EVPN BUM Flooding Reduction

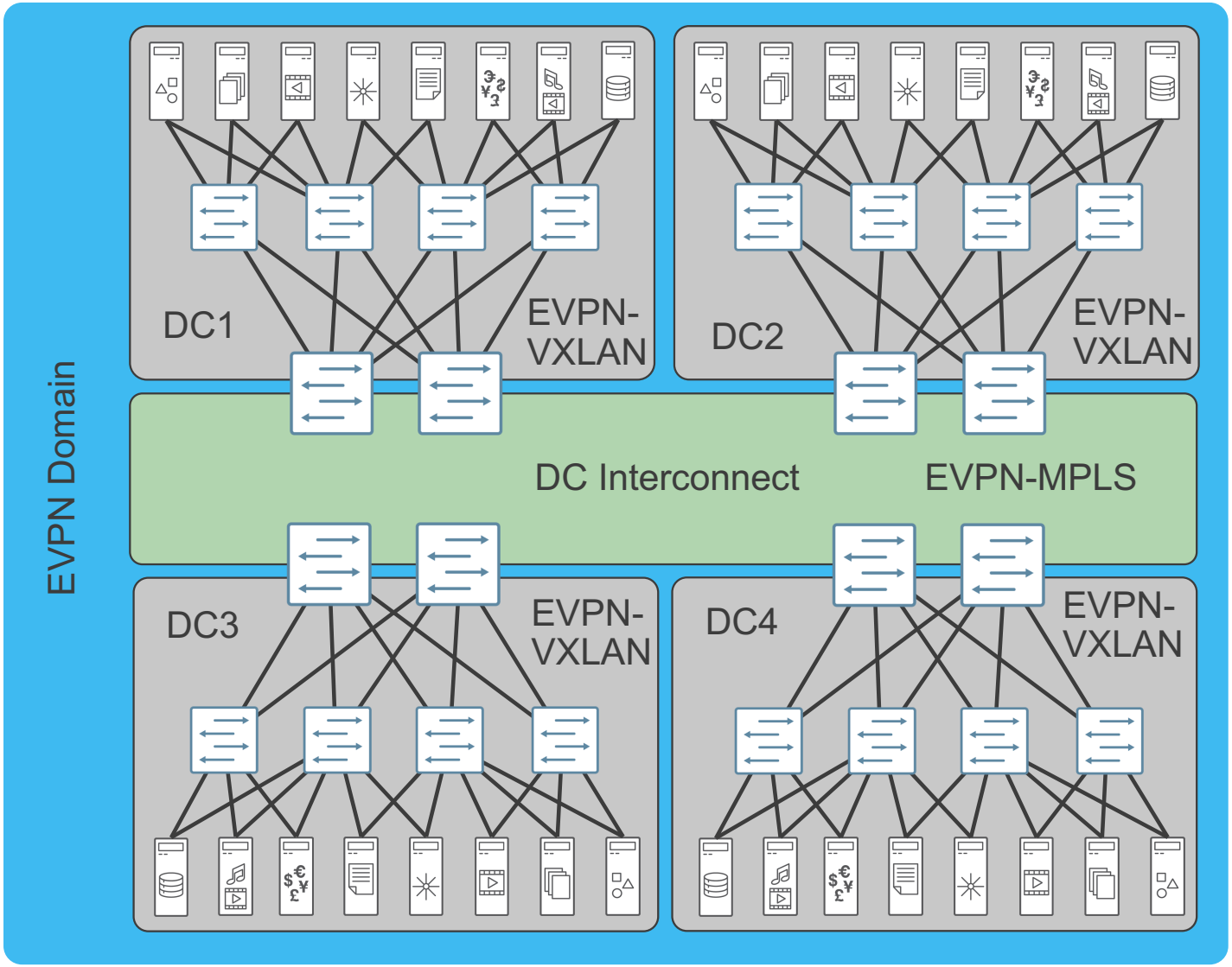
Krzysztof Grzegorz Szarkowicz, PLM
kszarkowicz@juniper.net

EVPN Introduction

- EVPN is getting traction in DC/Cloud deployments, replacing other (legacy) L2 architectures (i.e. VPLS)
- It has many benefits, like for example:
  - Unified, standardized control plane (BGP)
  - Unified, standardized A/A and A/S multi-homing
    - Multi-vendor interoperability
  - Near Hitless Host Mobility
  - Dramatic reduction of broadcast and multicast traffic
- This session covers the last bullet point in more details

EVPN in DC

- Mega DC
  - Many 100k hosts
- DCs are being interconnected
- It all results in large broadcast (flooding) domains



Session Agenda

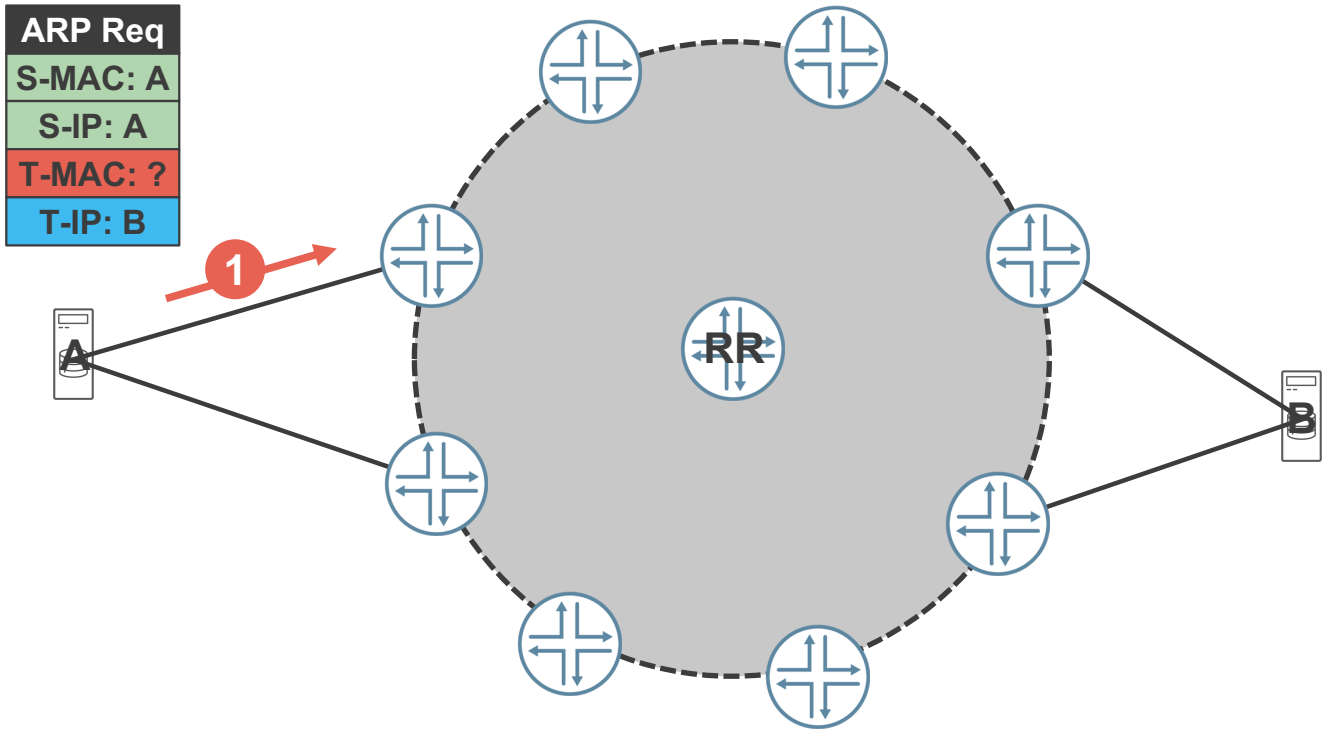
- ARP Flooding Reduction
- Multicast Flooding Reduction
- Efficient Replication of BUM Traffic
- Inter-Subnet Multicast

Broadcast Flooding

- Large broadcast flooding (e.g. ARP) might negatively impact DC operation
  - 600k hosts with 10 min ARP cache timeout → average 1k pps of ARP Requests
  - Routers connected to DC might need to process large number of ARPs
    - Typically, it happens in “slow path” (software processing)
    - Can cause heavy load on the router’s CPU
    - Typically limitation are low thousands per second
- Historically, some attempts have been made to address the problem:
  - RFC 6820: Address Resolution Problems in Large Data Center Networks
- EVPN brings holistic way to suppress ARP storms

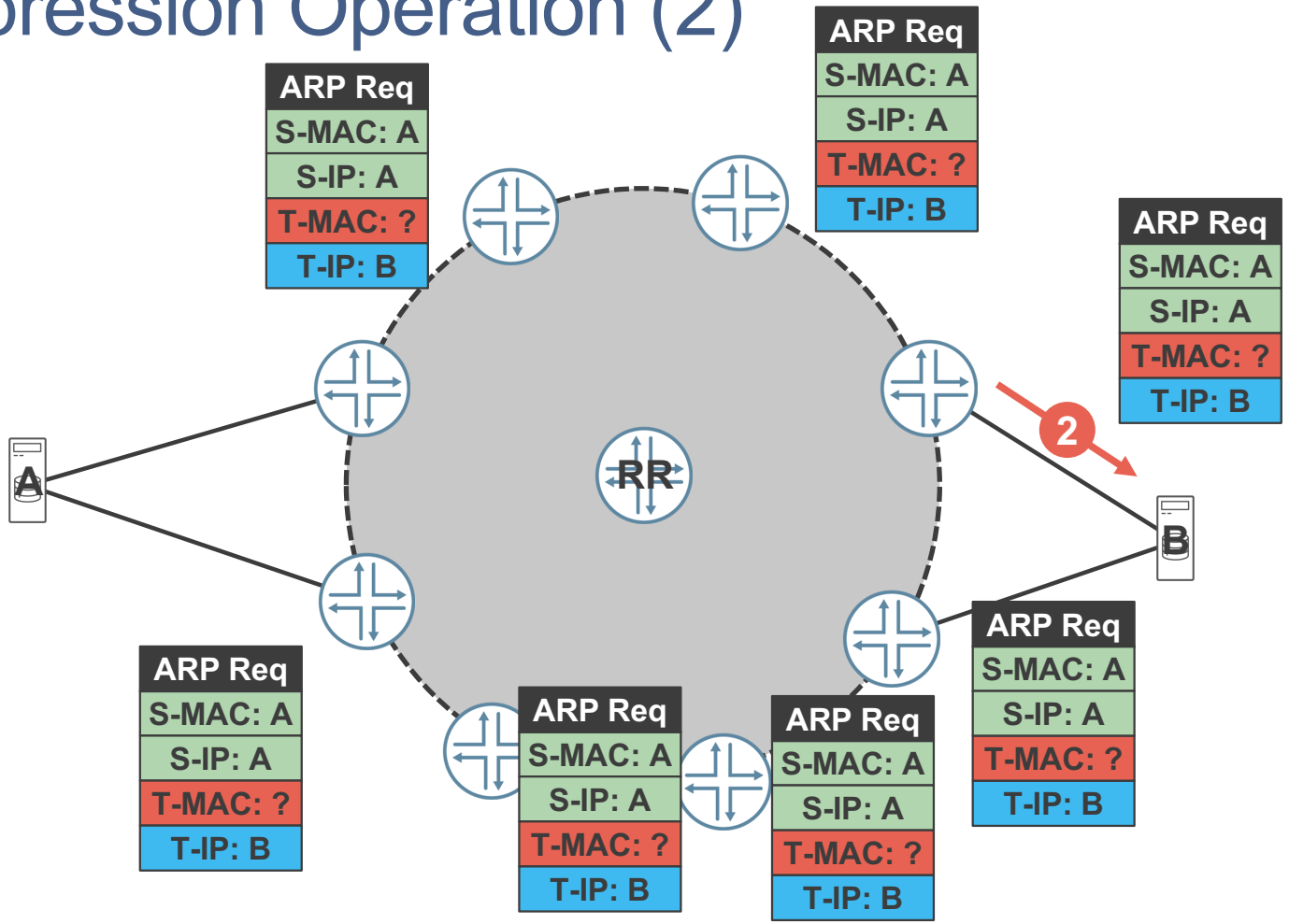
EVPN ARP Suppression Operation (1)

