Network Working Group
Internet Draft
Category: Informational

L. Yong
Huawei
M. Toy
Comcast
A. Isaac
Bloomberg
V. Manral
Hewlett-Packard
L. Dunbar
Huawei

Expires: November 2013 May 1, 2013

Use Cases for DC Network Virtualization Overlays

draft-ietf-nvo3-use-case-01

Abstract

This document describes the DC NVO3 use cases that may be potentially deployed in various data centers and apply to different applications. An application in a DC may be a combination of some use cases described here.

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 November, 2013.

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

Server Virtualization has changed IT industry in terms of efficiency, cost, and the speed in providing a new applications and/or services. However the problems in today's data center networks hinder the support of an elastic cloud service and dynamic virtual tenant networks [NVO3PRBM]. The goal of DC Network Virtualization Overlays, i.e. NVO3, is to decouple the communication among tenant systems from DC physical networks and to allow one physical network infrastructure to provide: 1) traffic isolation among tenant virtual networks over the same physical network; 2) independent address space in each virtual network and address isolation from the infrastructure's; 3) Flexible VM placement and move from one server to another without any of the physical network limitations. These characteristics will help address the issues that hinder true virtualization in the data centers [NVO3PRBM].

Although NVO3 enables a true virtualization environment, the NVO3 solution has to address the communication between a virtual network and a physical network. This is because 1) many DCs that need to provide network virtualization are currently running over physical networks, the migration will be in steps; 2) a lot of DC applications are served to Internet users which run directly on physical networks; 3) some applications are CPU bound like Big Data analytics and may not need the virtualization capability.

This document is to describe general NVO3 use cases that apply to various data centers. Three types of the use cases described here are:

- A virtual network connects many tenant systems within a Data Center and form one L2 or L3 communication domain. A virtual network segregates its traffic from others and allows the VMs in the network moving from one server to another. The case may be used for DC internal applications that constitute the DC East-West traffic.

- A DC provider offers a secure DC service to an enterprise customer and/or Internet users. In these cases, the enterprise customer may use a traditional VPN provided by a carrier or an IPsec tunnel over Internet connecting to a NVO3 network within a provider DC. This is mainly constitutes DC North-South traffic.

- o A DC provider may use NVO3 and other network technologies for a tenant network, construct different topologies or zones for a tenant network, and may design a variety of cloud applications that may require the network service appliance, virtual compute, storage, and networking. In this case, the NVO3 provides the networking functions for the applications.

The document uses the architecture reference model defined in [NVO3FRWK] to describe the use cases.

1.1. Contributors

Vinay Bannai
PayPal
2211 N. First St,
San Jose, CA 95131
Phone: +1-408-967-7784
Email: vbannai@paypal.com

Ram Krishnan
Brocade Communications
San Jose, CA 95134
Phone: +1-408-406-7890
Email: ramk@brocade.com

1.2. Terminology

This document uses the terminologies defined in [NVO3FRWK], [RFC4364]. Some additional terms used in the document are listed here.

CPE: Customer Premise Equipment

DMZ: Demilitarized Zone

DNS: Domain Name Service

NAT: Network Address Translation

VIRB: Virtual Integrated Routing/Bridging

Note that a virtual network in this document is a network virtualization overlay instance.

2. Basic Virtual Networks in a Data Center

A virtual network may exist within a DC. The network enables a communication among Tenant Systems (TSs) that are in a Closed User Group (CUG). A TS may be a physical server or virtual machine (VM) on a server. The network virtual edge (NVE) may co-exist with Tenant Systems, i.e. on an end-device, or exist on a different device, e.g. a top of rack switch (ToR). A virtual network has a unique virtual network identifier (may be local or global unique) for an NVE to properly differentiate it from other virtual networks.

The TSs attached to the same NVE are not necessary in the same CUG, i.e. in the same virtual network. The multiple CUGs can be constructed in a way so that the policies are enforced when the TSs in one CUG communicate with the TSs in other CUGs. An NVE provides the reachbility for Tenant Systems in a CUG, and may also have the policies and provide the reachbility for Tenant Systems in different CUGs (See section 4.2). Furthermore in a DC operators may construct many tenant networks that have no communication at all. In this case, each tenant network may use its own address space. Note that one tenant network may contain one or more CUGs.

