

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

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 September 14, 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

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.....	3
2. Terminology.....	4
3. Layer 2 Requirements in the Data Center.....	5
3.1. Layer 2 Requirement for VM Migration.....	5
3.2. Layer 2 Requirement for Load Balancing.....	5
3.3. Layer 2 Requirement for Active/Standby VMs.....	6
4. Cloud and Internet Data Centers with Virtualized Servers.....	6
5. ARP Issues in the Data Center.....	7
6. ARPs & VM Migration.....	9
7. Limitations of VLANs/Smaller Subnets in the Cloud Data Center.....	10
8. Why IETF Needs To Develop Solutions Instead of IEEE 802.....	10
9. Conclusion and Recommendation.....	10
10. Manageability Considerations.....	11
11. Security Considerations.....	11
12. IANA Considerations.....	11
13. Acknowledgments.....	11
14. References.....	11
Authors' Addresses.....	12
Intellectual Property Statement.....	12
Disclaimer of Validity.....	13

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 $20 \times 256 = 5120$ 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 than 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/windowsserver
2003/technologies/networking/tcpip03.mspx](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

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.

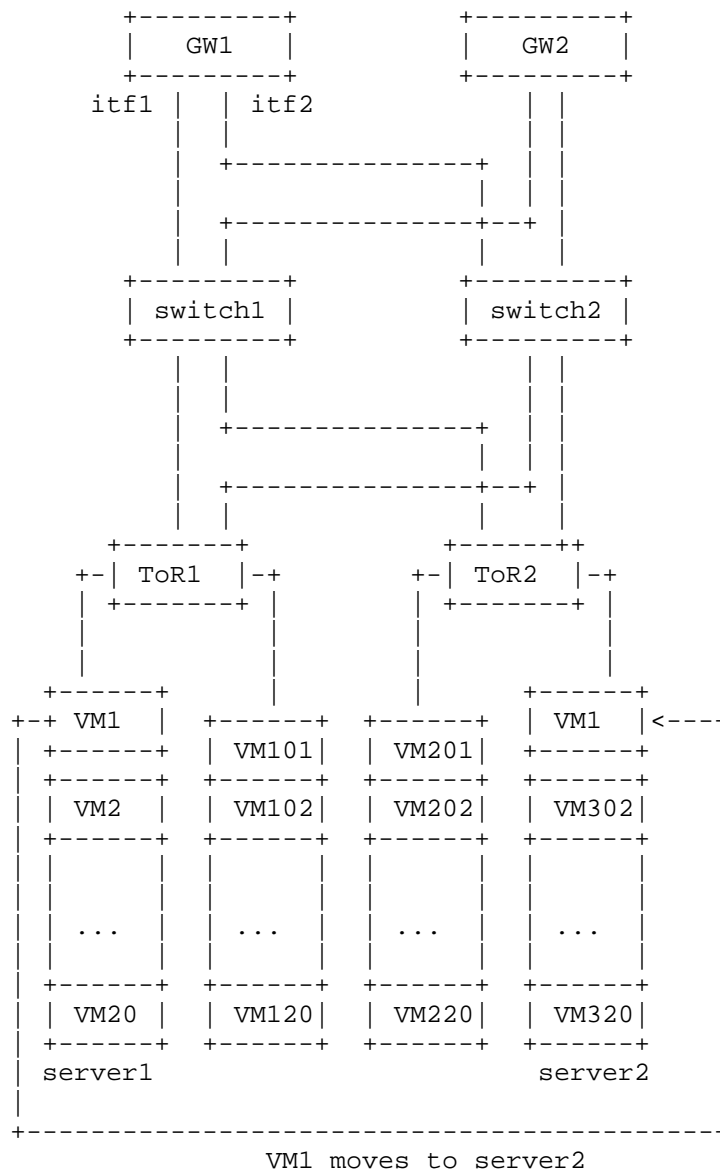


Figure 1 VM migration scenario

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 use 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

MPLS Working Group
Internet Draft
Intended status: Informational
Expires: April 2011

B. Mack-Crane
L. Dunbar
S. Hares
Huawei

October 12, 2010

IPv6 Neighbor Discovery Scalability for Large Data Centers
draft-mackcrane-armd-ipv6-nd-scaling-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 April 12, 2011.

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.

Abstract

Server virtualization allows one physical server to support many virtual machines (VMs) so that multiple hosts (20, 30, or hundreds) can be running from one physical platform. As virtual machines are introduced into a Data Center, the number of hosts within the data center can grow dramatically, which can have tremendous impact on the network and hosts.

This document provides an analysis of the scalability of IPv6 Neighbor Discovery (RFC 4861) in data centers with a 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. Network functions provided by IPv6 Neighbor Discovery.....	3
3. Basic ND protocol message use.....	4
3.1. Router Solicitation.....	4
3.2. Router Advertisement.....	4
3.3. Neighbor Solicitation.....	5
3.4. Neighbor Advertisement.....	5
3.5. Redirect.....	5
4. Some additional protocol activities.....	5
4.1. Duplicate Address Detection.....	5
4.2. Anycast and Proxy address resolution.....	6
4.3. Neighbor unreachability detection.....	6
4.4. Host-based Load Spreading.....	6
4.5. Router-based Load Spreading.....	6
4.6. Holding packets while address resolution occurs.....	7
5. Summary and conclusions.....	7
6. Manageability Considerations.....	7
7. Security Considerations.....	7
8. IANA Considerations.....	8
9. Acknowledgments.....	8
10. References.....	8
Authors' Addresses.....	8
Intellectual Property Statement.....	9
Disclaimer of Validity.....	9

1. Introduction

Server virtualization allows the sharing of the underlying physical machine (server) resources among multiple virtual machines (VMs), each running its own operating system. While Server 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). Large data centers may grow to support hundreds of thousands or even millions of hosts (VMs). Even though there may be enough link bandwidth to support the traffic volume from all those VMs, other issues associated with Layer 2, like frequent ARP broadcast by hosts, broadcast unknown, etc., can create problems for the network and hosts.

This document presents an initial analysis of the scalability of IPv6 Neighbor Discovery (RFC 4861) protocols in the context of a large data center network. Two network cases are considered: 1) a single L2 VLAN connecting a very large number of hosts and a relatively small number of routers, and 2) a core VLAN connecting a large number of routers and few, if any, hosts. The analysis presented here is a rough assessment of which protocol behaviors should scale well and which may present some concern. It does not provide hard numbers and is not based on any measurements in live networks.

2. Network functions provided by IPv6 Neighbor Discovery

The protocols described in RFC 4861 provide a variety of network functions used by IPv6 nodes to:

- find routers and discover link and network parameters,
- discover each other's presence,
- determine each other's link-layer addresses, and
- maintain information about the paths to active neighbors.

These functions are accomplished using five ICMP messages:

- Router Solicitation,
- Router Advertisement,
- Neighbor Solicitation,

- Neighbor Advertisement, and
- Redirect.

The first part of the analysis considers the basic ND protocol activities and how often each message is sent and to what L2 destination address to determine whether there is any concern that ND messages could take too much bandwidth or tax host processors with unnecessary work.

The second part of the analysis considers whether there may be scalability concerns related to other protocol behaviors mentioned in RFC 4861 for ancillary purposes, for example duplicate address detection.

3. Basic ND protocol message use

3.1. Router Solicitation

The Router Solicitation message is sent by nodes to discover routers on the LAN, effectively requesting routers to respond to the node with a Router Advertisement message. This message is sent to the all-routers multicast address and so is not seen by other hosts on the LAN.

A Router Solicitation message is generally sent when a node is first attached to (or comes up on) the LAN. The frequency of these events should be low and so both the traffic and processing load for Router Solicitation messages is expected to be negligible.

3.2. Router Advertisement

Router Advertisement messages are sent by routers periodically to the all-nodes multicast address to announce their presence on the LAN and advertise some link parameters. As long as there are not very many routers on the LAN this should not present much traffic or processing load. In the core case where the LAN is connecting many routers the traffic and processing load will increase with the number of routers and some measures may be needed to limit the traffic, either by reducing the transmission rate or disabling the protocol (if it is not needed in an all-router environment).

Router Advertisement is also unicast to the requesting node in response to a Router Solicitation message and, as noted above, this should not present a significant load.

3.3. Neighbor Solicitation

A Neighbor Solicitation message is sent by a node when that node has no (or a stale) cache entry for the L2 address for a particular next hop IPv6 address. This message is sent to a solicited-node multicast address which is manufactured from the next hop IPv6 address. A great advantage of using a solicited-node multicast address is that only the solicited neighbor node (or perhaps a very few more) will be subscribed to this address. Therefore the processing load for this message is restricted to a small number of nodes and is not likely to present a significant burden.

In general the frequency of Neighbor Solicitation messages will be related to the number of each node's communicating peers on the LAN. Since this number is directly related to the amount of traffic the LAN must support for communications in general the fraction consumed by Neighbor Solicitation should be very small.

