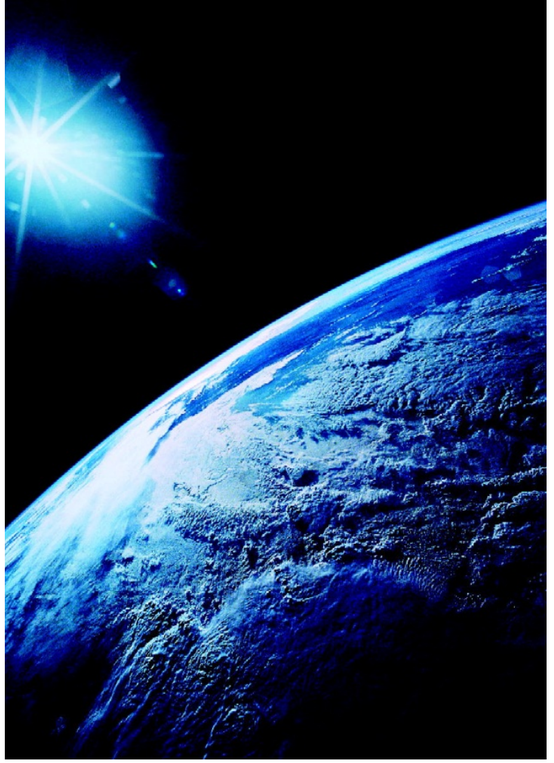
Cisco Connect

Praha, hotel Clarion
10. – 11. dubna 2013

Evoluce MPLS L2VPN



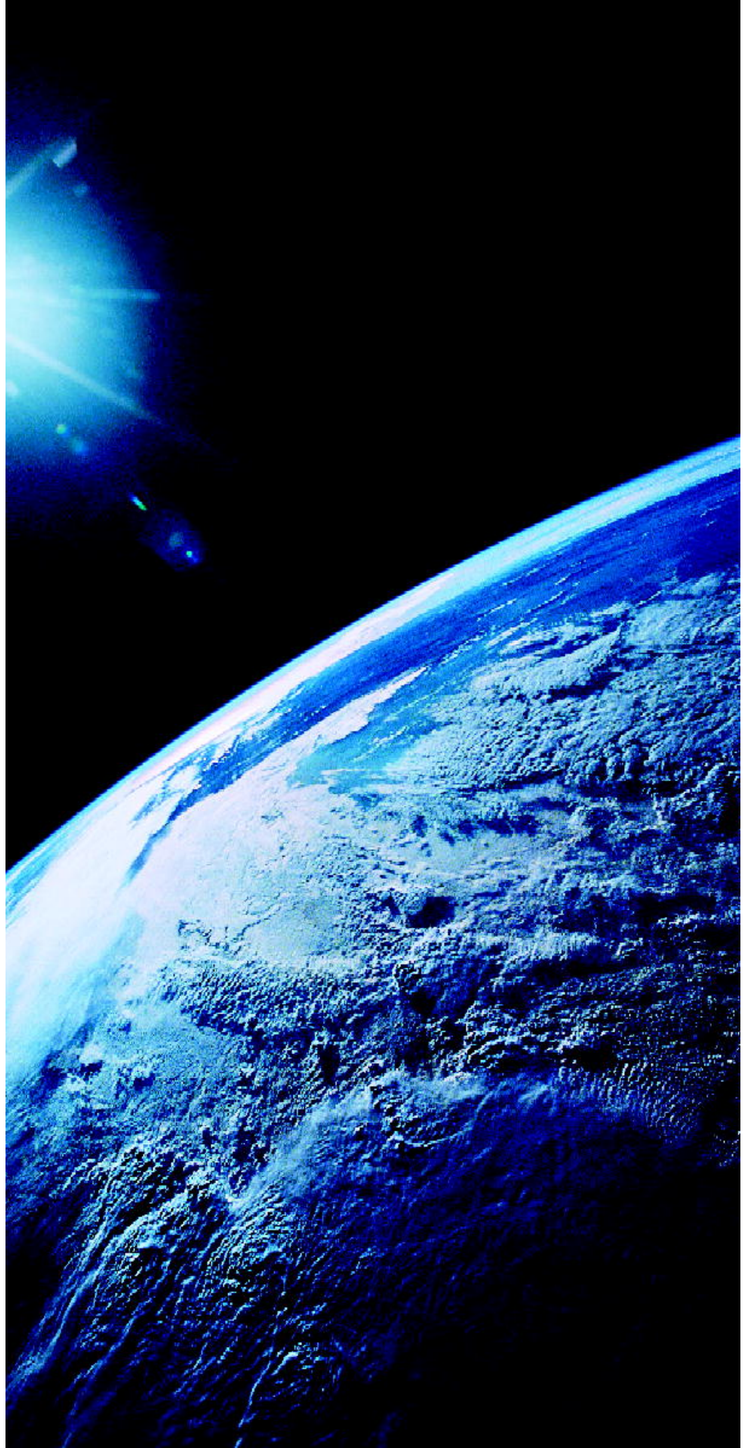
T-SP2/L2

Jiri Chaloupka – Cisco
Systems Engineer

Agenda

- Layer 2 VPN Motivation and Overview
- Virtual Private Wire Service (VPWS)
- Virtual Private LAN Service (VPLS)
- Next Generation L2VPN Solutions
  - E-VPN/PBB-EVPN
- E-VPN/PBB-EVPN – life of packet
- Sample Supported Access Topologies
- Summary
  - Comparison of L2VPN Solutions

