Independent Submission
Request for Comments: 7586
Category: Experimental
ISSN: 2070-1721

Y. Nachum

L. Dunbar
Huawei
I. Yerushalmi
T. Mizrahi
Marvell

June 2015

The Scalable Address Resolution Protocol (SARP) for Large Data Centers

Abstract

This document introduces the Scalable Address Resolution Protocol (SARP), an architecture that uses proxy gateways to scale large data center networks. SARP is based on fast proxies that significantly reduce switches' Filtering Database (FDB) table sizes and reduce impact of ARP and Neighbor Discovery (ND) 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 Virtual Machines (VMs) that can move across various physical locations.

Independent Submissions Editor Note

This is an Experimental document; that experiment will end two years after the RFC is published. At that point, the RFC authors will attempt to determine how widely SARP has been implemented and used.

IESG Note

The IESG notes that the problems described in RFC 6820 can already be addressed through the simple combination of existing standardized or other published techniques including Layer 2 VPN (RFC 4664), proxy ARP (RFC 925), proxy Neighbor Discovery (RFC 4389), IGMP and MLD snooping (RFC 4541), and ARP mediation for IP interworking of Layer 2 VPNs (RFC 6575).

Status of This Memo

This document is not an Internet Standards Track specification; it is published for examination, experimental implementation, and evaluation.

This document defines an Experimental Protocol for the Internet community. This is a contribution to the RFC Series, independently of any other RFC stream. The RFC Editor has chosen to publish this document at its discretion and makes no statement about its value for implementation or deployment. Documents approved for publication by the RFC Editor are not a candidate for any level of Internet Standard; see Section 2 of RFC 5741.

Information about the current status of this document, any errata, and how to provide feedback on it may be obtained at http://www.rfc-editor.org/info/rfc7586.

Copyright Notice

Table of Contents

1. Introduction

This document describes a proxy gateway technique, called the Scalable Address Resolution Protocol (SARP), which reduces switches' Filtering Database (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 the MAC address of their corresponding access node (or aggregation node). For example (Figure 1), when host A in the west site needs to communicate with host B, which is on the same VLAN but connected to a different access domain (east site), SARP requires host A to use the MAC address of SARP proxy 2, rather than the address of host B. 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; every switch only needs to be familiar with MAC addresses that reside in the current domain, and addresses of remote SARP proxy gateways. Therefore, the switches' FDB size is limited regardless of the number of access domains.

```
     +-------+     +-------+    _   __       +-------+     +-------+
     |       |     | SARP  |   / \_/  \_     | SARP  |     |       |
     |host A |<===>| proxy |<=>\_       \<==>| proxy |<===>|host B |
     |       |     |   1   |   /       _/    |   2   |     |       |
     +-------+     +-------+   \__   _/      +-------+     +-------+
                                  \_/
     <------West Site------>                 <------East Site------>
```

Figure 1: A Brief Overview of SARP

1.1. SARP Motivation

[RFC6820] discusses the impacts and scaling issues that arise in data center networks when subnets span across multiple Layer 2 / Layer 3 (L2/L3) boundary routers.

Unfortunately, when the combined number of VMs (or hosts) in all those subnets is large, it can lead to an explosion of the size of the switches' MAC address table and a heavy impact on network elements.

There are four major issues associated with subnets spanning across multiple L2/L3 boundary router ports:

1) Explosion of the size of the intermediate switches' MAC address table (FDB).

   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. Thus, those access switches are exposed to all MAC addresses across all VLANs.

   For example, for an access switch with 40 attached physical servers, where each server has 100 VMs, the access switch has 4,000 attached MAC addresses. If hosts/VMs can indeed be moved anywhere, the worst case for the Access Switch is when all those 4,000 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 address table potentially has 200 * 4,000 = 800,000 entries.

It is important to note that the example above is relevant regardless of whether IPv4 or IPv6 is used.

The example illustrates a scenario that is worse than what today's L2/L3 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) ARP/ND processing load impact on 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.

