Routed VPLS using BGP

draft-sajassi-l2vpn-rvpls-bgp-00.txt

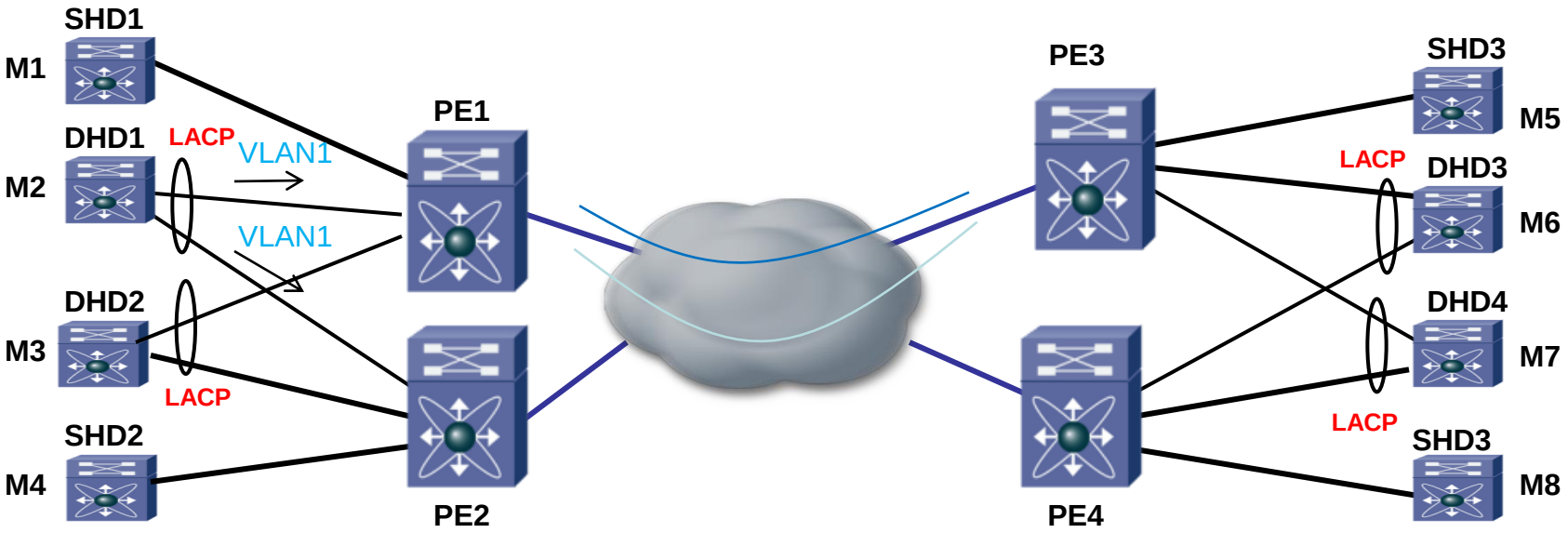
IETF 77, Anaheim, CA
March 2010

- Authors: Ali Sajassi, Samer Salam, Keyur Patel

Requirements

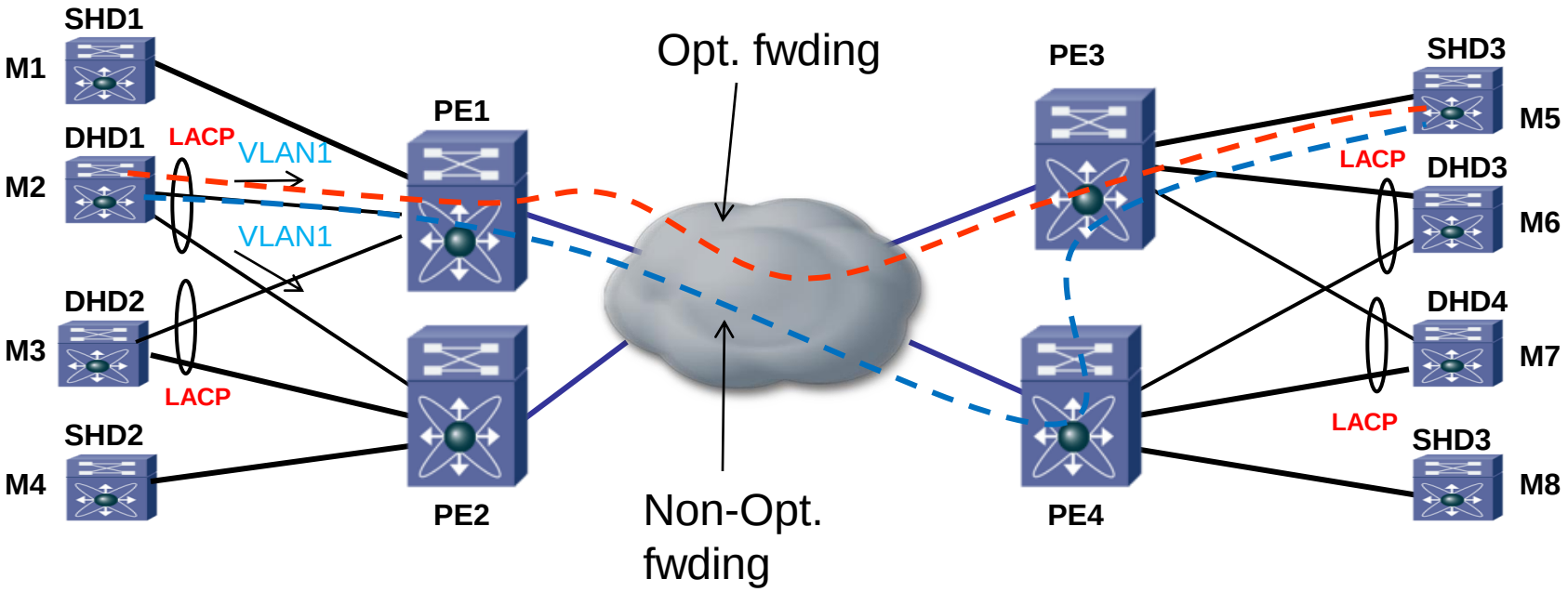
1. Load balancing on L2/L3/L4 flows
2. Flow-based multi-pathing
3. Geo-redundant PE nodes & optimum unicast forwarding
4. Flexible Redundancy grouping
5. Multicast optimization w/ MP2MP
   - e.g., PE1 & PE2 can be used for dual-homing one set of CEs and PE2& PE3 can be used for dual-homing different set of Ces – thus PE2 supporting multiple redundancy groups

