INTAREA Working Group
Internet Draft
Intended status: Proposed Standard
Expires: January 2014

Youval Nachum
Marvell
Linda Dunbar
Huawei
Ilan Yerushalmi
Tal Mizrahi
Marvell

July 15, 2013

Scaling the Address Resolution Protocol for Large Data Centers (SARP)
draft-nachum-sarp-06.txt

Abstract

This document introduces SARP, an architecture that uses proxy gateways to scale large data center networks. SARP is based on fast proxies that significantly reduce switches' FDB (MAC table) sizes and ARP/ND impact on network elements in an environment where hosts within one subnet (or VLAN) can spread over various locations. SARP is targeted for massive data centers with a significant number of VMs that can move across various physical locations.

Status of this Memo

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

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 January 15, 2014.

Table of Contents

1. Introduction

This document describes a proxy gateway technique, called Scalable Address Resolution Protocol (SARP), which reduces switches' Filtering Data Base (FDB) size and ARP/Neighbor Discovery impact on network elements in an environment where hosts within one subnet (or VLAN) can spread over various access domains in data centers.

The main idea of SARP is to represent all VMs (or hosts) under each access domain by their corresponding access (or aggregation) node's MAC address regardless whether the access (or aggregation) node is the VMs (hosts)' gateway or not. For example, when a host "a" under access domain "S" needs to communicate with peers on the same VLAN but connected to different access domains, SARP requires "a" to use remote access node's MAC address rather than peers' MAC addresses. By doing so, switches in each domain do not need to maintain a list of MAC addresses for all the VMs (hosts) in different access domains in their FDBs. Therefore, the switches' FDB size is limited regardless how VLAN is spread.

1.1. SARP Motivation

[RFC6820] has documented various impacts and scaling issues to data center networks when subnets span across multiple L2/l3 boundary routers.

Note: The L2/L3 boundary routers in this draft are capable of forwarding IEEE802.1 Ethernet frames (layer 2) without MAC header change. When subnets span across multiple ports of those routers, they are still under the category of a single link, or a multi-access link model recommended by [RFC4903]. They are different from the "multi-link" subnets described in [Multi-Link] and [RFC4903] which refer to a different physical media with the same prefix connected to a router and the layer 2 frames cannot be natively forwarded without header change.
Unfortunately, when the combined number of VMs (or hosts) in all those subnets is large, this can lead to switches' MAC table size

explosion and heavy impact on network elements. There are four major issues associated with subnets spanning across multiple L2/L3 boundary router ports:

1) Intermediate switches' MAC address table (FDB) explosion:
When hosts in a VLAN (or subnet) span across multiple access domains and each access domain has hosts belonging to different VLANs, each access switch has to enable multiple VLANs. Then, those access switches will be exposed to all MAC addresses among all the VLANs enabled.
For example, for an access switch with 40 physical servers attached, where each server has 100 VMs, there are 4000 hosts under the access switch. If indeed hosts/VMs can be moved anywhere, the worst case for the Access Switch is when all those 4000 VMs belong to different VLANs, i.e. the access switch has 4000 VLANs enabled. If each VLAN has 200 hosts, this access switch's MAC table potentially has 200*4000 = 800,000 entries.
It is important to note that the example above is relevant regardless of whether IPv4 or IPv6 are used.
The example illustrates a scenario that is worse than what today's L2/3 Gateway has to face. In today's environment where each subnet is limited to a few access switches, the number of MAC addresses the gateway has to learn is of a significantly smaller scale.

2) the ARP/ND processing load impact to the L2/L3 boundary routers;
All VMs periodically send NDs to their corresponding Gateway nodes to get gateway nodes' MAC addresses. When the combined number of VMs across all the VLANs is large, processing the responses to the ND requests from those VMs can easily exhaust the gateway's CPU utilization.
A L2/L3 boundary router could be hit with ARP/ND twice when the originating and destination stations are in different subnets attached to the same router and when those hosts do not communicate with external peers very frequently. The first hit is when the originating station in subnet-A initiates an ARP/ND request to the L2/L3 boundary router if the router's MAC is not in the host's cache; and the second hit is when the L2/L3

boundary router initiates an ARP/ND request to the target in subnet-B if the target is not in router's ARP/ND cache.

