Multicast DF Election for EVPN Based on bandwidth or quantity

draft-liu-bess-evpn-mcast-bw-quantity-df-election-00

Yisong Liu (Huawei)
Michael McBride (Huawei)

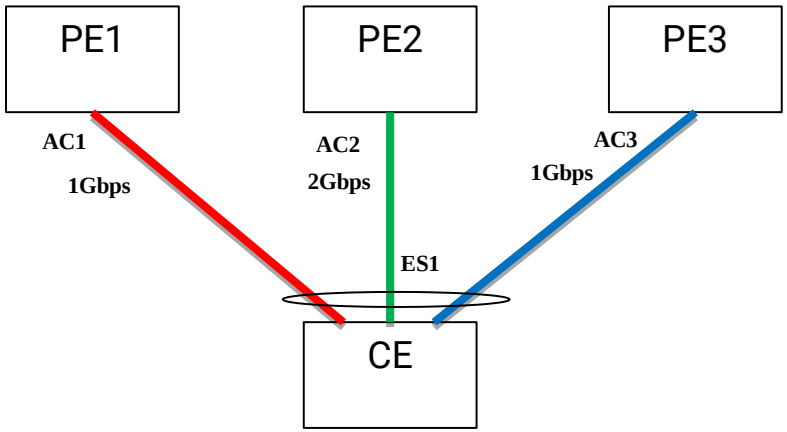
IETF104

Problem Statement

- RFC7432 defined basic method of DF election by a modulo algorithm
- draft-ietf-bess-evpn-df-election-framework
- ✓ defined DF election extended community, can use different DF election algorithms by notification
- ✓ Improve the basic DF election by a HRW algorithm
- draft-ietf-bess-evpn-per-mcast-flow-df-election proposes a method for DF election by enhancing the HRW algorithm, adding the source and group address of the multicast flow as hash factors

Problem Statement

- The relationship between the bandwidth of the multicast flows and the link capacity of different PEs, to the same CE device, is not considered in any of the current DF election algorithms.
- Result in severe bandwidth utilization of different links due to different bandwidth usage of multicast flows



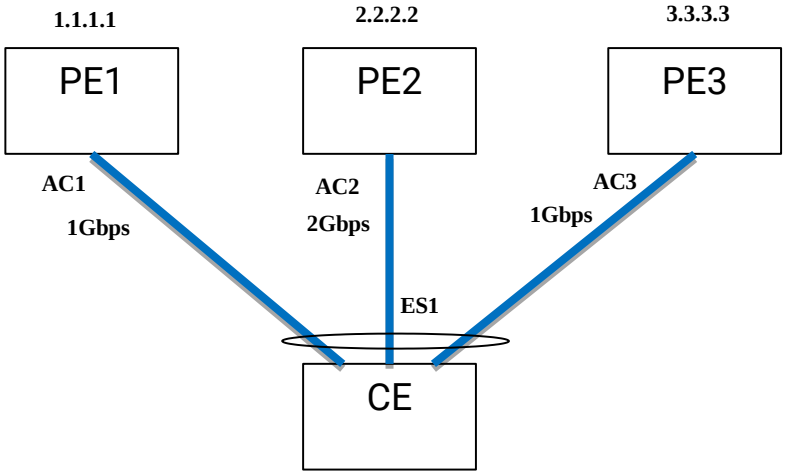
- AC1/AC2/AC3 belong to the same ES of multi-homed CE. The bandwidth of each link is as shown in the figure.
- For example, G1-G9 all join from CE， G1-G5 90Mbps per group， G6-G9 150Mbps per group. DF election for these groups：
  - AC1:G1,G6,G7,G8,G9: total 690Mbps
  - AC2:G2,G3： total 180Mbps
  - AC3:G4,G5： total 180Mbps

Solution Overview

- DF election based on bandwidth or state quantity
- ✓ The ratio of the current multicast flow bandwidth value to the link bandwidth weight is calculated according to the bandwidth weight of each multi-homed link, and the link with the smallest ratio is elected as the new multicast flow DF
- ✓ The ratio of the current multicast flow state quantity to the link bandwidth weight is calculated according to the bandwidth weight of each multi-homed link, and the link with the smallest ratio is elected as the new multicast flow DF
- In above 2 types of algorithms, if there are multiple PEs with the same calculated ratio, the DF is elected according to the method of maximum bandwidth weight of the link or maximum IP address of the EVPN peer
- In above 2 types of algorithms, draft-ietf-idr-link-bandwidth defines the link bandwidth extended community, it can be reused to transfer the link bandwidth value of the local ES to other multi-homed PEs

