

NVo3
Internet Draft
Intended status: Informational
Expires: December 2012

L. Dunbar
Huawei
June 28, 2012

Issues of Mobility in DC Overlay network

[draft-dunbar-nvo3-overlay-mobility-issues-00.txt](#)

Abstract

This draft describes the issues introduced by VM mobility in Data center overlay network.

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."

This Internet-Draft will expire on December 28, 2011.

Copyright Notice

Copyright (c) 2012 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.

Table of Contents

1.	Introduction	2
2.	Terminology	3
3.	Issues associated with Multicast in Overlay Network.....	3
4.	Issues associated with more than 4k Tenant Separation.....	4
4.1.	Collision of local VLAN Identifiers when VMs Move.....	7
4.1.1.	Local VIDs Managed by External Controller.....	10
4.1.2.	Local VIDs Managed by NVE	11
4.2.	Tenant Virtual Network separation at the physical gateway routers	11
5.	Summary and Recommendations.....	12
6.	Manageability Considerations.....	13
7.	Security Considerations.....	13
8.	IANA Considerations	13
9.	Acknowledgments	13
10.	References	13
	Authors' Addresses	14
	Intellectual Property Statement.....	14
	Disclaimer of Validity	14

[1.](#) Introduction

Overlay networks, such as VxLAN, NVGRE, etc, have been proposed to scale networks in Data Center with massive number of hosts as the result of server virtualization and business demand.

Overlay network can hide the massive number of VMs' addresses from the switches/routers in the core (i.e. underlay network).

One of the key requirements stated in [NVo3-problem] is the ability of moving VMs across wider range of locations, which could be

multiple server racks, PODs, or locations, without changing VM's IP/MAC addresses. That means the association of VMs to their corresponding NVE is changing as VMs migrate. This dynamic nature of VM mobility in Data Center introduces new challenges and complications to overlay networks.

This draft describes some of the issues introduced by VM migration in overlay environment. The purpose of the draft is to ensure those issues will be addressed by future solutions.

[2. Terminology](#)

CE: VPN Customer Edge Device

DC: Data Center

DA: Destination Address

EOR: End of Row switches in data center.

VNID: Virtual Network Identifier

NVE: Network Virtualization Edge

PE: VPN Provider Edge Device

SA: Source Address

ToR: Top of Rack Switch. It is also known as access switch.

VM: Virtual Machines

VPLS: Virtual Private LAN Service

[3. Issues associated with Multicast in Overlay Network](#)

Some data centers avoid the use of IP Multicast due, primarily, to the perceptions of configuration/protocol complexity and multicast scaling limits. There are also many data center operators for whom multicast is critical. Among the latter group, multicast is used for Internet Television (IPTV), market data, cluster load balancing, gaming, just to name a few.

The use of multicast in overlay environment can impose some issues to network when VMs move, in particular:

The association between multicast members to NVE becomes dynamic as VMs move. At one moment, all members of a multicast group could be attached to one NVE. At another moment, some members of the multicast group could be attached to different NVEs. Among VMs attached to one NVE, some can send, while others can only receive.

In addition, Overlay, which hides the VM addresses, introduces the IGMP snooping issue in the core. With NVE adding outer header to data frames from VMs (i.e. applications), multicast addresses are hidden from the underlay networks, making switches in the underlay network not being able to snoop on the IGMP reports from multicast members.

For unicast data frames, overlay network edge (e.g. TRILL edge) can learn the inner-outer address mapping by observing data frames passing by. Since multicast address is not placed in the inner-header's SA field of data frame, the learning approach for unicast won't work for multicast in overlay.

TRILL solves the multicast inner-outer address learning issues by creating common multicast trees in the TRILL domain. If TRILL's multicast approach is used for DC with VM mobility, the multicast states maintained by switches/routers in the underlay network have to change as VMs move, which means switches in the underlay network have to be aware of VMs mobility and change multicast states accordingly.