3) In IPv4, every end station in a subnet receives ARP broadcast messages from all other end stations in the subnet. IPv6 ND has eliminated this issue by using multicast.
However, most devices support a limited number of multicast addresses, due to multicast filtering scaling. Once the number of multicast addresses exceeds the multicast filter limit, the multicast addresses have to be processed by devices' CPU (i.e. the slow path).
It is less of an issue in DC without VM mobility because each port is only dedicated to one (or a few number of) VLANs. Thus, the number of multicast addresses hitting each port is significantly lower.

4) The ARP/ND messages are flooded to many physical link segments which can reduce the bandwidth utilization for user traffic; ARP/ND flooding is probably an insignificant issue in today's data center because the majority of data center servers are moving towards 1G or 10G ports. The bandwidth taken by ARP/ND, even when flooded to all physical links, becomes negligible compared to the link bandwidth. In addition, the IGMP/MLD snooping [RFC4541] can further reduce the ND multicast traffic to some physical link segments.

Statistics done by Merit Network [ARMD-Statistics] has shown that the major impact of a large number of mobile VMs in Data Centers is to the L2/L3 boundary routers, i.e., issue 2 above. A L2/L3 boundary router could be hit with ARP/ND twice when the originating and destination stations are in different subnets attached to the same router and those hosts do not communicate with external peers often enough. The first hit is when the originating station in subnet-A initiates an ARP/ND request to the L2/L3 boundary router if the router's MAC is not in the host's cache; and the second hit is when the L2/L3 boundary router initiates ARP/ND requests to the target in subnet-B if the target is not in router's ARP/ND cache.

Overlay approaches, e.g. [NVo3-PROBLEM], can hide hosts (VMs) addresses in the core but does not prevent the MAC table explosion problem (Issue 1) unless the NVE is on a server.

The scaling practices documented in [ARP-ND-PRACTICE] can only reduce some ARP impact to L2/L3 boundary routers in some scenarios, but not all.

In order to protect router CPUs from being overburdened by target resolution requests, some routers rate limit the target MAC resolution requests to CPU. When the rate limit is exceeded, the incoming data frames are dropped.

In traditional Data Centers, it is less of an issue because the number of hosts attached to one L2/L3 boundary router is limited by the number of physical ports of the switches/routers. When Servers are virtualized to support 30 plus VMs, the number of hosts under one router can grow 30 plus times. In addition, the traditional data center has each subnet nicely placed in a limited number of server racks, i.e., switches under router only need to deal with MAC addresses of those limited subnets. With subnets being spread across many server racks, the switches are exposed to VLAN/MAC of many subnets, greatly increasing the size of the FDB.

The solution proposed in this draft can eliminate or reduce the likelihood of inter-subnet data frames being dropped and reduce the host MAC addresses exposed to FDB on intermediate switches.

1.2. SARP Overview

SARP is a proxy gateway technique to reduce switches’ FDB (MAC table) sizes and ARP/ND impact on network elements in an environment where hosts within one subnet (or VLAN) can spread over various access domains in data centers.

Note: The Guidelines to proxy developers [RFC4389] have been carefully considered for the SARP protocols. Section 3.3 has demonstrated how SARP works when VMs are moved from one segment to another.

In order to enable VMs to be moved across greater number of servers while maintaining their MAC/IP addresses unchanged, the layer-2 network (e.g. VLAN) which interconnect those VMs may

spread across different server racks, different rows of server racks, or even different data centers.

For ease of description, let’s break the entire network which interconnects all those VMs into two segments: interconnecting segment and "access" segments. While the "Access" network is mostly likely Layer 2, the "interconnecting" segment might be not.

