

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-Draft BGP Convergence Methodology July 2011
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/lid-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) 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 Simplified BSD License.

This document may contain material from IETF Documents or IETF 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 3
2. Introduction (Benson Schliesser and Linda Dunbar) [ARMD-PANEL-NANOG52] 4

Internet-Draft	BGP Convergence Methodology	July 2011
3. Michael K. Smith, AdHost [Smith-ARMD-NANOG52]		5
4. Scott Whyte - "Data Centers: Inside the cloud" [2-ARMD-Whyte]		6
5. Igor Gashinsky - "Datacenter Scalability Panel" [Gahinsky-3-Y-Datacenter-scalability]		7
6. Jim Rees and Manish Karir (Merit Network Inc.) - "ARP Traffic Study" [MK-ARMD-NANOG52]		9
7. K.K. Ramakrishnan (AT&T Labs Research)		10
8. Final Questions		12
9. Security Considerations		12
10. IANA Considerations		12
11. References		12
11.1. Normative References		12
Author's Addresses		15

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. He presented on the characteristics of the Data Center. Igor is the principle architect at Yahoo.

Internet-Draft BGP Convergence Methodology July 2011
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].

Introduction (Benson Schliesser and Linda Dunbar) [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.

Michael K. Smith, AdHost [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.

Scott Whyte - "Data Centers: Inside the cloud" [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 load 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.

Igor Gashinsky - "Datacenter Scalability Panel" [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 ever 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.

Jim Rees and Manish Karir (Merit Network Inc.) - "ARP Traffic Study"
[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.]

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

Internet-Draft BGP Convergence Methodology July 2011
migration time, pause time, and network bandwidth, and b) CloudNets
use in disaster scenarios.

Questions: None

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.

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.

IANA Considerations

This document requires no IANA considerations.

References

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.

[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] Schiesser, B. & Dunbar, L. "ARMD Call for
Investigation", [http://www.ietf.org/id/draft-ietf-armd-
call-for-investigation-00.txt](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=MTgw
NiZuYW5vZzUy&nm=nanog52](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/Tuesda
y/Gahinsky-3-Y-Datacenter-scalability.pdf](http://www.nanog.org/meetings/nanog52/presentations/Tuesday/Gahinsky-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/Tuesda
y/Ramakrishnan-5-KK%20-att.pdf](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=201
1&sp=Karir&printvs=1](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](http://www.ietf.org/drafts/draft-karir-armd-statistics-00.txt), [early copy received on 7/1/2011].

Internet-Draft BGP Convergence Methodology July 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].

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

ARMD
Internet-Draft
Intended status: Informational
Expires: November 29, 2011

B. Schliesser
Cisco Systems, Inc.
L. Dunbar
Huawei Technologies
May 28, 2011

ARMD Call for Investigation
draft-ietf-armd-call-for-investigation-00

Abstract

This document is a call for investigation into the topic of address resolution in massive datacenters. It describes the intended work of the ARMD working group, providing both context and direction for investigating the issues outlined in the working group's charter.

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). Note that other groups may also distribute working documents as Internet-Drafts. The list of current Internet-Drafts is at <http://datatracker.ietf.org/drafts/current/>.

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."

This Internet-Draft will expire on November 29, 2011.

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 Simplified BSD License.

Table of Contents

1. Call for Investigation	3
1.1. Context	3
1.2. Questions	4
2. Acknowledgements	5
3. IANA Considerations	5
4. Security Considerations	5
Authors' Addresses	5

1. Call for Investigation

1.1. Context

Modern datacenters are increasingly used to support advanced services, such as multi-tenant hosting, cloud, and Internet-scale websites. Many of these datacenter facilities are being built to a much larger scale than previous generations. As a result, datacenter network infrastructure is being stressed in a number of dimensions and traditional limits to scale are being tested. One such aspect, being investigated by the ARMD working group, is the scaling of address resolution between the network (L3) and link (L2) layers of modern datacenter networks.

In many cases, datacenter operators are responsible for provisioning and running everything from routers, switches, load balancers, firewalls, servers, and storage infrastructure. Further, with the introduction of virtualization technology the capacity of these elements is increasing. Each physical device attached to the network may now represent multiple logical instances, and may expose those instances through unique MAC and/or IP addresses. For instance, with the introduction of Virtual Machine technology the operator can reduce wasted resources and achieve greater flexibility by instantiating multiple hosts on each physical server resource. Likewise virtual storage volumes, virtual routers, etc, may all exist in larger numbers.

This virtualization trend contributes to both the increased scale and management complexity of these datacenter environments. The flexibility of VM placement, including migration between different physical resources, has increased datacenter administrators' ability to instantiate VMs where the resources are, i.e. being able to relocate hosts from over-utilized servers to underutilized servers. There is a growing trend towards using resource-aware algorithms (e.g. evaluating energy, bandwidth, memory, CPU, etc) to determine placement that satisfies the processing and redundancy requirements of each VM while using the minimal number of physical resources. Fundamentally, such datacenter management tools are responsible for making trade-offs between different dimensions of scale, which can be difficult in very large and dynamic environments, and in all cases requires a significantly stronger understanding of platform capabilities.

In this environment, IP subnets can extend throughout multiple racks and/or rows in a data center, sometimes throughout multiple sites. There are cases, such as HPC and cloud datacenters, where the number of hosts in a single subnet (on a single segment) is growing. In addition to the more recent VM environment, traditional organic

growth of physical hosts can also cause L2 segments to be extended throughout massive datacenters. Availability / redundancy requirements, subnet size requirements (versus port density), and cost issues can all contribute to the growth and/or extension of segments.

The business demands and workloads for data centers have changed greatly over last 10 years, however some fundamental networking limitations remain unexplored. Even though deployed networks generally do work, this often is because they're designed around known limitations (and redesigned around newly discovered limitations as time goes on). There are datacenter networks that work fine until something changes, such as scale, at which time they're "fixed". Often the "fix" also introduces undesired limitations. An example of this is a datacenter that moves from a flat L2 topology to a L3 core with multiple segregated L2 domains due to scale limitations, and subsequently is unable to distribute clustered servers beyond the boundaries of the "pod" in which a VLAN scope can be configured.

