ARMD
Internet Draft
Intended status: Information Track
Expires: July 2012

L. Dunbar
Huawei
W. Kumari
Google
I. Gashinsky
Yahoo

January 3, 2012

BCP for ARP-ND Scaling for Large Data Centers

draft-dunbar-armd-arp-nd-scaling-bcp-00.txt

Status of this Memo

Copyright Notice

Abstract

This draft is intended to document some simple well established practices which can scale ARP/ND in data center environment.

Conventions used in this document

The key words "MUST", "MUST NOT", "REQUIRED", "SHALL", "SHALL NOT", "SHOULD", "SHOULD NOT", "RECOMMENDED", "MAY", and "OPTIONAL" in this document are to be interpreted as described in RFC-2119 0.

Table of Contents

1. Introduction

As described in [ARMD-Problems], the increasing trend of rapid workload shifting and server virtualization in modern data centers is requiring servers to be loaded (or re-loaded) with different hosts or applications at different times. Those different hosts loaded to one physical server may have different IP addresses, or even be in different IP subnets.
In order to allow a physical server to be re-loaded with hosts in different subnets, or VMs to be moved to different server racks without IP address re-configuration, the corresponding networks have to have multiple broadcast domains (many VLANs) on the interfaces of L2/L3 boundary routers and ToR switches. Unfortunately, this kind of network can lead to address resolution scaling issues, especially on the L2/L3 boundary routers, when the combined number of hosts in all those subnets is large.
This document describes some potential solutions which can minimize the ARP/ND scaling issues in a Data Center environment.

2. Terminology

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

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. It is also known as access switch.

UNA: IPv6's Unsolicited Neighbor Advertisement

VM: Virtual Machines

3. Potential Solutions to Scale Address Resolution in DC

The following solutions have been indicated by data center operators:

1) layer-3 connectivity to the access switch,

2) practices to scale ARP/ND in layer 2,

3) static ARP/ND entries,

4) DNS based approaches, and

5) Extensions to proxy ARP [RFC1027].

There is no single solution that fits all cases. This section suggests the best practices for each type of solution.

3.1. Layer 3 solution

This is referring to the network design with Layer 3 to the access switches.

As described in [ARMD-Problem], many data centers are designed this way, 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 all the way to servers, or even to the VMs. Then the ARP/ND broadcast/multicast

messages are further confined to the small number of hosts within the server, or none at all.

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

Disadvantage: The main disadvantage to this solution is that IP addresses have to be re-configured on switches when a server needs to be re-loaded with an application in different subnet, or VMs need to be moved to a different location.

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

3.2. Commonly practiced Layer 2 solution to scale address resolution

L2/L3 boundary routers can be heavily impacted by the ARP/ND broadcast/multicast messages in a Layer 2 domain which is mapped to one or multiple subnets (or VLANs) with combined large number of hosts in all subnets. This section describes some commonly used practices in reducing the ARP/ND processing required on L2/L3 boundary (or gateway) routers.

3.2.1. When a host needs to communicate with an external peer:

When the external peer is in a different subnet, the originating host needs to send ARP/ND requests to its default gateway router to get router's MAC address. If there are many subnets enabled on the gateway router with large combined number of hosts in all those subnets, the gateway router has to process a very large number of ARP/ND requests, which is CPU intensive.

Solution: For IPv4 networks, a common practice to alleviate this problem is to have the L2/L3 boundary router (or gateway router) send periodic gratuitous ARP messages, so that all the connected hosts can refresh their ARP caches. As the result, most hosts, if not all, won't send ARP messages to gateway routers when they need to communicate with external hosts.

However, IPv6 hosts are still required to send ND messages, via

unicast, to their default gateway router even with their gateway routers periodically sending Unsolicited Neighbor Advertisement. This is due to IPv6 requiring bi-directional path validation before a data packet can be sent.

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 change the ND protocol to allow data frames to be sent without requiring bidirectional frame validation.

3.2.2. When the L2/L3 boundary router receives an IP packet towards a host in one of its subnets:

When the source address is in a different subnet and the target is not in router's ARP/ND cache, the router usually holds the packet and triggers an ARP/ND request to make sure the target actually exists in its L2 domain. The router may need to send multiple ARP/ND requests until either a timeout is reached or an ARP/ND reply is received. After this the gateway router can forward the data packets towards the target's MAC address. This process is not only CPU intensive but also buffer intensive.

Solution: For IPv4 network, a common practice to alleviate this problem is by an L2/L3 boundary router (or gateway router) snooping ARP messages, so that its ARP cache can be refreshed with active hosts in its 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 subnets.

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

Advantage: Reduction of ARP requests which routers have to send upon receiving IPv4 packets and the amount of IPv4 data frames from external subnets 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 subnets and trigger ARP requests if the targets of the data packets either don't exist or are not very active.

Recommendation: Do not use with IPv6 or make protocol changes to IPv6's ND. For IPv4, if there are higher chance of routers receiving data packets towards non-existing or inactive targets, alternative approaches should be considered.

3.2.3. Hosts in two different subnets served by the router communicate with each other

The router will be hit twice under this scenario. Once for the originating host in subnet-A initiating ARP/ND request to the gateway (3.2.1 above); and the second for the gateway to initiate ARP/ND requests to the target in subnet-B (3.2.2 above).