A Tenant System may also be configured with multiple addresses and participate in multiple virtual networks, i.e. use different address in different virtual network. For examples, a TS is NAT GW; or a TS is a firewall server for multiple CUGs.

Network Virtualization Overlay in this context means the virtual networks over DC infrastructure network via a tunnel, i.e. a tunnel between any pair of NVEs. This architecture decouples tenant system address schema from the infrastructure address space, which brings a great flexibility for VM placement and mobility. This also makes the transit nodes in the infrastructure not aware of the existence of the virtual networks. One tunnel may carry the traffic belonging to different virtual networks; a virtual network identifier is used for traffic segregation in a tunnel.

A virtual network may be an L2 or L3 domain. An NVE may be a member of several virtual networks each of which is in L2 or L3. A virtual network may carry unicast traffic and/or broadcast/multicast/unknown traffic from/to tenant systems. An NVE may use p2p tunnels or a p2mp tunnel to transport broadcast or multicast traffic, or may use other mechanisms [NVO3MCAST].

It is worth to mention two distinct cases here. The first is that TS and NVE are co-located on a same end device, which means that the NVE can be made aware of the TS state at any time via internal API.

The second is that TS and NVE are remotely connected, i.e. connected via a switched network or point-to-point link. In this case, a protocol is necessary for NVE to know TS state.

One virtual network may have many NVE members each of which many TSs may attach to. TS dynamic placement and mobility results in frequent changes in the TS and NVE bindings. The TS reachbility update mechanism MUST be fast enough to not cause any service interruption. The capability of supporting a lot of TSs in a tenant network and a lot of tenant networks is critical for NVO3 solution.

If a virtual network spans across multiple DC sites, one design is to allow the corresponding NVO3 instance seamlessly span across those sites without DC gateway routers' termination. In this case, the tunnel between a pair of NVEs may in turn be tunneled over other intermediate tunnels over the Internet or other WANs, or the intra DC and inter DC tunnels are stitched together to form an end-to-end virtual network across DCs. The latter is described in section 3.2. Section 4.2 describes other options.

3. Interconnecting DC Virtual Network and External Networks

For customers (an enterprise or individuals) who want to utilize the DC provider's compute and storage resources to run their applications, they need to access their systems hosted in a DC through Internet or Service Providers' WANs. A DC provider may construct an NVO3 network which all the resources designated for a customer connect to and allow the customer to access their systems via the network. This, in turn, becomes the case of interconnecting a DC NVO3 network and external networks via Internet or WANs. Two cases are described here.

3.1. DC Virtual Network Access via Internet

A user or an enterprise customer connects securely to a DC virtual network via Internet. Figure 1 illustrates this case. A virtual network is configured on NVE1 and NVE2 and two NVEs are connected via an L3 tunnel in the Data Center. A set of tenant systems are attached to NVE1 on a server. The NVE2 resides on a DC Gateway device. NVE2 terminates the tunnel and uses the VNID on the packet to pass the packet to the corresponding VN GW entity on the DC GW. A user or customer can access their systems, i.e. TS1 or TSn, in the DC via Internet by using IPsec tunnel [RFC4301]. The IPsec tunnel is between the VN GW and the user or CPE at enterprise edge location. The VN GW provides IPsec functionality such as authentication scheme and encryption, as well as the mapping to the right virtual network entity on the DC GW. Note that 1) some VN GW functions such as

firewall and load balancer may also be performed by locally attached network appliance devices; 2) The virtual network in DC may use different address space than external users, then VN GW serves the NAT function.

