Address Resolution Scalability for VPN oriented Data Center Services

Linda Dunbar
Huawei Technologies, Plano, TX. USA
ldunbar@huawei.com

Abstract— Modern data centers tend to have very large number hosts due to business demand and technology advancement like server virtualization. Server virtualization has made it convenient to dynamically create VMs when the application requires more resources, and move VMs, either from overloaded servers, or to aggregate VMs onto fewer servers to save power when demand is light. This flexibility has made variety of virtual data center services possible. This paper introduces the address resolution scaling issues in modern data centers; especially the address resolution issues for VPN oriented Data Center services. The paper also describes some alternative solutions which can make the network scale.

Keywords-component; virtual machine (VM), Virtual Private Network (VPN).

I. 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. 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.

A. Server Virtualization

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.

B. 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 paper 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.

C. 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.

978-0-9836283-2-3 ⓒ 2011 ITC

This paper was peer reviewed by subject matter experts for publication in the Proceedings of DC–CaVES 2011

II. Large and flat Layer 2 Network in Data Centers

There is variety of network designs for data centers. Some data centers push Layer 3 all the way to access switches, i.e. IP to Top of Rack switches. While IP network has been proven to scale well, this kind of design has the following issues in data centers:

- When a VM needs to be moved from one Rack to another Rack, the boundary routers subnets have to be re-configured if the VM needs to maintain the same IP/MAC address. This makes live VM migration almost impossible. Maintaining same IP address is very often required in data centers.
- The number of hosts under one Layer 2 load balancer is limited to the number servers (&VMs) on one rack. When demand increases and rack has reached full capacity, it is difficult to add more hosts under the same load balancer. When demand decreases and need to consolidate hosts to smaller number of racks, all the hosts under one Layer 2 load balancer have to move together. Usually reconfiguration is needed.

Layer 3 all the way to ToR is easier in homogeneous data centers where applications running on servers/VMs are under the control of the data center operators.

Many data center operators, especially multi-tenant data centers, opt for pushing the L2/L3 boundary higher, i.e. having a bigger and flatter Layer 2 network to avoid the inconvenience and massive configuration issues associated with Layer 3 to ToR switches. The advantage of big and flat layer 2 is that VMs can move freely in the flat Layer 2 network without any router re-configuration. For multi-tenant data centers, some tenants (customers) may even require their Virtual Machines, placed in one or multiple sites, to be on one Layer 2. If the data center network has IP all the way to ToR switches, a lot of configurations (like mesh PWs for VPLS) would be required to satisfy those client requirements.

A. ARP/ND Issues for boundary routers in Data Centers

As described earlier, one major reason for large Layer 2 is to have the flexibility in moving VMs in wider space without re-configuration on routers or middleware boxes (e.g. Layer 2 load balancer). Large Layer 2 network can be under one big subnet (e.g. /16, /8, etc) or many smaller subnets. For security reasons or customer demand, one subnet (large or small) may be mapped to multiple VLANs.

When hosts need to communicate with hosts in different VLANs or outside the Data Center, they need to send ARP/ND to get the L2/L3 boundary router's MAC addresses. Routers are built with very powerful forwarding capability, but all the ARP/ND requests have to be responded by router's CPU which has limited speed. Most powerful routers built today can process up to 2000 ARPs per second. Statistics done by Merit Network [http://tools.ietf.org/html/draft-karir-armd-statistics-01] have shown that even 500 nodes in one Layer 2 can push routers to reach up to 50% CPU utilization in processing ARPs.

In addition, most routers have limited memory space to keep the ARP/ND entries for hosts attached. Some modern data centers are pushing the limit of ~100K hosts under L2/L3 boundary router or even higher. It is tremendous burden for routers.

To make things worse, many applications in modern data centers, especially in multi-tenant data centers, have both IPv4 and IPv6 stack on. That means routers have to process both ARP broadcast and ND multicast messages.

B. Wide spread broadcast and flooding in Data Centers

Traditional Layer 2 networks place hosts belonging to one VLAN closely together, so that broadcast messages among hosts in the VLAN are confined to a few ports on access switches.

