Network Virtualization Overlays Working Group
Internet-Draft
Intended status: Standards Track
Expires: December 30, 2013

L. Xia
Q. Wu
Huawei

June 28, 2013

Tenant system information discovery approaches Gap analysis
draft-wu-nvo3-mac-learning-arp-03

Abstract

This document analyzes various protocol solutions for tenant system information (e.g. MAC, IP, etc) discovery in the virtualization environment (e.g.,MAC in MAC, MAC in IP, IP in IP) and identifies the gap against NVO3 control plane and data plane requirements.

Status of this Memo

Copyright Notice

Table of Contents

1. Introduction

The tenant system information in this document is referred to as L2 address and L3 address of VM. As described in [I.D-ietf-nvo3-framework], for an L2 NVE, the NVE needs to be able to determine MAC addresses of the tenant system. For an L3 NVE, the NVE needs to be able to determine IP addresses of the tenant system.

This can be achieved mainly in 3 ways: data plane learning; ARP; control plane distribution (e.g. by BGP or IS-IS). This document analyzes various protocol solutions for tenant system information (e.g. MAC, IP, etc) discovery in the virtualization environment (e.g.,MAC in MAC, MAC in IP, IP in IP) and identifies the gap against NVO3 control plane and data plane requirements.

2. 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 RFC2119 [RFC2119].

3. Overview of tenant system information discovery in the virtualization domain using NVO3

Tenant system information discovery can be achieved either using dynamic data plane learning or ARP or control plane distribution. This document addresses how tenant system information discovery works in the overlay network enviroment. Figure 1 shows the NVO3 reference architecture for tenant system information discovery. The reference architecture assumes that:

- Tenant system A in DC site X wants to establish communication with tenant system B in the DC site Y.

- Tenant system A is connecting to VN by attaching to NVE X. Tenant System A knows IP address of Tenant System B using out of band means but does not know MAC address of Tenant System B.

- Tenant system B is connecting to VN by attaching to NVE Y. Tenant System B knows IP address of Tenant System A using out of band means but does not know MAC address of Tenant System A.

- NVE X associated with tenant system A doesn't know IP address and MAC address of tenant system B.

- NVE Y associated with tenant system B doesn't know IP address and MAC address of tenant system A.

```
                               ,---------.
                             ,'  Backend  `.
                            (      NVA       )
                             `.           ,'
                               `-+------+'
                               |         |
                          .--.,--. .--. ..
                         (    '           '.--.
                      .-.'        L3          '
                     (          Overlay       )
                      (                     '-'
                           .'--'._.'.-._.'.-._)
           NVE X =       //                \\    NVE Y =
     (MAC_X,IP_X) +------+              +-------+(MAC_Y,IP_Y)
                .-|NVE X |              | NVE Y |
              (   +------+--.         ( +-------+.--.
            .-.'              '     .-.'              '
           (    DC Site X     )    (    DC Site Y     )
            (             .'-'      (             .'-'
             '--'._.'.    )          '--'._.'.    )
                     '--'              /     '--'
                  /     \             /       \
               __/_      \           /_       _\__
        '--------'   '--------'   '--------'   '--------'
        : Tenant :   : Tenant :   : Tenant :   : Tenant :
        : SystemA:   : SystemC:   : SystemD:   : SystemB:
        '--------'   '--------'   '--------'   '--------'
          TSID=                                  TSID=
      (VNID,MAC_A,IP_A)                      (VNID,MAC_B,IP_B)
```

Figure 1: Example of NVO3 reference architecture for tenant system information discovery

3.1. Issues with tenant system information discovery in the virtualization domain using NVO3

Here we give an example of tenant system information discovery in large layer 2 domain using NVO3 using traditional approach for MAC address learning. The packet flow and control plane operation are as follows:

1. Tenant system A sends a broadcast ARP message to discover the MAC address of Tenant system B. The message contains IP_B in the ARP message payload.

2. The ARP proxy [RFC1027] in NVE X, receiving the ARP message and knowing source and destination are in the different subnet will encapsulate it with overlay header and outer header and flood it

on the overlay network for TSID = <VNID,IP_B,*>. VNID is included in the overlay header.

3. The ARP message will be processed by NVE Y which maintains mapping table matching TSID = <VNID,IP_B,*>. NVE Y, will forward the ARP message to tenant system B. Tenant System B sends ARP reply to tenant system A containing MAC_B.