Host ’A’ issues ARP Request to resolve IP address ‘B’



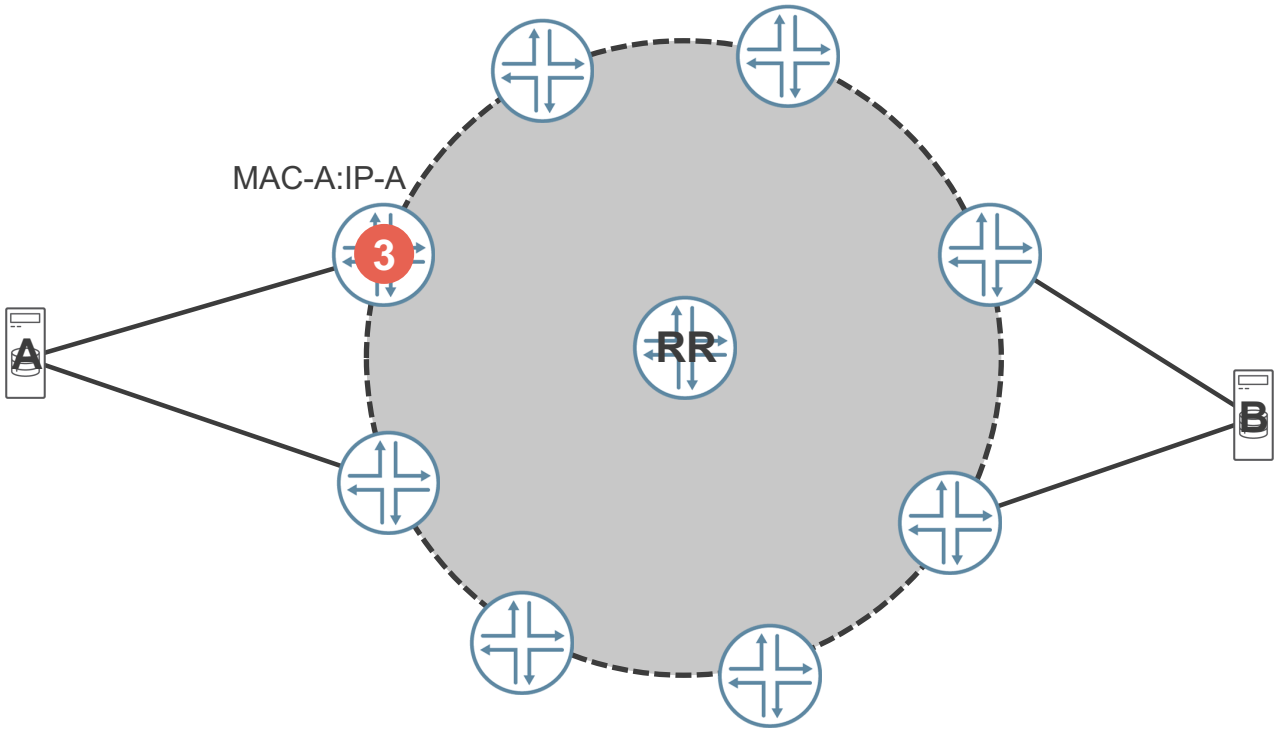
EVPN ARP Suppression Operation (2)

EVPN PE router, where ARP Request (with broadcast D-MAC) arrives, floods its via EVPN machinery, eventually arriving to host B



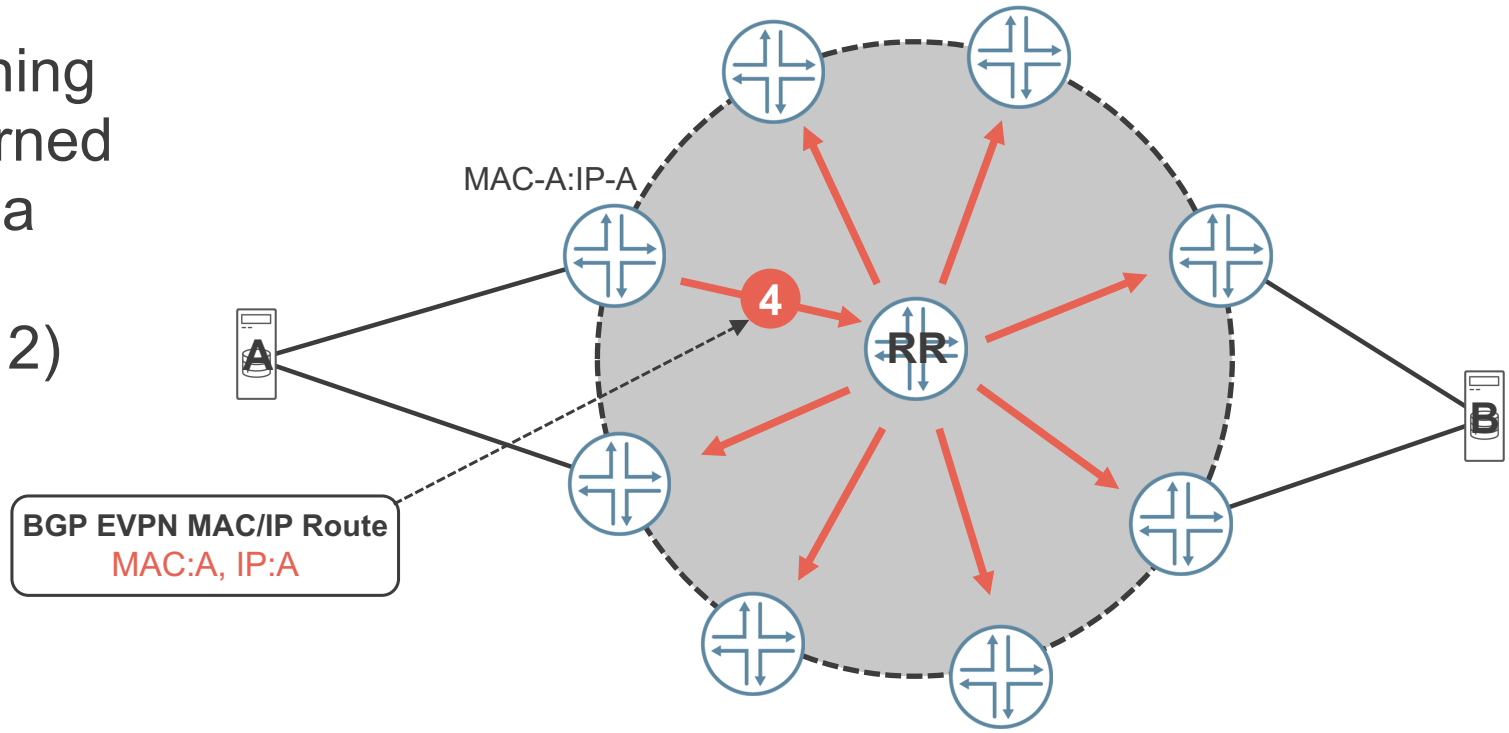
EVPN ARP Suppression Operation (3)

In the mean time, ingress EVPN PE intercepts ARP Request, learns MAC-A:IP-A association from it, and updates its EVPN database



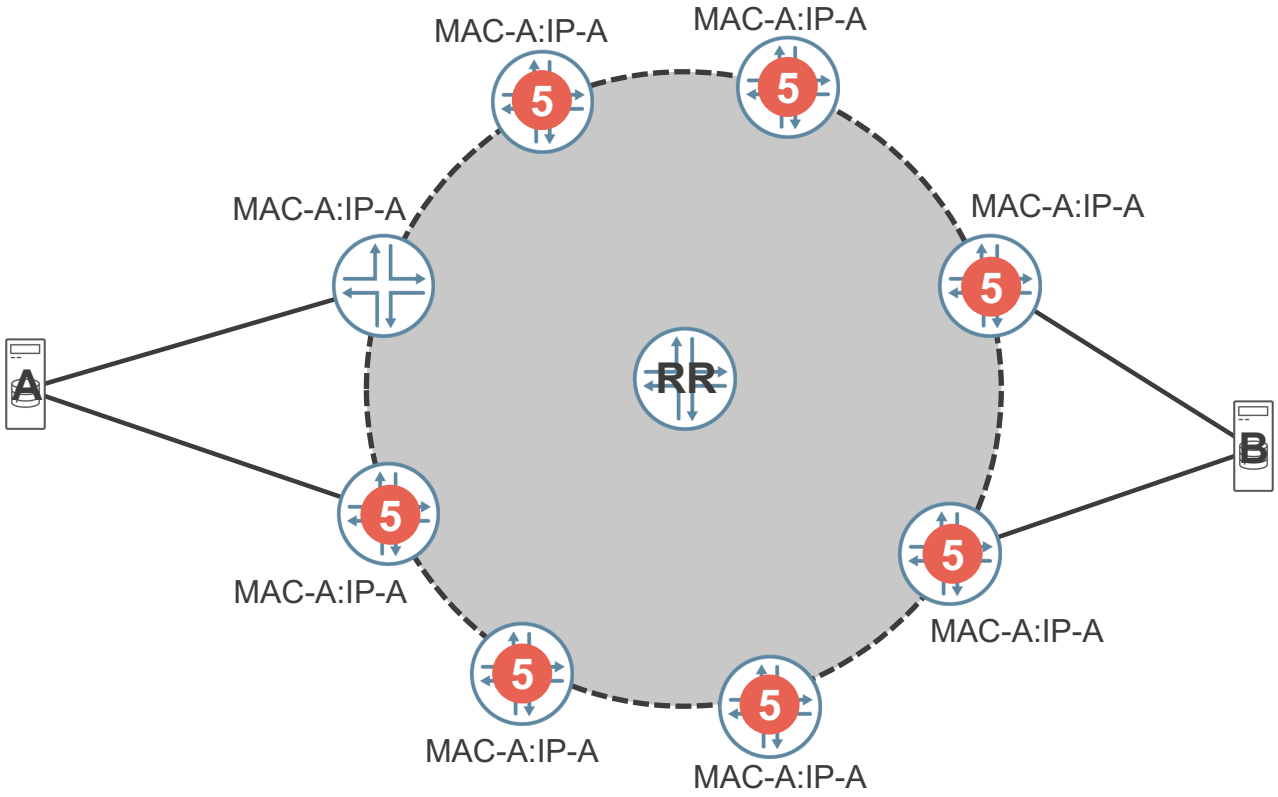
EVPN ARP Suppression Operation (4)

Ingress EVPN informs remaining PEs about learned MAC-A:IP-A via BGP EVPN MAC/IP (Type 2) Route