Requirements 1 & 2



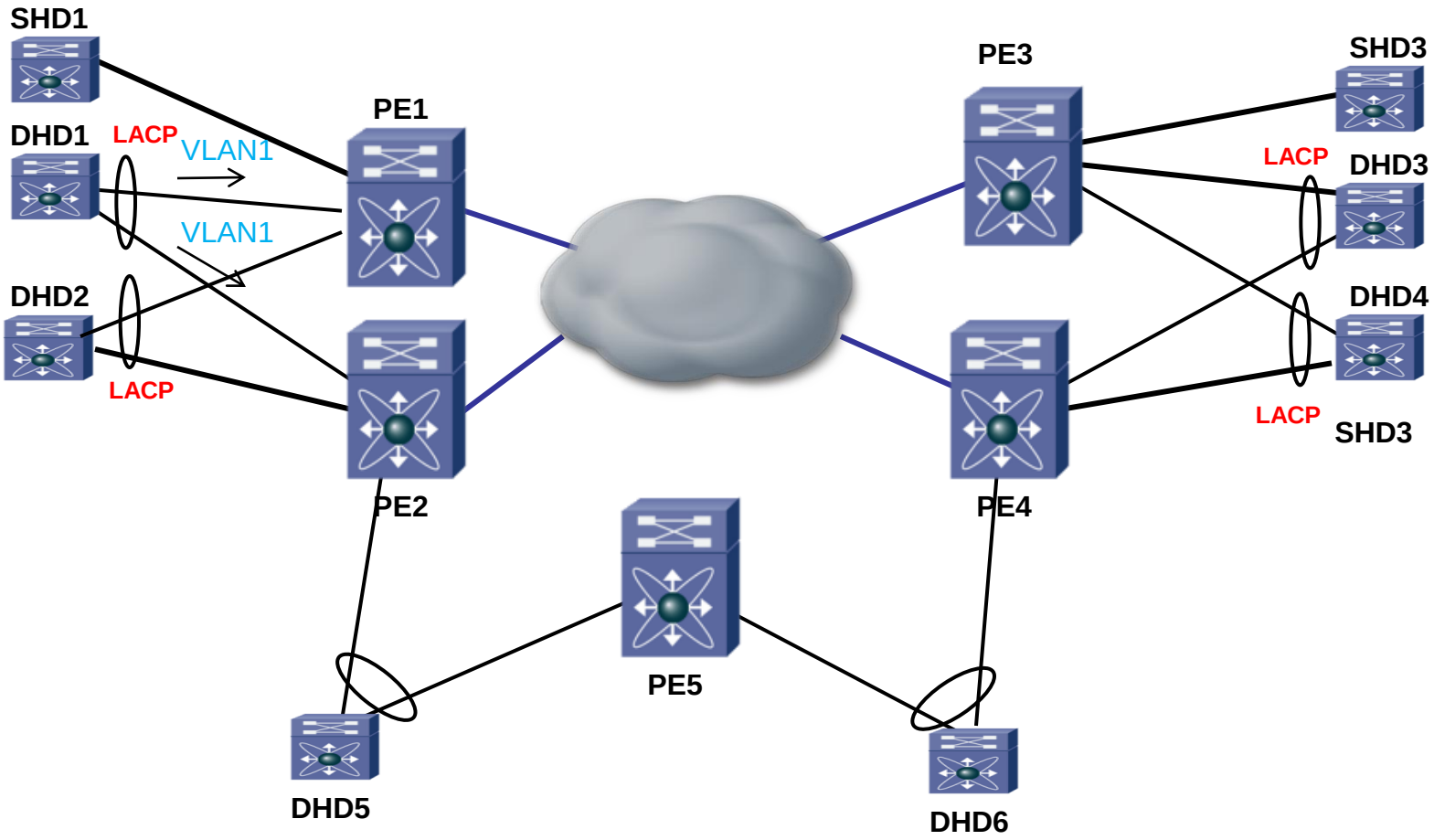
1. Per flow load balancing for a given VLAN
   - L2 flow (MAC DA, MAC SA)
   - L3 flow (IP DA, IP SA)
   - L4 flow (UDP/TCP source port, dest port)
   - Any combination of the above
2. Flow-based multi-pathing

Requirements 3



- Geo Redundancy and optimum unicast forwarding between any pair of CEs
  - single-homed CE to single-homed CE
  - single-homed CE to dual-homed CE
  - dual-homed CE to single-homed CE
  - dual-homed CE to dual-homed CE

