NVO3 working group
Internet Draft
Intended status: Standards Track
Expires: September 8, 2015

A. Ghanwani
Dell
L. Dunbar
Huawei
M. McBride
Ericsson
V. Bannai
Paypal
R. Krishnan
Brocade

March 9, 2015

A Framework for Multicast in NVO3
draft-ghanwani-nvo3-mcast-framework-00

Status of this Memo

This Internet-Draft is submitted in full conformance with the provisions of BCP 78 and BCP 79.

This Internet-Draft is submitted in full conformance with the provisions of BCP 78 and BCP 79. This document may not be modified, and derivative works of it may not be created, except to publish it as an RFC and to translate it into languages other than English.

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 July9,2015.

Copyright Notice

Copyright (c) 2015 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 Simplified BSD License.

Abstract

This document discusses a framework of supporting multicast traffic, , in a network that uses Network Virtualization using Overlays over Layer 3 (NVO3). Both infrastructure multicast and application-specific multicast are discussed. It describes the various mechanisms and considerations that can be used for delivering such traffic as well as the data plane and control plane considerations for each of the mechanisms.

Table of Contents

1. Introduction

Network virtualization using Overlays over Layer 3 (NVO3) is a technology that is used to address issues that arise in building large, multitenant data centers that make extensive use of server virtualization [PS].

This document provides a framework for supporting multicast traffic, in a network that uses Network Virtualization using Overlays over Layer 3 (NVO3). Both infrastructure multicast (ARP/ND, DHCP, mDNS, etc.) and application-specific multicast are considered. It describes the various mechanisms and considerations that can be used for delivering such traffic in networks that use NVO3.

The reader is assumed to be familiar with the terminology as defined in the NVO3 Framework document [FW] and NVO3 Architecture document [NVO3-ARCH].

1.1. Infrastructure multicast

Infrastructure multicast includes protocols such as ARP/ND, DHCP, and mDNS. It is possible to provide solutions for these that do not involve multicast in the underlay network. In the case of ARP/ND, an NVA can be used for distributing the mappings of IP address to MAC address to all NVEs, and the NVEs can respond to ARP messages from the TSs that are attached to it in a way that is similar to proxy-ARP. In the case of DHCP, the NVE can be configured to forward these messages using a helper function.

Of course it is possible to support all of these infrastructure multicast protocols natively if the underlay provides multicast transport. However, even in the presence of multicast transport, it may be beneficial to use the optimizations mentioned above to reduce the amount of such traffic in the network.

1.2. Application-specific multicast

Application-specific multicast traffic, which may be either Source-Specific Multicast (SSM) or Any-Source Multicast (ASM)[RFC 3569], has the following characteristics:

1. Receiver hosts are expected to subscribe to multicast content using protocols such as IGMP [RFC3376] (IPv4) or MLD (IPv6). Multicast sources and listeners participant in these protocols using addresses that are in the Tenant System address domain.

2. The list of multicast listeners for each multicast group is not known in advance. Therefore, it may not be possible for an NVA to get the list of participants for each multicast group ahead of time.

2. Acronyms

ASM: Any-Source Multicast

LISP: Locator/ID Separation Protocol

NVA: Network Virtualization Authority

NVE: Network Virtualization Edge

NVGRE: Network Virtualization using GRE

SSM: Source-Specific Multicast

STT: Stateless Tunnel Transport

VXLAN: Virtual eXtensible LAN

3. Multicast mechanisms in networks that use NVO3

In NVO3 environments, traffic between NVEs is transported using an encapsulation such as VXLAN [VXLAN], NVGRE [NVGRE], STT [STT], etc.

Besides the need to support the Address Resolution Protocol (ARP) and Neighbor Discovery (ND), there are several applications that require the support of multicast and/or broadcast in data centers [DC-MC]. With NVO3, there are many possible ways that multicast may be handled in such networks. We discuss some of the attributes of the following four methods: 1. No multicast support.

2. Replication at the source NVE.

3. Replication at a multicast service node.

4. IP multicast in the underlay.

These mechanisms are briefly mentioned in the NVO3 Framework [FW] and NVO3 architecture [NVO3-ARCH] document. This document attempts to provide more details about the basic mechanisms underlying each of these mechanisms and discusses the issues and tradeoffs of each.

We note that other methods are also possible, such as [EDGE-REP], but we focus on the above four because they are the most common.

3.1. No multicast support

In this scenario, there is no support whatsoever for multicast traffic when using the overlay. This can only work if the following conditions are met:

1. All of the traffic is unicast. traffic in the network and the only multicast/broadcast traffic is from ARP/ND protocols and flooding of frames with an unknown MAC destination address.

