INTERNET-DRAFT
Intended Status: Standards Track

Mingui Zhang
Yizhou Li
Huawei
Muhammad Durrani
Cisco
Margaret Cullen
Painless Security

Expires: April 18, 2016
October 16, 2015

Multi-homing of NVEs
draft-zhang-nvo3-active-active-nve-01.txt

Abstract

Multi-homing of NVEs is a desirable feature to NVO3 since it provides Tenant Systems with reliable connection, load-balance, etc. The major issue of the multi-homing of NVEs is that it introduces one-to-many mapping from Tenant System (inner) addresses to NVE (outer) addresses.

This document identifies the requirements for multi-homing of NVEs and specifies corresponding solutions to meet these requirements. The key is to let remote NVEs be aware of the group of NVEs that are offering multi-homing to Tenant Systems, therefore the remote NVE can cope with the one-to-many mapping correctly.

Table of Contents

1. Introduction

At the edge of the NOV3 network, a Tenant System (TS) can be attached to multiple NVEs or a single NVE that uses more than one underlay IP addresses. For the latter case, a typical implementation is that a NVE is located within a physical server with multiple Network Interface Cards (NICs) while each NIC is assigned a unique underlay IP address. Both above cases are considered as multi-homing of NVEs from the tenant system's point of view [I-D.ietf-nvo3-arch]. When this TS communicates with another TS attached to a remote NVE, this remote NVE will see one TS address being mapped to multiple NVE addresses.

In this document, the support of one-to-many mapping is achieved by extending either the management plane or the control plane of NVO3. The NVEs offering multi-homing form an Active Active NVE (AANVE) group and this group is explicitly identified. The remote NVE recognizes the AANVE group and installs into the mapping table an one-to-one mapping while keeps the one-to-many mapping at the management plane or control plane.

It's possible to extend the data plane to fix the one-to-many mapping issue. For example, an identifier of the AANVE group can be added to the NVO3 header. However, such an extension requires new hardware. This kind of solutions are interoperable with the solution defined in this document, but they are out the scope of this document.

The rest of the document is organized as follows. Section 2 defines the Acronyms and Terminology that will be used in this document. Section 3 describes the two scenarios of the multi-homing of NVEs in NVO3 networks. Section 4 lists the requirements and specifies the corresponding solutions. Security Considerations, IANA Considerations and References are given in the rest sections.

2. Acronyms and Terminology

2.1. Acronyms and Terms

NVO3: Network Virtualization Overlays
NVE: Network Virtualization Edge
VNI: Virtual Network Instance
VAP: Virtual Access Points
AANVE: Active-Active NVEs. The AANVE denotes the group of NVEs that offer active-active multi-homing VAPs to Tenant Systems.

2.2. Terminology

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].

Familiarity with [RFC7364], [RFC7365], [I-D.ietf-nvo3-arch], [I-D.ietf-nvo3-dataplane-requirements] and [RFC7348] is assumed in this document.

3. Scenarios for Multi-Homing of NVEs

According to [I-D.ietf-nvo3-arch], for the multi-homing of NVEs, one or multiple NVEs use more than one underlay IP addresses to encapsulate frames from a specific attached Tenant System. For the return traffic, the TS can be reached via any of these IP addresses.

Section 3.1 and Section 3.2 detail the two scenarios of multi-homing of NVEs as defined in [I-D.ietf-nvo3-arch]. A mixture of these two scenarios in a practical network is possible and allowable.