Requirements 4

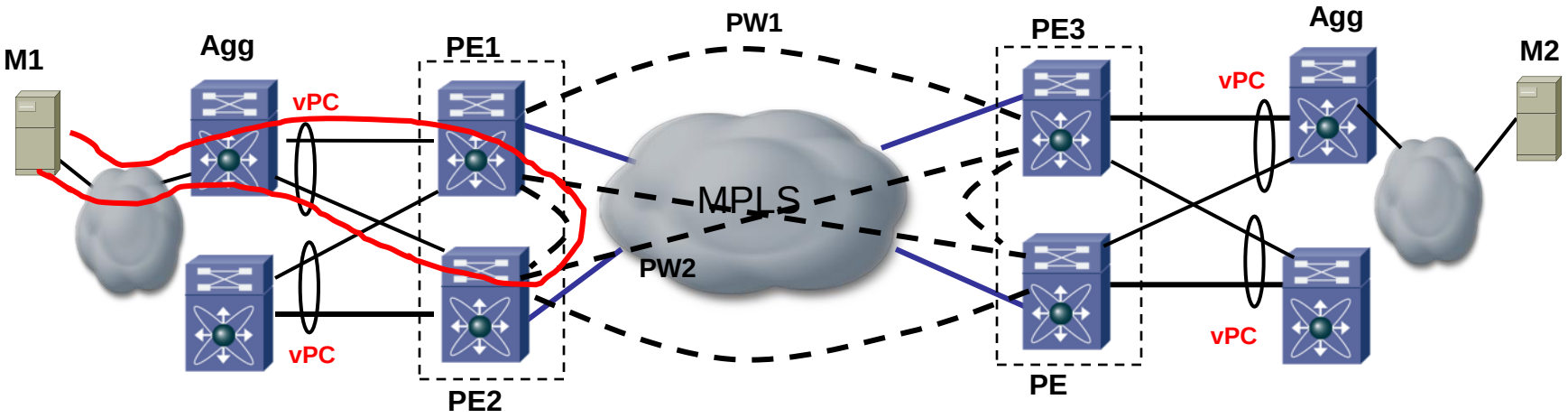


- Flexible Redundancy grouping

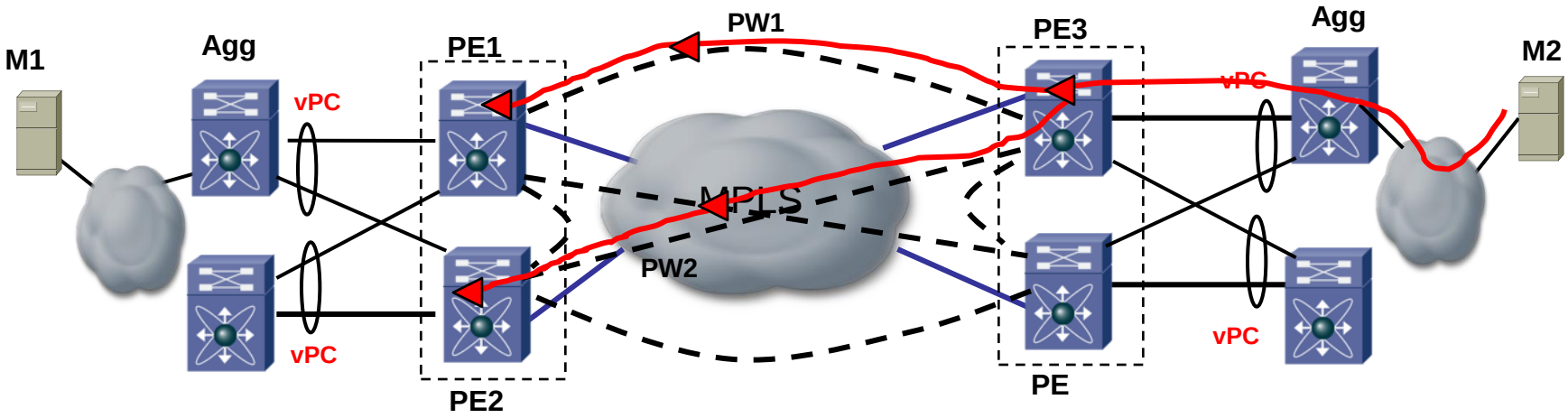
Issues

1. Forwarding Loops
2. Duplicate Frame Delivery
3. MAC Forwarding Table Instability
4. Source identification in MP2MP MDT

