A new Designated Forwarder Election for the EVPN

draft-mohanty-bess-evpn-df-election-02

IETF 94

Satya R. Mohanty

Ali Sajassi

Keyur Patel

Cisco Systems

John Drake

Juniper Networks

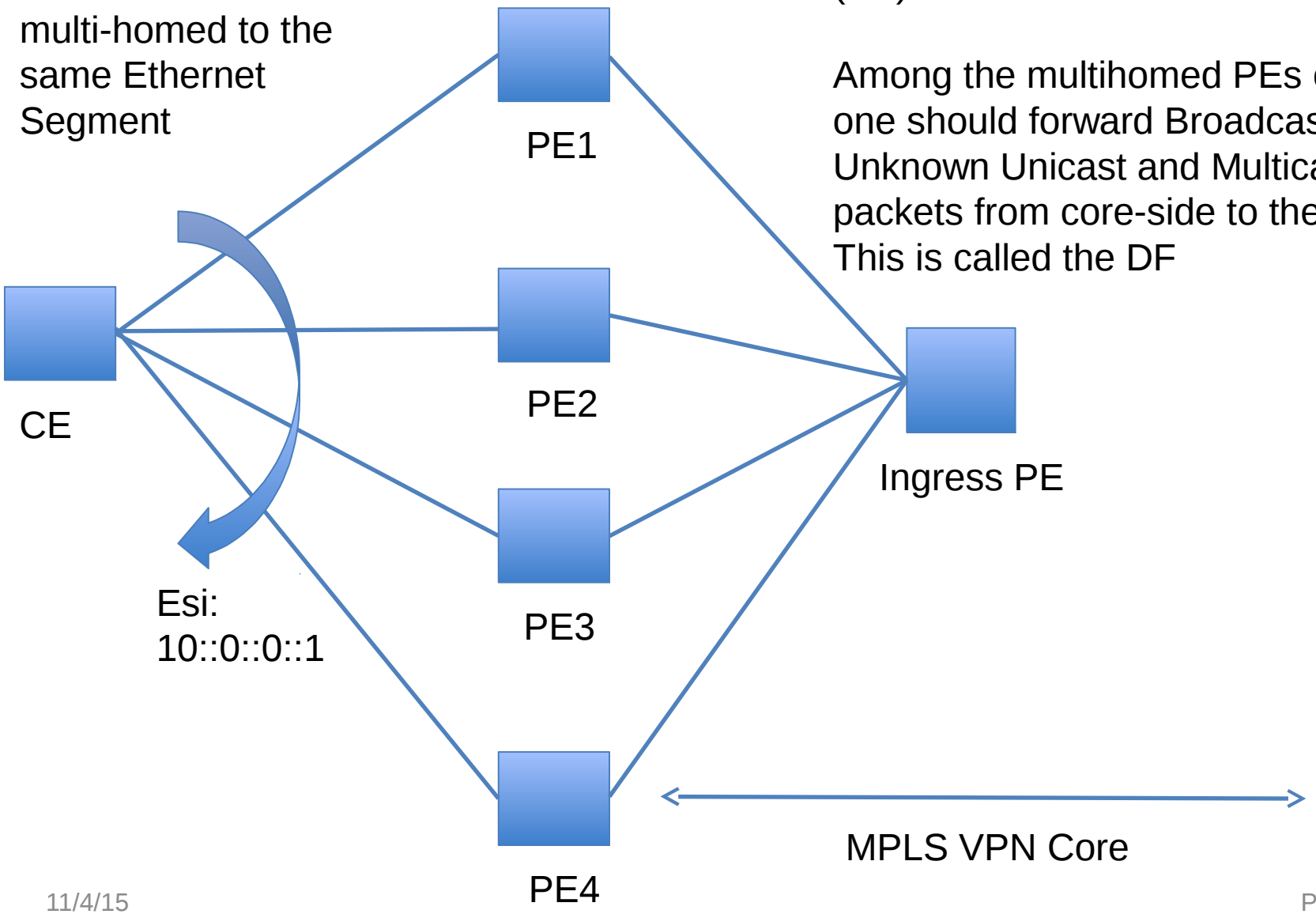
Antoni Przygienda

Ericcson

EVPN use-case where some PEs are multi-homed to the same Ethernet Segment

What is the Designated Forwarder (DF)?

