ARMD BOF
Internet Draft
Intended status: Informational Track
Expires: January 2012

M. Karir
J. Rees
Merit Network Inc.

October 20, 2011

Address Resolution Statistics

draft-karir-armd-statistics-01.txt

Status of this Memo

This Internet-Draft is submitted to IETF 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/ietf/1id-abstracts.txt

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

This Internet-Draft will expire on January 20, 2012.

Copyright Notice

Copyright (c) 2011 IETF Trust and the persons identified as the document authors. All rights reserved.

This document is subject to BCP 78 and the IETF Trust's Legal Provisions Relating to IETF Documents (http://trustee.ietf.org/license-info) in effect on the date of publication of this document. Please review these documents carefully, as they describe your rights and restrictions with respect to this document. Code Components extracted from this document must include Simplified BSD License text as described in Section 4.e of the Trust Legal Provisions and are provided without warranty as described in the BSD License.

Abstract

As large scale data centers continue to grow with an ever-increasing number of virtual and physical servers there is a need to re-evaluate performance at the network edge. Performance is often critical for large scale data center scale applications and it is important to minimize any unnecessary latency or load in order to streamline the operation of services at such large scales. To extract maximum performance from these applications it is important to optimize and tune all the layers in the data center stack. One critical area that requires particular attention is the link-layer address resolution protocol that maps an IP address with the specific hardware address at the edge of the network.

The goal of this document is to characterize this problem space in detail in order to better understand the scale of the problem as well as to identify particular scenarios where address resolution might have greater adverse impact on performance.

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

Data centers are a key part of delivering Internet scale applications. Performance at such large scales is critical as even a few milliseconds or microseconds of additional latency can result in loss of customer traffic. Data center design and network architecture is a key part of the overall service delivery plan. This includes not only determining the scale of physical and virtual servers but also optimizations to the entire data center stack including in particular the layer 3 and layer 2 architectures.

One aspect of data center design that has received some close attention is link-layer address resolution protocols such as Address Resolution Protocol (ARP - IPv4) and Neighbor Discovery (ND - IPv6). The goal of these protocols is to map an IP address of a destination node with the hardware address of the network interface for that node. This address resolution occurs at the edge of the network. In general, both ARP and ND are query/response protocols.

In order to maximize performance it is important to understand the behavior of these protocols at large scales. In particular, we need to understand what the performance implications of these protocols might be in terms of the number of additional messages that they generate as well the resulting load on devices on the network that must then process these messages.

2. Terminology

ARP: Address Resolution Protocol

ND: Neighbor Discovery

ToR: Top of Rack Switch

VM: Virtual Machines

3. Factors That Might Impact ARP/ND Performance

3.1. Number of Hosts

Every host on the network that attempts to send/receive traffic will produce some base level of ARP/ND traffic. The overall amount of ARP/ND traffic on the network will vary with the number of hosts. In the case of ARP, all address resolution request messages are broadcast and these will be received and processed by all nodes on the network. In the case of ND, address resolution messages are sent via multicast and therefore may have a lower overall impact on the network even though the number of messages exchanged is the same.

3.2. Traffic Patterns

The traffic pattern can have a significant impact on the level of ARP/ND traffic in the network. Therefore we would expect ARP/ND traffic pattern to vary significantly based on the data center design as well as the application mix. The traffic mix determines how many other nodes a given node needs to communicate with and how frequently. Both of these directly influence address discovery traffic on the network.

3.3. Network Events

Several specific network events can have a significant impact on ARP/ND traffic. One example of such an event is machine failure. If a host that is frequently accessed fails, it could result in much higher ARP/ND traffic as other hosts in the network continue to try to reach it by repeatedly sending out additional address resolution messages. Another example is Virtual Machine migration. If a VM is migrated to a system on a different switch, VLAN, or even geographically different data center, it can cause a significant shift in overall traffic patterns as well as ARP/ND traffic. Another particularly well-known network event that causes address resolution traffic spikes is a network scan. In a network scan, one or more hosts internal or external to the edge network attempt to connect to a large number of internal hosts in a very short period of time. This results in a sudden increase in the amount of address resolution traffic in the network.