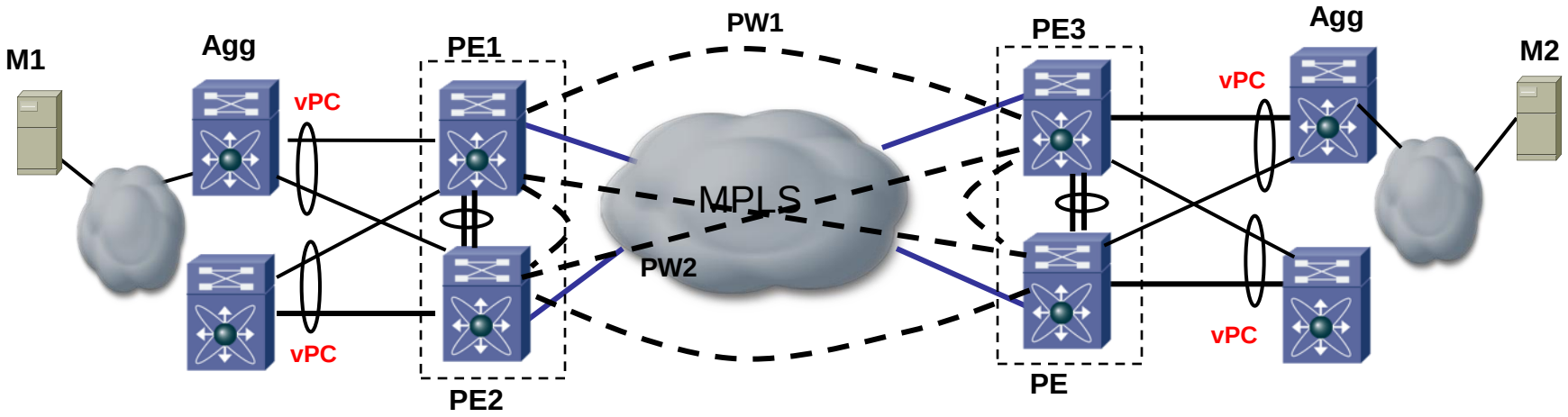
Issue 1: Looping of Traffic Flooded from PE



Issue 2: Duplicate Frames for Floods from the Core



Issue 3: MAC Flip-flopping over PW



- Agg nodes will load-balance traffic over Port Channel based on local algorithm, commonly:
  - L2: MAC SA, DA, both
  - L3: IP SA, DA, both
  - L4: Source Port, Destination Port, both
- Any LB algorithm that doesn't guarantee that a given MAC SA is consistently hashed to the same PE will cause MAC flip-flopping for the remote VPLS PEs.
  - e.g. traffic from M1 to M2 constantly moves between PW1 and PW2

R-VPLS: Routed VPLS

- A conceptually simple solution
  - Treat C-MACs as routable addresses and distribute them in BGP
  - The MAC address are learned in data-plane toward access as before but are distributed over MPLS/IP network using BGP
  - Receiving PE injects these MAC addresses into forwarding table with along with its associated adjacency
  - When multiple PE nodes advertise the same MAC, then multiple adjacency is created for that MAC address in the forwarding table
  - When forwarding traffic for a given unicast MAC DA, a hashing algorithm based on L2/L3/L4 hdr is used to pick one of the adjacencies for forwarding

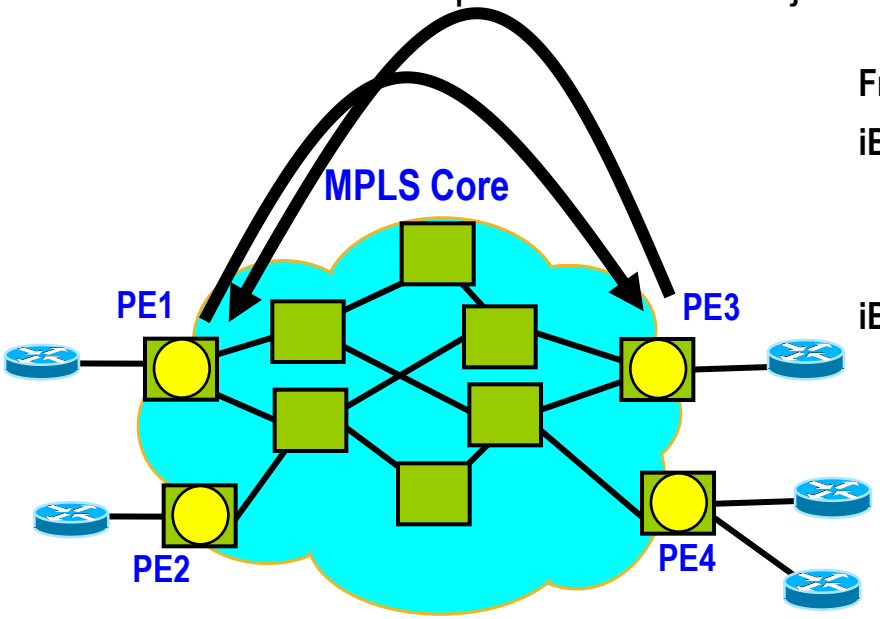
From PE1

iBGP L3-NLRI:

- **next-hop: n-PE1**
- **<C-IP1, L1>**

iBGP L2-NLRI

- **next-hop: n-PE1**
- **<C-MAC1, L2>**



From PE3

iBGP L3-NLRI:

- **next-hop: n-PE3**
- **<C-IP5, L1>**

iBGP L2-NLRI

- **next-hop: n-PE3**
- **<C-MAC3, L2>**

Description

- C-MAC SAs are learned over ACs. If it is a new entry, then it is sent to the control plane to be distributed via BGP with a MPLS label identifying VSI (similar to L3VPN where a label identifies the VRF)
  - A single MPLS label per VPLS instance is sufficient (e.g., MP2P connotation just like L3VPN)
- BGP NLRI is used with new AFI/SAFI to advertise these routable MACs to other PE
- In case of a dual-homed CE, when a MAC is learned by two PEs, then both PEs advertise the same MAC with different RDs
  - Remote PEs can install both paths for that MAC address
  - Remote PEs can use L2/L3/L4 hashing to pick among the BGP ECMP paths when forwarding based on that MAC address

Description – Cont.

- Known Unicast MACs
  - Forward them based on L2FIB entry
  - If there are multiple destinations, then select one based on L2/L3/L4 header hash
- Unknown Unicast MACs
  - Forwarding of these frames are optional
- Multicast/Broadcast MACs
  - Send these frames over MP2MP LSP or P2MP LSP or full-mesh of P2P PWs
  - Regardless of what LSP is used to send these frames, no MAC learning is performed when these frames arrive at egress PEs (thus resulting in simpler operation !!)
  - MH-ID is used to ensure that a single UNI in the multi-homed group is selected to sends the multicast/broadcast frames out toward the customers (thus avoiding possibility for any loop)