1.2. Questions

The initial goal of the ARMD working group is to document the limiting factors in address resolution scale and the problems associated with exceeding those limits. Subsequently, the working group will identify operational solutions to these problems (to be promoted as Best Current Practice) or will identify gaps in existing solutions (for exploration in subsequent work). Thus the ARMD working group is asked to consider the following questions:

1. What are the scaling characteristics of modern datacenter networks (e.g. "dimensions" of scale and their normal ranges) that are relevant to address resolution?
2. What are the operational problems related to address resolution in the modern datacenter environment?
3. What is the relationship between scaling characteristics of datacenter networks (question #1) and operational problems related to address resolution (question #2)?
4. What, if any, are alternative solutions to the operational problems of address resolution at massive scale?
5. What, if any, are the "gaps" in existing solutions?

2. Acknowledgements

The authors would like to thank Ron Bonica for his significant contributions to this text.

3. IANA Considerations

This memo includes no request to IANA.

4. Security Considerations

This document does not, by itself, introduce any specific security considerations. However, this document calls for further investigation into subject matter that may require significant consideration of security issues. It is anticipated that documents submitted in response to this call for investigation will include appropriate Security Considerations text.

Authors' Addresses

Benson Schliesser
Cisco Systems, Inc.

Email: bschlies@cisco.com

Linda Dunbar
Huawei Technologies

Email: ldunbar@huawei.com

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

M. Karir
J. Rees
Merit Network Inc.

July 10, 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/lid-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 10, 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.....	3
2. Terminology.....	3
3. Factors That Might Impact ARP/ND Performance.....	4
3.1. Number of Hosts.....	4
3.2. Traffic Patterns.....	4
3.3. Network Events.....	4
3.4. Address Resolution Implementations.....	4
3.5. Layer 2 Network Topology.....	5
4. Experiments and Measurements.....	5
4.1. Experiment Architecture.....	5
4.2. Impact of Number of Hosts.....	8
4.3. Impact of Traffic Patterns.....	8
4.4. Impact of Network Events.....	9
4.5. Implementation Issues.....	10
4.6. Experiment Limitations.....	10
5. Emulating Address Resolution Behavior.....	11
6. Conclusion and Recommendation.....	11
7. Manageability Considerations.....	11
8. Security Considerations.....	11
9. IANA Considerations.....	12

10. Acknowledgments.....	12
11. References.....	12
Authors' Addresses.....	12
Intellectual Property Statement.....	13
Disclaimer of Validity.....	13

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.

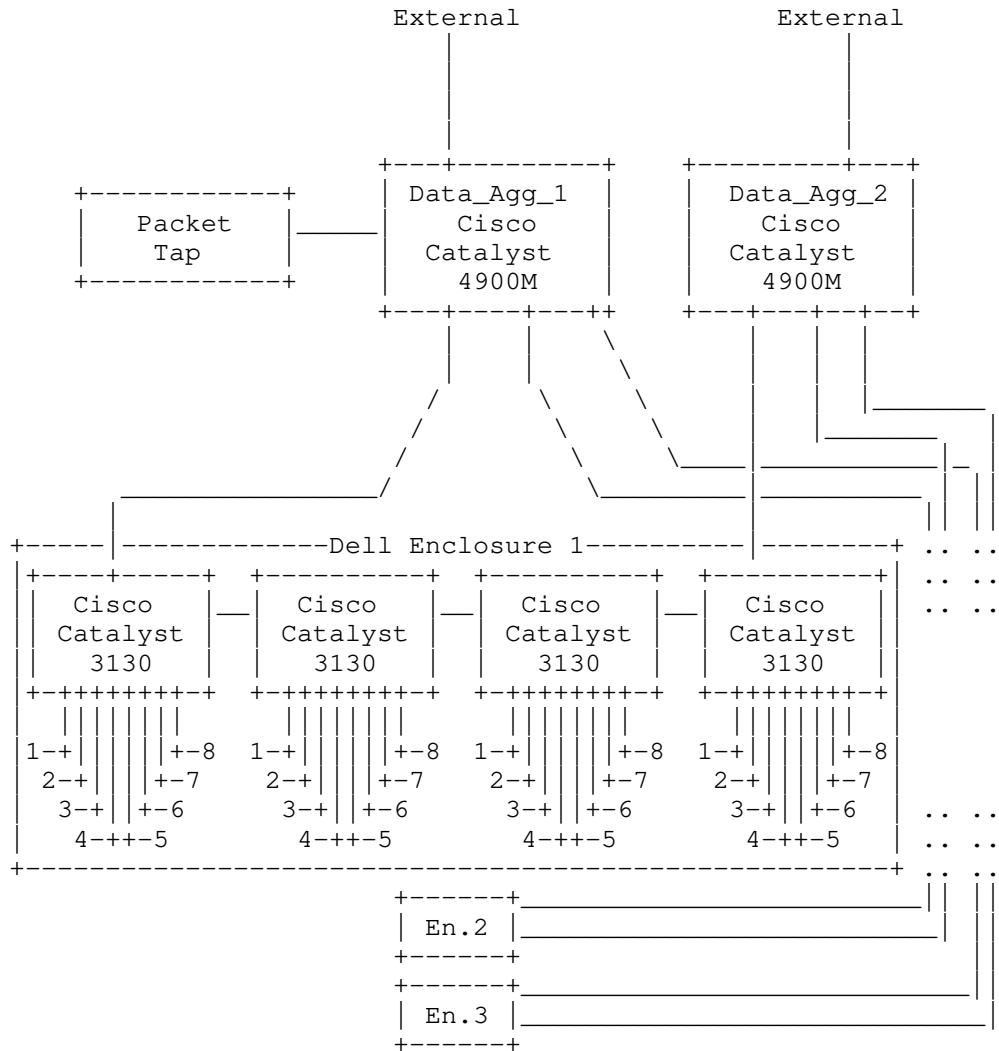
Using this resource pool we create a 500-node testbed based on Centos 5.5. We use custom command and control software that allows us to control these nodes for our experiments. This allows us to issue commands to all nodes to start/stop services and traffic generation scripts. We also use a custom traffic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



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 pattern 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](http://www.nanog.org/meetings/nanog52/presentations/Tuesday/Karir-4-ARP-Study-Merit%20Network.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.