3.4. Neighbor Advertisement

Neighbor Advertisement messages are sent in response to Neighbor Solicitation messages. They are unicast to the originator of the Neighbor Solicitation message and so the load presented in this case should, as with Neighbor Solicitation, be a small fraction of the traffic that must be supported on the LAN.

Unsolicited Neighbor Advertisement messages may also be sent to the all-nodes multicast address; however, as this may be done when a node's L2 address changes the frequency of these messages should be extremely low.

3.5. Redirect

Redirect messages are sent by routers to nodes to change the next hop that node is using to reach a particular destination. Although the likelihood of redirect depends on the network topology and other factors, it is not expected to present a significant load on either the network or hosts.

4. Some additional protocol activities

4.1. Duplicate Address Detection

Duplicate address detection as described in [ADDRCONF] involves sending a number of Neighbor Solicitation messages for the address to be checked (to that address's solicited-node multicast address). This is done before attempting to join the LAN using the address

being checked. Since this is an initialization procedure it is not expected to present a significant traffic or processing load during normal operation. It is also possible that address autoconfiguration will not be used in very large data centers.

4.2. Anycast and Proxy address resolution

Address resolution for Anycast addresses or addresses for which nodes are acting as a Proxy may solicit multiple Neighbor Advertisement messages in response. In this case of Anycast addresses the responses are sent with random delay so that the requesting node does not see an unmanageable burst of responses. The response traffic in this case may be greater but not likely a problem, and the additional processing load is only on the requesting node (which is in control of the rate of solicitation).

In a multi-site data center network it may be desirable to restrict the propagation of Anycast address resolution messages if it is desired that only responses local to the requesting node's site be delivered.

4.3. Neighbor unreachability detection

Neighbor unreachability detection relies on hints from higher layers to determine whether or not a given neighbor is still reachable. In some cases when connectivity is suspect and no higher layer hints are available, a Neighbor Solicitation message may be used to verify continued connectivity. This is not expected to be a common occurrence between hosts or hosts and routers (since higher layer hints are most likely available). Between routers there may not be higher layer hints available but there are likely other means to detect connectivity to router peers across the LAN making use of Neighbor Solicitation messages unnecessary.

4.4. Host-based Load Spreading

Host-based load spreading (e.g. RFC 4311) affects the selection of next hop router for particular packets. This may increase the number of routers a given host communicates with, but it is not expected to add significantly to neighbor discovery traffic or processing load.

4.5. Router-based Load Spreading

Router-based load spreading (i.e. the use of a NULL SA in a Router Advertisement message) requires hosts to solicit a next hop router address. This increases the number of solicitations for router addresses, but this should not be significant if the number of

routers on the LAN is small. This mechanism may be inappropriate (and unneeded) in a core LAN interconnecting a large number of routers and therefore not a concern in that case either.

4.6. Holding packets while address resolution occurs

In multi-site networks or virtualized networks in which the edge-to-edge delay may be increased over that in a normal (local) LAN, hold time for packets awaiting address resolution may increase significantly. This may be a concern depending on the percentage of packets that must wait for address resolution before being forwarded on the LAN.

5. Summary and conclusions

The following summarizes the analysis presented:

- IPv6 ND looks like it will scale well for the case of a large LAN with 1000s of hosts and a relatively small number of routers.
- For the case of a core LAN connecting a large number of routers there are some ND protocol behaviors that may not scale well but these are either optional or not needed between routers (i.e., there are other mechanisms available to the routers to accomplish the same end).
- Multi-site L2 networks may provide challenges for both holding time for packets while address resolution is carried out and address resolution for Anycast addresses (for example, if these are expected to select only local servers).
- The impact of network virtualization (many VLANs and virtual routers) on platforms that support many virtual networks has not been analyzed and may present additional scaling challenges.

6. Manageability Considerations

This document has no manageability considerations.

7. Security Considerations

This document adds no security considerations since it does not define any new protocol behaviors. However, it may be worthwhile to consider whether or not the size of an L2 network (as discussed here) presents any new security challenges. No analysis in this area is provided in this draft.

8. IANA Considerations

This document has no IANA considerations.

9. Acknowledgments

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

10. References

- [RFC4861] Narten, T., Nordmark, E., Simpson, W., and H. Soliman, "Neighbor Discovery for IP version 6 (IPv6)", RFC 4861, September 2007.
- [RFC4862] Thomson, S., Narten, T., and T. Jinmei, "IPv6 Stateless Address Autoconfiguration", RFC 4862, September 2007.

Authors' Addresses

Ben Mackcrane
Huawei Technologies
1700 Alma Drive, Suite 500
Plano, TX 75075, USA
Phone: (630) 810 1132
Email: tmackcrane@huawei.com

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 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.

Working Group: ARMD
Intended Status: Informational
Internet Draft

Himanshu Shah
Ciena Corp

Anoop Ghanwani
Brocade

Expiration Date: April 27, 2012

Nabil Bitar
Verizon

October 28, 2011

ARP Broadcast Reduction for Large Data Centers
draft-shah-armd-arp-reduction-02.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/lid-abstracts.html>

The list of Internet-Draft Shadow Directories can be accessed at
<http://www.ietf.org/shadow.html>

This Internet-Draft will expire on April 27, 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.

Abstract

With advent of server virtualization technologies, a host is able to support multiple Virtual Machines (VMs) in a single physical machine. Data Centers can leverage these capabilities to instantiate on the order of 10s to 100s of VMs in a single server with current technology. It is conceivable that this number can be much higher in the future. Each VM operates as an independent IP host with a set of Virtual Network Interface Cards (vNICs), each having its own MAC address and mapping to a physical Ethernet interface. These physical servers are typically installed in a rack with their Ethernet interfaces connected to a top-of-the-rack (ToR) switch. The ToR switches are interconnected through End-of-the-Row (EoR) or aggregation switches which are in turn connected to core switches.

As discussed in [ARP-Problem] the host VMs use ARP broadcasts to find other host VMs and use periodic (broadcast) Gratuitous ARPs to refresh their IP to MAC address binding in other VM hosts. Such broadcasts in a large data center with potentially thousands of VM hosts in a Layer 2 based topology can overwhelm the network.

This memo proposes mechanisms to reduce the number of broadcasts that are sent throughout the network. This is done by having the ToRs intelligently process ARP and frames, rather than simply broadcasting them throughout the broadcast domain.

While this document addresses ARP, the Neighbor Discovery mechanisms used by the IPv6 hosts that make use of multicast rather than broadcast also pose similar issues in the Data Center. The solutions defined herein should be equally applicable to hosts running IPv6. The details will be specified in a subsequent revision.

Conventions

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 [RFC 2119].

Table of Contents

Copyright Notice	1
Abstract	2
1.0 Overview	3
1.1 Terminology	5
2.0 Configuration	6
3.0 Building the ARP tables	6
3.1 ARP Requests	6
3.2 ARP Reply	7
3.3 Gratuitous ARP	7
3.4 Host movement	8
4.0 Conclusion	9
5.0 Security Considerations	10
6.0 Acknowledgments	10
7.0 References	10
7.1 Normative References.....	10
7.2 Informative References	10
8.0 Author's Address	11

1.0 Overview

The traditional topology in a data center consists of racks of servers connected to top-of-rack (ToR) switches, which connect to aggregation switches, which in turn connect to core switches. The network architecture typically combines Layer 2 and Layer 3. In some architectures, Layer 2 is terminated at the ToR, with Layer 3 being run in the aggregation and core devices. In other architectures, Layer 2 may be extended all the way to the aggregation switch. The primary concerns that have influenced network architectures in the data center have been keeping broadcast domains manageable and spanning tree domains contained.

Moving forward, these traditional network architectures are being challenged due to emerging technologies such as server virtualization.

The effect of server virtualization in the data center brings some challenges. Because of virtualization, the number of hosts that the network sees increases dramatically - 10 to 100 times the number of physical servers. These virtual hosts are referred to as Virtual machines (VMs). VMs offer server mobility wherein a VM can be relocated to run on a different physical server. In order for the mobility to be non-disruptive to other hosts that have communication in progress with the VM being moved, the VM must retain its MAC address and IP address. Because of the requirement to retain the MAC and IP address, it is desirable to develop network architectures that would offer the least restrictions in terms of server mobility.

As an example, in a network architecture where TOR switches terminate the L2 domain, the range of mobility would be restricted to a single ToR switch. It would be more preferable to allow the flexibility of moving the VM anywhere within the data center, or perhaps even a different data center.

Technologies such as TRILL [TRILL] overcome some of the issues of spanning trees because which traditional Layer 2 topologies have been constrained. However, because of virtualization there are 2 specific problems that are introduced with respect to broadcast traffic.

1. A larger number of hosts. A single physical server now hosts multiple virtual machines taking the scale factor to a different level. If each VM has the same number of broadcasts as a physical server, the amount of broadcast traffic has increased 10 to greater than 100 times.
2. If the Layer 2 domains are extended to go across data centers, then broadcast traffic will now go across the backbone. If Layer 2 was terminated at the ToR switch, the increase in broadcast traffic would be been restricted to a single ToR switch, but as discussed earlier, this restriction is not desirable.

