August 2022

SUPPORTING L2 MULTICAST IN L2VNI-ONLY DEPLOYMENTS

Rajeev Kumar

Sanjay Hooda

SUPPORTING L2 MULTICAST IN L2VNI-ONLY DEPLOYMENTS

AUTHORS:
Rajeev Kumar
Sanjay Hooda

ABSTRACT

Techniques are presented herein that support an optimized Layer 2 (L2) multicast capability in software-defined access (SDA) fabric environments by modifying Internet Group Management Protocol (IGMP) snooping, IGMP forwarding, and multicast forwarding behavior in such environments. Such modifications help to address multiple customer requirements where customers need multicast traffic to be forwarded at L2, especially for Internet of Things (IoT) devices. Aspects of the presented techniques encompass modifying multicast router (mrouter) learning, handling IGMP control traffic, and handling sources and receivers.

DETAILED DESCRIPTION

Under a software-defined access (SDA) architecture, multicast communication may be supported only in Layer 3 (L3) SDA deployments and not in Layer 2 (L2) only deployments. In L2-only deployments, a multicast communication is treated as broadcast traffic and it is distributed to the entire fabric using fabric flood (i.e., broadcast-underlay).

Flooding a multicast communication to potentially thousands of switches at a site leads to a replication of the multicast traffic to places where it is not required (and where it is ultimately dropped), the utilization of bandwidth, and central processing unit (CPU) spikes as all of the traffic is travelling everywhere. These issues render such a solution non-deployable for customers who have multicast use cases.

To address the type of challenge that was described above, techniques are presented herein that support making a L2 multicast work like a L3 multicast in fabric environments, thus avoiding the flooding of a multicast communication to every switch in a network.

The presented techniques support modifying the multicast components in a L2 virtual extensible local area network (VXLAN) network identifier (L2VNI) environment. Specifically, aspects of the presented techniques encompass modifying multicast router

(mrouter) learning, handling Internet Group Management Protocol (IGMP) control traffic, handling sources and receivers, etc. The below narrative describes and illustrates the details for how each of these components work in tandem to optimize L2 multicast communication in fabric environments.

A first component that may be modified according to the techniques presented herein encompasses mrouter learning in a fabric environment. Traditionally, an mrouter learns behind an interface. In fabric based L2 multicast environments (if the desire is to optimize L2 multicast), such an approach is not sufficient. To optimize for L2 multicast, the key for an mrouter location becomes the combination (interface, Routing Locator (RLOC, behind which the mrouter exists)). Such a (interface, RLOC) combination allows the fabric edge (FE) nodes to send IGMP joins, IGMP queries, IGMP leaves, etc. to the mrouter that is located behind the RLOC thus removing the need to flood for such messages.

According to aspects of the techniques presented herein, mrouter learning encompasses a number of activities. For example, under a first activity, switches – upon enabling IGMP snooping (which may be a default behavior) – may be modified to support mrouter learning over a L2 Locator/ID Separation Protocol (L2LISP) interface and behind an RLOC.

Under a second activity, when a switch virtual interface (SVI) outside of a L2 fabric is Protocol-Independent Multicast (PIM)-enabled, it keeps sending periodic general queries (GQs). Under current behavior, such queries are flooded in the virtual local area network (VLAN). Each FE snoops those GQs to learn about an mrouter interface on a L2LISP0.x interface and, additionally, store the RLOC of the FE and the L2 backup designated router (L2BDR) behind which the mrouter is located.

Under a third activity, in addition to the mrouter (L2LISP0.x, RLOC), a CPU is also added as an mrouter port. Under a fourth activity, in a pure L2 multicast deployment it is possible to have a querier also in place of an SVI outside border or behind any fabric node. Under such scenarios, the above-described behavior will still work for a pure L2 environment.

Figure 1, below, depicts elements of mrouter learning according to aspects of the techniques presented herein and reflective of the above discussion.