The SARP proxies are located at the boundaries where the "Access" segment connects to its "Interconnecting" segment. The boundary node could be a Hypervisor virtual switch, a Top of Rack switch, an Aggregation switch (or end of row switch), or a data center core switch. 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. SARP significantly reduces the ARP/ND transmissions over the "interconnection" network. The ARP/ND broadcast/multicast messages are bounded by the SARP proxies.

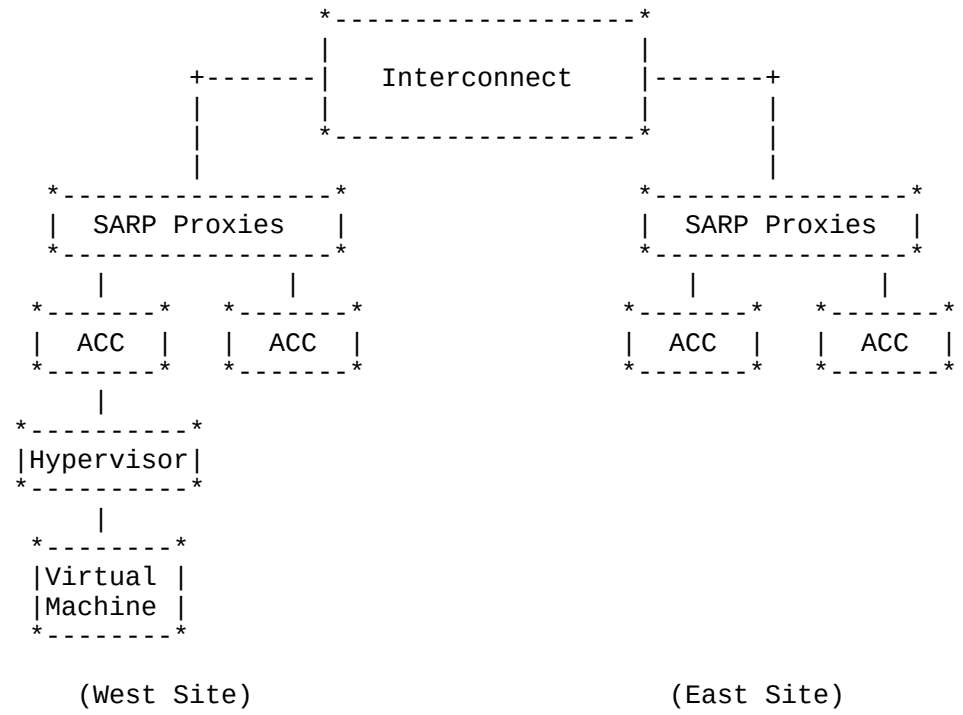

Figure 1 SARP Networking Architecture Example.

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).