3.1. Attached to One NVE with Multiple IP Addresses

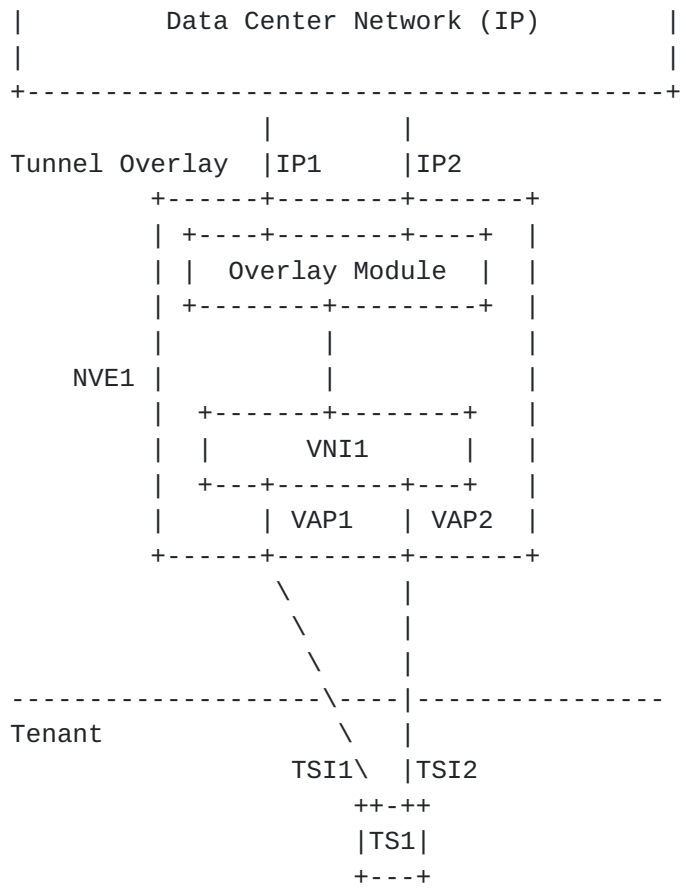


Figure 3.1: The NVE uses multiple underlay IP addresses for one TS.

In this scenario, the TS is connected to a single NVE. However, this NVE owns multiple underlay IP addresses and any of these IP addresses might be used, in an active-active way, for the TS. The multi-path of the underlay can be used to achieve load-balance.

The NVE can either be deployed on the TOR or co-located with the TS on a server. When the NVE is deployed on the TOR, the TS might be connected to a bridge which is in turn connected to the NVE using Link Aggregation; it might also be directly connected to the NVE with NIC teaming.