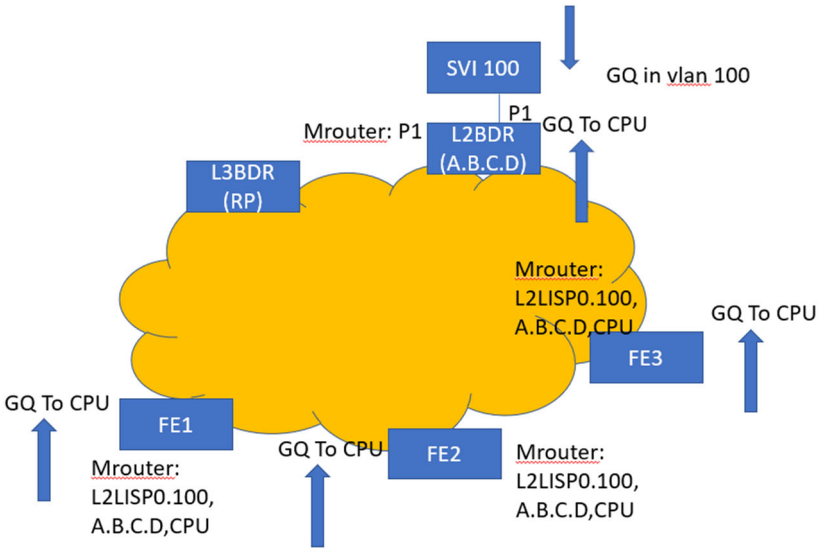

Figure 1: mrouter Learning on Various L2 Fabric Nodes

A second component that may be modified according to the techniques presented herein encompasses client IGMP join processing behavior. According to aspects of the presented techniques, such modifications encompass a number of activities.

Under a first activity, an IGMP join may be encapsulated in a unicast VXLAN packet and sent to the RLOC behind which the mrouter exists. Such behavior allows an mrouter to process the IGMP join messages so that it may begin forwarding multicast traffic towards the client. Under a second activity, based on an IGMP join message entries may be created so that a client as receiver and a client as source are properly handled. A detailed description of such handling is provided below in connection with the description of a third component that may be modified according to the techniques presented herein.

Under a third activity, an IGMP snooping entry:

(V,G) -> P1, (L2LISP0.x, RLOC), (L2LISP0.x, G')

may be created where P1 is a local port on the fabric edge and the combination (L2LISP0.x, RLOC) represents the mrouter entry. The combination (L2LISP0.x, G') characterizes other receivers in the fabric, where G' is the mapped underlay group for the overlay group G. In essence, when an FE receives an IGMP join (from, for example, port P1 on VLAN V) for G, it forwards the join message to the mrouter over (L2LISP0.x, RLOC) through software encapsulation and installs a group entry in VLAN V.

Under a fourth activity, a fabric edge also joins the underlay group G'. Such an underlay join will result in the creation of an underlay PIM entry (for G') as follows:

(*,G') -> Intf towards RP 'A', L2LISP0 L2LISP Decap 'F'

The above-described behavior results in pulling the traffic from a multicast source to the multicast receivers. Figure 2, below, depicts elements of such an approach according to aspects of the techniques presented herein and reflective of the above discussion.

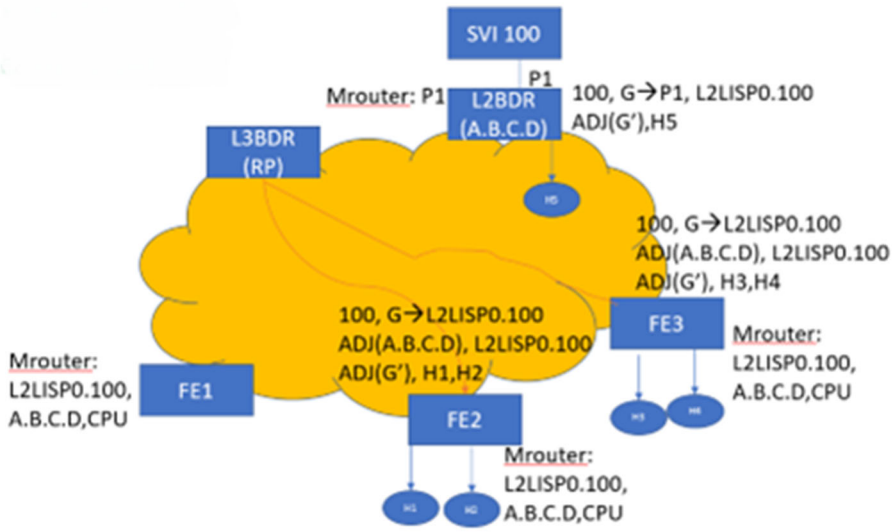

Figure 2: Receivers on Various L2 Fabric Nodes

In Figure 2, above, the elements H1, H2, H3, H4, and H5 represent receivers for G; G' represents a mapped group of "G" in an underlay; and the combination (*, G') maps to the elements FE1, FE2, and RP (in the underlay).

