ARMD Working Group
Internet Draft
Intended status: Informational
Expires: January 2012

Susan Hares
Huawei
July 1, 2011

NANOG 52 Operators Perspective
draft-hares-armd-nanog52-00.txt

Abstract

Data Centers are growing in number of physical and virtual machines. The scaling of broadcast domains impacts the scale of basic Address resolution protocols (ARP and ND).

The ARMD working (http://tools.ietf.org/wg/armd/charters) has been charter to examine the details of this problem. Part of the examination was to ask operators of data centers and researchers to provide details on the scope of the problem. The ARMD chairs (Benson Schliesser (Cisco) and Linda Dunbar (Huawei) held a panel session at NANOG 52 to report initial findings.

The researchers on the panel were: Manish Karir (Merit), and K.K. Ramakrishnan (AT&T Research). The Operators on this panel were from Google (Scott Whyte), Yahoo (Igor Gashinsky), and Adhost (Michael K. Smith).

This memo brings into IETF format notes taken at the panel session. Any errors in the summary are the author's. The presentations for the session are listed at the ARMD track at:

http://www.nanog.org/meetings/nanog52/agenda.php

However, an audio recording was not made. This document is an informational RFC whose intent is to record a moment in time.

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 3, 2009.

Copyright Notice

Copyright (c) 2010 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 Simplified BSD License

This document may contain material from IETF Documents or IET Contributions published or made publicly available before November 10, 2008. The person(s) controlling the copyright in some of this material may not have granted the IETF Trust the right to allow
modifications of such material outside the IETF Standards Process.

Without obtaining an adequate license from the person(s) controlling the copyright in such materials, this document may not be modified outside the IETF Standards Process, and derivative works of it may not be created outside the IETF Standards Process, except to format it for publication as an RFC or to translate it into languages other than English.

Table of Contents

1. Introduction

Volunteering gets you into interesting places in the IETF and NANOG. I volunteered to take notes at NANOG 52's ARMD session. I believed that NANOG 52 was recording the audio recording of the talks, and my notes would simply help the ARMD panel chairs. However, the audio recording is not up on the NANOG 52 web site. The chairs have asked me to make my notes available to the wider IETF community who could not attend.

The NANOG session had the following agenda:

* Overview (Benson Schliesser and Linda Dunbar, ARMD co-chairs),
* Michael K. Smith (Adhost),
* Scott Whyte (Google),
* Igor Gashinsky (Yahoo!),
* Manish Karir (Merit), and
* K.K. Ramakrishnan (AT&T Research).

The notes follow this agenda, but to prepare the reader we will introduce the speakers ahead of time. Benson Schliesser is the co-chair of ARMD. In the past, Benson worked at a service provider who had large Data Center deployments. Linda Dunbar is the second co-chair of ARMD. Linda's background comes from teams working on developing next-generation Data centers within the Corporate or Enterprise space.

Michael K. Smith comes from Adhost Internet, LLC which is a Web hosting company based in Seattle. Scott Whyte is a "network engineer" at Google Google. He presented on the characteristics of the Data Center. Igor is the principle architect at Yahoo.

Manish Karir is Director of Research and Development at Merit network. His past research interests include DARPA funded control plane research, Homeland Security funded PREDICT project on botnets, and BGP [MK-BIO].

K.K. Ramakrishnan is a AT&T Research investigating making cloud storage and computing resources available in transparent and seamless fashion. He is also examining "large scale XML-based information dissemination" [KK-Bio].

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 [RFC2119].

2. Overview(Benson Schliesser and Linda Dunbar)

Presentation([ARMD-PANEL-NANOG52])

ARMD is a working group examining “Address Resolution for the Massive numbers of hosts in the Data Center” [ARMD-Charter]. Address resolution includes IPv4 ARP [RFC826] and IPv6 Neighbor Discovery [RFC2461]. The focus of the working group is to determine the impact of ARP and ND in the real network.

[Editor’s note] The working group is considering the body of work included in the ARP and ND protocols. For ARP this includes the IPv4 Address Conflict Resolution [RFC5227].