An 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 1 initiates an ARP/ND request to the L2/L3 boundary router. The second hit is when the L2/L3 boundary router initiates an ARP/ND request to the target in subnet 2 if the target is not in the 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 the scaling of multicast filtering. Once the number of multicast addresses exceeds the multicast filter limit, the multicast addresses have to be processed by the devices' CPUs (i.e., the slow path).

It is less of an issue in data centers without VM mobility, since each port is only dedicated to one (or a small 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 that can reduce the bandwidth utilization for user traffic.

ARP/ND flooding is, in most cases, an insignificant issue in today's data center networks, as the majority of data center servers are shifting towards 1G or 10G Ethernet ports. The bandwidth used by ARP/ND, even when flooded to all physical links, becomes negligible compared to the link bandwidth. Furthermore, IGMP and Multicast Listener Discovery (MLD) snooping [RFC4541] can further reduce the ND multicast traffic to some physical link segments.

Statistics gathered by Merit Network [ARMDStats] have shown that the major impact of a large number of VMs in data centers is on the L2/L3 boundary routers, i.e., issue 2 above. An L2/L3 boundary router could be hit with ARP/ND twice when 1) the originating and destination stations are in different subnets attached to the same router, and 2) those hosts do not communicate with external peers often enough.

Overlay approaches, e.g., [RFC7364], can hide addresses of hosts (VMs) in the core, but they do not prevent the MAC address table explosion problem (issue 1) unless the Network Virtualization Edge (NVE) is on a server.

The scaling practices documented in [RFC7342] can only reduce some ARP impact on 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 the router’s CPU. When the rate limit is exceeded, the incoming data frames are dropped. In traditional data centers, this issue is less significant, since 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+ VMs, the number of hosts under one router can grow by a factor of 30+. Furthermore, in traditional data center networks, each subnet is neatly bound to a limited number of server racks, i.e., switches only need to be familiar with MAC addresses of hosts that reside in this small number of subnets. In contemporary data center networks, as subnets are spread across many server racks, switches are exposed to VLAN/MAC addresses of many subnets, greatly increasing the size of switches’ FDB tables.

The solution proposed in this document can eliminate or reduce the likelihood of inter-subnet data frames being dropped and reduce the number of host MAC addresses that intermediate switches are exposed to, thus reducing switches’ FDB table sizes.

1.2. SARP Overview

The SARP approach uses proxy gateways to address the problems discussed above.

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

In order to enable VMs to be moved across servers while ensuring their MAC/IP addresses remain unchanged, the Layer 2 network (e.g., VLAN) that interconnects those VMs may spread across different server racks, different rows of server racks, or even different data center sites.

A multisite data center network is comprised of two main building blocks: an interconnecting segment and an access segment. While the access network is, in most cases, a Layer 2 network, the interconnecting segment is not necessarily a Layer 2 network.

The SARP proxies are located at the boundaries where the access segment connects to its interconnecting segment. The boundary node can 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 2 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 interconnecting network.

```
                            *-------------------*
                            |                   |
                    +-------| Interconnecting   |-------+
                    |       |     network       |       |
                    |       *-------------------*       |
                    |                                   |
           *-----------------*                  *----------------*
           |  SARP Proxies   |                  |  SARP Proxies  |
           *-----------------*                  *----------------*
              |           |                        |           |
          *-------*   *-------*                *-------*   *-------*
          |Access |   |Access |                |Access |   |Access |
          *-------*   *-------*                *-------*   *-------*
              |
         *----------*
         |Hypervisor|
         *----------*
              |
          *--------*
          |Virtual |
          |Machine |
          *--------*

             (West Site)                          (East Site)
```

               Figure 2: SARP: Network 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 2 depicts), data center core, or data center aggregation (denoted by "Agg" in the figure). SARP can also be implemented by the hypervisor (as Figure 3 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 2 depicts).