A third component that may be modified according to the techniques presented herein encompasses source handling, and itself comprises two different scenarios. As described and illustrated below, a first scenario entails the absence of an existing receiver and a second scenario entails the presence of an existing receiver.

Under the first scenario (reflecting the absence of an existing receiver), a source attaches on a fabric node which does not have programmed an IGMP snooping entry on (VLAN, G) as there are no local receivers. In this case of no existing receiver, only the mrouter configuration is present in the fabric edge. The traffic from the source is forwarded to the mrouter RLOC (which was already discovered through mrouter learning). Additionally, a copy of same is also forwarded to a CPU to create a snooping entry.

The CPU then programs a snooping entry: "(VLAN, G) -> (L2LISP0.x, RLOC), (L2LISP0.x, G')" into the IGMP snooping table. Once such a snooping entry is created, the traffic that is generated by the source hits the snooping entry, a copy is sent to the mrouter RLOC, and a second copy is placed on the underlay group G' for all of the other receivers to receive as multicast traffic.

For completeness, it is important to note that to send traffic in the underlay an entry is created: "(local RLOC, G') -> Null0 'A', intf towards tree 'F'" that indicates that the fabric edge is receiving the multicast traffic to Null0, where this multicast traffic is forwarded to the underlay.

The deletion of the snooping entry: "(VLAN, G) -> (L2LISP0.x, RLOC), (L2LISP0.x, G')" is important and may occur when the source has no longer sent any multicast traffic for some time interval (which, for example, may be a default value of three minutes). Such an action may be triggered by the expiry of the entry: "(local RLOC, G') -> Null0 'A', intf towards tree 'F'." In essence, upon the detection of the expiration of (localRLOC, G') the overlay IGMP snooping entry may be deleted (but only if no local interface is present in the snooping group entry).

Figure 3, below, depicts elements of such an approach according to aspects of the techniques presented herein and reflective of the above discussion.

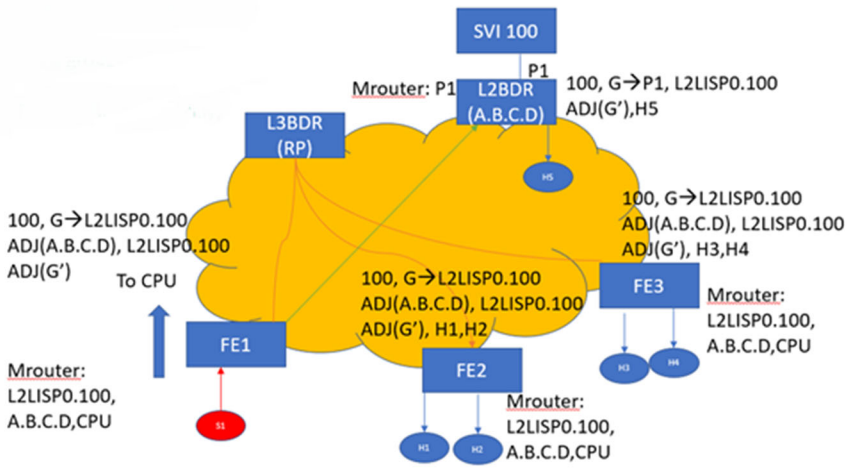

Figure 3: Source on a Node Without Any Receivers

In Figure 3, above, the element S1 represents a source for the group G; the elements H1, H2, H3, H4, and H5 represent receivers for G; G' represents a mapped group of "G" in an underlay; and the combination (*, G') maps to the elements FE1, FE2, and RP (in the underlay).

Under the second scenario (reflecting the presence of an existing receiver), a source attaches to a fabric node which already has (VLAN, G) programmed (i.e., there are one or more existing receivers on the FE). As there is an existing receiver, a snooping entry already exists on the fabric node. That existing snooping entry may be of the form: "(VLAN, G) -> (L2LISP0.x, RLOC), (L2LISP0.x, G'), P1, P2, …." where P1, P2, etc. are receiver ports on the fabric edge.

As the snooping entry is already present, the traffic hits this snooping entry resulting in a copy being sent to all of the local receivers that are connected on ports P1, P2, etc. A copy is also sent to the mrouter RLOC in the overlay, such copy encapsulated in the VXLAN with the destination as the RLOC of the mrouter. This allows the multicast traffic to be received outside of the fabric (e.g., an external network that is served by the mrouter). Additionally, one copy is placed on underlay group G'. That copy will be received by the receivers in the fabric which are connected to other fabric edges.

