January 24, 2019

ENABLING INSTANTANEOUS SWITCH FROM DEFAULT MULTICAST DISTRIBUTION TREE TO DATA MULTICAST DISTRIBUTION TREE FOR MOBILITY AND MULTIHOMING IN ETHERNET VIRTUAL PRIVATE NETWORK TENANT ROUTED MULTICAST

Raunak Banthia

Manoj Pandey

Kesavan Thiruvenkatasamy

Saif Mohammed

Vikas Ramesh Kamath

ENABLING INSTANTANEOUS SWITCH FROM DEFAULT MULTICAST DISTRIBUTION TREE TO DATA MULTICAST DISTRIBUTION TREE FOR MOBILITY AND MULTIHOMING IN ETHERNET VIRTUAL PRIVATE NETWORK TENANT ROUTED MULTICAST

AUTHORS:
Raunak Banthia
Manoj Pandey
Kesavan Thiruvenkatasamy
Saif Mohammed
Vikas Ramesh Kamath

ABSTRACT

Techniques are described for providing an optimized way of switching from a default Multicast Distribution Tree (MDT) to a data MDT to handle two key use-cases in Ethernet Virtual Private Network (EVPN) / Tenant Routed Multicast (TRM). The first key use-case is an Ethernet Segment Identifier (ESI) failover case over a multihomed network in an EVPN fabric. The second use-case is host mobility.

DETAILED DESCRIPTION

With Tenant Routed Multicast (TRM), once data traffic for a given Source exceeds a particular threshold value, the Provider Edge (PE) device (First-Hop Router (FHR)) originates Selective P-Multicast Service Interface (S-PMSI) Auto-Discovery (AD) route (Type-3). The S-PMSI route carries a tunnel attribute to indicate to peers that the Source has transitioned from using an inefficient tree called the default Multicast Distribution Tree (MDT) to using a more efficient tree called the data MDT. The data MDT is more efficient because when using the default MDT, multicast packets are routed to all PEs that belong to the Virtual Routing and Forwarding (VRF) entity, whether they have receivers or not. However, when using the data MDT, multicast packets are routed only to those PEs that have receivers attached to them.

The threshold criterion to originate the S-PMSI route means that the Source is monitored for a certain period of time called the evaluation period. It is only after the Source has been sending data at a high traffic rate during the entire evaluation period that the S-PMSI route is originated. The evaluation period may be several minutes (e.g., 3-6 minutes) because in order to ascertain that the Source is indeed sending traffic at a high

Published by Technical Disclosure Commons, 2019

rate before transitioning the source from the default MDT to the data MDT. Thus, there is an inherent cost of waiting when transitioning from the default MDT to the data MDT.

There are two problems with this approach.

First, a Source moves from one PE to another PE. Upon the move, the evaluation period must end in order to confirm that the traffic threshold has been exceeded even though this exercise was already done earlier on the older PE. Thus, upon every move, the Source would need to go from the data MDT to the default MDT and after the evaluation period has concluded successfully, move back to the data MDT. Since the evaluation period can be long, this means that the Source can end up using the inefficient default MDT for the entire evaluation period instead of switching to a more efficient data MDT right away.

Second, a similar problem occurs when an Ethernet Virtual Private Network (EVPN) Ethernet Segment Identifier (ESI)-based multi-homing is present. For instance, the traffic from the Source may hash to one of the ESI peers. That peer would go through the evaluation period and then originate the S-PMSI route. If the Source's packet hashes to any of the other ESI peers (e.g., because the earlier peer went down or the ESI link to that peer went down), then the new peer would have to wait for the entire evaluation period before concluding that the source is sending data at a high rate.

Figure 1 below illustrates an example network where the multicast source S1 is two-way multi-homed to PE1 and PE2. Here, the Source moves from one PE to another. In this case, it is necessary to wait for three minutes for the PE3 to originate a S-PMSI A-D (type 3) route. This behavior may be optimized.