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

            (West Site)                             (East Site)
```

Figure 3: SARP Deployment Options

1.4. Comparison with Existing Solutions

The IETF has developed several mechanisms to address issues associated with Layer 2 networks over multiple geographic locations, for example, Layer 2 VPN [RFC4664], proxy ARP [RFC925] [ProxyARP], proxy Neighbor Discovery [RFC4389], IGMP and MLD snooping [RFC4541], and ARP mediation for IP interworking of Layer 2 VPNs [RFC6575].

However, all those solutions work well when hosts within one subnet are placed together under one access domain, so that the intermediate switches in each access domain are only exposed to host addresses from a limited number of subnets. SARP is to provide a solution when hosts within one subnet are spread across multiple access domains, and each access domain has hosts from many subnets. Under this environment, the intermediate switches in each access domain are exposed to combined hosts of all the subnets that are enabled by the access domain.

2. Terms and Abbreviations Used in This Document

ARP: Address Resolution Protocol [ARP]

FDB: Filtering Database, which is used for Layer 2 switches [802.1Q]. Layer 2 switches flood data frames when the Destination Address (DA) is not in the FDB, whereas routers drop data frames when the DA is not in the Forwarding Information Base (FIB). That is why the FDB is used for Layer 2 switches.

FIB: Forwarding Information Base

Hypervisor: a software layer that creates and runs virtual machines on a server

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 [ND]. In this document, ND also refers to Neighbor Solicitation, Neighbor Advertisement, and Unsolicited Neighbor Advertisement messages defined by RFC 4861.

NS: IPv6 ND’s Neighbor Solicitation

SARP Proxy: The components that participate in SARP

UNA: IPv6 ND’s Unsolicited Neighbor Advertisement [ND]

VM: Virtual Machine

3. SARP: Theory of Operation

3.1. Control Plane: ARP/ND

This section describes the ARP/ND procedure scenarios. The first scenario addresses a case where both the source and destination VMs reside in the same access segment. In the second scenario, the source VM is in the 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 (Figure 4), the address resolution process is as described in [ARP] and [ND]; host A sends an ARP request or an IPv6 Neighbor Solicitation (NS) to learn the IP-to-MAC mapping of host B, and it receives a reply from host B with the IP-D to MAC-D mapping.

```
     +-------+      _  __        +-------+      _  __
     |host A |     / \_/  \_     | SARP  |     / \_/  \_
     | IP-S  |<--->\_access \<==>| proxy |<===>\_interc.\
     | MAC-S |     /network_/    |   1   |     /network_/
     +-------+  +->\__    _/     +-------+     \__    _/
                |     \_/                         \_/
     +-------+  |
     |host B |<-+
     | IP-D  |
     | MAC-D |
     +-------+

     <--------------West Site------------>
```

Figure 4: SARP: Two Hosts in the Same Access Segment

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.

