Working Group: ARMD
Intended Status: Informational
Internet Draft

Expiration Date: April 27, 2012

Himanshu Shah
Ciena Corp

Anoop Ghanwani
Brocade

Nabil Bitar
Verizon

October 28, 2011

ARP Broadcast Reduction for Large Data Centers
draft-shah-armd-arp-reduction-02.txt

Status of this Memo

This Internet-Draft is submitted 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/1id-abstracts.html

The list of Internet-Draft Shadow Directories can be accessed at http://www.ietf.org/shadow.html

This Internet-Draft will expire on April 27, 2012

Copyright Notice

Abstract

With advent of server virtualization technologies, a host is able to support multiple Virtual Machines (VMs) in a single physical machine. Data Centers can leverage these capabilities to instantiate on the order of 10s to 100s of VMs in a single server with current technology. It is conceivable that this number can be much higher in the future. Each VM operates as an independent IP host with a set of Virtual Network Interface Cards (vNICs), each having its own MAC address and mapping to a physical Ethernet interface. These physical servers are typically installed in a rack with their Ethernet interfaces connected to a top-of-the-rack (ToR) switch. The ToR switches are interconnected through End-of-the-Row (EoR) or aggregation switches which are in turn connected to core switches.

As discussed in [ARP-Problem] the host VMs use ARP broadcasts to find other host VMs and use periodic (broadcast) Gratuitous ARPs to refresh their IP to MAC address binding in other VM hosts. Such broadcasts in a large data center with potentially thousands of VM hosts in a Layer 2 based topology can overwhelm the network.

This memo proposes mechanisms to reduce the number of broadcasts that are sent throughout the network. This is done by having the ToRs intelligently process ARP and frames, rather than simply broadcasting them throughout the broadcast domain.

While this document addresses ARP, the Neighbor Discovery mechanisms used by the IPv6 hosts that make use of multicast rather than broadcast also pose similar issues in the Data Center. The solutions defined herein should be equally applicable to hosts running IPv6. The details will be specified in a subsequent revision.

Conventions

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 [RFC 2119].

Table of Contents

1.0 Overview

The traditional topology in a data center consists of racks of servers connected to top-of-rack (ToR) switches, which connect to aggregation switches, which in turn connect to core switches. The network architecture typically combines Layer 2 and Layer 3. In some architectures, Layer 2 is terminated at the ToR, with Layer 3 being run in the aggregation and core devices. In other architectures, Layer 2 may be extended all the way to the aggregation switch. The primary concerns that have influenced network architectures in the data center have been keeping broadcast domains manageable and spanning tree domains contained.

Moving forward, these traditional network architectures are being challenged due to emerging technologies such as server virtualization.

The effect of server virtualization in the data center brings some challenges. Because of virtualization, the number of hosts that the network sees increases dramatically - 10 to 100 times the number of physical servers. These virtual hosts are referred to as Virtual machines (VMs). VMs offer server mobility wherein a VM can be relocated to run on a different physical server. In order for the mobility to be non-disruptive to other hosts that have communication in progress with the VM being moved, the VM must retain its MAC address and IP address. Because of the requirement to retain the MAC and IP address, it is desirable to develop network architectures that would offer the least restrictions in terms of server mobility.

As an example, in a network architecture where TOR switches terminate the L2 domain, the range of mobility would be restricted to a single ToR switch. It would be more preferable to allow the flexibility of moving the VM anywhere within the data center, or perhaps even a different data center.

Technologies such as TRILL [TRILL] overcome some of the issues of spanning trees because which traditional Layer 2 topologies have been constrained. However, because of virtualization there are 2 specific problems that are introduced with respect to broadcast traffic.

1. A larger number of hosts. A single physical server now hosts multiple virtual machines taking the scale factor to a different level. If each VM has the same number of broadcasts as a physical server, the amount of broadcast traffic has increased 10 to greater than 100 times.
2. If the Layer 2 domains are extended to go across data centers, then broadcast traffic will now go across the backbone. If Layer 2 was terminated at the ToR switch, the increase in broadcast traffic would be been restricted to a single ToR switch, but as discussed earlier, this restriction is not desirable.