When the last local receiver leaves the group, the IGMP snooping entry is not immediately deleted (only the local port is removed from the group). The process that was described above in connection with the first scenario may be followed with waiting for the underlay (localRLOC, G') to age out after which the snooping entry may be deleted.

Figure 4, below, depicts elements of such an approach according to aspects of the techniques presented herein and reflective of the above discussion.

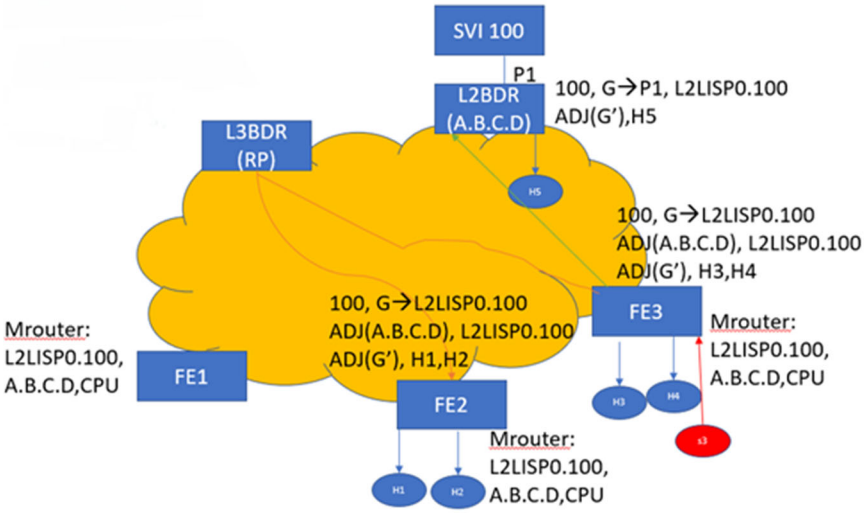

Figure 4: Source on a Node With Receivers

In Figure 4, above, the element S3 represents a source for the group G; the elements H1, H2, H3, H4, and H5 represent receivers for G; G' represents a mapped group of "G" in an underlay; and the combination (*, G') maps to the elements FE1, FE2, and RP (in the underlay).

Figure 5, below, depicts elements of an alternative arrangement (comprising a source that is located outside of a fabric) according to aspects of the techniques presented herein and reflective of the above discussion.

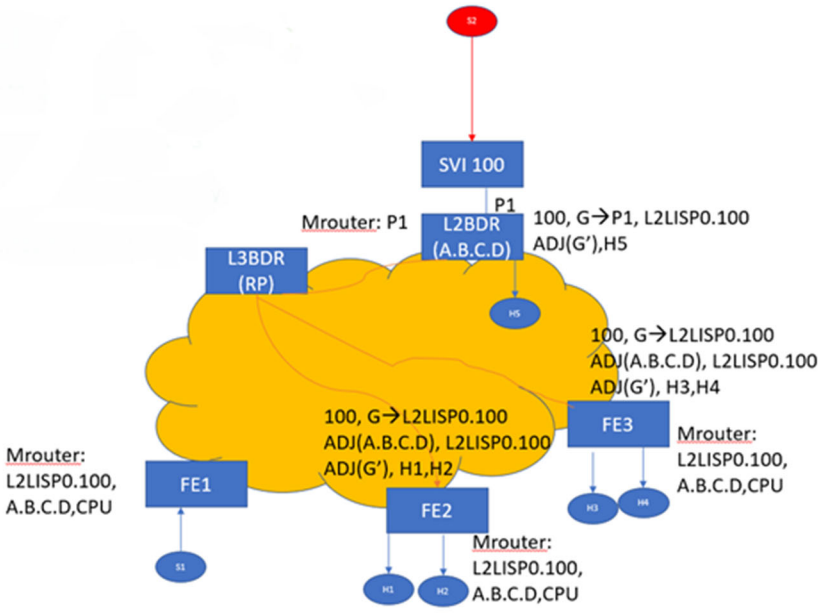

Figure 5: Source Outside of a Fabric

In Figure 5, above, the element S2 represents a source for the group G; the elements H1, H2, H3, H4, and H5 represent receivers for G; G' represents a mapped group of "G" in an underlay; and the combination (*, G') maps to the elements FE1, FE2, and RP (in the underlay).

Figure 6, below, depicts elements of a L2 multicast integrated view according to aspects of the techniques presented herein and reflective of the above discussion.