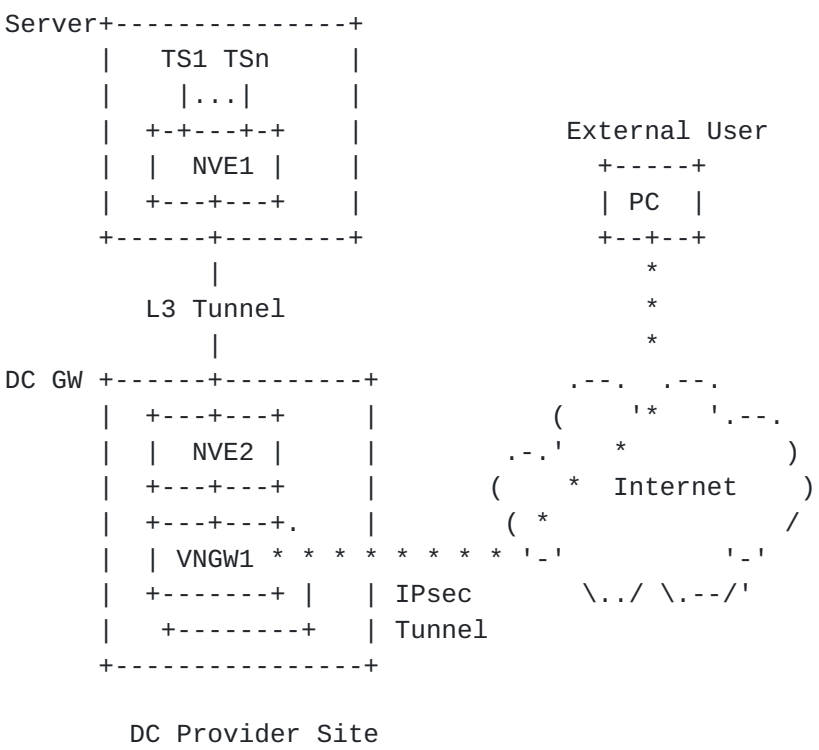


Figure 1 DC Virtual Network Access via Internet

3.2. DC VN and Enterprise Sites interconnected via SP WAN

An Enterprise company would lease some DC provider compute resources to run some applications. For example, the company may run its web applications at DC provider sites but run backend applications in their own DCs. The Web applications and backend applications need to communicate privately. DC provider may construct a NVO3 network to connect all VMs running the Enterprise Web applications. The enterprise company may buy a p2p private tunnel such as VPWS from a SP to interconnect its site and the NVO3 network in provider DC site. A protocol is necessary for exchanging the reachability between two peering points and the traffic are carried over the tunnel. If an enterprise has multiple sites, it may buy multiple p2p tunnels to form a mesh interconnection among the sites and DC provider site. This requires each site peering with all other sites for route distribution.

Another way to achieve multi-site interconnection is to use Service Provider (SP) VPN services, in which each site only peers with SP PE site. A DC Provider and VPN SP may build a NVO3 network (VN) and VPN independently. The VN provides the networking for all the related TSes within the provider DC. The VPN interconnects several enterprise sites, i.e. VPN sites. The DC provider and VPN SP further connect the VN and VPN at the DC GW/ASBR and SP PE/ASBR. Several options for the interconnection of the VN and VPN are described in RFC4364 [RFC4364]. In Option A with VRF-LITE [VRF-LITE], both DC GW and SP PE maintain the routing/forwarding table, and perform the table lookup in forwarding. In Option B, DC GW and SP PE do not maintain the forwarding table, it only maintains the VN and VPN identifier mapping, and exchange the identifier on the packet in the forwarding process. In option C, DC GW and SP PE use the same identifier for VN and VPN, and just perform the tunnel stitching, i.e. change the tunnel end points. Each option has pros/cons (see RFC4364) and has been deployed in SP networks depending on the applications. The BGP protocols may be used in these options for route distribution. Note that if the provider DC is the SP Data Center, the DC GW and PE in this case may be on one device.

This configuration allows the enterprise networks communicating to the tenant systems attached to the VN in a provider DC without interfering with DC provider underlying physical networks and other virtual networks in the DC. The enterprise may use its own address space on the tenant systems attached to the VN. The DC provider can manage the VMs and storage attachment to the VN for the enterprise customer. The enterprise customer can determine and run their applications on the VMs. See section 4 for more.

The interesting feature in this use case is that the VN and compute resource are managed by the DC provider. The DC operator can place them at any location without notifying the enterprise and WAN SP because the DC physical network is completely isolated from the carrier and enterprise network. Furthermore, the DC operator may move the VMs assigned to the enterprise from one sever to another in the DC without the enterprise customer awareness, i.e. no impact on the enterprise 'live' applications running these resources. Such advanced features bring DC providers great benefits in serving these kinds of applications but also add some requirements for NVO3 [NVO3PRBM].

