NVO3 working group   A. Ghanwani
Internet Draft   Dell
Intended status: Informational   L. Dunbar
Expires: November 8, 2018   M. McBride
                                                              Huawei
                                                            V. Bannai
                                                              Google
                                                           R. Krishnan
                                                                 Dell

                                                     October 5, 2017

             A Framework for Multicast in Network Virtualization Overlays
                      draft-ietf-nvo3-mcast-framework-10

Status of this Memo

Copyright Notice

Abstract

   This document provides a framework of supporting multicast traffic
   in a network that uses Network Virtualization Overlays (NVO3). Both
   infrastructure multicast and application-specific multicast are
   discussed. It describes the various mechanisms 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)[RFC7365]
is a technology that is used to address issues that arise in
building large, multitenant data centers that make extensive use of
server virtualization [RFC7364].

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 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 [RFC7365] and NVO3 Architecture
document [RFC8014].

1.1. Infrastructure multicast

Infrastructure multicast is a capability needed by networking
services, such as Address Resolution Protocol (ARP), Neighbor
Discovery (ND), Dynamic Host Configuration Protocol (DHCP),
multicast Domain Name Server (mDNS), etc. RFC3819 Section 5 and 6
have detailed description for some of the infrastructure multicast
[RFC3819]. It is possible to provide solutions for these that do
not involve multicast in the underlay network. In the case of
ARP/ND, a network virtualization authority (NVA) can be used for
distributing the mappings of IP address to MAC address to all
network virtualization edges (NVEs). The NVEs can then trap ARP
Request/ND Neighbor Solicitation messages from the TSs (Tenant
System) that are attached to it and respond to them, thereby
eliminating the need to for broadcast/multicast of such messages.
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 are originated and consumed
   by user applications. The Application-specific multicast, which can
   be either Source-Specific Multicast (SSM) or Any-Source Multicast
   (ASM)[RFC3569], has the following characteristics:

      1. Receiver hosts are expected to subscribe to multicast content
         using protocols such as IGMP [RFC3376] (IPv4) or MLD [RFC2710]
         (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.

 1.3. Terminology clarification

2. Acronyms & Terminology

   In this document, the terms host, tenant system (TS) and virtual
   machine (VM) are used interchangeably to represent an end station
   that originates or consumes data packets.

   ASM: Any-Source Multicast

   IGMP: Internet Group Management Protocol

   LISP: Locator/ID Separation Protocol

   MSN: Multicast Service Node

   RLOC: Routing Locator

   NVA: Network Virtualization Authority

   NVE: Network Virtualization Edge

   NVGRE: Network Virtualization using GRE

PIM: Protocol-Independent Multicast

SSM: Source-Specific Multicast

TS: Tenant system

VM: Virtual Machine

VN: Virtual Network

VTEP: VxLAN Tunnel End Points

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 Virtual eXtensible Local Area Network (VXLAN)
[RFC7348,VXLAN-GPE], Network Virtualization Using Generic Routing
Encapsulation (NVGRE) [RFC7637], Geneve [Geneve], Generic UDP
Encapsulation (GUE) [GUE], etc.

What makes NVO3 different from any other network is that some NVEs,
especially the NVE implemented on server, might not support PIM or
other native multicast mechanisms. They might just encapsulate the
data packets from VMs with an outer unicast header. Therefore, it is
important for networks using NVO3 to have mechanisms to support
multicast as a network capability for NVEs, to map multicast traffic
from VMs (users/applications) to an equivalent multicast capability
inside the NVE, or to figure out the outer destination address if
NVE does not support native multicast (e.g. PIM) or IGMP.

Besides the need to support ARP and 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 methods are briefly mentioned in the NVO3 Framework [RFC7365] and NVO3 architecture [RFC8014] document. This document provides more details about the basic mechanisms underlying each of these methods and discusses the issues and trade-offs 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 method can only work if the following conditions are met:

1. All of the application traffic in the network is unicast traffic and the only multicast/broadcast traffic is from ARP/ND protocols.