The broadcast as such in Layer 2 networks has far reaching impacts; i.e. wastage in network bandwidth as well as CPU resources used by all the VMs while processing superfluous ARP broadcasts (IPv6 gets rid of the latter by running ND as a multicast service rather than a broadcast service).

The solution presented here attempts to minimize negative effects of ARP broadcasts. The solution requires the first hop Ethernet switches, typically ToR, to maintain an ARP table learned from the ARP PDUs received by the switch and selectively propagates the ARP to, or proxy-responds on behalf of, the remote peer. These types of ARP processing principles are well known and used/described in L2VPN Working Group documents such as [ARP-Mediation] and [IPLS]. The ARP proxy response differs from that described in [RFC1027] as the ARP response contains MAC address of the destination and not that of the switch as is suggested in [RFC 1027].

The following sections describe the details of ARP snooping, learning and maintaining ARP tables, using the learned information to limit broadcast propagation and proxy (the response) on behalf of the remote peers.

1.1 Terminology

ToR switch	Top-of-Rack switch. An Ethernet switch installed at the top of a rack of servers which provides network connectivity to those servers.
Downlink	The Ethernet link between the ToR switch and a directly connected host/server in the rack.
Uplink	The network-facing Ethernet connection in the ToR switch. Typically, the uplinks from ToRs connect to end-of-row or aggregation switches.
EoR switch	End-of-Row switch. An Ethernet switch which aggregates traffic from multiple racks. Also commonly referred to as an aggregation switch. Uplinks from the ToR connects to EoR switches and uplinks from EoR switches in turn connect to core switches.
Host/Server	A host or server running the IP protocol. This could be a physical entity or a logical entity (such as a Virtual Machine) in a physical host. The term server refers to its role in data center. Both terms are used interchangeably and refer to an IP end station.
Local hosts	Used in the context of a ToR switch to denote the VM hosts connected to a ToR switch on the downlink, i.e. directly connected hosts.
Remote hosts	Used in the context of a ToR switch to denote the hosts that are accessible via the uplink of the ToR switch.
VM	Virtual Machine. This is a logical instance of a host that operates independently in a physical host and has its own IP and MAC addresses. The VM architecture allows efficient use of physical host resources (such as multiple CPU cores).

2.0 Configuration

It is assumed that ARP reduction methodologies that are defined in this document will be limited to ToR switches. The maximum benefit of restraining ARP broadcasts in the network is achieved by the first hop switches (the ones directly connected to the hosts) without placing additional burden on second or third tier switches.

First, the ToR switches would need to be configured in order to enable the ARP reduction feature. Every Ethernet interface needs to be identified as either a downlink or uplink within the context of this feature. The ARP reduction feature treats ARP frames received from downlink or uplink differently as described in the following sections.

In addition the operator may optionally configure various ARP reduction related parameters such as:

- . ARP aging timer,
- . size of the ARP table,
- . static entries of IP to MAC address, etc.

3.0 Building the ARP tables

When ARP reduction is enabled, the ToR switch will monitor all ARP traffic transiting the switch (regardless of uplink port or downlink port) and will process any ARP PDUs in the following manner:

- . ARP Request PDUs must be redirected to control plane CPU.
- . Gratuitous ARP PDUs (ARP Reply PDU with a broadcast MAC DA) must be redirected to control plane CPU.
- . Other ARP Reply PDUs (ARP Reply PDU with a unicast MAC DA) should be bi-casted; one copy sent to control plane CPU and other copy forwarded out normally.

3.1 ARP Requests

The ToR examines the source IP and the source hardware address (MAC address) in the ARP Request. The source IP and MAC address association is learned, or is updated/refreshed if already learned. The destination IP address is searched in the ARP table. If an entry exists, the associated MAC address from the table is used to prepare a unicast ARP Reply PDU. The same MAC address is used as the source MAC address in the MAC header, as well as for the target hardware address, in the unicast ARP Reply PDU.

If the destination IP address in the request is not present in the ARP table, then the original ARP request PDU is broadcast to all the switch ports that are member of the same VLAN except the source port that the Request was received from. However, if the requested

(destination) IP address is present in the ARP table, a unicast ARP Reply PDU is prepared as described above and sent to the switch port from which the ARP Request was received and original ARP request PDU is dropped.

The intent is to prevent propagation of ARP Request PDU broadcasts as much as possible using the information present in the ARP table. The following observations can be made from such behavior.

- . Most of the ARP requests from the local hosts of a ToR switch for the local hosts of the ToR switch can be prevented.
- . Most of the ARP requests from the remote hosts of a ToR switch for the local hosts of the ToR switch can be prevented from getting forwarded on downlinks or other uplinks of the ToR switch.
- . Many of the ARP requests from the local hosts of a ToR switch for the remote hosts of the ToR switch can be prevented from being forwarded on uplinks if the remote host IP to MAC association is known to the ToR switch.

3.2 ARP Reply

The unicast ARP Reply is examined to learn/update the ARP table for source and destination IP/MAC address association, but is also forwarded out as a normal frame.

3.3 Gratuitous ARP

Gratuitous ARP is a broadcast ARP Reply PDU with destination IP address set to the IP address of the sender and target hardware address set to the MAC address of the sender. It is typically used by the IP hosts (including VMs) to keep its association fresh in peer's ARP cache.

The ToR switch should process Gratuitous ARP in the following manner.

- . Learn/update/refresh the ARP table entry.
- . If the IP address is new, or exists but with a different hardware address, then the Gratuitous ARP PDU is forwarded out; otherwise the PDU is discarded.

The goal for handling of the Gratuitous ARP PDU received from the downlinks (i.e. local hosts) is to avoid propagating it into the 'network' (i.e. to uplinks), unless there is a new association.

By suppressing the propagation of Gratuitous ARP PDUs, the peer IP hosts will end up aging out the corresponding ARP table entries. This will result in generation of the broadcast ARP Requests by those IP hosts if they need to continue to communicate with the IP host whose Gratuitous ARPs were obstructed. The handling of the ARP Request, as described above, by the first hop ToR switch will be able to respond to this request based on the ARP cache maintained in the ToR switch. In essence, presence of large ARP tables with longer age out times compensates for the smaller ARP table present in the

IP hosts and eliminates the need for periodic use of Gratuitous ARPs in order to refresh the ARP table in the IP hosts.

3.4 Host movement

As mentioned earlier, server virtualization technology allows movement of VMs to different physical servers. The flexibility to move VMs is one of the key benefits of server virtualization. The VM movement could be manual (operator initiated) or may be done automatically in reaction to demands placed by the application users. The important point is that in either case, VM movement is not transparent and is made known to the network.

There is ongoing work in IEEE 802.1 standards organization (IEEE 802.1Qbg) to coordinate/communicate the presence and capabilities of the VMs to the directly connected network switch.

VMs typically retain their MAC and IP address, and as such, there would be little impact to the ARP table maintained by the ARP reduction mechanism described herein. However, the ARP reduction mechanism would benefit from knowing if a VM is completely decommissioned so that the ToR can remove the ARP entry it has for that VM in a timely fashion, rather than waiting for it to timeout.

3.5 Applicability to environments with overlay transport

Recently, there have been multiple proposals for using overlay transport technologies such as VXLAN [VXLAN] and NVGRE [NVGRE]. These proposals allow the network operator to build the network using L2 or L3 technologies while building an L2-overlay on top of that. As such, while they address the issue of network design, they do not eliminate the need for a mechanism to reduce the amount of broadcast traffic that may have to traverse the core, if there are VMs of the same tenant on servers attached to different ToR switches.

One of the ways for the overlay transport proposals to address this issue would be to implement the mechanism discussed in this document at the point where the overlay encapsulation and decapsulation is performed (i.e. in the virtual switch).

3.6 Scaling Considerations

Depending on the number of hosts in the networks, the ARP table can be quite large. Although it is possible to implement some of the mechanisms for ARP reduction as described in this document in hardware in the forwarding plane, the number of ARP entries may favor maintaining the ARP table in the control plane memory.

3.7 Miscellaneous Issues

Because of the distributed nature of the mechanisms described herein, there are a few additional issues that warrant consideration from the network operator.

Earlier in the document, we had mentioned the configuration of a timer for ARP entries. A longer timer for holding on to ARP entries helps with reduction of broadcasts. However, the risk of having a "too large timer" can cause problems in certain situations. Consider the following scenario. Host A is attached to ToR switch #1, and host B is attached to ToR switch #2. If host B issues an ARP request for host A, if the entry is available at switch #2, then switch #2 would send the ARP Reply on behalf of host A. It is possible that host A is no longer available, but there is no way for switch #2 to know this, and it would continue to respond on behalf of host A, until its entry for host A has timed out. In this case, it is easy to see that a smaller timer would be beneficial. Additionally, since host B has an ARP age timer, it means that host B would find out about host A's unavailability only after its entry has aged, which would be after it has aged out of switch #2.