3.4. Address Resolution Implementations

As with any other protocol, the activity of address resolution protocols such as ARP/ND can vary significantly with specific implementations as well as the default settings for various protocol parameters. ARP cache timeout is a common parameter that has a

direct impact on the amount of address resolution traffic. Older versions of Microsoft Windows would use a default value of 2 minutes for this parameter, however Windows Vista and Windows 2008 implementations changed this to be a random value between 15 seconds and 45 seconds. This parameter defaults to 60 seconds for Linux and 20 minutes for FreeBSD. The default value for Cisco routers and switches is 4 hours. For ND, one relevant parameter is the prefix stale time, which determines when old entries can be aged out. This value is 30 days for Cisco, and 60 seconds for Linux. The overall address resolution traffic in a data center will vary based on the mix of various ARP implementations that are present.

3.5. Layer 2 Network Topology

The layer 2 network topology within a data center can also influence the impact of various address resolution protocols. While ARP traffic is broadcast and must be processed by all nodes within that broadcast domain, a well designed layer 2 topology can limit the size of the broadcast domain and the amount of address resolution traffic. ND traffic on the other hand is multicast and might potentially increase the load on the directly connected layer 2 switch if the traffic pattern spans across broadcast domains.

4. Experiments and Measurements

4.1. Experiment Architecture

In an attempt to quantify address resolution issues in a data center environment we have run experiments in our own data center, which is used for production services. We were able to leverage unused capacity for our experiments. The data center topology is fairly simple. There are a pair of redundant access switches which pass traffic to and from the data center. These switches connect to the top of the rack switches which in turn connect to blade switches in our Dell blade chassis. The entire hardware platform is managed via VMware's vCloud Director. In total we have access to 8 blades of resources on a single chassis, which is roughly 3TB of disk, 200GB of RAM and 100GHz of CPU. The network available to us is a /22 network block of IPv4 space and a /64 of IPv6 address space in a flat topology.

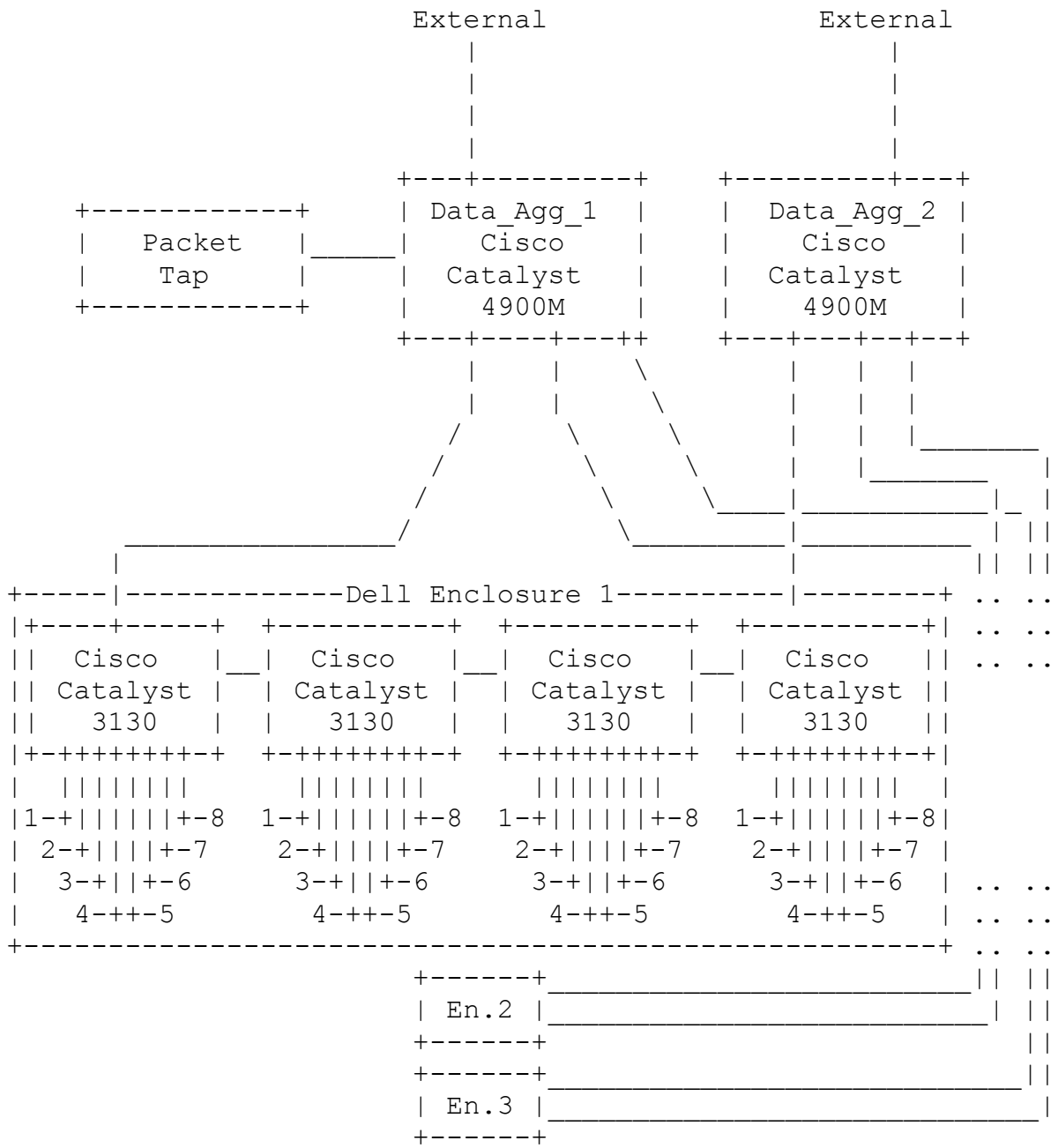
Using this resource pool we create a 500-node testbed based on Centos 5.5. We use custom command and control software that allows us to control these nodes for our experiments. This allows us to issue commands to all nodes to start/stop services and traffic generation scripts. We also use a custom traffic generator agent in