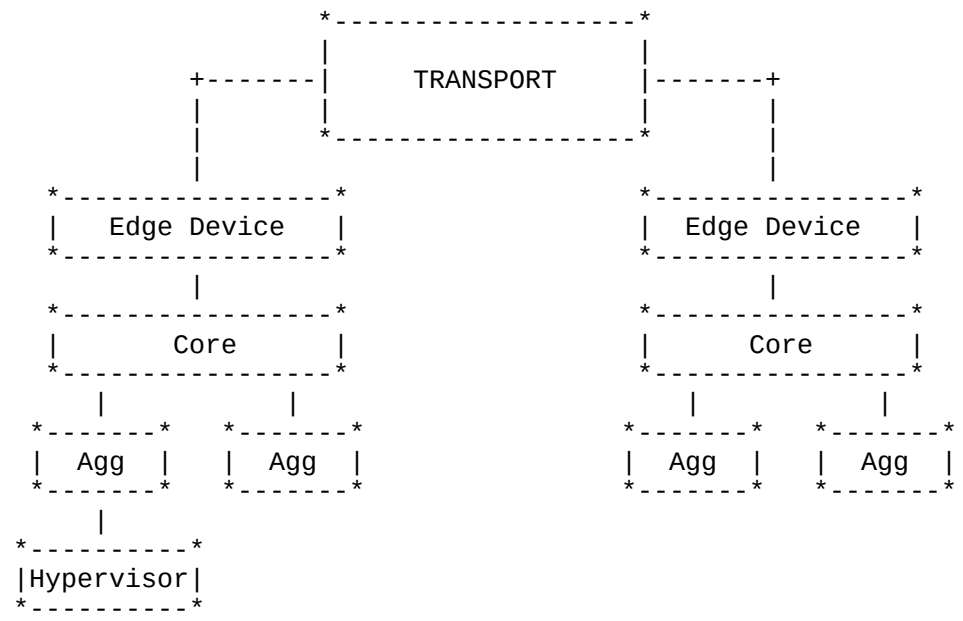

Figure 2 SARP deployment options.

2. Terms and Abbreviations Used in this Document

ARP: Address Resolution Protocol

FDB: Filtering Data Base, which is used for Layer-2 switches (IEEE802.1Q). Layer 2 switches flood data frames when DA is not in FDB, whereas routers drop data frames when the DA is not in the Forwarding Information Base (FIB). That is why Filtering Data Base (FDB) is used for Layer 2 switches.

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

NA: IPv6 ND's Neighbor Advertisement

ND: IPv6 Neighbor Discovery Protocol. In this document, ND also refers to Neighbor Solicitation, Neighbor Advertisement, Unsolicited Neighbor Advertisement messages defined by RFC4861

NS: IPv6 ND's Neighbor Solicitation

SARP Proxy: The components that participates in the SARP protocol.

UNA: IPv6 ND's Unsolicited Neighbor Advertisement

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 Access Segment. In the second scenario, the source VM is local Access Segment and the destination VM is located at the remote Access Segment.

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

3.1.1. ARP/NS Request for a Local VM

When source and destination VMs are located at the same Access Segment, the Address Resolution process is as described in [ARP] and [ND]. When the VM sends an ARP request or IPv6's Neighbor Solicitation (NS) 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/NS Request for a Remote VM

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

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

When the source VM sends an ARP/NS request to find out the IP to MAC mapping of a remote VM, if the local SARP proxy doesn't have the ARP cache for the target IP address or the cache entry has expired, the ARP/NS request is propagated to all Access Segments which might have VMs in the same virtual network as the originating VM, including the east Access Segment.

The destination VM responds to the ARP/NS request and transmits an ARP reply (IPv4) or Neighbor Advertisement (IPv6) 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 or NA message's source MAC-D to MAC-E and forwards the modified ARP reply or NA 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 IP<->MAC cache of the Remote VMs. By doing so, it significantly reduces the volume of the ARP/ND transmission over the network.

When the west SARP proxy caches the IP<-> MAC mapping entries for remote VMs, the timers for the entries to expire should be set relatively small to prevent stale entries due to remote VMs being moved or deleted. For environment where VMs move more frequently, it is not recommended for SARP Proxy to cache the IP<-> MAC mapping entries of remote VMs.

3.1.3. Gratuitous ARP and Unsolicited Neighbor Advertisement (UNA)

Hosts (or VMs) send out Gratuitous ARP (IPv4) and Unsolicited Neighbor Advertisement - UNA (IPv6) for other nodes to refresh IP<->MAC entries in their cache.

The local SARP processes the Gratuitous ARP or UNA in the same way as the ARP reply or IPv6 NA, i.e. replace the source MAC with its own MAC.

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 can either 1) simply forward the data frame to MAC-E, or 2)replace 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.

It is recommended for west SARP proxy to replace Source Address with its own if the "interconnecting segment" has address learning enabled. Otherwise nodes in the "interconnecting segment" can’t learn the address of the switch on which west SARP proxy is running unless the switch sends out frames periodically.