Overall, the VM mobility in overlay environment make multicast more complicated for switches/routers in the underlay network and for NVEs.

[4.](#) Issues associated with more than 4k Tenant Separation

The [\[NVo3-framework\]](#) has a good figure showing the logical network seen by each tenant. There are L2 domains being connected by L3 infrastructure. Each tenant can have multiple virtual networks, which are identified IEEE802.1Q compliant 12 bits VLAN ID, under its logical routers (Rtr). Any VMs communicating with peers in different subnets, either within DC or outside DC, will have their L2 MAC address destined towards its local Router (Rtr in the figure below).

The overlay introduced by [NVo3-problem] makes the core (i.e. the underlay network) switches/routers forwarding tables not be impacted when VMs belonging to different tenants are placed or moved to anywhere, as shown in the Figure below (copied from [NVo3-framework]).

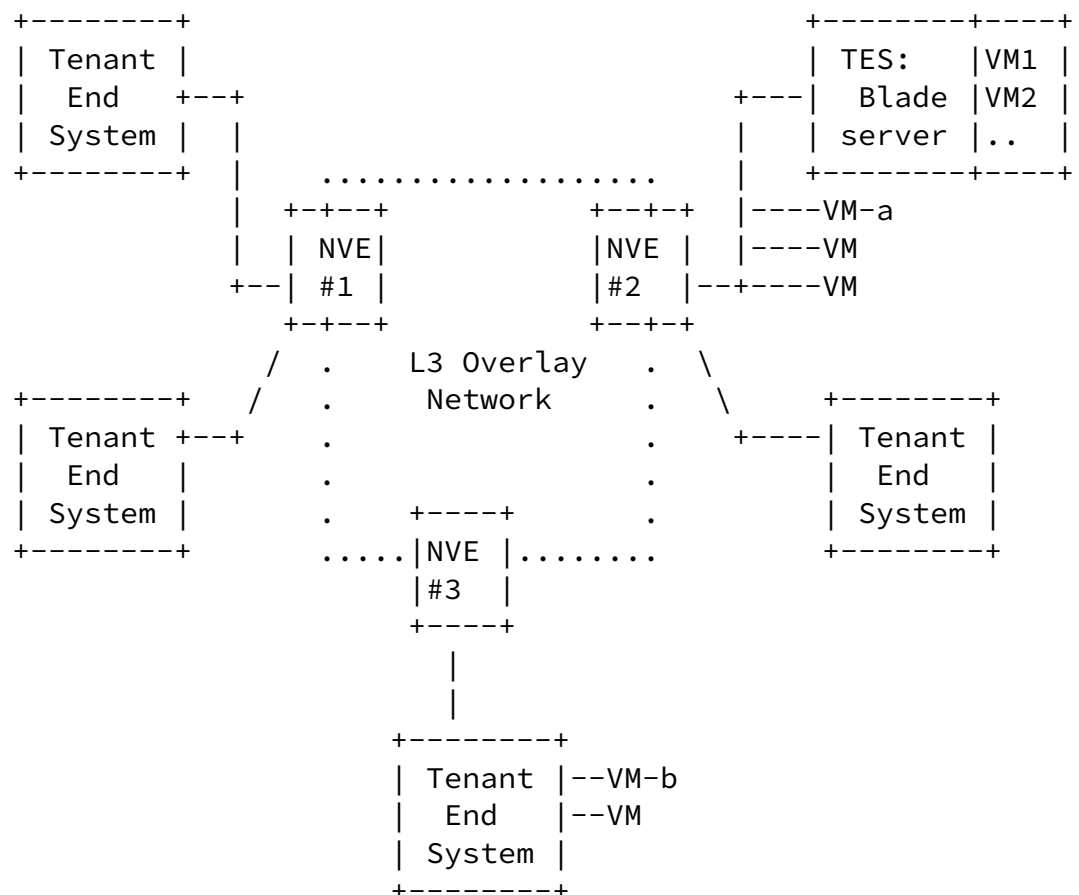


Figure 2: Overlay example

For client traffic "VM-a" to "VM-b", the ingress NVE encapsulates the client payload with an outer header which includes at least egress NVE as DA, ingress NVE as SA, and a VNID. The VNID is a 24-bits identifier proposed by [\[NVo3-Problem\]](#) to separate tens of thousands of tenant virtual networks. When the egress NVE receives the data frame from its ports facing the underlay network, the egress NVE decapsulates the outer header and then forward the decapsulated data frame to the attached VMs.