3.2. Attached to Multiple NVEs

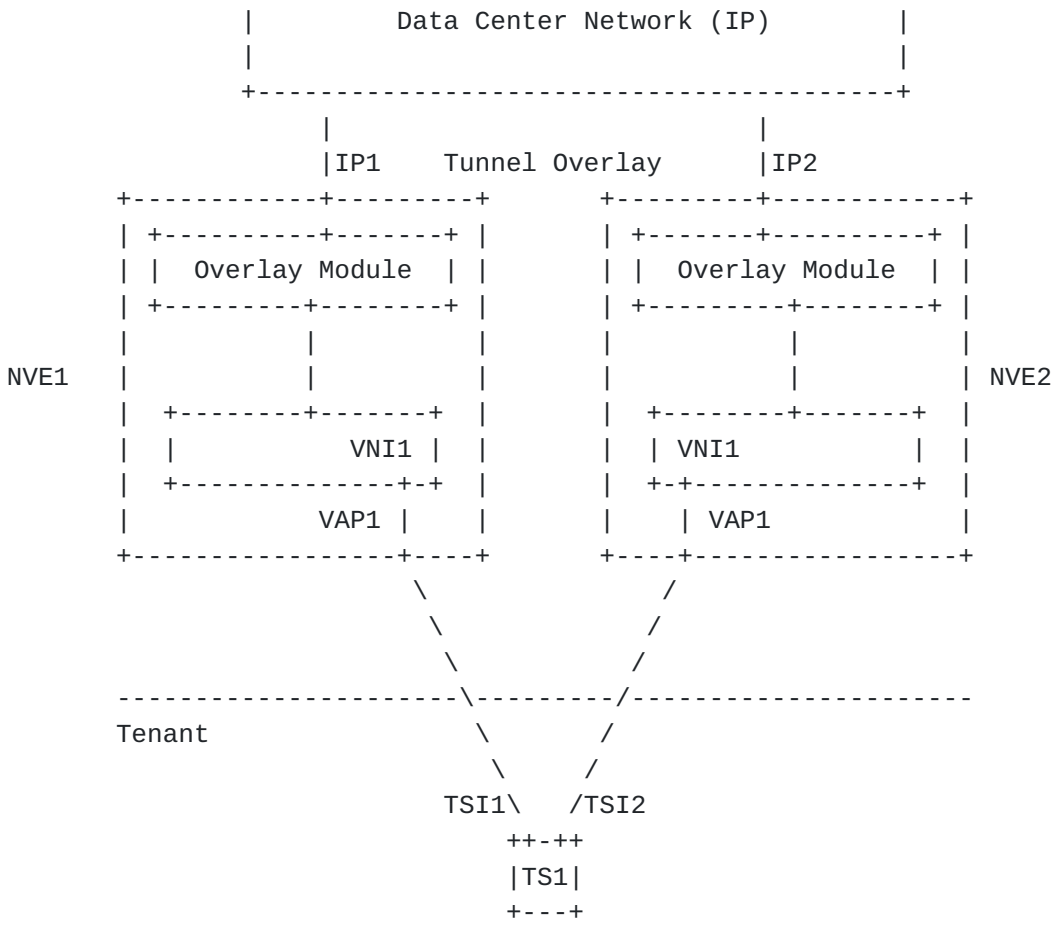


Figure 3.2: The TS is attached to multiple NVEs.

In this scenario, the TS is attached to multiple NVEs which have their own underlay IP addresses. These NVEs work in an active-active mode for this TS.

There is no reason to have the NVEs and the TS co-located on the same server, since there would be a better choice to implement multiple

NVEs as a single one in that case. It would be normal to have these NVEs deployed on TORs and it's possible that these NVEs are deployed in a single TOR. The TS may directly use NIC teaming to connect to these NVEs. Also, it may be connected to a bridge and this bridge is connected to these NVEs using MC-LAG or DRNI [802.1AX].

4. Requirements and Solutions

The key issue of multi-homing of NVEs is that it introduces one-to-many mapping. However, an one-to-many mapping does not have to be installed into the NVE address mapping table. Instead, this one-to-many mapping may be kept at the management plane or control plane.

Explicit identification of the multi-homing is the key to solve the one-to-many mapping issue. With this identification, remote NVEs are able to recognize the NVEs participating an AANVE. Only one of those multiple underlay IP addresses will be installed into the NVE address mapping table.

This section describes the requirements of the multi-homing of NVEs and specifies the corresponding solutions.

4.1. No Flip-Flop

Since traffic from one TS may be observed, by a remote NVE, coming from more than one NVE addresses. It is REQUIRED that this NVE does not flip-flop among these NVE addresses. Otherwise, the TS will see packet disorder for the returning traffic.

The remote NVE maintains a per-VN table (data plane) of mappings from Tenant System (inner) addresses to NVE (outer) addresses (See Section 4.2 of [I-D.ietf-nvo3-arch]). It is REQUIRED that this remote NVE keeps in this table a unique outer address per inner address.

From the control plane or management plane [I-D.zhang-nvo3-yang-active-active-cfg], the remote NVE learns the set of NVE addresses in an AANVE. It locally selects one out of this address set to be kept in the mapping table. It is RECOMMENDED that the one with the least-cost path to the remote NVE is selected.

Note that the NVE address installed into the mapping table does not have to be equal to the one carried in the packet. Take the scenario in Section 3.2 as an example, even if NVE1 encapsulates the frame from TS1 and sends the packet to the remote NVE, the remote NVE may install NVE2's IP address into the mapping table. In that case, the return traffic will be sent to NVE2.

4.2. No Duplication

No duplication will happen to packets ingressed by the AANVE and unicast packets egressed by AANVE. Nevertheless, if more than one NVEs out of the AANVE group egress a copy of a multicast packet, the TS will see packet duplication. Therefore, it is REQUIRED that a unique NVE is appointed as the multicast egress NVE. This unique multicast egress NVE can appointed by configuration as specified in [I-D.zhang-nvo3-yang-active-active-cfg], where the NVE with the highest priority is appointed. It can also be appointed according to a selection method agreed on by all NVEs of the AANVE group [802.1AX].

The remote NVE may send serial unicast packets instead of multicast packets to each member of a multicast group. For this special case, the remote NVE MUST not send more than one copy to a AANVE group.

