

MPLS Working Group
Internet Draft
Intended status: Standard Track
Expires: January 2011

L. Dunbar
Huawei
S. Hares
Huawei
July 2, 2010

Scalable Address Resolution for Large Data Center Problem Statements
draft-dunbar-arp-for-large-dc-problem-statement-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 January 2, 2009.

Copyright Notice

Copyright (c) 2010 IETF Trust and the persons identified as the document authors. All rights reserved.

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

Internet-Draft ARP for Large DC Problem Statement

July 2010

Abstract

Virtual machines, or virtualized servers, basically allow one physical server to support multiple hosts (20, 30, or hundreds of). As virtual machines are introduced to Data center, the number of hosts within one data center can grow dramatically, which could create tremendous impact to networks and hosts.

This document describes reasons why it is still desirable to have virtual machines in Data Center to be in one Layer 2 network and potential problems this type of Layer 2 network will face. The goal is to justify why it is necessary for IETF to create a working group to work on interoperable and scalable solutions for data center(s) with large number of virtual machines.

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.....	4
3.	Reasons for Virtual Machines in Data Center to stay in Layer 24	
3.1.	Load balance requires group of hosts on same Layer 2.....	4
3.2.	Redundancy requires both active and standby VM on same Layer 2.....	5
3.3.	VM mobility requires them on same Layer 2.....	5
4.	Cloud Computing Service.....	6
5.	Problems facing Layer 2 with large number of hosts.....	7
5.1.	Address Resolution creates significant burden to hosts...	8
5.2.	Large amount of MAC addresses to be learnt by intermediate switches.....	9
5.3.	More chances of unknown broadcast.....	10
6.	Why dividing one Layer 2 into many smaller subnets is not enough?	10
7.	Why IETF needs to develop solutions instead of relying on IEEE802	11
8.	Conclusion and Recommendation.....	11

9. Manageability Considerations.....	12
10. Security Considerations.....	12
11. IANA Considerations.....	12
12. Acknowledgments.....	12
13. References.....	13

Authors' Addresses.....	13
Intellectual Property Statement.....	13
Disclaimer of Validity.....	14

[1. Introduction](#)

Virtual machines are created by server virtualization, which allows the sharing of the underlying physical machine (server) resources among different virtual machines, each running its own operating system. Server virtualization is the key enabler for Cloud Computing services, such as Amazon's EC2 service. Virtual Machine also makes it possible for virtual desktop services, which allow servers in data center(s) to provide virtual desktops to millions of end users.

Servers virtualization provides numerous benefits, including higher utilization rates, improved IT productivity, increased data security, reduced user downtime, and even significant power conservation, and the promise of a more flexible and dynamic computing environment. As a result, many organizations are highly motivated to incorporate server virtualization technologies into their data centers. In fact, ESG research indicates that virtualization is being widely adopted in production environments.

While Servers Virtualization is a great technology for flexible management of server resources, it does impose great challenges to networks which interconnect all the servers in data center(s).

For a typical tree structured Layer 2 network, with one or two aggregation switches connected to a group of Top of Rack (ToR) switches and each ToR switch connected to a group of physical servers (hosts), the number of servers connected in this network is limited to the switches' port count. If ToR switch has 20 downstream ports, there are only 20 servers or hosts connected to the ToR switch. If the Aggregation Switch has 256 ports connecting to ToR switch, there could be up to $20 \times 256 = 5120$ hosts connected to one aggregation switch

when the servers are not virtualized.

When Virtual Machines are introduced to servers, one server can support hundreds of VMs. Hypothetically, if one server supports up to 100 VMs, the same ToR switches and Aggregation switch as above can support up to 512000 hosts. Even if there is enough bandwidth on links to support the traffic volume from all those VMs, other issues associated with Layer 2, like frequent ARP broadcast by hosts, unknown broadcast, are creating a lot of challenges to the network and hosts.

Dunbar

Expires January 2, 2011

[Page 3]

Internet-Draft ARP for Large DC Problem Statement

July 2010

[2. Terminology](#)

Bridge: IEEE802.1Q compliant device. In this draft, Bridge is used interchangeably with Layer 2 switch.