When "VM-b" is on the same subnet (or VLAN) as "VM-a" and located within the same data center, the corresponding egress NVE is usually on a virtual switch in a server, on a ToR switch, or on a blade switch.

When "VM-b" is on a different subnet (or VLAN), the corresponding egress NVE should be next to (or located on) the logical Rtr (Figure 1), which is most likely located on the data center gateway router(s).

[4.1](#). Collision of local VLAN Identifiers when VMs Move

Since the VMs attached to one NVE could belong to different virtual networks, the traffic under each NVE have to be identified by local network identifiers, which is usually VLAN if VMs are attached to NVE access ports via L2.

To support tens of thousands of virtual networks, the local VID associated with client payload under each NVE has to be locally significant. If ingress NVE simply encapsulates an outer header to data frames received from VMs and forward the encapsulated data frames to egress NVE via underlay network, the egress NVE can't simply decapsulate the outer header and send the decapsulated data frames to attached VMs as done by TRILL. Egress NVE needs to convert the VID carried in the data frame to a local VID for the virtual network before forwarding the data frame to the VMs attached.

In VPLS, operator has to configure the local VIDs under each PE to specific VPN instances. In VPLS, the local VID mapping to VPN instance ID doesn't change very much. In addition, most likely CE is not shared by multiple tenants, so the VIDs on one physical port of PE to CE are only for one tenant. For rare occasion of multiple tenants sharing one CE, the CE can convert the tuple [local customer VIDs & Tenant Access Port] to the VID designated by VPN operator for each VPN instance on the shared link between CE port and PE port. For example, in the figure below, the VIDs under CE#21 and the VIDs under CE#22 can be duplicated as long as the CEs can convert the local VIDs from their downstream links to the VIDs given by the VPN operators for the links between PE and CEs.