ARMD
Internet Draft
Intended status: Informational
Expires: September 2011

Y. Li
Huawei Technologies
March 11, 2011

Problem statement on address resolution in virtual machine migration
draft-liyz-armd-vm-migration-ps-01.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/ietf/lid-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 September 11, 2011.

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.

Abstract

VM migration is one of the key features provided by larger scaled virtualized data center. Various optimizations for address resolution in such network are expected to be provided by ARMD. This draft describes the problems that are introduced by VM migration. It is expected that solutions provided by ARMD would address these problems.

Table of Contents

1. Introduction	2
2. Conventions used in this document.....	5
3. Some dimensions to consider in supporting VM migration	5
4. ARP Problems in address resolution in VM migration.....	5
5. Security Considerations.....	9
6. IANA Considerations	9
7. Conclusions	9
8. References	10
8.1. Normative References.....	10
8.2. Informative References.....	10
9. Acknowledgments	10

1. Introduction

When virtualization is used in data center, it makes the server management more flexible and consequently more complex. One of the reasons is it would be much easier to move a VM (virtual machine) without the service interruption among physical servers. It is called VM migration. VM migration may occur due to server pool re-arrangement for maintenance, relocation, energy saving, load balancing, utilization optimization and other management purposes.

Figure 1 shows a typical VM migration scenario within a data center. VM1 moves from server 1 to server 2. VM migration is under control of the virtual machine management tools. It is known in advance by VM manager that where the VM would be moved to. Movement could occur between different servers of the same rack or across different racks or even across data centers.

The assumptions of VM migration include

- o VM does not change its MAC and IP address after migration

- o Service provided by VM should not be interrupted. Some packet loss may be observed at the moment of migration; however it should be recoverable by upper layer protocol and should not cause connection termination.

VM itself has no knowledge about its movement and therefore it should not be expected that VM would do anything special to accommodate the migration. On the other hand, hypervisor in a server participates in the whole migration process. Hypervisor in the destination server knows when the migration finishes and usually it will send certain data or control packet to signal the network entities that VM migration completes and it is ready to receive packets at the new location. Such signaling packet may be gratuitous ARP request, gratuitous ARP reply or reverse ARP depending on different implementation.

It has been shown in [I-D. dunbar-arp-for-large-dc-problem-statement] that there are basically two types of approaches used in virtualized larger layer 2 data center to solve the scaling issue,

1. Address translation: map raw flat MAC address to some hierarchical or manageable MAC address.
2. Address encapsulation: use additional header to encapsulate the frame/packet.

Either address translation or encapsulation could be performed by address registration or source address learning. In any case, VM live migration is a fundamental scenario to handle. The following sections talk about the problems caused by VM migration.

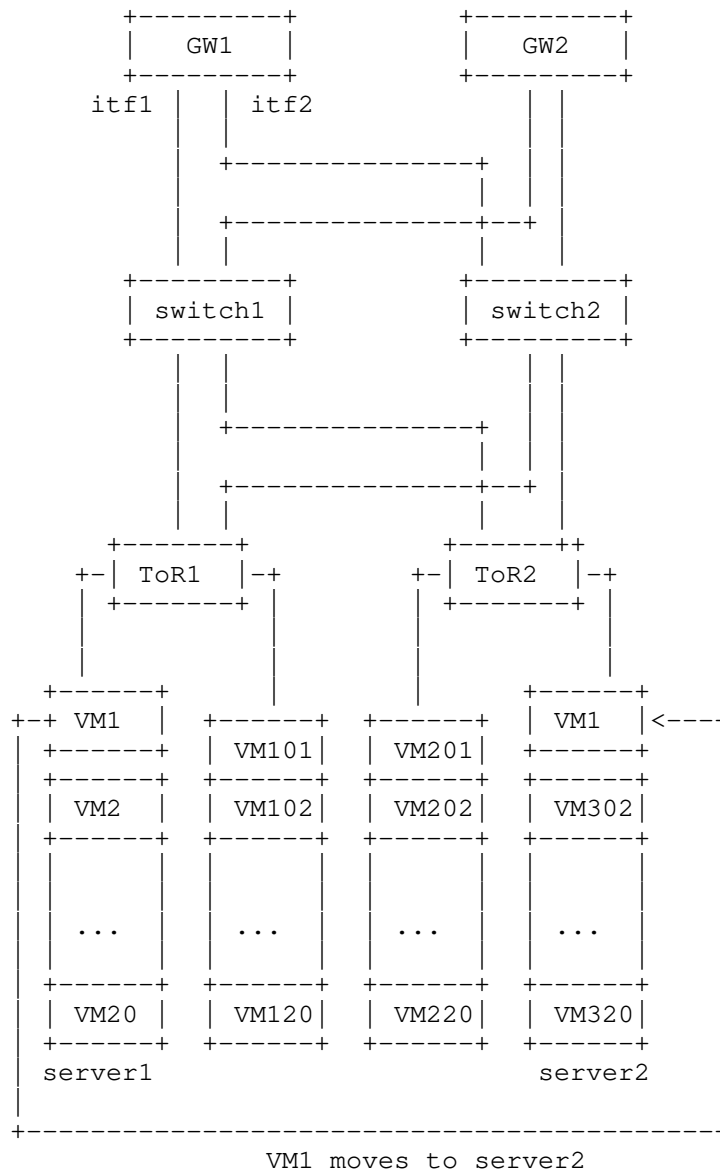


Figure 1 VM migration scenario

2. 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 [RFC2119].