4. NVE X processes ARP reply message and populates the mapping table with the received entry, then sends it to Tenant System A that includes MAC_B and IP_B of Tenant System B.

5. Tenant system A learns MAC_B from the ARP rely message and can now send a packet to Tenant system B by including MAC_B, and IP_B, as destination addresses.

The issues with tenant system information discovery are as follows:

o The demand on the forwarding table capacity at each NVE is increased compared to non-virtualized environments since layer 2 network is no longer constrained to small local network and has a need for millions of hosts.

o If Address resolution protocol is used for control plane learning, it may cause excessive flooding since ARP packets need to be flooded over the whole overlay network. the ARP/ND processing load imposes great challenge on L2/L3 boundary routers.

o Dynamic data plane learning implies that flooding of unknown destinations be supported and hence implies that broadcast and/or multicast be supported or that ingress replication, which may cause excessive flooding issue and lead to significant scalability limitations.

o A control plane protocol (e.g., BGP) that carries both MAC and IP addresses eliminates the need for ARP, however some NVEs or DC Gateways may not support complex control plane protocol, for example, BGP protocol.

4. Related work for Tenant system information discovery

Currently, 3 main solutions or their combination can be used to perform the tenant system information discovery. They are dynamic data plane learning, ARP, control plane distribution (including two options: centralized or distributed). Additionally, the ARP proxy [RFC1027] mechanism can be used for preventing the ARP flooding in the core network and limiting the MAC table size of NVEs and hosts. Here is a brief analysis of them and the associated protocols are discussed.

4.1. SPB and TRILL

Shortest Path Bridging (SPB) [SPB] and TRILL [TRILL] are two different methods of IS-IS based overlay that operates over L2 Ethernets. They all use the MAC in MAC encapsulation and have the same default MAC address learning method:

- o Using IS-IS extension for outer MAC address distribution over the SPB area or TRILL campus network;

- o Using ARP or data plane snooping for inner MAC address learning of locally attached hosts.

- o In addition, the TRILL maybe use [draft-ietf-trill-directory-framework] distributes the inner MAC address between all the RBriges

In the centralized approach, TRILL may use TRILL ESADI to distribute the inner MAC address between all the RBridges however SPB doesn't support ESADI distribution mechanism. In the distributed approach, SPB and TRILL may use combination of the above 3 methods.

4.2. ARMD and SARP

The ARMD WG examined data center scaling issues with a focus on address resolution and developed a problem statement document [RFC6820]. In this document, the scaling issues of MAC address learning related to the overlay-based approach are listed as followed:

ARP processing on Routers: This issue mainly concerns about the significant amount of ARP traffic or BUM packets traffic in large L2 broadcast domains and its impact to the routers. Finally, some optimized method are proposed;

IPv6 Neighbor Discovery has the similar issue as ARP processing on router;

MAC Address Table Size Limitations at Switches: This issue mainly concerns on the MAC Address Table Size Limitations when the VM number is very large in the Virtualized data center environment.

In order to tackle the above problems, SARP [SARP] seamlessly supports Layer 2 network virtualization services over the overlay network and significantly reduces their complexity in terms of table size and performance. The overlay networks are only required to map MAC addresses of the SARP proxies, instead of MAC address of the destination end host, to the correct tunnel.

4.3. BGP/MPLS IP VPNs - Distributed control plane distribution

BGP/MPLS IP VPNs [RFC4364] provides IP Virtual Private Networks (VPNs) for its customers and support VPN traffic isolation, address overlapping and separation between customer networks. The BGP/MPLS control plane is used to distribute both the VPN labels and the tenant system IP addresses that are used to identify the customer. However BGP/MPLS IP VPN doesn't support interconnection with Data Center (DC) overlay networks and provide a virtual end to end tenant network service to tenant systems in the BGP/MPLS IPVPN.It also has the scalability related problems when IP addresses of a large number of VMs need to be propagated in control plane in the Virtualized data center environments.

For an L3 overlay node, the overlay node only needs to determine IP addresses of the tenant system but doesn't need to know the MAC address of the destination system since overlay tunnels the L3 traffic from the tenant system in an encapsulated format to the final destination and doesn't care about the MAC address of destination end system for the inner L3 packet. Therefore overlay node can answer any address resolution query with its own MAC address or one virtual MAC address. In [I.D-ietf-l3vpn-end-system], NVE uses XMPP to exchange information with the tenant system and answer the address resolution query from tenant system with a virtual router MAC address.