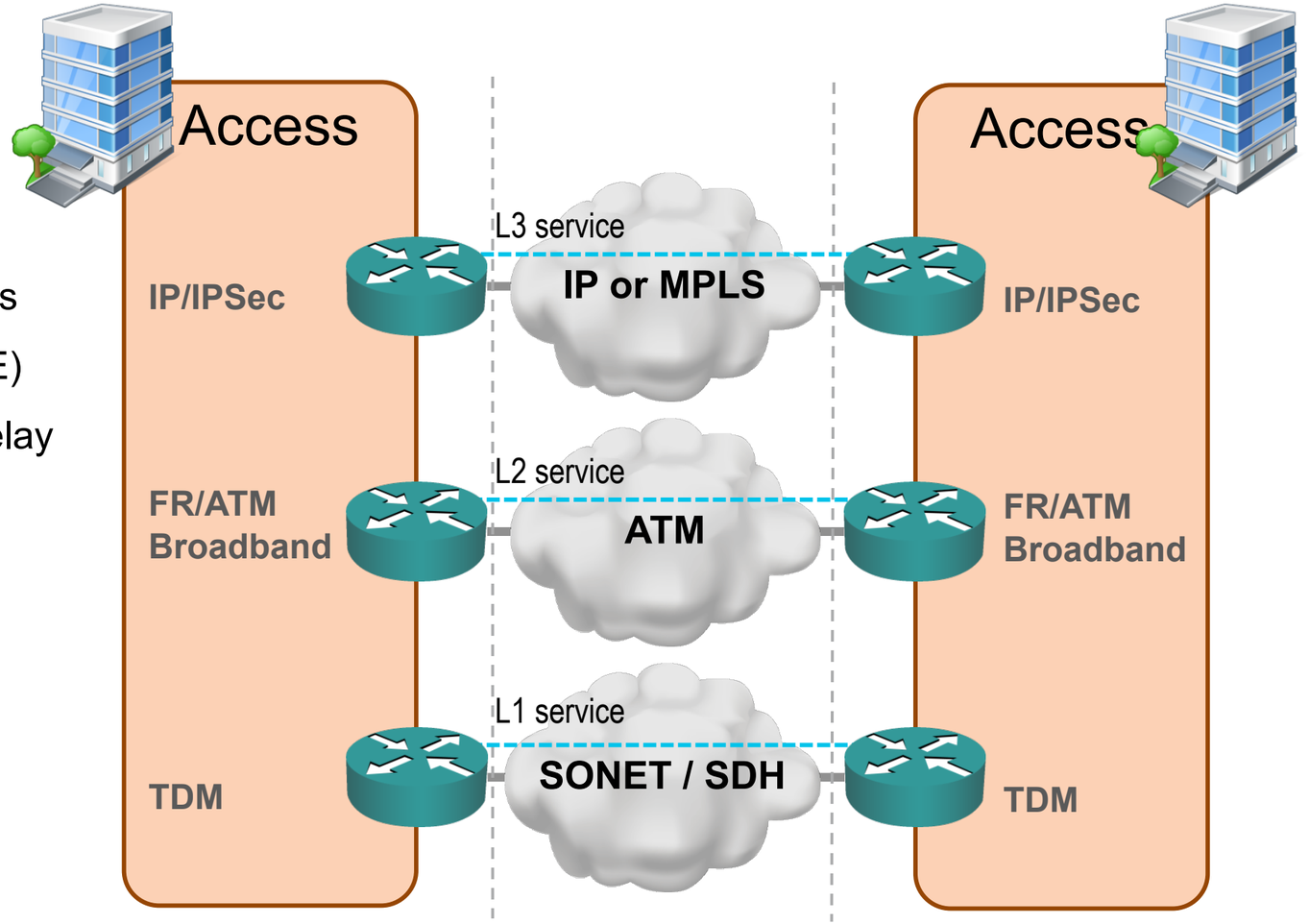
L2VPN Motivation and Overview



Motivation for L2VPNs

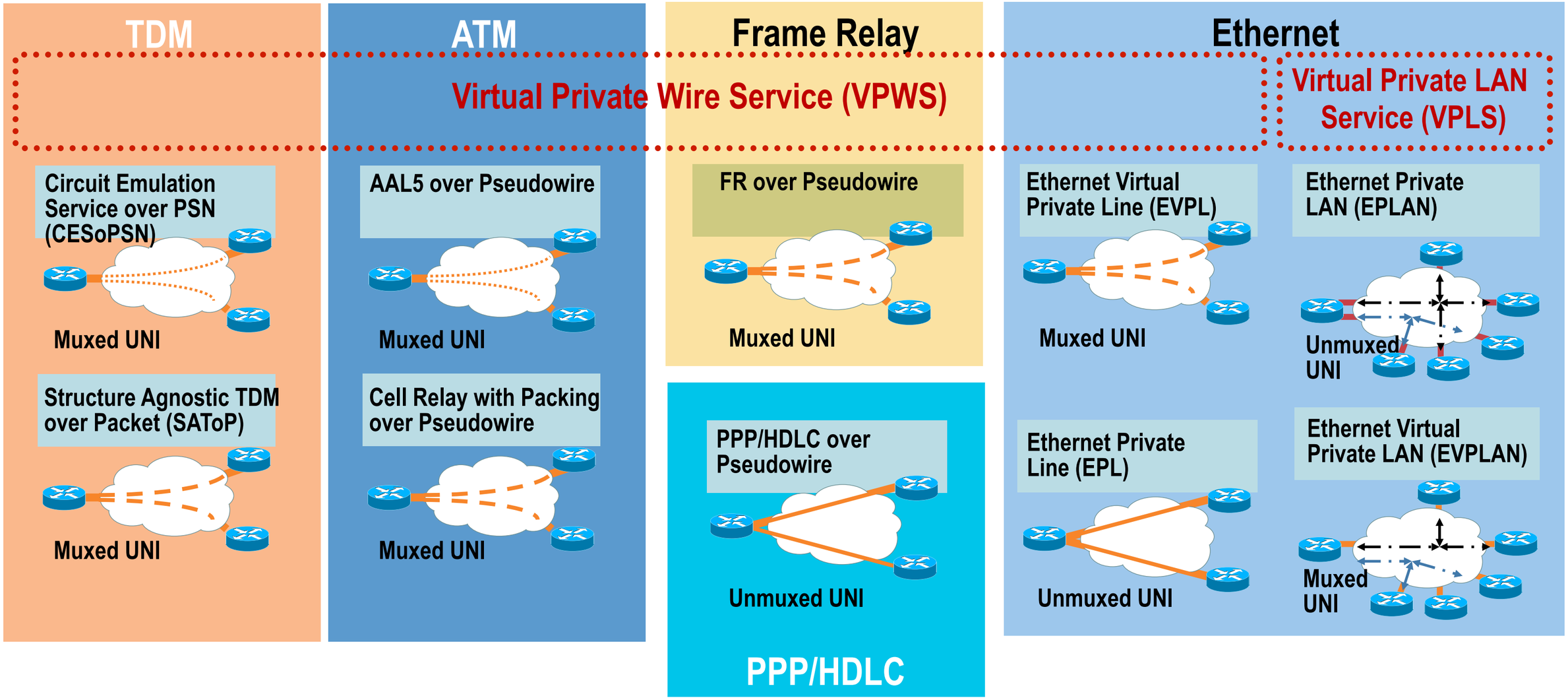
Old and New Drivers

- Network Consolidation
  - Multiple access services (FR, ATM, TDM) required multiple core technologies
- Enterprise Ethernet WAN Connectivity
  - Ethernet well understood by Enterprise / SPs
  - CAPEX (lower cost per bit) / Growth (100GE)
  - Layer 2 VPN replacement to ATM/Frame Relay
  - Internet / Layer 3 VPN access (CE to PE)
- Data Center Interconnection (DCI)
- Mobile Backhaul Evolution
  - TDM /PDH to Dual/Hybrid to All-packet (IP/ Ethernet)
  - Single (voice + data) IP/Ethernet mobile backhaul universally accepted solution