```
  +-------+     +-------+     _  __       +-------+     +-------+
  |host A |     | SARP  |    / \_/  \_    | SARP  |     |host B |
  | IP-S  |<===>|proxy 1|<=>\_        \<==>|proxy 2|<===>| IP-D  |
  | MAC-S |     | MAC-W |    /        _/   | MAC-E |     | MAC-D |
  +-------+     +-------+    \__    _/    +-------+     +-------+
                                \_/
  <------West Site------>                 <------East Site------>
```

Figure 5: SARP: Two Hosts That Reside in Different Segments

In the example illustrated in Figure 5, the source VM is located at the west access segment and the destination VM is located at the east access segment.

When host A sends an ARP/NS request to find out the IP-to-MAC mapping of host B:

1. If SARP proxy 1 does not have IP-D in its ARP cache, the ARP/NS request is propagated to all access segments that might have VMs in the same virtual network as the originating VM, including the east access segment.

2. As SARP proxy 1 forwards the ARP/NS message, it replaces the source MAC address, MAC-S, with its own MAC address, MAC-W. Thus, all switches that reside in the interconnecting segment are not exposed to MAC-S.

3. The ARP/NS request reaches SARP proxy 2.

4. If SARP proxy 2 does not have IP-D in its ARP cache, the ARP/NS request is forwarded to the east access network. Host B responds with an ARP reply (IPv4) or a Neighbor Advertisement (IPv6) to the request with MAC-D.

5. When the response message reaches SARP proxy 2, it replaces MAC-D with MAC-E; thus, the response reaches SARP proxy 1 with MAC-E.

6. As SARP proxy 1 forwards the response to host A, it replaces the destination address from MAC-W to MAC-S.

SARP Proxy ARP/ND Cache

SARP proxies maintain a cache of the IP-to-MAC mapping. This cache is based on ARP/ND messages that are sent by hosts and traverse the SARP proxies.

In steps 1 and 4 above, if the SARP proxy has IP-D in its ARP cache, it responds with MAC-E, without forwarding the ARP/NS request.

This caching approach significantly reduces the volume of the ARP/ND transmission over the network and reduces the round-trip time of ARP/ND requests.

When the west SARP proxy caches the IP-to-MAC mapping entries for remote VMs, the expiration timers should be set to relatively low values to prevent stale entries due to remote VMs being moved or deleted. In environments where VMs move more frequently, it is not recommended for SARP proxies to cache the IP-to-MAC mapping entries of remote VMs.

3.1.3. Gratuitous ARP and Unsolicited Neighbor Advertisement (UNA)

Hosts (or VMs) send out Gratuitous ARP (IPv4) [TcpIp] and Unsolicited Neighbor Advertisement (UNA) (IPv6) messages to allow other nodes to refresh IP-to-MAC entries in their caches.

The local SARP proxy processes the Gratuitous ARP or UNA message in the same way as the ARP reply or IPv6 NA, i.e., replaces the MAC addresses in the same manner.

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 (Figure 4), the data plane is unaffected by SARP; 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 (Figure 5) traverse the SARP proxy of both sites.

A packet sent from host A to host B undergoes the following procedure:

1. Host A sends a packet to IP-D, and based on its ARP table, it uses the MAC addresses {MAC-E, MAC-S}.

2. SARP proxy 1 receives the packet and replaces the source MAC address, such that the packet includes {MAC-E, MAC-W}.

3. SARP proxy 2 receives the packet and replaces the destination MAC address, and the packet is sent to host B with {MAC-D, MAC-W}.

SARP proxy 1 replaces the source MAC address with its own, since switches in the interconnecting segment are only familiar with SARP proxy MAC addresses and are not familiar with host addresses.

Note: it is a common security practice in data center networks to use access lists, allowing each VM to communicate only with a list of authorized peer VMs. In most cases, such access control lists are based on IP addresses and, hence, are not affected by the MAC address replacement in SARP.

3.3. VM Migration

3.3.1. VM Local Migration

When a VM migrates locally within its access segment, SARP does not require any special behavior. VM migration is resolved entirely by the Layer 2 mechanisms.

3.3.2. VM Migration from One Site to Another

This section focuses on a scenario where a 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 locations:

- origin site (west site)

- destination site (east site)

- other sites

