

Network working group
Internet Draft
Category: Informational

L. Yong
L. Dunbar
Huawei

Expires: March 2013

December 11, 2012

NV03 Framework and Data Plane Requirement Addition
draft-yong-nvo3-frwk-dpreq-addition-00

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 March 11, 2013.

Copyright Notice

Copyright (c) 2009 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.

Internet-Draft

FRWK and DP Req. Addition

December 2012

Abstract

This document describes some additional functions and requirements for NV03 framework [[NV03FRWK](#)] and data plane requirements [[DPREQ](#)]. These additions are necessary in supporting VM communication and mobility.

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.....	3
2.	New NVE Service Type.....	3
2.1.	Why a New NVE Service Type.....	3
2.2.	L2-3 NVE Providing IP Routing/Bridging-like Service (Framework Addition).....	4
2.3.	L2-3 VNI (Data Plane Requirement Addition).....	4
3.	Tenant System Mobility.....	5
3.1.	Background.....	5
3.2.	NVE Functions for TS Mobility (Framework Addition).....	6
3.2.1.	Tenant System Mobility.....	6
3.2.2.	Tenant Multicast Traffic.....	6
3.2.3.	The Policy Associated With TS.....	7
3.3.	Tenant System Mobility (Data Plane Requirement Addition)..	7
4.	Security Considerations.....	8
5.	IANA Considerations.....	8
6.	Acknowledgements.....	8
7.	References.....	8
7.1.	Normative References.....	8
7.2.	Informative References.....	8

1. Introduction

NV03 framework [[NVO3FRWK](#)] and data plane requirement [[DPREQ](#)] documents specify the network virtualization overlay framework and data plane requirements, which aims on an architecture to support the network virtualization overlay in DC [[NVO3PRBM](#)]. The main application of NV03 is to support multi-tenant networks on a common infrastructure, where a tenant virtual network may contain one or more subnets [[HYPERV](#)]. However, current framework specifies two NVE service types. Neither of them naturally supports the communication among the VMs when some VMs are on the same subnet and other on different. Second, one of the key aspects of NV03 is to support the virtual machines (VM) mobility. However, neither document mentions VM mobility nor specifies any function and/or requirements in supporting VM mobility. This document addresses the two additions.

To use the terminologies specified in the framework document, this document refers VM mobility as to Tenant System mobility or TS mobility.

2. New NVE Service Type

2.1. Why a New NVE Service Type

A virtual machine on a server behaves like a physical server to an application or guest OS on it. This means that any frame from/to a virtual machine is an Ethernet frame, just as a frame from/to a physical server.

L2 NVE service type specified in the framework [[NVO3FRWK](#)] provides Ethernet LAN like service where multiple Tenant Systems appear to interconnected by an LAN environment over a set of L3 tunnels. However, from the host (physical servers or VMs) perspective, only the hosts on the same subnet can communicate in an LAN network. This implies that L2 NVE service type only applies to a single subnet.

L3 NVE service type [[NVO3FRWK](#)] provides a virtualized IP routing and forwarding like IETF IP VPN. The IP VPN emulates a route domain and

provides forwarding and routing among TSes that are the same and/or different subnets. IETF IP VPN has the assumption that the Layer 3 MUST be implemented between a PE and a CE, which means between an NVE and a TS in this context. This assumption does not fit to the case where an NVE attached by the multiple TSes that are on the same subnet where the TSes uses bridging mechanism for the communication.

To support TSes, regardless on the same or different subnets, communicating in an L2 environment, this document suggests adding a new L2-3 NVE Service Type. Suggested Text for the framework and data plane requirement documents is in [section 2.2](#) and [section 2.3](#), respectively.

2.2. L2-3 NVE Providing IP Routing/Bridging-like Service (Framework Addition)

L2-3 NVE is similar to IRB function on a router [[CIRB](#)] device today. It supports the TSes attached to the NVE (locally or remotely) to communicate with each other when they are in a same route domain, i.e. a tenant virtual network. The NVE provides per tenant virtual switching and routing instance with address isolation and L3 tunnel encapsulation across the core. The L2-3 NVE supports the bridging among TSes that are on the same subnet and the routing among TSes that are on the different subnets.

2.3. L2-3 VNI (Data Plane Requirement Addition)

L2-3 VNIs MUST provide virtualized IP routing and bridging. L2-3 VNI MUST support per-tenant forwarding instance with IP and MAC address isolation and L3 tunneling for interconnecting instances of the same VNI on NVEs. L2-3 VNI MUST perform the virtual bridging for the Tenant Systems that are on the same subnet and the IP routing for the Tenant Systems that are on the different subnets. L2-3 VNI MUST support L2/3 gateway function.