FDB: Filtering Database for Bridge or Layer 2 switch

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

Aggregation switch: a Layer 2 switch which connects a group of ToR switches. It is also known as End of Row switch in data center.

VM: Virtual Machines

[3. Reasons for Virtual Machines in Data Center to stay in Layer 2](#)

Two application scenarios for Virtual Machines deployment are considered here:

- One Data Center with large amount of Virtual Machines
- Cloud Computing Service (Infrastructure as a Service) which requires a large amount of virtual hosts.

[3.1. Load balance requires group of hosts on same Layer 2](#)

Server load balancing is a technique to distribute workload evenly across two or more servers, in order to get optimal resource utilization, minimize response time, and avoid overload. Using

multiple servers with load balancing, instead of a single one, also increase reliability through redundancy. One of the most common applications of load balancing is to provide a single Internet service from multiple servers, sometimes known as a server farm. Commonly, load-balanced systems include popular web sites, large Internet Relay Chat networks, high-bandwidth File Transfer Protocol sites, NNTP servers and DNS servers.

The load balancer typically sits in-line between the client and the hosts that provide the services the client wants to use. Some load balancer requires hosts to have a return route that points back to the load balancer so that return traffic will be processed through it on its way back to the client.

However, for applications with relative smaller amount of traffic going into server and relative large amount of traffic from server, it is desirable to allow reply data from servers go directly to clients without going through the Load Balancer. In this kind of design, called Direct Server Return, all servers in the cluster have same IP addresses as the load balancer. External requests from clients are directly sent to the Load Balancer, which distributes the request to appropriate host among the cluster based on their load. Any of those servers send reply directly out to clients, who see the same IP address regardless which server handles the requests. Under this design it is necessary for Load Balancer and the cluster of hosts to be on same Layer 2 network so that they communicate with each other via their MAC addresses.

[3.2](#). Redundancy requires both active and standby VM on same Layer 2

For redundant servers (or VMs) serving same applications, both Active and Standby servers (VMs) need to have keep-alive messages between them. Since both Active and Standby servers (VMs) might have same IP address, the only way to achieve this is via Layer 2 keep-alive because the Active and Stand-by will have different MAC Addresses.

The VRRP (virtual router redundancy protocol) ([RFC 3768](#)) is designed to increase the availability of the default gateway servicing hosts on the same subnet. Two or more physical routers are then configured to stand for the virtual router, with only one doing the actual routing at any given time. It is necessary for the group of physical

routers serving one virtual router to be on same subnet as their hosts.

3.3. VM mobility requires them on same Layer 2

VM mobility is referring to moving virtual machines from one server to another. To fully realize the benefits of a virtualized server environment, the ability to seamlessly move VMs within a resource pool and still guarantee application performance and security is a must. Mobility adds tremendous value because it enables organizations to:

- Rapidly scale out new applications - Creating golden copies of a VM allows organizations to introduce new applications into a resource pool in dramatically less time. This accelerates the time to market for new applications.
- Balance workloads - The ability to dynamically adjust workloads within a resource pool will enable companies to optimize

Dunbar

Expires January 2, 2011

[Page 5]

Internet-Draft ARP for Large DC Problem Statement

July 2010

performance and minimize power and cooling costs during off hours by dynamically adjusting VM locations.

- Deliver high levels of availability - Mobility guarantees that even in the event of physical infrastructure failure, applications can be quickly moved to another physical resource within that pool, dramatically minimizing downtime and eliminating the need for dedicated redundant infrastructure.
- Recover from a site disaster - This involves the ability to quickly migrate VMs to a remote secondary location in order for operations to resume. Recovering from VMs significantly reduces the amount of time required vs. manually reloading servers from bare metal.

One of the key requirements for VM mobility is the VM maintaining the same IP address after moving to the new location so that its operation can be continued in the new location.

For a VM to maintain the same IP address when moving from Server A to Server B, it is necessary for both servers to be on the same subnet. If Server A and Server B are on different subnets, they will have

different gateway routers. In the subnet where Server A is in, e.g. Subnet A, the VM sends ARP broadcast requests for target IP addresses in the same subnet. For the target IP address not in the Subnet A, the VM sends data frame to default gateway router. When this VM is moved to Server B, if server B is in a different subnet than Server A, e.g. Subnet B, then this VM couldn't even forward data to its default gateway and can't find MAC addresses for other hosts in subnet A.