3. Some dimensions to consider in supporting VM migration

When we investigate the impact on ARP traffic by VM migration in data center, there are several dimensions to examine.

- o Network topology. VM can be moved within a single layer 2 domain in current practice. The range of the domain restricts the movement. Therefore position of default gateways normally determines the size of the layer 2 network as they terminate the layer 2 traffic and handle the layer 3 traffic. If the default gateway is aligned with ToR, VM can only migrate within the same rack. If the default gateway is aligned with core switches, VM can be moved within the whole network. Therefore larger sized layer 2 network is more preferred considering VM migration.
- o Protocol used at layer 2. Traditionally STP is used. In order to enjoy more efficient use of all links and faster convergence and support multipathing for fat tree structure based data center, routing based layer 2 protocol like TRILL or SPB are expected to be used in data center. They both provide additional encapsulation at the edge switches and make the core nodes simpler at the forwarding plane. Different operational recommendation may be needed for each.

4. ARP Problems in address resolution in VM migration

Take figure 1 as example. During the process of VM1 movement, other hosts may still keep sending data packet to VM1. The switches including ToR1 have no knowledge that VM1 is going to move. All the packets still go to server 1 as normal. At the moment VM1 stops receiving packet from server 1, the incoming packet could be lost as the destination becomes a black hole to other hosts. After a short while, VM1 should be able to receive the packet from its new location server 2. It is very common that hypervisor at server 2 will flood a gratuitous ARP request/reply for VM1 to inform the whole broadcast domain about VM1's new location.

In traditional switches, there is no ARP table. Only routers/gateways keep the ARP table. In some of the approaches, switches have the ARP

cache for local host and/or remote host. We will study the impact for both.

4.1 No ARP message to indicate VM having left a server.

Gratuitous ARP is a message to inform others a new node coming up for free. It is used for IP/MAC correspondence announcement. At same time, switches perform source MAC address learning to know the MAC/port/vlan correspondence. However there is no gratuitous ARP "leave" message to make others forget the previous learned source address and location information. Aging is a normal way to delete the cached information. Black hole may last as long as aging out time.

There are several ways to make it up.

- o Operationally if the VM sends out the gratuitous ARP or reverse ARP right after the migration, and the message is not lost, it will fresh the ARP table entry on gateways and switches. It is the most common way given that migration process, i.e. the time from VM stopping receiving frame at old location to VM starting receiving frame at new location, is very short and the frame lost is rare.
- o In virtualized system architecture, virtual machine management tool like vCenter knows a VM is going to move at management level. Therefore it is possible to delete the stale cache through management plane and it needs collaboration between virtual machine manager and network manager.
- o Use some lightweight keepalive mechanism to guarantee the freshness of the local ARP entry. It is called ARP detection in some implementations. It decreases the possibility of re-issuing gratuitous ARP for silent hosts. If an ARP entry becomes invalid, some specific message needs to be flooded to let remote switches "forget" the entry if switch also has the ARP cache for remote hosts.

4.2 Uncertainty of ARP message type after VM migration.

Currently there is no standard behavior defined for hypervisor in VM migration. Hypervisor may send gratuitous ARP request/reply and even reverse ARP after migration completes. The reason for sending the signaling message is to inform the switches and gateways about the new location of VM1 and make them have the correct entry for interface/port in the ARP/MAC table.

However, there are a large variety of ARP implementations. We have tested on one of switches in market on various ARP messages; the result is in figure 2.

The testing scenario is as follows. VM1 moves from server 1 to server 2 which connect to GW1 via interface 1 and interface 2 accordingly. Before migration, ARP table of GW1 has the entry to include IP/MAC of VM1 and its outgoing interface is itf1. After migration, hypervisor of server 2 may flood ARP or other signaling message; it is also possible that it keeps silent and does not send out any signaling packet in which case black hole problem would become more significant. The expected result should be GW1 updates its ARP table entry to correlate VM1 with interface 2 (itf2) as soon as possible when VM finishes migration.

#	packet sent aft VM1 migration	Is VM1's interface updated to itf2 on GW1?
1	std gratuitous ARP	Y
2	broadcast ARP reply	N
3	RARP	N
4	ARP request with GW1 as target IP	Y
5	ARP request with other host as target IP	N
6	unicast ARP reply with GW1 as destination	Y
7	unicast ARP reply with other host as destination	N

Figure 2 Test result of GW ARP table update in VM migration

There are various implementations of switches and hypervisors. Figure 2 shows one example that depending on the type of ARP message sent by hypervisor and handling of switch, result may not be always as what we expect.

It is recommended that interface number for an ARP table entry on gateway should be updated for any ARP messages including ARP request/reply and reverse ARP no matter if the frame is destined for itself.

4.3 ARP message unreliable delivery

Gratuitous ARP from an end host is normally sent three times in order to survive from frame loss. However it is hard to 100% avoid ARP frame loss. Some analysis says a typical congestion is about 10-20 seconds which is longer than 3 retries of gratuitous ARP. In case the ARP frames are lost after VM migration, the gateway is not able to correctly update the corresponding interface number in ARP table entry. For inbound traffic from gateway, the gateway will keep sending it to the old location which is a black hole. It is noted that the ARP table will not be updated by data frames. Hence even the VM sends out data frame from new location, gateway will not update the relevant entry of ARP table.

For internal traffic within data center, if switches do not have any ARP cache, MAC/port correspondence will be updated accordingly along the path. As most of the data traffic should be bidirectional, MAC table should be correctly updated after a short while. Everything should be ok. On the other hand, if switches have ARP caching table, situation would be more completed depending on where the frame is lost, if switches cache remote ARP entry.

If ARP table is updated by data frames in addition ARP frame, it will solve most of the problems here. However, it may bring some performance and security issue.