4. DC Applications Using NVO3

NVO3 brings DC operators the flexibility in designing and deploying different applications in an end-to-end virtualization environment, where the operators not need worry about the constraints of the

physical network configuration in the Data Center. DC provider may use NVO3 in various ways and also use it in the conjunction with physical networks in DC for many reasons. This section highlights some use cases but not limits to.

4.1. Supporting Multi Technologies and Applications in a DC

Most likely servers deployed in a large data center are rolled in at different times and may have different capacities/features. Some servers may be virtualized, some may not; some may be equipped with virtual switches, some may not. For the ones equipped with hypervisor based virtual switches, some may support VxLAN [VXLAN] encapsulation, some may support NVGRE encapsulation [NVGRE], and some may not support any types of encapsulation. To construct a tenant virtual network among these servers and the ToR switches, it may construct one virtual network overlay and one virtual network w/o overlay, or two virtual networks overlay with different implementations. For example, one virtual network overlay uses VxLAN encapsulation and another virtual network w/o overlay uses traditional VLAN or another virtual network overlay uses NVGRE.

The gateway device or virtual gateway on a device may be used. The gateway participates in to both virtual networks. It performs the packet encapsulation/decapsulation and may also perform address mapping or translation, and etc.

A data center may be also constructed with multi-tier zones. Each zone has different access permissions and run different applications. For example, the three-tier zone design has a front zone (Web tier) with Web applications, a mid zone (application tier) with service applications such as payment and booking, and a back zone (database tier) with Data. External users are only able to communicate with the web application in the front zone. In this case, the communication between the zones MUST pass through the security GW/firewall. The network virtualization may be used in each zone. If individual zones use the different implementations, the GW needs to support these implementations as well.

4.2. Tenant Network with Multi-Subnets or across multi DCs

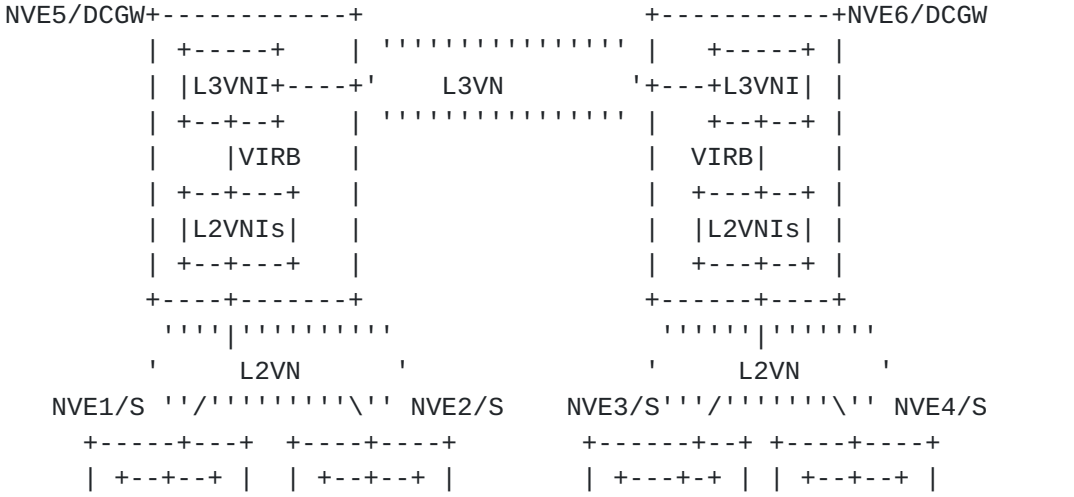
A tenant network may contain multiple subnets. DC operators may construct multiple tenant networks. The access policy for inter-subnets is often necessary. To benefit the policy management, the policies may be placed at some designated gateway devices only. Such design requires the inter-subnet traffic MUST be sent to one of the gateways first for the policy checking. However this may cause traffic hairpin on the gateway in a DC. It is desirable that an NVE

can hold some policy and be able to forward inter-subnet traffic directly. To reduce NVE burden, the hybrid design may be deployed, i.e. an NVE can perform forwarding for the selected inter-subnets and the designated GW performs for the rest. For example, each NVE performs inter-subnet forwarding for a tenant, and the designated GW is used for inter-subnet traffic from/to the different tenant networks.

