Independent Submission
Request for Comments: 7342
Category: Informational
ISSN: 2070-1721

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

August 2014

Practices for Scaling ARP and Neighbor Discovery (ND) in Large Data Centers

Abstract

This memo documents some operational practices that allow ARP and Neighbor Discovery (ND) to scale in data center environments.

Status of This Memo

This document is not an Internet Standards Track specification; it is published for informational purposes.

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/rfc7342.

Table of Contents

1. Introduction

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

As described in [RFC6820], the increasing trend of rapid workload shifting and server virtualization in modern data centers requires servers to be loaded (or reloaded) with different Virtual Machines (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 allow VMs to be moved to different server racks without IP address reconfiguration, the networks need to enable multiple broadcast domains (many VLANs) on the interfaces of L2/L3 boundary routers and Top-of-Rack (ToR) switches and allow some subnets to span multiple router ports.

Note: L2/L3 boundary routers as discussed in this document are capable of forwarding IEEE 802.1 Ethernet frames (Layer 2) without a Media Access Control (MAC) header change. When subnets span multiple ports of those routers, they still fall under the category of "single-link" subnets, specifically the multi-access link model

recommended by [RFC4903]. They are different from the "multi-link" subnets described in [Multi-Link] and RFC 4903, which refer to different physical media with the same prefix connected to one router. Within the "multi-link" subnet described in RFC 4903, Layer 2 frames from one port cannot be natively forwarded to another port without a header change.

Unfortunately, when the combined number of VMs (or hosts) in all those subnets is large, this can lead to address resolution (i.e., IPv4 ARP and IPv6 ND) scaling issues. There are three major issues associated with ARP/ND address resolution protocols when subnets span multiple L2/L3 boundary router ports:

1) The ARP/ND messages being flooded to many physical link segments, which can reduce bandwidth utilization for user traffic.

2) The ARP/ND processing load impact on the L2/L3 boundary routers.

3) In IPv4, every end station in a subnet receiving ARP broadcast messages from all other end stations in the subnet. IPv6 ND has eliminated this issue by using multicast.

Since the majority of data center servers are moving towards 1G or 10G ports, the bandwidth taken by ARP/ND messages, even when flooded to all physical links, becomes negligible compared to the link bandwidth. In addition, IGMP/MLD (Internet Group Management Protocol and Multicast Listener Discovery) snooping [RFC4541] can further reduce the ND multicast traffic to some physical link segments.

As modern servers’ computing power increases, the processing taken by a large amount of ARP broadcast messages becomes less significant to servers. For example, lab testing shows that 2000 ARP requests per second only takes 2% of a single-core CPU server. Therefore, the impact of ARP broadcasts to end stations is not significant on today’s servers.

Statistics provided by Merit Network [ARMD-Statistics] have shown that the major impact of a large number of mobile VMs in a data center is on the L2/L3 boundary routers, i.e., issue 2 above.

This memo documents some simple practices that can scale ARP/ND in a data center environment, especially in reducing processing loads to L2/L3 boundary routers.

2. Terminology

This document reuses much of the terminology from [RFC6820]. 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 document, the term "Bridge" is used interchangeably with "Layer 2 switch"

DC: Data Center

DA: Destination Address

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

EoR: End-of-Row switches in a data center

NA: IPv6 Neighbor Advertisement

ND: IPv6 Neighbor Discovery [RFC4861]

NS: IPv6 Neighbor Solicitation

SA: Source Address

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

UNA: IPv6 Unsolicited Neighbor Advertisement

VM: Virtual Machine

Subnet: Refers to the multi-access link subnet referenced by RFC 4903

3. Common DC Network Designs

Some common network designs for a 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 configures Layer 3 to the access switches, effectively making the access switches the L2/L3 boundary routers for the attached VMs.

As described in [RFC6820], 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 a Layer 3 infrastructure 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 of this network design occurs during VM movement. During VM movement, either VMs need an address change or switches/routers need a configuration change when the VMs are moved to different locations.

Summary: This solution is more suitable to data centers that have a static workload and/or network operators who can reconfigure 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 for reducing the ARP/ND processing required on L2/L3 boundary routers.

5.1.1. Communicating with a Peer in a Different Subnet

Scenario: When the originating end station doesn't have its default gateway MAC address in its ARP/ND cache and needs to communicate with a peer in a different subnet, it 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 that don't have the gateway MAC address in their ARP/ND caches, the gateway router has to process a very large number of ARP/ND requests. This is often CPU intensive, as ARP/ND messages are usually processed by the CPU (and not in hardware).

Note: Any centralized configuration that preloads the default MAC addresses is not included in this scenario.

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 a result, most (if not all) end stations will not need to send ARP requests for the gateway routers when they need to communicate with external peers.

For the above scenario, 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) because IPv6 requires bidirectional path validation.

Advantage: This practice results in a reduction of ARP requests to be processed by the L2/L3 boundary router for IPv4.