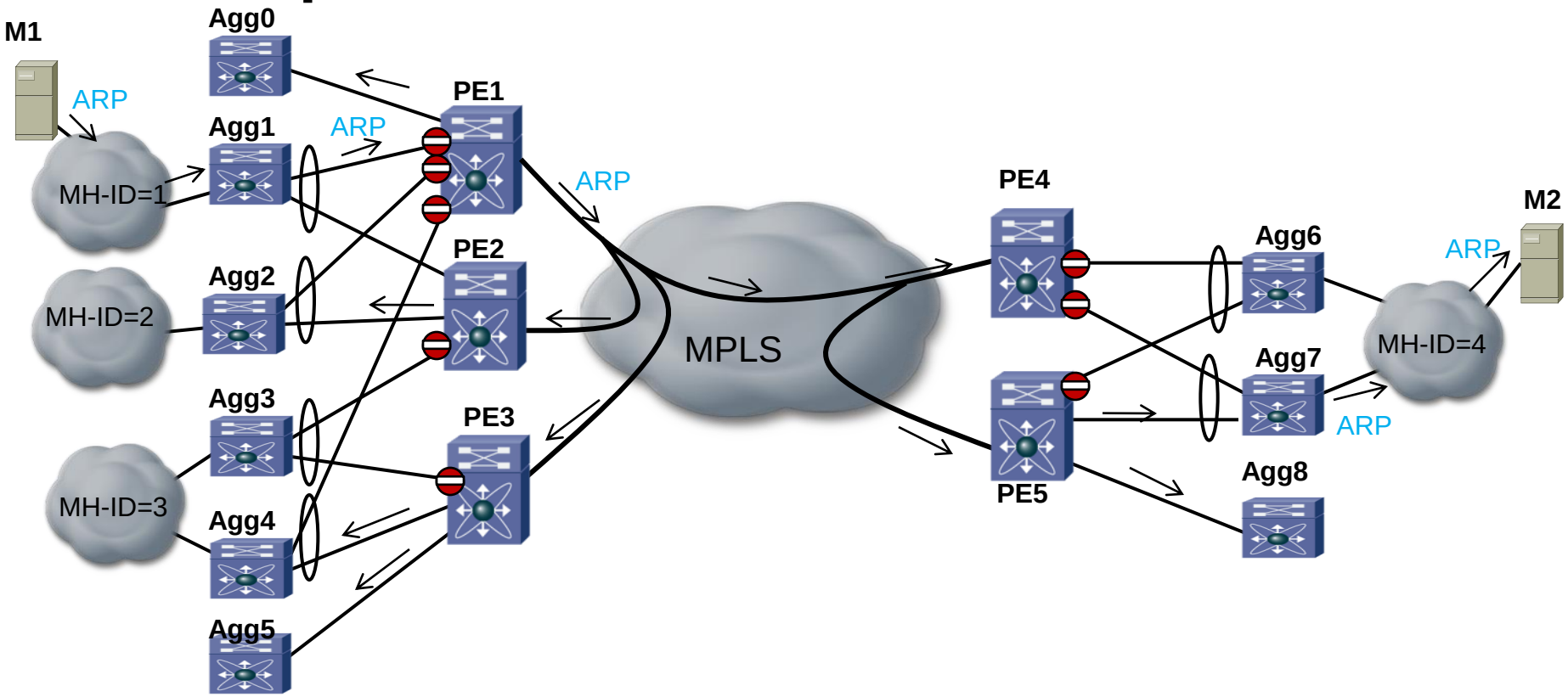
Operation: General

- Perform VPLS auto-discovery as before and setup a single MP2MP tunnel per VPLS instance instead of full mesh of PWs
- When a PE receives a broadcast/multicast frame, it distributes the MAC-SA via BGP (if new) to all other PEs and sends the frame over MP2MP LSM
- On the far-end PE just forward the frame over local ACs (no learning)
- If a PE receives a frame with unknown MAC DA, then simply discard the frame by default (or optionally forward it)
- If a PE receives a frame with known MAC DA, then forward it using MP2P label associated with that VPLS instance (VSI)
- If P2MP tunnels are needed for some customer flows, set them up as required

Operation: Loop Prevention

- In order to prevent loop for multicast/broadcast frames, the following simple mechanism is used:
  - For a multi-homed DHD or DHN with several active ACs, only a single AC can be a designated forwarder for the multicast/broadcast traffic
  - MH-ID & DF procedure is used per draft-l2vpn-vpls-multihoming to select a single DF in a group of ACs
  - All multicast/broadcast Ethernet frames are marked with a MH-ID label to identify the source multi-homed site
  - A PE that receives a multicast/broadcast frame from the WAN, it filters out that frame over an AC whose MH-ID matches the one in the received frame