The traditional picture of ARP or ND in a switching environment is a few hosts attached to a switch. The modern datacenters are buildings the size of 2 square city blocks with rows and rows of equipment. Many data centers host multiple tenants physically and virtual. The dynamic network environment includes Virtual Machine (VM) mobility and the ability to provide backup (1-1 or n-machines to 1).

The modern data center resembles a highly elastic weather balloon. The data center size allows massive number of hosts and large numbers of subnets. This scale inflates the weather balloon’s reach and the number of address resolutions needed. Server virtualizations have made it easier to build highly dense Virtual machine clusters in the data center, and then move them around flexibility. Igor commented that the algorithms used by the server and virtual machine people helped enable this growth.

The goal of ARMD is to identify how the scaling of address resolution between the network (L3) and the Link (L2) layers of modern datacenter networks. The “identification” includes how the growth of the number of hosts impacts hosts, servers, routers, switches, and link by the transmission or processing of Address Resolution Messages (ARP or ND).

The working is handling a “call for investigation” described in [ARMD-Investigate]. The key questions are:

* What are the scaling characteristics of Address Resolution and what operational problems does this impact?

* What are the alternative solutions to address these issues?

* Are there gaps?

The investigation is looking at ARP, ND, and the combination of ARP/ND in dual stacks.

The NANOG session is to let data center operators describe the environment address resolution exists in and any issues with the ARP traffic being broadcast (or multicast) and the multicast ND traffic.

This session also looks to researchers to examine the theoretical maximum, minimums, and norms for a variety of situations found in the data center. These situations include a cluster of virtual hosts, 2+ clusters of hosts connected by a switch, real hosts connected by switches, and other scenarios. One question the theoretical discussions might ask is "why does Layer 2 still exist in the data center" or "Why does layer 3 still exist in the data center?"

Another part of the general question is the sizing for data center. What are the size ranges and traffic load ranges for different data centers? How important is Host placement and movement? When and how does the Address Resolution need to occur, and what is gratuitous resolution.

3. Michael K. Smith, AdHost

Presentation ([Smith-ARMD-NANOG52])

ADhost provides co-location, hosting, and cloud servers. We work at medium size due to the demands of our customer base. We support a combination of layer 2 and layer 3 due to their demands.

Let's take the example of 5 racks at Layer 2. Two of the racks are in one site, and another site. The customer's application requires the connection at layer 2. We enable the customer's applications to run easily in our datacenter.

Questions for Michael Smith:

1. Why do you not use layer 3? [author (?)]

Answer: Our customers have an application that requires running at layer 2. We

2. Do they want to see a virtual network or can they use a virtual layer 3 network? [Author (?)]

Answer: Business reasons cause the customer to want to run their layer 2 application native.

3. Does L2VPN help you provide this support? (Ron Bonica)

Answer: I still need to carry the ARP or ND information across the Layer 2 VPN.

4. Why not go to Layer 3? [Ron Boncia]

Answer: Layer 3 is more expensive, and does not fit the customer needs.

Igor comment: All traffic needs to be both Layer 2 and Layer 3. It is when you get to the L2/L3 translation that it becomes problem?

5. Why do you not give a direct connection? [author (?)]

The cost of the fiber network is a problem.

4. Scott Whyte – "Data Centers: Inside the cloud"

Presentation:[2-ARMD-Whyte]

Data centers have the following different roles: hosting, managed services, campus data center, and large data center. At the campus level of data centers there is no homogeneity [i.e., heterogeneous]. At large data centers such as Google, there can be a very homogeneous deployment of equipment. At Google, the Data Center is the OS for the application.

The workloads on data centers can be virtualized machines, centralized applications, distributed application or the "big compute" process. These workloads balance the "timesharing" of the workload versus the effort involved in parallelizing the workload.

For virtualized machines, we examine if the workload is tough at 50, 500, 5000 or larger. The centralized application creates an image of centralized hardware within the data center. In the

distributed application, the process abstracts away the hardware, software, and OS into once virtual application. The "big compute" is an interesting application we continue to study. We are looking into whether the parallelization double or triple the information passed, and impacts network control protocols such as ARP, ND, and others.

