INTERNET-DRAFT
Intended Status: Informational
Expires: August 12, 2014

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

February 13, 2014

Multicast Issues in Networks Using NVO3
draft-ghanwani-nvo3-mcast-issues-01

Status of this Memo

Copyright Notice

to this document.

Abstract

This memo discusses issues with supporting multicast traffic in a network that uses Network Virtualization using Overlays over Layer 3 (NVO3). It describes the various mechanisms that may be used for multicast and discusses some of the considerations with supporting multicast applications in networks that use NVO3.

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 is focused specifically on the problem of supporting multicast in networks that use NVO3. Because of the requirement of multi-destination delivery, multicast traffic poses some unique challenges.

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

2. Multicast mechanisms in networks that use NVO3

In NVO3 environments, traffic between NVEs is transported using a tunnel 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, but other methods are also possible.

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] document. This document attempts to fill in some more details about the basic mechanisms underlying each of these mechanisms and discusses the issues and tradeoffs of each.

2.1 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. In other words, there are no multicast applications in the network and the only multicast traffic is due to ARP/ND and due to flooding of frames with an unknown MAC destination address.

2. A network virtualization authority (NVA) is used at the NVE to determine the MAC address-to-NVE mapping and to determine the MAC address-to-IP address bindings. In other words, there is no data plane learning, and address resolution requests via ARP/ND that are issued by the VMs must be resolved by the NVE that they are attached to.

With this approach, certain multicast/broadcast applications such as DHCP can be supported by use of a helper function in the NVE.

The main issues that need to be addressed with this mechanism are the handling of hosts for which a mapping does not already exist in the NVA. This issue can be particularly challenging if such end systems are reachable through more than one NVE.

2.2 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 example, in the case of an ARP broadcast or an ND multicast, the source NVE must know the IP addresses of all the remote NVEs where there are members of the tenant subnet in question.

The obvious drawback with this method is that we have multiple copies of the same packet that will traverse any common links that are along the path to each of the destination NVEs. 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, especially if it is implemented in software.

Note that this method is similar to what was used in VPLS [VPLS] prior to extensive support of MPLS multicast [MPLS-MC].

2.3 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. The multicast service node, 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].

Unlike the method described in Section 2.2, 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 multicast service node to each of the egress NVEs. Additional issues that are introduced with this method include the availability of the multicast service node, methods to scale the services offered by the multicast service node, 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.

2.4 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). The NVE would be required to participate in the underlay as a host using IGMP/MLD in order for the underlay to learn about the groups that the NVE participates in.

With this method, there are none of the issues with the methods described in Sections 2.2.

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.

2.5 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.

3. 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 2.1 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.

4. IP multicast applications in the overlay

When IP multicast is implemented in the overlay (i.e. the tenant traffic is IP multicast), there are a few issues that need to be addressed.

First, in all cases where L2 virtual network interfaces (VNIs) are present, the NVE would need to support IGMP/MLD snooping in order to prevent delivery of packets to tenant systems that are not interested in receiving them.

Second is the issue of how the groups are setup and mapped to tunnels

in the underlay. This can be accomplished entirely by an NVA if the mechanisms described in Section 2.2 or Section 2.3 are used, with the NVE just participating in snooping of IGMP messages from the tenant systems. If the method of Section 2.4 is used, then a mechanism must be provide for mapping the tenant IP multicast address to an IP multicast address for use in the underlay, and the NVE would be required to translate the information from the snooped IGMP/MLD messages from the tenant systems into corresponding requests for the underlay.

Third, when using the scheme described in Section 2.3, it may be useful to have the multicast service node support the IGMP querier function.

Fourth, if the IP multicast traffic is contained within a single virtual network (VN), then the schemes described herein are sufficient. If, on the other hand, the IP multicast traffic needs to traverse VNs, then the routing mechanisms at the NVE need to offer IP multicast forwarding. Once again, depending on how the groups are setup -- whether by an NVA or some other entity -- the forwarding tables at the NVE that has L3 virtual network interfaces (VNIs) would need to be setup by that entity.

5. Summary

This document has identified various mechanisms for supporting 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.

6. Security Considerations

This is an informational document, and as such, does not introduce any new security considerations beyond what may be present in proposed solutions.

7. IANA Considerations

This draft does not have any IANA considerations.

8. References

8.1 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.

8.2 Informative References

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

[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.

[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.

Authors' Addresses

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

Linda Dunbar

Huawei
Email: ldunbar@huawei.com

Vinay Bannai
Paypal
Email: vbannai@paypal.com

Ram Krishnan
Brocade
Email: ramk@brocade.com