L. Dunbar
Huawei
W. Kumari
Google
Igor Gashinsky
Yahoo

Internet Draft
Intended status: Informational
Expires: July 2013

January 31, 2013

Practices for scaling ARP and ND for large data centers

draft-dunbar-armd-arp-nd-scaling-practices-05

Status of this Memo

Copyright Notice

Abstract

This draft documents some operational practices that allow ARP/ND to scale in data center environments.

Table of Contents

1. Introduction

This draft documents some operational practices that allow ARP/ND to scale in data center environments.

As described in [ARMD-Problem], the increasing trend of rapid workload shifting and server virtualization in modern data centers requires servers to be loaded (or re-loaded) with different VMs or applications at different times. Different VMs residing on one physical server may have different IP addresses, or may even be in different IP subnets.

In order to allow a physical server to be loaded with VMs in different subnets, or VMs to be moved to different server racks without IP address re-configuration, the networks need to enable multiple broadcast domains (many VLANs) on the interfaces of L2/L3 boundary routers and ToR switches. Unfortunately, when the combined number of VMs (or hosts) in all those subnets is large, this can lead to address resolution scaling issues, especially on the L2/L3 boundary routers.

This draft documents some simple practices which can scale ARP/ND in data center environment.

2. Terminology

This document reuses much of terminology from [ARMD-Problem]. Many of the definitions are presented here to aid the reader.

ARP: IPv4 Address Resolution Protocol [RFC826]

Aggregation Switch: A Layer 2 switch interconnecting ToR switches

Bridge: IEEE802.1Q compliant device. In this draft, Bridge is used interchangeably with Layer 2 switch.

DC: Data Center

DA: Destination Address

End Station: VM or physical server, whose address is
either a destination or the source of a data frame.

EOR: End of Row switches in data center.

NA: IPv6's Neighbor Advertisement

ND: IPv6's Neighbor Discovery [RFC4861]

NS: IPv6's Neighbor Solicitation

SA: Source Address

ToR: Top of Rack Switch (also known as access switch).

UNA: IPv6's Unsolicited Neighbor Advertisement

VM: Virtual Machines

3. Common DC network Designs

Some common network designs for data center include:

1) Layer 3 connectivity to the access switch,

2) Large Layer 2, and

3) Overlay models.

There is no single network design that fits all cases. The following sections document some of the common practices to scale Address Resolution under each network design.

4. Layer 3 to Access Switches

This network design makes Layer 3 configured to the access switches; effectively making the access switches the L2/L3 boundary routers for the attached VMs.

As described in [ARMD-Problem], many data centers are architected so that ARP/ND broadcast/multicast messages are confined to a few ports (interfaces) of the access switches (i.e. ToR switches).

Another variant of the Layer 3 solution is Layer 3 infrastructur configured all the way to servers (or even to

the VMs), which confines the ARP/ND broadcast/multicast messages to the small number of VMs within the server.

Advantage: Both ARP and ND scale well. There is no address resolution issue in this design.

Disadvantage: The main disadvantage to this network design occurs during VM movement. During VM movement, either VMs need address change or switches/routers need configuration change when the VMs are moved to different locations.

Summary: This solution is more suitable to data centers which have static workload and/or network operators who can re-configure IP addresses/subnets on switches before any workload change. No protocol changes are suggested.

5. Layer 2 practices to scale ARP/ND

5.1. Practices to alleviate APR/ND burden on L2/L3 boundary routers

The ARP/ND broadcast/multicast messages in a Layer 2 domain can negatively affect the L2/L3 boundary routers, especially with a large number of VMs and subnets. This section describes some commonly used practices in reducing the ARP/ND processing required on L2/L3 boundary routers.

5.1.1. Communicating with a peer in a different subnet

When the communicating peer is in a different subnet, the originating end station needs to send ARP/ND requests to its default gateway router to resolve the router's MAC address. If there are many subnets on the gateway router and a large number of end stations in those subnets, the gateway router has to process a very large number of ARP/ND requests. This is often CPU intensive as ARP/ND are usually processed by the CPU (and not in hardware).