Another issue that can be somewhat problematic could be the inconsistency of tables in switches. Once again, consider a scenario similar to the one described above with 2 hosts each connected to its respect ToR switch. Let the ARP entries at both A and B be learned by both switches. Now assume that the IP address on host A changes. This change is signaled to switch #1 which in turn broadcasts the message on its uplink. Now, if this message is discarded due to network congestion or signal integrity issues, then switch #2 will not learn about the change and will continue to respond to host B's ARP Requests for host A's old IP address with stale information. This lasts until the ARP entry for A times out at Switch #2.

4.0 Conclusion

Based on the procedures described in this document, it is possible for ToR switches in the data center to contain ARP broadcasts significantly. The solution is based on well known, non-intrusive procedures and strives to curtail broadcasts that are increasingly becoming a cause for concern in the data centers. In essence, ToR switches facilitate the offloading of the extended ARP table management from the IP hosts to itself. The ARP table timeout can be tuned higher by the operator based on the available switch resources and network traffic behavior. The larger capacity of the ARP table directly translates to more effective subduing of the ARP broadcasts.

5.0 Security Considerations

The details of the security aspects will be addressed in future revision.

6.0 Acknowledgments

This document resulted from discussions with Linda Durbar (Huawei), Sue Hares (Huawei), and T Sridhar (VMware). We would like to acknowledge their contribution to this work.

7.0 References

7.1 Normative References

[ARP] D. Plummer, "An Ethernet Address Resolution Protocol: Or Converting Network Protocol Addresses to 48.bit Ethernet Addresses for Transmission on Ethernet Hardware," RFC 826, STD 37.

[ARP-Problem] T. Narten, "Problem Statement for ARMD," work in progress, <draft-ietf-armd-problem-statement>.

7.2 Informative References

[ARP-Mediation] H. Shah et al., "ARP Mediation for IP interworking in Layer 2 VPN," work in progress, <draft-ietf-l2vpn-arp-mediation>.

[IPLS] H.Shah et al., "IP-only LAN service," work in progress, <draft-ietf-l2vpn-ipls>.

[PROXY-ARP] J. Postel, "Multi-LAN Address Resolution," RFC 925.

[RFC1027] Smoot et al., "Using ARP to Implement Transparent Subnet Gateways".

[VXLAN] M. Mahalingam et al., "VXLAN: A Framework for Overlaying Virtualized Layer 2 Networks over Layer 3 Networks", work in progress, <draft-mahalingam-dutt-dcops-vxlan>.

[NVGRE] M. Sridharan et al., " NVGRE: Network Virtualization using Generic Routing Encapsulation", work in progress, <draft-sridharan-virtualization-nvgre>.

8.0 Author's Address

Himanshu Shah
Ciena Corp
Email: hshah@ciena.com

Anoop Ghanwani
Brocade
Email: anoop@alumni.duke.edu

Nabil Bitar
Verizon
Email: nabil.n.bitar@verizon.com

Network working group
Internet Draft
Category: Informational

X. Xu
Huawei Technologies

S. Hares

Y. Fan
China Telecom

C. Jacquenet
France Telecom

Expires: January 2014

July 15, 2013

Virtual Subnet: A L3VPN-based Subnet Extension Solution

draft-xu-virtual-subnet-11

Abstract

This document describes a Layer3 Virtual Private Network (L3VPN)-based subnet extension solution referred to as Virtual Subnet, which can be used as a kind of Layer3 network virtualization overlay approach for data center interconnect.

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 January 15, 2014.

Copyright Notice

Copyright (c) 2013 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.

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].

Table of Contents

1. Introduction	4
2. Terminology	6
3. Solution Description.....	6
3.1. Unicast	6
3.1.1. Intra-subnet Unicast	6
3.1.2. Inter-subnet Unicast	7
3.2. Multicast	9
3.3. CE Host Discovery	9
3.4. ARP/ND Proxy	10
3.5. CE Host Mobility	10
3.6. Forwarding Table Scalability	10
3.6.1. MAC Table Reduction on Data Center Switches	10
3.6.2. PE Router FIB Reduction	11
3.6.3. PE Router RIB Reduction	12
3.7. ARP/ND Cache Table Scalability on Default Gateways	14
3.8. ARP/ND and Unknown Uncast Flood Avoidance	14
3.9. Path Optimization	14
4. Considerations for Non-IP traffic	15
5. Security Considerations	15
6. IANA Considerations	15
7. Acknowledgements	15
8. References	15

8.1. Normative References	15
8.2. Informative References	15
Authors' Addresses	16

1. Introduction

For business continuity purposes, Virtual Machine (VM) migration across data centers is commonly used in those situations such as data center maintenance, data center migration, data center consolidation, data center expansion, and data center disaster avoidance. It's generally admitted that IP renumbering of servers (i.e., VMs) after the migration is usually complex and costly at the risk of extending the business downtime during the process of migration. To allow the migration of a VM from one data center to another without IP renumbering, the subnet on which the VM resides needs to be extended across these data centers.

In Infrastructure-as-a-Service (IaaS) cloud data center environments, to achieve subnet extension across multiple data centers in a scalable way, the following requirements SHOULD be considered for any data center interconnect solution:

1) VPN Instance Space Scalability

In a modern cloud data center environment, thousands or even tens of thousands of tenants could be hosted over a shared network infrastructure. For security and performance isolation purposes, these tenants need to be isolated from one another. Hence, the data center interconnect solution SHOULD be capable of providing a large enough Virtual Private Network (VPN) instance space for tenant isolation.

2) Forwarding Table Scalability

With the development of server virtualization technologies, a single cloud data center containing millions of VMs is not uncommon. This number already implies a big challenge for data center switches, especially for core/aggregation switches, from the perspective of forwarding table scalability. Provided that multiple data centers of such scale were interconnected at layer2, this challenge would be even worse. Hence an ideal data center interconnect solution SHOULD prevent the forwarding table size of data center switches from growing by folds as the number of data centers to be interconnected increases. Furthermore, if any kind of L2VPN or L3VPN technologies is used for interconnecting data centers, the scale of forwarding tables on PE routers SHOULD be taken into consideration as well.

3) ARP/ND Cache Table Scalability on Default Gateways

[RFC6820] notes that the Address Resolution Protocol (ARP)/Neighbor Discovery (ND) cache tables maintained by data center default gateways in cloud data centers can raise both scalability and security issues. Therefore, an ideal data center interconnect solution SHOULD prevent the ARP/ND cache table size from growing by multiples as the number of data centers to be connected increases.

4) ARP/ND and Unknown Unicast Flood Suppression or Avoidance

It's well-known that the flooding of Address Resolution Protocol (ARP)/Neighbor Discovery (ND) broadcast/multicast and unknown unicast traffic within a large Layer2 network are likely to affect performances of networks and hosts. As multiple data centers each containing millions of VMs are interconnected together across the Wide Area Network (WAN) at layer2, the impact of flooding as mentioned above will become even worse. As such, it becomes increasingly desirable for data center operators to suppress or even avoid the flooding of ARP/ND broadcast/multicast and unknown unicast traffic across data centers.

5) Path Optimization

A subnet usually indicates a location in the network. However, when a subnet has been extended across multiple geographically dispersed data center locations, the location semantics of such subnet is not retained any longer. As a result, the traffic from a cloud user (i.e., a VPN user) which is destined for a given server located at one data center location of such extended subnet may arrive at another data center location firstly according to the subnet route, and then be forwarded to the location where the service is actually located. This suboptimal routing would obviously result in the unnecessary consumption of the bandwidth resources which are intended for data center interconnection. Furthermore, in the case where the traditional VPLS technology [RFC4761, RFC4762] is used for data center interconnect and default gateways of different data center locations are configured within the same virtual router redundancy group, the returning traffic from that server to the cloud user may be forwarded at layer2 to a default gateway located at one of the remote data center premises, rather than the one placed at the local data center location. This suboptimal routing would also unnecessarily consume the bandwidth resources which are intended for data center interconnect.

This document describes a L3VPN-based subnet extension solution referred to as Virtual Subnet (VS), which can meet all of the

requirements of cloud data center interconnect as described above. Since VS mainly reuses existing technologies including BGP/MPLS IP VPN [RFC4364] and ARP/ND proxy [RFC925][RFC1027][RFC4389], it allows those service providers offering IaaS public cloud services to interconnect their geographically dispersed data centers in a much scalable way, and more importantly, data center interconnection design can rely upon their existing MPLS/BGP IP VPN infrastructures and their experiences in the delivery and the operation of MPLS/BGP IP VPN services.

Although Virtual Subnet is described as a data center interconnection solution in this document, there is no reason to assume that this technology couldn't be used within data centers.

