Draft Objective

Optimally handle scenarios with unequal PE-CE link bandwidth distribution within a multi-homed EVPN-LAG bundle:

- Load-balance overlay unicast flows “unequally” in proportion to each PE’s link bandwidth share in a LAG
- Load-share DF role “unequally” in proportion to each PE’s link bandwidth share in a LAG

*Both overlay unicast and BUM flows load-balanced in proportion to PE-CE link bandwidth share in a LAG*
Overlay Load Balancing in proportion to PE-CE link bandwidth share in a LAG

- Hash 2/3rd flows to ES-1 via PE1
- Hash 1/3rd flows to ES-1 via PE2

Load-balance Distribution in proportion to ES link-bandwidth share

Unicast flows

L2 stretch
DF Role Load Sharing in proportion to PE-CE link bandwidth share in a LAG

PE1

PE2

PE-x

CE1

CE-x

ESI-1

L2 stretch

200Gb

100Gb

ES-1 DF for 2/3rd EVIs

ES-1 DF for 1/3rd EVIs

RT-4

BUM Replication across all EVIs

BGP-EVPN

ESI-1

ES-1 DF for 1/3rd EVIs

ES-1 DF for 2/3rd EVIs
Updates

• Expanded scope to include both unicast and BUM flows
• Detailed Procedures added to influence DF election based on link bandwidth share for each DF election algorithm (DF Type) (section 4):
  • Type 0: Default DF Election
  • Type 1: HRW algorithm
  • Type 2: Preference algorithm
  • Type 4: HRW per-multicast flow DF election
• Added applicability to RT-5 for a routed overlay use case (section 6)
• Clarified scope to be limited to “provisioned” available bandwidth as opposed to “real-time” available bandwidth (section 5)
• Allow BGP link-bandwidth attribute to be signaled to eBGP neighbors for inter-AS support (section 3.1)
• Collaboration and contributions from additional co-authors
Unicast Traffic Load-Balancing

- Local PE
  - Advertises per-ESI link-band-width attribute as part of per-ESI EAD RT-1
- Remote PE
  - ESI Path-list is computed in proportion to received link-band-width attribute from each PE

DF Election

- New “BW” capability bit (28) in DF Election Extended-Community indicates desire to augment specified DF election algorithm to be “BW aware” as specified in section 4 of this draft
- Local PE
  - Advertises additional per-ES link-band-width attribute with per-ES RT-4
- Remote PE
  - Type 0 (service carving): Candidate PE list computed in proportion to bandwidth share
  - Type 1 and 4 (HRW): Candidate hash computations for each PE in proportion to it’s bandwidth share
  - Type 2 (Preference): additional link-band-width tie-breaker based on PE’s bandwidth share
Draft Status

• Ready for WG adoption
Weighted Multi-Path Procedures for EVPN All-Active Multi-Homing
(draft-malhotra-bess-evpn-unequal-lb-04)

Thank You

Neeraj Malhotra (Arrcus), Ali Sajassi (Cisco)
Jorge Rabadan (Nokia), John Drake (Juniper)
Samir Thoria (Cisco), Avinash Lingala (AT&T)
BACKUP
Prior Art

- RFC 7432 EVPN All-Active Multi-Path procedures (aliasing, mass withdraw)
  - Enable overlay Equal Cost Multi-Path
  - Overlay flows load-balanced “equally” across a set of all-active multi-homing PEs
- RFC 7432 EVPN “per-service” DF election
  - Per-service DF role ”equally” distributed across a set of multi-homing PEs
Problem - Unicast ECMP

- PE1
- PE2
- CE1
- PE-x
- BGP-EVPN
- EVI-1, ESI-1
- EVI-2, ESI-1
- EVI-x, ESI-1
- Unicast flows to ESI-1
- 1/2 flows to ES-1

1/2 flows to ES-1
1/2 flows to ES-1
Problem - *Sub-optimal* Unicast ECMP – asymmetric access BW distribution

```
PE1
PE2
CE1

200Gb

100Gb

ESI-1

ECMP

EVI-1, ESI-1

EVI-2, ESI-1

......

EVI-x, ESI-1

ESI-1 -> PE1

ESI-1 -> PE2

1/2 flows to ES-1

BGP-EVPN

Unicast flows to ESI-1
```

1/2 flows to ES-1
Problem - BUM Flows – DF Service Carving

ESI-1 DF for ~1/2 EVIs

RT-4 based Service Carving

ESI-1 DF for ~1/2 EVIs

BUM Replication across all EVIs

PE1

PE2

PE-x

CE1

CE-x

100Gb

100Gb

100Gb

BGP-EVPN

L2 stretch
Problem - *Sub-optimal* BUM Flows – DF Service Carving – asymmetric access BW

ESI-1 DF for ~1/2 EVIs

RT-4 based Service Carving

ESI-1 DF for ~1/2 EVIs

BUM Replication across all EVIs

PE1

PE2

PE-x

CE1

CE-x

200Gb

100Gb

L2 stretch

BGP-EVPN