ARMD BOF
Internet Draft
Intended status: Informational
Expires: September 2011

L. Dunbar
S. Hares
Huawei
M. Sridharan
N. Venkataramaiah
Microsoft
B. Schliesser
Cisco Systems

March 14, 2011

Address Resolution for Large Data Center Problem Statement
draft-dunbar-armd-problem-statement-01.txt

Status of this Memo

Copyright Notice

Abstract

Modern data center networks face a number of scale challenges. One such challenge for so-called "massive" data center networks is address resolution, such as is provided by ARP and/or ND. This document describes the problem of address resolution in massive data centers. It discusses the network impact of various data center technologies including server virtualization, illustrates reasons why it is still desirable to have multiple hosts on the same Layer 2 data center network, and describes potential address resolution problems this type of Layer 2 network will face.

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.

Table of Contents

1. Introduction

Modern data center networks face a number of scale challenges, especially as they reach sizes and densities that are "massive" relative to historical norms. One such challenge is the effective and efficient performance of address resolution, such as is provided by ARP and/or ND.

The fundamental issue challenging address resolution in massive data centers is the need to grow both the number and density of logical Layer 2 segments while retaining flexibility in the physical location of host attachment. This problem has historically been bounded by physical limits on data center size, as well as practical considerations in the physical placement of server resources. However, the increasing popularity of server virtualization technology (e.g. in support of "cloud" computing), the trend toward building physically massive data center facilities, and the logical extension of network segments across traditional geographic boundaries is driving an increase of the number of addresses in the modern data center network.

1.1. Server Virtualization

Server virtualization allows the sharing of the underlying physical machine (server) resources among multiple virtual machines, each running its own operating system. Server virtualization is the key enabler to data center workload agility, i.e. allowing any server to host any applications 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. However, server virtualization also stresses the data center network by enabling the creation of many more network hosts (accompanied by their network interfaces and addresses) within the same physical footprint.

Further, in order to maximize the benefits of server virtualization, VM placement algorithms (e.g. based on efficiency, capacity, redundancy, security, etc) may be designed in such a way that increases both the range and density of Layer 2 segments. For instance, these algorithms may satisfy the processing requirements of each VM while requiring the minimal number of physical servers and switching devices, simultaneously spreading the VM hosts across a diverse and redundant infrastructure. Such an algorithm may potentially result in a large number of diverse Layer 2 segments

attached to each physical host, as well as a larger number and range of data center-wide Layer 2 segments. With this, and similar types of VM assignment algorithm, subnets tend to extend throughout the network and ARP/ND traffic associated with each subnet is likely to traverse a significant number of links and switches in the network.

1.2. Physically Massive Facilities

Regardless of server virtualization technology, in recent years the physical facility of a data center has been seen to grow larger. There are inherent efficiencies in constructing larger data center buildings, infrastructure, and networks. As data center operators pursue these physical efficiencies, the address resolution problem described by this document becomes more prevalent. Physically massive data centers may face address resolution scale challenges simply due to their physical capacity. Combined with server virtualization, the host and address density of these facilities is historically unmatched.

1.3. Geographically Extended Network Segments

The modern data center network is influenced by the demands of flexibility due to cloud computing, demands of redundancy due to regulatory or enterprise uptime requirements, as well as demands on topology due to security and/or performance. In support of these demands and others, VPN and physical network extensions (including both Layer 3 and Layer 2 extensions) increase the data center network scope beyond physical and/or geographical boundaries.

As such, the number of addresses that are present on a single Layer 2 segment may be greater than the number of hosts physically or logically present within the data center itself. Combined with

physically massive data center facilities and server virtualization, this trend results in a potential for massive numbers of addresses per Layer 2 segment, beyond any historical norm, truly challenging address resolution protocols such as ARP and/or ND.

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.

CUG: Closed User Group

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.

VM: Virtual Machines

VPN: Virtual Private Network

3. Layer 2 Requirements in the Data Center

3.1. Layer 2 Requirement for VM Migration

VM migration refers to moving virtual machines from one physical server to another. Current technology even allows for the real-time migration of VM hosts in a "live" state. Seamlessly moving VMs within a resource pool is the key to achieve efficient server utilization and data center agility.

One of the key requirements for VM migration is the VM maintaining the same IP address and MAC address after moving to the new location, so that its operation can be continued in the new location. This requirement is even more stringent in the case of "live" migrations, for which ongoing stateful connections must be maintained. Thus, in absence of new technology, VMs can only be migrated among servers on the same Layer 2 network.