Disadvantage: This practice doesn't reduce ND processing on the L2/L3 boundary router for IPv6 traffic.

Recommendation: If the network is an IPv4-only network, then this approach can be used. For an IPv6 network, one needs to consider the work described in [RFC7048]. Note: ND and Secure Neighbor Discovery (SEND) [RFC3971] use the bidirectional nature of queries to detect and prevent security attacks.

5.1.2. L2/L3 Boundary Router Processing of Inbound Traffic

Scenario: When an L2/L3 boundary router receives a data frame destined for a local subnet and the destination is not in the 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/ND cache. When the rate is exceeded, the incoming traffic whose target is not in the ARP/ND cache is dropped.

For an IPv4 network, another common practice to alleviate pain caused by 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. [RFC6820] Section 7.1 provides a full description of this problem.

For IPv6 end stations, routers are supposed to send Router Advertisements (RAs) unicast even if they have snooped UNAs/NSs/NAs from those stations. Therefore, this practice allows an L2/L3 boundary to send unicast RAs to the target instead of multicasts. [RFC6820] Section 7.2 has a full description of this problem.

Advantage: This practice results in a reduction of the number of ARP requests that routers have to send upon receiving IPv4 packets and the number of IPv4 data frames from external peers that routers have to hold due to targets not being in the ARP cache.

Disadvantage: The amount of ND processing on routers for IPv6 traffic is not reduced. 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. In this case, IPv4 processing or IPv4 buffers are not reduced.

Recommendation: If there is a higher chance of routers receiving data packets that are destined for nonexistent or inactive targets, alternative approaches should be considered.

5.1.3. Inter-Subnet Communications

The router could be hit with ARP/ND requests twice when the originating and destination stations are in different subnets attached to the same router and those hosts don't 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 (Section 5.1.1 above), 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 the router's ARP/ND cache (Section 5.1.2 above).

Again, practices described in Sections 5.1.1 and 5.1.2 can alleviate some problems in some IPv4 networks.

For IPv6 traffic, the practices described above don't reduce the ND processing on L2/L3 boundary routers.

Recommendation: Consider the recommended approaches described in Sections 5.1.1 and 5.1.2. However, any solutions that relax the bidirectional requirement of IPv6 ND disable the security that the two-way ND communication exchange provides.

5.2. Static ARP/ND Entries on Switches

In a data center 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: When some VMs are added, deleted, or moved, many switches' static entries need to be updated. In a data center with virtualized servers, those events can happen frequently. For example, for an event of one VM being added to one server, if the subnet of this VM spans 15 access switches, all of them need to be updated. Network management mechanisms (SNMP, the Network Configuration Protocol (NETCONF), or proprietary mechanisms) are available to provide updates or incremental updates. However, there is no well-defined approach for switches to synchronize their content with the management system for efficient incremental updates.

Recommendation: Additional work may be needed within IETF working groups (e.g., NETCONF, NVO3, I2RS, etc.) to get prompt incremental updates of static ARP/ND entries when changes occur.

5.3. ARP/ND Proxy Approaches

RFC 1027 [RFC1027] specifies one ARP Proxy approach referred to as "Proxy ARP". However, RFC 1027 does not discuss a scaling mechanism. Since the publication of RFC 1027 in 1987, many variants of Proxy ARP have been deployed. RFC 1027's Proxy ARP technique allows a gateway to return its own MAC address on behalf of the target station.

[ARP_Reduction] describes a type of "ARP Proxy" that allows a ToR switch to snoop ARP requests and return the target station's MAC if the ToR has the information in its cache. However, [RFC4903] doesn't recommend the caching approach described in [ARP_Reduction] because such a cache prevents any type of fast mobility between Layer 2 ports and breaks Secure Neighbor Discovery [RFC3971].

IPv6 ND Proxy [RFC4389] specifies a proxy used between an Ethernet segment and other segments, such as wireless or PPP segments. ND Proxy [RFC4389] doesn't allow a proxy to send NA messages on behalf of the target to ensure that the proxy does not interfere with hosts moving from one segment to another. Therefore, the ND Proxy [RFC4389] doesn't reduce the number of ND messages to an L2/L3 boundary router.

Bottom line, the term "ARP/ND Proxy" has different interpretations, depending on vendors and/or environments.

Recommendation: For IPv4, even though those Proxy ARP variants (not RFC 1076) have been used to reduce ARP traffic in various environments, there are many issues with caching.

The IETF should consider making proxy recommendations for data center environments as a transition issue to help DC operators transitioning to IPv6. Section 7 of [RFC4389] ("Guidelines to Proxy Developers") should be considered when developing any new proxy protocols to scale ARP.

5.4. Multicast Scaling Issues

Multicast snooping (IGMP/MLD) has different implementations and scaling issues. [RFC4541] notes that multicast IGMPv2/v3 snooping has trouble with subnets that include IGMPv2 and IGMPv3. [RFC4541] also notes that MLDv2 snooping requires the use of either destination MAC (DMAC) address filtering or deeper inspection of frames/packets to allow for scaling.