L2-3 VNI MUST NOT change Tenant System communication mechanism in a route domain, i.e. a tenant virtual network, and not violate Tenant Systems communication rules. Tenant System communication rules are if Tenant Systems are on the same subnet, they are bridged directly; if Tenant Systems are on different subnets, they MUST communicate through a router. A tenant system uses the ARP/ND protocol to

discover other tenant system MAC addresses if they are on the same subnet; a tenant system sends a packet to a known gateway if the destination of the packet is on different subnet from the sender TS; a tenant system uses ARP/ND protocol to find the gateway MAC address.

Forwarding table entries provide mapping information between MAC/IP and L3 Tunnel destination addresses. Such entries MAY be populated by a control or management plane.

The L2-3 VNI MUST support the ARP protocol at virtual access points (VAPs) and a default VGW MAC address.

In the case of L2-3 VNI, when the packet is forwarded from one subnet to another subnet, inner TTL field and outer TTL field process MUST be the same as described in L3 VNI section.

When tenant multicast is supported, L2-3 VNI SHOULD also be possible to select whether the NVE provides optimized multicast trees inside the VNI for individual tenant multicast groups or whether the default VNI multicast tree is used, where all the NVEs of the corresponding VNI are members, is used.

[3. Tenant System Mobility](#)

3.1. Background

NV03 generic reference model specifies that a Tenant System can be attached to an NVE locally or remotely. The local means that a TS and the NVE are resident in the same device, e.g. server. The remote means a TS attached to the NVE via a point-to-point connection or a switched network, e.g. Ethernet.

When an NVE is local, the state of Tenant System can be provided without protocol assistance. This implies that when Tenant System state changes, the NVE is immediately aware of the changes. When an NVE is remote, the state of the Tenant System needs to be exchanged via a data or control plane protocol, or via a management entity.

VM mobility further requires support of hot and cold move [[VMMOVE](#)]. In the hot move, the moving is seamless to the application that runs on the moved TS, which implies that the existing connectivity

between the moved TS and other TSes that the moved TS communicates with MUST be maintained while the TS is moved regardless if these TSes are on the same or different subnets.

When a TS and NVE are resident in the same device, the TS moves from one NVE, NVE1 to another NVE, NVE2. NVE2 instantly knows the TS address, state, and etc. However other NVEs that other TSes attach to and have the connectivity with the moved TS MUST be also aware of the TS new location, i.e. NVE2 location, and NVE2 MUST be also aware of these NVE locations in order to maintain the connectivity.

When a TS and NVE are remotely attached, TS moving only applies when a TS attaches to the NVE via a switched network, i.e. L2 physical and/or virtual network. [\[VMMOVE\]](#) In addition of the actions in the local NVE case (mentioned above), when an NVE is remote, the state of the Tenant System needs to be exchanged via a data or control plane protocol, or via a management entity.

Other two cases are a TS moved away from a local NVE and to a remote NVE and vice versa.

To support TS mobility, this document suggests adding a new section in the NV03 framework and data plane requirement documents and the suggested text is in [section 3.2](#) and 3.3, respectively.

3.2. NVE Functions for TS Mobility (Framework Addition)

3.2.1. Tenant System Mobility

If an NVE (say ingress NVE) is responsible to notify other NVEs (egress NVEs) regarding a new moved TS attaching to it. If the ingress NVE is not yet on the tenant virtual network that the moved TS belongs to, the NVE MUST establish the membership to the virtual network first and create a virtual access point (VAP) to associate to the virtual network. The NVE MUST send a notification about the TS to other egress NVEs that has the same membership. This can be done via data plane or control plane. Upon receiving the notification from an ingress NVE, an egress NVE has to update its VNIs that are associate to the same membership. If an NVE is remote, the VNI MUST send the new TS address notification to the access networks via the virtual access points (VAPs).

Note that if the ingress NVE is L2-3 NVE, and if it is not yet on the same tenant virtual network subnet as the moved TS belongs to, the NVE MUST establish the membership to the virtual subnet network first and create a VAP to associate with it. The NVE MUST send a notification about the TS to other egress NVEs that has the same membership. If an NVE is remote, they MUST only send the notification to the access networks that are on the same tenant virtual subnet as the moved TS is on.

Note that, when a TS moves away from an NVE and it is the last TS attached to the NVE belong to the tenant virtual network, the NVE MAY delete the membership of the tenant virtual network.