Service Offerings

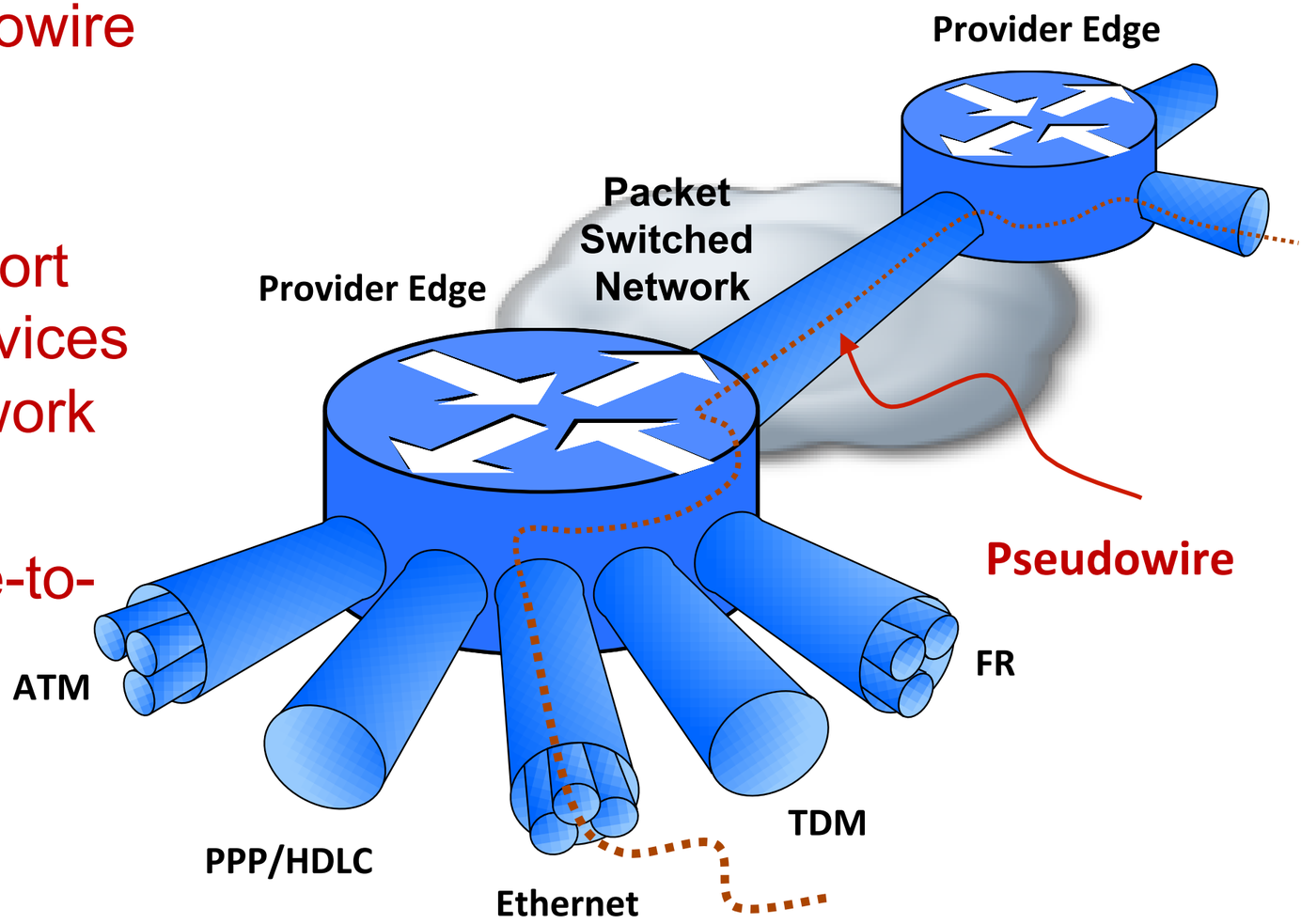
L2VPN Transport Services



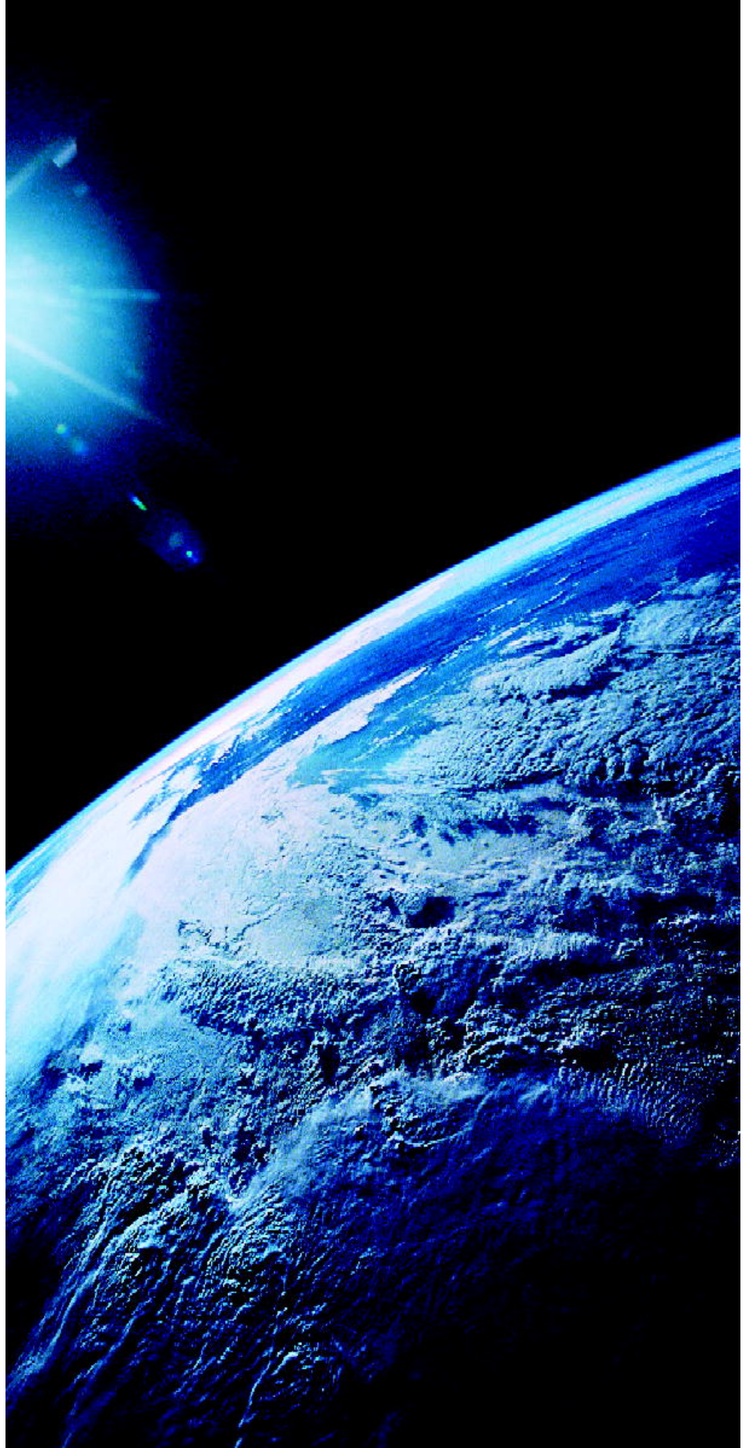
Layer 2 VPN Enabler

The Pseudowire

- L2VPNs are built with Pseudowire (PW) technology
- PWs provide a common intermediate format to transport multiple types of network services over a Packet Switched Network (PSN)
- PW technology provides Like-to-Like transport and also Interworking (IW)

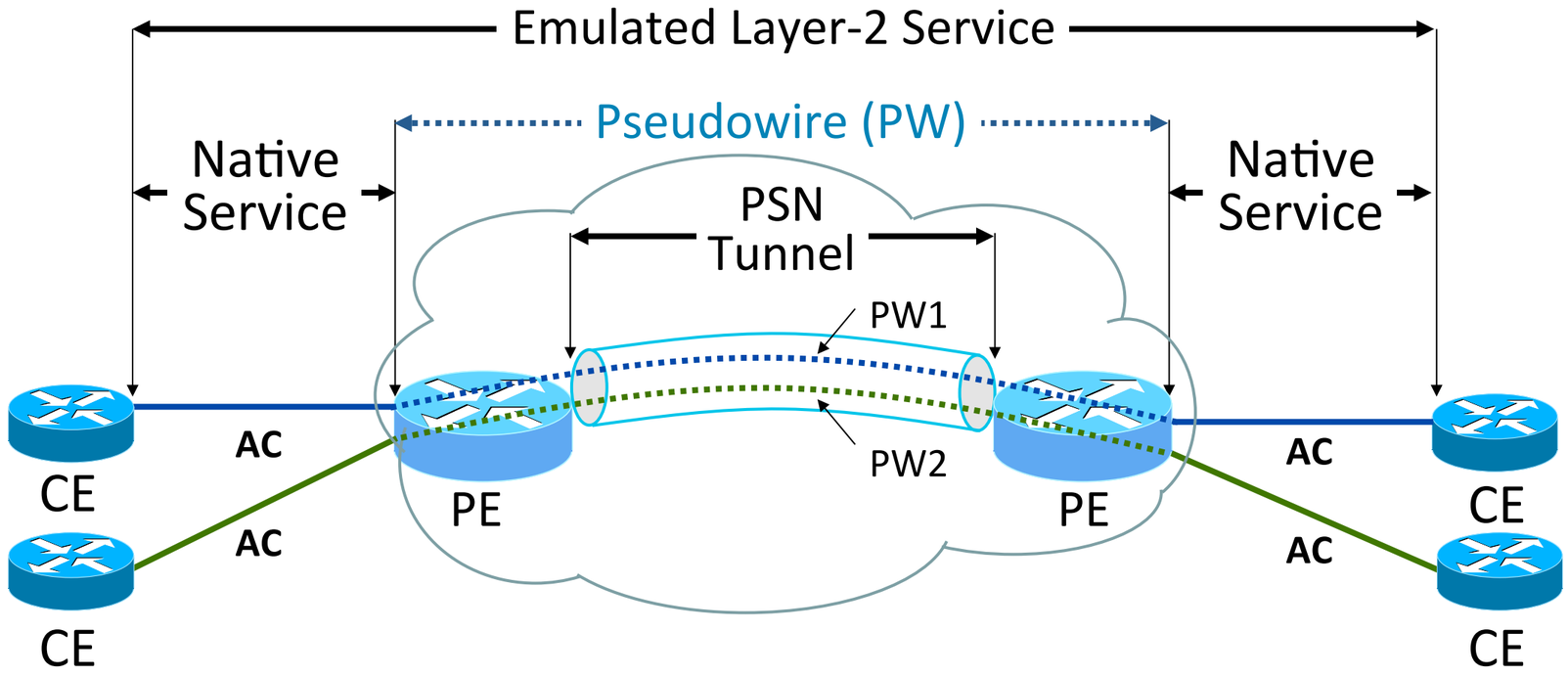


Virtual Private Wire Service (VPWS)
Overview



Pseudowire Reference Model

- Any Transport Over MPLS (AToM) is Cisco's implementation of VPWS for IP/MPLS networks
- An Attachment Circuit (AC) is the physical or virtual circuit attaching a CE to a PE
- Customer Edge (CE) equipment perceives a PW as an unshared link or circuit



Ref: RFC 3985 Pseudo Wire Emulation Edge-to-Edge (PWE3) Architecture, March 2005

Layer 2 Transport over MPLS

Control Connection

- Targeted LDP session / BGP session / Static
  - Used for VC-label negotiation, withdrawal, error notification

The “emulated circuit” has three (3) layers of encapsulation

Tunnelling Component

- Tunnel header (Tunnel Label)
  - To get PDU from ingress to egress PE
  - MPLS LSP derived through static configuration (MPLS-TP) or dynamic (LDP or RSVP-TE)

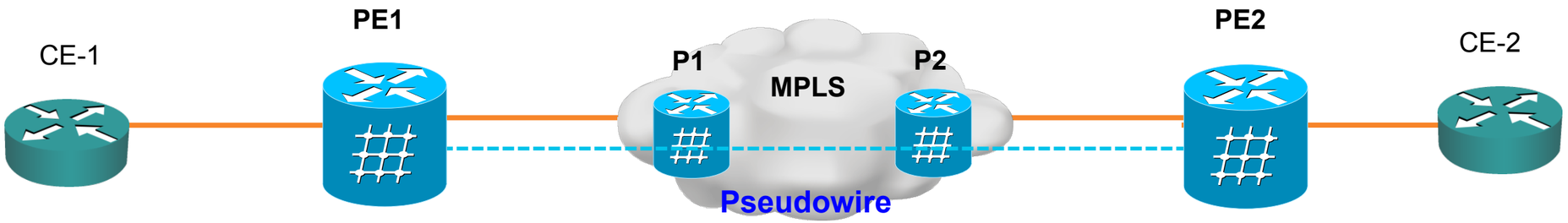
Demultiplexing Component

- Demultiplexer field (VC Label)
  - To identify individual circuits within a tunnel
  - Could be an MPLS label, L2TPv3 header, GRE key, etc.