The unique characteristics of the data center workloads are varying tolerances for latency, bandwidth needs, storage needs, and the compute resources. Processing workloads may be able to deal with "oversubscription", varying availability of resources, shedding low for power requirements, and auto deployment to various servers.

The large-scale data centers must focus on being efficient and effective in power/cooling, workload placement, and resource management. Protocol improvements or upgrades can help efficiency and effectiveness.

Questions:

1. Are these characteristic of workload for inter-data center or intra-data center? [Lucy Yong]

[Scott] We are discuss the intra-data center case.

2. Are these characteristics how you quantify the scale?

[Scott] This is a characteristic that is specific data centers we have examined.

Benson's comment: This is one type of questions we are trying to investigate. What types of dimensions need to be focused on to scale the Data Center? We are trying to get specifics for a specific type of data center.

5. Igor Gashinsky – "Datacenter Scalability Panel"

Presentation: [Gahinsky-3-Y-Datacenter-scalability]

Scott Whyte did a nice job of describing the general issues.

Today warehouse data centers are being built that can accommodate over 120,000 physical servers. Each server packs a lot of processing cores with 24 cores. With a decent virtualization processing, this allows 20 Virtual Machines

(VMs) per server. This means with a 120,000 machines in a data center, that's 2.4 million VMs. And that's only today.

The future data center has 10Gig Ethernet to the Server. DAS (directly-attached storage) left the [data center] building a long time ago. Network-attached storage is on its way out, and cloud storage is the new "in." This means that every server will contain both a storage device and a compute node.

To get the best utilization of all those resources, we (Yahoo) need to be able to place a VM anywhere, any time. The VM must be able to be migrated where every need it. To accomplish this we need a "flat" network with a very low oversubscription ration. Our target oversubscription ration is 2:1.

This means our network needs to be a flat layer 2 network to support IP/VM mobility. The rack switches need to be 40 ports of 10Gig Ethernet and 200Gig throughput with 10/40/100G uplinks. The core switches need to have 300+ 40/100G ports. The control plane scalability needs ot hold (and move) 2.4 M VMs. This means a movement of 2.4 million MAC address, 2.4 million IPv4 address, and 4.8 million (2.4 *2) IPv6 addresses.

So, What's the problem? We need core switches with 300+ 40/100G ports. The movement of the MAC address (2.4 million), the IPv4 addresses (2.4 million), and 4.8 IPv6 address is not doable using current techniques.

What about Segmentation of the Network? The largest VM domain that we can scale now is 10,000 (10K) servers. The 10K server times 20 Virtual Machines (VMs) per box means we have domains of 200,000 VMs. This still does not help.

We are looking for a better way. So what are our options?

Option 1: Overlay a logical network on top of a physical network. This shifts the control plane scalability into the server/vSwitch.

Option 2: Find a lighter way to scale the current network. The means better learning mechanisms for addresses and IP addresses; and better CAM scalability.

There has been a lot of research into "programmable data centers" such as monsoon, Seattle, VL2, Moose, and openflow. However, no single of these "programmable data centers"

addresses all the issues. Some of these want to change host stacks. Others want to change everything in the Internet.

What is a possible solution? Perhaps we could "program" the data center without modifying the host stack and addressing. In large-scale deployments companies have very extensive Inventory Management systems, and they already know: a) the location of every server, b) the switch and port every server is plugged into, and c) the IP and MAC addresses of every server. Why is the network bothering to learn it every X seconds, instead of having the inventory management systems simply program this.

This solution solves the network discovery scalability issues.

Discussion:

1. What about mobility? [Dave Meyers]

Suppose there is a VM and a VM server. If an automated system kicks off an automated move, it updates the data plane servers.

2. What about the network behind the distributed server? (author (?))

[Igor G.] The distributed servers get thousands of queries from servers and stay in sync. The network vendors cannot get two line cards to stay in sync.

[Dave Meyers] The distributed systems vendors solved this problem, and it is being pulled into networking gear.

