Multicast VPN Upstream Designated Forwarder Selection

draft-wang-bess-mvpn-upstream-df-selection-02

Fanghong Duan@Huawei

Siyu Chen@Huawei

IETF 115

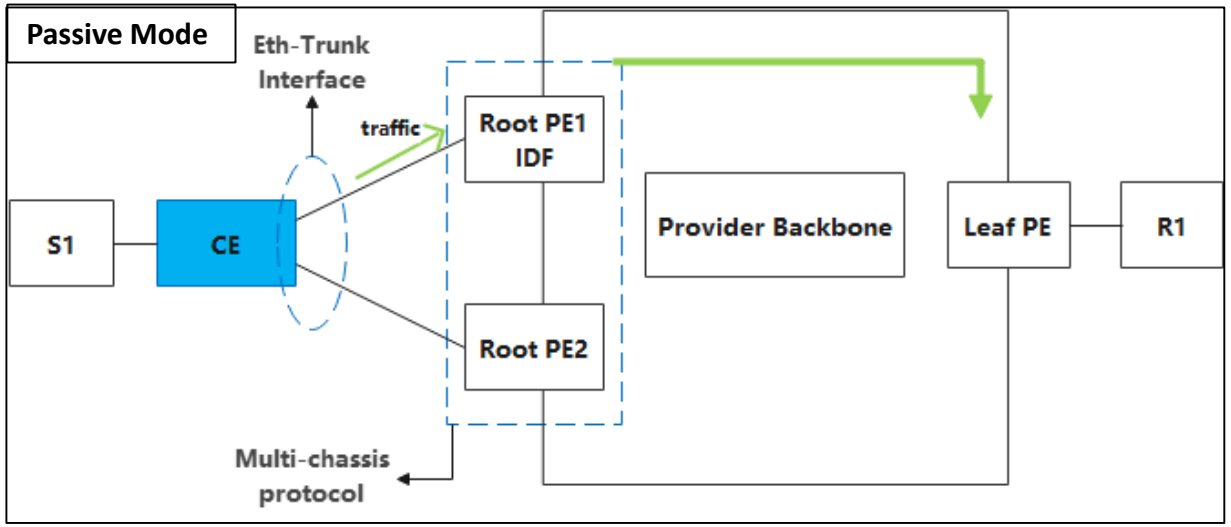
Nov. 2022

Background

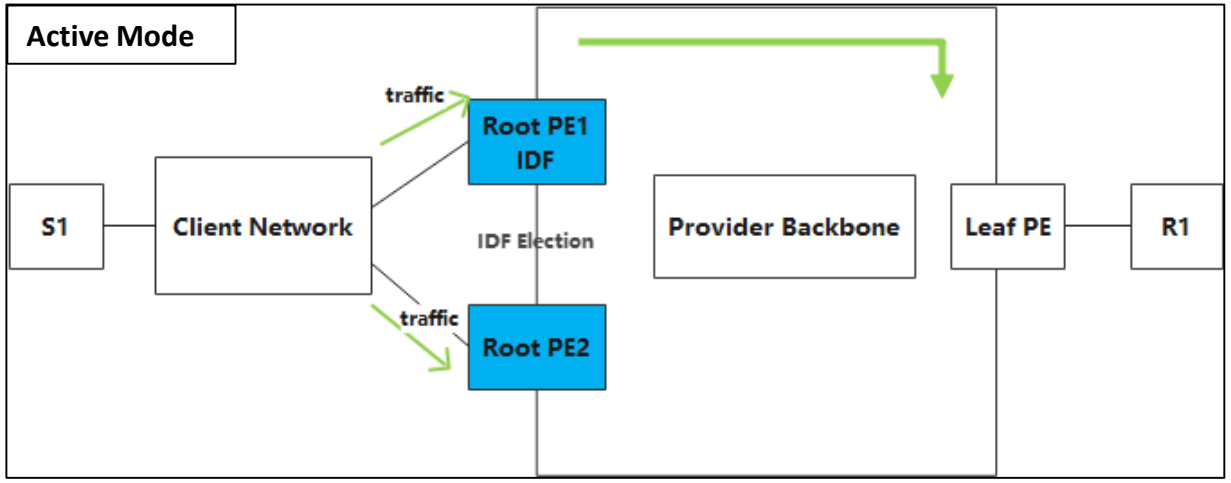
- Compared with “Hot Root Standby”, ”Warm Root Standby” avoids steady traffic redundancy and saves bandwidth
- [RFC9026] defines that UMH selection is conducted by leaf PEs based on Provider-Tunnel Status
- But,
  - Due to inconsistency of the primary PE considered by root and leaf PE, failover time cannot reach same level as “hot root standby”.
  - No endogenous mechanism to discover failure of primary PE.
  - Inconsistencies of transient unicast routing, P-Tunnel status, etc. -> Unstable “Warm Root Standby”.
  - All multicast traffic use the same primary and standby PE. Cannot perform load balancing.
- In previous versions of draft in IETF113&IETF114:
  - Upstream Designated Forwarder(DF) Selection by VRRP. →**This draft defines endogenous method for IDF election and fast failover.**
  - Upstream DF Selection Extended Community. → **IDF Negotiation Community and BFD Discriminator Attribute.**
  - Downstream PEs advertise C-multicast Route to both Primary and Standby upstream PEs and accept traffic from both sides.
  - Downstream PE performs “Anycast Reverse Path Forwarding(RPF) Checking”.