4.4 Duplicate address detection

Gratuitous ARP is also used for duplicate address detection. For example, in Windows NT 4.0 with Service Pack 3 or higher installed, a statically addressed Windows NT computer will perform a gratuitous ARP up to 3 times: 1 time when the TCP/IP stack initializes, and 2 more times after .5 and 1 second intervals, if no response is received. Whenever a statically configured IP address is changed, Windows NT sends a single gratuitous ARP. If Windows NT receives a response to a gratuitous ARP, it disables the interface that issued the gratuitous ARP, generates an event (event ID 26), and generates a pop-up dialog box on the console warning the user that a duplicate IP address has been detected resulting in the shutdown of the affected interface. For DHCP leased address, Windows NT sends a single gratuitous ARP.

VM migration normally takes time in magnitude of second depending on the amount of memory to be copied over at the last stage. If another VM starts up and tries to use the same IP address of the migrated VM right within its migration process, there will be no duplicate address detected. Therefore the new VM can safely uses that IP address. Then after the migrated VM completes the movement, there will be duplicated IP address running at same time or migrated VM will block itself from using that IP address. Neither behavior is desired.

5. Security Considerations

It may not be easy to tell if an ARP sent from a new location is really for a migrated VM or it is a spoofed one. With VM migration, some security mechanisms are not applicable any more, like:

- o MAC locking: locking a MAC address to a specific physical port of the switch.
- o DHCP snooping: binding IP/MAC by snooping DHCP ACK to port of switch. VM does not send DHCP request again after migration. Some mechanism should be introduced to move the binding to the new port in migration case.

VM migration itself does not introduce more risk to ARP messages. However some existing solutions to solve ARP security issues may wrongly treat ARP after migration as illegal one.

6. IANA Considerations

This document requires no IANA actions.

7. Conclusions

VM migration brings extra problem to larger scale virtualized data center. Any solution in ARMD, like directory based address resolution, distributed caching, or specially designed control protocol, should consider the VM migration carefully. It is suggested to include the information from the draft in the problem statement of impact on address resolution for massive number of hosts in the data center.

8. References

8.1. Normative References

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

8.2. Informative References

[I-D. dunbar-arp-for-large-dc-problem-statement]Dunbar, L. and Hares,
S., " Scalable Address Resolution for Large Data Center Problem
Statements", draft-dunbar-arp-for-large-dc-problem-statement-00, July
2010.

9. Acknowledgments

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

Authors' Addresses

Li Yizhou
Huawei Technologies
101 Software Avenue,
Nanjing 210012
China

Phone: +86-25-56622310
Email: liyizhou@huawei.com

Internet Engineering Task Force
Internet-Draft
Intended status: Informational
Expires: January 6, 2012

T. Narten
IBM
July 5, 2011

Problem Statement for ARMD
draft-narten-armd-problem-statement-00

Abstract

This document examines problems related to the massive scaling of data centers. Our initial scope is relatively narrow. Specifically, we focus on address resolution (ARP and ND) within a single L2 broadcast domain, in which all nodes are within the same physical data center. From an IP perspective, the entire L2 network comprises one IP subnet or IPv6 "link". Data centers in which a single L2 network spans multiple geographic locations are out-of-scope.

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). Note that other groups may also distribute working documents as Internet-Drafts. The list of current Internet-Drafts is at <http://datatracker.ietf.org/drafts/current/>.

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."

This Internet-Draft will expire on January 6, 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 Simplified BSD License.

Table of Contents

1. Introduction	3
2. Background	3
3. Out-of-Scope Topics	4
4. Address Resolution	5
5. Summary	5
6. Acknowledgements	6
7. IANA Considerations	6
8. Security Considerations	6
Author's Address	6

1. Introduction

This document examines problems related to the massive scaling of data centers. Our initial scope is relatively narrow. Specifically, we focus on address resolution (ARP and ND) within a single L2 broadcast domain, in which all nodes are within the same physical data center. From an IP perspective, the entire L2 network comprises one IP subnet or IPv6 "link". Data centers in which a single L2 network spans multiple geographic locations are out-of-scope.

This document is intended to support the ARMD WG identify its work areas. The scope of this document intentionally starts out narrow, mirroring the ARMD WG charter. Expanding the scope requires careful thought, as the topic of scaling data centers generally has an almost unbounded potential scope. It is important that this group restrict itself to considering problems that are widespread and that it has the ability to solve.

2. Background

Large, flat L2 networks have long been known to have scaling problems. As the size of an L2 network increases, the level of broadcast traffic from protocols like ARP increases. Large amounts of broadcast traffic pose a particular burden because every device (switch, host and router) must process and possibly act on such traffic. In addition, large L2 networks can be subject to "broadcast storms". The conventional wisdom for addressing such problems has been to say "don't do that". That is, split the L2 network into multiple separate networks, each operating as its own L3/IP subnet. Unfortunately, this conflicts in some ways with the current trend of virtualized systems.

Server virtualization is fast becoming the norm in data centers. With server virtualization, each physical server supports multiple virtual servers, each running its own operating system, middleware and applications. Virtualization is a key enabler of workload agility, i.e. allowing any server to host any application and providing the flexibility of adding, shrinking, or moving services among the physical infrastructure. Server virtualization provides numerous benefits, including higher utilization, increased data security, reduced user downtime, and even significant power conservation, along with the promise of a more flexible and dynamic computing environment.

The greatest flexibility in VM management occurs when it is possible to easily move a VM from one place within the data center to another. Unfortunately, movement of services within a data center is easiest

when movement takes place within a single IP subnet, that is, within a single L2 broadcast domain. Typically, when a VM is moved, it retains such state as its IP address. That way, no changes on the either the VM itself, or on clients communicating with the VM are needed. In contrast, if a VM moves to a new IP subnet, its address must change, and clients may need to be made aware of that change. From a VM management perspective, life is much simpler if all servers are on a single large L2 network.