4.3. No Echo

If a multicast packet is ingressed by one NVE out of an AANVE group, forwarded across the network, and then received by another NVE in the same group, it is important that the second NVE does not egress this packet. Otherwise, a forwarding loop will occur. Hence, the NVE in an AANVE group MUST NOT egress a multicast packet which ingressed by any other NVE in the same AANVE group.

For the case that the NVE in an AANVE group uses serial unicast instead of multicast, there is not reason for this NVE to send a copy to another NVE in the same AANVE since the TS attached to the AANVE already has copy.

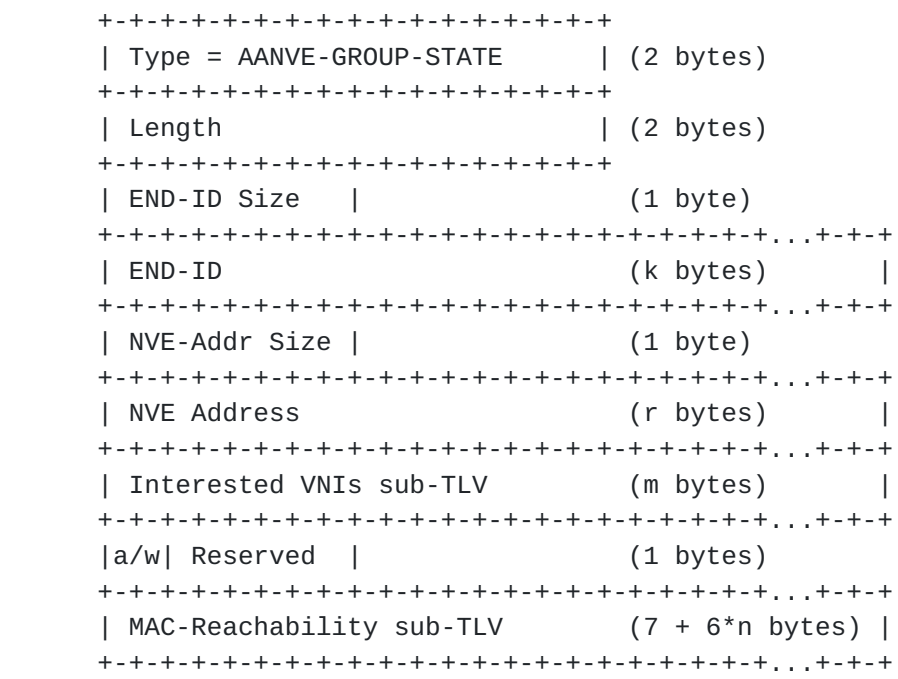
4.4. No Black-hole

If an NVE senses that the VAP connecting to the TS fails as defined in Section 4.5 of [I-D.ietf-nvo3-arch], this NVE MUST notify remote NVEs. Otherwise, packets to the TS which is attached to this NVE will be subject to loss (a.k.a. black-hole).

The state of the AANVE group MUST be refreshed immediately and reflected in the configuration or a control message (See Section 4.6). When the remote NVE obtains the new state of the AANVE group, it MUST update its local mapping table immediately. Usually this update means the withdraw of a batch of TS addresses.

4.5. Discovery of Multi-Homing NVEs

NVEs in an AANVE group SHOULD be either be discovered through configuration [I-D.zhang-nvo3-yang-active-active-cfg] or through exchanging of the control message specified as follows.



- Type: AANVE Group State TLV (NVO3 APPsub-TLV type tbd1)

- Length: The total bytes of the AANVE Group State not including the Type and Length fields.

- END-ID Size: The length k of the END-ID in bytes.

- END-ID: The ID of the active-actively connected end device [RFC7365] which is k bytes long. This ID identifies the AANVE group. If the LAALP is an MC-LAG or DRNI, it is the 8-byte ID specified in Clause 6.3.2 in [802.1AX].

- NVE-Addr Size: The length r of the NVE Address in bytes.

- NVE Address: The NVE address that the originating NVE is used for the AANVE group. This field is useful because the originating NVE might own multiple underlay IP addresses.

- Interested VNIs sub-TLV: This sub-TLV specifies the VNIs that the originating NVE is interested in. In these VNIs multi-homing will be offered by the originating NVE. The VNIs may be given as any form of the following.