IDF Negotiation Mode

- Passive Mode
  - CE selects one member interface to forward traffic
  - Ingress Designated Forwarder(IDF) PE is decided by CE
  - Root PE accept the IDF role passively



- Active Mode
  - Client network contains one or more CEs
  - Interfaces are not bundled
  - Each root PEs can receive multicast traffic
  - Only one root PE send traffic to leaf

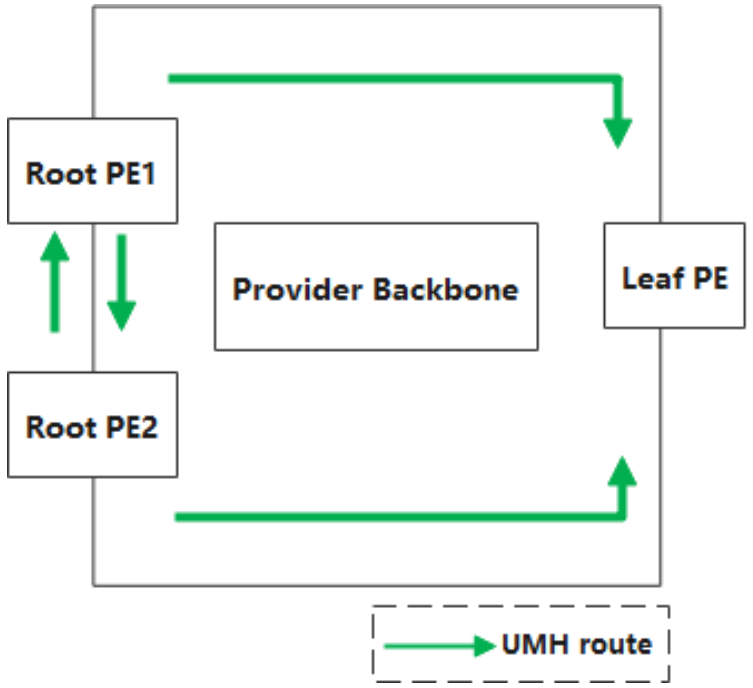


MVPN Extensions

- IDF negotiation Community
  - Format:
    - Carried in UMH routes (to client multicast source)
    - To be allocated from “BGP Well-known Communities” registry for each mode
  - **Function:**
    - Notify other root PEs to perform IDF election
    - A symbol for leaf PE to add root PE to anycast RPF checklist
- BFD Discriminator Attribute
  - **Format:**
    - Carried in UMH routes
    - Reuses the format defined in RFC 9026

