ARMD
Internet Draft
Intended status: Information Track
Expires: December 2011

Ning So
Verizon
L. Dunbar
Huawei

June 30, 2011

Address Resolution Requirements for VPN-oriented Data Center Services
draft-so-armd-vdcs-ar-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/1id-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 December 30, 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

VPN-oriented data center services seamlessly integrate the computing and storage resources in data centers and the users together with the traditional VPN services. This draft describes the address resolution issues and requirements induced by those services.

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

VPN-oriented Data Center Services (VDCS) integrate the virtual resources in data centers and user together using VPN as the common link. This kind of service is attractive to customers who often do not want to use public Internet to access data center resources. VDCS also have more restrictive requirements on what and how the virtualized data center resources can be shared. In addition, it provides a common service operational management framework using VPN as the central control point(s).

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.

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

VDCS: VPN oriented data center services

VM: Virtual Machines

VPN: Virtual Private Network

VPN-o-CS: VPN oriented Computing Service

3. VDCS service description

Many data centers offer virtualized services today, allowing clients to lease virtual data center resources without actually owning any physical servers or storage devices. However, majority of those services do not include network infrastructure. Intra-data center, inter-data center networks, and the networks connecting users to data centers are designed and operated separately from the data center server/storage systems. It is difficult for customers to integrate the leased virtual data center resources with their own internal data center resources, and make those leased resources appearing as if they come from their internal infrastructure.

VDCS has the following characteristics:

A secure collection of servers and/or virtual machines spanning one or more data centers.

All the applications running on the Virtual resources in network provider’s data centers are connected with the enterprise’s VPN in the same way as applications running over enterprise’s internal data centers. Therefore, the enterprises can treat those resources as if they are from their internal data centers.

Provide the VPN equivalent level of traffic segregation and privacy for those virtual resources attached to the VPN.

Make the virtual resources’ location known to VPN customers.

Created by network provider with no end host configuration.

Allow VMs and user devices using VDCS associated with one VPN to be partitioned into multiple subnets while still retain the detailed knowledge of each other.

Allow VPN clients to use private IP addresses (IPv4 or IPv6) for VDCS.

3.1. Components of VDCS

There are many components in VDCS system, including (but not limited to):

Network back office support systems, such as provisioning, billing, and etc,

VPN management systems such as monitoring, reporting, trouble shooting, and etc.

Data center resource monitoring systems, which include monitoring the utilization of servers and storage devices in data centers

Data center resource management systems, which include VMs placement to servers and racks based on the criteria associated with VMs.

Others.

This draft only focuses on networking (switching and routing) related components within VDCS framework.

3.2. Networking related components in support of VDCS

In the figure below, Vx represents a VM or a server belonging to VPN-x. The data center depicted in the figure has VMs belonging to 5 different VPNs, VPN-1, VPN-2, VPN-3, VPN-4, and VPN-5. Most data centers have many rows of server racks. Each rack holds many servers and has 1 or 2 Top of Rack (ToR) switches. Each server can have many VMs. The ToRs can be connected to aggregation switches/routers, which are then connected to Data Center gateway switches/routers. In some data centers, ToRs may be directly connected to Data Center gateway switches/routers.

It is essential to segregate traffic from VMs belonging to different VPNs within one data center and across multiple data centers. VLAN is usually used to segregate traffic from different VPNs within one data center. However, when a data center needs to house virtual machines belonging to more than 4095 VPNs, alternative segregation methods have to be used.