  (1) VNIs listed individually;
  (2) Ranges with Upper/Lower bounds;
  (3) One or multiple bit-maps of VNIs.

- a/w: Append or withdraw. The values of these two bits indicates the following actions on the MAC addresses given by the subsequent MAC-Reachability sub-TLV.

  0 - null MAC addresses

  The MAC-Reachability sub-TLV in the AANVE Group State TLV MUST be ignored.

  1 - listed MAC addresses withdraw

  The MAC addresses listed by the MAC-Reachability sub-TLV MUST be withdrawn.

  2 - append listed MAC addresses

  The MAC addresses listed by the MAC-Reachability sub-TLV SHOULD be installed to the mapping table by the receiving NVE.

  3 - withdraw all MAC addresses

  Within the interested VNIs listed by the Interested VNIs sub-TLV, all MAC addresses attached to the originating NVE MUST be withdrawn.

- MAC-Reachability sub-TLV: The AANVE Group State TLV value contains the MAC-Reachability TLV defined in [RFC6165] as a sub-TLV.

- Reserved: Flags reserved for future use. These MUST be sent as zero and ignored on receipt.

5. Security Considerations

This document raises no new security issues.

6. IANA Considerations

IANA is requested to allocated the IS-IS Application Identifier (tbd2) under the Generic Information TLV (#251) [RFC6823] for NVO3. Then, IANA is requested to assign a type (tbd1) under this NVO3 GENINFO TLV for the NVO3 APPsub-TLV specified in Section 4.5.

7. References

7.1. Normative References

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

[802.1AX] IEEE, "IEEE Standard for Local and Metropolitan Area Networks - Link Aggregation", 802.1AX-2014, 24 December 2014.

[I-D.zhang-nvo3-yang-active-active] M. Zhang, J. Xia, "YANG Data Model for NVO3 Active Active Access", draft-zhang-nvo3-yang-active-active, working in progress.

[RFC6165] Banerjee, A. and D. Ward, "Extensions to IS-IS for Layer-2 Systems", RFC 6165, April 2011.

[RFC6823] Ginsberg, L., Previdi, S., and M. Shand, "Advertising Generic Information in IS-IS", RFC 6823, December 2012.

7.2. Informative References

[I-D.ietf-nvo3-arch] D. Black, J. Hudson, et al, "An Architecture for Overlay Networks (NVO3)", draft-ietf-nvo3-arch, work in progress.

[RFC7364] Narten, T., Ed., Gray, E., Ed., Black, D., Fang, L., Kreeger, L., and M. Napierala, "Problem Statement: Overlays for Network Virtualization", RFC 7364, October 2014.

[RFC7365] Lasserre, M., Balus, F., Morin, T., Bitar, N., and Y. Rekhter, "Framework for Data Center (DC) Network Virtualization", RFC 7365, October 2014.

[RFC7348] Mahalingam, M., Dutt, D., Duda, K., Agarwal, P., Kreeger, L., Sridhar, T., Bursell, M., and C. Wright, "Virtual eXtensible Local Area Network (VXLAN): A Framework for Overlaying Virtualized Layer 2 Networks over Layer 3 Networks", RFC 7348, August 2014.

[I-D.ietf-nvo3-dataplane-requirements] Nabil Bitar, Marc Lasserre, et al, "NVO3 Data Plane Requirements", draft-ietf-nvo3-dataplane-requirements, working in progress.

Author's Addresses

Mingui Zhang
Huawei Technologies
No. 156 Beiqing Rd. Haidian District,
Beijing 100095
China

EMail: zhangmingui@huawei.com

Yizhou Li
Huawei Technologies
101 Software Avenue,
Nanjing 210012
China

Phone: +86-25-56625409
EMail: liyizhou@huawei.com

Muhammad Durrani
Cisco Systems, Inc.
170 West Tasman Dr.
San Jose, CA 95134
USA

EMail: mdurrani@cisco.com

Margaret Cullen
Painless Security
14 Summer St. Suite 202
Malden, MA 02148 USA

EMail: margaret@painless-security.com