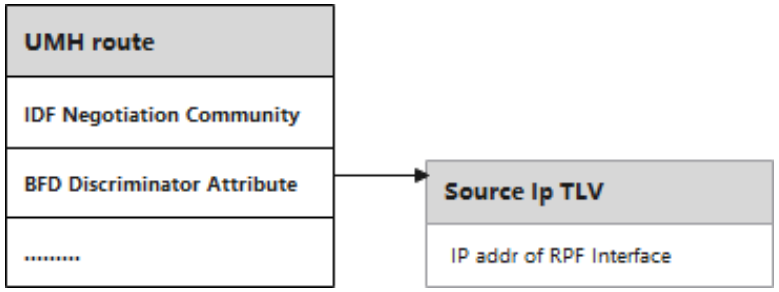
```
 0                   1                   2                   3
 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1
+-+-+-+-+-+-+-+-+
|    BFD Mode   |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|                       BFD Discriminator                       |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
~                         Optional TLVs                         ~
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
```

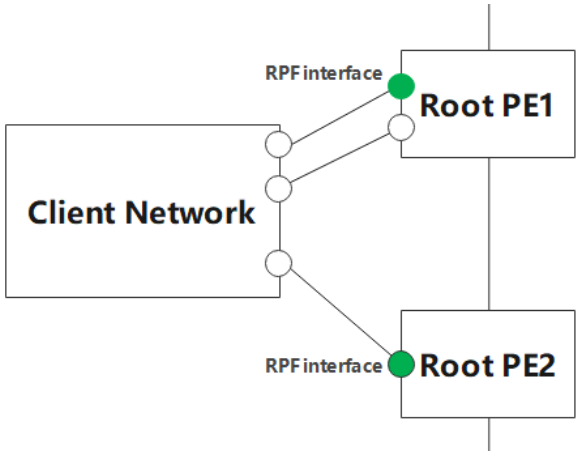
    - BFD Mode: Redefined as unicast BFD session type, value is 2
    - Source IP optional TLV: Mandatory
  - **Function:**
    - Establish a BFD session to detect the failure of primary IDF PE



IDF Election Procedure

- Root PEs originate UMH routes:



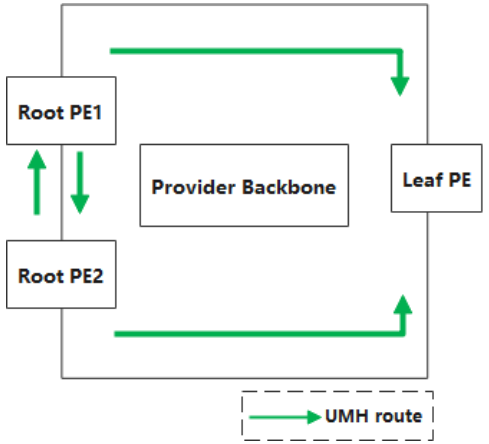


  - IP address of Source IP TLV: establish a BFD session to do fast tracking of IDF failure

- Leaf PEs:
  - Originate distinct C-multicast routes to each root PEs
  - Installs P-Tunnels into anycast RPF tunnel checklist
  - Traffic received from each P-Tunnel in checklist is valid

IDF Election Procedure(cont.)

- Endogenous Mechanism for IDF Election
  - Each root PEs learn prefix of source
  - RDs of multi-homed root PEs for a same MVPN are distinct
  - Root PEs originate VPN route (to client multicast source) with originator IP address of PE
  - VPN route sent to other root PEs and leaf PEs
  - Each PE builds an ordered list of IP addresses of all root PEs in ascending IP order



- **Election method a): All C-Gs use one primary IDF**

  Election occurs upon receiving all UMH routes of other PEs
  - PE Index represents its position, 0 corresponding to the lowest IP address
  - **IDF: PE with Index 0; Standby IDF: PE with index 1**

Ordered list for all C-G:

Index	IP	Role
0	1.1.1.1	**Primary IDF**
1	2.2.2.2	**Standby IDF**
2	3.3.3.3	Common root PE
……	……	……
N	N.N.N.N	Common root PE

IDF Election Procedure(cont.)