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. When the destination VM receives the packet, the Source Address field would be the MAC address of the VM on the west side or the MAC address of the west side SARP proxy,

Noted: it is common for data center network to have security policies to enforce some VMs can communicate with each other, and some VMs can’t. Most likely, those policies are configured by VM’s IP addresses. Even though the originating VM’s MAC address might be lost when the packet arrives at the destination VM, the originating VM’s IP address is still present in the data packets for security policy to be enforced.

Noted: for the option which doesn’t need west SARP to change source MAC of the data frames, the originating VM’s MAC will be

present when the data frames arrive at the destination VMs. Therefore, this option is valuable when hosts/VMs need to validate source VMs MAC addresses to comply any policies imposed.

Noted: Most hosts/VMs refresh its IP<->MAC mapping cache, with the Source MAC and Source IP of a received data frame. For the option which west SARP changes data frame's source MAC with its own MAC address, the destination VM's IP<->MAC cache can be refreshed with the valid mapping of the Source-VM-IP <->West-SARP-MAC. For the option of West SARP not changing source MAC, the destination VM has to turn off the learning of IP<->MAC mapping from the received data frames.

3.3. VM Migration

3.3.1. VM Local Migration

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

3.3.2. VM Migration from One Site to Another

In our example, the VM migrates from the west site to the east site while maintaining its MAC and IP addresses.

VM migration might affect 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 is before migration. It is the west site in our example.

Before the VM (IP=IP-D, MAC=MAC-D) is moved, 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. VMs on any other "Access Segments" will have ARP entry of (IP-D to MAC-W) mapping where MAC-W is the MAC address of the SARP proxy on the West Access Segment.

After the VM (IP-D) in the West Site moves to East Site, if there is gratuitous ARP (IPv4) or Unsolicited Neighbor Advertisement (IPv6) sent out by the destination hypervisor for the VM (IP-D), then the IP<->MAC mapping cache of VMs on all Access Segments will be updated by (IP-D to MAC-E) where MAC-E is the MAC address of the SARP proxy on the East Site. If there isn't any gratuitous ARP or Unsolicited Neighbor Advertisement sent out by the destination hypervisor, the IP<->MAC cache on the VMs in west site (and other sites) will eventually aged out.

Until IP<->MAC mapping cache 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 VM manager sending out a fake gratuitous ARP or Unsolicited Neighbor Advertisement on behalf of destination Hypervisor, shorter aging timer configured for IP<->MAC cache table, 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.

Before any gratuitous ARP or Unsolicited Neighbor Advertisement messages are sent out by the destination hypervisor, all VMs at the east site (and all other sites) might have (IP-D to MAC-W) mapping in their IP<->MAC mapping cache. IP<->MAC mapping cache is updated by aging or by a gratuitous ARP or UNA message sent by the destination hypervisor. Until IP<->MAC mapping caches are updated, the source VMs from the east site continue to send packets to MAC-W. This can be resolved by VM manager sending out a fake gratuitous ARP/UNA immediately after the VM migration, or 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 destination 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.

3.3.2.1. Impact to IP<->MAC Mapping Cache Table of VMs being moved

When a VM (IP-D) is moved from one site to another site, its IP<->MAC mapping entries for VMs located at the other sites (i.e. neither east site nor west site) are still valid, even though most Guest OSs (or VMs) will refresh their IP<->MAC cache after migration.

The VM (IP-D)’s IP<->MAC mapping entries for VMs located at east site, if not refreshed after migration, can be kept with no change until the ARP aging time since they are mapped to MAC-E. All traffic originated from the VM (IP-D) in its new location to VMs located at the east site traverses the SARP proxy of the east Site. The ARP/UNA sent by the SARP proxy of the east site or by the VMs on east side can always refresh the corresponding entries in the VM (IP-D)’s IP<->MAC cache .

The VM (IP-D)’s ARP entries (i.e. IP to MAC mapping) for VMs located at west sites will not be changed either until the ARP entries age out or new data frames are received from the remote sites. Since 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 by the east SARP proxy having mapping entries for VMs in the west side. 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.