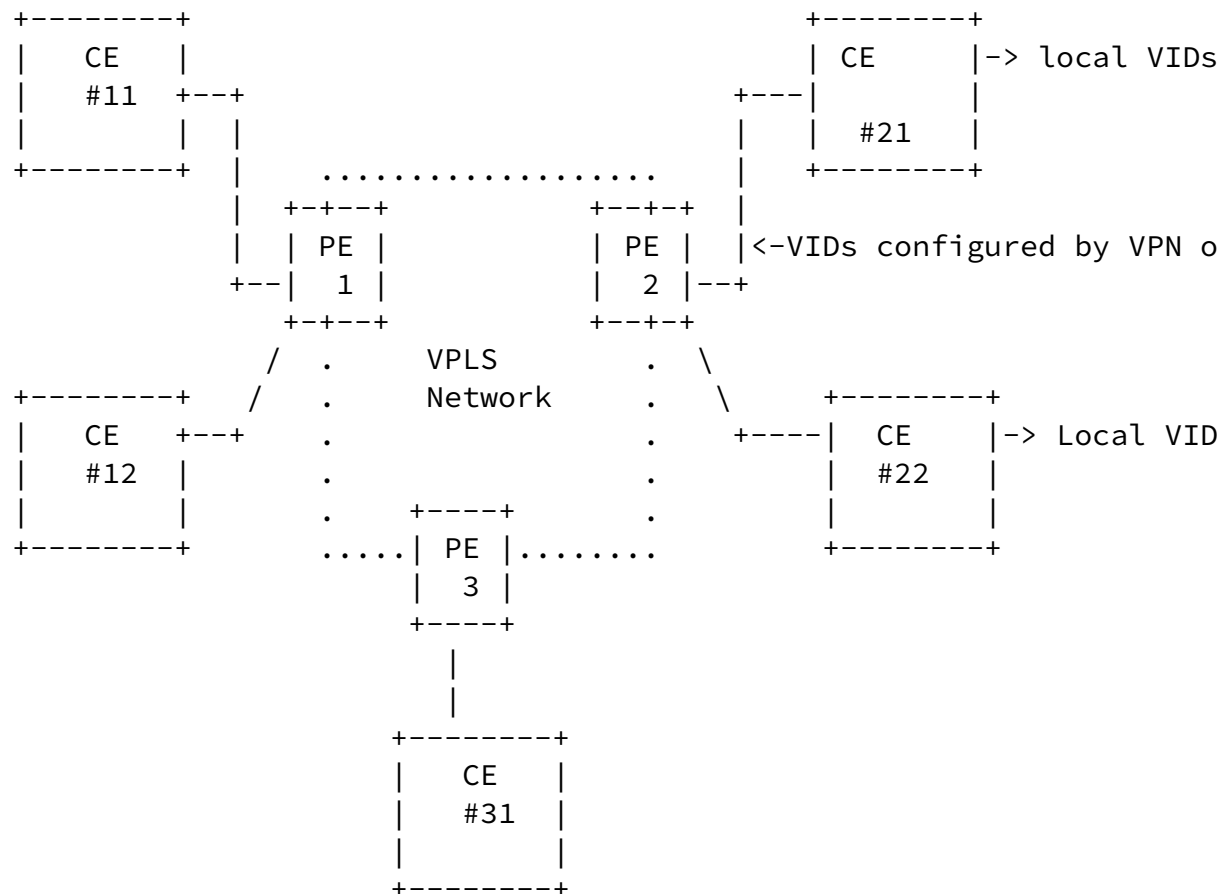


Figure 3: VPLS example

When all VMs under one virtual network are moved away from a NVE, the local VID, which was designated for this virtual network, might need to be used for different virtual network whose VMs are moved in later.

In the Figure below, the NVE#1 may have local VID #100~#200 assigned to some virtual networks attached. The NVE#2 may have local VID

#100~#150 assigned to different virtual networks. With VNID encoded in the outer header of data frames, the traffic in the L3 Overlay Network is strictly separated.