The virtual machines in data center can be connected to VDCS via L2VPN or L3VPN. For VMs belonging to L3VPN, the data center gateway router and the VPN PE router have to maintain detailed VRF tables that contain all the VM IP addresses associated with the each VPN. For VMs belonging to L2VPN, the data center gateway switch and the VPN edge switch have to maintain detailed Learned MAC Table that contains all the VM MAC addresses associated with each VPN.

```
   ------------------------------------+-----------------+
              Layer 2 based            |                 |
   +--------+                          |                 |
   |V1|V1|V3|----+                     |                 |
   +--------+    |         +--------+  |  +-----------+  |
   +--------+    +---------| DC GW  |--|--|           |  |
   |V2|V1|VM|--------------|Switch/ |  |  |           |  |
   +--------+    +---------|Router  |--|--|  VPN Edge/|  |
   +--------+    |         |        |--|--|  Switched |  |
   |V2|V4|V5|----+         |        |--|--|    router |  |
   +--------+    +---------|        |--|--|           |  |
                 | +-------+--------+  |  |           |  |
   +--------+    | |                   |  |           |  |
   |V2|V2|V2|----+ |                   |  |           |  |
   +--------+      |                   |  |           |  |
   +--------+      |                   |  |           |  |
   |V4|V1|V1|------+                   |  |           |  |
   +--------+                          |  + ----------+  |
                                       |                 |
   ------------------------------------+-----------------+
```

Figure 1 VMs and Network in Data Center

When VMs belonging to one VPN are partitioned into multiple subnets, it is necessary to have VLANs or other mechanisms to segregate traffic from different subnets belonging to one VPN.

4. Address resolution Scaling Issue for VDCS

4.1. Address Resolution for VMs attached to L2VPN

Before severs in a data center are instantiated with VMs for a particular VPLS L2VPN for the very first time (i.e. there is no VMs in the data center belonging to the L2VPN yet), the data center gateway router (CE router) should have the base VPLS configured already, which means a full mesh of pseudo-wires between L2VPN PEs already exist. The CE should have an attachment circuit (AC) built for the VPLS service between CE and PE.

At the time of VDCS instantiation, the new VMs’ MAC addresses are learned and added to the CE and PE’s MAC Table, so they can be learned by other switches and end stations already on the L2VPN in multiple sites as if they are on one LAN.

When a host or a VM in a data center needs to communicate with another host/VM in the L2VPN, an ARP (IPv4) or a ND(IPv6) is flooded to all PWs and all ACs (except the one from which the request is coming from).

Under this scenario, all VMs’ MAC addresses belonging to a particular L2VPN are visible to each other. And the L2VPN’s PEs and VSIs have to learn and maintain the MAC and VLAN addresses for all the hosts/VMs associated with this L2VPN. This may leads to address table scalability problems for data center VSI and L2VPN PE.

For example, assuming there are 1000 L2VPNs with hosts/VMs residing in this data center. That translates to 1000 VSIs on the CE, with each VSI containing the entire MAC and VLAN mapping for all the switches and end-stations associated with all the L2VPNs. This requires a very large amount of memory for the data center gateway switch/router using current technology.

```
   ------------------------------------------------------+
              Layer 2 based                              |
   +--------+                          |                 |
   |V1|V1|V3|----+                     |                 |
   +--------+    |         +--------+  |  +-----------+  |
   +--------+    +---------| DC GW  |--|--|  +----+   |  |
   |V2|V1|VM|--------------|        |  |  |  |VSI1|   |  |
   +--------+    +---------|        |--|--|  +----+   |  |
   +--------+    |         |        |--|--|  +----+   |  |
   |V2|V4|V5|----+         |        |--|--|  |VSI2|   |  |
   +--------+    +---------|        |--|--|  +----+   |  |
                 | +-------+--------+  |  |  |VSI3|   |  |
   +--------+    | |                   |  |  +----+   |  |
   |V2|V2|V2|----+ |                   |  |  |VSI4|   |  |
   +--------+      |                   |  |  +----+   |  |
   +--------+      |                   |  |  |VSI5|   |  |
   |V4|V1|V1|------+                   |  |  +----+   |  |
   +--------+                          |  + ----------+  |
                                       |       PE        |
   ------------------------------------+-----------------+
```

Figure 2 L2VPN associated VMs in Data Center

4.2. Address Resolution for VMs attached to L3VPN

When severs in a data center are instantiated with VMs for a particular L3VPN for the very first time (i.e. there were no VMs in the data center belonging to the L3VPN yet), it assumes that all the necessary L3VPN configuration has already been completed on the data center gateway router (CE) and the L3VPN edge router (PE). There are two scenarios for VMs attached to L3VPN:

> Scenario 1: all the VMs belonging to the L3VPN client are added as a separate site for the L3VPN. Under this scenario, the provider data center becomes the additional site (or peers) to the L3VPN.