Layer 2 Encapsulation

- Emulated VC encapsulation (Control Word)
  - Information on enclosed Layer 2 PDU
  - Implemented as a 32-bit control word

VPWS Forwarding Plane Processing



Traffic direction

VC and Tunnel label imposition

Push
Push

Tunnel label swapping through MPLS cloud

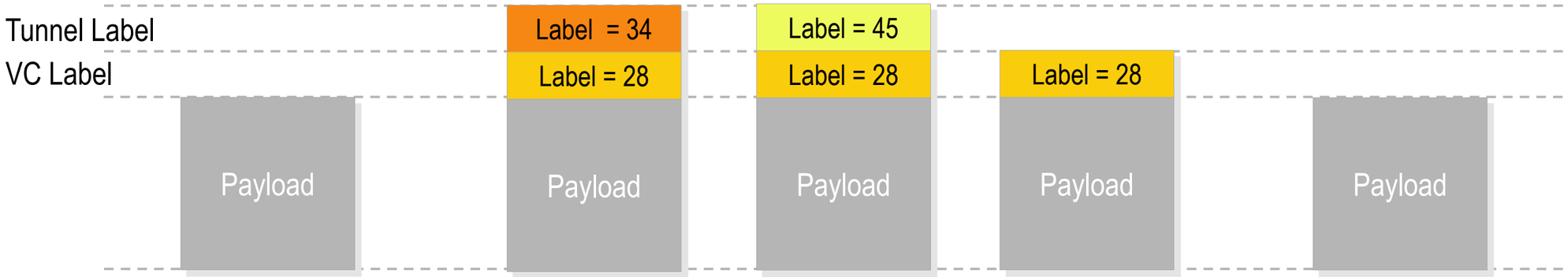
Swap

Penultimate Hop Popping (PHP)

Pop

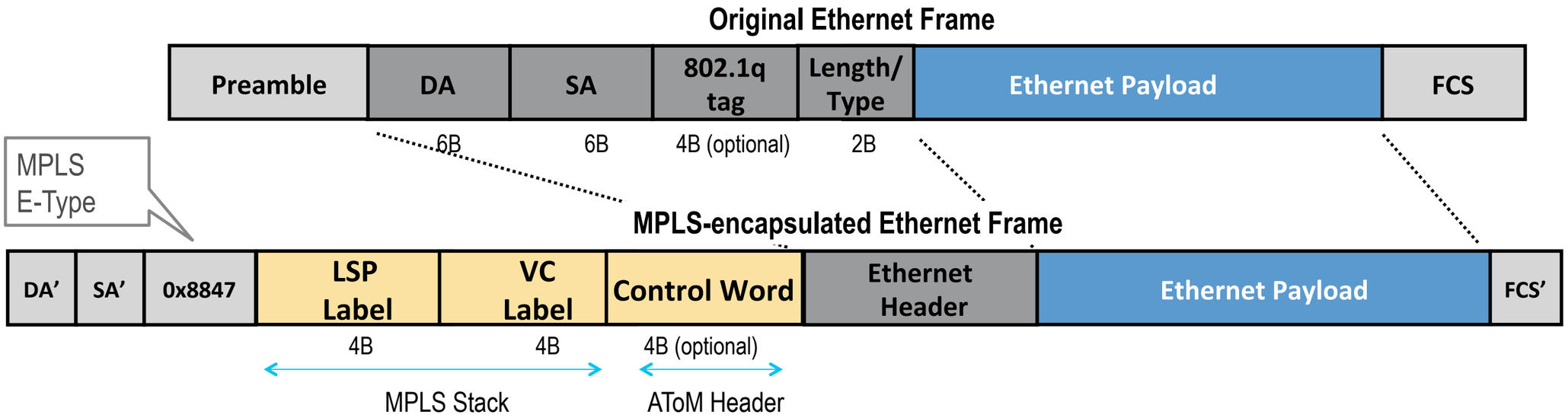
VC label disposition

Pop



How Are Ethernet Frames Transported?

- Ethernet frames transported without Preamble, Start Frame Delimiter (SFD) and FCS
- Two (2) modes of operation supported:
  - Ethernet VLAN mode (VC type 0x0004) – created for VLAN over MPLS application
  - Ethernet Port / Raw mode (VC type 0x0005) – created for Ethernet port tunneling application



Ethernet PW VC Type Negotiation

Cisco IOS

- Cisco devices by default will generally attempt to bring up an Ethernet PW using VC type 5
- If rejected by remote PE, then VC type 4 will be used
- Alternatively, Cisco device can be manually configured to use either VC type 4 or 5

```
7604-2(config-pw-class)#interworking ?
  ethernet  Ethernet interworking
  ip        IP interworking
  vlan      VLAN interworking

7604-2#show running-config
pseudowire-class test-pw-class-VC4
 encapsulation mpls
 interworking vlan
!
pseudowire-class test-pw-class-VC5
 encapsulation mpls
 interworking ethernet
```

Ethernet PW VC Type Negotiation

Cisco IOS-XR

- Cisco devices by default will generally attempt to bring up an Ethernet PW using VC type 5
- If rejected by remote PE, then VC type 4 will be used
- Alternatively, Cisco device can be manually configured to use either VC type 4 or 5

```
RP/0/RSP0/CPU0:ASR9000-2(config-l2vpn-pwc-
mpls)#transport-mode ?
  ethernet  Ethernet port mode
  vlan      Vlan tagged mode
RP/0/RSP0/CPU0:ASR9000-2(config-l2vpn-pwc-
mpls)#transport-mode vlan ?
  passthrough  passthrough incoming tags

RP/0/RSP0/CPU0:ASR9000-2#show running-config l2vpn
l2vpn
 pw-class test-pw-class-VC4
  encapsulation mpls
   transport-mode vlan

 pw-class test-pw-class-VC4-passthrough
  encapsulation mpls
   transport-mode vlan passthrough

 pw-class test-pw-class-VC5
  encapsulation mpls
   transport-mode ethernet
```

Introducing Cisco EVC Framework

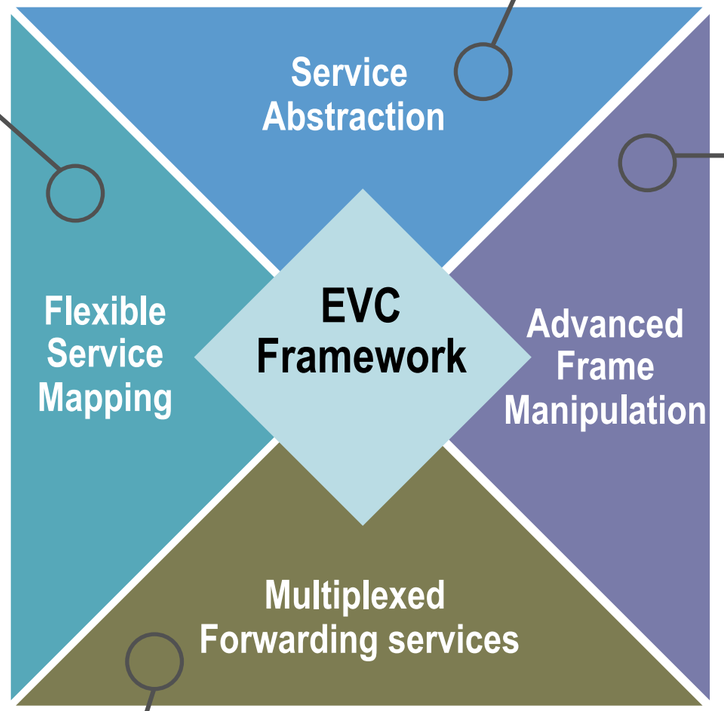
Functional Highlights

Ethernet Service Layer

- Ethernet Flow Point (EFP)
- Ethernet Virtual Circuit (EVC)
- Bridge Domain (BD)
- Local VLAN significance

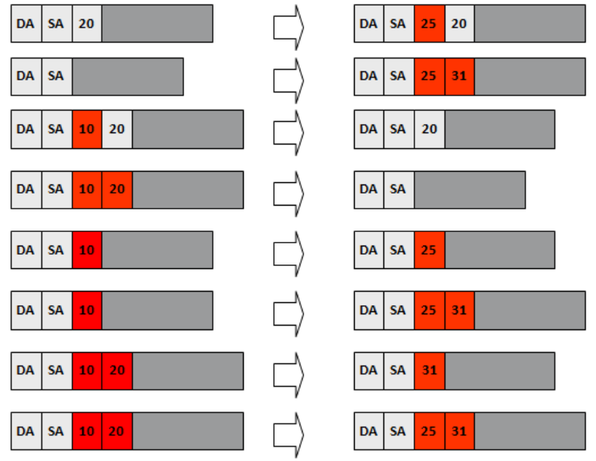
Flexible service delimiters

- Single-tagged, Double-tagged
- VLAN Lists, VLAN Ranges
- Header fields (COS, Ethertype)