[Igor G.] It is now getting pulled into networking gear so it is 3 years before it will be available as a commercial project. It is not the distributed system vendors or the network vendors fault. Both attempted solutions and the distributed systems got it first. It is just that the networking vendors must now upgrade to the solution.

6. Jim Rees and Manish Karir (Merit Network Inc.) – "ARP Traffic Study"

Presentation:[MK-ARMD-NANOG52]

Manish presented the traffic study which attempted to understand ARP behavior under various conditions. The

methodology looks to combine observing ARP behavior in data centers with simulated environments, and emulators. Since data center environments vary, the emulator will be able to mimic a variety of environments.

Merit's study plans the following steps: a) observe the ARP behavior in medium size data center deployments, b) recreate the same ARP behavior in simulated environments, c) build a model of ARP/ND based on experiments and collect data from model, and d) build scalable ARP/ND emulator for large scale experiments which can mimic various environments, e) evaluate operations of software and protocols, f) propose solutions (if possible).

Manish has written up the study in full in [Karir-ARMD]. This document will only provide the Question/Answer period discussion.

Questions:

1. (Igor) Where are doing the ARP generation? Is this all on one server or across a switch?

[Manish] It is only on one server, and does not cross the switch. We tried to limit the restricts that ARP would face with switches.

[Igor] Your experiment doesn't test the switch traffic, but just the data center devices.

[Manish] This is correct. Should it test cross switch traffic?

[Igor] If you created two subnets, it would test the switch where there are problems. The ARP may be massively optimized.

[Editor: Igor's actual words were "may have the hell optimized out of it", but for our cross-cultural English speakers I have provided a more generic translation.]

7. K.K. Ramakrishnan (AT&T Labs Research)

Presentation:[KK-Ramakrishnan-NANOG52]

K.K. presented research on the CloudNet which is an Enterprise Ready Virtual Private clouds. This research work is joint work with Timothy Woods, Jacobus Van der merwe, and Prashant Shenoy.

K.K. Ramakrishnan and colleagues are examining how to make computing and storage resource location transparent for enterprises and general computing.

This transparency looks to provide secure and flexible migration for the for the application while minimizing the performance impact. This would allow quick recovery during disaster where computing must be quickly transfer to a remote location that does not fate-share with the original data center. An example of a disaster is a flood or a tornado affecting a data center.

K.K.’s work defines private virtual clouds (VPC) as a secure collection of server, storage, network resources spanning one or more cloud data centers. This secure collection is “seamlessly” connected to one or more enterprise sites via VPNs. These VPNs can be L2 or L3 MPLS based VPNs.
The benefit of the VPC for each enterprise customer is isolation of network and compute resources per application, and the simplification of deployment. The VPCs benefit service providers by providing control over resource reservation, and simplifying management of multiple data centers.

One example of the VPC is AT&T Cloud Net which has a cloud manager that talks to the network manager handling the VPNs (L2 MPLS, L3 MPLS or others). The Cloud manager manages VPN assignments, and allocates computation and storage resources. The network manager reserves VPN resources, and creates and/or configures VPN endpoints.

The CloudNet Cloud manager works with an IRSCP entity. The IRSCP entity acts as a route server. The IRSCP send to the network manager new route-targets for L2/L3 MPLS VPN connections. The Cloud Manager also dynamically configures logical CE routers on the customer side with VLAN and L2/L3 MPLS configurations. The IRSCP rewrites Route-targets to create the VPN membership.

Storage migration is done via: a) asynchronous couple of disk storage to remote site initially, and b) synchronous copy of incremental updates during subsequent live memory migration. The live memory migration needs to balance multiple requirements of the total time for migration, the pause time (quiescent time for final migration), and the amount of data transferred (bandwidth).

Ramakrishnan's full slide set is available at [KK-Ramakrishnan-NANOG52]. His algorithm work includes: a) algorithms to optimize migration time, pause time, and network bandwidth, and b) CloudNets use in disaster scenarios.