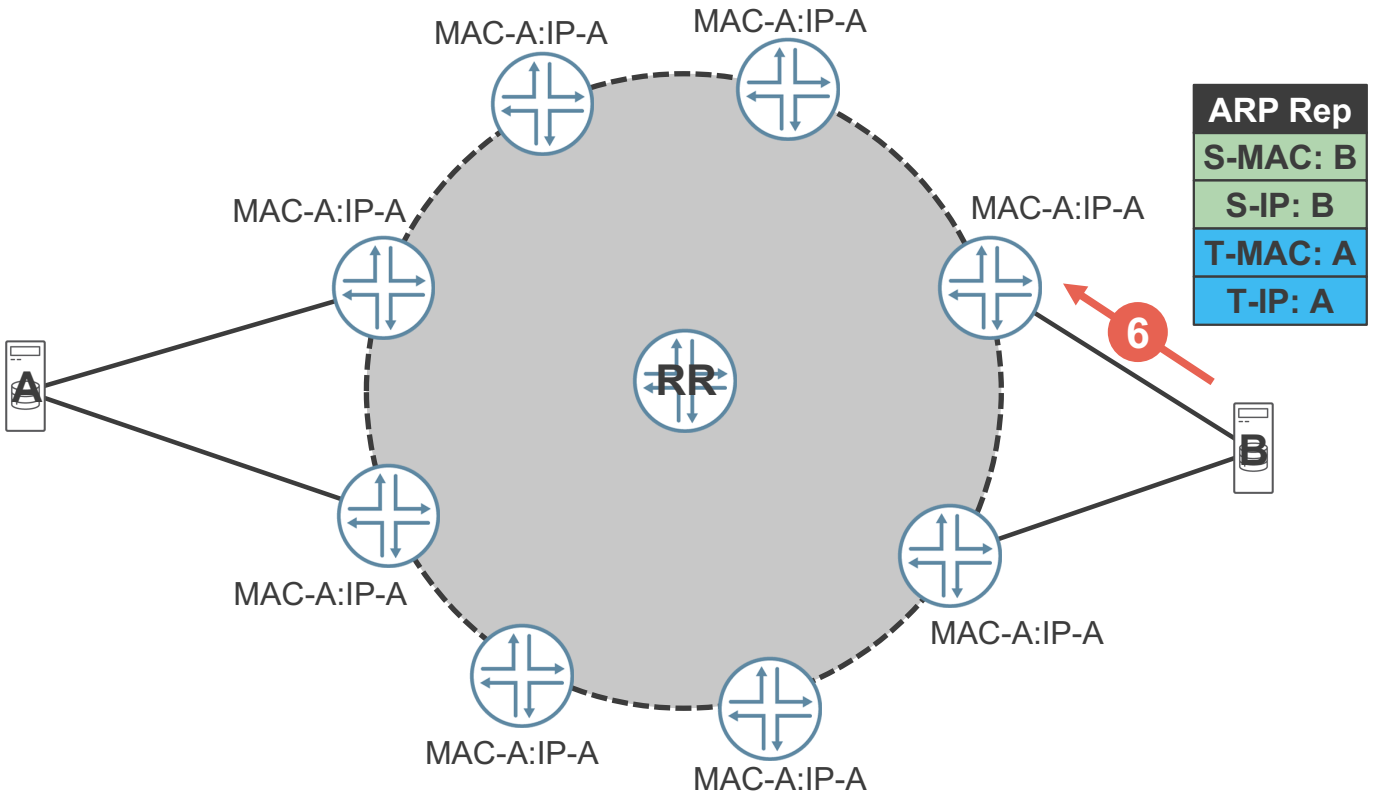
EVPN ARP Suppression Operation (5)

Remaining EVPN PEs update their EVPN database with MAC-A:IP-A association learned from ingress PE. Eventually, all PEs know about MAC-A:IP-A



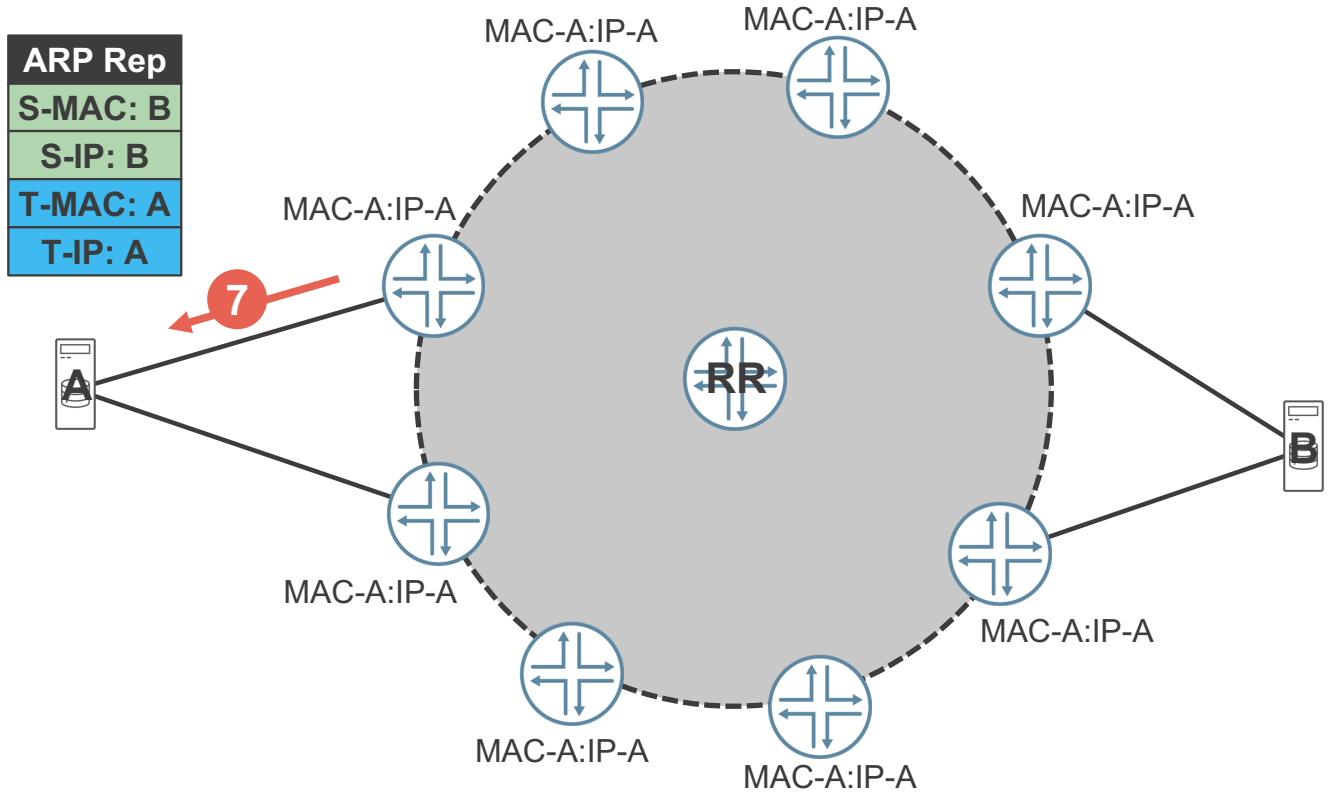
EVPN ARP Suppression Operation (6)

Host-B answers with ARP Reply



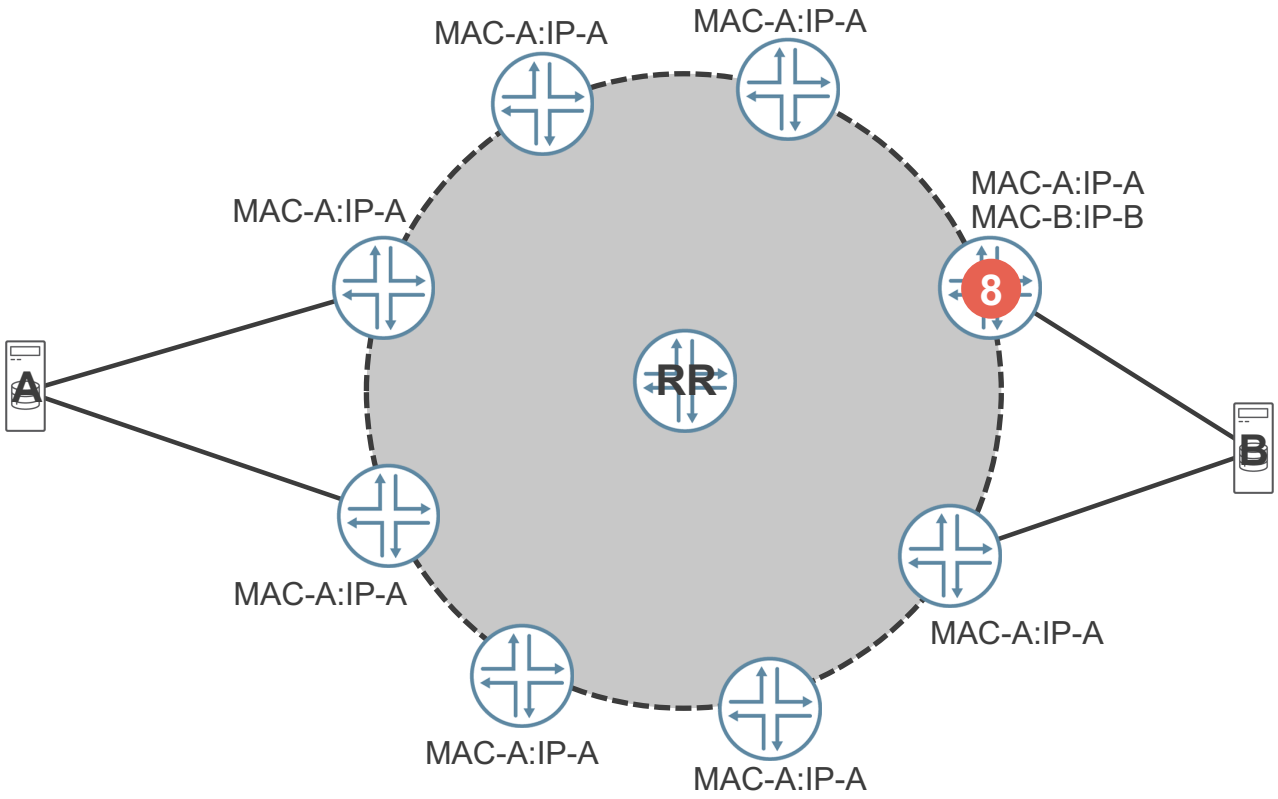
EVPN ARP Suppression Operation (7)

EVPN PE router, where ARP Reply arrives, has already MAC-A entry in its EVPN database, so ARP Reply is unicasted (not broadcasted) via EVPN machinery, and eventually arrives at Host-A