2. A network virtualization authority (NVA) is used by the NVEs to determine the mapping of a given Tenant System's MAC/IP address to its NVE. In other words, there is no data plane learning. Address resolution requests via ARP/ND that are issued by the Tenant Systems must be resolved by the NVE that they are attached to.

With this approach, it is not possible to support application-specific multicast. However, certain multicast/broadcast applications such as DHCP can be supported by use of a helper function in the NVE.

The main drawback of this approach, even for unicast traffic, is that it is not possible to initiate communication with a Tenant System for which a mapping to an NVE does not already exist with the NVA. This is a problem in the case where the NVE is implemented in a physical switch and the Tenant System is a physical end station that has not registered with the NVA.

3.2. Replication at the source NVE

With this method, the overlay attempts to provide a multicast service without requiring any specific support from the underlay, other than that of a unicast service. A multicast or broadcast transmission is achieved by replicating the packet at the source NVE, and making copies, one for each destination NVE that the multicast packet must be sent to.

For this mechanism to work, the source NVE must know, a priori, the IP addresses of all destination NVEs that need to receive the packet. For the purpose of ARP/ND, this would involve knowing the IP addresses of all the NVEs that have Tenant Systems in the virtual network instance (VNI) of the Tenant System that generated the request. For the support of application-specific multicast traffic, a method similar to that of receiver-sites registration for a particular multicast group described in [LISP-Signal-Free] can be used. The registrations from different receiver-sites can be merged at the NVA, which can construct a multicast replication-list inclusive of all NVEs to which receivers for a particular multicast group are attached. The replication-list for each specific multicast group is maintained either by the NVA.

The receiver-sites registration is achieved by egress NVEs performing the IGMP/MLD snooping to maintain state for which attached Tenant Systems have subscribed to a given IP multicast group. When the members of a multicast group are outside the NVO3 domain, it is necessary for NVO3 gateways to keep track of the remote members of each multicast group. The NVEs then communicate these mappings to the NVA. Even if the membership is not communicated to the NVA, if it is necessary to prevent hosts attached to an NVE that have not subscribed to a multicast group from receiving the multicast traffic, the NVE needs to maintain the multicast group membership.

In the absence of IGMP/MLD snooping, the traffic would be delivered to all hosts that are part of the VNI.

This method requires multiple copies of the same packet to all NVEs that participate in the VN. If, for example, a tenant subnet is

spread across 50 NVEs, the packet would have to be replicated 50 times at the source NVE. This also creates an issue with the forwarding performance of the NVE.

Note that this method is similar to what was used in VPLS [VPLS] prior to support of MPLS multicast [MPLS-MC]. While there are some similarities between MPLS VPN and the NVO3 overlay, there are some key differences:

- The CE-to-PE attachment in VPNs is somewhat static, whereas in a DC that allows VMs to migrate anywhere, the TS attachment to NVE is much more dynamic.

- The number of PEs to which a single VPN customer is attached in an MPLS VPN environment is normally far less than the number of NVEs to which a VNI's VMs are attached in a DC.

  When a VPN customer has multiple multicast groups, [RFC6513] "Multicast VPN" combines all those multicast groups within each VPN client to one single multicast group in the MPLS (or VPN) core. The result is that messages from any of the multicast groups belonging to one VPN customer will reach all the PE nodes of the client. In other words, any messages belonging to any multicast groups under customer X will reach all PEs of the customer X. When the customer X is attached to only a handful of PEs, the use of this approach does not result in excessive wastage of bandwidth in the provider's network.

  In a DC environment, a typical server/hypervisor based virtual switch may only support 10's VMs (as of this writing). A subnet with N VMs may be, in the worst case, spread across N vSwitches. Using "MPLS VPN multicast" approach in a such a scenario would require the creation of a Multicast group in the core for this VNI to reach all N NVEs. If only small percentage of this client's VMs participate in application specific multicast, a great number of NVEs will receive multicast traffic that is not forwarded to any of their attached VMs, resulting in considerable wastage of bandwidth.

Therefore, the Multicast VPN solution may not scale in DC environment with dynamic attachment of Virtual Networks to NVEs and greater number of NVEs for each virtual network.

3.3. Replication at a multicast service node

With this method, all multicast packets would be sent using a unicast tunnel encapsulation to a multicast service node (MSN). The MSN, in turn, would create multiple copies of the packet and would deliver a copy, using a unicast tunnel encapsulation, to each of the NVEs that are part of the multicast group for which the packet is intended.

This mechanism is similar to that used by the ATM Forum's LAN Emulation [LANE] specification [LANE].