That is why most VM mobility systems, such as VMware's vMotion, require all hosts in one Layer 2.

[4.](#) Cloud Computing Service

Cloud Computing service, like Amazon's Elastic Compute Cloud (Amazon EC2), allows users (clients) to create their own virtual servers and virtual subnets. There are many potential services which Cloud Computing services could offer to their clients. Here are just some examples of those services:

- A client can request a group of virtual servers and their associated virtual subnet,
- A client can specify policies among multiple subnets they purchased,

Dunbar

Expires January 2, 2011

[Page 6]

Internet-Draft ARP for Large DC Problem Statement

July 2010

- A client can specify preferred geographic locations for some of the virtual servers they purchased,
- A client can specify redundancy criteria, like two virtual servers on two different physical servers or in two different locations, etc.

In order to efficiently support those services, network has to support virtual subnet, i.e. one Layer 2 network spanning across multiple locations, in addition to large amount of hosts in Layer 2.

It is possible for Cloud Computing service to have a network design that each virtual subnet is mapped one independent Layer 2 network. But this kind of design would require huge amount of administrative and planning work to properly partition servers, switches to appropriate Layer 2 network. The problem is that virtual servers and virtual subnets purchased by clients change all the time. It is a lot

of administrative work to change Layer 2 network partition each time there is a client request. Having a large amount of virtual machines in one Layer 2 network can simplify some aspects of Cloud Computing service design and management.

5. Problems facing Layer 2 with large number of hosts

In Layer 2 network, hosts can be attached and re-attached at any location on the network. Hosts use ARP (Address Resolution Protocol) to find the corresponding MAC address of a target host. ARP is a protocol that uses the Ethernet broadcast service for discovering a host's MAC address from its IP address. For host A to find the MAC address of a host B on the same subnet with IP Address B-IP, host A broadcasts an ARP query packet containing B as well as its own IP address (A-IP) on its Ethernet interface. All hosts in the same subnet receive the packet. Host B, whose IP address is B-IP, replies (via unicast) to inform A of its MAC address. A will also record the mapping between B-IP and B-MAC.

Even though all hosts maintain the MAC to target IP address mapping locally to avoid repetitive ARP broadcast message for the same target IP address, all hosts age out their learnt MAC to IP mapping very frequently. For Microsoft Windows (versions XP and server 2003), the default ARP cache policy is to discard entries that have not been used in at least two minutes, and for cache entries that are in use, to retransmit an ARP request every 10 minutes. So hosts send out ARP very frequently.

In addition to broadcast messages sent from hosts, Layer 2 switches also broadcast received packet if the destination address is unknown.

All Layer 2 switches learn MAC address of data frames which traverse through the switches. Layer 2 switches also age out their learnt MAC addresses in order to limit the number of entries in their Filtering Database (FDB). When a switch receives packet with an unknown MAC address, it broadcast this packet to all ports which are enabled for the corresponding VLAN.

The flooding and broadcast have worked well in the past when the Layer 2 network is limited to a smaller size. Most Layer 2 networks limit the number of hosts to be less than 200, so that broadcast storm and flooding can be kept in a smaller domain.

5.1. Address Resolution creates significant burden to hosts

When a Layer 2 network has tens of thousands of hosts, the frequent ARP broadcast messages from all those hosts clearly present a significant burden, especially to hosts. Many of today's layer 2 switches, even with hundreds of ports, can forward 10G traffic at line rate. But they don't need to process ARP requests, they just forward them. It is the host who needs to process every ARP message that circulates in the network.

[Scaling Ethernet] of Carnegie Mellon did a study on the number of ARP queries received at a workstation on CMU's School of Computer Science LAN over a 12 hour period on August 9, 2004. At peak, the host received 1150 ARPs per second, and on average, the host received 89 ARPs per second. During the data collection, 2,456 hosts were observed sending ARP queries. [Scaling Ethernet] expects that the amount of ARP traffic will scale linearly with the number of hosts on the LAN. For 1 million hosts, it is expected to have 468,240 ARPs per second or 239 Mbps of ARP traffic to arrive at each host at peak, which is more than enough to overwhelm a standard 100 Mbps LAN connection. Ignoring the link capacity, forcing hosts to handle an extra half million packets per second to inspect each ARP packet would impose a prohibitive computational burden.