Solution based on multicast bandwidth

- Each PE obtains the link bandwidth values of the other multi-homed PEs according to the extended community of the Link bandwidth, and calculates the link bandwidth weight ratio AC1:AC2:AC3=1:2:1
- When the CE sends an IGMP join to one of the PEs, like PE1, PE1 advertises the PE2, PE3 by the EVPN IGMP Join Synch route （ Type 7 ）
- Each PE calculates the ratio of the current multicast flows bandwidth to the link bandwidth weight. The one PE in PE1, PE2, PE3, which has the smallest ratio, is elected as the DF of the new multicast flow. When the smallest ratios of more than one PE are the same, for example the PE with the maximum EVPN peer IP address is elected as the DF

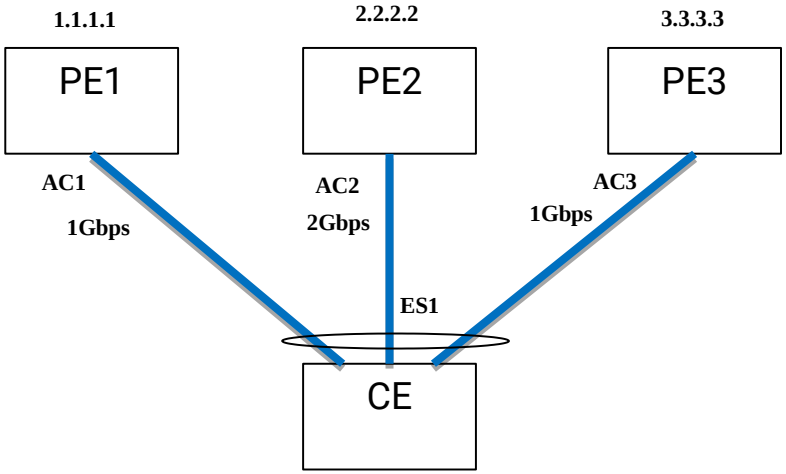


- G1-G5 : 90Mbps
- G6-G9 : 150Mbps
- AC1： G7,G5 total 240Mbps
- AC2: G2,G8,G6,G9 total 540Mbps
- AC3： G1,G3,G4 total 270Mbps

Group	PE1 Current Bandwidth Ratio	PE2 Current Bandwidth Ratio	PE3 Current Bandwidth Ratio	Elected DF
G1	0/1	0/2	0/1	PE3
G2	0/1	0/2	90/1	PE2
G7	0/1	90/2	90/1	PE1
G8	150/1	90/2	90/1	PE2
G3	150/1	240/2	90/1	PE3
G6	150/1	240/2	180/1	PE2
G5	150/1	390/2	180/1	PE1
G4	240/1	390/2	180/1	PE3
G9	240/1	390/2	270/1	PE2

Solution based on multicast state quantity

- Each PE obtains the link bandwidth values of the other multi-homed PEs according to the extended community of the Link bandwidth, and calculates the link bandwidth weight ratio AC1:AC2:AC3=1:2:1
- When the CE sends an IGMP join to one of the PEs, like PE1, PE1 advertises the PE2, PE3 by the EVPN IGMP Join Synch route （ Type 7 ）
- Each PE calculates the ratio of the current multicast flows state quantity to the link bandwidth weight. The one PE in PE1, PE2, PE3, which has the smallest ratio, is elected as the DF of the new multicast flow. When the smallest ratios of more than one PE are the same, for example the PE with the maximum EVPN peer IP address is elected as the DF