The following are the possible ways for the MSN to get the membership information for each multicast group:

- The MSN can obtain this information by snooping the IGMP/MLD messages from the Tenant Systems and/or sending query messages to the Tenant Systems. In order for MSN to snoop the IGMP/MLD messages between TSs and their corresponding routers, the NVEs that TSs are attached have to encapsulate a special outer header, e.g. outer destination being the multicast server node. See Section 3.3.2 for detail.

- The MSN can obtain the membership information from the NVEs that snoop the IGMP/MLD messages. This can be done by having the MSN communicate with the NVEs, or by having the NVA obtain the information from the NVEs, and in turn have MSN communicate with the NVA.

Unlike the method described in Section 3.2, there is no performance impact at the ingress NVE, nor are there any issues with multiple copies of the same packet from the source NVE to the multicast service node. However there remain issues with multiple copies of the same packet on links that are common to the paths from the MSN to each of the egress NVEs. Additional issues that are introduced with this method include the availability of the MSN, methods to scale the services offered by the MSN, and the sub-optimality of the delivery paths.

Finally, the IP address of the source NVE must be preserved in packet copies created at the multicast service node if data plane learning is in use. This could create problems if IP source address reverse path forwarding (RPF) checks are in use.

3.4. IP multicast in the underlay

In this method, the underlay supports IP multicast and the ingress NVE encapsulates the packet with the appropriate IP multicast address in the tunnel encapsulation header for delivery to the desired set of NVEs. The protocol in the underlay could be any variant of Protocol Independent Multicast (PIM), or protocol dependent multicast, such as [ISIS-Multicast].

If an NVE connects to its attached TSs via Layer 2 network, there are multiple ways for NVEs to support the application specific multicast:

- The NVE only supports the basic IGMP/MLD snooping function, let the TSs routers handling the application specific multicast. This scheme doesn't utilize the underlay IP multicast protocols.
-
- The NVE can act as a pseudo multicast router for the directly attached VMs and support proper mapping of IGMP/MLD's messages to the messages needed by the underlay IP multicast protocols.

With this method, there are none of the issues with the methods described in Sections 3.2.

With PIM Sparse Mode (PIM-SM), the number of flows required would be (n*g), where n is the number of source NVEs that source packets for the group, and g is the number of groups. Bidirectional PIM (BIDIR-PIM) would offer better scalability with the number of flows required being g.

In the absence of any additional mechanism, e.g. using an NVA for address resolution, for optimal delivery, there would have to be a separate group for each tenant, plus a separate group for each multicast address (used for multicast applications) within a tenant.

Additional considerations are that only the lower 23 bits of the IP address (regardless of whether IPv4 or IPv6 is in use) are mapped to the outer MAC address, and if there is equipment that prunes multicasts at Layer 2, there will be some aliasing. Finally, a mechanism to efficiently provision such addresses for each group would be required.

There are additional optimizations which are possible, but they come with their own restrictions. For example, a set of tenants may be restricted to some subset of NVEs and they could all share the same outer IP multicast group address. This however introduces a problem of sub-optimal delivery (even if a particular tenant within the group of tenants doesn't have a presence on one of the NVEs which another one does, the former's multicast packets would still be delivered to that NVE). It also introduces an additional network management burden to optimize which tenants should be part of the same tenant group (based on the NVEs they share), which somewhat dilutes the value proposition of NVO3 which is to completely decouple the overlay and physical network design allowing complete freedom of placement of VMs anywhere within the data center.

Multicast schemes such as BIER (Bit Index Explicit Replication) may be able to provide optimizations by allowing the underlay network to provide optimum multicast delivery without requiring routers in the core of the network to main per-multicast group state.

3.5. Other schemes

There are still other mechanisms that may be used that attempt to combine some of the advantages of the above methods by offering multiple replication points, each with a limited degree of replication [EDGE-REP]. Such schemes offer a trade-off between the amount of replication at an intermediate node (router) versus

performing all of the replication at the source NVE or all of the replication at a multicast service node.

4. Simultaneous use of more than one mechanism

While the mechanisms discussed in the previous section have been discussed individually, it is possible for implementations to rely on more than one of these. For example, the method of Section 3.1 could be used for minimizing ARP/ND, while at the same time, multicast applications may be supported by one, or a combination of, the other methods. For small multicast groups, the methods of source NVE replication or the use of a multicast service node may be attractive, while for larger multicast groups, the use of multicast in the underlay may be preferable.

5. Other issues

5.1. Multicast-agnostic NVEs

Some hypervisor-based NVEs do not process or recognize IGMP/MLD frames; i.e. those NVEs simply encapsulate the IGMP/MLD messages in the same way as they do for regular data frames.

