draft-malhotra-bess-evpn-unequal-lb-04

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

IETF 102, July 2018
Montreal

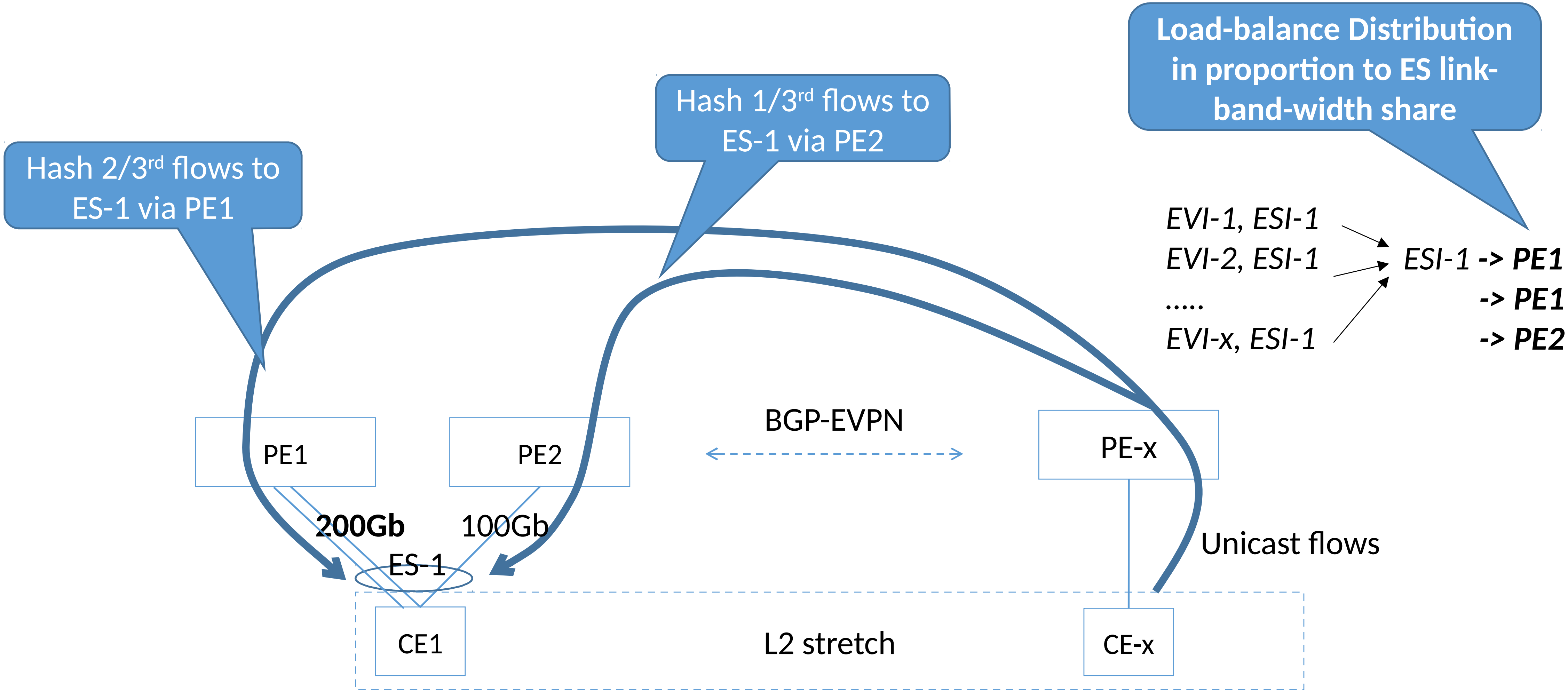
Draft Objective

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

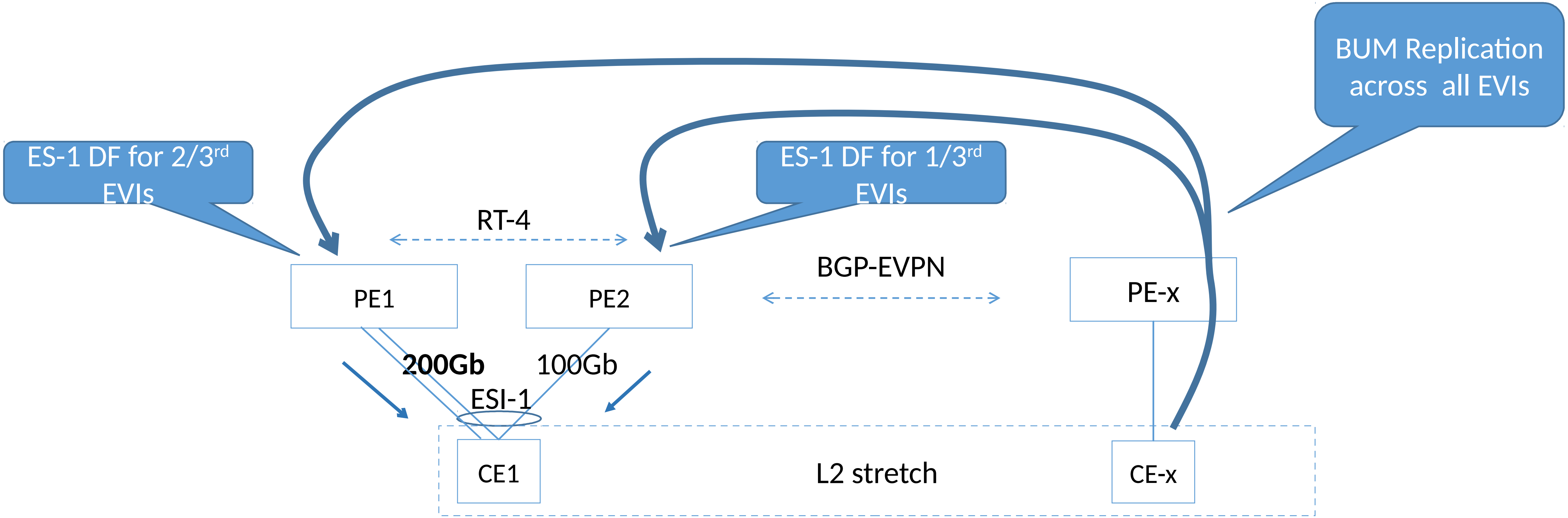
Hash 1/3rd flows to ES-1 via PE2

Load-balance Distribution in proportion to ES link-band-width share

EVI-1, ESI-1
EVI-2, ESI-1
.....