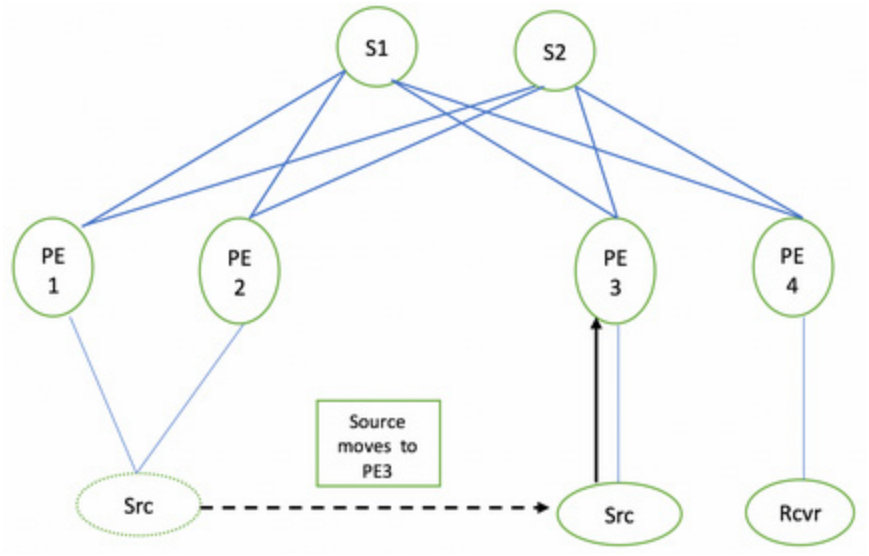

Figure 1

As illustrated in Figure 2 below, the data traffic may hash onto one of the multi-homed PE (e.g., PE1). When data traffic from this multicast source exceeds a certain threshold, PE1 may wait for some time (e.g., three minutes) to ensure the Source is consistently sending high multicast traffic above the threshold and is not bursty. After three minutes, PE1 may originate the S-PMSI A-D (type 3) route which goes to each of the PE's wherever the Multicast Virtual Private Network (MVPN) import Route Target (RT) is configured.

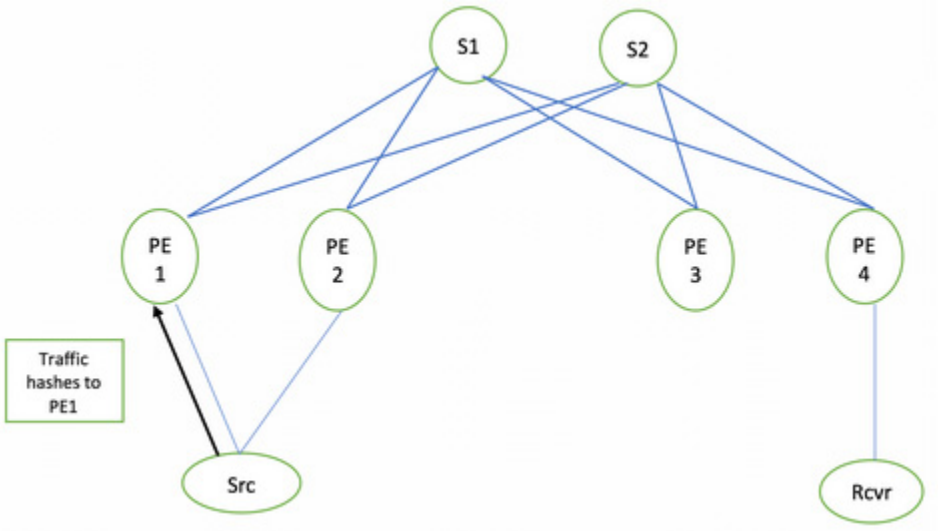

Figure 2

Figure 3 below illustrates PE1 going down. In this case, traffic may hash onto the other PE (PE2). As such, it is necessary to wait again for three minutes for PE2 to originate the S-PMSI A-D (type 3) route. This behavior may also be optimized.