The broadcast as such in Layer 2 networks has far reaching impacts; i.e. wastage in network bandwidth as well as CPU resources used by all the VMs while processing superfluous ARP broadcasts (IPv6 gets rid of the latter by running ND as a multicast service rather than a broadcast service).

The solution presented here attempts to minimize negative effects of ARP broadcasts. The solution requires the first hop Ethernet switches, typically ToR, to maintain an ARP table learned from the ARP PDUs received by the switch and selectively propagates the ARP

to, or proxy-responds on behalf of, the remote peer. These types of ARP processing principles are well known and used/described in L2VPN Working Group documents such as [ARP-Mediation] and [IPLS]. The ARP proxy response differs from that described in [RFC1027] as the ARP response contains MAC address of the destination and not that of the switch as is suggested in [RFC 1027].

The following sections describe the details of ARP snooping, learning and maintaining ARP tables, using the learned information to limit broadcast propagation and proxy (the response) on behalf of the remote peers.

1.1 Terminology

ToR switch	Top-of-Rack switch. An Ethernet switch installed at the top of a rack of servers which provides network connectivity to those servers.
Downlink	The Ethernet link between the ToR switch and a directly connected host/server in the rack.
Uplink	The network-facing Ethernet connection in the ToR switch. Typically, the uplinks from ToRs connect to end-of-row or aggregation switches.
EoR switch	End-of-Row switch. An Ethernet switch which aggregates traffic from multiple racks. Also commonly referred to as an aggregation switch. Uplinks from the ToR connects to EoR switches and uplinks from EoR switches in turn connect to core switches.
Host/Server	A host or server running the IP protocol. This could be a physical entity or a logical entity (such as a Virtual Machine) in a physical host. The term server refers to its role in data center. Both terms are used interchangeably and refer to an IP end station.
Local hosts	Used in the context of a ToR switch to denote the VM hosts connected to a ToR switch on the downlink, i.e. directly connected hosts.

Remote hosts	Used in the context of a ToR switch to denote the hosts that are accessible via the uplink of the ToR switch.
VM	Virtual Machine. This is a logical instance of a host that operates independently in a physical host and has its own IP and MAC addresses. The VM architecture allows efficient use of physical host resources (such as multiple CPU cores).

2.0 Configuration

It is assumed that ARP reduction methodologies that are defined in this document will be limited to ToR switches. The maximum benefit of restraining ARP broadcasts in the network is achieved by the first hop switches (the ones directly connected to the hosts) without placing additional burden on second or third tier switches.

First, the ToR switches would need to be configured in order to enable the ARP reduction feature. Every Ethernet interface needs to be identified as either a downlink or uplink within the context of this feature. The ARP reduction feature treats ARP frames received from downlink or uplink differently as described in the following sections.

In additional the operator may optionally configure various ARP reduction related parameters such as:

- ARP aging timer,
- size of the ARP table,
- static entries of IP to MAC address, etc.

3.0 Building the ARP tables

When ARP reduction is enabled, the ToR switch will monitor all ARP traffic transiting the switch (regardless of uplink port or downlink port) and will process any ARP PDUs in the following manner:

- ARP Request PDUs must be redirected to control plane CPU.
- Gratuitous ARP PDUs (ARP Reply PDU with a broadcast MAC DA) must be redirected to control plane CPU.
- Other ARP Reply PDUs (ARP Reply PDU with a unicast MAC DA) should be bi-casted; one copy sent to control plane CPU and

other copy forwarded out normally.

3.1 ARP Requests

The ToR examines the source IP and the source hardware address (MAC address) in the ARP Request . The source IP and MAC address association is learned, or is updated/refreshed if already learned. The destination IP address is searched in the ARP table. If an entry exists, the associated MAC address from the table is used to prepare a unicast ARP Reply PDU. The same MAC address is used as the source MAC address in the MAC header, as well as for the target hardware address,in the unicast ARP Reply PDU.