3.2. Layer 2 Requirement for Network Services

Many network services such as firewalls and load balancers must be in-line with network traffic in order to function correctly. As such, Layer 2 networks often provide a form of traffic engineering for steering traffic through these devices for a given subnet or segment.

Further, even in some cases where the network service need not be in-line for all traffic, it must be connected on a common Layer 2 segment in order to function. One such common application is load

balancing (providing a single Internet service from multiple servers) with Layer 2 Direct Server Return. While a traditional load balancer typically sits in-line between the client and the hosts that provide the services to the client, for applications with relative smaller amount of traffic going into servers and relative large amount of traffic from servers, it is sometimes desirable to allow reply data from servers go directly to clients without going through the Load Balancer. In this kind of 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.

3.3. Layer 2 Requirement for Active/Standby VMs

For redundant servers (or VMs) serving redundant instances of the same applications, both Active and Standby servers (VMs) need to share keep-alive messages between them. Further, the mechanism for failing over from Active to Standby may be facilitated by assumption of a shared MAC address and/or some kind of ARP/ND announcement. When the Active server fails/is taken out of service, the switch over to the Standby would be transparent if they are on the same Layer 2 network.

4. Cloud and Internet Data Centers with Virtualized Servers

Cloud Computing service often allows subscribers to create their own virtual hosts and virtual subnets which are housed within the cloud providers' data centers. Network service providers may also extend existing VPNs to connect with VMs that are hosted by servers in the provider's data center(s). This is often realized by grouping hosts belonging to one subscriber's VPN into distinct segregated subnets in the data center(s). This design for a multi-tenant data center network typically requires the secure segregation of different customers' VMs and hosts.

Further, these client subnets in the data center could have client-specific IP addresses, which could lead to possible overlaps in address spaces. In this scenario, it is very critical to segregate traffic among different client subnets (or VPNs) in data center. As a result, within a cloud data center there may be a larger number of distinct Layer 2 segments as well as a larger demand for host density within each Layer 2 segment.

Cloud/Internet Data Centers have the following special properties:

. Massive number of hosts

Consider 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. The number of servers connected in this network is limited to the port count of the ToR switches. For example, if a ToR switch has 20 downstream ports, there are only 20 servers or hosts connected to it. If the aggregation switch has 256 ports connecting to ToR switches, there could be up to 20*256=5120 hosts connected to one aggregation switch when the servers are not virtualized.

When servers are virtualized, one server can support tens or hundreds of VMs. Hypothetically, if one server supports up to 100 VMs, the same ToR switches and Aggregation switch as above would need to support up to 512000 hosts. Even if there is enough bandwidth on the links to support the traffic volume from all those VMs, other issues associated with Layer 2, like frequent ARP broadcast by hosts and flooding due to unknown DA, create

challenges to the network.

. Massive number of client subnets or Closed User Groups co-existing in the data center, with each subnet having their own IP addresses

In the example of VPN (L2VPN or L3VPN)extended with virtual machines residing in Service Provider data centers, each VPN would require an unique subnet for its associated VMs in the data center. Due to large number of VPNs being deployed today, those types of services can require a large number of subnets to be supported by the data center.

. Hosts (VMs) migrate from one location to another

When data center is virtualized, physical resource and logical hosts/contents are separated. One application could be loaded to any Virtual Machines on any servers, and could be migrated to different locations during the continuous process of minimizing the physical resources consumed in data center(s).

As discussed earlier, this migration requires the VMs to maintain the same IP and MAC addresses. The association to their corresponding subnet (or VPN) should not be changed either.

5. ARP/ND Issues in the Data Center

Traditional Layer 2 networks placed hosts belonging to one subnet (or VLAN) closely together, so that broadcast messages among hosts in the subnet are confined to the access switches. However this kind

of network design puts a lot of constraints on where VMs can be placed and can lead to very unbalanced utilization of data center resources.

In data center with virtualized servers, data center administrators may want to leverage the flexibility of server virtualization to place VMs in such a way that satisfies the processing requirements of each VM but require the minimal number of physical servers and switching devices. When those types of VM placement algorithms are used, hosts can be attached and re-attached at any location on the network. IPv4 hosts use ARP (Address Resolution Protocol-RFC826) to find the corresponding MAC address of a target host. IPv4 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) on its Ethernet interface. All hosts in the same subnet receive the packet. Host B, whose IP address is B , replies (via unicast) to inform A of its MAC address. A will also record the mapping between B 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, 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 flood received data frames if the destination MAC address is unknown.