Among the multihomed PEs only one should forward Broadcast, Unknown Unicast and Multicast packets from core-side to the CE. This is called the DF



Current DF election: All E-tags change DF, even when their DF did not go down

				when PE1 is down	
PE	PE IP address	Ord	Ethernet tag DF	Ord	Ethernet tag DF
PE1	192.0.2.1	0	892,896		
PE2	192.0.2.2	1	893	0	891,894
PE3	192.0.2.3	2	894	1	892,895
PE4	192.0.2.4	3	891,895	2	893,896

Proposed DF election: E-tag whose DF did not go down does not change.

PE	PE IP address	Ethernet tag DF	Ethernet tag DF
			when PE1 down
PE1	192.0.2.1	894	
PE2	192.0.2.2	892,893,895	892,893,895
PE3	192.0.2.3	891	891,894
PE4	192.0.2.4	896	896

For each Tag, the PE with ordinal == (V mod N) becomes DF

(V mod N) for tag/IP combinations

Eth Tag		891	892	893	894	895	896
ip address	ordinal						
192.0.2.1	0	3	0	1	2	3	0
192.0.2.2	1	3	0	1	2	3	0
192.0.2.3	2	3	0	1	2	3	0
192.0.2.4	3	3	0	1	2	3	0

Eth Tag		891	892	893	894	895	896
ip address	ordinal						
~~192.0.2.1~~							
192.0.2.2	0	0	1	2	0	1	2
192.0.2.3	1	0	1	2	0	1	2
192.0.2.4	2	0	1	2	0	1	2

When 192.0.2.1 goes down, all Tags change DF

For each Tag, the PE with IP with the greatest hash becomes DF

Hashes for tag/IP combinations

Eth Tag	891	892	893	894	895	896
ip address						
192.0.2.1	1030724564	501370518	227039903	786483140	769731393	1512711410
192.0.2.2	1443204555	1651686021	1683927472	166013787	2115159210	338879529
192.0.2.3	1474980878	599428380	1224551449	772514622	104185799	588040224
192.0.2.4	441543909	1306804267	1063370714	75805525	1254959328	1729765511

- When PE 192.0.2.1 goes down, hashes do not change.
- Only the Tag that used this PE for DF gets a new greatest hash.
- The second highest hash becomes the new highest hash, therefore it is the Backup DF.
- PE coming up is the reverse of PE going down.

Highest Random Weight

- Every PE computes hash ***H(***Pei, vj), for every Pei which is a DF participant
- Pek corresponding to highest value of ***H*** is the DF for vlan vj

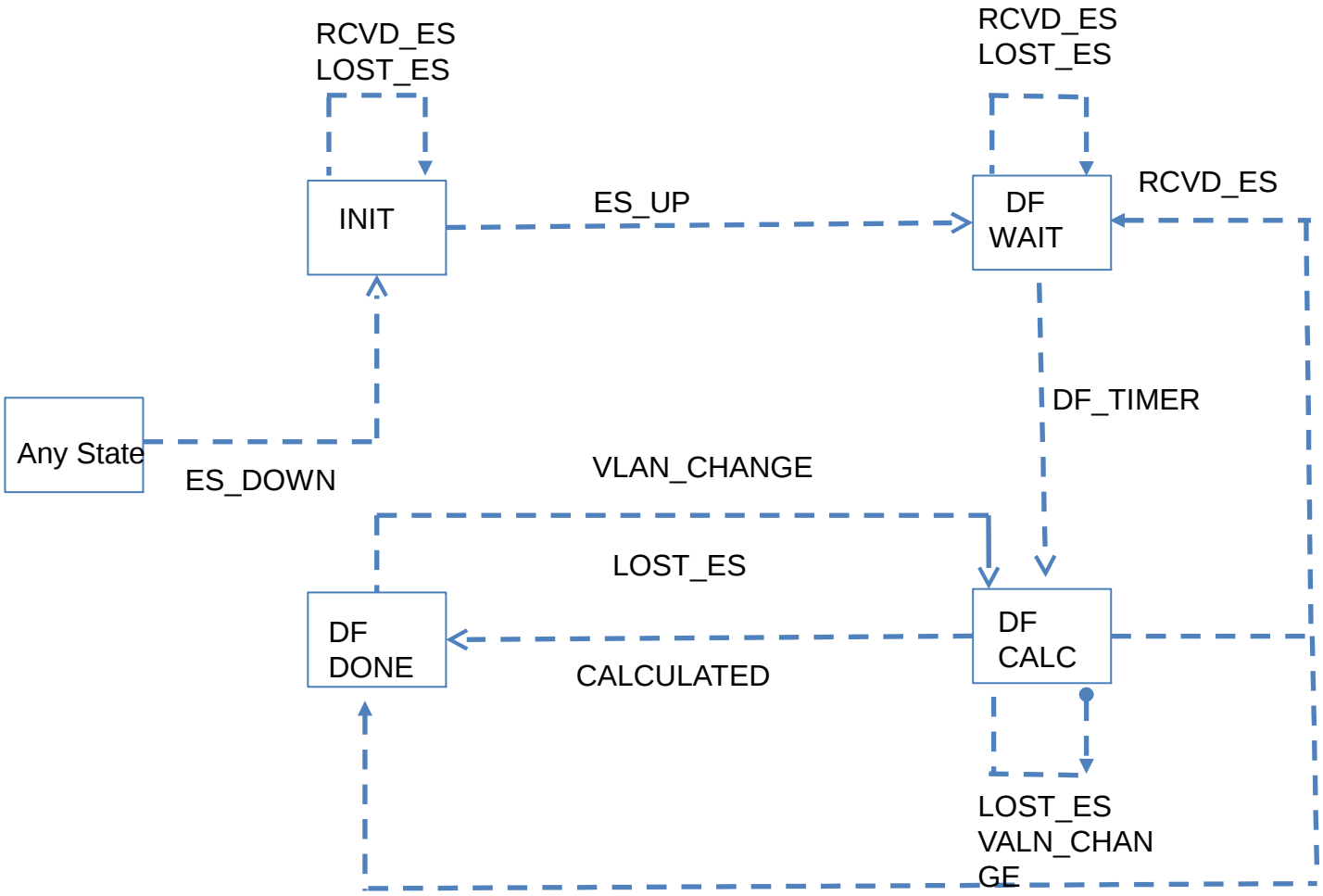
Suggested hash function
H = (1103515245 * ((1103515245 * Si + 12345) XOR CRC32(v)) + 12345)
Computed in modulo 0x7FFFFFFF arithmetic
Where
Si = IP address of PE
v = Ethernet Tag

