ARMD BOF
Internet Draft
Intended status: Standard Track
Expires: April 2011

L. Dunbar
S. Hares
Huawei
Murari Sridharan
Narasimhan Venkataramaiah
Microsoft
T Sridhar
Force 10
October 18, 2010

Address Resolution for Large Data Center Problem Statement
draft-dunbar-armd-problem-statement-00.txt

Status of this Memo

Copyright Notice

Abstract

Server virtualization enables one physical server to support multiple virtual machines (VMs) so that multiple virtual hosts (20, 30, or hundreds of) can be running on one physical server. As virtual machines are introduced to the data center, the number of hosts within one data center can grow dramatically, resulting in significant impact on the network.

This document describes reasons why it is still desirable to have virtual machines in the data center to be in one Layer 2 network and potential problems this type of Layer 2 network will face. The goal is to outline the problem area for the IETF to create a working group. This working group will 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

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 agility, i.e. allowing any server to host any application and providing the flexibility of adding, shrinking, or moving services within the physical infrastructure. Server virtualization is also the key element for Cloud Computing services, such as Amazon's EC2 service, and virtual desktop services, which allow servers in data center(s) to provide virtual desktops to millions of end users.

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. As a result, many organizations are highly motivated to incorporate server virtualization technologies into their data centers.

While server virtualization is an enabler for flexible management of server resources, it does impose significant challenges to networks which interconnect all the servers in data center(s).

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 (hosts). 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 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 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, unknown flooding, create challenges for the network.

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

EOR: End of Row switches in data center.

FDB: Filtering Database for Bridge or Layer 2 switch

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. 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. Thus, VMs can only be migrated among servers on the same Layer 2 network.

3.2. Layer 2 Requirement for Load Balancing

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. The 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 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, 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 same applications, both Active and Standby servers (VMs) need to have keep-alive messages between them. 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, like Amazon's Elastic Compute Cloud (Amazon EC2) and Virtual Private Cloud (Amazon VPC), allows users (clients) to create their own virtual hosts and virtual subnets which are housed by VMs in the cloud providers' data center.

Telecom service providers may also extend their existing VPNs to accommodate client VMs that the service provider hosts on its on physical servers. This could be realized by client "subnets" in the data center.

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.

Cloud/Internet Data Centers have the following special properties:

Massive number of hosts

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 Private Cloud VPN (L2VPN or L3VPN) with virtual hosts residing in Service Provider data centers, each VPN could also include PEs (Provider Edge switch/router) at traditional Customer Locations.

Hosts (VMs) migrate from one location to another

Physical resource and logical hosts/contents are separated, i.e. one user's application could be loaded to any Virtual Machines on any servers, and could be migrated to different locations for efficient server and storage management.

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 change either.

5. ARP Issues in the Data Center

In a Layer 2 network, 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 IP well as its own IP address (A) on its Ethernet interface. All hosts IP 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 IP also record the mapping between B and B-MAC. IP

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. All Layer 2 switches learn the source 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 a packet with an unknown MAC address, it floods 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. A common scenario is for Layer 2 networks to limit the number of hosts to be less then 200, so that broadcast storms and flooding can be restricted to a smaller domain.

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.

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

Cloud 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 to form a large Layer 2 network, Address Resolution process, has to be enhanced to support large scale data center and service agility

Therefore, we recommend IETF to create a working group to develop interoperable solutions for Address Resolution for Massive amount of hosts in Data Center (ARMD).

10. Manageability Considerations

This document does not add additional manageability considerations.

11. Security Considerations

This document has no additional requirement for security.

12. IANA Considerations

13. Acknowledgments

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

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

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

T Sridhar
Force 10 Networks
350 Holger Way,
San Jose, CA 95134, USA
Phone:
Email: tsridhar@force10networks.com

Intellectual Property Statement

Disclaimer of Validity

Acknowledgment

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