- **Election method b) :Different C-G can use different IDF (Load Balancing)**
  - Election occurs upon root PEs receiving C-multicast join of corresponding C-G
  - **IDF: PE with index i, i = (C-G mod N)**
  - **A new ordered list without the elected primary IDF**
  - **Standby IDF: PE j, j = (C-G mod (N-1))**

When (C-G mod N) = 1:

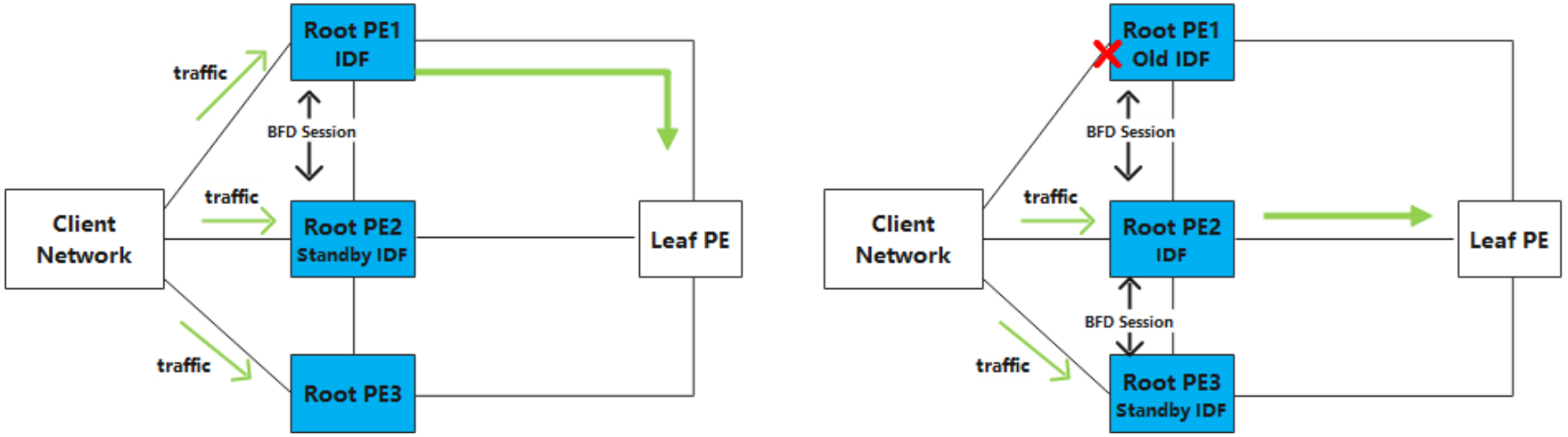
Index	IP	Role
0	1.1.1.1	Common root PE
1	2.2.2.2	**Primary IDF**
2	3.3.3.3	Common root PE
……	……	……
N	N.N.N.N	Common root PE

When (C-G mod (N-1)) = 0:

Index	IP	Role
0	1.1.1.1	**Standby IDF**
2	3.3.3.3	Common root PE
……	……	……
N-1	N.N.N.N	Common root PE

Failure detection and fast failover

- Endogenous mechanism for Active IDF Mode
- Detect **failure of IDF node or client facing link** of IDF quickly
- Standby IDF: Initializes a BFD session
    - Destination IP address: from Source IP TLV of BFD Discriminator Attribute of IDF



- If obsoleted IDF PE recovers and it needs to failback:
    - Obsoleted PE establishes multicast path towards SDR
    - When failback time expires, running IDF establishes the BFD session with the obsoleted PE
    - Running IDF stops sending multicast traffic and obsoleted IDF become the new IDF
    - New IDF sends multicast traffic to leaf