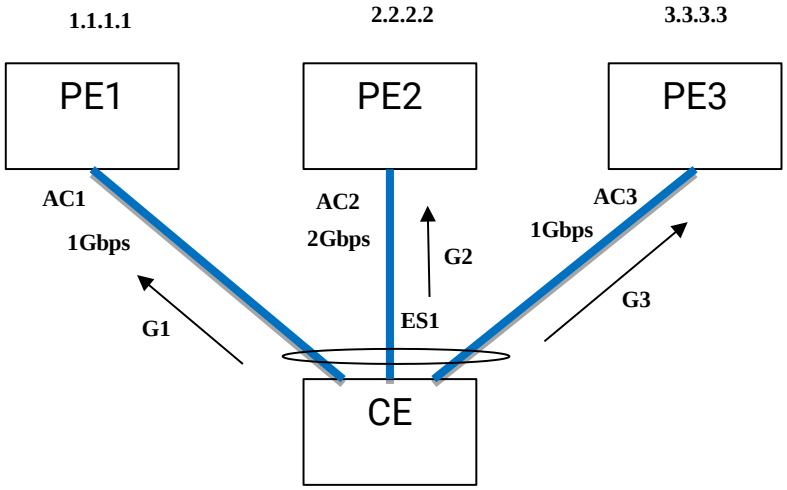


- G1-G10 : No specific bandwidth value, the uniform bandwidth is 1
- AC1： G7,G5 total 2 states
- AC2: G2,G8,G6,G4 total 4 states
- AC3： G1,G3,G9 total 3 states

Group	PE1 Current Bandwidth Ratio	PE2 Current Bandwidth Ratio	PE3 Current Bandwidth Ratio	Elected DF
G1	0/1	0/2	0/1	PE3
G2	0/1	0/2	1/1	PE2
G7	0/1	1/2	1/1	PE1
G8	1/1	1/2	1/1	PE2
G3	1/1	2/2	1/1	PE3
G6	1/1	2/2	2/1	PE2
G5	1/1	3/2	2/1	PE1
G4	2/1	3/2	2/1	PE2
G9	2/1	4/2	2/1	PE3

Inconsistent Timing among Multi-homed PEs

- The inconsistent processing timing of the same multicast group joining process among PEs may cause electing different DFs. For example:



- Multicast group G1, G2, and G3 join packets are sent from the CE to PE1, PE2 and PE3.
- PE1 calculates the DF of G1, while PE2 calculates the DF of G2, and PE calculates the DF of G3, and at this moment each PE has not received the EVPN Join Synch route.
- PE1, PE2 and PE3 select the link with the largest bandwidth, such as AC2, so that the same DF PE2 may be elected for G1, G2, and G3.
- After receiving the EVPN Join Synch route sent by PE2, PE1 may calculate the DF of G2 as PE3, which is inconsistent with the calculation result of PE2.

- EVPN Join Synch routes need to carry elected DF information in the route advertisement as the extended community called Multicast DF Extended Community, which can make the DF information for a given multicast flow state between PEs consistent

Decrease of Multi-homed PEs

- The multicast flows destined to the failed PE need to be in a specific order to reassign the DF
  - for example, the group and source address ascending order
- The DF election calculation is completed by one of the specified multi-homing PEs, and the specified calculated PE can be selected according to the link bandwidth weight value or the IP address of the EVPN peer
- The specified PE needs to advertise each DF election result to the other multi-homed PEs by the EVPN Join Synch route carrying the Multicast DF Extended Community
- If a new multicast join is received in the above calculation process, the DF election calculation of the new multicast flow is still completed by the PE receiving the multicast join packet (not by the specified PE)

Increase of Multi-homed PEs

- Option 1: No active adjustment.
- ✓ The DF of the subsequent new multicast flow is elected to make approximately equalized among multi-homed PEs
- Option 2: Active adjustment
- ✓ Each time consider calculating the ratio of using the DF election algorithm, the multicast entries, whose ratio of the existing multi-homed PE is the largest, are migrated to the new PE.
- ✓ The multicast entries are migrated in descending order of multicast flow bandwidth or in ascending order of the group and source address until the ratio of the new PE is greater than the existing smallest ratio of other multi-homed PEs.
- ✓ The calculation is performed by one specific PE among the multi-homed PEs, selected according to the link bandwidth weight value or the IP address of the EVPN peer.
- ✓ Start a timer to suppress the synchronization process from the new PE to other existing PE's for a while, to avoid receiving the existing multicast join as a new one

Next Step

- Questions and comments are welcomed
- Seeking co-authors involved