EVI-x, ESI-1

ESI-1 ***-> PE1***
-> PE1
-> PE2

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



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

Solution Summary

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

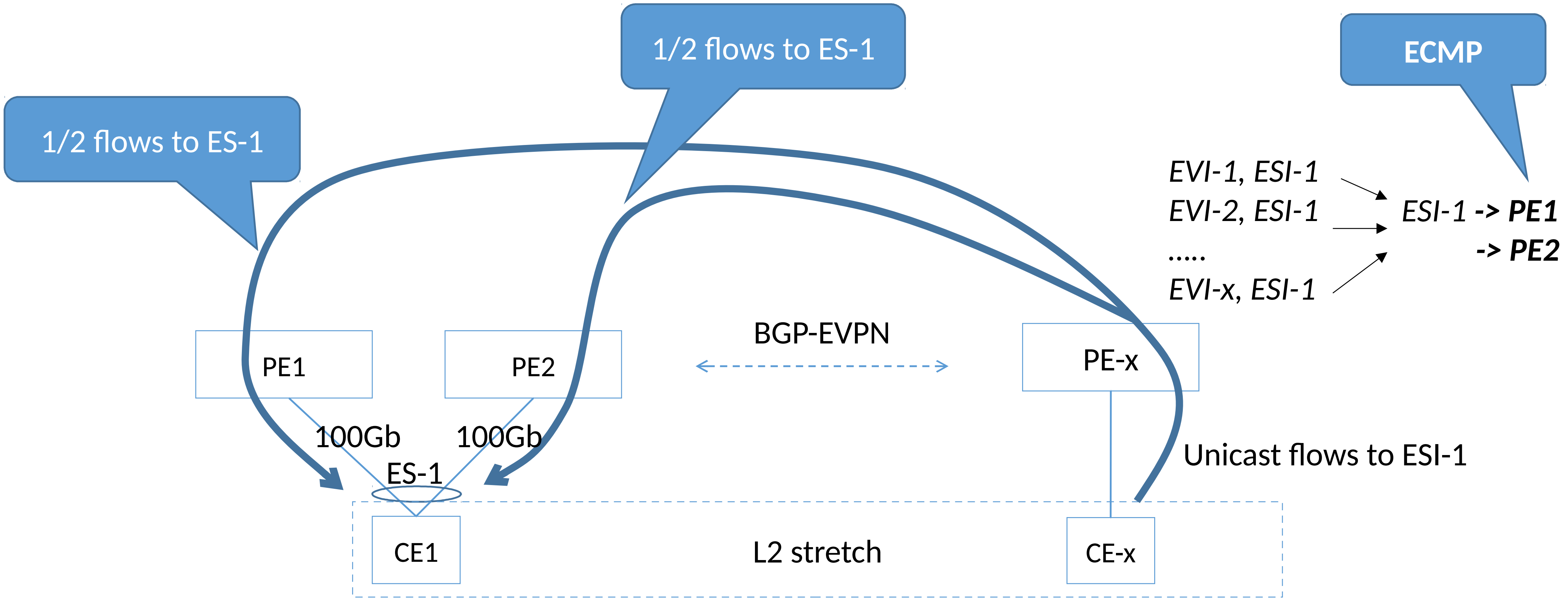
Neeraj Malhotra (Arrcus) , Ali Sajassi (Cisco)
Jorge Rabadan (Nokia), John Drake (Juniper)
Samir Thoria (Cisco), Avinash Lingala (AT&T)