With virtualization, a single server now hosts multiple VMs, each having its own IP address. Consequently, the number of addresses per machine (and hence per subnet) is increasing, even if the number of physical machines stays constant. Today, it is not uncommon to support 10 VMs per physical server. In a few years, the number will likely reach 100 VMs per physical server.

In the past, services were static in the sense that they tended to stay in one physical place. A service installed on a machine would stay on that machine because the cost of moving a service elsewhere was generally high. Moreover, services would tend to be placed in such a way as to encourage communication locality. That is, servers would be physically located near the services they accessed most heavily. The network traffic patterns in such environments could thus be optimized, in some cases keeping significant traffic local to one network segment. In these more static and carefully managed environments, it was possible to build networks that approached scaling limitations, but did not actually cross the threshold.

Today, with VM migration becoming increasingly common, traffic patterns are becoming more diverse and changing. In particular, there can easily be less locality of network traffic as services are moved for such reasons as reducing overall power usage (by consolidating VMs and powering off idle machine) or to move a virtual service to a physical server with more capacity or a lower load. In today's changing environments, it is becoming more difficult to engineer networks as traffic patterns continually shift as VMs move around.

In summary, both the size and density of L2 networks is increasing, with the increased deployment of VMs putting pressure on creating ever larger L2 networks. Today, there are already data centers with 120,000 physical machines. That number will only increase going forward. In addition, traffic patterns within a data center are changing.

3. Out-of-Scope Topics

At the present time, the following items are out-of-scope for this

document.

Cloud Computing - Cloud Computing is broad topic with many definitions. Without a clear (and probably narrow) scoping of what aspect of Cloud Computing to include in this effort, it will remain out-of-scope.

L3 Links - ARP and ND operate on individual links. Consequently, this effort is currently restricted to L2 networks

Geographically Extended Network Segments - Geographically separated L2 networks introduce their own complexity. For example, the bandwidth of links may be reduced compared to the local LAN, and round-trip delays become more of a factor. At the present time, such scenarios are out-of-scope.

VPNs - It is assumed that L2 VLANs are commonly in use to segregate traffic. At the present time, it is unclear how that impacts the problem statement for ARMD. While the limit of a maximum of 4095 VLANs may be a problem for large data centers, addressing it is out-of-scope for this document. L3 VPNs, are also out-of-scope, as are all L3 scenarios.

4. Address Resolution

In IPv4, ARP performs address resolution. To determine the link-layer address of a given IP address, a node broadcasts an ARP Request. The request is flooded to all portions of the L2 network, and the node with the requested IP address replies with an ARP response. ARP is an old protocol, and by current standards, is sparsely documented. For example, there are no clear requirement for retransmitting ARP requests in the absence of replies. Consequently, implementations vary in the details of what they actually implement.

From a scaling perspective, there are two main problems with ARP. First, it uses broadcast, and any network with a large number of attached hosts will result in a large amount of broadcast ARP traffic. The second problem is that it is not feasible to change host implementations of ARP - current implementations are too widely entrenched, and any changes to host implementations of ARP would take years to become sufficient deployed to matter.

5. Summary

This document outlines the scope of the problem the ARMD effort is intended to address. It intentionally begins with a very narrow

scope of kind of data center ARMD is focusing on. The scope can be expanded, but only after identifying shared aspects of data centers that can be clearly defined and scoped.

6. Acknowledgements

7. IANA Considerations

8. Security Considerations

Author's Address

Thomas Narten
IBM

Email: narten@us.ibm.com

ARMD
Internet Draft
Intended status: Information Track
Expires: December 2011

Ning So
Verizon
L. Dunbar
Huawei
June 30, 2011

Address Resolution Requirements for VPN-oriented Data Center
Services
draft-so-armd-vdcs-ar-00.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/lid-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 December 30, 2011.

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

VPN-oriented data center services seamlessly integrate the computing and storage resources in data centers and the users together with the traditional VPN services. This draft describes the address resolution issues and requirements induced by those services.

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	2
2. Terminology	3
3. VDCS service description.....	3
3.1. Components of VDC	4
3.2. Networking related components in support of VDCS	5
4. Address resolution Scaling Issue for VDCS.....	6
4.1. Address Resolution for VMs attached to L2VPN.....	6
4.2. Address Resolution for VMs attached to L3VPN.....	7
5. Conclusion and Recommendation.....	9
6. Manageability Considerations.....	9
7. Security Considerations.....	9
8. IANA Considerations	9
9. Acknowledgments	9
10. References	9
Authors' Addresses	10
Intellectual Property Statement.....	10
Disclaimer of Validity	11

1. Introduction

VPN-oriented Data Center Services (VDCS) integrate the virtual resources in data centers and user together using VPN as the common link. This kind of service is attractive to customers who often do not want to use public Internet to access data center resources. VDCS also have more restrictive requirements on what and how the virtualized data center resources can be shared. In addition, it provides a common service operational management framework using VPN as the central control point(s).

2. Terminology

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.

FDB: Filtering Database for Bridge or Layer 2 switch

SA: Source Address

ToR: Top of Rack Switch. It is also known as access switch

VDCS: VPN oriented data center services

VM: Virtual Machines

VPN: Virtual Private Network

VPN-o-CS: VPN oriented Computing Service

3. VDCS service description

Many data centers offer virtualized services today, allowing clients to lease virtual data center resources without actually owning any physical servers or storage devices. However, majority of those services do not include network infrastructure. Intra-data center, inter-data center networks, and the networks connecting users to data centers are designed and operated separately from the data center server/storage systems. It is difficult for customers to integrate the leased virtual data center resources with their own internal data center resources, and make those leased resources appearing as if they come from their internal infrastructure.

VDCS has the following characteristics:

A secure collection of servers and/or virtual machines spanning one or more data centers.

All the applications running on the Virtual resources in network provider's data centers are connected with the enterprise's VPN in the same way as applications running over enterprise's internal data centers. Therefore, the enterprises can treat those resources as if they are from their internal data centers.