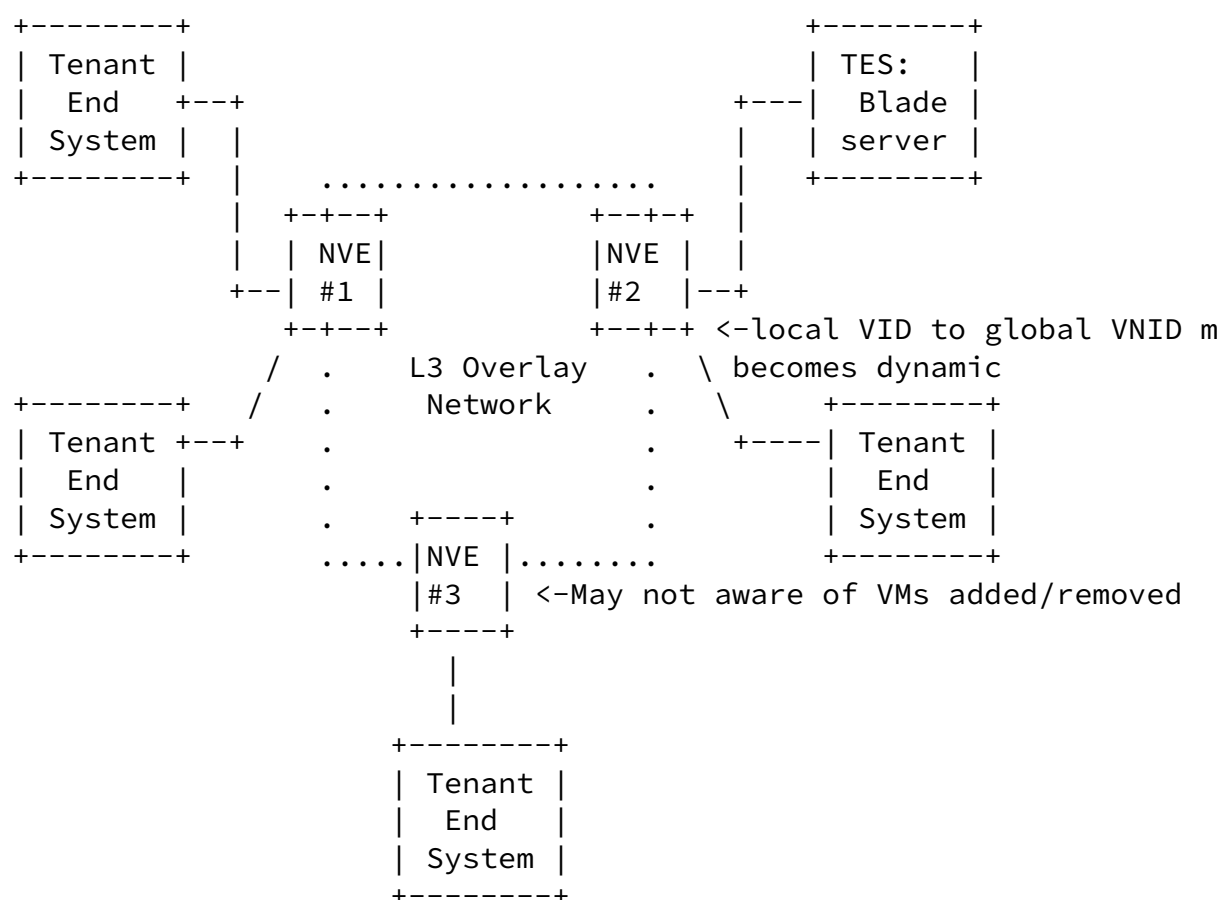


Figure 4: Overlay example

When some VMs associated with Virtual Network X using VID 120 under NVE1 are moved to NVE2, a new VID must be assigned for the Virtual Network X under NVE2.

It gets complicated when the local VIDs are tagged by none-NVE devices, e.g. VMs themselves, blade server switches, or virtual switches within servers.

The devices which add VID to untagged frames need to be informed of the local VID. If data frames from VMs already have VID encoded in data frames, then there has to be a mechanism to notify the first switch port facing the VMs to convert the VID encoded by the VMs to the local VID which is assigned for the virtual network under the new NVE. That means when a VM is moved to a new location, its immediate adjacent switch port has to be informed of local VID to convert the VID encoded in the data frames from the VM.

NVE will need the mapping between local VID and the VNID to be used to face L3 underlay network.

Dunbar December 28, 2012 [Page 9]

Internet-Draft

Mobility Issues in Overlay

Nov 1, 2011

[4.1.1](#). Local VIDs Managed by External Controller

Most likely the VM assignment to a physical location is managed by a non-networking entity, e.g. VM Manager or a Server Manager. NVEs may not be aware of VMs being added or deleted unless NVEs have a north bound interface to a controller which can communicate with VM/server Manager(s).

When NVE can be informed of VMs being added/deleted and their associated tenant virtual networks via its controller, NVE should be able to get the specific VNID from its controller for untagged data frames arriving at its Virtual Access Points [VNo3-framework 3.1.1].

Since local VIDs under each NVE are really locally significant, it might be less confusing to egress NVE if ingress NVE remove the local VID attached to the data frame. So that egress NVE always has to assign its own local VID to data frame before sending the decapsulated data frame to attached VMs.