VLAN Header operations - VLAN Rewrites

- POP
- PUSH
- SWAP

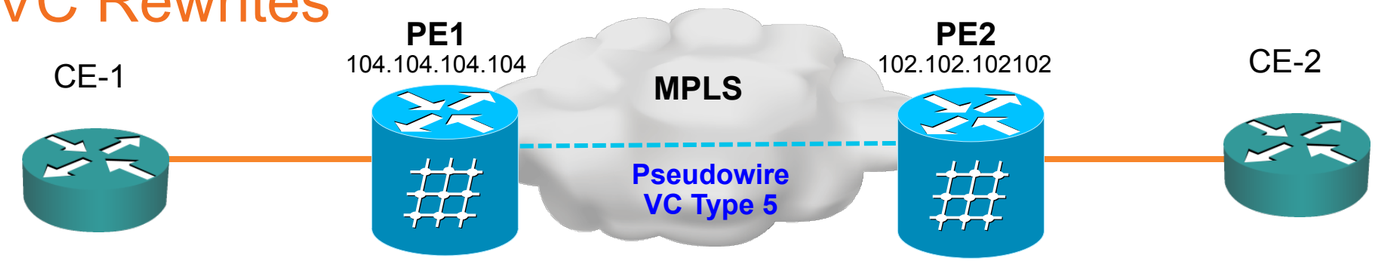


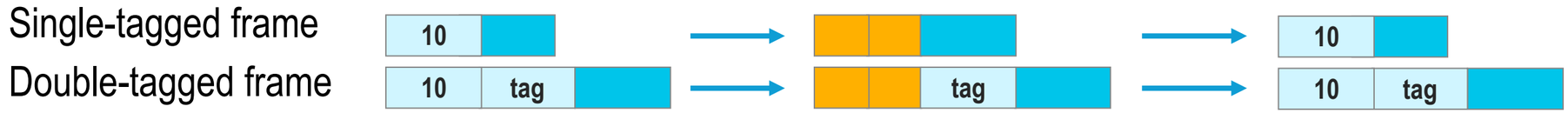
ANY service – ANY port

- Layer 2 Point-to-Point
- Layer 2 Multipoint
- Layer 3

Encapsulation Adjustment Considerations

VC 5 and EVC Rewrites





IOS-XR

```
l2vpn
 pw-class class-VC5
  encapsulation mpls
   transport-mode ethernet

 xconnect group Cisco
  p2p xc-sample-1
   interface GigabitEthernet0/0/0/2.100
   neighbor 102.102.102.102 pw-id 111
    pw-class class-VC5
```

```
interface GigabitEthernet0/0/0/2.100 l2transport
 encapsulation dot1q 10
 rewrite ingress tag pop 1 symmetric
```

- POP VLAN 10
- No Push of Dummy tag (VC 5)

- No service-delimiting vlan expected (VC 5)
- PUSH VLAN 10

IOS

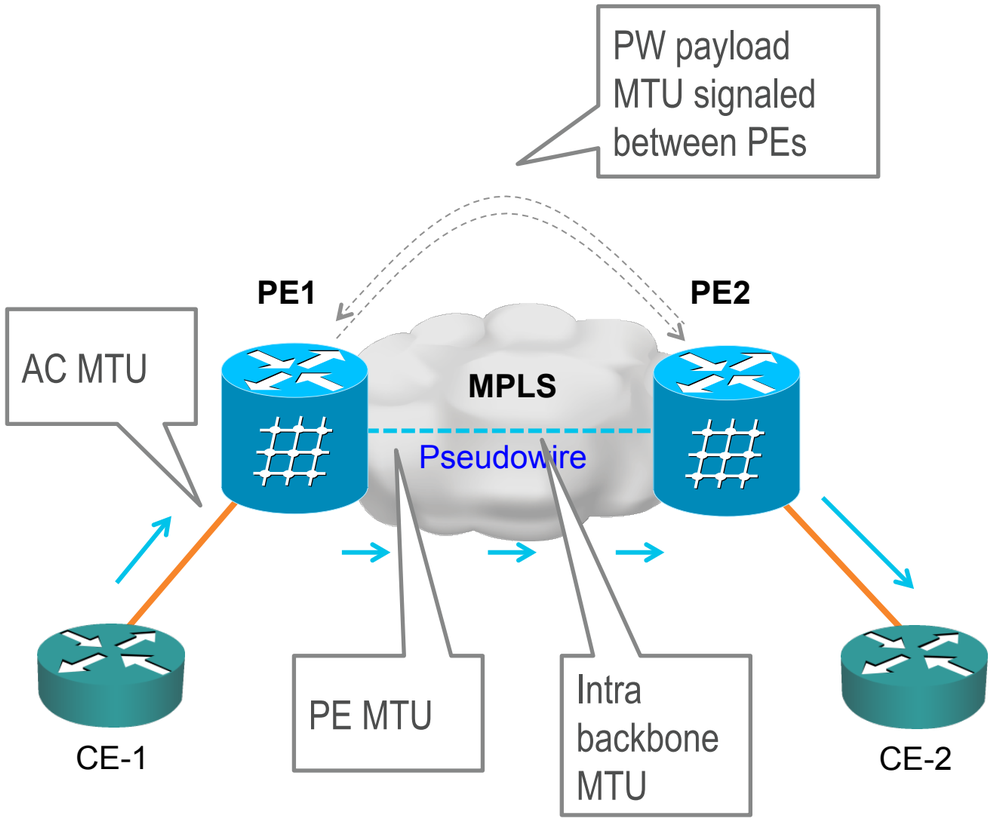
```
pseudowire-class class-VC5
 encapsulation mpls
 interworking ethernet
```

```
interface GigabitEthernet2/2
 service instance 3 ethernet
 encapsulation dot1q 10
 rewrite ingress tag pop 1 symmetric
 xconnect 104.104.104.104 111 encap mpls pw-class class-VC5
```

MPLS label

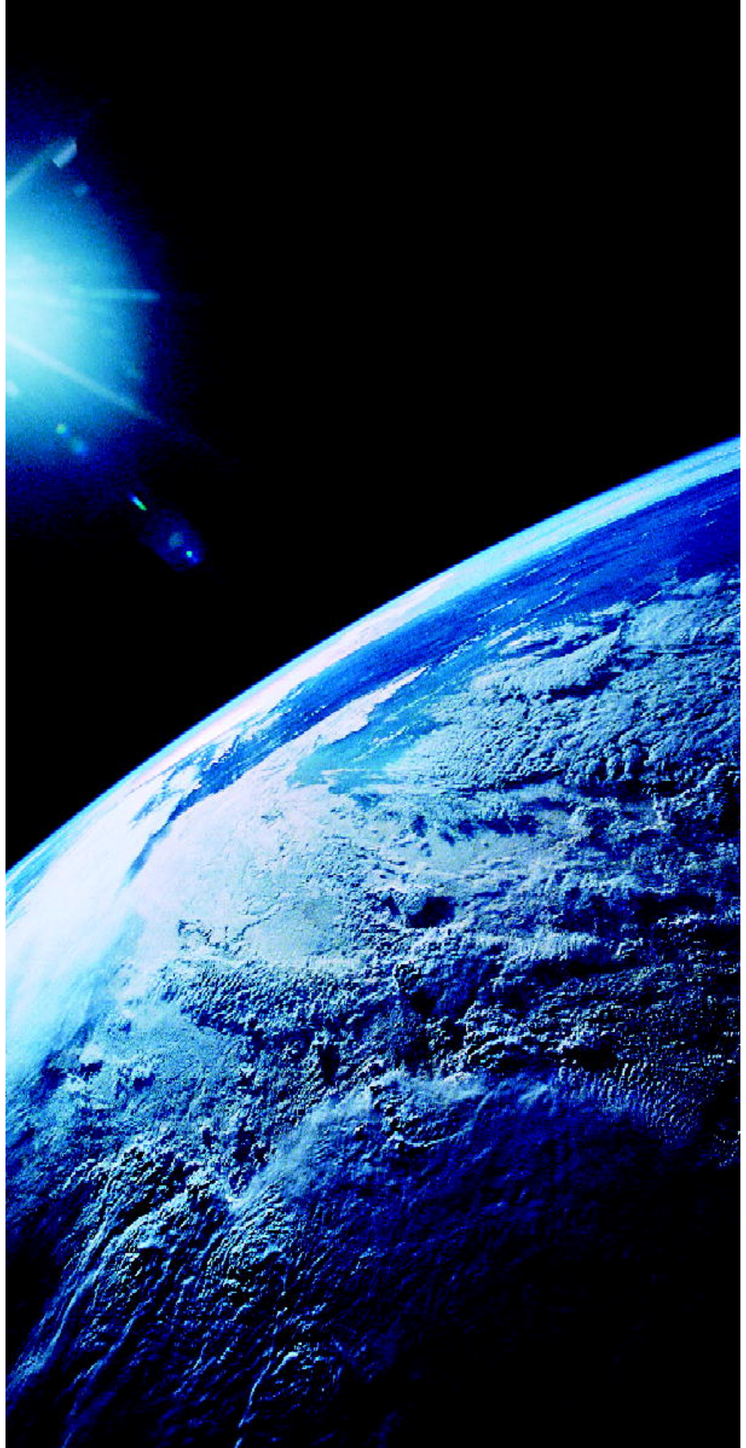
MTU Considerations

- No payload fragmentation supported
- Incoming PDU dropped if MTU exceeds AC MTU
- PEs exchange PW payload MTU as part of PW signaling procedures
  - Both ends must agree to use same value for PW to come UP
  - PW MTU derived from AC MTU
- No mechanism to check Backbone MTU
  - MTU in the backbone must be large enough to carry PW payload and MPLS stack



Virtual Private LAN Service (VPLS)