Provide the VPN equivalent level of traffic segregation and privacy for those virtual resources attached to the VPN.

Make the virtual resources' location known to VPN customers.

Created by network provider with no end host configuration.

Allow VMs and user devices using VDCS associated with one VPN to be partitioned into multiple subnets while still retain the detailed knowledge of each other.

Allow VPN clients to use private IP addresses (IPv4 or IPv6) for VDCS.

3.1. Components of VDCS

There are many components in VDCS system, including (but not limited to):

Network back office support systems, such as provisioning, billing, and etc,

VPN management systems such as monitoring, reporting, trouble shooting, and etc.

Data center resource monitoring systems, which include monitoring the utilization of servers and storage devices in data centers

Data center resource management systems, which include VMs placement to servers and racks based on the criteria associated with VMs.

Others.

This draft only focuses on networking (switching and routing) related components within VDCS framework.

3.2. Networking related components in support of VDCS

In the figure below, Vx represents a VM or a server belonging to VPN-x. The data center depicted in the figure has VMs belonging to 5 different VPNs, VPN-1, VPN-2, VPN-3, VPN-4, and VPN-5. Most data centers have many rows of server racks. Each rack holds many servers and has 1 or 2 Top of Rack (ToR) switches. Each server can have many VMs. The ToRs can be connected to aggregation switches/routers, which are then connected to Data Center gateway switches/routers. In some data centers, ToRs may be directly connected to Data Center gateway switches/routers.

It is essential to segregate traffic from VMs belonging to different VPNs within one data center and across multiple data centers. VLAN is usually used to segregate traffic from different VPNs within one data center. However, when a data center needs to house virtual machines belonging to more than 4095 VPNs, alternative segregation methods have to be used.

The virtual machines in data center can be connected to VDCS via L2VPN or L3VPN. For VMs belonging to L3VPN, the data center gateway router and the VPN PE router have to maintain detailed VRF tables that contain all the VM IP addresses associated with the each VPN. For VMs belonging to L2VPN, the data center gateway switch and the VPN edge switch have to maintain detailed Learned MAC Table that contains all the VM MAC addresses associated with each VPN.

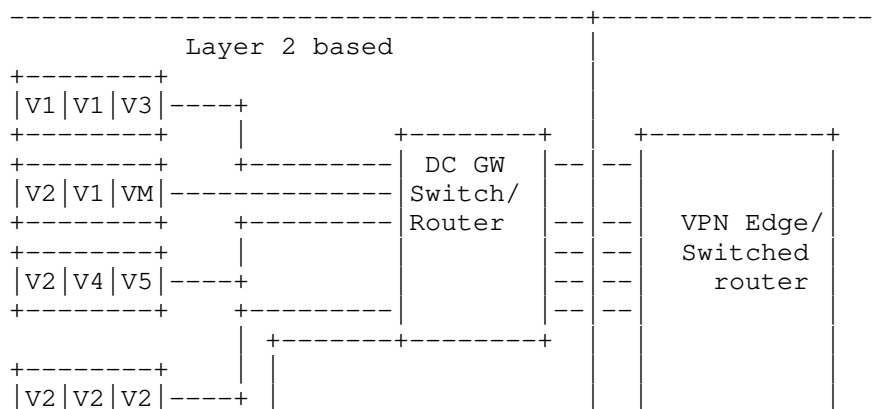




Figure 1 VMs and Network in Data Center

When VMs belonging to one VPN are partitioned into multiple subnets, it is necessary to have VLANs or other mechanisms to segregate traffic from different subnets belonging to one VPN.

4. Address resolution Scaling Issue for VDCS

4.1. Address Resolution for VMs attached to L2VPN

Before servers in a data center are instantiated with VMs for a particular VPLS L2VPN for the very first time (i.e. there is no VMs in the data center belonging to the L2VPN yet), the data center gateway router (CE router) should have the base VPLS configured already, which means a full mesh of pseudo-wires between L2VPN PEs already exist. The CE should have an attachment circuit (AC) built for the VPLS service between CE and PE.

At the time of VDCS instantiation, the new VMs' MAC addresses are learned and added to the CE and PE's MAC Table, so they can be learned by other switches and end stations already on the L2VPN in multiple sites as if they are on one LAN.

When a host or a VM in a data center needs to communicate with another host/VM in the L2VPN, an ARP (IPv4) or a ND(IPv6) is flooded to all PWs and all ACs (except the one from which the request is coming from).

Under this scenario, all VMs' MAC addresses belonging to a particular L2VPN are visible to each other. And the L2VPN's PEs and VSIs have to learn and maintain the MAC and VLAN addresses for all the hosts/VMs associated with this L2VPN. This may lead to address table scalability problems for data center VSI and L2VPN PE.

For example, assuming there are 1000 L2VPNs with hosts/VMs residing in this data center. That translates to 1000 VSIs on the CE, with

each VSI containing the entire MAC and VLAN mapping for all the switches and end-stations associated with all the L2VPNs. This requires a very large amount of memory for the data center gateway switch/router using current technology.