BACKUP

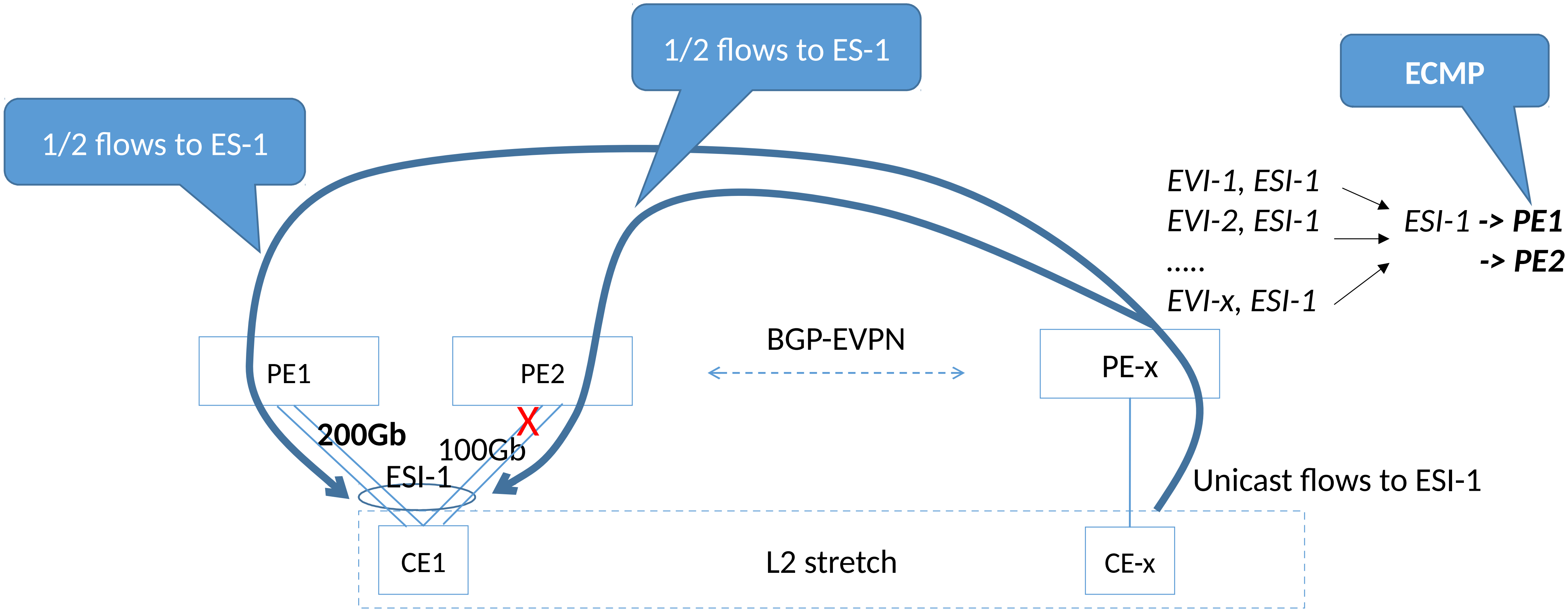
Prior Art

- 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

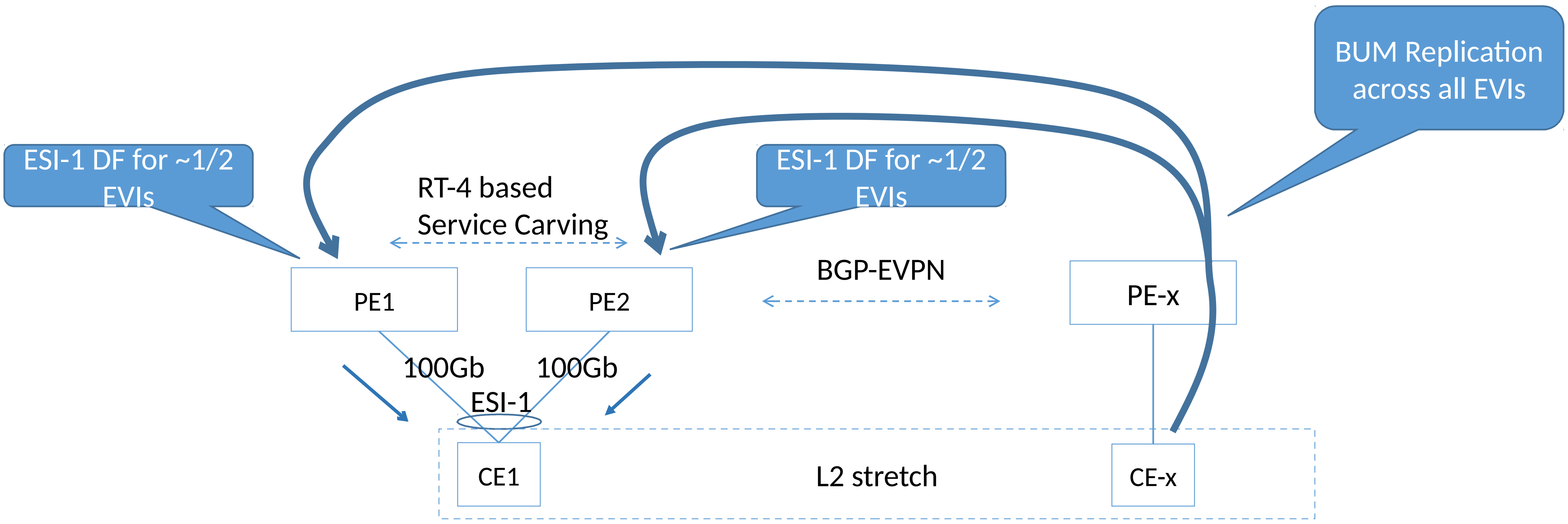
Problem - Unicast ECMP



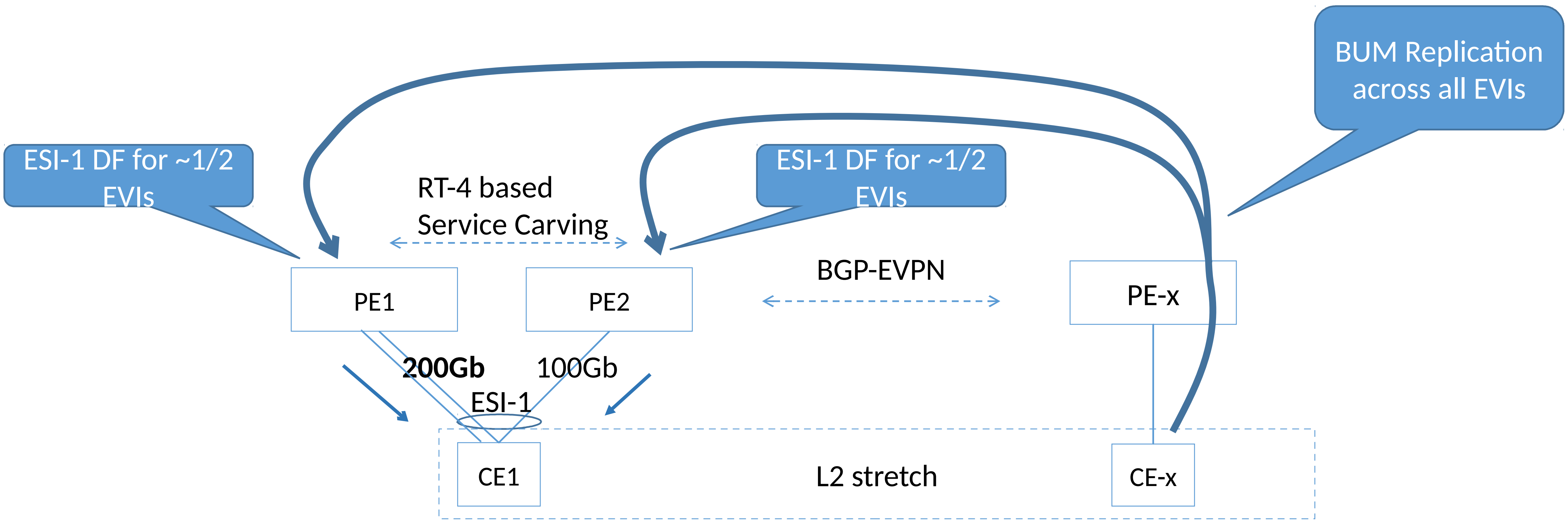
Problem - *Sub-optimal* Unicast ECMP – asymmetric access BW distribution



Problem - BUM Flows – DF Service Carving



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


BUM Replication across all EVIs
ESI-1 DF for ~1/2 EVIs
RT-4 based Service Carving
ESI-1 DF for ~1/2 EVIs
BGP-EVPN
PE1
PE2
PE-x
200Gb
100Gb
ESI-1
CE1
L2 stretch
CE-x