Operational Scenario: ARP

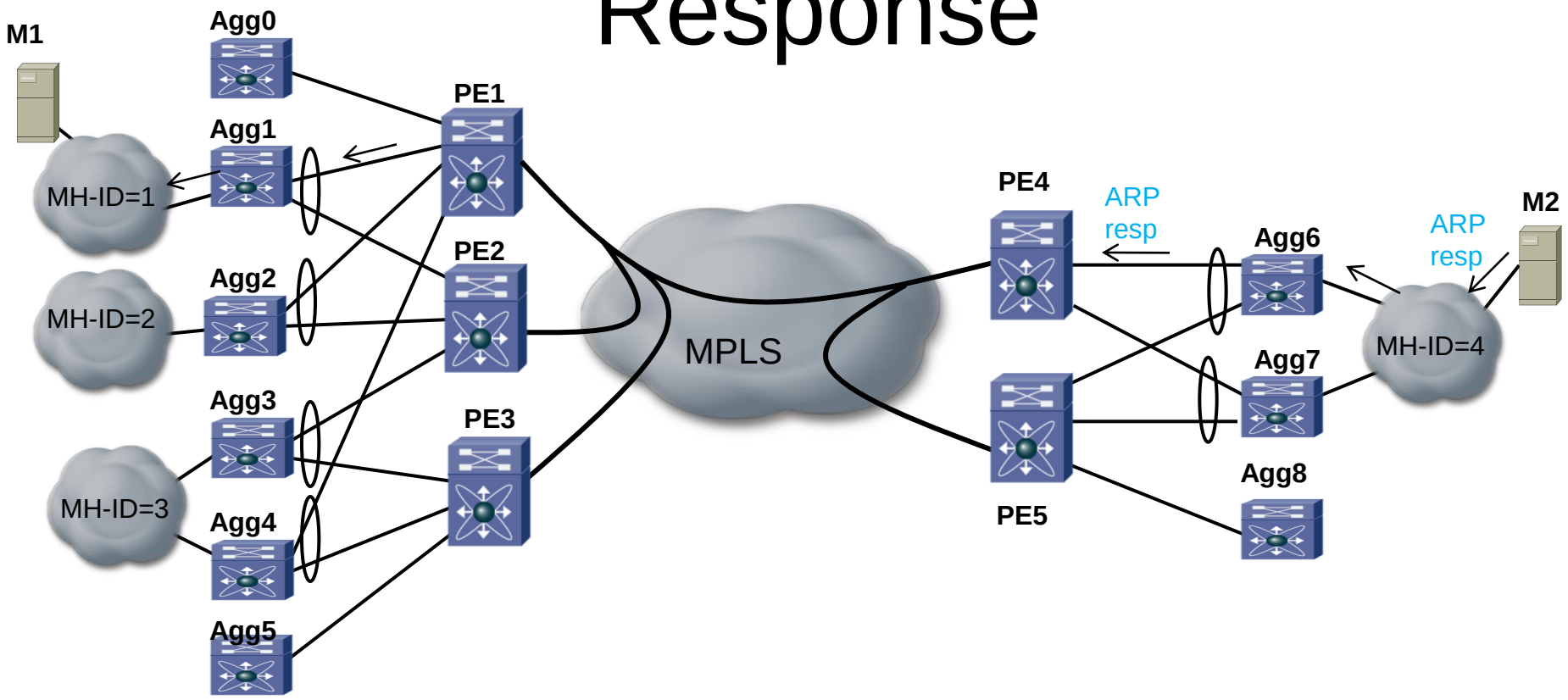


- Host M1 sends an ARP message with MAC SA = M1 and MAC DA=bcast
- PE1 learns M1 over its Agg1-PE1 AC and distributes it via BGP to other PE devices
- All other PE devices learn that M1 sits behind PE1

Operational Scenario: ARP – cont.

- PE1 also sends this ARP message over all its local ACs that are not blocked (for mcast/bcast) as well as it sends it over MP2MP LSP associated with that VPLS instance
  - Only a single AC per MH-ID can be a designated forwarder (DF) to send (but not receive) mcast/bcast messages to the customer site
  - Any AC in the group (per MH-ID) can receive mcast/bcast messages
- PE2 receives the ARP message but it drops it at its Agg1-PE2 AC even though this AC is a DF for MH-ID=1 because
  - – MH-ID of the frame matches the MH-ID of the AC
- PE2 and all other PEs send this ARP message over its non-blocked ACs (for mcast/bcast frames)
  - – Where the MH-ID of the frame is different from that of the ACs

Operational Scenario: ARP Response



- Host M2 sends an ARP response with MAC SA = M2 and MAC DA = M1
- PE4 learns M2 over its Agg6-PE4 AC and distributes it via BGP to other PE devices
- All other PE devices learn that M2 sits behind PE4

Operational Scenario: ARP Response – cont.

- Since PE4 already knows that M1 sits behind PE1, it forwards the frame to PE1
  - If PE4 has two BGP ECMP for M1 (e.g., both PE1 & PE2 have already advertise M1), then it uses a hash based on L2/L3/L4 header to decide which of the two PEs to forward the frame to
- Upon receiving the frame, PE1 does a MAC lookup and forwards the frame to Agg1-PE1 AC

Next steps

- Solicit feedback from the working group
- Progress this work item toward WG draft