To detect address conflict and refresh hosts address in a Layer 2 network, many types of hosts and almost all Virtual Machines send out gratuitous ARP on regular basis. The Gratuitous ARP could mean either gratuitous ARP request or gratuitous ARP reply. Gratuitous in this case means a request/reply that is not normally needed according to the ARP specification ([RFC 826](#)) but could be used in some cases. A gratuitous ARP request is an Address Resolution Protocol request packet where the source and destination IP are both set to the IP of the machine issuing the packet and the destination MAC is the broadcast address ff:ff:ff:ff:ff:ff. Ordinarily, no reply packet will

occur. A gratuitous ARP reply is a reply to which no request has been made.

All the Gratuitous ARP messages also need to be processed by all hosts in the Layer 2 domain.

Handling up to 1000~2000 ARP requests per second is almost the high

limit for any hosts. With more than 20K hosts in one Layer 2 domain, the amount of ARP broadcast messages, plus other broadcast messages such as DHCP, can create too much burden to be handled by hosts.

5.2. Large amount of MAC addresses to be learnt by intermediate switches

Ethernet's non-hierarchical flat layer 2 MAC addressing makes it not possible for any types for address summarization. MAC addresses, plus their VLAN IDs, have to be placed in switch's FDB without any abbreviation, not like IP addresses which only need proper prefix to be stored in router's forwarding table.

One advantage of Ethernet switches is that it can forward much more addresses than its FDB entries. When a data frame's destination address is not present in a switch's FDB entry, the switch just flood this data frame to all ports which are enabled for the corresponding VLAN. That is why Ethernet switches can have a relative small FDB size, which is one of the key reasons that Ethernet switches can be built at much lower cost than routers. To improve efficiency of the FDB, Ethernet switches frequently age out learnt MAC addresses which haven't been in use for a while and always replace older MAC entries with newly learnt MACs when the FDB is full.

When servers in data center are virtualized, each server can host tens or hundreds of virtual machines. Each virtual machine can be a host to an application, which has its own IP address and MAC address. With the same type and number of network equipments, i.e. ToR switches and Aggregation switches, the number of hosts can grow dramatically in this network. When the number of hosts grows, the number of MAC addresses to be learnt by Layer 2 switches grows too. For an example of tree shaped Layer 2 network with one core switch connected to 3 aggregation switches, each aggregation switch connected to 25 ToR switches, and each ToR switch connected to 25 physical servers, if each server supports 50 virtual machines, there will be $50 \times 25 \times 25 \times 3 = 93750$ hosts in this network.

Typical bridges support in the range of 16 to 32K MAC Addresses, with some supporting 64K. With external memory (TCAM), bridges can support up to 512K to 1M MAC addresses. But TCAM is expensive, which will defeat the low cost advantage of Layer 2 switches. This problem is

especially severe for Top of Rack switches, which are supposed to be very low cost.

In summary, the low cost ToR switches usually don't have enough FDB entries for all VM's MAC addresses in the domain.

[5.3.](#) More chances of unknown broadcast

When the number of hosts, MAC addresses, are above the switches FDB size, learnt MAC addresses in the FDB are aged out faster, which will increase the chances of switch's FDB not having the entry for the received packet's destination address, which then causes the packet being flooded.

When the spanning tree topology changes (e.g. when a link fails), a bridge clears its cached station location information because a topology change could lead to a change in the spanning tree, and packets for a given source may arrive on a different port on the bridge. As a result, during periods of network convergence, network capacity drops significantly as the bridges fall back to flooding for all hosts.

[6.](#) Why dividing one Layer 2 into many smaller subnets is not enough?

Subnet (VLAN) can partition one Layer 2 network into many virtual Layer 2 domains. All the broadcast messages are confined within one subnet (VLAN). Subnet (VLAN) has worked well when each server serving one single host. The server will not receive broadcast messages from hosts in other subnets (VLANs).