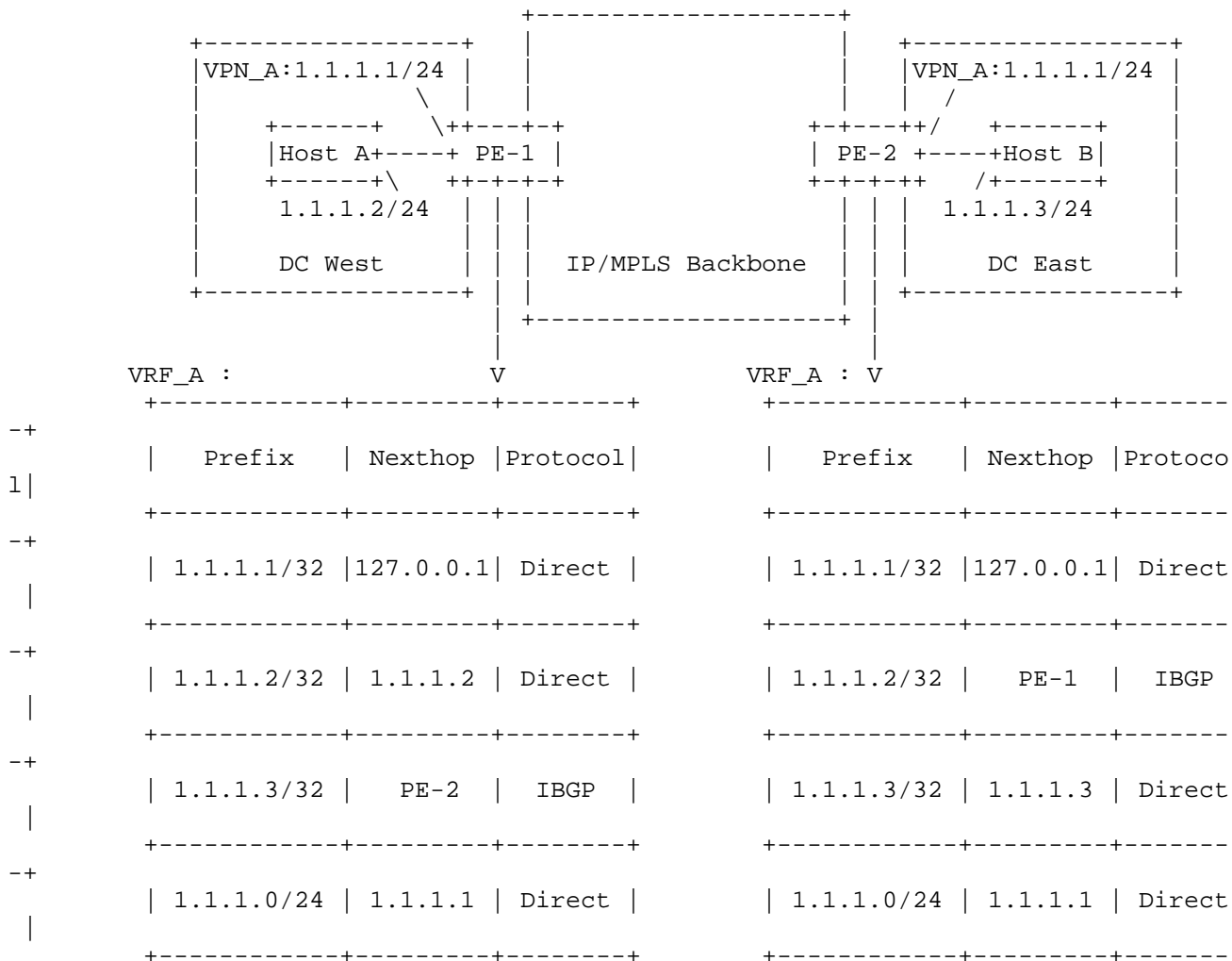
2. Terminology

This memo makes use of the terms defined in [RFC4364], [RFC2338] [MVPN] and [VA-AUTO].

3. Solution Description

3.1. Unicast

3.1.1. Intra-subnet Unicast



-+

Figure 1: Intra-subnet Unicast Example

Xu, et al.

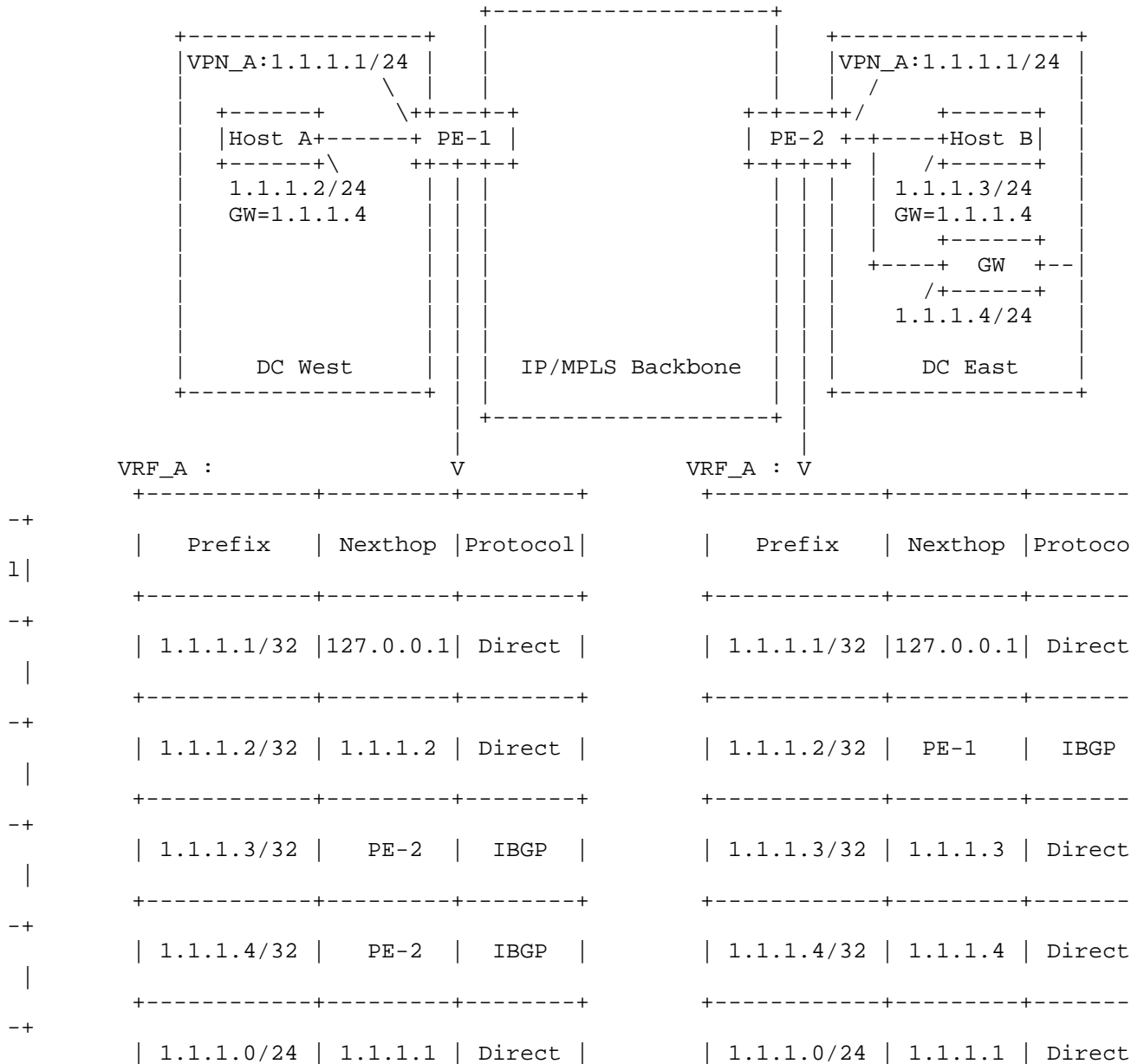
Expires January 15, 2014

[Page 6]

As shown in Figure 1, two CE hosts (i.e., Hosts A and B) belonging to the same subnet (i.e., 1.1.1.0/24) are located at different data centers (i.e., DC West and DC East) respectively. PE routers (i.e., PE-1 and PE-2) which are used for interconnecting these two data centers create host routes for their local CE hosts respectively and then advertise them via L3VPN signaling. Meanwhile, ARP proxy is enabled on VRF attachment circuits of these PE routers.

Now assume host A sends an ARP request for host B before communicating with host B. Upon receiving the ARP request, PE-1 acting as an ARP proxy returns its own MAC address as a response. Host A then sends IP packets for host B to PE-1. PE-1 tunnels such packets towards PE-2 which in turn forwards them to host B. Thus, hosts A and B can communicate with each other as if they were located within the same subnet.