A tenant network may span across multiple Data Centers in distance. DC operators may want an L2VN within each DC and L3VN between DCs for a tenant network. L2 bridging has the simplicity and endpoint awareness while L3 routing has advantages in policy based routing, aggregation, and scalability. For this configuration, the virtual L2/L3 gateway can be implemented on DC GW device. Figure 2 illustrates this configuration.

Figure 2 depicts two DC sites. The site A constructs an L2VN with NVE1, NVE2, and NVE3. NVE1 and NVE2 reside on the servers where the tenant systems are created. NVE3 resides on the DC GW device. The site Z has similar configuration with NVE3 and NVE4 on the servers and NVE6 on the DC GW. An L3VN is configured between the NVE5 at site A and the NVE6 at site Z. An internal Virtual Integrated Routing and Bridging (VIRB) is used between L2VNI and L3VNI on NVE5 and NVE6. The L2VNI is the MAC/NVE mapping table and the L3VNI is the IP prefix/NVE mapping table. A packet to the NVE5 from L2VN will be decapsulated and converted into an IP packet and then encapsulated and sent to the site Z.

Note that both the L2VNs and L3VN in Figure 2 are encapsulated and carried over within DC and across WAN networks, respectively.



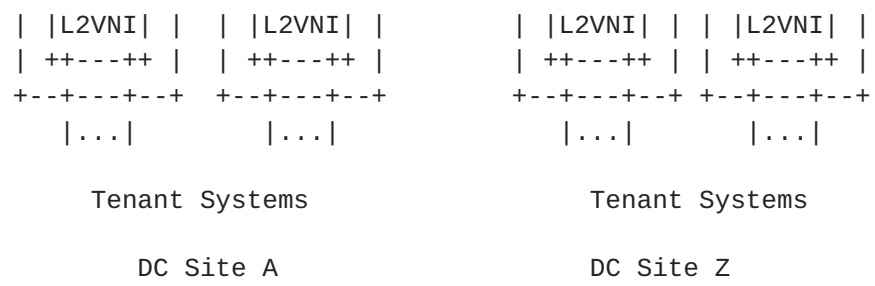


Figure 2 Tenant Virtual Network with Bridging/Routing

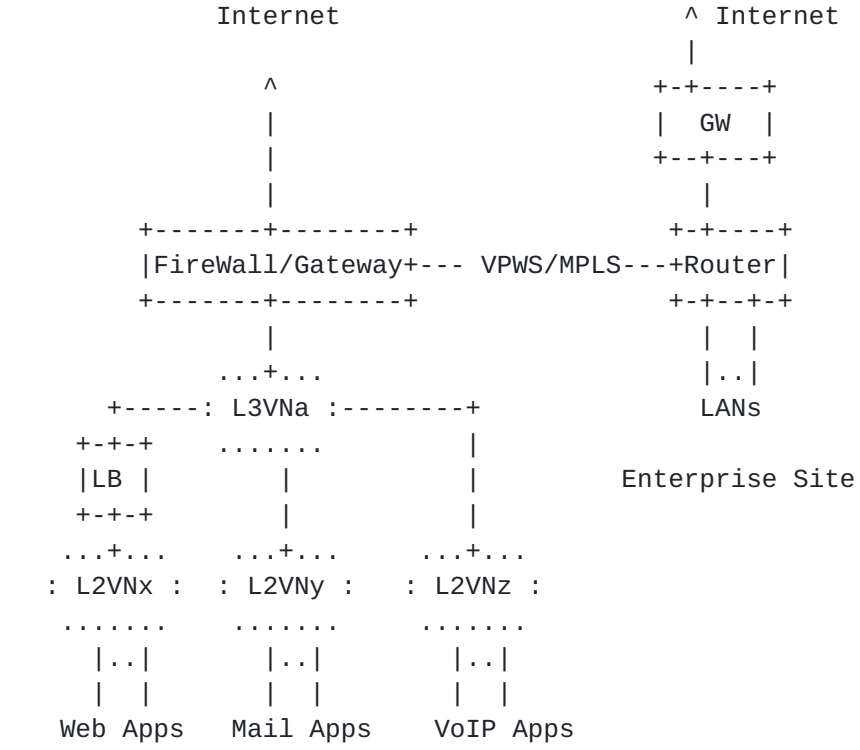
4.3. Virtual Data Center (vDC)