Important property that ensures DF for a vlan does not move among unchanged PEs:

- The hash does not depend on the number of PEs

Two Updates from previous version

- State Machine for DF Election is proposed
- Section 7.6 of RFC7432 describes how the value of the ES-Import Route Target for ESI types 1, 2, and 3 can be auto-derived by using the high-order six bytes of the nine byte ESI value. This document extends the same auto-derivation procedure to ESI types 0, 4, and 5



EVPN Designated Forwarder Election Finite State Machine

States:

INIT: Initial State

DF WAIT: State in which the participants waits for enough information to perform the DF election for the EVI/ESI/VLAN combination.

DF CALC: State in which the new DF is recomputed.

DF DONE: State in which the according DF for the EVI/ESI/VLAN combination has been elected.

Events:

ES_UP: The ESI has been locally configured as 'up'.

ES_DOWN: The ESI has been locally configured as 'down'.

VLAN_CHANGE: The VLANs configured in a bundle that uses the ESI changed. This event is necessary for VLAN bundles only.

DF_TIMER: DF Wait timer has expired.

RCVD_ES: A new or changed Ethernet Segment Route is received in a BGP REACH UPDATE. Receiving an unchanged UPDATE MUST NOT trigger this event.

LOST_ES: A BGP UNREACH UPDATE for a previously received Ethernet Segment route has been received. If an UNREACH is seen for a route that has not been advertised previously, the event MUST NOT be triggered.

CALCULATED: DF has been successfully calculated.

ACTIONS:

1. ANY STATE on ES_DOWN: (i)stop DF timer (ii) assume non-DF for local PE

2. INIT on ES_UP: (i)Do nothing

3. INIT on RCVD_ES, LOST_ES: (i)Do nothing

4. DF_WAIT on entering the state: (i) start DF timer if not started or expired (ii) assume non-DF for local PE

5. DF_WAIT on RCVD_ES, LOST_ES: Do nothing

6. DF_WAIT on DF_TIMER: Do nothing

7. DF_CALC on entering or re-entering the state: (i) rebuild according list and hashes and perform election (ii) FSM generates CALCULATED event against itself

8. DF_CALC on LOST_ES or VLAN_CHANGE: Do nothing

9. DF_CALC on RCVD_ES: do nothing

10. DF_CALC on CALCULATED: (i) mark election result for VLAN or bundle

11. DF_DONE on exiting the state: (i)if RFC7432 election or new election and lost primary DF then assume non-DF for local PE for VLAN or VLAN bundle.

12. DF_DONE on VLAN_CHANGE or LOST_ES: Do nothing

Next step:
Request Working Group Adoption

Thanks!!!