3.1.2. Inter-subnet Unicast



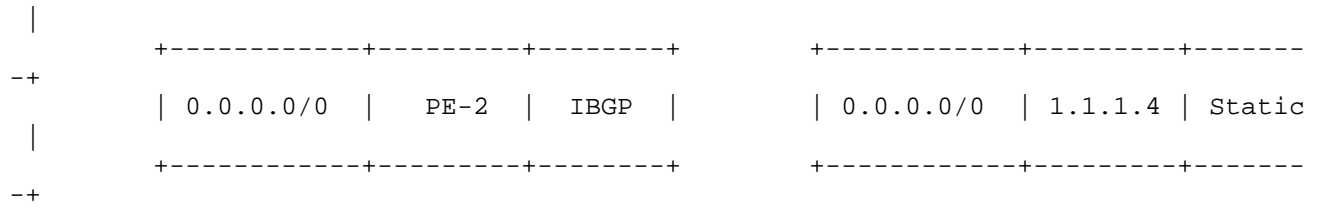
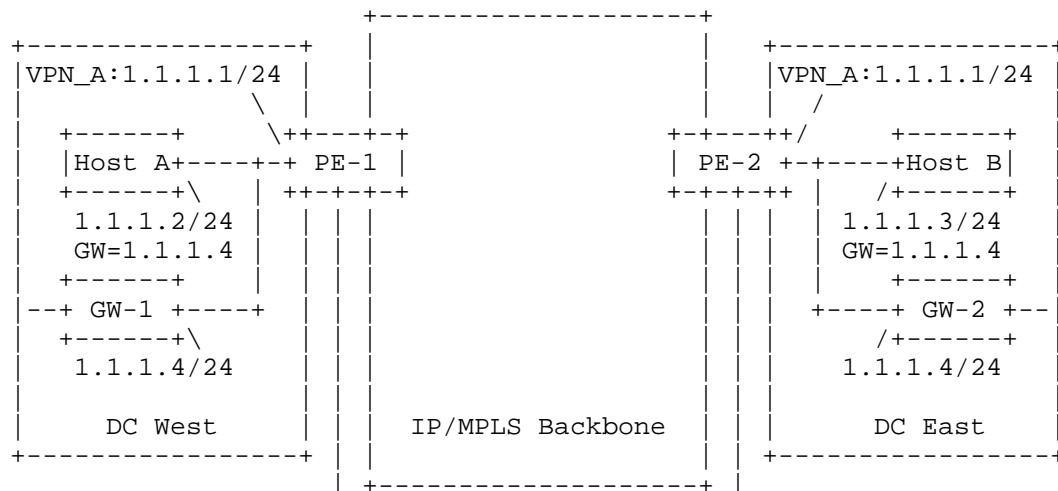


Figure 2: Inter-subnet Unicast Example (1)

As shown in Figure 2, only one data center (i.e., DC East) is deployed with a default gateway (i.e., GW). PE-2 which is connected to GW would either be configured with or learn from GW a default route with next-hop being pointed to GW. Meanwhile, this route is distributed to other PE routers (i.e., PE-1) as per normal [RFC4364] operation. Assume host A sends an ARP request for its default gateway (i.e., 1.1.1.4) prior to communicating with a destination host outside of its subnet. Upon receiving this ARP request, PE-1 acting as an ARP proxy returns its own MAC address as a response. Host A then sends a packet for Host B to PE-1. PE-1 tunnels such packet towards PE-2 according to the default route learnt from PE-2, which in turn forwards that packet to GW.



	VRF_A :	V			VRF_A :	V		
	+-----+	+-----+	+-----+		+-----+	+-----+	+-----+	
-+ 1		Prefix	Nexthop	Protocol		Prefix	Nexthop Proto	
	+-----+	+-----+	+-----+		+-----+	+-----+	+-----+	
-+ 		1.1.1.1/32	127.0.0.1	Direct		1.1.1.1/32	127.0.0.1 Direct	
	+-----+	+-----+	+-----+		+-----+	+-----+	+-----+	
-+ 		1.1.1.2/32	1.1.1.2	Direct		1.1.1.2/32	PE-1 IBGP	
	+-----+	+-----+	+-----+		+-----+	+-----+	+-----+	
-+ 		1.1.1.3/32	PE-2	IBGP		1.1.1.3/32	1.1.1.3 Direct	
	+-----+	+-----+	+-----+		+-----+	+-----+	+-----+	
-+ 		1.1.1.4/32	1.1.1.4	Direct		1.1.1.4/32	1.1.1.4 Direct	
	+-----+	+-----+	+-----+		+-----+	+-----+	+-----+	
-+ 		1.1.1.0/24	1.1.1.1	Direct		1.1.1.0/24	1.1.1.1 Direct	
	+-----+	+-----+	+-----+		+-----+	+-----+	+-----+	
-+ 		0.0.0.0/0	1.1.1.4	Static		0.0.0.0/0	1.1.1.4 Static	

+-----+-----+-----+ +-----+-----+-----+
-+

Figure 3: Inter-subnet Unicast Example (2)

As shown in Figure 3, in the case where each data center is deployed with a default gateway, CE hosts will get ARP responses directly from their local default gateways, rather than from their local PE routers when sending ARP requests for their default gateways.

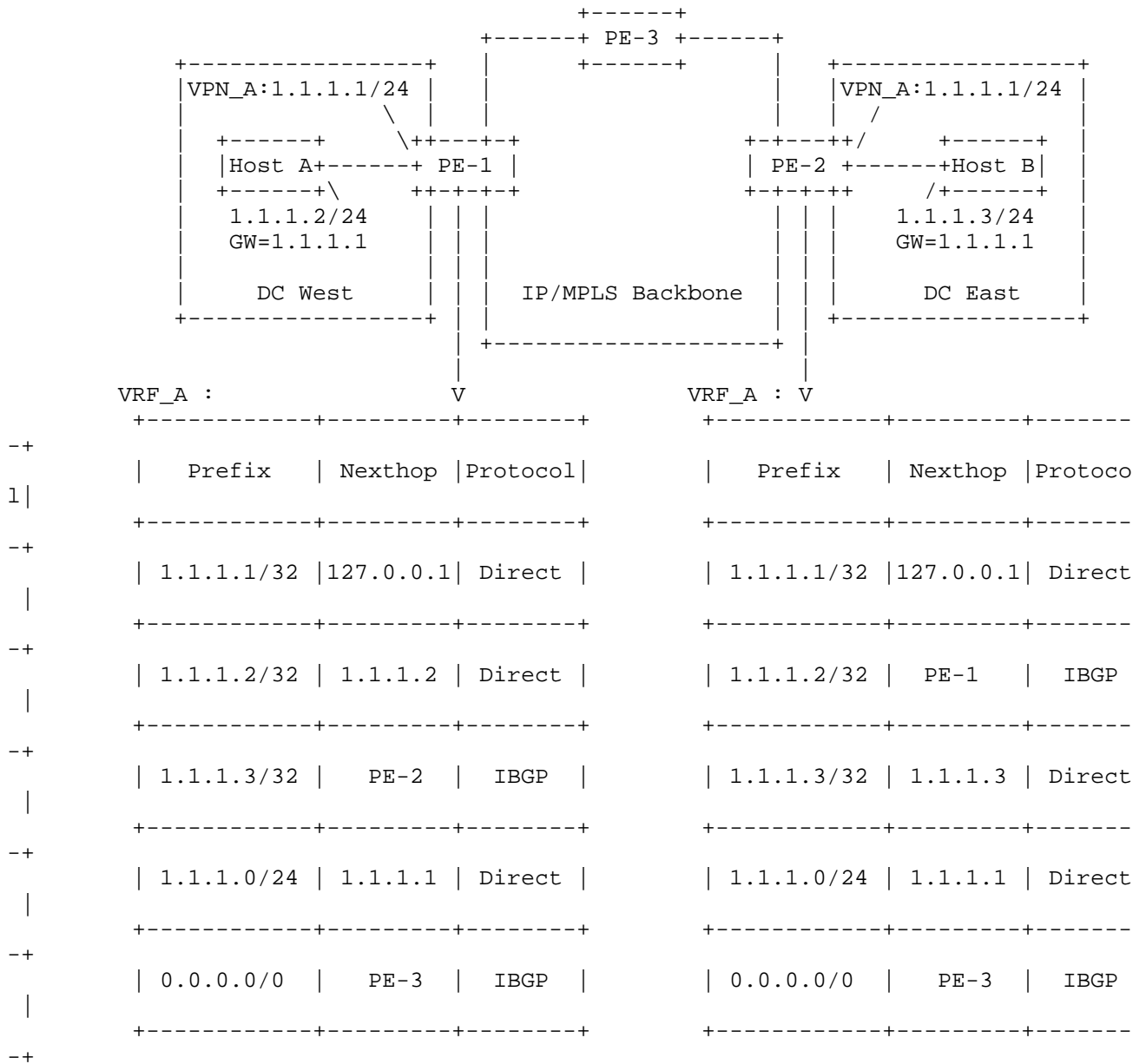


Figure 4: Inter-subnet Unicast Example (3)

Alternatively, as shown in Figure 4, PE routers themselves could be directly configured as default gateways of their locally connected CE hosts as long as these PE routers have routes for outside networks.

3.2. Multicast

To support IP multicast between CE hosts of the same virtual subnet, MVPN technology [MVPN] could be directly reused. For example, PE routers attached to a given VPN join a default provider multicast distribution tree which is dedicated for that VPN. Ingress PE routers, upon receiving multicast packets from their local CE hosts, forward them towards remote PE routers through the corresponding default provider multicast distribution tree.