2. An NVA is used by the NVEs to determine the mapping of a given Tenant System's (TS'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 TSs 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 TS for which a mapping to an NVE does not already exist in the NVA. This is a problem in the case where the NVE is implemented in a physical switch and the TS 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 TSs in the virtual network
(VN) of the TS 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 by
the NVA. Note: Using LISP-signal-free does not necessarily mean the
head-end (i.e. NVE) must do replication. If the mapping database
(i.e. NVA) indicates that packets are encapsulated to multicast
RLOCs, then there is no replication happening at the NVE.

The receiver-sites registration is achieved by egress NVEs
performing the IGMP/MLD snooping to maintain state for which
attached TSs 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 and NVO3 gateways then communicate
the multicast groups that are of interest to the NVA. If the
membership is not communicated to the NVA, and 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 would
need to maintain multicast group membership information.

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

In multi-homing environments, i.e., in those where a TS is attached
to more than one NVE, the NVA would be expected to provide
information to all of the NVEs under its control about all of the
NVEs to which such a TS is attached. The ingress NVE can choose any
one of the egress NVEs for the data frames destined towards the TS.

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. Obviously, this approach creates more
traffic to the network that can cause congestion when the network

load is high. This also creates an issue with the forwarding
performance of the NVE.

Note that this method is similar to what was used in Virtual Private
LAN Service (VPLS) [RFC4762] prior to support of Multi-Protocol
Label Switching (MPLS) multicast [RFC7117]. While there are some
similarities between MPLS Virtual Private Network (VPN) and NVO3,
there are some key differences:

- The Customer Edge (CE) to Provider Edge (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 VN's VMs are attached in a DC.

  When a VPN customer has multiple multicast groups, "Multicast VPN"
  [RFC6513] 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 such a scenario would
  require the creation of a Multicast group in the core for this VN
  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 from the ingress NVE 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 Asynchronous Transfer
   Mode (ATM) Forum's LAN Emulation (LANE) specification [LANE]. The
   MSN is similar to the RP (Rendezvous Point) in PIM SM, but different
   in that the user data traffic are carried by the NVO3 tunnels.

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

   - The MSN can obtain this membership information from the IGMP/MLD
     report messages sent by TSs in response to IGMP/MLD query messages
     from the MSN. The IGMP/MLD query messages are sent from the MSN to
     the NVEs, which then forward the query messages to TSs attached to
     them. An IGMP/MLD query messages sent out by the MSN to an NVE is
     encapsulated with the MSN address in the outer source address
     field and the address of the NVE in the outer destination address
     field. The encapsulated IGMP/MLD query messages also has a VNID
     for a virtual network (VN) that TSs belong in the outer header and
     a multicast address in the inner destination address field. Upon
     receiving the encapsulated IGMP/MLD query message, the NVE
     establishes a mapping "MSN address" <-> "multicast address",
     decapsulates the received encapsulated IGMP/MLD message, and
     multicast the decapsulated query message to TSs that belong to the
     VN under the NVE. A IGMP/MLD report message sent by a TS includes
     the multicast address and the address of the TS. With the proper
     "MSN Address" <-> "Multicast-Address" mapping, the NVEs can
     encapsulate all multicast data frames to the "Multicast-Address"
     with the address of the MSN in the outer destination address
     field.

- The MSN can obtain the membership information from the NVEs that
  have the capability to establish multicast groups by snooping
  native IGMP/MLD messages (p.s. the communication must be specific
  to the multicast addresses), or by having the NVA obtain the
  information from the NVEs, and in turn have MSN communicate with
  the NVA. This approach requires additional protocol between MSN
  and NVEs.

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 a 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. Unfortunately, many vendors still do not fully
   support BIDIR or have limitations on its implementation. RFC6831
   [RFC6831] has good description of using SSM as an alternative to
   BIDIR if the VTEP/NVE devices have a way to learn of each other's IP
   address so that they could join all SSM SPT's to create/maintain an
   underlay SSM IP Multicast tunnel solution.