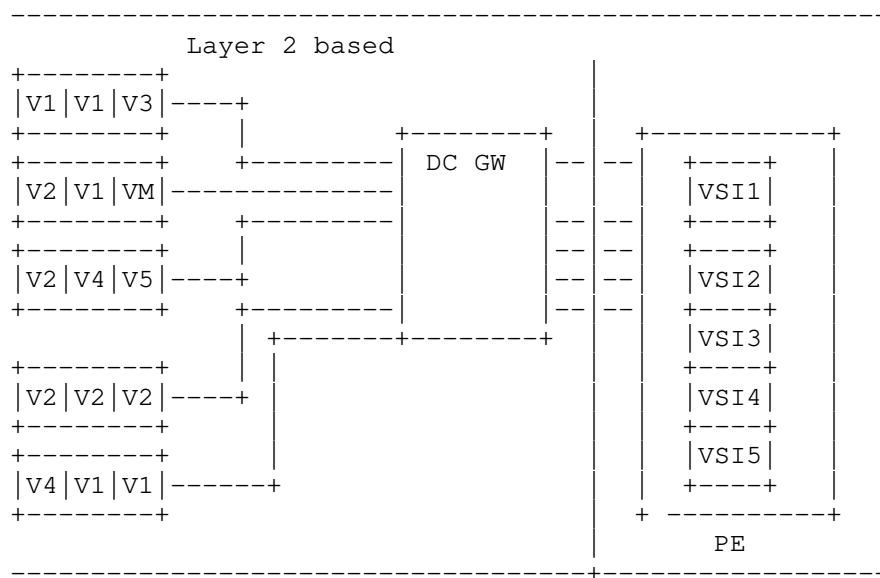


Figure 2 L2VPN associated VMs in Data Center

4.2. Address Resolution for VMs attached to L3VPN

When servers in a data center are instantiated with VMs for a particular L3VPN for the very first time (i.e. there were no VMs in the data center belonging to the L3VPN yet), it assumes that all the necessary L3VPN configuration has already been completed on the data center gateway router (CE) and the L3VPN edge router (PE). There are two scenarios for VMs attached to L3VPN:

Scenario 1: all the VMs belonging to the L3VPN client are added as a separate site for the L3VPN. Under this scenario, the provider data center becomes the additional site (or peers) to the L3VPN.

Scenario 2: Hosts or applications in client's own data centers (or premises) see those VMs attached to L3VPN as if they are from the same subnets. Under this scenario, the traditional "subnet" concept is broken. VMs in the data center have to be connected to their designated sites as if they are in one subnet.

Under scenario 1, the APR/ND broadcast/multicast requests are terminated at the CE. Similar to the condition described in the last section on VMs attached to L2VPN, all IP addresses associated with all L3VPNs in the data center have to be learned and maintained at the CE and the L3VPN PE router.

This can require a very large amount of memory on the CE and PE router using today's technology, especially when the CE and the PE routers are hosting both L2VPN and L3VPN simultaneously. The amount of memory requirement is even larger if those VMs addresses can't be aggregated.

In addition, it is possible that IP addresses for VMs belonging to different VPNs could be duplicated.

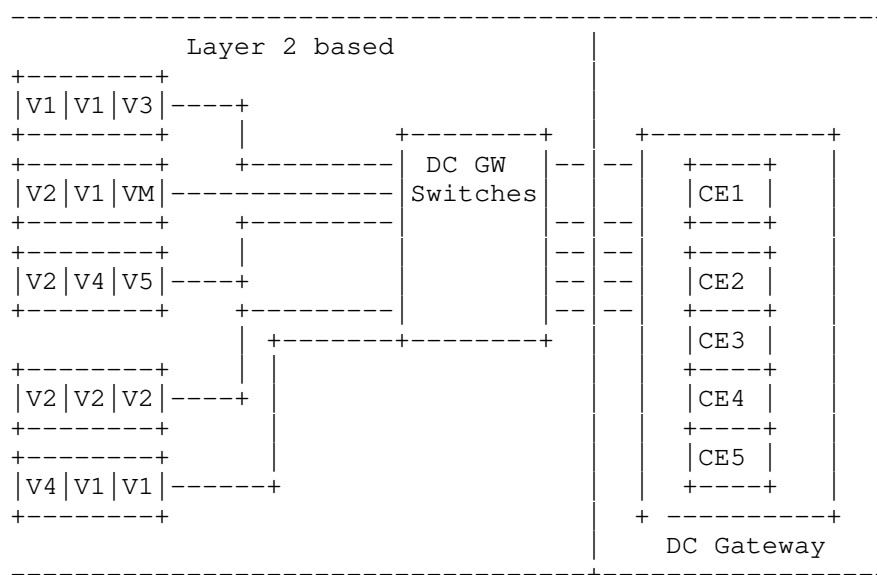


Figure 3 L3VPN associated VMs in Data Center

Under the Scenario 2, the ARP/ND messages from the VMs in the data center have to be flooded to the corresponding sites to which those VMs belonging. The data center gateway routers (CEs or PEs) have to do both L2VPN and L3VPN.

5. Conclusion and Recommendation

Future data center can scale up to millions of virtual machines. Theoretically, network service provider can make their data centers hosting VMs for all of their VPN clients. Using current technology, it is very difficult for routers in data center and at network edge facing the data center to maintain all the VSIs or VRFs needed for the huge number of VPNs and the VPN-associated VMs being deployed.

Therefore, we recommend ARMD WG to investigate alternative solutions on address resolution and address scalability issues to make data center gateway routers capable of supporting the VPN oriented data center services.

6. Manageability Considerations

This document does not add additional manageability considerations.

7. Security Considerations

This document has no additional requirement for security.

8. IANA Considerations

9. 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.

10. References

- [VDCS] So, et al, "Requirement and Framework for VPN-Oriented Data Center Services", draft-so-vdcs-00, June 2011.
- [ARP] D.C. Plummer, "An Ethernet address resolution protocol." RFC826, Nov 1982.

[Microsoft Windows] "Microsoft Windows Server 2003 TCP/IP implementation details."
<http://www.microsoft.com/technet/prodtechnol/windowsserver2003/technologies/networking/tcpip03.msp>, June 2003.

[Scaling Ethernet] Myers, et. al., " Rethinking the Service Model: Scaling Ethernet to a Million Nodes", Carnegie Mellon University and Rice University

[Cost of a Cloud] Greenberg, et. al., "The Cost of a Cloud: Research Problems in Data Center Networks"

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

Authors' Addresses

Ning So
Verizon Inc.
2400 N. Glenville Ave.,
Richardson, TX75082
ning.so@verizonbusiness.com

Linda Dunbar
Huawei Technologies
5340 Legacy Drive, Suite 175
Plano, TX 75024, USA
Phone: (469) 277 5840
Email: ldunbar@huawei.com

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.