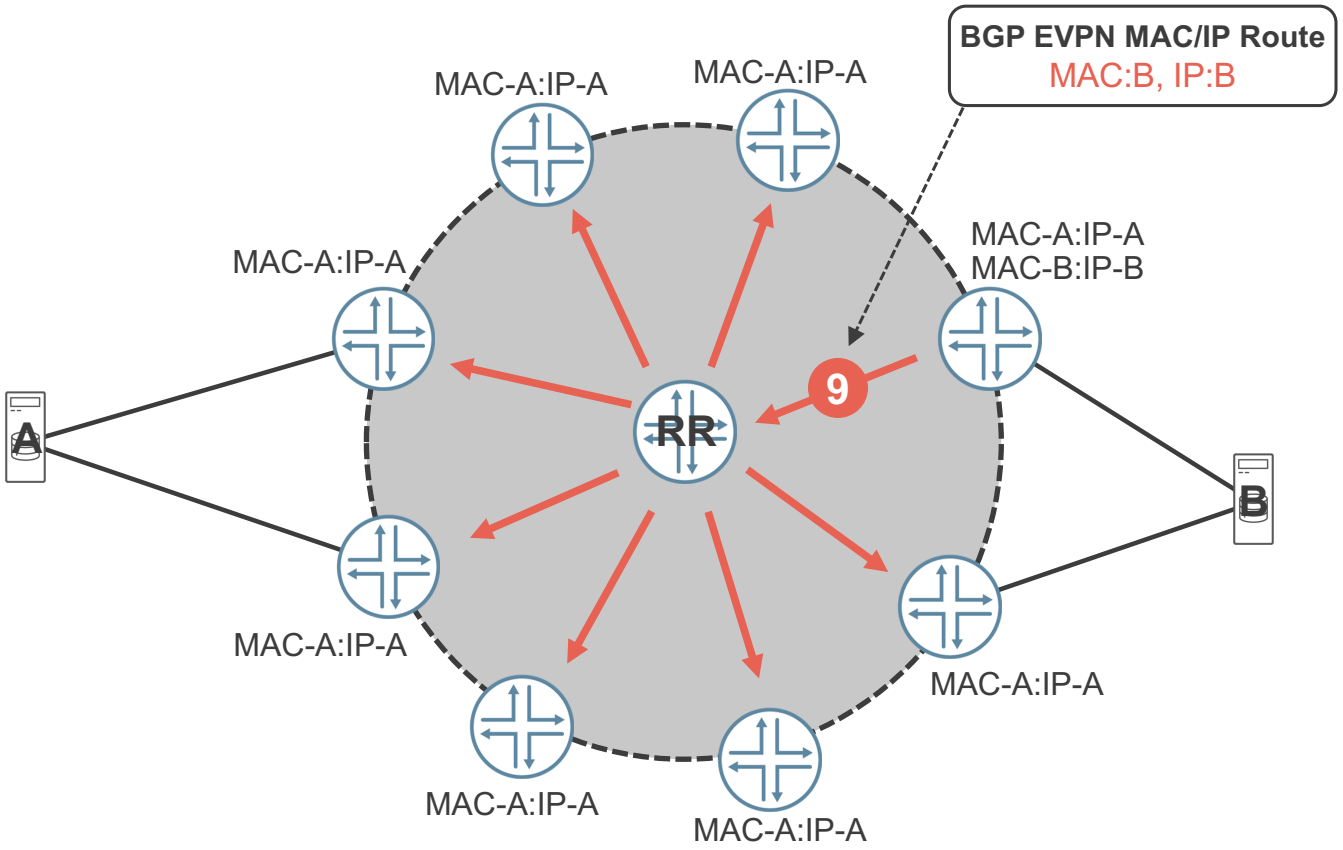
EVPN ARP Suppression Operation (8)

In the mean time, EVPN PE intercepts ARP Reply, learns MAC-B:IP-B association from it, and updates its EVPN database



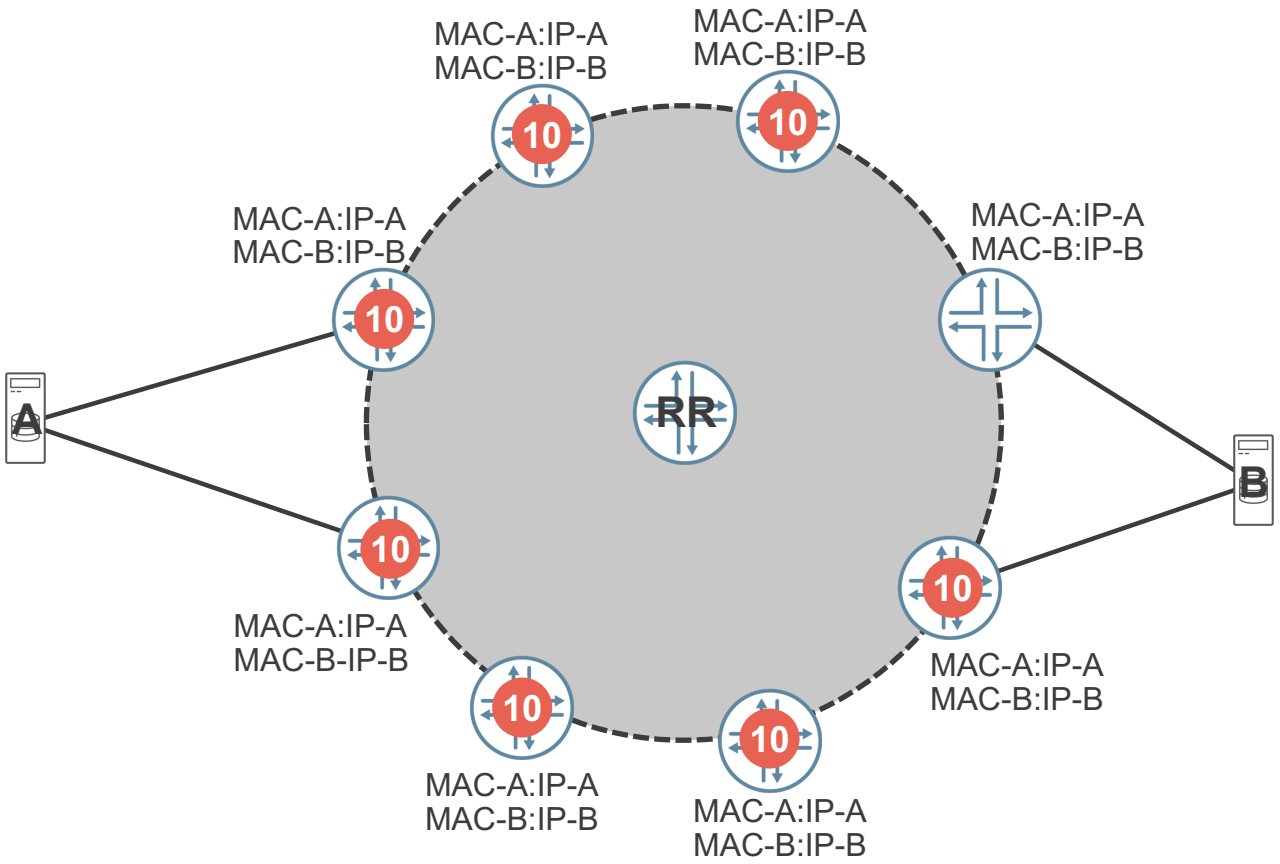
EVPN ARP Suppression Operation (9)

Ingress EVPN informs remaining PEs about learned MAC-B:IP-B via BGP EVPN MAC/IP (Type 2) Route



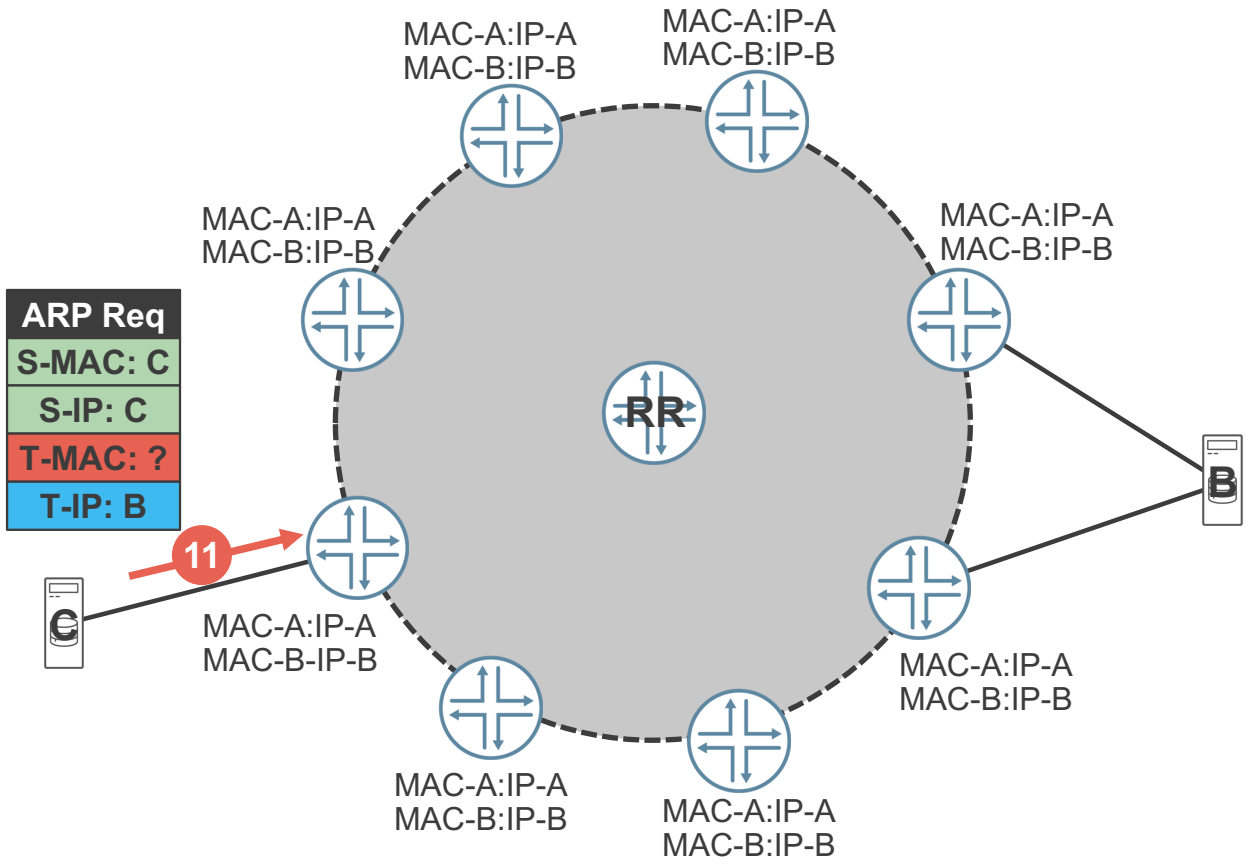
EVPN ARP Suppression Operation (10)

Remaining EVPN PEs update their EVPN database with MAC-B:IP-B association learned from ingress PE. Eventually, all PEs know about MAC-A:IP-A and MAC-B:IP-B



EVPN ARP Suppression Operation (11)

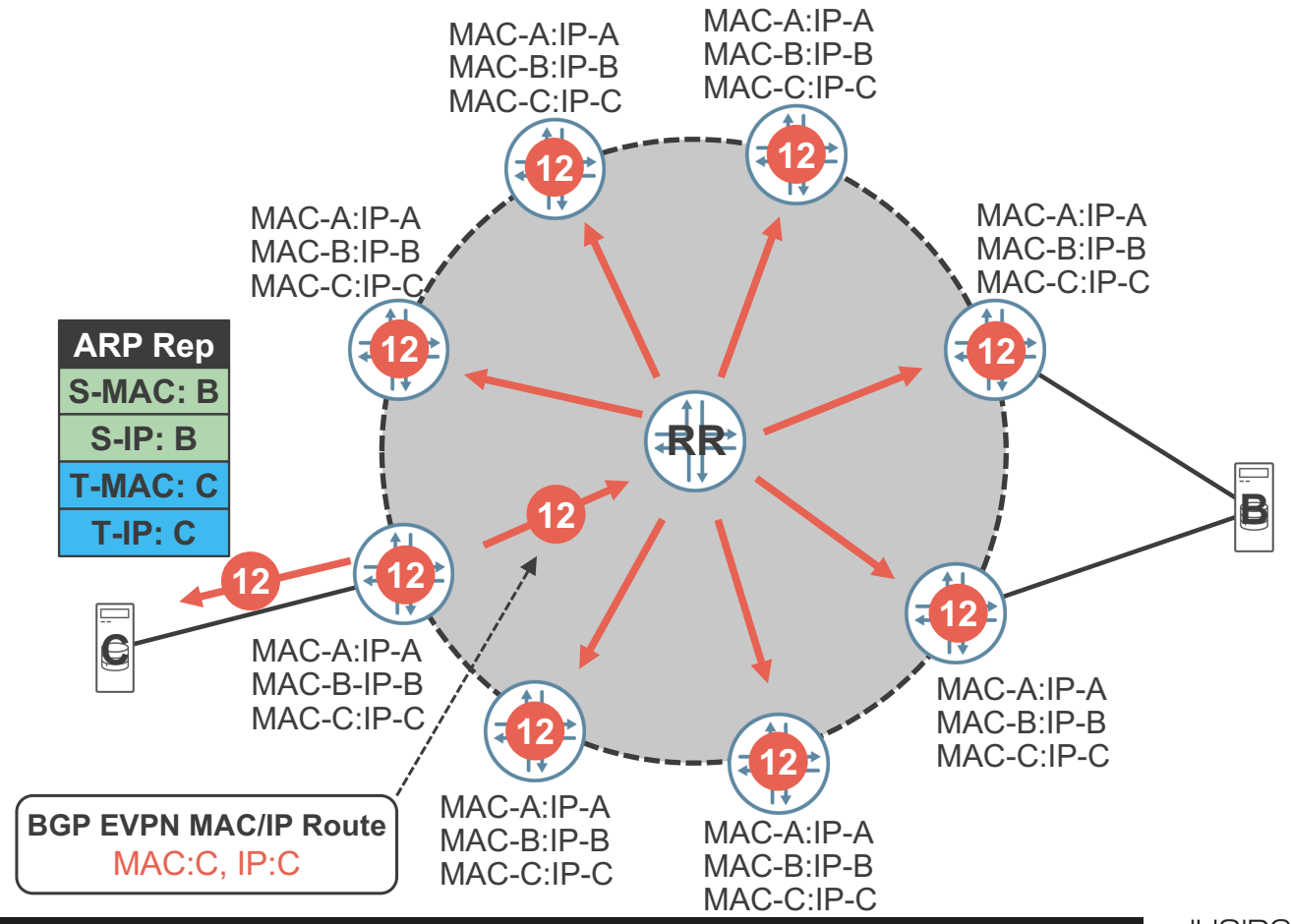
Host ’C’ issues ARP Request to resolve IP address ‘B’