When one physical server is supporting 100 plus Virtual Machines, i.e. 100 plus hosts, most likely the virtual hosts on one server are on different subnets (VLANs). If there are 50 subnets (VLANs) enabled on the switch port to the server, the server has to handle all the ARP broadcast messages on all 50 subnets (VLANs). When virtual hosts are added or deleted from a server, the switch port to the server may end up enabling more VLANs than the number of subnets actually active on the server. Therefore, the amount of ARP messages to be processed by each server is still too much.

For Cloud Computing Service, the number of virtual hosts and virtual subnets can be very high. It might not be possible to limit the number of virtual hosts in each subnet.

[7.](#) Why IETF needs to develop solutions instead of relying on IEEE802

Here are the reasons that IETF need to develop solutions:

- Client of Cloud Computing services may request redundancy across two geographical locations. They may want two VMs in one Virtual Subnet to be in two locations -> Most likely it is the IP/MPLS networks which interconnect multiple locations
- The two redundant VMs may have same IP address with one being Active and other one being Standby. The Active and Stand-by need to have keep-alive messages between them. The only way to achieve this is via Layer 2 keep-alive because the Active and Stand-by will have different MAC Addresses -> Require the two VMs on same Layer 2
- It is desirable for all hosts (VMs) of one Virtual Subnet to be in one Layer 2 network for efficient multicast and broadcast among them.
- Client may request hosts (VMs) in one Virtual Subnet to be on different locations to have faster response for their applications. -> Require IP/MPLS to interconnect
- Hosts can be added to one Virtual Subnet at different time. It is possible that newly added hosts have to be placed at a different site due to computing & storage resource availability -> Require IP/MPLS to interconnect.

[8.](#) Conclusion and Recommendation

When there are tens of thousands of VMs in one Data Center, we have concluded that:

- It is necessary to restrain the ARP storm and broadcast storm initiated by (unpredictable) servers and applications into a confined domain.
- It is necessary to have a way to restrain Layer 2 switches from learning tens of thousands of MAC addresses.
- It is necessary for reduce the amount of un-known addresses arriving at any Layer 2 switches to prevent large amount of un-known broadcast in one Layer 2.

For Cloud services which offer virtual hosts and virtual subnets, we have concluded that:

Internet-Draft ARP for Large DC Problem Statement

July 2010

- It is necessary to have a more scalable address resolution protocol for Layer 2 Network which spans across multiple locations.
- It is desirable to constrain MAC addresses in each site from being learnt by other sites. This is to allow traditional Layer 2 switches, which have limited amount of address space for forwarding table, to function properly and to minimize unknown broadcast by those switches.

Therefore, we recommend IETF to create a working group:

- To develop scalable address resolution protocols for data center with large amount of hosts and Layer 2 spanning across multiple locations,
- To develop mechanism to scope the broadcast messages, beyond ARP and DHCP, to minimize impact to each layer 2 domain by broadcast storms from other layer 2 domains,
- To have a scalable inter Layer 2 domain protocol, like BGP, for each domain's gateways to exchange hosts' reachability information among each other, and
- To identify mechanisms for proper handling of multicast messages among hosts in one Virtual Subnet which spans across multiple locations.

[9](#). Manageability Considerations

This document does not add additional manageability considerations.

[10](#). Security Considerations

This document has no additional requirement for a change to the security models of MPLS-Ping and MPLS-Ping-Enhanced.

[11](#). IANA Considerations

A future revision of this document will present requests to IANA for codepoint allocation.

12. Acknowledgments

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

Dunbar

Expires January 2, 2011

[Page 12]

Internet-Draft ARP for Large DC Problem Statement

July 2010

13. References

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

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

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

Authors' Addresses

Linda Dunbar
Huawei Technologies
1700 Alma Drive, Suite 500
Plano, TX 75075, USA
Phone: (972) 543 5849
Email: ldunbar@huawei.com

Sue Hares
Huawei Technologies
2330 Central Expressway,
Santa Clara, CA 95050, USA
Phone:
Email: shares@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

Dunbar

Expires January 2, 2011

[Page 13]

Internet-Draft ARP for Large DC Problem Statement

July 2010

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.

Dunbar

Expires January 2, 2011

[Page 14]