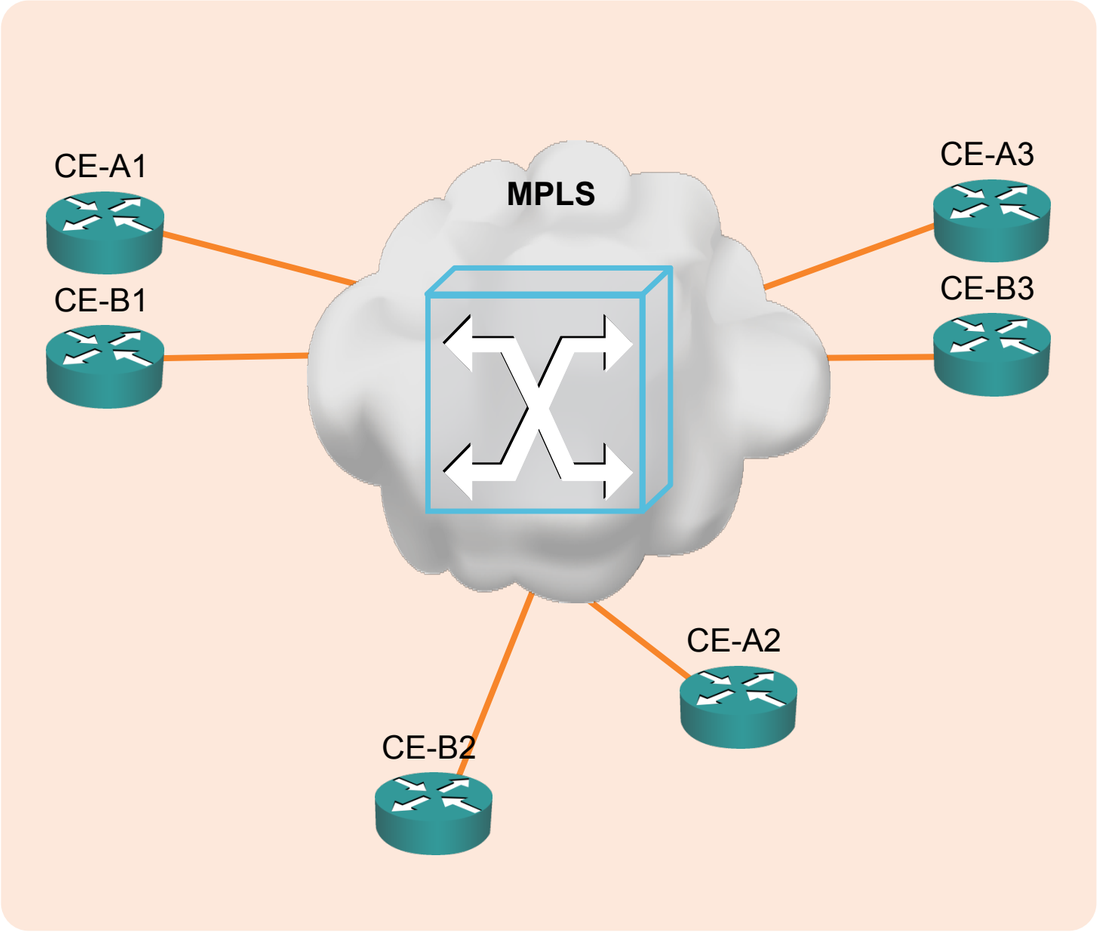
Overview



Virtual Private LAN Service

Overview

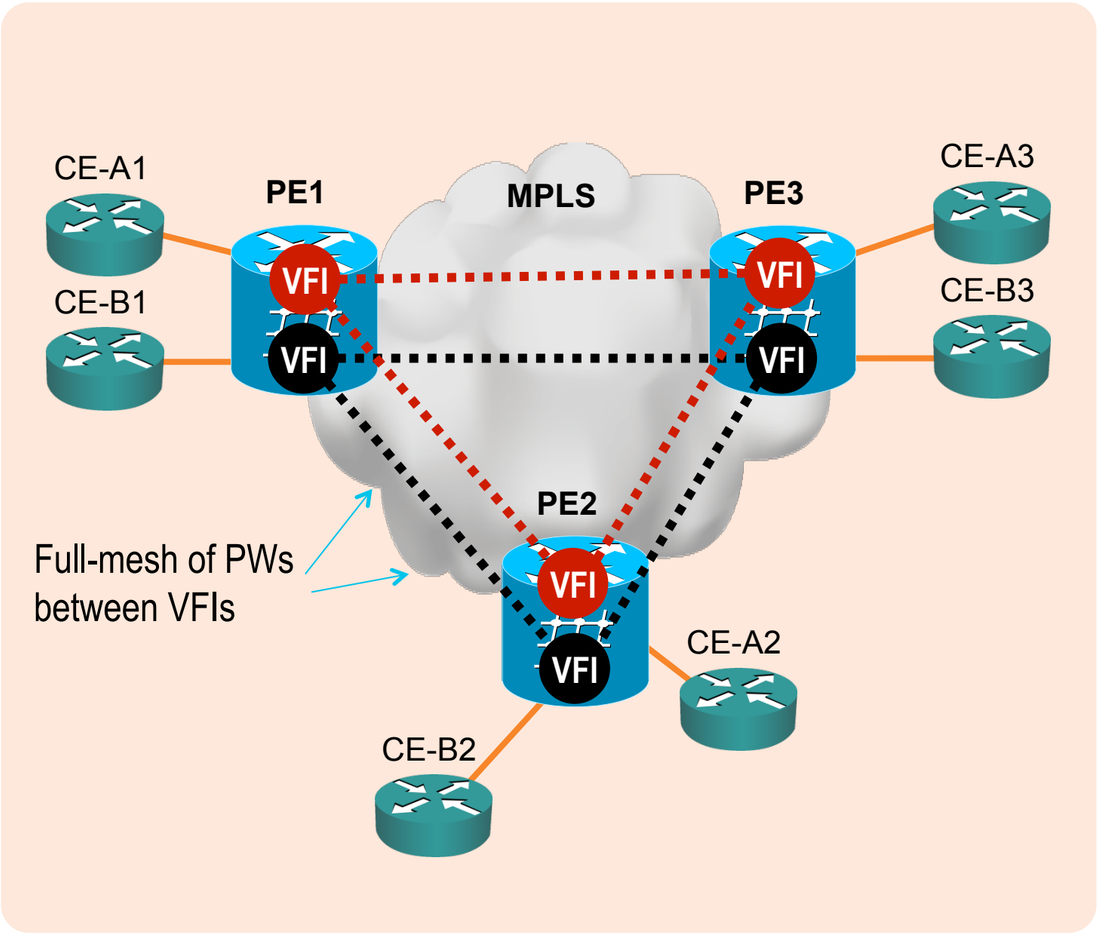
- Defines Architecture to provide Ethernet Multipoint connectivity sites, as if they were connected using a LAN
- VPLS operation emulates an IEEE Ethernet switch
- Two (2) signaling methods
  - RFC 4762 (LDP-Based VPLS)
  - RFC 4761 (BGP-Based VPLS)



Virtual Private LAN Service

Reference Model

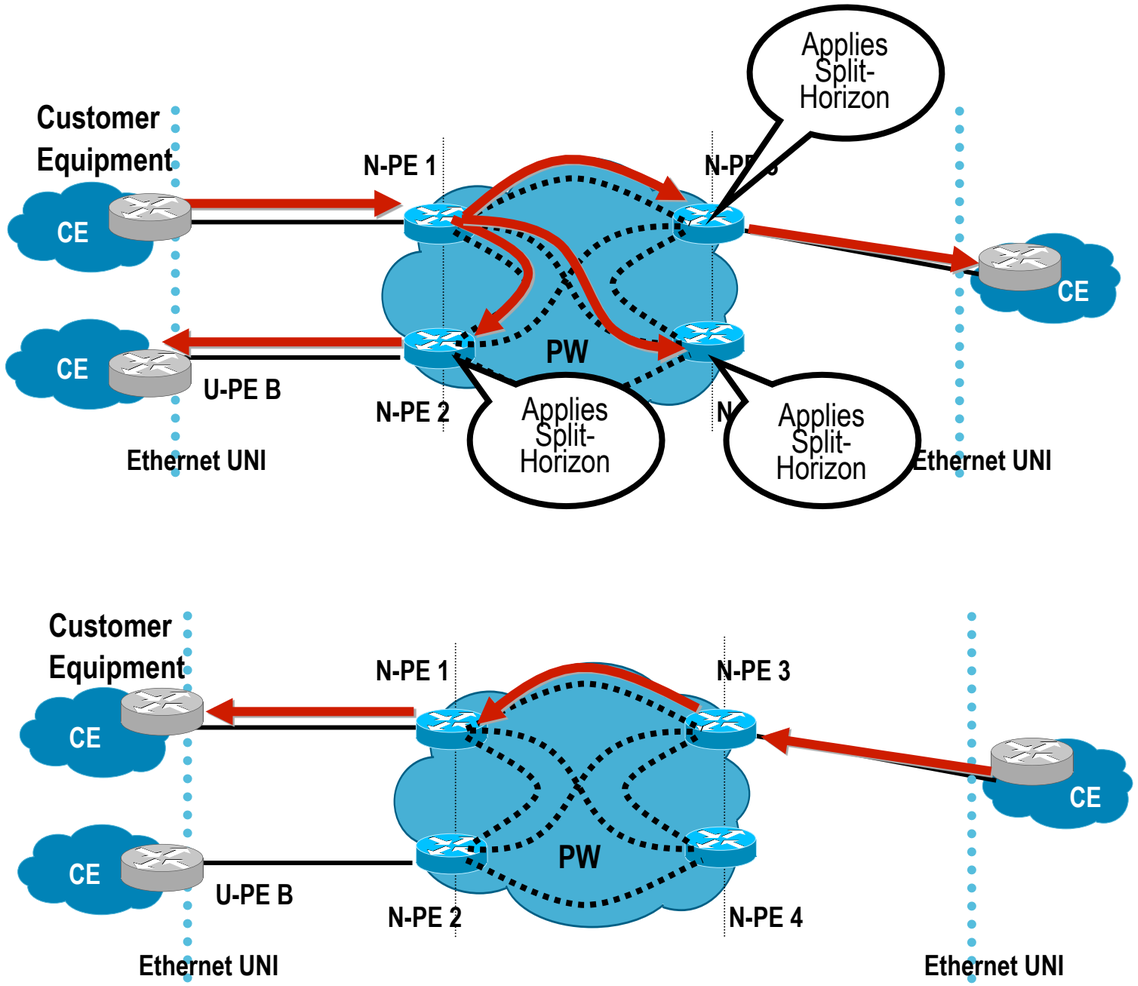
- VFI (Virtual Forwarding Instance)
  - Also called VSI (Virtual Switching Instance)
  - Emulates L2 broadcast domain among ACs and VCs
  - Unique per service. Multiple VFIs can exist same PE
- AC (Attachment Circuit)
  - Connect to CE device, it could be Ethernet physical or logical port
  - One or multiple ACs can belong to same VFI
- VC (Virtual Circuit)
  - EoMPLS data encapsulation, tunnel label used to reach remote PE, VC label used to identify VFI
  - One or multiple VCs can belong to same VFI
  - PEs must have a full-mesh of PWs in the VPLS core