Server virtualization has made it convenient to dynamically create VMs when the application requires more resources, and move VMs, either from overloaded servers, or to aggregate VMs onto fewer servers to save power when demand is light. This may lead to hosts belonging to same VLAN (or subnet) being placed under different locations (racks or rows). When hosts belonging to one VLAN are placed in multiple places and one Rack has hosts belonging to multiple VLANs, switches have to enable all those VLANs on many ports, even on many aggregation switches' ports. Under this kind of configuration, all broadcast messages and unknown DA flooding within one VLAN will traverse many backbone links and switches.

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. Some Linux based applications have shorter timeout values for ARP/ND entries.

During transition periods, some hosts might be temporarily taken out of service. Then, there will be lots of ARP/ND request broadcast/multicast 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/ND messages from various hosts being flooded across the network.

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.

C. VLANs exhaustion in Data Centers

Many small & medium size multi-tenant data centers needs thousands of VLANs because each client may require 5~10 VLANs for various reasons (e.g. security or traffic segregation, etc). Large Data Centers will need more than 4095 VLANs. So, simple VLAN partitioning is no longer enough to segregate traffic among all those subnets. To enforce traffic segregation for different clients, some types of encapsulation have to be implemented, such as TRILL or IEEE802.1aq (SPB). IEEE802.1aq and TRILL can make VLANs locally significant, which greatly increase the number of segregated segments in the large Layer 2 network.

D. Solutions and Optimization

Data center network topology is based on racks, rows. Hosts assignment to Servers, Racks, and Rows is orchestrated by Server/VM Management system, not random. That means switches can easily get information on where target hosts are attached from directory either embedded or acquirable externally. For some data centers using TRILL or IEEE802.1aq (SPB), the directory assisted mapping can eliminate flooding across the backbone. [TRILL-Directory] describes an optimization method for TRILL. A similar approach can also be used for IEEE802.1aq SPB.

The solution described in [TRILL-Directory] is especially useful in virtualized data center environment where VMs migrate all the time. If migrated VMs send out gratuitous ARP (IPv4) or Unsolicited Neighbor Advertisement (IPv6) from the new location, those gratuitous broadcast messages have to flood to backbone and all edge switches (TRILL edge or IEEE802.1aq Edge) with the corresponding VLANs enabled. If the migrated VMs don't send out gratuitous ARP (or ND) from the new location, packets towards those migrated VMs will be sent to the wrong egress TRILL or IEEE802.1aq edge nodes, which not only wastes bandwidth but also causes those data packets being black holed.

The benefits of using directory assistance include:

- The Directory enforced MAC&VLAN <-> Edge mapping table can determine if a frame needs to be forwarded across IEEE802.1aq or TRILL domain.

  When multiple RBridge edge ports are accessible from a server (hosts/VMs), a directory assisted RBridge edge won't flood frames with an unknown DA to all to other RBridge ports. Therefore, there is no need to designate an Appointed Forwarder among all the RBridge Edge ports connected to a Bridge LAN, which enables all RBridge ingress ports to forward traffic.

- Directory assisted approach can not only eliminate the flooding within IEEE802.1aq or RBridge domain (unknown learning), but also reduce the flooding on the bridged LAN attached to edge ports.
- Reduce the amount of MAC&VLAN <-> Edge mapping maintained by edge switches. No need for an edge switch to keep the MAC entries for hosts which don't communicate with hosts attached to the edge.

There can be two different models for IEEE802.1aq Edge or TRILL edge switches in data center to be assisted by Directory services:

- Push Model:

  Directory Server(s) push down the host-address (MAC&VLAN) <-> Edge mapping for all the hosts which might communicate with hosts attached to the edge switch.

  There are multiple ways to narrow down the smallest set of remote hosts which communicate with hosts attached to an edge switch. A very simple approach: for VLAN #i enabled on one of edge switch port(s), MAC/IP entries for hosts in VLAN #i will not be pushed down to the switch if there are no hosts belonging to VLAN #i attached to the switch.