The flooding and broadcast have worked well in the past when hosts belonging to one subnet (or VLAN) are placed closely together. A common scenario is for Layer 2 networks to limit the number of hosts in one subnet to be less then 200, so that broadcast storms and flooding can be restricted to a smaller domain when all the hosts are confined to small number of ports on access switches. When subnets are spanning across multiple ToR switches or EoR switches, each subnet's broadcast messages and flooding will be exposed to the backbone links and switches of entire Data Center network. Then, the network will experience the similar problems as one big flat Layer 2 network. With large number of hosts in data centers with virtualized servers, the amount of broadcast messages and flooding over the backbone links can take away huge amount of bandwidth.

As indicated in Reference [Scaling Ethernet], 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. The report 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 at peak, which is more than enough to overwhelm a standard 100 Mbps LAN connection. Ignoring the link capacity, forcing servers to handle an extra half million packets per second to inspect each ARP packet would impose a prohibitive computational burden.

6. ARPs & VM Migration

In general, there are more flooding and more ARP messages when VMs migrate.VM migration in Layer 2 environments will require updating the Layer 2 (MAC) FDB in the individual switches in the data center to ensure accurate forwarding. Consider a case where a VM migrates across racks. The migrated VM often sends out a gratuitous ARP broadcast when it comes up at the new location. This is flooded by the TOR switch at the new rack to the entire network. The TOR at the old rack is not aware of the migration until it receives this gratuitous ARP. So it continues to forward frames to the port where it learnt the VM's MAC address from before, leading to black holing of traffic. The duration of this black holing period may depend upon the topology. It may be longer if the VM has moved to a rack in a different data center connected to this data center over Layer 2.

During transition periods, some hosts might be temporarily taken out of service. Then, there will be lots of ARP request broadcast messages repetitively transmitted from hosts to those temporarily out of service hosts. Since there is no response from those target hosts, switches do not learn their path, which will cause ARP messages from various hosts being flooded across the network.

Simple VLAN partitioning is no longer enough to segregate traffic among tens of thousands of subnets (or Closed User Groups) within a data center. Some types of encapsulation have to be used, like MAC-in-MAC or TRILL, to further isolate the traffic belonging to different subnets. When encapsulation is performed by TOR, VMs migration can cause more broadcast messages and more data frames being flooded in the network due to new TOR not knowing the destination address of the outer header of the encapsulation.

Therefore, it is very critical to have some types of ARP optimization or extended ARP reply for TOR switches which perform the encapsulation. This can involve knowledge of the target TOR address,

so that the amount of flooding among TOR switches due to unknown destination can be dramatically reduced.

7. Limitations of VLANs/Smaller Subnets in the Cloud Data Center

Large data centers might need to support more subnets or VLANs than 4095. So, simple VLAN partitioning is no longer enough to segregate traffic among all those subnets. To enforce traffic segregation among all those subnets, some types of encapsulation have to be implemented.

As the result of continuous VM migration, hosts in one subnet (VLAN) may start with being close together and gradually being relocated to various places.

When one physical server is supporting more than 100 Virtual Machines, i.e. >100 hosts, it may start with serving hosts belonging to smaller number of VLANs. But gradually, as VM migration proceeds, hosts belonging to different VLANs may end up being loaded to VMs on this server. Consider a case when 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). The amount of ARP to be processed by each server is still too much.

8. Why IETF Needs To Develop Solutions Instead of IEEE 802

ARP involves IP to MAC mapping, which traditionally has been standardized by IETF, e.g. RFC826.

9. Conclusion and Recommendation

When there are tens of thousands of VMs in one Data Center or multiple data centers interconnected by a common Layer 2 network, Address Resolution has to be enhanced to support large scale data center and service agility

Therefore, we recommend that the IETF engage in the study of this address resolution scale problem and, if appropriate, the development of interoperable solutions for address resolution in massive data center networks.

10. Manageability Considerations

This document does not add additional manageability considerations.

11. Security Considerations

This document discusses a number of topics with their own security concerns, such as address resolution mechanisms including ARP and/or ND as well as multi-tenant data center networks, but creates no additional requirement for security.

12. IANA Considerations

This document creates no additional IANA considerations.

13. Acknowledgments

Many thanks to T. Sridhar for his contributions to the text.

14. References

[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.mspx, 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

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

Murari Sridharan
Microsoft Corporation
muraris@microsoft.com

Narasimhan Venkataramaiah
Microsoft Corporation
One Microsoft Way
Redmond, WA 98052-6399 USA
Phone : 425-707-4328
Email : narave@microsoft.com

Benson Schliesser
Cisco Systems, Inc.
Phone:
Email: bschlies@cisco.com

Intellectual Property Statement

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.