Virtual Private LAN Service

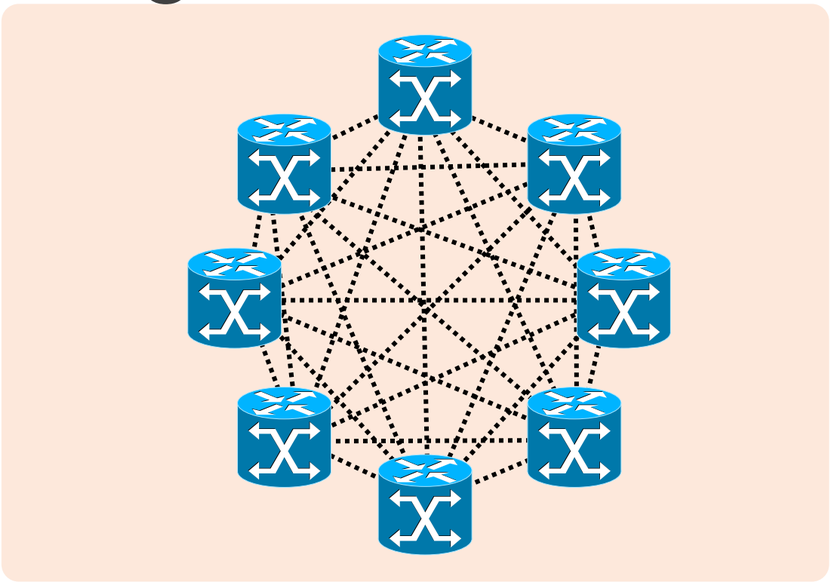
Operation

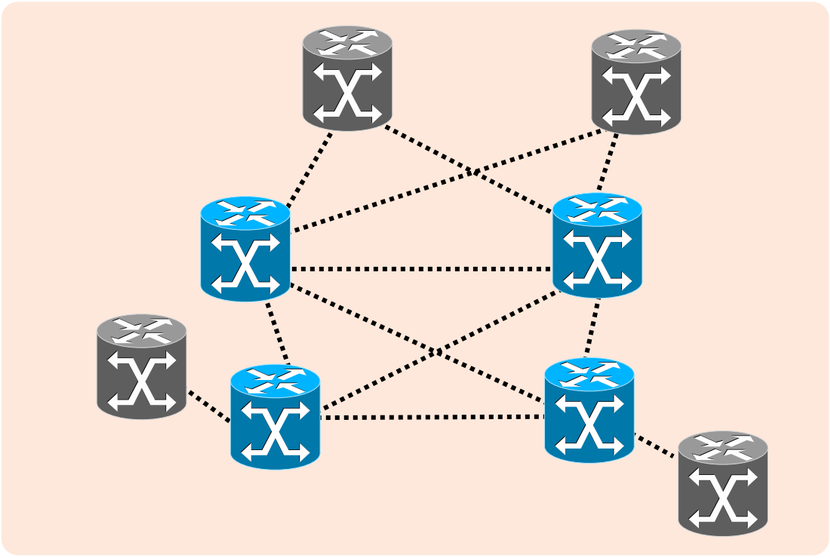
- Flooding / Forwarding
  - Forwarding based on destination MAC addresses
  - Flooding (Broadcast, Multicast, Unknown Unicast)
- MAC Learning/Aging/Withdrawal
  - Dynamic learning based on Source MAC and VLAN
  - Refresh aging timers with incoming packet
  - MAC withdrawal upon topology changes
- Split-Horizon and Full-Mesh of PWs for loop-avoidance in core
  - SP does not run STP in the core



Why H-VPLS? Improved Scaling

- Flat VPLS
  - Potential signaling overhead
  - Packet replication at the edge
  - Full PW mesh end-end
- Hierarchical-VPLS
  - Minimizes signaling overhead
  - Packet replication at the core only
  - Full PW mesh in the core