  Whenever there is any change in Host-Address <-> Edge mapping, which can be triggered by hosts being moved, de-commissioned, or temporarily out of service due to maintenance, an incremental update can be sent to the switches which are impacted by the change.

  Under this model, switches can simply drop a data frame (instead of flooding) if the destination address can't be found in the host-address <-> Edge mapping table.

- Pull model:

  Under this model, an edge switch (RBridge Edge or IEEE802.1aq Edge) can simply intercept all ARP/ND requests and forward them to the Directory Server(s) which has the information of how each MAC&VLAN is mapped to its corresponding edge switch(es).

  The reply from the Directory Server can be the standard ARP/ND reply with an extra field showing the edge switch address. The switch can cache the mapping.

  If a switch receives an unknown MAC-DA, it could drop the data frame as in the Push Model, or it can query the directory server. If there is no response from the directory server, the switch can drop the frame.

III. VPN ORIENTED DATA CENTER SERVICES AND ITS ADDRESS RESOLUTION ISSUES

VPN-oriented Data Center Services [VPN-o-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.

A. 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.

B. 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 paper only focuses on networking (switching and routing) related components within VDCS framework.

C. Networking related components in support of VDCS

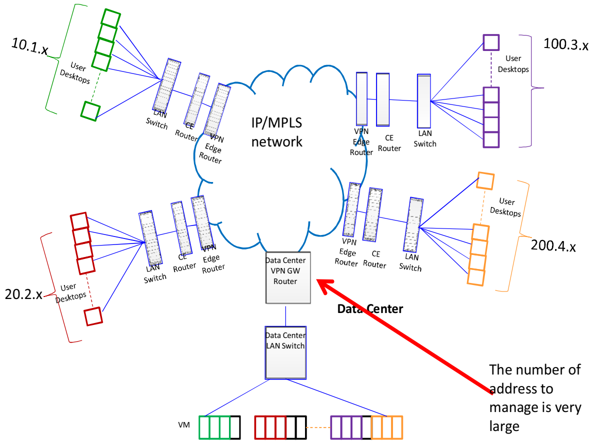
In Figure-1, Vx represents a VM or a server belonging to VPN-x. The data center depicted in the figure has VMs belonging to 5 different VPNs, VPN-1, VPN-2, VPN-3, VPN-4, and VPN-5. Most data centers have many rows of server racks. Each rack holds many servers and has 1 or 2 Top of Rack (ToR) switches. 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.

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 within Data Center for VDCS

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.

There are two basic address resolution problems associated with VDCS: scalability and address conflict. Figure-2 shows that the amount of addresses to be maintained by the gateway router is huge.



The number of address to manage is very large

Figure 2: VDCS Global View

D. *Address Resolution for VMs attached to L2VPN*

Before severs 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 leads 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.

E. *Address Resolution for VMs attached to L3VPN*

When severs 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.

IV. Conclusion

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 to maintain and process all the ARP/ND entries and all the VSIs or VRFs needed for the huge number of hosts and VPNs.

IETF ARMD WG has been created to investigate address resolution scaling issues in data centers and alternative solutions to make data center gateway routers capable of supporting business demands.

Acknowledgment

Many thanks to Ning So (Verizon), Benson Schliesser (Cisco), Radia Perlman (Intel), Igor Gashinsky (Yahoo), Huawei's Susan Hares, Peter Ashwood Smith, Donald Eastlake, Ben MackCrane, and Robert Sultan in shaping the problem domain and solutions.

References

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

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

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

[4] Dunbar, et. al. "Directory Assisted RBridge edge", draft-dunbar-trill-directory-assisted-edge-01.txt, July 2011.

[5] So, et. al. "Requirement and Framework for VPN-Oriented Data Center Services", draft-so-vdcs-01.txt, July 2011.

[6] So, et. al. "Address Resolution Requirements for VPN-oriented Data Center Services" draft-so-armd-vdcs-ar-00, July 2011.

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

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