If the destination IP address in the request is not present in the ARP table, then the original ARP request PDU is broadcast to all the switch ports that are member of the same VLAN except the source port that the Request was received from. However, if the requested

(destination) IP address is present in the ARP table, a unicast ARP Reply PDU is prepared as described above and sent to the switch port from which the ARP Request was received and original ARP request PDU is dropped.

The intent is to prevent propagation of ARP Request PDU broadcasts as much as possible using the information present in the ARP table. The following observations can be made from such behavior.

- Most of the ARP requests from the local hosts of a ToR switch for the local hosts of the ToR switch can be prevented.
- Most of the ARP requests from the remote hosts of a ToR switch for the local hosts of the ToR switch can be prevented from getting forwarded on downlinks or other uplinks of the ToR switch.
- Many of the ARP requests from the local hosts of a ToR switch for the remote hosts of the ToR switch can be prevented from being forwarded on uplinks if the remote host IP to MAC association is known to the ToR switch.

3.2 ARP Reply

The unicast ARP Reply is examined to learn/update the ARP table for source and destination IP/MAC address association, but is also forwarded out as a normal frame.

3.3 Gratuitous ARP

Gratuitous ARP is a broadcast ARP Reply PDU with destination IP

address set to the IP address of the sender and target hardware address set to the MAC address of the sender. It is typically used by the IP hosts (including VMs) to keep its association fresh in peer's ARP cache.

The ToR switch should process Gratuitous ARP in the following manner.

- Learn/update/refresh the ARP table entry.
- If the IP address is new, or exists but with a different hardware address, then the Gratuitous ARP PDU is forwarded out; otherwise the PDU is discarded.

The goal for handling of the Gratuitous ARP PDU received from the downlinks (i.e. local hosts) is to avoid propagating it into the 'network' (i.e. to uplinks), unless there is a new association.

By suppressing the propagation of Gratuitous ARP PDUs, the peer IP hosts will end up aging out the corresponding ARP table entries. This will result in generation of the broadcast ARP Requests by those IP hosts if they need to continue to communicate with the IP host whose Gratuitous ARPs were obstructed. The handling of the ARP Request, as described above, by the first hop ToR switch will be able to respond to this request based on the ARP cache maintained in the ToR switch. In essence, presence of large ARP tables with longer age out times compensates for the smaller ARP table present in the

IP hosts and eliminates the need for periodic use of Gratuitous ARPs in order to refresh the ARP table in the IP hosts.

3.4 Host movement

As mentioned earlier, server virtualization technology allows movement of VMs to different physical servers. The flexibility to move VMs is one of the key benefits of server virtualization. The VM movement could be manual (operator initiated) or may be done automatically in reaction to demands placed by the application users. The important point is that in either case, VM movement is not transparent and is made known to the network.

There is ongoing work in IEEE 802.1 standards organization (IEEE 802.1Qbg) to coordinate/communicate the presence and capabilities of the VMs to the directly connected network switch.

VMs typically retain their MAC and IP address, and as such, there would be little impact to the ARP table maintained by the ARP reduction mechanism described herein. However, the ARP reduction mechanism would benefit from knowing if a VM is completely

decommissioned so that the ToR can removed the ARP entry it has for that VM in a timely fashion, rather than waiting for it to timeout.

3.5 Applicability to environments with overlay transport

Recently, there have been multiple proposals for using overlay transport technologies such as VXLAN [VXLAN] and NVGRE [NVGRE]. These proposals allow the network operator to build the network using L2 or L3 technologies while building an L2-overlay on top of that. As such, while they address the issue of network design, they do not eliminate the need for a mechanism to reduce the amount of broadcast traffic that may have to traverse the core, if there are VMs of the same tenant on servers attached to different ToR switches.

One of the ways for the overlay transport proposals to address this issue would be to implement the mechanism discussed in this document at the point where the overlay encapsulation and decapsulation is performed (i.e. in the virtual switch).