MLDv2 snooping needs to be re-examined for scaling within the DC. Efforts such as IGMP/MLD explicit tracking [IGMP-MLD-Tracking] for downstream hosts need to provide better scaling than IGMP/MLDv2 snooping.

6. Practices to Scale ARP/ND in Overlay Models

There are several documents on using overlay networks to scale large Layer 2 networks (or avoid the need for large L2 networks) and enable mobility (e.g., [L3-VM-Mobility], [VXLAN]). Transparent Interconnection of Lots of Links (TRILL) and IEEE 802.1ah (Mac-in-Mac) are other types of overlay networks that can scale Layer 2.

Overlay networks hide the VMs' addresses from the interior switches and routers, thereby greatly reducing the number of addresses exposed to the interior switches and router. The overlay edge nodes that perform the network address encapsulation/decapsulation still handle all remote stations' addresses that communicate with the locally attached end stations.

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 node(s) in processing resolving target stations' physical addresses (MAC or IP) and the overlay edge address within the data center.

Here are two approaches that can be used to minimize this problem:

1. Use static mapping as described in Section 5.2.

2. Have multiple L2/L3 boundary nodes (i.e., routers), with each handling a subset of stations' addresses that 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 that 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 of 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:

- Relax the bidirectional requirement of IPv6 ND in some environments. However, other issues will be introduced when the bidirectional requirement of ND is relaxed. Therefore, it is necessary to have performed a comprehensive study of possible issues prior to making those changes.

- Create an incremental "update" scheme for efficient static ARP/ND entries.

- Develop IPv4 ARP/IPv6 ND Proxy standards for use in the data center. Section 7 of [RFC4389] ("Guidelines to Proxy Developers") should be considered when developing any new proxy protocols to scale ARP/ND.

- Consider scaling issues with IGMP/MLD snooping to determine whether or not new alternatives can provide better scaling.

8. Security Considerations

This memo 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.

Security is a major issue for data center environments. Therefore, security should be seriously considered when developing any future protocol extensions.

9. Acknowledgements

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

10. References

10.1. Normative References

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

[RFC826] 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, November 1982.

[RFC1027] Carl-Mitchell, S. and J. Quarterman, "Using ARP to implement transparent subnet gateways", RFC 1027, October 1987.

[RFC3971] Arkko, J., Ed., Kempf, J., Zill, B., and P. Nikander, "SEcure Neighbor Discovery (SEND)", RFC 3971, March 2005.

[RFC4389] Thaler, D., Talwar, M., and C. Patel, "Neighbor Discovery Proxies (ND Proxy)", RFC 4389, April 2006.

[RFC4541] Christensen, M., Kimball, K., and F. Solensky, "Considerations for Internet Group Management Protocol (IGMP) and Multicast Listener Discovery (MLD) Snooping Switches", RFC 4541, May 2006.

[RFC4861] Narten, T., Nordmark, E., Simpson, W., and H. Soliman, "Neighbor Discovery for IP version 6 (IPv6)", RFC 4861, September 2007.

[RFC4903] Thaler, D., "Multi-Link Subnet Issues", RFC 4903, June 2007.

[RFC6820] Narten, T., Karir, M., and I. Foo, "Address Resolution Problems in Large Data Center Networks", RFC 6820, January 2013.

10.2. Informative References

[ARMD-Statistics]
Karir, M. and J. Rees, "Address Resolution Statistics", Work in Progress, July 2011.

[ARP_Reduction]
Shah, H., Ghanwani, A., and N. Bitar, "ARP Broadcast Reduction for Large Data Centers", Work in Progress, October 2011.

[IGMP-MLD-Tracking]
Asaeda, H., "IGMP/MLD-Based Explicit Membership Tracking Function for Multicast Routers", Work in Progress, December 2013.

[L3-VM-Mobility]
Kumari, W. and J. Halpern, "Virtual Machine mobility in L3 Networks", Work in Progress, August 2011.

[Multi-Link]
Thaler, D. and C. Huitema, "Multi-link Subnet Support in IPv6", Work in Progress, June 2002.

[RFC1076] Trewitt, G. and C. Partridge, "HEMS Monitoring and Control Language", RFC 1076, November 1988.

[RFC7048] Nordmark, E. and I. Gashinsky, "Neighbor Unreachability Detection Is Too Impatient", RFC 7048, January 2014.

[VXLAN] Mahalingam, M., Dutt, D., Duda, K., Agarwal, P., Kreeger, L., Sridhar, T., Bursell, M., and C. Wright, "VXLAN: A Framework for Overlaying Virtualized Layer 2 Networks over Layer 3 Networks", Work in Progress, April 2014.

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
USA

EMail: warren@kumari.net

Igor Gashinsky
Yahoo
45 West 18th Street 6th floor
New York, NY 10011
USA

EMail: igor@yahoo-inc.com