EVPN ARP Suppression Operation (12)

EVPN PE already has an entry for MAC-B:IP-B, so it

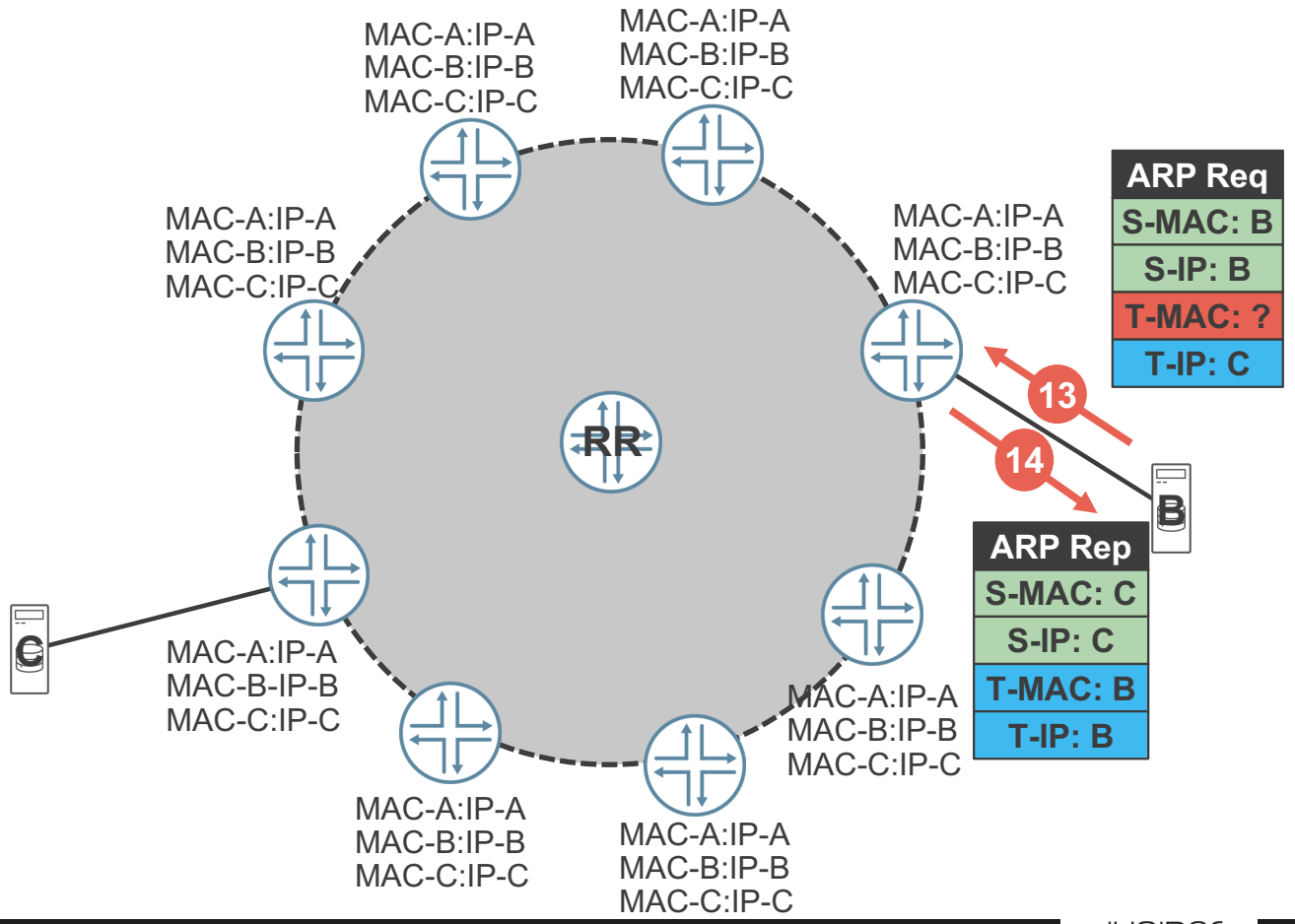
- sends ARP Reply to host C
- Learns MAC-C:IP-C
- Informs remaining PEs about MAC-C:IP-C



EVPN ARP Suppression Operation (13, 14)

When ARP cache on Host-B expires, Host-B issues ARP Request

- suppressed on PE
- PE sends immediate ARP Reply
- No update in EVPN BGP machinery required



EVPN ND Suppression

- ND suppression follows similar concepts to ARP suppression, hence not discussed explicitly in this session

Session Agenda

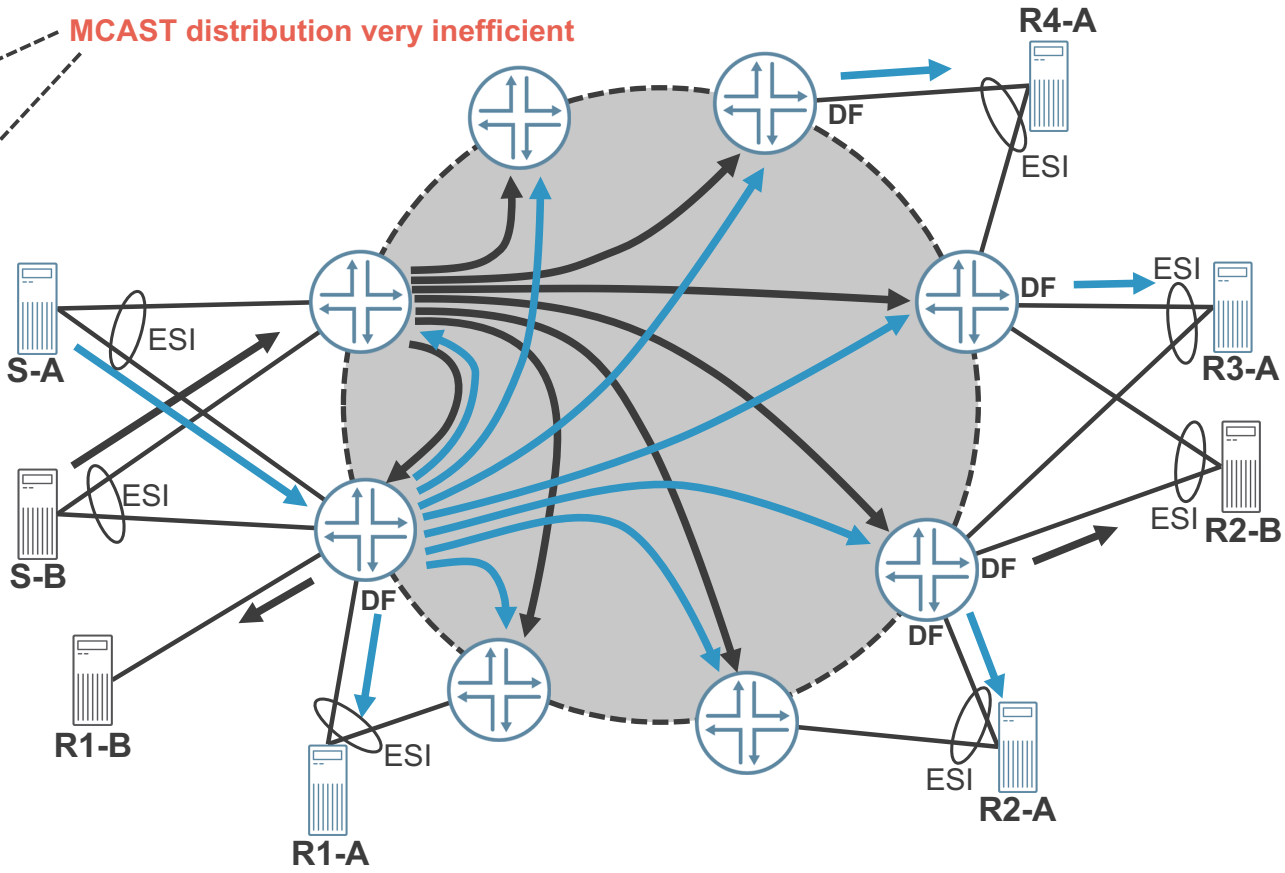
- ARP Flooding Reduction
- Multicast Flooding Reduction
- Efficient Replication of BUM Traffic
- Inter-Subnet Multicast

Basic EVPN Multicast Distribution (1)

Multicast is delivered from ingress PE to **all** egress PEs participating in given EVPN via **ingress replication**