   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 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 Indexed Explicit Replication)
[BIER-ARCH] 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 maintain 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 (e.g. 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 the MSN to 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, the MSN may not always be aware of the client specific multicast addresses. In order to perform multicast filtering, the MSN has to snoop the IGMP/MLD messages between TSs and their corresponding routers to maintain the multicast membership. In order for the MSN to snoop the IGMP/MLD messages between TSs and their router, the 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 TSs' 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 (as described in Section 3.3), so that the multicast membership can be updated promptly. Otherwise, if there are changes of VMs attachment to NVEs, within 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 is described n NVO3 Architecture (RFC8014).

8. IANA Considerations

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

9. References

9.1. Normative References

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

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

[RFC7364] Narten, T. et al., "Problem statement: Overlays for network virtualization", October 2014.

[RFC7365] Lasserre, M. et al., "Framework for data center (DC) network virtualization", October 2014.

[RFC8014] Narten, T. et al.," An Architecture for Overlay Networks (NVO3)", RFC8014, Dec. 2016.

9.2. **Informative References**

   [RFC2710] S. Deering et al, "Multicast Listener Discovery (MLD) for
             IPv6", Oct 1999.

   [RFC3569] S. Bhattacharyya, Ed., "An Overview of Source-Specific
             Multicast (SSM)", July 2003.

   [RFC3819] P. Harn et al., "Advice for Internet Subnetwork
             Designers", July 2004.

   [RFC4762] Lasserre, M., and Kompella, V. (Eds.), "Virtual Private
             LAN Service (VPLS) using Label Distribution Protocol (LDP)
             signaling," January 2007.

   [RFC6831] Farinacci, D. et al., "The Locator/ID Seperation Protocol
             (LISP) for Multicast Environments", Jan, 2013.

   [RFC7117] Aggarwal, R. et al., "Multicast in VPLS," February 2014.

   [RFC7348] Mahalingam, M. et al., " Virtual eXtensible Local Area
             Network (VXLAN): A Framework for Overlaying Virtualized
             Layer 2 Networks over Layer 3 Networks", August 2014.

   [RFC7365] M. Lasserre, et al. "Framework for Data Center (DC)
             Network Virtualization", Oct 2014.

   [RFC7637] Garg P. and Wang, Y. (Eds.), "NVGRE: Network
             Vvirtualization using Generic Routing Encapsulation",
             September 2015.

   [BIER-ARCH]
             Wijnands, IJ. (Ed.) et al., "Multicast using Bit Index
             Explicit Replication," <draft-ietf-bier-architecture-03>,
             January 2016.

   [DC-MC] McBride, M. and Lui, H., "Multicast in the data center
             overview," <draft-mcbride-armd-mcast-overview-02>, work in
             progress, July 2012.

   [EDGE-REP]

            Marques P. et al., "Edge multicast replication for BGP IP
            VPNs," <draft-marques-l3vpn-mcast-edge-01>, work in
            progress, June 2012.

   [Geneve]
            Gross, J. and Ganga, I. (Eds.), "Geneve: Generic Network
             Virtualization Encapsulation", <draft-ietf-nvo3-geneve-
            01>, work in progress, January 2016.

   [GUE]
            Herbert, T. et al., "Generic UDP Encapsulation", <draft-
             ietf-nvo3-gue-02>, work in progress, December 2015.

   [ISIS-Multicast]
            Yong, L. et al., "ISIS Protocol Extension for Building
            Distribution Trees", <draft-yong-isis-ext-4-distribution-
            tree-03>, work in progress, October 2014.

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

    [LISP-Signal-Free]

            Moreno, V. and Farinacci, D., "Signal-Free LISP
            Multicast", <draft-ietf-lisp-signal-free-multicast-01>,
            work in progress, April 2016.

    [VXLAN-GPE]

            Kreeger, L. and Elzur, U. (Eds.), "Generic Protocol
            Extension for VXLAN", <draft-ietf-nvo3-vxlan-gpe-02>, work
            in progress, April 2016.

10. Acknowledgments

   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
   Huawei Technologies
   Email: mmcbride7@gmail.com

   Vinay Bannai
   Google
   Email: vbannai@gmail.com

   Ram Krishnan
   Dell
   Email: ramkri123@gmail.com