```
     +-------+     +-------+    _   __       +-------+     +-------+
     |host A |     | SARP  |   / \_/  \_     | SARP  |     |host A |
     | IP-D  |<===>|proxy 1|<=>\_       \<==>|proxy 2|<===>| IP-D  |
     | MAC-D |     | MAC-W |   /       _/    | MAC-E |     | MAC-D |
     +-------+     +-------+   \__   _/      +-------+     +-------+
                                  \_/
     <------West Site------>                 <------East Site------>
           Origin Site                          Destination Site
```

Figure 6: SARP: Host A Migrates from West Site to East Site

Origin Site

The origin site is the site where the VM resides before the migration (west site).

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 -> MAC-D mapping. VMs on other access segments have an ARP entry of IP-D -> 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 the east site, if a Gratuitous ARP (IPv4) or an Unsolicited Neighbor Advertisement (IPv6) message is sent out by the destination hypervisor on behalf of the VM (IP-D), then the IP-to-MAC mapping cache of the VMs in all access segments is updated by IP-D -> MAC-E, where MAC-E is the MAC address of the SARP proxy on the east site. If no Gratuitous ARP or UNA message is sent out by the destination hypervisor, the IP-to-MAC cache on the VMs in the west site (and other sites) is eventually aged out.

Until the IP-to-MAC mapping cache tables are updated, the source VMs from the west site continue sending packets locally to MAC-D, and switches at the west site are still configured with the old location of MAC-D. This transient condition can be resolved by having the VM manager send out a fake Gratuitous ARP or UNA message on behalf of the destination Hypervisor. Another alternative is to have a shorter aging timer configured for the IP-to-MAC cache table.

Destination Site

The destination site is the site to which the VM migrated, i.e., the east site in Figure 6.

Before any Gratuitous ARP or UNA messages are sent out by the destination hypervisor, all VMs at the east site (and all other sites) might have an IP-D -> MAC-W mapping in their IP-to-MAC mapping cache. The IP-to-MAC mapping cache is updated by aging or by a Gratuitous ARP or UNA message sent by the destination hypervisor. Until the IP-to-MAC mapping caches are updated, VMs from the east site continue to send packets to MAC-W. This can be resolved by having the VM manager send out a fake Gratuitous ARP or UNA message immediately after the VM migration or by redirecting the packets from the SARP proxy of the east site back 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 -> MAC-W mapping. The 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 to an IP-D -> MAC-E mapping. Until ARP tables are updated, VMs from other sites continue sending packets to MAC-W.

3.3.2.1. Impact on IP-to-MAC Mapping Cache Table of Migrated VMs

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

The migrated VM's IP-to-MAC mapping entries for VMs located at the east site, if not refreshed after migration, can be kept with no change until the ARP aging time, as these entries are mapped to MAC-E. All traffic originated from the migrated VM in its new location to VMs located at the east site traverses the SARP proxy of the east site. That SARP proxy can redirect the traffic back to the corresponding destinations on the east site. Furthermore, an ARP/UNA message sent by the SARP proxy of the east site or by the VMs on the east site can refresh the corresponding entries in the migrated VM's IP-to-MAC cache.

The migrated VM's ARP entries for VMs located at the west site remain unchanged until either 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 (during the transient period before the ARP entries age out). This transient behavior is avoided if the SARP proxy has the destination IP address in its ARP cache, and, upon receiving a packet with an unknown destination MAC address, it could send a Gratuitous ARP or UNA message to the migrated VM.

Note that overlay networks providing 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

Multicast and broadcast traffic is forwarded by SARP proxies as follows:

- o SARP proxies modify the source MAC address of multicast and broadcast packets as described in Section 3.2.
- o SARP proxies do not modify the destination MAC address of multicast and broadcast packets.

3.5. Non-IP Packet

The L2/L3 boundary routers in the current document are capable of forwarding non-IP IEEE 802.1 Ethernet frames (Layer 2) without changing the MAC headers. 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 differ from the "multi-link" subnets described in [MultLinkSub] and [RFC4903], which refer to a different physical media with the same prefix connected to a router, where the Layer 2 frames cannot be natively forwarded without changing the headers.