Egress PE delivers/blocks MCAST to local receivers based on

- DF/non-DF state
- Local IGMP membership state

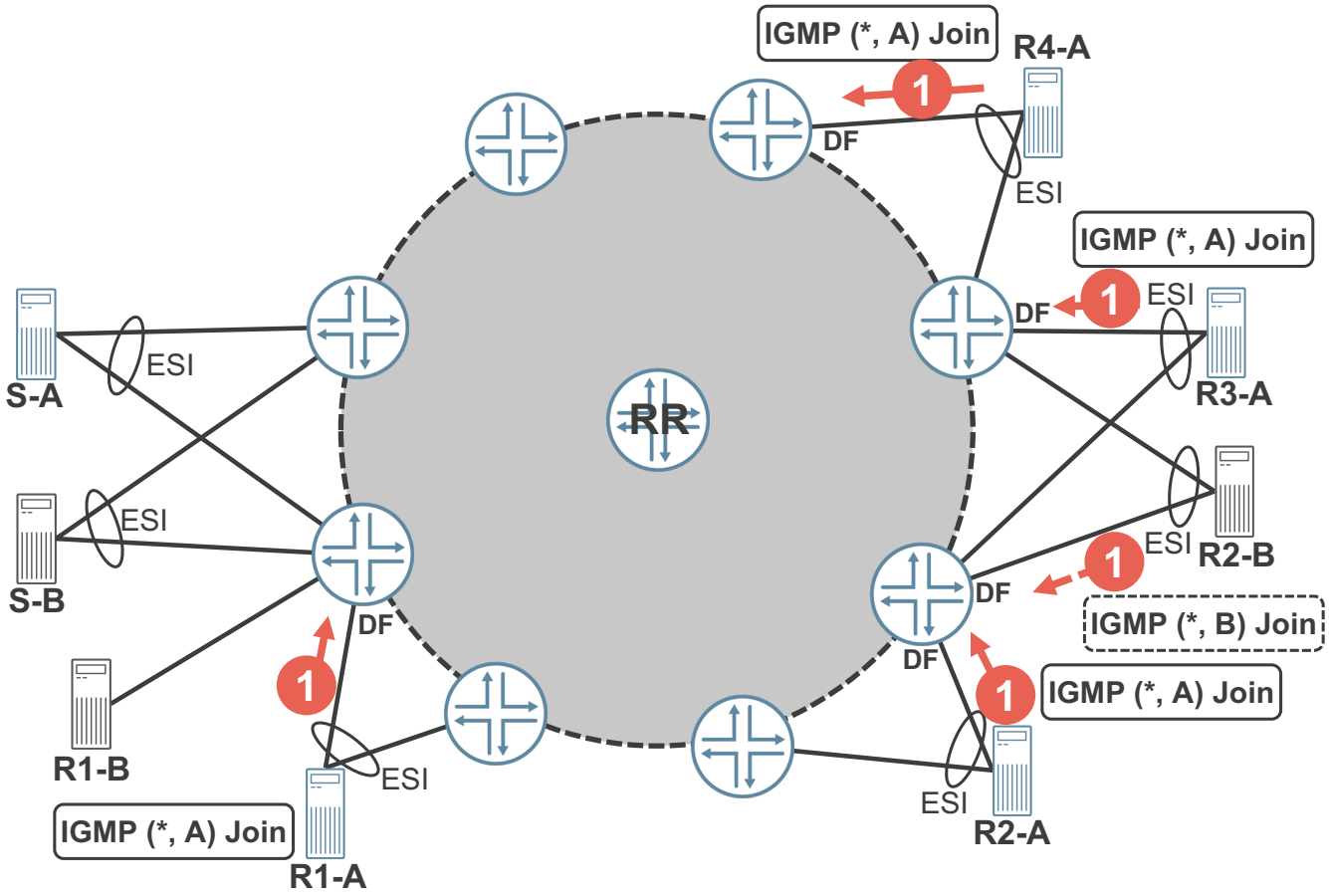


Basic EVPN Multicast Distribution (2)

- Two aspects of inefficient MCAST distribution in basic EVPN deployments
  - MCAST distributed to all PEs
    - EVPN creates states basic on
      - Data plane or PE-CE control plane (for traffic received from CE)
        - » IGMP
      - PE-PE BGP EVPN control plane (for traffic received via EVPN core)
        - » BGP EVPN extensions required to accomplish that → SMET (Type 6) Route
  - Ingress replication
    - More efficient replication methods required
      - P2MP (i.e. PIM, mLDP, RSVP, BIER)
      - Assisted Replication

Selective Multicast Ethernet Tag (SMET) Route (1)

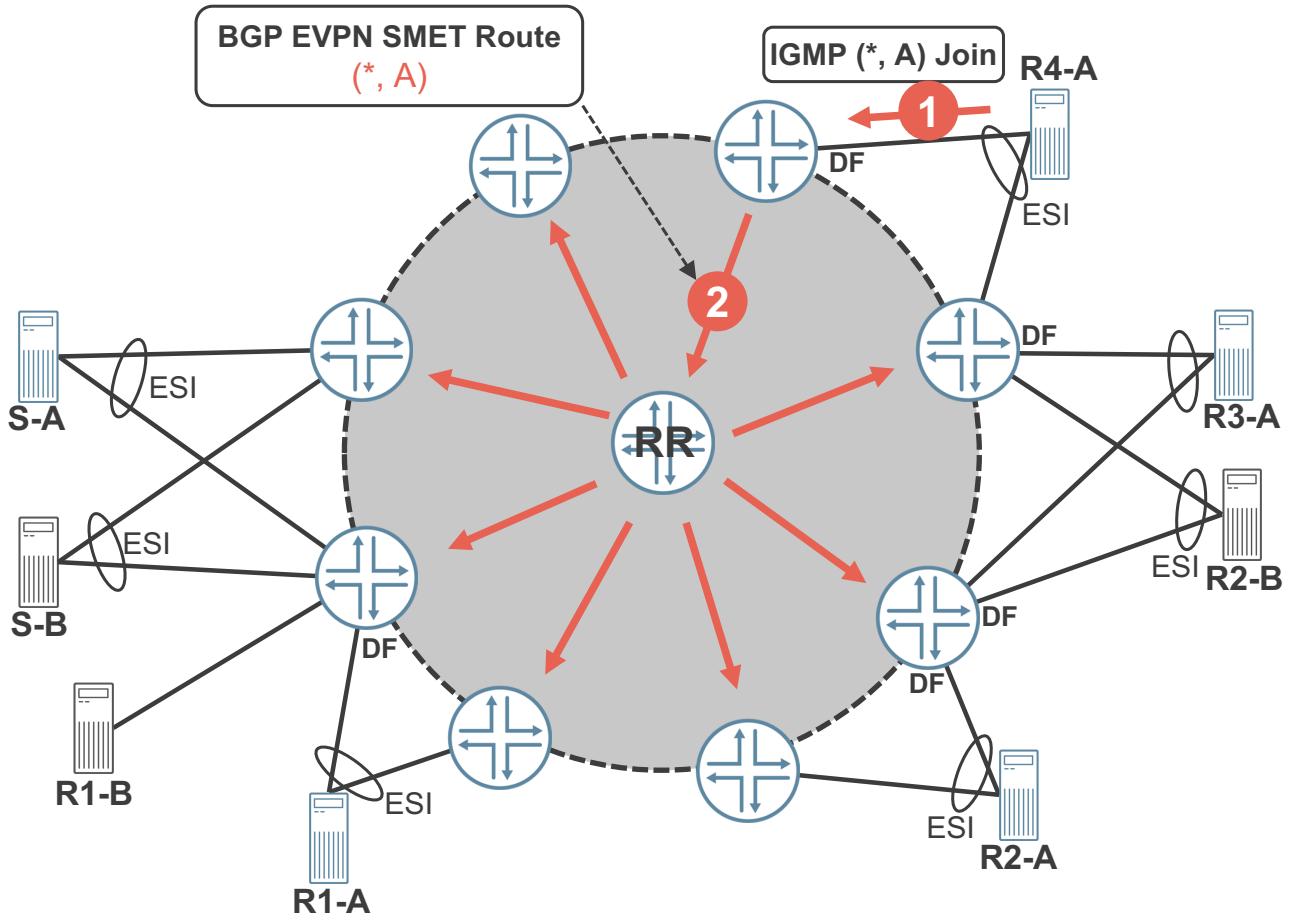
Receives reports the willingness to receive MCAST traffic via standard IGMP (v1/v2/v3) Group Membership (“*Join*”) messages



Selective Multicast Ethernet Tag (SMET) Route (2)

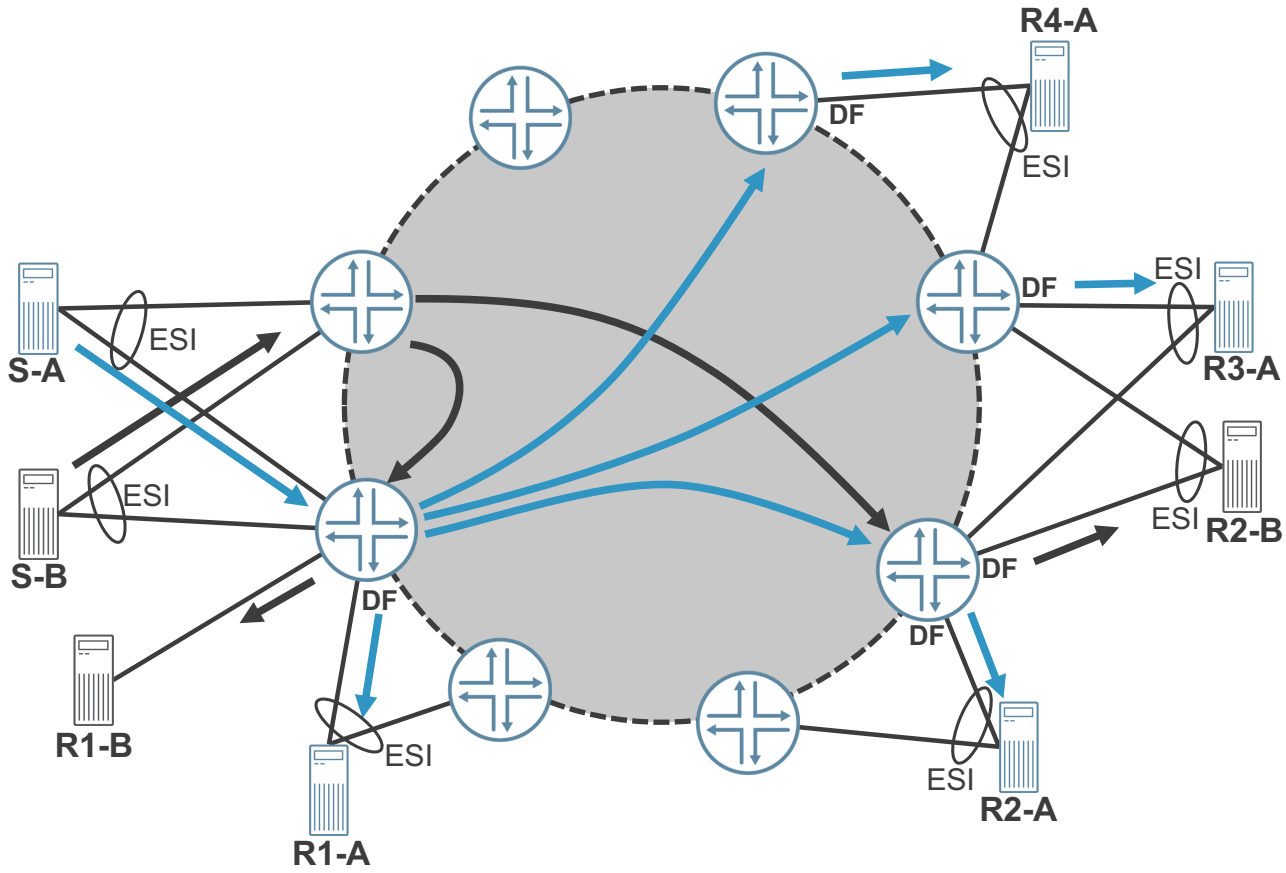
First hop PEs convert IGMP Group Membership messages to BGP EVPN Selective Multicast Ethernet Tag (SMET) messages (Type 6)

- Only R4-A shown, as an example
- Based on that information, all involved PEs are aware, where multicast receivers for specific MCAST flows reside



Selective Multicast Ethernet Tag (SMET) Route (3)