Enterprise DC's today may often use several routers, switches, and network appliance devices to construct its internal network, DMZ, and external network access. A DC Provider may offer a virtual DC service to an enterprise customer and run enterprise applications such as website/emails as well. Instead of using many hardware devices to do it, with the network virtualization overlay technology, DC operators may build such vDCs on top of a common network infrastructure for many such customers and run network service applications per a vDC basis. The net service applications such as firewall, DNS, load balancer can be designed per vDC. The network virtualization overlay further enables potential for vDC mobility when customer moves to different locations because tenant systems and net appliances configuration can be completely decouple from the infrastructure network.

Figure 3 below illustrates one scenario. For the simple illustration, it only shows the L3VN or L2VN as virtual and overlay routers or switches. In this case, DC operators construct several L2 VNs (L2VNx, L2VNy, L2VNz) in Figure 3 to group the end tenant systems together per application basis, create an L3VNa for the internal routing. A net device (may be a VM or server) runs firewall/gateway applications and connects to the L3VNa and Internet. A load Balancer (LB) is used in L2VNx. A VPWS p2p tunnel is also built between the gateway and enterprise router. The design runs Enterprise Web/Mail/Voice applications at the provider DC site; lets the users at Enterprise site to access the applications via the VPN tunnel and Internet via a gateway at the Enterprise site; let Internet users access the applications via the gateway in the provider DC.

The Enterprise customer decides which applications are accessed by intranet only and which by both intranet and extranet; DC operators then design and configure the proper security policy and gateway function. Furthermore DC operators may use multi-zones in a vDC for the security and/or set different QoS levels for the different applications based on customer applications.

This use case requires the NVO3 solution to provide the DC operator an easy way to create a VN and NVEs for any design and to quickly assign TSs to a VNI on a NVE they attach to, easily to set up virtual topology and place or configure policies on an NVE or VMs that run net services, and support VM mobility. Furthermore, DC operator needs to view the tenant network topology and know the tenant node capability and is able to configure a net service on the tenant node. DC provider may further let a tenant to manage the vDC itself.



firewall/gateway and Load Balancer (LB) may run on a server or VMs

Figure 3 Virtual Data Center by Using NVO3

5. OAM Considerations

NVO3 brings the ability for a DC provider to segregate tenant traffic. A DC provider needs to manage and maintain NVO3 instances. Similarly, the tenant needs to be informed about underlying network failures impacting tenant applications or the tenant network is able to detect both overlay and underlay network failures and builds some resiliency mechanisms.

Various OAM and SOAM tools and procedures are defined in [IEEE 802.1ag], [ITU-T Y.1731], [RFC4378], [RFC5880], [ITU-T Y.1564] for L2 and L3 networks, and for user, including continuity check, loopback, link trace, testing, alarms such as AIS/RDI, and on-demand and periodic measurements. These procedures may apply to tenant overlay networks and tenants not only for proactive maintenance, but also to ensure support of Service Level Agreements (SLAs).

As the tunnel traverses different networks, OAM messages need to be translated at the edge of each network to ensure end-to-end OAM.

6. Summary

The document describes some general potential use cases of NVO3 in DCs. The combination of these cases should give operators flexibility and capability to design more sophisticated cases for various purposes.

DC services may vary from infrastructure as a service (IaaS), platform as a service (PaaS), to software as a service (SaaS), in which the network virtualization overlay is just a portion of an application service. NVO3 decouples the service construction/configurations from the DC network infrastructure configuration, and helps deployment of higher level services over the application.