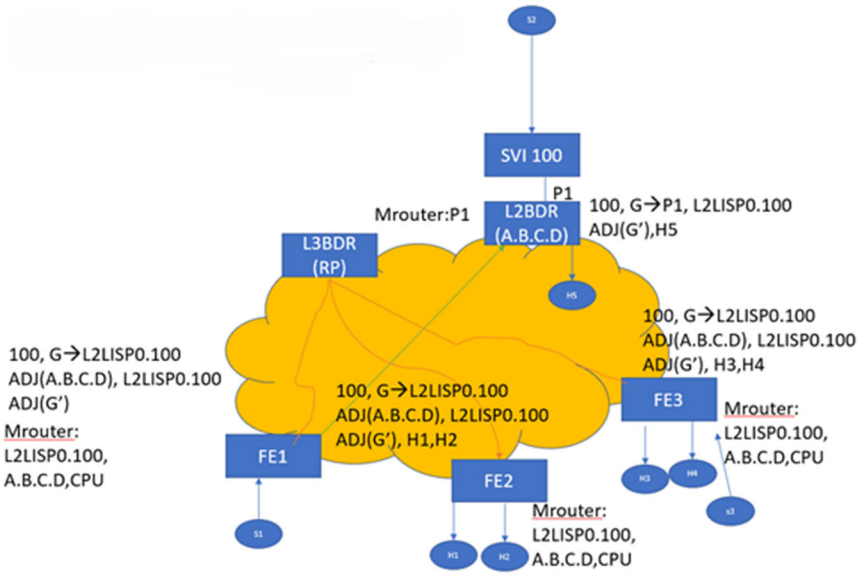

Figure 6: L2 Multicast Integrated View

A fourth component that may be modified according to the techniques presented herein encompasses GQ and Group Specific Query (GSQ) handling.

GQ and GSQ messages are generated by a multicast router (e.g., an SVI) or by a querier. A GQ is flooded in the fabric using a broadcast-underlay (this is also used for mrouter learning in the fabric). A GSQ is forwarded on the shared tree by a L2BDR and all of the receivers receive a copy.

A fifth component that may be modified according to the techniques presented herein encompasses IGMP control packet handling of a L2BDR (e.g., a L2 border where an mrouter is connected). According to aspects of the presented techniques, such modifications encompass a number of activities.

Under a first activity, when a fabric edge receives an IGMP join it encapsulates the join in a VXLAN and it sends it towards the L2BDR. The L2BDR decapsulates the IGMP join and then forwards it to the mrouter. Under a second activity, as the IGMP join is received over an L2LISP0.x interface, a group entry is installed on the L2BDR pointing towards a shared underlay multicast tree. Such an entry may be of the form: "(VLAN, G) -> (L2LISP0.x, G')" and ensures that even if the source is external, the traffic from outside the fabric can reach the receivers using the underlay G' entry.

Under a third activity, the IGMP leaves on the L2BDR may be tracked and if a leaf is a last leaf, then the entry may be marked for deletion. Such an entry may be deleted after the (localRLOC, G') entry expires.

It is important to note that the techniques presented herein, as described and illustrated in the above narrative, are generic in nature and do not depend upon the Locator/ID Separation Protocol (LISP), as used in an SDA. Aspects of the presented techniques employ LISP only for reachability information. Consequently, the presented techniques are not SDA-specific. However, in the case of other overlay technologies (such as a Border Gateway Protocol (BGP)-Ethernet virtual private network (EVPN)) an alternate means exists for distributing multicast-related information using BGP (as each node is connected through BGP for control plane exchanges).

Additionally, existing attempts at addressing the challenge that was described above may optimize multicast traffic distribution to unwanted nodes through new messages in BGP (such as to all of the sources and receivers through explicit tracking). In contrast, aspects of the techniques presented herein enable a L2 multicast capability in an overlay technology without modifying any existing protocols (such as BGP or LISP).

Further, as described above the techniques presented herein may be generalized for any overlay technology since the main concept of a distribution of multicast traffic in overlay technologies involves the use of Layer 3 multicast which does not work in a pure L2 environment. However, other technologies may already have a modified protocol to address such a case (e.g., L2 Tenant Routed Multicast (L2TRM) in BGP-EVPN).

In summary, techniques have been presented herein that support an optimized L2 multicast capability in SDA fabric environments by modifying IGMP snooping, IGMP forwarding, and multicast forwarding behavior in such environments. Such modifications help to address multiple customer requirements where customers need multicast traffic to be forwarded at L2, especially for Internet of Things (IoT) devices. Aspects of the presented techniques encompass modifying mrouter learning, handling IGMP control traffic, and handling sources and receivers.