3.4. Multicast and Broadcast

To be added in a future version of this document

3.5. Non IP packet

To be added in a future version of this document

3.6. IP<->MAC caching on SARP Proxy

ARP/NS Requests for a VM located at a remote site require flooding messages over the interconnecting network to all sites which have enabled the virtual network on which the VM belongs to. This scenario is described in details at 3.1.2. In such cases, SARP caching can reduce the number of ARP/ND transmissions over interconnecting networks.

In the example presented at section 3.1.2. 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 or Neighbor Solicitation request to discover the IP to MAC mapping of the remote VM, the request can be intercepted by the west SARP proxy.

The west SARP proxy learns or refreshes the source IP to source MAC mapping and looks up the IP to MAC translation of the destination IP. If the destination IP entry is found and is valid, the west SARP proxy replies with an ARP reply or Neighbor Advertisement without propagating the packet to other sites. Otherwise, the packet is propagated to all sites which have the virtual network enabled including the east site.

The propagated ARP/NS request is intercepted again by the east SARP proxy. It learns or refreshes the source IP to source MAC mapping and looks up the destination IP to MAC translation. If the destination IP entry is found and is valid the SARP proxy replies with an ARP reply or NA without propagating the ARP request to the east site. Otherwise, the ARP/NS request is broadcasted within the east site.

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

The east side SARP proxy intercepts the ARP reply or NA and learns or refreshes the Destination IP to Destination MAC mapping, replace the source MAC with its own MAC before sending the ARP reply or NA to the west SARP proxy (so that requesting VM can learn the IP-D to MAC-E mapping).

The West SARP Proxy intercepts the ARP reply or NA and learns or refreshes the Destination IP to Destination MAC mapping and propagates the ARP reply to the source VM.

The SARP proxies maintain an ARP caching table of IP to MAC mapping for all recent ARP/NS requests and replies. This table allows the SARP proxy to respond with low latency to the ARP/NS requests sent locally and avoid the broadcast transmissions of such requests over the transport network and all over the broadcast domains at the remote sites.

3.7. High availability and load balancing

The SARP proxy is located at the boundary where the local Layer 2 infrastructure connects to the interconnecting network. All traffic from the local site to the remote sites traverses the SARP proxy. The SARP proxy is subject to high availability and bandwidth requirements.

The SARP architecture supports multiple SARP proxies connecting a single site to the transport network. In SARP architecture all proxies can be active and can backup one another. The SARP architecture is robust and allows the network administrator to allocate proxies according to the bandwidth and high availability requirements.

Traffic is segregated between SARP proxies by using VLANs. An SARP proxy is the Master-SARP proxy of a set of VLANs and the Backup-SARP proxy of another set of VLANs.

For example the SARP proxies of the west site (as Figure 1 depicts) are SARP proxy-1 and SARP proxy-2. The west site supports VLAN-1 and VLAN-2 while SARP proxy-1 is the Master SARP proxy of VLAN-1 and the Backup proxy of VLAN-2 and SARP proxy-2 is the Master SARP proxy of VLAN-2 and the Backup SARP proxy of VLAN-1. Both proxies are members of VLAN-1 and VLAN-2.

The Master SARP proxy updates its Backup proxy with all the ARP reply messages. The Backup SARP proxy maintains a backup database to all the VLANs that it is the Backup SARP proxy.

The Master and the Backup SARP proxies maintain a keepalive mechanism. In case of a failure the Backup proxy becomes the Master SARP proxy. The failure decision is per VLAN. When the Master and the Backup proxies switchover, the backup SARP proxy can use the MAC address of the Master SARP proxy. The backup SARP proxy sends locally a gratuitous ARP message with the MAC address of the Master SARP proxy to update the forwarding tables on the

local switches. The backup SARP proxy also updates the remote SARP proxies on the change.

3.8. SARP Interaction with Overlay networks

SARP interaction with overlay networks providing L2 network virtualization (such as IP, VPLS, Trill, 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. Conclusions

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 trigger 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.