In order to propagate tenant system information to the whole overlay network environment, [I.D-ietf-l3vpn-end-system]use Route Server to gather VPN membership on each Forwarder and IP addresses that are currently associated with each virtual interface of tenant system and advertise them to the BGP speaker. In addition, BGP speaker also can interact with Route Server to generate tenant system information update to the upstream end systems.

4.4. BGP/MPLS Ethernet VPNs and PBB-EVPN

Ethernet Virtual Private Networks (E-VPNs) [I-D.ietf-l2vpn-evpn] provide an emulated L2 service in which each tenant has its own Ethernet network over a common IP or MPLS infrastructure. PBB-EVPN [I-D.ietf -l2vpn-pbb-evpn] is a combined solution of PBB and E-VPN. They all use BGP for MAC address distribution over the core MPLS/IP network, and use ARP or data plane snooping for MAC address learning of locally attached hosts. In other words, the mapping table information <VNID,IP_A,NVE_X> should be distributed to all the remote overlay nodes that belong to the same VN. After that,the tenant system information<VNID,IP_A, MAC_X> is distributed from remote overlay nodes to all the remote tenant system. When all the tenant system information is populated, overlay nodes will process the packet from each tenant system and perform a lookup operation in its map table for the destination TSID=<VNID,IP_B> and determine which tunnel the packet needs to be sent to.

The analysis of their MAC address learning methods is as followed:

Pros:

- o ARP broadcast Suppression: They all construct ARP caches on the PEs and synchronize them either via BGP or data plane snooping. The PEs act as ARP proxies for locally attached hosts, thereby preventing repeated ARP broadcast over the core MPLS/IP network;

- o Comparing E-VPN, PBB-EVPN reduces the number of BGP MAC advertisement routes, provide C-MAC address mobility, confine the scope of C-MAC address learning to only active flows, offer per site policies and avoid C-MAC address flushing on topology changes.

Con: An E-VPN PE sends a BGP MAC Advertisement Route per customer/ client MAC (C-MAC) address. This will raise the scalability related problems in the case of Virtualized data center environments where the number of virtual machines (VMs) is very large.

4.5. VPLS - ARP + data plane learning

VPLS is an L2 VPN technology. VPLS uses the ARP and data plane learning for L2 tenant system information discovery, and not advertised and distributed via a BGP/LDP control plane. The analysis of this method is as followed:

Pros:

o Reducing complexity and work burden of the control plane by decreasing the control packets;

o MAC address learning based on active flows can save the space of MAC mapping table.

Cons:

o PE will learn all active MACs over the associated PW by BUM flooding of data plane. But, some active MACs is not destined to the PE;

o Unlike the active MAC withdraw mechanism in control plane, PE cannot flush MAC address real-time in data plane, when host MACs behind the PE are changed.

4.6. LISP - Centralized control plane distribution

LISP[RFC6830] essentially provides an IP over IP overlay where the internal addresses are end station Identifiers and the outer IP addresses represent the location of the end station within the core IP network topology. [draft-maino-nvo3-lisp-cp-02] discusses L2 over L3 LISP Encapsulation and proposes a LISP Mapping System for ARP resolution to eliminate the flooding of ARP traffic and further reduce the need for multicast in the underlay network. This system relies on mapping system for tenant system information distribution and involves MAP-request/MAP-Response message exchange between overlay node and mapping system. With introduced LISP Mapping system, the scalability is improved for tenant system information discovery. the packet flow and control plane operation are as follows:

o Tenant System A sends a broadcast ARP message to discover the MAC address of Tenant system B. The message contains IP_B in the ARP message payload.

o NVE X as an ARP proxy, receiving the ARP message and knowing source and destination are in the different subnet[RFC1027], but rather than flooding it on the overlay network, sends a Map-Request(i.e.,LISP signaling) to the backend LISP mapping system (i.e.,NVA) that maintains mapping information for entire overlay network for TSID = <VNID,IP_B,*>.

o The Map-Request is reflected by the backend LISP Mapping system to NVE Y, that will send a Map-Reply back to NVE X containing the mapping TSID=<VNID,IP_B,MAC_B>. Alternatively, depending on the Backend LISP Mapping system configuration, the backend LISP

Mapping system may send directly a Map-Reply to NVE X.