Questions: None

8. Final Questions

1. What is the output of ARMD WG? [Igor G.]

Benson: It is the description of the problem, and the potential solutions.

2. Is it a general or a specific design that you are trying to capture.

[Ron Bonica, AD OPS] The purpose is to discuss what is not scaling, and what are potential alternatives for ARP or ND.

9. Security Considerations

This draft has no security considerations.

This draft only provides notes for the NANOG 52 ARMD session. It is not intended for deployment in any network or virtual process (organic or silicon) for long periods of time, but should only engender thinking. Of course, thinking can be the challenge to any security issue.

10. IANA Considerations

This document requires no IANA considerations.

11. References

11.1. Normative References

[RFC826] Plummer, D.C.,"An Ethernet Address Resolution Protocol", RFC 826, November 1982.

[RFC2461] Narten, T., Nordmark, E., Simpson, W, "Neighbor Discovery for IP Version 6 (IPv6), December 1998.

[RFC4098] Cheshire, S. "IP Address conflict Detection", RFC5227, July 2008.

[RFC2119] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, March 1997.

11.2 [Informative References]

[ARMD-charter] ARMD-WG, "Address Resolution for Massive Numbers of Hosts in the Data Center (ARMD)", online: http://tools.ietf.org/wg/armd/charters [accessed: 7/1/2011].

[ARMD-Investigate] Schilesser, B. & Dunbar, L. "ARMD Call for Investigation", http://www.ietf.org/id/draft-ietf-armd-call-for-investigation-00.txt

[ARND-PANEL-NANOG52] Schliesser, B. & Dunbar, L. "ARMD Panel at NANOG 52", online: http://www.nanog.org/meetings/nanog52/abstracts.php?pt=MTgwNiZuYW5vZzUy&nm=nanog52 [accessed: 7/1/2011].

[Gahinsky-3-Y-Datacenter-scalability] Gahinsky, I. "Data Center Scalability Panel", online: http://www.nanog.org/meetings/nanog52/presentations/Tuesday/Gashinsky-3-Y-Datacenter-scalability.pdf [accessed: 7/11/2011].

[KK-bio] "K.K. Ramkrishnan's Home Page", online: http://www2.research.att.com/~kkrama/ [accessed: 7/1/2011].

[KK-Ramakrishnan-NANOG52] Ramkrishnan, K.K. (2011). "CloudNet: Enterprise Ready Virtual Private Clouds", online: http://www.nanog.org/meetings/nanog52/presentations/Tuesday/Ramakrishnan-5-KK%20-att.pdf

[MK-Bio] "Manish Karir Biography" as referenced in "Merit: Not Just Your Internet Service Provider - RADB and Merit." Online: [http://www.merit.edu/events/mmc/abstracts.php?mamdate=2011&sp=Karir&printvs=1] [accessed:7/1/2011].

[MK-ARMD] Karir, M, and Reese, J. "Address Resolution Statistics" [unpublished, publishing pending at: http://www.ietf.org/drafts/draft-karir-armd-statistics-00.txt, [early copy received on 7/1/2011].

[MK-ARMD-NANOG52] Karir, M, and Reese, J. "ARP Traffic Study", NANOG52, ARMD panel, online: http://www.nanog.org/meetings/nanog52/presentations/Tuesday/Karir-4-ARP-Study-Merit%20Network.pdf,[accessed: 7/1/2011].

[Smith-ARMD-NANOG52] Smith, M.K. "Adhost Internet, LLC.)", [online: http://www.nanog.org/meetings/nanog52/presentations/Tuesday/Smith-1-Drawing-%20Adhost.pdf [accessed: 7/1/2011]

[Whyte-ARMD-NANOG52] Whyte, S. "Data Centers", online: http://www.nanog.org/meetings/nanog52/presentations/Tuesday/2-ARMD-Whyte.pdf [accessed: 7/1/2011].

Author's Addresses

Susan Hares
Huawei Technologies (USA)
2330 Central Expressway
Santa Clara, CA 95050
Phone: +408-330-4581
Cell: +1-734-604-0332
Email shares@huawei.com