5. Security Considerations

The SARP proxies are located at the boundaries where the local Layer 2 infrastructure connects to its Layer 2 cloud. The SARP proxies interoperate with overlay network protocols that extend the Layer-2 subnet across data centers or between different systems within a data center.

SARP control plane and data plane are traversed by the overlay network hence SARP does not expose the network to additional security threats.

SARP proxies may be exposed to Denial of Service (DoS) attacks by means of ARP/ND message flooding. Thus, the SARP proxies must have sufficient resources to support the SARP control plane without making the network more vulnerable to DoS than without SARP proxies.

SARP adds security to the data plane by hiding all the local layer 2 MAC addresses from potential attacker located at the remote clouds. The only MAC addresses that are exposed at remote sites are the MAC addresses of the SARP proxies.

6. IANA Considerations

There are no IANA actions required by this document.

RFC Editor: please delete this section before publication.

7. References

7.1. 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.

[GratuitousARP] S. Cheshire, "IPv4 Address Conflict Detection", RFC 5227, July 2008.

[IGMP-MLD-tracking] H. Aseda, and N. Leymann, "IGMP/MLD-Based Explicit Membership Tracking Function for Multicast Routers" (http://tools.ietf.org/html/draft-ietf-pim-explicit-tracking-02), Oct, 2012.

[RFC826] D.C. Plummer, "An Ethernet address resolution protocol." RFC826, Nov 1982.

[RFC1027] Mitchell, et al, "Using ARP to Implement Transparent Subnet Gateways" (http://datatracker.ietf.org/doc/rfc1027/)

[RFC4389] Thaler, et al, "Neighbor Discovery Proxies (ND Proxy)", RFC4389, April 2006.

[RFC4541] Christensen, et al, "Considerations for Internet Group Management Protocol (IGMP) and Multicast Listener Discovery (MLD) Snooping Switches", RFC 4541, May 2006

[RFC4861] Narten, et al, "Neighbor Discovery for IP version 6 (IPv6)", RFC4861, Sept 2007

[RFC4903] Thaler, "Multilink Subnet Issues", RFC4903, July 2007.

[RFC6820] Narten, et al, "Address Resolution Problems in Large Data Center Networks", RFC6820, Jan 2013.

7.2. Informative References

[Impatient-NUD] E. Nordmark, I. Gashinsky, "draft-ietf-6man-impatient-nud"

[ARMD-Statistics] M. Karir, J. Rees, "Address Resolution Statistics", draft-karir-armd-statistics-01.txt (expired), July 2011. https://datatracker.ietf.org/doc/draft-karir-armd-statistics/

[ARP_Reduction] Shah, et al, "ARP Broadcast Reduction for Large Data Centers", draft-shah-armd-arp-reduction-02.txt (expired), Oct 2011. https://datatracker.ietf.org/doc/draft-shah-armd-arp-reduction/

[ARP-ND-PRACTICE] Dunbar, Kumari, Gashinsky, "Practices for scaling ARP and ND for large data centers", draft-dunbar-armd-arp-nd-scaling-practices-06, Feb 2013

[NVo3-PROBLEM] Narten, T., Gray, E., Black, D., Fang, L., Kreeger, L., Napierala, M., "Problem Statement: Overlays for Network Virtualization", draft-ietf-nvo3-overlay-problem-statement, work in progress, May 2013.

[Multi-Link] Thaler, et al, "Multi-link Subnet Support in IPv6", draft-ietf-ipv6-multi-link-subnets-00.txt (expired), Dec 2002. https://datatracker.ietf.org/doc/draft-ietf-ipv6-multilink-subnets/

8. Acknowledgments

We want to thank Ted Lemon in providing many rounds of valuable comments and suggestions to the draft.

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

Authors' Addresses

Youval Nachum
Email: youval.nachum@gmail.com

Linda Dunbar
Huawei Technologies
5430 Legacy Drive, Suite #175
Plano, TX 75024, USA
Phone: (469) 277 5840
Email: ldunbar@huawei.com

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

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