More details about how to support multicast and broadcast in VS will be explored in a later version of this document.

3.3. CE Host Discovery

PE routers SHOULD be able to discover their local CE hosts and keep the list of these hosts up to date in a timely manner so as to ensure

the availability and accuracy of the corresponding host routes originated from them. PE routers could accomplish local CE host discovery by some traditional host discovery mechanisms using ARP or ND protocols. Furthermore, Link Layer Discovery Protocol (LLDP) described in [802.1AB] or VSI Discovery and Configuration Protocol (VDP) described in [802.1Qbg], or even interaction with the data center orchestration system could also be considered as a means to dynamically discover local CE hosts.

3.4. ARP/ND Proxy

Acting as ARP or ND proxies, PE routers SHOULD only respond to an ARP request or Neighbor Solicitation (NS) message for the target host when there is a corresponding host route in the associated VRF and the outgoing interface of that route is different from the one over which the ARP request or the NS message arrived.

In the scenario where a given VPN site (i.e., a data center) is multi-homed to more than one PE router via an Ethernet switch or an Ethernet network, Virtual Router Redundancy Protocol (VRRP) [RFC5798] is usually enabled on these PE routers. In this case, only the PE router being elected as the VRRP Master is allowed to perform the ARP/ND proxy function.

3.5. CE Host Mobility

During the VM migration process, the PE router to which the moving VM is now attached would create a host route for that CE host upon receiving a notification message of VM attachment while the PE router to which the moving VM was previously attached would withdraw the corresponding host route when receiving a notification message of VM detachment. Meanwhile, the latter PE router could optionally broadcast a gratuitous ARP/ND message on behalf of that CE host with source MAC address being one of its own. In the way, the ARP/ND entry of that moved CE host which has been cached on any local CE host would be updated accordingly.

3.6. Forwarding Table Scalability

3.6.1. MAC Table Reduction on Data Center Switches

In a VS environment, the MAC learning domain associated with a given virtual subnet which has been extended across multiple data centers is partitioned into segments and each segment is confined within a single data center. Therefore data center switches only need to learn local MAC addresses, rather than learning both local and remote MAC addresses.

3.6.2. PE Router FIB Reduction

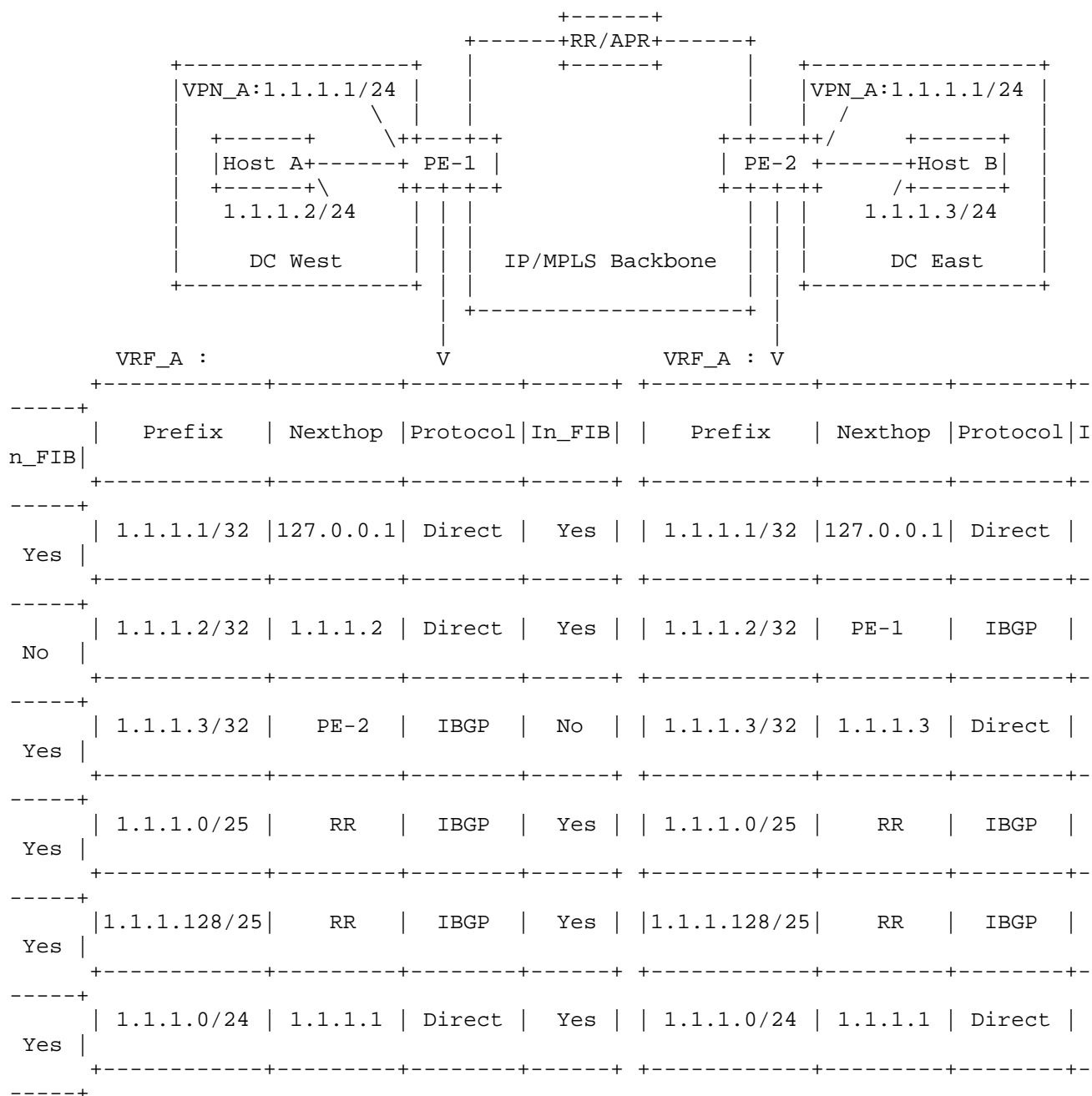


Figure 5: FIB Reduction Example

To reduce the FIB size of PE routers, Virtual Aggregation (VA) [VA-AUTO] technology can be used. Take the VPN instance A shown in Figure 5 as an example, the procedures of FIB reduction are as follows:

- 1) Multiple more specific prefixes (e.g., 1.1.1.0/25 and 1.1.1.128/25) corresponding to the prefix of virtual subnet (i.e., 1.1.1.0/24) are configured as Virtual Prefixes (VPs) and a Route-Reflector (RR) is configured as an Aggregation Point Router (APR) for these VPs. PE routers as RR clients advertise host routes for their own local CE hosts to the RR which in turn, as an APR, installs those host routes into its FIB and then attach the "can-suppress" tag to those

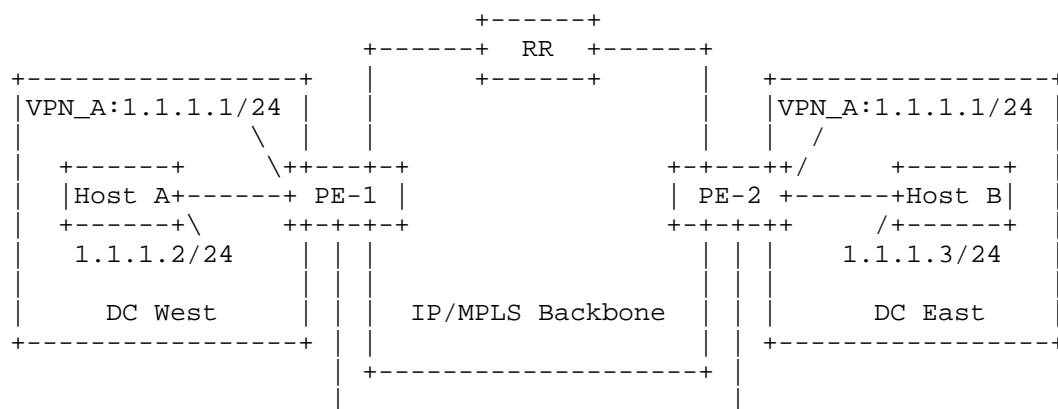
host routes before reflecting them to its clients.

- 2) Those host routes which have been attached with the "can suppress" tag would not be installed into FIBs by clients who are VA-aware since they are not APRs for those host routes. In addition, the RR as an APR would advertise the corresponding VP routes to all of its

clients, and those of which who are VA-aware in turn would install these VP routes into their FIBs.