By default, TSs router periodically sends IGMP/MLD query messages to all the hosts in the subnet to trigger the hosts that are interested in the multicast stream to send back IGMP/MLD reports. In order for MSN get the updated multicast group information, the MSN can also send the IGMP/MLD query message comprising a client specific multicast address, encapsulated in an overlay header to all the NVEs to which the TSs in the VN are attached.

However, MSN may not always be aware of the client specific multicast addresses. Then MSN has to snoop the IGMP/MLD messages between TSs and their corresponding routers to maintain the multicast membership. In order for MSN to snoop the IGMP/MLD messages between TSs and their router, NVA needs to configure the NVE to send copies of the IGMP/MLD messages to the MSN in addition to the default behavior of sending them to the TSs' routers; e.g. the NVA has to inform the NVEs to encapsulate data frames with DA being 224.0.0.2 (destination address of IGMP report) to TSs' router and MSN.

This process is similar to "Source Replication" described in Section 3.2, except the NVEs only replicate the message to TS's router and MSN.

5.2. Multicast membership management for DC with VMs

For data centers with virtualized servers, VMs can be added, deleted or moved very easily. When VMs are added, deleted or moved, the NVEs to which the VMs are attached are changed.

When a VM is deleted from an NVE or a new VM is added to an NVE, the VM management system should notify the MSN to send the IGMP/MLD query messages to the relevant NVEs, so that the multicast membership can be updated promptly. Otherwise, if there are changes of VMs attachment to NVEs, then for the duration of the configured default time interval that the TSs routers use for IGMP/MLD queries, multicast data may not reach the VM(s) that moved.

6. Summary

This document has identified various mechanisms for supporting application specific multicast in networks that use NVO3. It highlights the basics of each mechanism and some of the issues with them. As solutions are developed, the protocols would need to consider the use of these mechanisms and co-existence may be a consideration. It also highlights some of the requirements for supporting multicast applications in an NVO3 network.

7. Security Considerations

This draft does not introduce any new security considerations beyond what may be present in proposed solutions

8. IANA Considerations

This document requires no IANA actions. RFC Editor: Please remove this section before publication.

9. References

9.1. Normative References

[PS] Lasserre, M. et al., "Framework for DC network virtualization", work in progress, January 2014.

[FW] Narten, T. et al., "Problem statement: Overlays for network virtualization", work in progress, July 2013.

[NVO3-ARCH] Narten, T. et al.," An Architecture for Overlay Networks (NVO3)", work in progress, Feb 2014

[RFC3376] B. Cain, et al, "Internet Group Management Protocol, Version 3", Oct 2002.

[RFC6513] Rosen, E. et al., "Multicast in MPLS/BGP IP VPNs". Feb 2012.

9.2. Informative References

[VXLAN] Mahalingam, M. et al., "VXLAN: A framework for overlaying virtualized Layer 2 networks over Layer 3 networks," work in progress. [AG: Replace with RFC.]

[NVGRE] Sridharan, M. et al., "NVGRE: Network virtualization using Generic Routing Encapsulation," work in progress.

[STT] Davie, B. and Gross J., "A stateless transport tunneling protocol for network virtualization," work in progress.

[DC-MC] McBride M., and Lui, H., "Multicast in the data center overview," work in progress.

[ISIS-Multicast] L. Yong, et al, "ISIS Protocol Extension For Building Distribution Trees", work in progress. Oct 2013.

[VPLS] Lasserre, M., and Kompella, V. (Eds), "Virtual Private LAN Service (VPLS) using Label Distribution Protocol (LDP) signaling," RFC 4762, January 2007.

[MPLS-MC] Aggarwal, R. et al., "Multicast in VPLS," work in progress.

[LANE] "LAN emulation over ATM," The ATM Forum, af-lane-0021.000, January 1995.

[EDGE-REP] Marques P. et al., "Edge multicast replication for BGP IP VPNs," work in progress, June 2012.

[RFC 3569] S. Bhattacharyya, Ed., "An Overview of Source-Specific Multicast (SSM)", July 2003.

[LISP-Signal-Free] V. Moreno & D. Farinacci, "Signal-Free LISP Multicast", work in progress. Dec 2014.

10. Acknowledgments

We want to thank Dino Farinacci for comments and suggestions to this draft.

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

Authors' Addresses

Anoop Ghanwani
Dell
Email: anoop@alumni.duke.edu

Linda Dunbar
Huawei Technologies
5340 Legacy Drive, Suite 1750
Plano, TX 75024, USA
Phone: (469) 277 5840
Email: ldunbar@huawei.com

Mike McBride
Ericsson
mike.mcbride@ericsson.com

Vinay Bannai
Paypal
Email: vbannai@paypal.com

Ramki Krishnan
Brocade
Email: ramk@brocade.com