- o NVE X populates the mapping table with the received entry, and sends an ARP-Agent Reply to Tenant System A that includes MAC_B and IP_B.

- o Tenant system A learns MAC_B from the ARP message and can now send a packet to Tenant system B by including MAC_B, and IP_B, as destination addresses.

5. Gap Analysis and Discuss

The following table compares several tenant system information discovery methods from different aspects under the same network topology and scale.

TS Discovery method	Forwarding table size	Packets flooding impact	Control plane Distribution support	Directory Support
SPB &TRILL	Mediaum	Medium	Yes	Trill:Yes SPB:No
ARMD&SARP	Small	Medium	No	No
LISP + ARP proxy	Medium	Medium	Yes	LISP Mapping System
BGP/MPLS IP VPN	Large	Large	Yes	No
BGP/MPLS Ethernet VPN	Large	Large	Yes	No
VPLS + ARP proxy	Medium	Small	Yes	No

Table 1: The comparison between several tenant system

information discovery methods

6. Conclusion

There are three ways for tenant system information discovery, data plane learning and control plane ARP learning and control plane distribution. In large layer 2 domain, the MAC address can not be simply learnt by looking at the outer layer 2 header, instead, Deeper parsing inner Ethernet header is required. However it also introduces a lot of processing overhead. In order to address this issue, the control plane distribution is proposed, and used to carry both MAC address and IP address and eliminate the above data plane learning issue. However distribution protocol is needed. How distribution protocol is used to propagate tenant system information and mapping table information in large scale and in a more efficient way is still under study.

7. IANA Considerations

This document has no actions for IANA.

8. Security Considerations

TBC.

9. Normative References

[ESADI] Eastlake, D., "TRILL (Transparent Interconnection of Lots of Links): ESADI (End Station Address Distribution Information) Protocol", ID draft-ietf-trill-esadi-02, February 2013.

[I-D.ietf-l2vpn-evpn] Sajassi, A. and R. Aggarwal, "BGP MPLS Based Ethernet VPN", ID draft-ietf-l2vpn-evpn-03, February 2013.

[I-D.ietf-l2vpn-pbb-evpn] Sajassi, A., "PBB-EVPN", ID draft-ietf-l2vpn-pbb-evpn-04, April 2013.

[I-D.ietf-trill-directory-framework] Dunbar, L. and D. Eastlake, "TRILL (Transparent Interconnection of Lots of Links): Edge Directory Assistance Framework", ID draft-ietf-trill-directory-framework-05, April 2013.

[I.D-ietf-l3vpn-end-system] Marques, P., "BGP-signaled end-system IP/VPNs", ID draft-maino-nvo3-lisp-cp-02, April 2013.

[I.D-ietf-nvo3-framework] Lasserre, M., "Framework for DC Network Virtualization", ID draft-ietf-nvo3-framework-00, September 2012.

[RFC1027] Carl-Mitchell, S., "Using ARP to Implement Transparent Subnet Gateways", October 1987.

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

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

[RFC6325] Perlman, R., "RBridges: Base Protocol Specification", July 2011.

[RFC6820] Farinacci, D., "The Locator/ID Separation Protocol (LISP)", January 2013.

[RFC6830] Farinacci, D., "The Locator/ID Separation Protocol (LISP)", January 2013.

[SARP] Dunbar, L. and I. Yerushalmi, "Scaling the Address

Resolution Protocol for Large Data Centers (SARP)", ID draft-nachum-sarp-04, February 2013.

[SPB] "IEEE standard for local and metropolitan area networks: Media access control (MAC) bridges and virtual bridged local area networks -- Amendment 20: Shortest path bridging", IEEE 802.1aq, June 2012.

[draft-maino-nvo3-lisp-cp]
Maino, F. and R. Aggarwal, "LISP Control Plane for Network Virtualization Overlays", ID draft-maino-nvo3-lisp-cp-02, April 2013.

Authors' Addresses

Liang Xia
Huawei
101 Software Avenue, Yuhua District
Nanjing, Jiangsu 210012
China

Email: frank.xialiang@huawei.com

Qin Wu
Huawei
101 Software Avenue, Yuhua District
Nanjing, Jiangsu 210012
China

Email: bill.wu@huawei.com