order to generate both internal and external traffic via wget commands to various hosts.

The command and control software uses UDP broadcast messages for communication so that no additional address resolution messages are generated that might affect our measurements. Each of the 500 nodes is given a list of other nodes that it must contact at the beginning of an experiment. This is used to affect the traffic patterns for a given experiment. In addition each experiment determines traffic rate by specifying the inter-communication delay between attempts to contact other nodes. The shorter the duration the more the traffic that will be generated. The nodes all run dual IPv4/IPv6 stacks.

A packet tap attached to a monitor port on the access switch allows us to monitor the arrival rate of ARP and ND requests and replies. We also monitor the CPU load on the access switch at two-second intervals via SNMP queries [STUDY].

Figure 1. shows our experimental setup.


External
External
Packet Tap
Data_Agg_1 Cisco Catalyst 4900M
Data_Agg_2 Cisco Catalyst 4900M
Dell Enclosure 1
Cisco Catalyst 3130
Cisco Catalyst 3130
Cisco Catalyst 3130
Cisco Catalyst 3130
En.2
En.3

4.2. Impact of Number of Hosts

One of the most simple experiments is to determine the overall baseline load that is generated on a given network segment when a varying number of hosts are active. While the absolute numbers might vary on a large number of factors, what we are interested in here is how the traffic scales as different numbers of hosts are brought online given all other factors being held constant. Our experiment therefore simply changes the number of active hosts in our experiment setup from one run to the next and we measure address resolution traffic on the network. The number of hosts is increased from 100 to 500 in steps of 100. The results indicate that address resolution traffic scales in a linear fashion with the number of hosts in the network. This linear scaling applies both to ARP as well as ND traffic though raw ARP traffic rate was considerably higher than ND traffic rate. For our parameters the rate varied from 100 to 250pps of ARP traffic and from 25pps to 200pps for ND traffic. There is a clear spike in CPU load on the access switch in the beginning of each experiment, which can reach almost 40 percent. We were not able to discern any increase in this spike across experiments.

4.3. Impact of Traffic Patterns

Traffic patterns can have a significant impact on the amount of address resolution traffic in the network. In order to study this in detail we constructed two distinct experiments, the first of which simply increased the rate at which nodes were attempting to communicate with each other, while the second experiment controlled the number of active versus inactive nodes in the traffic exchange matrix.