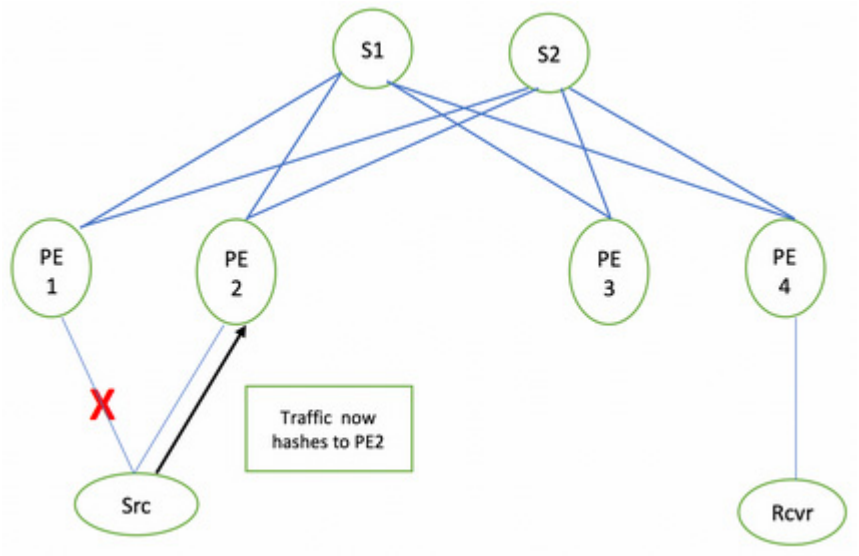

Figure 3

Accordingly, provided herein are two solutions at the FHR PE to handle this inefficiency. First, if a PE already has an S-PMSI AD route for a given (S, G) and then finds that the Source has become local, it should treat the presence of this route as a signal that the local Source is already high-traffic. Accordingly, the new PE should immediately originate an S-PMSI. Second, however, there can be cases where the S-PMSI AD route might get withdrawn before the Source becomes local. To handle these cases, the PE device may maintain a cache of the S-PMSI AD route once Border Gateway Protocol (BGP) gets a withdrawal. The entries in this cache may be time-stamped and purged after a certain time-period. If the PE finds that the Source has become local and it finds an entry in its cache, then it should treat that as a signal to know that the local Source is already high-traffic. Accordingly, the new PE should immediately originate an S-PMSI.

While one of the use-cases that benefits from this scheme is EVPN multi-homing, this approach is generic in nature and applies to non-multi-homed cases as well.

These techniques enable instantaneous switch from the default MDT to the data MDT for a mobility use case in the EVPN domain**.** Frequent source movement between PEs/pods is very common in the data-center environment. When a source / Virtual Machine (VM) / container moves from one PE to another PE, transitioning from the data MDT to

the default MDT is not desirable in the data-center environment. For example, this could lead to short term over-subscription/bombardment of the default MDT.

The traditional MVPN provides the ability to support multicast over a Layer 3 VPN. TRM uses MVPN signaling in the EVPN domain to provide optimized multicast forwarding between Layer 2 segments of the same Broadcast Domain (BD) or a different BD of the same tenant domain stretched across multiple PEs.

In a large data-center, there may be a large set of pods (e.g., greater than 25) with each pod having large number (e.g., greater than 150) of leaf nodes/PEs. In data centers, the VM/container/source mobility is a key requirement. Source movement from one PE to another PE within each pod, across pods, or between data-centers happens quite often. This mobility use case scenario is not applicable for traditional MVPN (e.g., Layer 3 VPN).

The techniques described herein offer a solution to seamlessly transition a VM from one PE to another PE without altering the underlying multicast tree. Enabling an immediate switch for all traffic is not acceptable in the high scale data-center environment. This requires large encapsulation/decapsulation entries to be programmed in the leaf devices, which do not support a large scale set. If overlapping groups are used, that may mitigate the scaling issue, but that impacts optimal forwarding, since the leaf device could draw unnecessary traffic for which it does not have a receiver.

Also, for some high bandwidth traffic, an immediate switch is not desirable. In fact, it is desirable to switch only after some finite time period. If source movement occurs frequently for such traffic streams (which is very common in data-centers), it is necessary to wait for that finite time after each VM move. The techniques described herein may help overcome this problem.

In summary, techniques are described for providing an optimized way of switching from a default MDT to a data MDT to handle two key use-cases in EVPN/TRM. The first key use-case is an ESI failover case over a multihomed network in an EVPN fabric. The second use-case is host mobility.