Solution: For IPv4 networks, a practice to alleviate this problem is to have the L2/L3 boundary router send periodic gratuitous ARP [GratuitousARP] messages, so that all the connected end stations can refresh their ARP caches. As the result, most (if not all) end stations will not need to ARP for the gateway routers when they need to communicate with external peers.

However, due to IPv6 requiring bi-directional path validation Ipv6 end stations are still required to send unicast ND messages to their default gateway router (even with those routers periodically sending Unsolicited Neighbor Advertisements).

Advantage: Reduction of ARP requests to be processed by L2/L3 boundary router for IPv4.

Disadvantage: No reduction of ND processing on L2/L3 boundary router for IPv6 traffic.

Recommendation: Use for IPv4-only networks, or make change to the ND protocol to allow data frames to be sent without requiring bi-directional frame validation. Some work in progress in this area is [Impatient-NUD]

5.1.2. L2/L3 boundary router processing of inbound traffic

When a L2/L3 boundary router receives a data frame destined for a local subnet and the destination is not in router's ARP/ND cache, some routers hold the packet and trigger an ARP/ND request to resolve the L2 address. The router may need to send multiple ARP/ND requests until either a timeout is reached or an ARP/ND reply is received before forwarding the data packets towards the target's MAC address. This process is not only CPU intensive but also buffer intensive.

Solution: To protect a router from being overburdened by resolving target MAC addresses, one solution is for the router to limit the rate of resolving target MAC addresses for inbound traffic whose target is not in the router's ARP cache. When the rate is exceeded, the incoming traffic whose target is not in the ARP cache is dropped.

For an IPv4 network, another common practice to alleviate this problem is for the router to snoop ARP messages between other hosts, so that its ARP cache can be refreshed with active addresses in the L2 domain. As a result, there is an increased likelihood of the router's ARP cache having the IP-MAC entry when it receives data frames from external peers.

For IPv6 end stations, routers are supposed to send ND unicast even if it has snooped UNA/NS/NA from those stations. Therefore, this practice doesn't help IPv6 very much.

Advantage: Reduction of the number of ARP requests which routers have to send upon receiving IPv4 packets and the number of IPv4 data frames from external peers which routers have to hold.

Disadvantage: The amount of ND processing on routers for IPv6 traffic is not reduced. Even for IPv4, routers still need to hold data packets from external peers and trigger ARP requests if the targets of the data packets either don't exist or are not very active.

Recommendation: This scheme doesn't work with IPv6. For IPv4, if there is higher chance of routers receiving data packets towards non-existing or inactive targets, alternative approaches should be considered.

5.1.3. Inter subnets communications

The router will be hit with ARP/ND twice when the originating and destination stations are in different subnets attached to the same router. Once when the originating station in subnet-A initiates ARP/ND request to the L2/L3 boundary router (5.1.1 above); and the second time when the L2/L3 boundary router to initiates ARP/ND requests to the target in subnet-B (5.1.2 above).

Again, practices described in 5.1.1 and 5.1.2 can alleviate problems in IPv4 network, but don't help very much for IPv6.

Advantage: reduction of ARP processing on L2/L3 boundary routers for IPv4 traffic.

For IPv6 traffic, there is no reduction of ND processing on L2/L3 boundary routers.

Recommendation: Consider the recommended approaches described in 5.1.1 & 5.1.2.

5.2. Static ARP/ND entries on switches

In a datacenter environment the placement of L2 and L3 addressing may be orchestrated by Server (or VM) Management

System(s). Therefore it may be possible for static ARP/ND entries to be configured on routers and / or servers.

Advantage: This methodology has been used to reduce ARP/ND fluctuations in large scale data center networks.

Disadvantage: There is no well-defined mechanism for devices to get prompt incremental updates of static ARP/ND entries when changes occur.