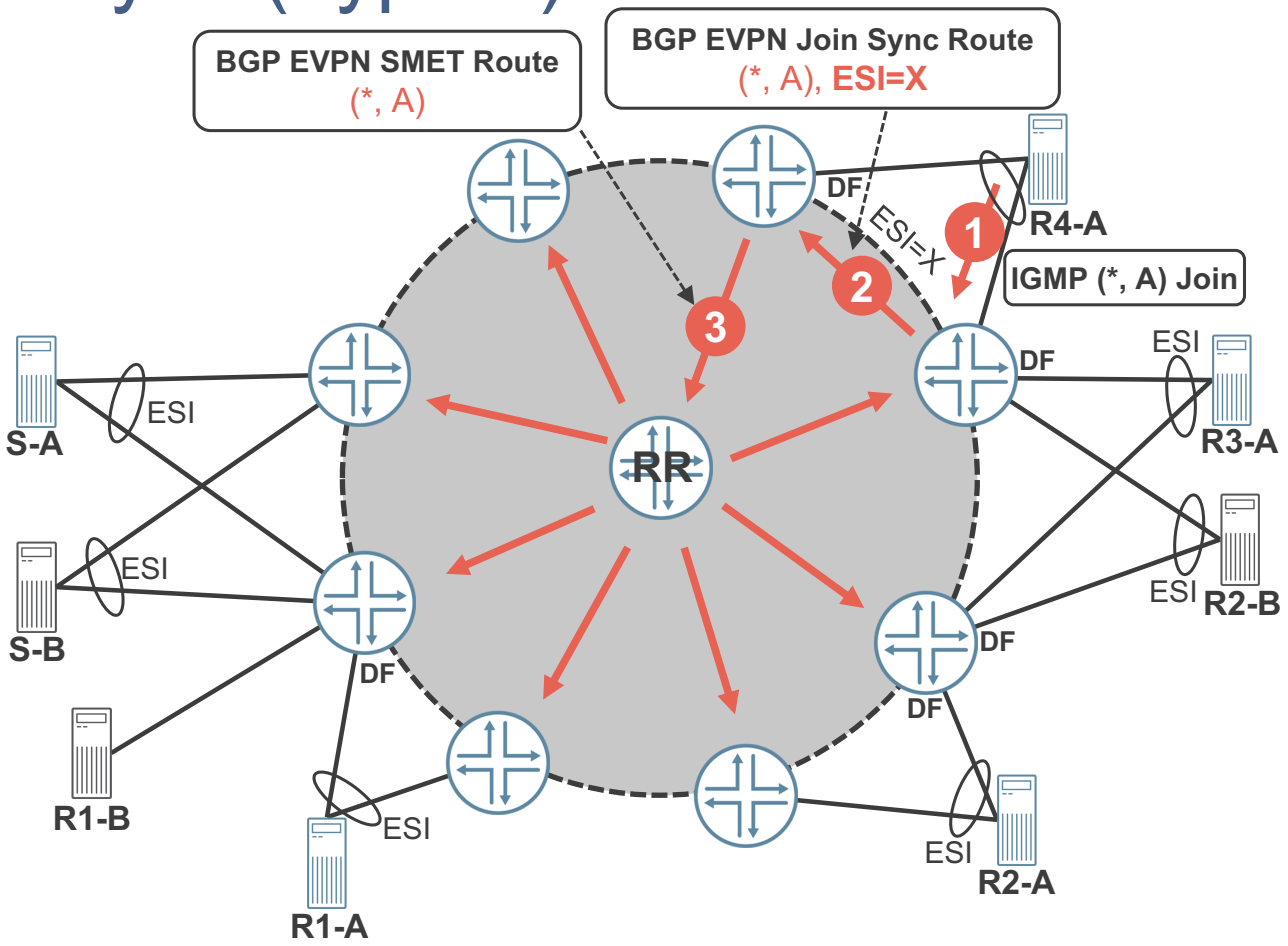
Based on BGP EVPN SMET (Type 6) Route, PEs with attached sources can send MCAST flows to specific PEs only



BGP EVPN Join Sync (Type 7) Route
BGP EVPN Leave Sync (Type 8) Route

In EVPN A/A multi-homing

1) IGMP Join/Leave might arrive to non-DF
2) It is converted to EVPN Join/Leave Sync (Type 7/8) Route
3) SMET (Type 6) Route announced by DF only based on local IGMP Join or EVPN Join



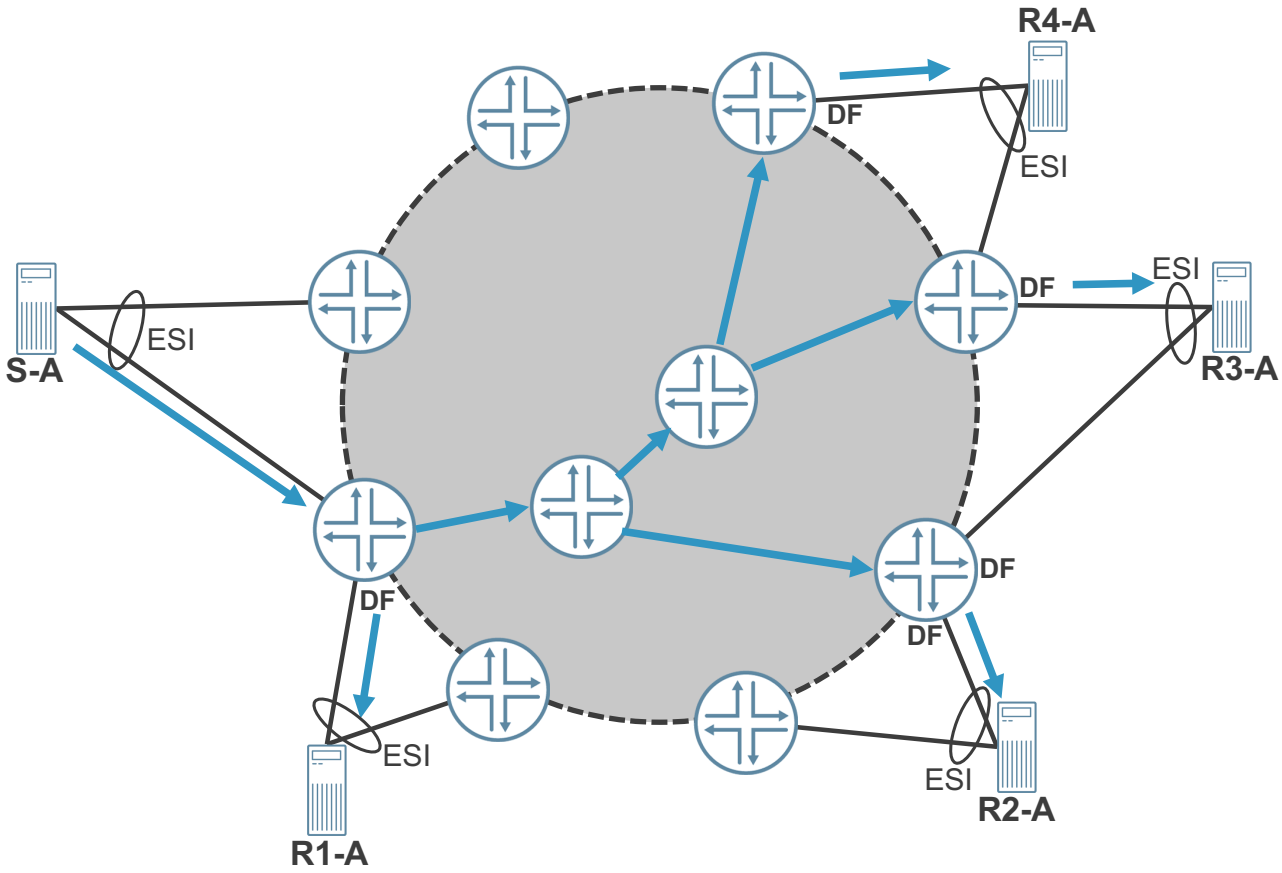
Session Agenda

- ARP Flooding Reduction
- Multicast Flooding Reduction
- Efficient Replication of BUM Traffic
- Inter-Subnet Multicast

EVPN P2MP Multicast Distribution

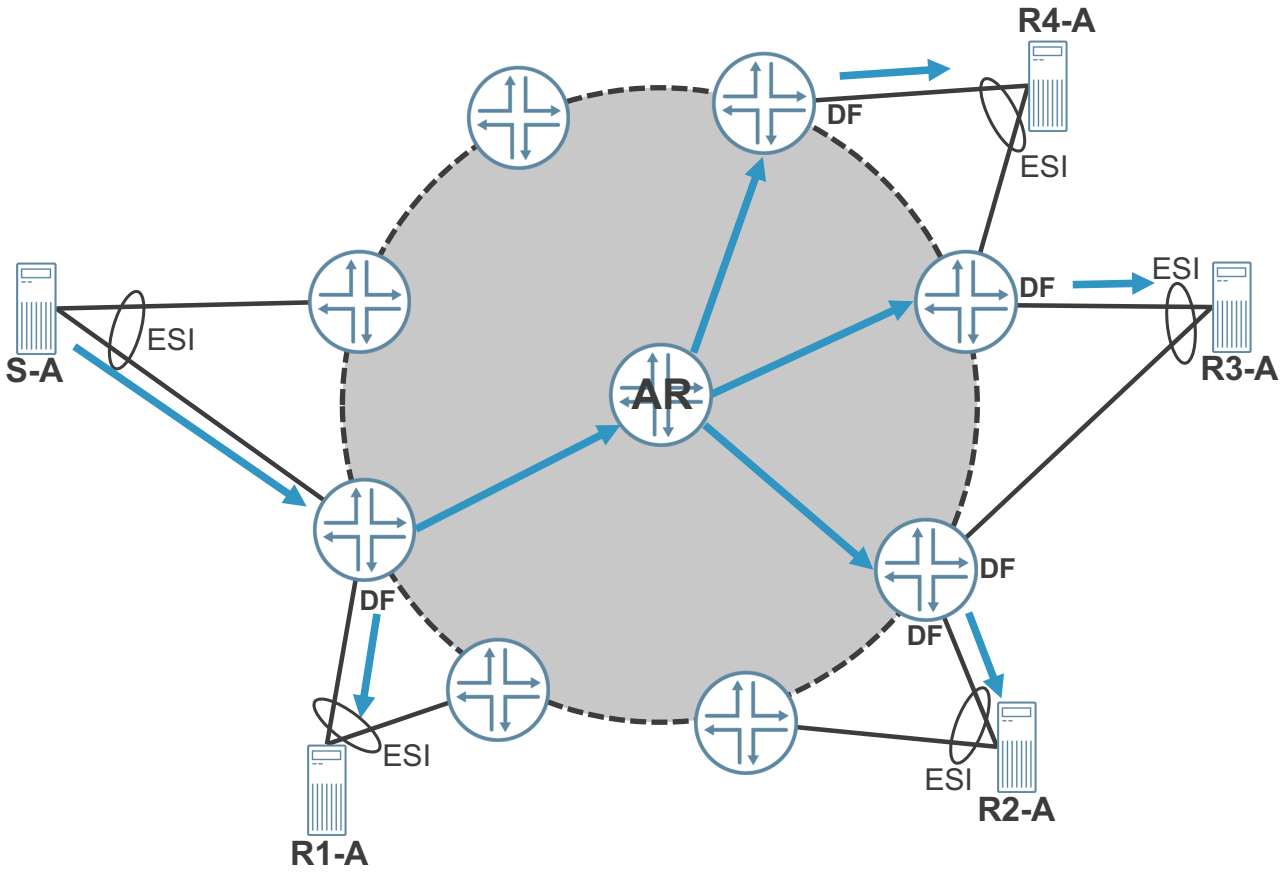
BUM frames are replicated on transit nodes, according to the P2MP structure

- Universally deployable in any arbitrary topology
- Requires consistent P2MP support on all nodes
- Information about P2MP tunnel distributed via Provider Multicast Service Interface (PMSI) attribute in the Inclusive Multicast Ethernet Tag (Type 3) EVPN Route