NVO3's underlying network provides the tunneling between NVEs so that two NVEs appear as one hop to each other. Many tunneling technologies can serve this function. The tunneling may in turn be tunneled over other intermediate tunnels over the Internet or other WANs. It is also possible that intra DC and inter DC tunnels are stitched together to form an end-to-end tunnel between two NVEs.

A DC virtual network may be accessed via an external network in a secure way. Many existing technologies can help achieve this.

NVO3 implementation may vary. Some DC operators prefer to use centralized controller to manage tenant system reachbility in a tenant network, other prefer to use distributed protocols to advertise the tenant system location, i.e. attached NVEs. For the migration and special requirement, the different solutions may apply to one tenant network in a DC. When a tenant network spans across multiple DCs and WANs, each network administration domain may use different methods to distribute the tenant system locations. Both control plane and data plane interworking are necessary.

7. Security Considerations

Security is a concern. DC operators need to provide a tenant a secured virtual network, which means one tenant's traffic isolated from the other tenant's traffic and non-tenant's traffic; they also need to prevent DC underlying network from any tenant application attacking through the tenant virtual network or one tenant application attacking another tenant application via DC networks. For example, a tenant application attempts to generate a large volume of traffic to overload DC underlying network. The NVO3 solution has to address these issues.

8. IANA Considerations

This document does not request any action from IANA.

9. Acknowledgements

Authors like to thank Sue Hares, Young Lee, David Black, Pedro Marques, Mike McBride, David McDysan, Randy Bush, and Uma Chunduri for the review, comments, and suggestions.

10. References

10.1. Normative References

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

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

[IEEE 802.1ag] "Virtual Bridged Local Area Networks - Amendment 5: Connectivity Fault Management", December 2007.

[ITU-T G.8013/Y.1731] OAM Functions and Mechanisms for Ethernet based Networks, 2011.

[ITU-T Y.1564] "Ethernet service activation test methodology", 2011.

[RFC4378] Allan, D., Nadeau, T., "A Framework for Multi-Protocol Label Switching (MPLS) Operations and Management (OAM)", RFC4378, February 2006

[RFC4301] Kent, S., "Security Architecture for the Internet Protocol", rfc4301, December 2005

[RFC5880] Katz, D. and Ward, D., "Bidirectional Forwarding Detection (BFD)", rfc5880, June 2010.

10.2. Informative References

[NVGRE] Sridharan, M., "NVGRE: Network Virtualization using Generic Routing Encapsulation", draft-sridharan-virtualization-nvgre-02, work in progress.

[NVO3PRBM] Narten, T., etc "Problem Statement: Overlays for Network Virtualization", draft-ietf-nvo3-overlay-problem-statement-02, work in progress.

[NVO3FRWK] Lasserre, M., Motin, T., and etc, "Framework for DC Network Virtualization", draft-ietf-nvo3-framework-02, work in progress.

[NVO3MCAST] Ghanwani, A., "Multicast Issues in Networks Using NVO3", draft-ghanwani-nvo3-mcast-issues-00, work in progress.

[VRF-LITE] Cisco, "Configuring VRF-lite", http://www.cisco.com

[VXLAN] Mahalingam,M., Dutt, D., etc "VXLAN: A Framework for Overlaying Virtualized Layer 2 Networks over Layer 3 Networks", draft-mahalingam-dutt-dcops-vxlan-03.txt, work in progress.

Authors' Addresses

Lucy Yong
Huawei Technologies,
5340 Legacy Dr.
Plano, TX 75025

Phone: +1-469-277-5837
Email: lucy.yong@huawei.com

Mehmet Toy

Comcast
1800 Bishops Gate Blvd.,
Mount Laurel, NJ 08054

Phone : +1-856-792-2801
E-mail : mehmet_toy@cable.comcast.com

Aldrin Isaac
Bloomberg
E-mail: aldrin.isaac@gmail.com

Vishwas Manral
Hewlett-Packard Corp.
3000 Hanover Street, Building 20C
Palo Alto, CA 95014

Phone: 650-857-5501
Email: vishwas.manral@hp.com

Linda Dunbar
Huawei Technologies,
5340 Legacy Dr.
Plano, TX 75025 US

Phone: +1-469-277-5840
Email: linda.dunbar@huawei.com