sites

Origin site:

The Origin site is the site where the VM started its connections before the migration, west site in our example.

All VMs at the west site that have an ARP entry of IP-D in their ARP table have the (IP-D to MAC-D) mapping. ARP mapping is updated by aging or by a gratuitous ARP message sent by the new hypervisor of the migrating VM and modified by the SARP proxy of the east site with (IP-D to MAC-E) mapping. Until ARP tables are updated, the source VMs from the west site continue sending packets to MAC-D. Switches at the west site are still configured with the old location of MAC-D. This can be resolved by MAC table aging or by redirecting the packets to the proxy SARP of the west site.

Destination Site:

The destination site is the site to which the VM migrated, the east site in our example.

All VMs at the east site that have an ARP entry of IP-D in their ARP table have the (IP-D to MAC-W) mapping. ARP mapping is updated by aging or by a gratuitous ARP message sent by the hypervisor (IP-D to MAC-D) mapping. Until ARP tables are updated, the source VMs from the

west site continue to send packets to MAC-W. This can be resolved by redirecting the packets from the SARP proxy of the east site to the migrated VM by updating the destination MAC of the packets to MAC-D.

Other Sites:

All VMs at the other sites that have an ARP entry of IP-D in their ARP table have the (IP-D to MAC-W) mapping. ARP mapping is updated by aging or by a gratuitous ARP message sent by the new hypervisor of the migrated VM and modified by the SARP proxy of the east site (IP-D to MAC-E) mapping. Until ARP tables are updated, the source VMs from the west site continue sending packets to MAC-W. This can be resolved by redirecting the packets from the SARP proxy of the west site to the SARP proxy of the east site by updating the destination MAC of the packets to MAC-E.

3.4.1. ARP/ND Table of Mobile VMs

The ARP table of the mobile VMs migrating from the west site to the east site includes the following types of VMs:

- Origin site (west site)

- Destination site (east site)

- Other Sites inhabitants

The IP to MAC mapping of VMs located at the other sites is unaffected by the migration.

The IP to MAC mapping of VMs located at east site can be kept with no change until the ARP aging time since they are mapped to MAC-E. All traffic from the migrated VM to VMs located at the east site traverses the SARP proxy of the east Site. This can be mitigated by ARP advertisement sent by the SARP proxy of the east site or by the hypervisor.

IP to MAC mapping of VMs located at west sites can be kept with no change until the ARP entries age out. All MAC addresses of the VMs located at the west site are unknown at the east site. All unknown traffic from the VM is intercepted by the SARP proxy of the east site and forwarded to the SARP proxy of the west site (just for ARP aging time). This can be resolved earlier by the east SARP proxy. Upon receiving unknown packets, it can update the migrating VM with the new IP to MAC mapping by sending a modified gratuitous ARP with (IP-D to MAC-W) mapping.

Note that overlay networks providing the Layer 2 network virtualization services configure their Edge Device MAC aging timers to be greater than the ARP request interval.

3.5. Multicast and Broadcast

To be added in a future version of this document

3.6. Non IP packet

To be added in a future version of this document

3.7. ARP caching

To be added in a future version of this document

3.8. SARP Interaction with Overlay networks

SARP interaction with overlay networks providing L2 network virtualization (such as IP, VPLS, OTV, NVGRE and VxLAN) is efficient and scalable.

The mapping of SARP to overlay networks is straightforward. The VM does the destination IP to SARP proxy MAC mapping. The mapping of the proxy MAC to its correct tunnel is done by the overlay networks. SARP significantly scales down the complexity of the overlay networks and transport networks by reducing the mapping tables to the number of SARP proxies.

4. Security Considerations

Security considerations will be added in a future version of this document.

5. IANA Considerations

There are no IANA actions required by this document.

RFC Editor: please delete this section before publication.

6. References

6.1. Normative References

[ARP] Plummer, D., "An Ethernet Address Resolution Protocol", RFC 826, November 1982.

[ND] Narten, T., Nordmark, E., Simpson, W., and H. Soliman, "Neighbor Discovery for IP version 6 (IPv6)", RFC 4861, September 2007.

6.2. Informative References

[ARMD] Narten, T., Karir, M., Foo, I., " Problem Statement for ARMD", draft-ietf-armd-problem-statement, February 2012.

7. Acknowledgments

This document was prepared using 2-Word-v2.0.template.dot.

Authors' Addresses

Youval Nachum
Marvell
6 Hamada St.
Yokneam, 20692 Israel
Email: youvaln@marvell.com

Tal Mizrahi
Marvell
6 Hamada St.
Yokneam, 20692 Israel
Email: talmi@marvell.com

Ilan Yerushalmi
Marvell
6 Hamada St.
Yokneam, 20692 Israel
Email: yilan@marvell.com