If, for whatever reason, it is necessary to have local VID in the data frames before encapsulating outer header of EgressNVE-DA/ IngressNVE-SA /VNID, NVE should get the specific local VID from the

external Controller for those untagged data frames coming to each Virtual Access Point.

If the data frame is tagged before reaching the NVE's Virtual Access Point (e.g. tagged data frames from VMs) and NVE is more than one hop away from VMs, the first (virtual) port facing the VMs has to be informed by the external controller of the new local VID to replace the VID encoded in the data frames. For reverse direction, i.e. data frames coming from core towards VMs, the first switching port facing VMs have to convert the VIDs encoded in the data frames to the VIDs used by VMs.

The IEEE802.1Qbg's VDP protocol (Virtual Station Interface (VSI) discovery and configuration protocol) requires hypervisor to send VM profile upon a new VM is instantiated. However, not all hypervisors support this function.

[4.1.2](#). Local VIDs Managed by NVE

If NVEs don't have interface to any controllers which can be informed of VMs being added to or deleted from NVEs, then NVEs have to learn new VMs/VLANs being attached, figure out to which tenant virtual network those VMs/VLANs belong, and/or age out VMs/VLANs after a specified timer expires. Network management system has to assist NVEs in making the decision, even if the network management system doesn't have interface to VM/server managers.

When NVE receives a data frame with a new VM address (e.g. MAC) in a tagged data frame from its Virtual Access Point, the new VM could be from an existing local virtual network, from a different virtual network (being brought in as the VM being added in), or from an illegal VM.

Upon NVE learns a new VM being added, either by learning a new MAC address or a new VID, it needs its management system to confirm the validity of the new VID and/or new address. If the new address or VID is from invalid or illegal source, the data frame has to be dropped.

4.2. Tenant Virtual Network separation at the physical gateway routers

When a VM communicates with peers in a different subnets, data frames will be sent to the tenant logical Router (Rtr1 or Rtr2 in the Figure 1). Very often, the logical routers of all tenants in a data center are just logical entities (e.g. VRF) on the gateway router(s). That means that all the VLANs for all tenants will be terminated at the Data Center Gateway router(s), as shown in the figure below.

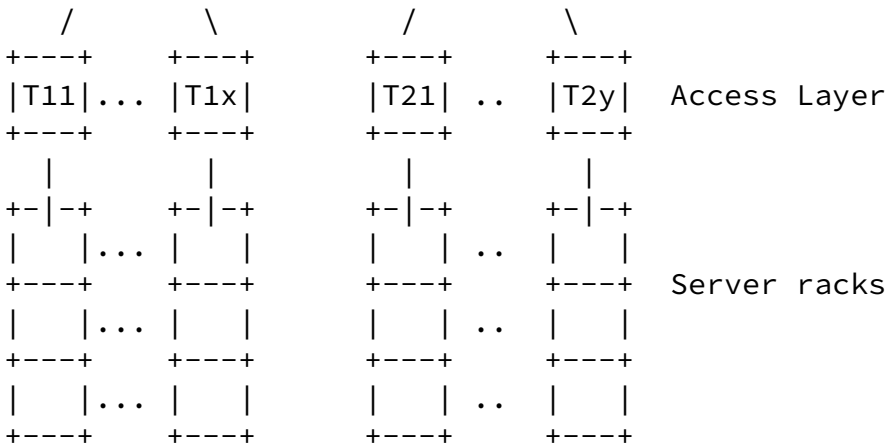
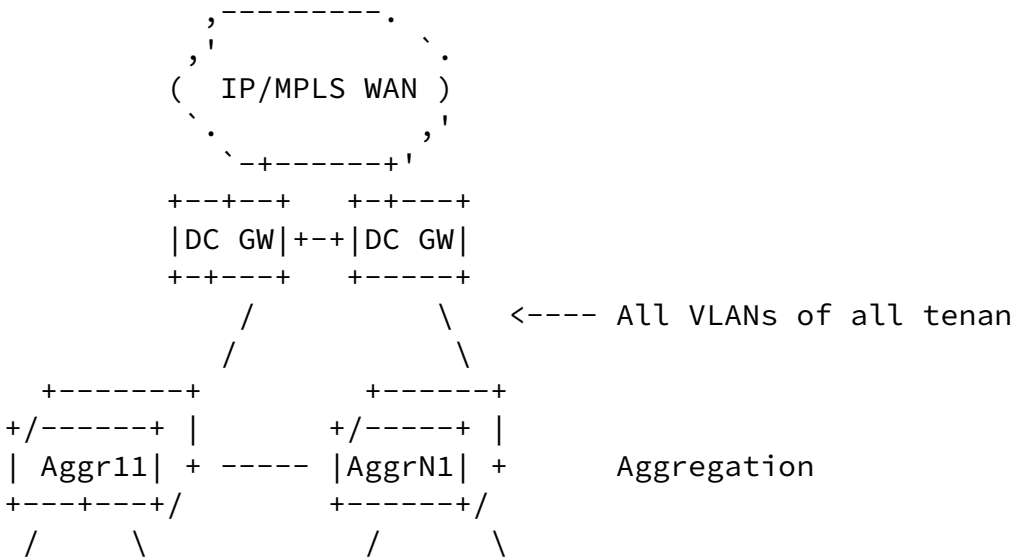