The first experiment uses all 500 nodes in our experiment and increases the traffic load for each run by reducing the wait time between communication events. The wait time is reduced from 50 seconds to 1 second over a series of 6 runs by roughly halving the duration for each run. All other parameters remain the same across experiment runs. Therefore the only factor we are varying is the total number of nodes a single node will attempt to communicate within a given interval of time. Once again we observe a linear scaling in ARP traffic volumes ranging from 200pps for the slowest experiment to almost 1800pps for the most aggressive experiment. The linear trend also holds for ND traffic, which increases from 50pps to 1400pps across different runs.

The goal of the second experiment is to determine the impact of active versus inactive hosts in the network. An inactive host in this context means one for which an IP address has been assigned, but there is nothing at that address so that ARP requests and all other packets are ignored. All 500 hosts are involved in traffic initiation. The pool of targets for this traffic starts out being the same 500 hosts that are initiating. In subsequent runs we vary the ratio of active to inactive target hosts, from 500/0 to 400/100 in steps of 100. This experiment showed roughly a 60% increase (220-360 pps) in traffic for the IPv4 (ARP) case and about an 80% increase (160-290 pps) for the IPv6 case.

In a slight variation on the second experiment all 500 nodes attempt to contact all other hosts plus an additional varying number of inactive hosts in steps of 100 up to a maximum of 400. In this experiment we see a slight linear increase as the total number of nodes in the traffic matrix increases for both ARP and ND.

We ran these experiments for IPv4 only, IPv6 only, and simultaneous IPv4 and IPv6. ARP and ND traffic seemed to be independent of each other. That is, the ARP and ND traffic rates and switch CPU load depend on the presented traffic load, not on the presence of other traffic on the network.

One final experiment attempted to determine what the maximum additional load of ARP/ND traffic might be in our setup. For this purpose we configured our experiment to use all 500 nodes to communicate with all 500 other nodes one at a time as fast as possible. We were able to observe ARP traffic peak of up to 4000pps and a maximum CPU load of 65% on the access switch.

4.4. Impact of Network Events

Network scanning is commonly understood to cause significant address resolution activity on the edge of the network. Using our experimental setup we attempted to repeatedly scan our network both from the outside as well as within. In each case we were able to generate ARP traffic spikes of up to 1400pps and ND traffic spikes of 1000pps. These are also accompanied by a corresponding spike in CPU load at the access switch.

Node failures in a network also have the ability to significantly impact address resolution traffic. This effect depends on the particular traffic patter and the number of other hosts that are attempting to communicate with the failed node. All nodes will repeatedly attempt to perform address resolution for the failed node and this can lead to significant increase in ARP/ND traffic. We are

able to show this via a simple experiment that creates 400 active nodes which all attempt to communicate with nodes in a separate group of 80 nodes. For each experiment run we then shutdown hosts in the target group of 80 nodes in batches of 10 each. We are able to demonstrate that ARP traffic actually increases in this scenario from an overall rate of 200pps to 300pps.

Another network event that might result in significant changes in address resolution traffic is the migration of VMs in a data center. We attempted to replicate this scenario in our somewhat limited environment by placing one of our 8 blades in maintenance mode, which forced all 36 VMs on that blade to migrate to other blades. However, as our entire experimental infrastructure is located within a single rack we do not notice any changes in ARP traffic during this event.

Many hypervisors remove the problem of virtual machine migration by assigning a MAC address to a VM, and then a kernel switching module handles all address resolution, accepting and sending packets for all the MAC addresses of its virtual machines through a determined host interface. In other words, the hypervisor responds to the appropriate traffic for the VMs it contains. It behaves as a router for the Layer 2 traffic it is exposed to.

4.5. Implementation Issues

Protocol implementations and default parameter values can also have a significant impact on the behavior of address resolution traffic in the network. Parameters such as cache timeout values in particular determine when cached entries are removed or need to be accessed to ensure they are not stale. Though these parameters are unlikely to be modified the variation in these for different systems can impact ARP/ND traffic when different systems are present on a given network in varying numbers. Our experimental setup did not explore this issue of mixed environments or sensitivity of ARP/ND traffic to the various protocols parameters.

4.6. Experiment Limitations

Our experimental environment though fairly typical in the hardware and software aspects probably only represents a very limited small data center configuration. It is difficult to thoroughly instrument very large environments and even smaller experimental environments in a lab might not be very representative. We believe our architecture is fairly representative and provides us with useful insights regarding the scale and trends of address resolution traffic in a data center.