Again, practices described in 3.2.1 and 3.2.2 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.

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

Recommendation: do not use with IPv6 or consider other approaches.

3.3. Static ARP/ND entries on switches

In a data center environment, applications placement to servers, racks, and rows may be orchestrated by Server (or VM) Management System(s). Therefore it is possible for static ARP/ND entries to be downloaded to switches, routers or servers.

Advantage: This methodology has been used to reduce ARP/ND

fluctuations in large scale deployments.

Disadvantage: There is no well defined mechanism for switches to get static ARP/ND entries, to get prompt update of static ARP/ND entries when changes occur, or to perform certain steps when switches go through reset.

Recommendation: The IETF should create a well-defined mechanism (or protocols) for switches or servers to get static ARP/ND entries.

3.4. DNS based solution

This solution is best suited to environments where applications resolve the address of things they need to connect to via DNS, and periodically refresh these addresses. While this solution is very well known, and extensively used, it is mainly appropriate for stateless services, or for services that have a large number of short

lived connections. While elegant, it may not be appropriate for generic host migration.
. When a VM is to be moved to a new location, here are the steps in getting the IP addresses:

- Instantiate the service on a VM in a distant rack. The new VM gets a new IP address
- Change the address of the service in DNS
- Wait for the DNS TTL to expire. While you are waiting, watch the number of connections to the new VM increase and the number of connections to the old VM decrease.
- Wait a little longer. When the number of connections to the old VM reaches zero, shut down the old VM.

Advantage: DNS is existing technology and this is a well-known, commonly practiced technique.

Disadvantage: This approach is not suitable for multi-tenant scenarios, or when the data center operators does not have full control of the applications.

Recommendation: Limited use to where the data-center operators are in control of the entire application and runs the DNS. More appropriate for service migration than host / VM migration..

3.5. ARP/ND Proxy approaches

RFC1027 specifies one ARP proxy approach. Since RFC1027, which was published 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 host. Another technique, also called "ARP Proxy" is for a ToR switch to snoop ARP requests and return the target hosts MAC if it knows it. .

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

Disadvantage: Proxy ARP protocol [RFC1027] was developed prior to the concepts of VLANs and for hosts which don't support subnets, and does not provide the scaling.

Recommendation: Revise RFC1027 with VLAN support and scalability for the Data Center Environment.

3.6. Overlay models

There are several drafts on using overlay networks to scale large layer 2 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 hosts' addresses form the interior switches and routers. The Overlay Edge nodes which perform the network address encapsulation/decapsulation still see all remote hosts addresses which communicate with hosts attached locally.

For a large data center with tens of thousands of applications communicating with peers outside the data center, all those applications' IP addresses are visible to external peers. When a great number of VMs move freely within a data center, all those VMs' IP addresses might not be aggregated very nicely on gateway routers, causing forwarding table size exploding.

When the Gateway router receives a data frame from external peers destined to a target within the data center, routers need to resolve

target's MAC address and the Overlay Edge node's address in order to perform the proper overlay encapsulation.

Therefore, the overlay network will have a bottleneck at the Gateway router(s) in processing resolving target hosts' 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 3.3.

2. Have multiple gateway nodes (i.e. routers), with each handling a subset of hosts addresses which are visible to external peers, e.g. Gateway #1 handles a set of prefixes, Gateway #2 handles another subset of prefixes, etc. This architecture assumes that each gateway have enough downstream ports to be connected to all server racks.

   If each server rack is allowed to instantiate hosts/applications with any IP addresses, or allowing any VM to move anywhere without re-configuring IP/MAC addresses, each gateway has to resolve addresses which are potentially located on any server rack. The address resolution processing for each gateway can still be very heavy.

4. Summary and Recommendations

This memo describes some best practices which can alleviate impact of address resolution to L2/L3 gateway routers.

In the Data Center, no single solution fits all deployments. This memo has summarized five different technologies and the advantages and disadvantages about all of these practices.

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

Extend IPv6 ND method,

Create a "download" static ARP/ND entry protocol,

Revise Proxy ARP [1027] for use in the data center.

5. Manageability Considerations

This text gives recommendations for best practices in order to improve manageability of DC.

6. Security Considerations

Security will be addressed in a separate document.

7. IANA Considerations

None.

8. Acknowledgments

We want to acknowledge the following people for their valuable inputs to this draft: K.K.Ramakrishnan.

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

9. References

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

[DC-ARCH] Karir,et al, "draft-karir-armd-datacenter-reference-arch"

[ARMD-Problem] Narten, "draft-ietf-armd-problem-statement" in progress, Oct 2011.

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

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

Intellectual Property Statement

---

users of this specification can be obtained from the IETF on-line IPR repository at http://www.ietf.org/ipr

The IETF invites any interested party to bring to its attention any copyrights, patents or patent applications, or other proprietary rights that may cover technology that may be required to implement any standard or specification contained in an IETF Document. Please address the information to the IETF at ietf-ipr@ietf.org.

Disclaimer of Validity

Acknowledgment

Funding for the RFC Editor function is currently provided by the Internet Society.