Scenario 2: Hosts or applications in client’s own data centers (or premises) see those VMs attached to L3VPN as if they are from the same subnets. Under this scenario, the traditional "subnet" concept is broken. VMs in the data center have to be connected to their designated sites as if they are in one subnet.

Under scenario 1, the APR/ND broadcast/multicast requests are terminated at the CE. Similar to the condition described in the last section on VMs attached to L2VPN, all IP addresses associated with all L3VPNs in the data center have to be learned and maintained at the CE and the L3VPN PE router.

This can require a very large amount of memory on the CE and PE router using today’s technology, especially when the CE and the PE routers are hosting both L2VPN and L3VPN simultaneously. The amount of memory requirement is even larger if those VMs addresses can’t be aggregated.

In addition, it is possible that IP addresses for VMs belonging to different VPNs could be duplicated.

```
   ------------------------------------------------------+
              Layer 2 based            |                 |
   +--------+                          |                 |
   |V1|V1|V3|----+                     |                 |
   +--------+    |         +--------+  |  +-----------+  |
   +--------+    +---------| DC GW  |--|--|  +----+   |  |
   |V2|V1|VM|--------------|Switches|  |  |  |CE1 |   |  |
   +--------+    +---------|        |--|--|  +----+   |  |
   +--------+    |         |        |--|--|  +----+   |  |
   |V2|V4|V5|----+         |        |--|--|  |CE2 |   |  |
   +--------+    +---------|        |--|--|  +----+   |  |
                 | +-------+--------+  |  |  |CE3 |   |  |
   +--------+    | |                   |  |  +----+   |  |
   |V2|V2|V2|----+ |                   |  |  |CE4 |   |  |
   +--------+      |                   |  |  +----+   |  |
   +--------+      |                   |  |  |CE5 |   |  |
   |V4|V1|V1|------+                   |  |  +----+   |  |
   +--------+                          |  + ----------+  |
                                       |    DC Gateway   |
   ------------------------------------+-----------------+
```

Figure 3 L3VPN associated VMs in Data Center

Under the Scenario 2, the ARP/ND messages from the VMs in the data center have to be flooded to the corresponding sites to which those VMs belonging. The data center gateway routers (CEs or PEs) have to do both L2VPN and L3VPN.

5. Conclusion and Recommendation

Future data center can scale up to millions of virtual machines. Theoretically, network service provider can make their data centers hosting VMs for all of their VPN clients. Using current technology, it is very difficult for routers in data center and at network edge facing the data center to maintain all the VSIs or VRFs needed for the huge number of VPNs and the VPN-associated VMs being deployed.

Therefore, we recommend ARMD WG to investigate alternative solutions on address resolution and address scalability issues to make data center gateway routers capable of supporting the VPN oriented data center services.

6. Manageability Considerations

This document does not add additional manageability considerations.

7. Security Considerations

This document has no additional requirement for security.

8. IANA Considerations

9. Acknowledgments

We want to acknowledge the following people for their valuable inputs to this draft: K.K.Ramakrishnan.

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

10. References

[VDCS] So, et al, "Requirement and Framework for VPN-Oriented Data Center Services", draft-so-vdcs-00, June 2011.

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

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

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

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

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

Authors' Addresses

Ning So
Verizon Inc.
2400 N. Glenville Ave.,
Richardson, TX75082
ning.so@verizonbusiness.com

Linda Dunbar
Huawei Technologies
5340 Legacy Drive, Suite 175
Plano, TX 75024, USA
Phone: (469) 277 5840
Email: ldunbar@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.