One very significant limitation that we came across in our experiments was the problems of using all 500 nodes in a high load scenario. When all 500 nodes were active simultaneously our architecture would run into a bottleneck while accessing disk storage. This limitation also prevents us from attempting to scale our experiments for more than 500 nodes. This also limited us in what experiments we could run at the maximum possible load.

Our experimental testbed shared infrastructure, including network access switches, with production equipment. This limited our ability to stress the network to failure, and our ability to try changes in switch configuration.

5. Scaling Up: Emulating Address Resolution Behavior on Larger Scales

Based on the data collected from our experiments we have built an ARP/ND traffic emulator that has the ability to generate varying amounts of address resolution traffic on a network with varying address ranges. This gives us the ability to scale beyond 500 VM nodes in our experiments. Our software emulator can be used to directly test the impact of such traffic on nodes and switches in the network at much larger scales.

Preliminary results show a good match between the testbed and the emulator for both traffic rates and switch load over a wide range of presented traffic load. We have calibrated the emulator from the testbed data and will use the emulator to run experiments at scales that would otherwise be impractical in the real network available to us.

6. Conclusion and Recommendation

In this document we have described some of our experiments in determining the actual amount of address resolution traffic on the network under a variety of conditions for a simple small data center topology. We are able to show that ARP/ND traffic scales linearly with the number of hosts in the network as well as the traffic interconnection matrix. In addition we also study the impact of network events such as scanning, machine failure and VM migrations on address resolution traffic. We were able to show that even in a small data center with only 8 blades and 500 virtual hosts, ARP/ND traffic can reach rates of thousands of packets per second, and switch CPU loads can reach 65% or more.

We are able to utilize the data from our experiments to build a software based ARP/ND traffic emulation engine that has the ability to generate address resolution traffic at even larger scales. The

goal of this emulation engine is to allow us to study the impact of this traffic on the network for large data centers.

7. Manageability Considerations

This document does not add additional manageability considerations.

8. Security Considerations

This document has no additional requirement for security.

9. IANA Considerations

None.

10. Acknowledgments

We want to acknowledge the following people for their valuable discussions related to this draft: Igor Gashinsky, Kyle Creyts, Warren Kumari.

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

11. References

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

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

[STUDY] Rees, J., Karir, M., "ARP Traffic Study." MANOG52, June 2011. URL http://www.nanog.org/meetings/nanog52/presentations/Tuesday/Karir-4-ARP-Study-Merit Network.pdf

Authors' Addresses

Manish Karir
Merit Network Inc.
1000 Oakbrook Dr, Suite 200
Ann Arbor, MI 48104, USA
Phone: 734-527-5750
Email: mkarir@merit.edu

Jim Rees
Merit Network Inc.
100 Oakbrook Dr, Suite 200
Ann Arbor, MI 48104, USA
Phone: 734-527-5751
Email: rees@merit.edu

Intellectual Property Statement

The IETF Trust takes no position regarding the validity or scope of any Intellectual Property Rights or other rights that might be claimed to pertain to the implementation or use of the technology described in any IETF Document or the extent to which any license under such rights might or might not be available; nor does it represent that it has made any independent effort to identify any such rights.

Copies of Intellectual Property disclosures made to the IETF Secretariat and any assurances of licenses to be made available, or the result of an attempt made to obtain a general license or permission for the use of such proprietary rights by implementers or 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

All IETF Documents and the information contained therein are provided on an "AS IS" basis and THE CONTRIBUTOR, THE ORGANIZATION HE/SHE REPRESENTS OR IS SPONSORED BY (IF ANY), THE INTERNET SOCIETY, THE IETF TRUST AND THE INTERNET ENGINEERING TASK FORCE DISCLAIM ALL WARRANTIES, EXPRESS OR IMPLIED, INCLUDING BUT NOT LIMITED TO ANY WARRANTY THAT THE USE OF THE INFORMATION THEREIN WILL NOT INFRINGE ANY RIGHTS OR ANY IMPLIED WARRANTIES OF MERCHANTABILITY OR FITNESS FOR A PARTICULAR PURPOSE.

Acknowledgment

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