Recommendation: The IETF should consider creating standard mechanism (or protocols) for switches or servers to get incremental static ARP/ND entries updates.

5.3. ARP/ND Proxy approaches

RFC1027 [RFC1027] specifies one ARP proxy approach. Since the publication of RFC1027 in 1987 there have been many variants of ARP proxy being deployed. The term ''ARP Proxy'' is a loaded phrase, with different interpretations depending on vendors and/or environments. RFC1027's ARP Proxy is for a Gateway to return its own MAC address on behalf of the target station. Another technique, also called ''ARP Proxy'' is for a ToR switch to snoop ARP requests and return the target station's MAC if the ToR has the information.

Advantage: Proxy ARP [RFC1027] and its variants have allowed multi-subnet ARP traffic for over a decade.

Disadvantage: Proxy ARP protocol [RFC1027] was developed for hosts which don't support subnets.

Recommendation: Revise RFC1027 with VLAN support and make it scale for Data Center Environment.

6. Practices to scale ARP/ND in Overlay models

There are several drafts on using overlay networks to scale large layer 2 networks (or avoid the need for large L2 networks) and enable mobility (e.g. draft-wkumari-dcops-l3-vmmobility-00, draft-mahalingam-dutt-dcops-vxlan-00). TRILL and IEEE802.1ah (Mac-in-Mac) are other types of overlay network to scale Layer 2.

Overlay networks hide the VMs' addresses from the interior switches and routers, thereby greatly reduces the number of addresses exposed to the interior switches and router. The Overlay Edge nodes which perform the network address

encapsulation/decapsulation still handle all remote stations addresses which communicate with stations attached locally.

For a large data center with many applications, these applications' IP addresses need to be reachable by external peers. Therefore, the overlay network may have a bottleneck at the Gateway devices(s) in processing resolving target stations' physical address (MAC or IP) and overlay edge address within the data center.

Here are some approaches being used to minimize the problem:

1. Use static mapping as described in Section 5.2.

2. Have multiple gateway nodes (i.e. routers), with each handling a subset of stations addresses which are visible to external peers, e.g. Gateway #1 handles a set of prefixes, Gateway #2 handles another subset of prefixes, etc.

7. Summary and Recommendations

This memo describes some common practices which can alleviate the impact of address resolution on L2/L3 gateway routers.

In Data Centers, no single solution fits all deployments. This memo has summarized some practices in various scenarios and the advantages and disadvantages about all of these practices.

In some of these scenarios, the common practices could be improved by creating and/or extending existing IETF protocols. These protocol change recommendations are:

- Extend IPv6 ND method,

- Create a incremental ''update'' schemes for static ARP/ND entries,

- Revise Proxy ARP [RFC1027] for use in the data center.

8. Security Considerations

This draft documents existing solutions and proposes additional work that could be initiated to extend various

IETF protocols to better scale ARP/ND for the data center environment. The security of future protocol extension will be discussed in their respective documents.

9. IANA Considerations

This document does not request any action from IANA.

10. Acknowledgements

We want to acknowledge ARMD WG and the following people for their valuable inputs to this draft: Susan Hares, Benson Schliesser, T. Sridhar, Ron Bonica, Kireeti Kompella, and K.K.Ramakrishnan.

11. References

11.1. Normative References

[ARMD-Problem] Narten, ''Problem Statement for ARMD''(http://datatracker.ietf.org/doc/draft-ietf-armd-problem-statement/); Aug 2012

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

[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/)

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

11.2. Informative References

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

Authors' Addresses

Linda Dunbar
Huawei Technologies
5340 Legacy Drive, Suite 175
Plano, TX 75024, USA
Phone: (469) 277 5840
Email: ldunbar@huawei.com

Warren Kumari
Google
1600 Amphitheatre Parkway
Mountain View, CA 94043
US
Email: warren@kumari.net

Igor Gashinsky
Yahoo
45 West 18th Street 6th floor
New York, NY 10011
Email: igor@yahoo-inc.com