3.6 Scaling Considerations

Depending on the number of hosts in the networks, the ARP table can be quite large. Although it is possible to implement some of the mechanisms for ARP reduction as described in this document in hardware in the forwarding plane, the number of ARP entries may favor maintaining the ARP table in the control plane memory.

3.7 Miscellaneous Issues

Because of the distributed nature of the mechanisms described herein, there are a few additional issues that warrant consideration from the network operator.

Earlier in the document, we had mentioned the configuration of a timer for ARP entries. A longer timer for holding on to ARP entries helps with reduction of broadcasts. However, the risk of having a "too large timer" can cause problems in certain situations. Consider the following scenario. Host A is attached to ToR switch #1, and host B is attached to ToR switch #2. If host B issues an ARP request for host A, if the entry is available at switch #2, then switch #2 would send the ARP Reply on behalf of host A. It is possible that host A is no longer available, but there is no way for switch #2 to know this, and it would continue to respond on behalf

of host A, until its entry for host A has timed out. In this case, it is easy to see that a smaller timer would be beneficial. Additionally, since host B has an ARP age timer, it means that host B would find out about host A's unavailability only after its entry has aged, which would be after it has aged out of switch #2.

Another issue that can be somewhat problematic could be the inconsistency of tables in switches. Once again, consider a scenario similar to the one described above with 2 hosts each connected to its respect ToR switch. Let the ARP entries at both A and B be learned by both switches. Now assume that the IP address on host A changes. This change is signaled to switch #1 which in turn broadcasts the message on its uplink. Now, if this message is discarded due to network congestion or signal integrity issues, then switch #2 will not learn about the change and will continue to respond to host B's ARP Requests for host A's old IP address with stale information. This lasts until the ARP entry for A times out at Switch #2.

4.0 Conclusion

Based on the procedures described in this document, it is possible for ToR switches in the data center to contain ARP broadcasts significantly. The solution is based on well known, non-intrusive procedures and strives to curtail broadcasts that are increasingly becoming a cause for concern in the data centers. In essence, ToR switches facilitate the offloading of the extended ARP table management from the IP hosts to itself. The ARP table timeout can be tuned higher by the operator based on the available switch resources and network traffic behavior. The larger capacity of the ARP table directly translates to more effective subduing of the ARP broadcasts.

5.0 Security Considerations

The details of the security aspects will be addressed in future revision.

6.0 Acknowledgments

This document resulted from discussions with Linda Durbar (Huawei), Sue Hares (Huawei), and T Sridhar (VMware). We would like to acknowledge their contribution to this work.

7.0 References

7.1 Normative References

[ARP] D. Plummer, "An Ethernet Address Resolution Protocol: Or Converting Network Protocol Addresses to 48.bit Ethernet Addresses for Transmission on Ethernet Hardware," RFC 826, STD 37.

[ARP-Problem] T. Narten, "Problem Statement for ARMD," work in progress, <draft-ietf-armd-problem-statement>.

7.2 Informative References

[ARP-Mediation] H. Shah et al., "ARP Mediation for IP interworking in Layer 2 VPN," work in progress, <draft-ietf-l2vpn-arp-mediation>.

[IPLS] H.Shah et al., "IP-only LAN service," work in progress, <draft-ietf-l2vpn-ipls>.

[PROXY-ARP] J. Postel, "Multi-LAN Address Resolution," RFC 925.

[RFC1027] Smoot et al., "Using ARP to Implement Transparent Subnet Gateways".

[VXLAN] M. Mahalingam et al., "VXLAN: A Framework for Overlaying Virtualized Layer 2 Networks over Layer 3 Networks"," work in progress, <draft-mahalingam-dutt-dcops-vxlan>.

[NVGRE] M. Sridharan et al., " NVGRE: Network Virtualization using Generic Routing Encapsulation", work in progress, <draft-sridharan-virtualization-nvgre>.

8.0 Author's Address

Himanshu Shah
Ciena Corp
Email: hshah@ciena.com

Anoop Ghanwani
Brocade

Email: anoop@alumni.duke.edu

Nabil Bitar
Verizon
Email: nabil.n.bitar@verizon.com