3.6. 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 the SARP architecture, all proxies can be active and can back up one another. The SARP architecture is robust and allows network administrators to allocate proxies according to 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, assume the SARP proxies of the west site 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 SARP 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 SARP 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 of.

The Master and the Backup SARP proxies maintain a keepalive mechanism. In case of a failure, the Backup SARP proxy becomes the Master SARP proxy. The failure decision is per VLAN. When the Master and the Backup SARP proxies switch over, 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.

3.7. SARP Interaction with Overlay Networks

SARP can be used over overlay networks, providing L2 network virtualization (such as IP, Virtual Private LAN Service (VPLS), Transparent Interconnection of Lots of Links (TRILL), Overlay Transport Virtualization (OTV), Network Virtualization using GRE (NVGRE), and Virtual eXtensible Local Area Network (VXLAN)). The mapping of SARP to overlay networks is straightforward; the VM does the mapping of the destination IP to the SARP proxy MAC address. 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

SARP proxies are located at the boundaries of access networks, where the local Layer 2 infrastructure connects to its Layer 2 cloud. 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 does not expose the network to security threats beyond those that exist whether or not SARP is present.

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

SARP adds security to the data plane in terms of network reconnaissance, by hiding all the local Layer 2 MAC addresses from potential attackers located at the interconnecting network and significantly limiting the number of addresses exposed to an attacker at a remote site.

5. References

5.1. Normative References

[ARP] Plummer, D., "Ethernet Address Resolution Protocol: Or Converting Network Protocol Addresses to 48.bit Ethernet Address for Transmission on Ethernet Hardware", STD 37, RFC 826, DOI 10.17487/RFC0826, November 1982, <http://www.rfc-editor.org/info/rfc826>.

[ND] Narten, T., Nordmark, E., Simpson, W., and H. Soliman, "Neighbor Discovery for IP version 6 (IPv6)", RFC 4861, DOI 10.17487/RFC4861, September 2007, <http://www.rfc-editor.org/info/rfc4861>.

[ProxyARP] Carl-Mitchell, S. and J. Quarterman, "Using ARP to implement transparent subnet gateways", RFC 1027, DOI 10.17487/RFC1027, October 1987, <http://www.rfc-editor.org/info/rfc1027>.

[RFC925] Postel, J., "Multi-LAN address resolution", RFC 925, DOI 10.17487/RFC0925, October 1984, <http://www.rfc-editor.org/info/rfc925>.

[RFC4389] Thaler, D., Talwar, M., and C. Patel, "Neighbor Discovery Proxies (ND Proxy)", RFC 4389, DOI 10.17487/RFC4389, April 2006, <http://www.rfc-editor.org/info/rfc4389>.

[RFC4541] Christensen, M., Kimball, K., and F. Solensky, "Considerations for Internet Group Management Protocol (IGMP) and Multicast Listener Discovery (MLD) Snooping Switches", RFC 4541, DOI 10.17487/RFC4541, May 2006, <http://www.rfc-editor.org/info/rfc4541>.

[RFC4664] Andersson, L., Ed., and E. Rosen, Ed., "Framework for Layer 2 Virtual Private Networks (L2VPNs)", RFC 4664, DOI 10.17487/RFC4664, September 2006, <http://www.rfc-editor.org/info/rfc4664>.

[RFC6575] Shah, H., Ed., Rosen, E., Ed., Heron, G., Ed., and V. Kompella, Ed., "Address Resolution Protocol (ARP) Mediation for IP Interworking of Layer 2 VPNs", RFC 6575, DOI 10.17487/RFC6575, June 2012, <http://www.rfc-editor.org/info/rfc6575>.

5.2. Informative References

[802.1Q] IEEE, "IEEE Standard for Local and metropolitan area networks -- Bridges and Bridged Networks", IEEE Std 802.1Q.

[ARMDStats] Karir, M., and J. Rees, "Address Resolution Statistics", Work in Progress, draft-karir-armd-statistics-01, July 2011.

[MultLinkSub] Thaler, D., and C. Huitema, "Multi-link Subnet Support in IPv6", Work in Progress, draft-ietf-ipv6-multi-link-subnets-00, June 2002.

[RFC4903] Thaler, D., "Multi-Link Subnet Issues", RFC 4903, DOI 10.17487/RFC4903, June 2007, <http://www.rfc-editor.org/info/rfc4903>.

[RFC6820] Narten, T., Karir, M., and I. Foo, "Address Resolution Problems in Large Data Center Networks", RFC 6820, DOI 10.17487/RFC6820, January 2013, <http://www.rfc-editor.org/info/rfc6820>.

[RFC7342] Dunbar, L., Kumari, W., and I. Gashinsky, "Practices for Scaling ARP and Neighbor Discovery (ND) in Large Data Centers", RFC 7342, DOI 10.17487/RFC7342, August 2014, <http://www.rfc-editor.org/info/rfc7342>.

[RFC7364] Narten, T., Ed., Gray, E., Ed., Black, D., Fang, L., Kreeger, L., and M. Napierala, "Problem Statement: Overlays for Network Virtualization", RFC 7364, DOI 10.17487/RFC7364, October 2014, <http://www.rfc-editor.org/info/rfc7364>.

[TcpIp] Stevens, W., "TCP/IP Illustrated, Volume 1: The Protocols", Addison-Wesley, 1994.

Acknowledgments

The authors thank Ted Lemon, Eric Gray, and Adrian Farrel for providing valuable comments and suggestions for the document.

Authors' Addresses

Youval Nachum
EMail: youval.nachum@gmail.com

Linda Dunbar
Huawei Technologies
5430 Legacy Drive, Suite #175
Plano, TX 75024
United States
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