Figure 5: Data Center Physical topology

Gateway routers can mitigate the overwhelming number of virtual network instances by integrating NVE function within the router(s). That requires routers to map VNID to VRF directly if routers' outbound to external network is VPN based. That requires routers to support tens of thousands of VRF instances, which can be challenging to routers.

Data center can also use multiple gateway routers, with each handling a subset of tenants in data centers. That means that each tenant's VMs are only reachable by their designated routers or router ports. With the typical DC design shown in Figure 5, the number of server racks reachable by each gateway router is limited by the number of router ports enabled for the tenant virtual networks. That means the range of locations where each tenant's VMs can be moved across are limited.

When VMs in data center communicates with external peers, data frames have to go through gateway. Even though majority of data centers have much more east west traffic volume than north south traffic volume, majority (as high as 90%) of applications (hosted on servers or VMs) in a data center still communicate with external peers. Just the volume of north south traffic is much less in many data centers.

[5. Summary and Recommendations](#)

Overlay network can hide individual VMs addresses, making switches/routers in the core scalable. However overlay introduces other challenges, especially when VMs move across wide range of NVEs. This draft is to identify those issues introduced by mobility in

Dunbar December 28, 2012 [Page 12]

Internet-Draft

Mobility Issues in Overlay

Nov 1, 2011

overlay environment, to ensure that they will be addressed by future solutions.

[6. Manageability Considerations](#)

[7. Security Considerations](#)

Security will be addressed in a separate document.

[8. IANA Considerations](#)

None.

9. Acknowledgments

We want to acknowledge the following people for their valuable comments to this draft: David Black, Ben MackCrane, Peter AshwoodSmith, Lucy Yong and Young Lee.

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

10. References

[NVo3-Problem] Narten, et al, "Problem Statement: Overlays for Network Virtualization." Draft-narten-nvo3-overlay-problem-statement-02, June 2012.

[NVo3-framework] Lasserre, et al, "Framework for DC Network Virtualization". Draft-lasserre-nvo3-framework-02, June 2012

[IEEE802.1Qbg] "MAC Bridges and Virtual Bridged Local Area Networks - Edge Virtual Switch". IEEE802.1Qbg/D2.2, Feb, 2012. Work in progress

[ARMD-Problem] Narten,et al "[draft-ietf-armd-problem-statement](#)" in progress, Oct 2011.

[ARMD-Multicast] McBride, Lui, "[draft-mcbride-armd-mcast-overview-01](#)", in progress, March 10, 2012

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

Authors' Addresses

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.