3.2.2. Tenant Multicast Traffic

If a tenant application on a set of TSes needs to send broadcast or multicast traffic among them, the NVE multicast and broadcast capability can facilitate such forwarding [[NVO3FRWK](#)]. To support VM mobility, when one of the TSes is moved from one NVE (say NVE1) to another (say NVE2) in hot mode, the NVE2 has to know which multicast groups that the TS is associated with. If NVE2 is local, such information can be available to the NVE2 via some API; if NVE2 is remote, such information can be available to the NVE2 via data plane,

control plane, or management entity [[NVO3FRWK](#)]. If NVE2 is need to learn Tenant Multicast Groups that a moved TS is on, the NVEs MUST be able to send a query message to the moved TS; The TS response which groups it is on.

Once NVE2 knows which multicast groups that the new attached TS is associated with, NVE2 MUST bind itself to these multicast groups if it is not on yet. Furthermore, NVE2 MAY (if not yet) have to bind the overlay multicast groups to one or more underlying multicast tree if it uses the underlay multicast trees to delivery overlay multicast traffic. An NVE MUST provide these capabilities, if it supports tenant multicast traffic, to ensure tenant application seamlessly running while a Tenant System is moved.

Similarly, when NVE1 knows a TS moved away and being the last one on the tenant virtual network, NVE1 MAY unbind itself to the corresponding multicast group. Furthermore, if this is the last multicast group on the NVE1, NVE1 MAY unbind the multicast group to the shared multicast tree if used.

3.2.3. The Policy Associated With TS

An NVE provides the policy based forwarding and routing [[NVO3FRWK](#)] [[DPREQ](#)]. When a TS is moved from one NVE (say NVE1) to another (say NVE2), the NVE2 has to apply to the same set of policy to the TS as well. If TS related policies are specified in the TS service profile that is moved along with the TS, and the file can be passed to the NVE2 via API if the TS is locally attached to or via a data plane or control plane protocol, or a management entity if remotely attached to. An NVE MUST be able to automatically install these policies at the VAP that a new TS attaches to. NVE1 MUST automatically delete the policies that are applied to the moved TS only.

3.3. Tenant System Mobility (Data Plane Requirement Addition)

If the data plane learning is used to populate the forwarding table[[DPREQ](#)], an NVE (local or remote) MUST be able to send a notification message to all the NVEs that are the membership of the tenant virtual network that the TS belongs to, e.g. ARP gratuitous message. The notification MUST contain the TS address and tenant VN ID. Upon receiving the notification message, an NVE MUST update the corresponding VNI indicated in the NV03 overlay header. If the receiving NVE is remote, the NVE MUST send a notification to the local access networks that is on the same subnet as of one indicated in the NV03 overlay header via the VAPs.

[4.](#) Security Considerations

When a Tenant System is moved from one NVE to another, automatic virtual network membership creation on an NVE may leave some security concern. Either certain authentication is needed for an NVE to accept a new TS or management entity assisted process is used to ensure the security.

Supporting TS mobility brings a new challenge for NV03 is discussed in [[NVO3PRBM](#)].

[5.](#) IANA Considerations

The document does not require any IANA action.

[6.](#) Acknowledgements

Thank Weiguo Hao for the review and input to the draft.

[7.](#) References

7.1. Normative References

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

7.2. Informative References

[DPREQ] Bitar, N., and etc, "NVO3 Data Plane Requirement", [draft-bl-nvo3-dataplane-requirements-03.txt](#), November 2012

[CIRB] Cisco, "Understanding and Configuring VLAN Routing and Bridging on a Router Using the IRB Feature", Doc. ID 17054

[HYPERV] Microsoft, "Hyper-V Network Virtualization Packet Flow", September 2012

[NVO3FRWK] LASSERRE, M., Motin, T., and etc, "Framework for DC Network Virtualization", [draft-ietf-nvo3-framework-01](#), October 2012

[NVO3PRBM] Narten, T., and etc "Problem Statement: Overlays for Network Virtualization", [draft-ietf-nvo3-overlay-problem-statement-01](#), October 2012

[VMMOVE] Rakhter, Y., and etc, "Network-related VM Mobility Issue", [draft-rekhter-nvo3-vm-mobility-issues-03.txt](#), Sept. 2012

Authors' Addresses

Lucy Yong
Huawei USA
5340 Legacy Drive
Plano, TX 75025
U.S.A

Phone: 469-277-5837

Email: lucy.yong@huawei.com

Linda Dunbar
Huawei USA
5340 Legacy Drive
Plano, TX 75025
U.S.A

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