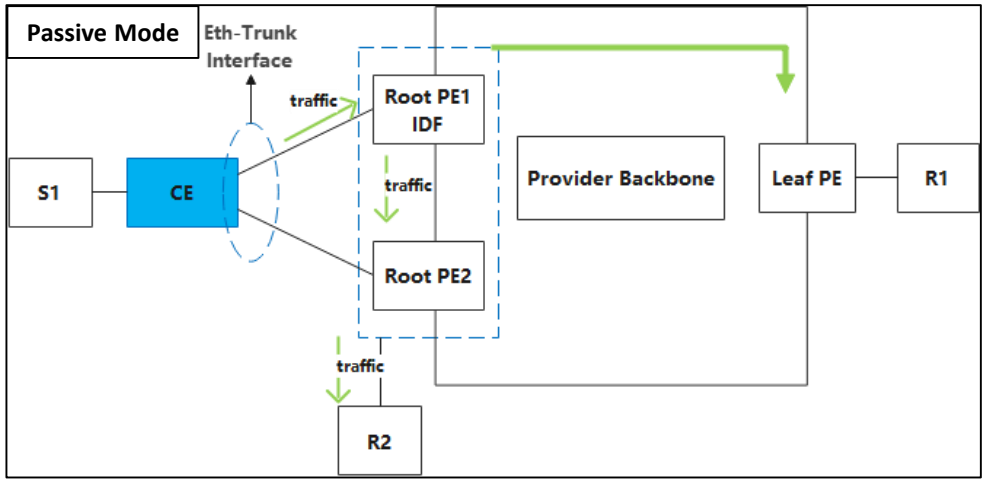
Data forwarding

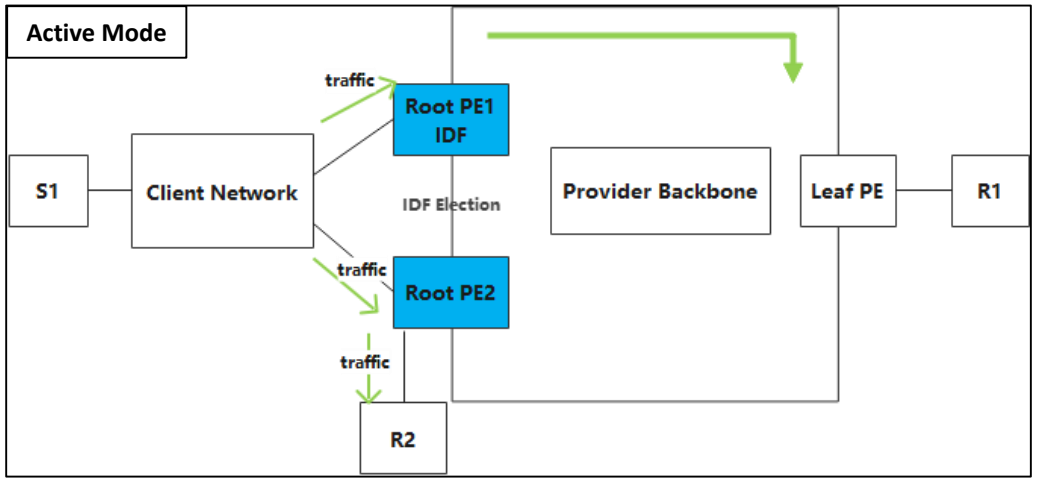
- Root PEs

  Passive mode, root PE has local receivers:

  - When one PE is selected as IDF, the other PEs may have local receivers
  - PEs need perform anycast RPF checking on client facing interface or IDF PE oriented P-tunnels when receiving traffic
  - Unidirectional forwarding: send traffic only to local receivers

  Active mode: All root PEs can send traffic to local receivers, but only primary IDF send data to leaf PEs





- Leaf PEs
  - Install each P-Tunnel into anycast RPF checklist for corresponding multicast flow (C-S, C-G)
  - Accept traffic from each root PEs
  - Accept traffic from standby IDF without latency

Next Steps

- Comments and discussion.

Thanks