- 3) Upon receiving a packet from a local CE host, if no matching host route found, the ingress PE router will forward the packet to the RR according to one of the VP routes learnt from the RR, which in turn forwards the packet to the relevant egress PE router according to the host route learnt from that egress PE router. In a word, the FIB table size of PE routers can be greatly reduced at the cost of path stretch. Note that in the case where the RR is not available for transferring L3VPN traffic between PE routers for some reason (e.g., the RR is implemented on a server, rather than a router), the APR function could actually be performed by a given PE router other than the RR as long as that PE router has installed all host routes belonging to the virtual subnet into its FIB. Thus, the RR only needs to attach a "can-suppress" tag to the host routes learnt from its clients before reflecting them to the other clients. Furthermore, PE routers themselves could directly attach the "can-suppress" tag to those host routes for their local CE hosts before distributing them to remote peers as well.
- 4) Provided a given local CE host sends an ARP request for a remote CE host, the PE router that receives such request will install the host route for that remote CE host into its FIB, in case there is a host route for that CE host in its RIB and has not yet been installed into the FIB. Therefore, the subsequent packets destined for that remote CE host will be forwarded directly to the egress PE router. To save the FIB space, FIB entries corresponding to remote host routes which have been attached with "can-suppress" tags would expire if they have not been used for forwarding packets for a certain period of time.

3.6.3. PE Router RIB Reduction



Internet-Draft	Virtual Subnet			July 2013			
VRF_A :	V			VRF_A : V			
	+	+	+	+	+	+	+
1		Prefix	Nexthop Protocol		Prefix	Nexthop Protoco	
	+	+	+	+	+	+	+
		1.1.1.1/32	127.0.0.1 Direct		1.1.1.1/32	127.0.0.1 Direct	
	+	+	+	+	+	+	+
		1.1.1.2/32	1.1.1.2 Direct		1.1.1.3/32	1.1.1.3 Direct	
	+	+	+	+	+	+	+
		1.1.1.0/25	RR IBGP		1.1.1.0/25	RR IBGP	
	+	+	+	+	+	+	+
		1.1.1.128/25	RR IBGP		1.1.1.128/25	RR IBGP	
	+	+	+	+	+	+	+
		1.1.1.0/24	1.1.1.1 Direct		1.1.1.0/24	1.1.1.1 Direct	
	+	+	+	+	+	+	+

Figure 6: RIB Reduction Example

To reduce the RIB size of PE routers, BGP Outbound Route Filtering (ORF) mechanism is used to realize on-demand route announcement. Take the VPN instance A shown in Figure 6 as an example, the procedures of RIB reduction are as follows:

- 1) PE routers as RR clients advertise host routes for their local CE hosts to a RR which however doesn't reflect these host routes by default unless it receives explicit ORF requests for them from its clients. The RR is configured with routes for more specific subnets (e.g., 1.1.1.0/25 and 1.1.1.128/25) corresponding to the virtual subnet (i.e., 1.1.1.0/24) with next-hop being pointed to Null0 and then advertises these routes to its clients via BGP.
- 2) Upon receiving a packet from a local CE host, if no matching host route found, the ingress PE router will forward the packet to the RR according to one of the subnet routes learnt from the RR, which in turn forwards the packet to the relevant egress PE router according to the host route learnt from that egress PE router. In a word, the RIB table size of PE routers can be greatly reduced at the cost of path stretch.
- 3) Just as the approach mentioned in section 3.6.2, in the case where the RR is not available for transferring L3VPN traffic between PE routers for some reason, a PE router other than the RR could advertise the more specific subnet routes as long as that PE router has installed all host routes belonging to that virtual subnet into its FIB.
- 4) Provided a given local CE host sends an ARP request for a remote CE host, the ingress PE router that receives such request will request the corresponding host route from its RR by using the ORF

mechanism (e.g., a group ORF containing Route-Target (RT) and prefix information) in case there is no host route for that CE host in its RIB yet. Once the host route for the remote CE host is

learned from the RR, the subsequent packets destined for that CE host would be forwarded directly to the egress PE router. Note that the RIB entries of remote host routes could expire if they have not been used for forwarding packets for a certain period of time. Once the expiration time for a given RIB entry is approaching, the PE router would notify its RR not to pass the updates for corresponding host route by using the ORF mechanism.

3.7. ARP/ND Cache Table Scalability on Default Gateways

In case where data center default gateway functions are implemented on PE routers of the VS as shown in Figure 4, since the ARP/ND cache table on each PE router only needs to contain ARP/ND entries of local CE hosts, the ARP/ND cache table size will not grow as the number of data centers to be connected increases.

3.8. ARP/ND and Unknown Unicast Flood Avoidance

In VS, the flooding domain associated with a given virtual subnet that has been extended across multiple data centers, has been partitioned into segments and each segment is confined within a single data center. Therefore, the performance impact on networks and servers caused by the flooding of ARP/ND broadcast/multicast and unknown unicast traffic is alleviated.

3.9. Path Optimization

Take the scenario shown in Figure 4 as an example, to optimize the forwarding path for traffic between cloud users and cloud data centers, PE routers located at cloud data centers (i.e., PE-1 and PE-2), which are also data center default gateways, propagate host routes for their local CE hosts respectively to remote PE routers which are attached to cloud user sites (i.e., PE-3).

As such, traffic from cloud user sites to a given server on the virtual subnet which has been extended across data centers would be forwarded directly to the data center location where that server resides, since traffic is now forwarded according to the host route for that server, rather than the subnet route.

Furthermore, for traffic coming from cloud data centers and forwarded to cloud user sites, each PE router acting as a default gateway would forward the traffic received from its local CE hosts according to the best-match route in the corresponding VRF. As a result, traffic from data centers to cloud user sites is forwarded along the optimal path as well.

4. Considerations for Non-IP traffic

Although most traffic within and across data centers is IP traffic, there may still be a few legacy clustering applications which rely on non-IP communications (e.g., heartbeat messages between cluster nodes). To support those few non-IP traffic (if present) in the Virtual Subnet solution, the approach following the idea of "route all IP traffic, bridge non-IP traffic" could be considered as an enhancement to the original Virtual Subnet solution.

Note that more and more cluster vendors are offering clustering applications based on Layer 3 interconnection.

5. Security Considerations

This document doesn't introduce additional security risk to BGP/MPLS L3VPN, nor does it provide any additional security feature for BGP/MPLS L3VPN.

6. IANA Considerations

There is no requirement for any IANA action.

7. Acknowledgements

Thanks to Dino Farinacci, Himanshu Shah, Nabil Bitar, Giles Heron, Ronald Bonica, Monique Morrow, Rajiv Asati and Eric Osborne for their valuable comments and suggestions on this document.

8. References

8.1. Normative References

[RFC2119] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, March 1997.

8.2. Informative References

[RFC4364] Rosen. E and Y. Rekhter, "BGP/MPLS IP Virtual Private Networks (VPNs)", RFC 4364, February 2006.

[MVPN] Rosen. E and Aggarwal. R, "Multicast in MPLS/BGP IP VPNs", draft-ietf-l3vpn-2547bis-mcast-10.txt, Work in Progress, January 2010.

- [VA-AUTO] Francis, P., Xu, X., Ballani, H., Jen, D., Raszuk, R., and L. Zhang, "Auto-Configuration in Virtual Aggregation", draft-ietf-grow-va-auto-05.txt, Work in Progress, December 2011.
- [RFC925] Postel, J., "Multi-LAN Address Resolution", RFC-925, USC Information Sciences Institute, October 1984.
- [RFC1027] Smoot Carl-Mitchell, John S. Quarterman, "Using ARP to Implement Transparent Subnet Gateways", RFC 1027, October 1987.
- [RFC4389] D. Thaler, M. Talwar, and C. Patel, "Neighbor Discovery Proxies (ND Proxy) ", RFC 4389, April 2006.
- [RFC5798] S. Nadas., "Virtual Router Redundancy Protocol", RFC 5798, March 2010.
- [RFC4761] Kompella, K. and Y. Rekhter, "Virtual Private LAN Service (VPLS) Using BGP for Auto-Discovery and Signaling", RFC 4761, January 2007.
- [RFC4762] Lasserre, M. and V. Kompella, "Virtual Private LAN Service (VPLS) Using Label Distribution Protocol (LDP) Signaling", RFC 4762, January 2007.
- [802.1AB] IEEE Standard 802.1AB-2009, "Station and Media Access Control Connectivity Discovery", September 17, 2009.
- [802.1Qbg] IEEE Draft Standard P802.1Qbg/D2.0, "Virtual Bridged Local Area Networks -Amendment XX: Edge Virtual Bridging", Work in Progress, December 1, 2011.
- [RFC6820] Narten, T., Karir, M., and I. Foo, "Problem Statement for ARMD", RFC 6820, January 2013.

Authors' Addresses

Xiaohu Xu
Huawei Technologies,
Beijing, China.
Phone: +86 10 60610041
Email: xuxiaohu@huawei.com

Susan Hares
Email: shares@ndzh.com

Internet-Draft

Virtual Subnet

July 2013

Yongbing Fan
Guangzhou Institute, China Telecom
Guangzhou, China.
Phone: +86 20 38639121
Email: fanyb@gsta.com

Christian Jacquenet
France Telecom
Rennes
France
Email: christian.jacquenet@orange.com