EVPN Assisted Replication

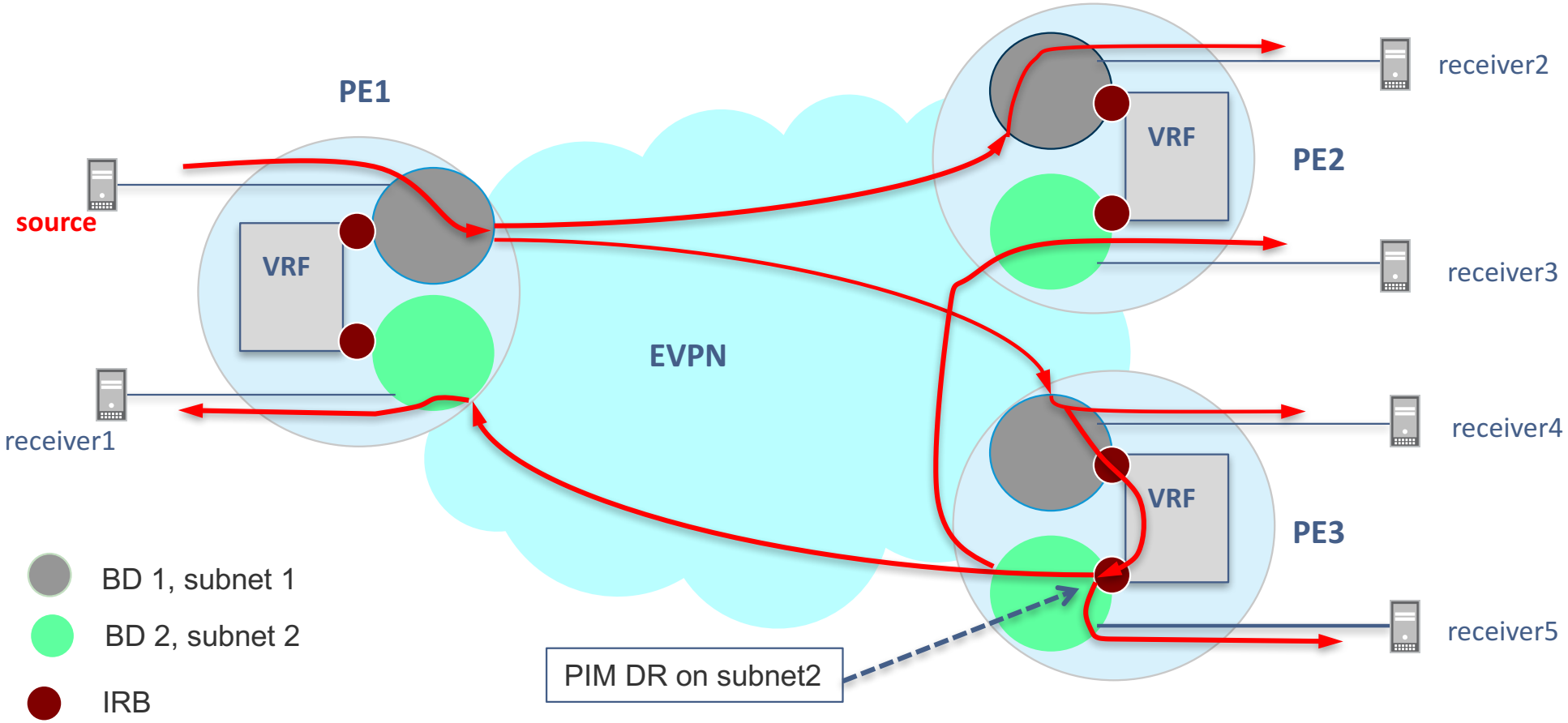
- Referred often as “Optimized Ingress Replication”
- Selected (powerful) nodes are designated to perform replication
- Typically suitable to NVO/DC (Leaf/Spine) designs, with powerful Spines, and low performance Leafs



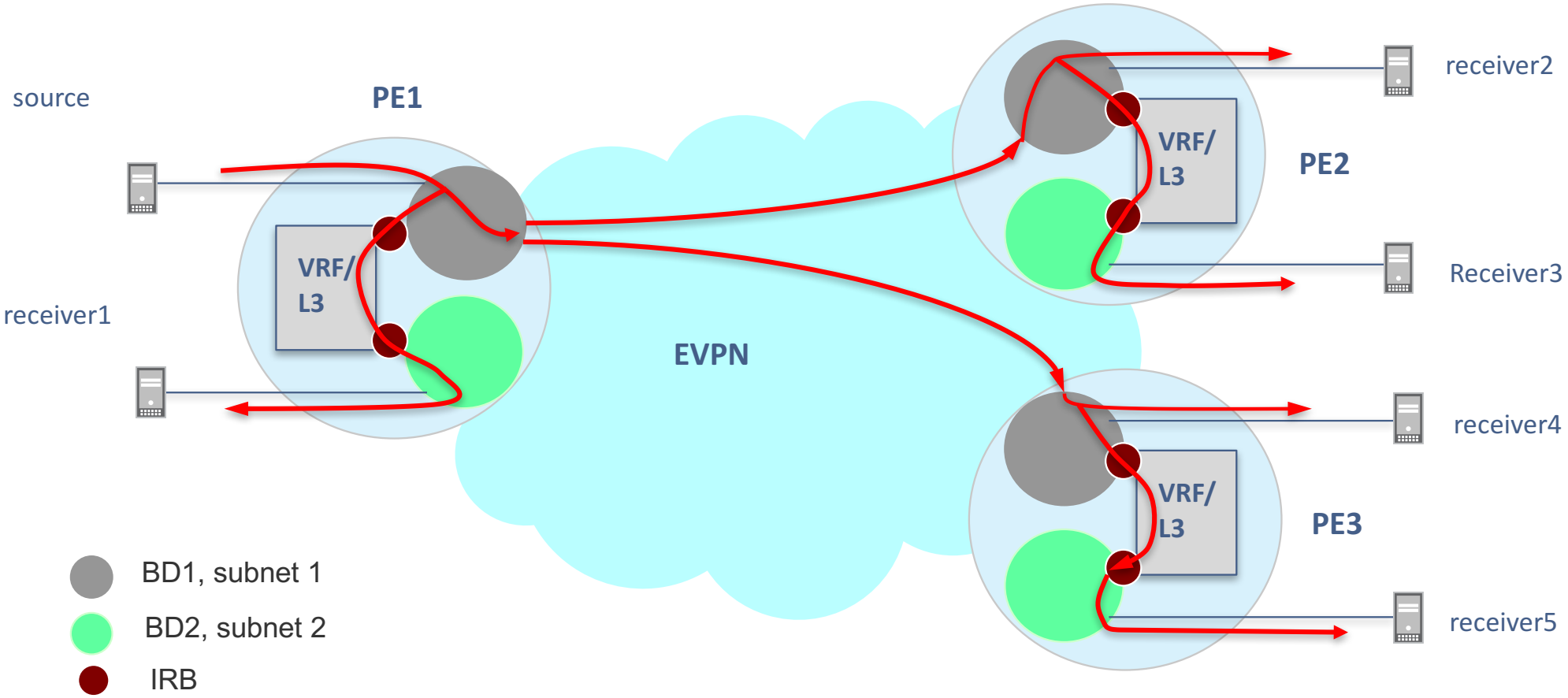
Session Agenda

- ARP Flooding Reduction
- Multicast Flooding Reduction
- Efficient Replication of BUM Traffic
- Inter-Subnet Multicast

EVPN Legacy Inter-Subnet Multicast

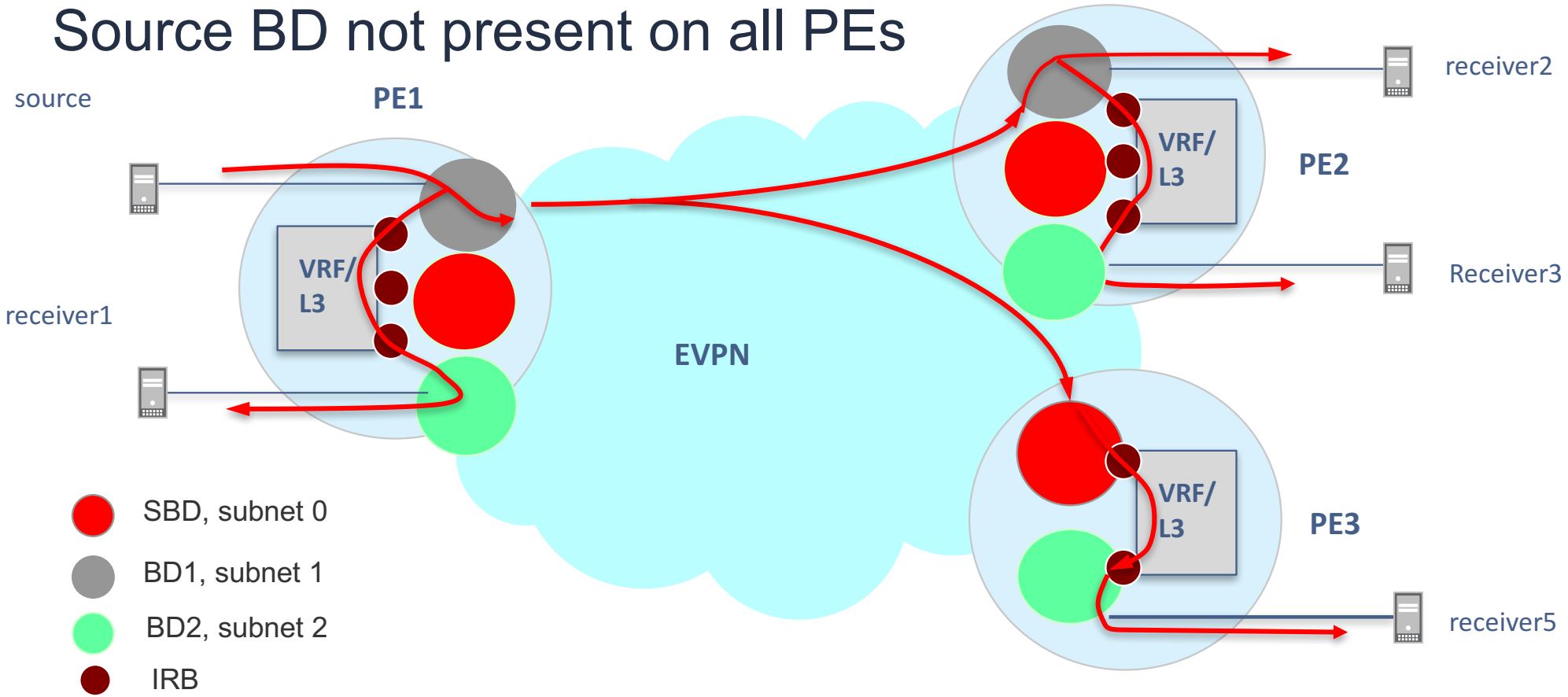


EVPN Optimized Inter-Subnet Multicast



EVPN Optimized Inter-Subnet Multicast

Source BD not present on all PEs



Summary – Standardization Status

Feature	Specification	EVPN Route Types Involved
ARP/ND Flooding Reduction (APR/ND Snooping/Proxy)	RFC 7432, Section 10	Type 2: MAC/IP Advertisement Route
Multicast Flooding Reduction (IGMP/MLD Snooping/Proxy)	draft-ietf-bess-evpn-igmp-mld-proxy-01	Type 6: Selective Multicast Ethernet Tag Route Type 7: IGMP Join Synch Route Type 8: IGMP Leave Synch Route
P2MP BUM Trees	RFC 7432, Section 16.2 → RFC 7117	Type 3: Inclusive Multicast Ethernet Tag Route
Assisted Replication	draft-ietf-bess-evpn-optimized-ir-03	Type 3: Inclusive Multicast Ethernet Tag Route Type 11: Leaf Auto-Discovery (AD) route
Optimized Inter-Subnet Multicast	draft-ietf-bess-evpn-irb-mcast-00	Type 3: Inclusive Multicast Ethernet Tag Route Type 6: Selective Multicast Ethernet Tag Route Type 10: S-PMSI Auto-Discovery (AD) route
Multicast Flooding Reduction (PIM Snooping/Proxy)	draft-skr-bess-evpn-pim-proxy-01	Type 6: Selective Multicast Ethernet Tag Route Type 7: IGMP/PIM Join Synch Route Type <tbd>: Multicast Router Discovery (MRD) Route Type <tbd>: PIM RPT-Prune Route Type <tbd>: PIM RPT-Prune Join Synch Route
DHCP Flooding Reduction (DHCP Snooping/Proxy)	draft-surajk-evpn-access-security-00	Type <tbd>: DHCP Snoop Advertisement Route


THANK YOU
Krzysztof Grzegorz Szarkowicz, PLM
kszarkowicz@juniper.net