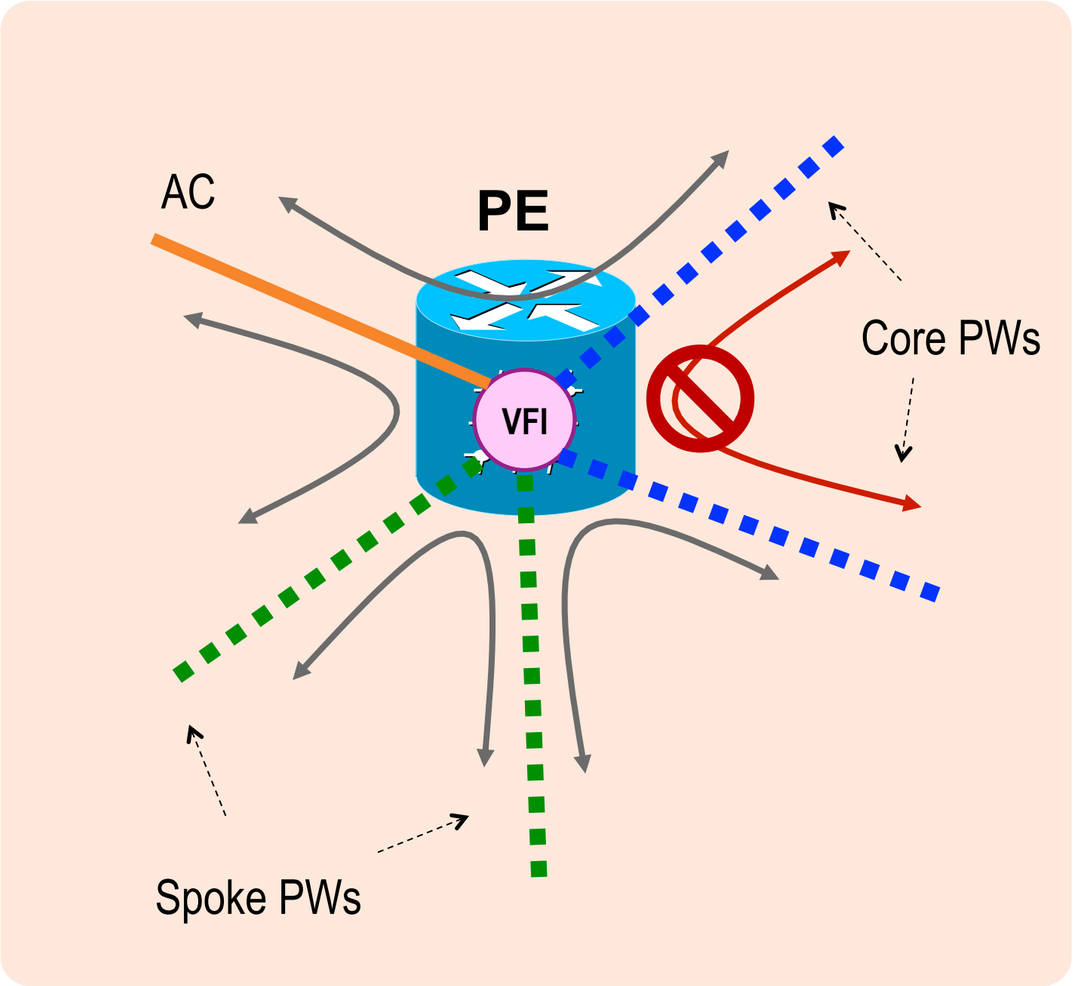




VPLS Operation

Loop Prevention

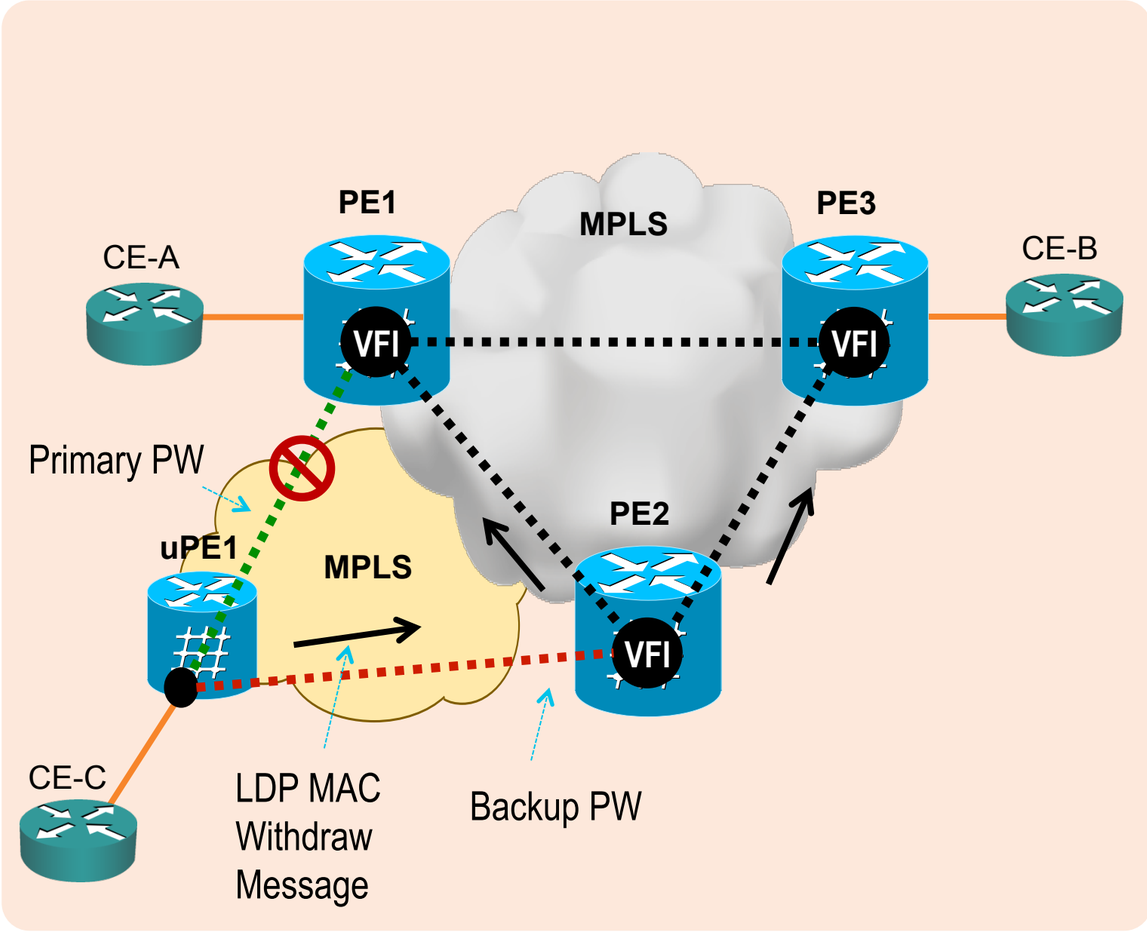
- Core PW – Split Horizon ON
- Spoke PW – Split Horizon OFF (default)
- Split-Horizon Rules
  - Forwarding between Spoke PWs
  - Forwarding between Spoke and Core PWs
  - Forwarding between ACs and Core / Spoke PWs
  - Forwarding between ACs
  - Blocking between Core PWs



VPLS Operation

MAC Address Withdrawal

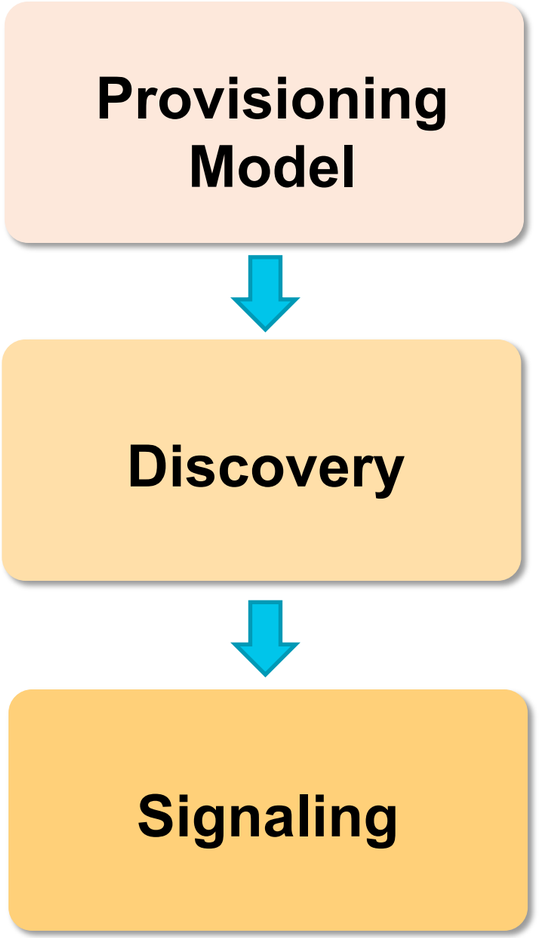
- Remove (flush) dynamic MAC addresses upon Topology Changes
  - Faster convergence – avoids blackholing
  - Uses LDP Address Withdraw Message (RFC 4762)
- H-VPLS dual-home example
  - U-PE detects failure of Primary PW
  - U-PE activates Backup PW
  - U-PE sends LDP MAC address withdrawal request to new N-PE
  - N-PE forwards the message to all PWs in the VPLS core and flush its MAC address table



VPWS / VPLS

An abstraction

- Provisioning Model
  - What information needs to be configured and in what entities
  - Semantic structure of the endpoint identifiers (e.g. VC ID, VPN ID)
- Discovery
  - Provisioning information is distributed by a "discovery process“
  - Distribution of endpoint identifiers
- Signaling
  - When the discovery process is complete, a signaling protocol is automatically invoked to set up pseudowires (PWs)



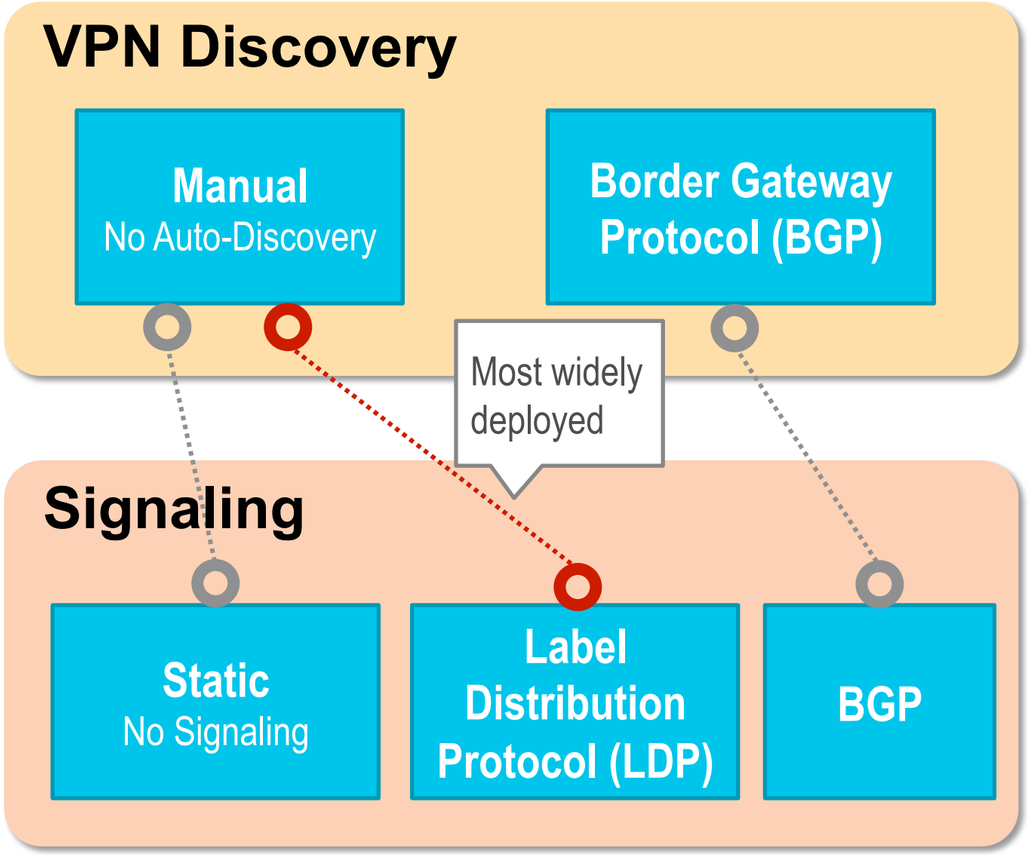
VPWS

Discovery and Signaling Alternatives

- VPWS Signaling
  - LDP-based (RFC 4447)
  - BGP-based

    draft-kompella-l2vpn-l2vpn – expired 2012

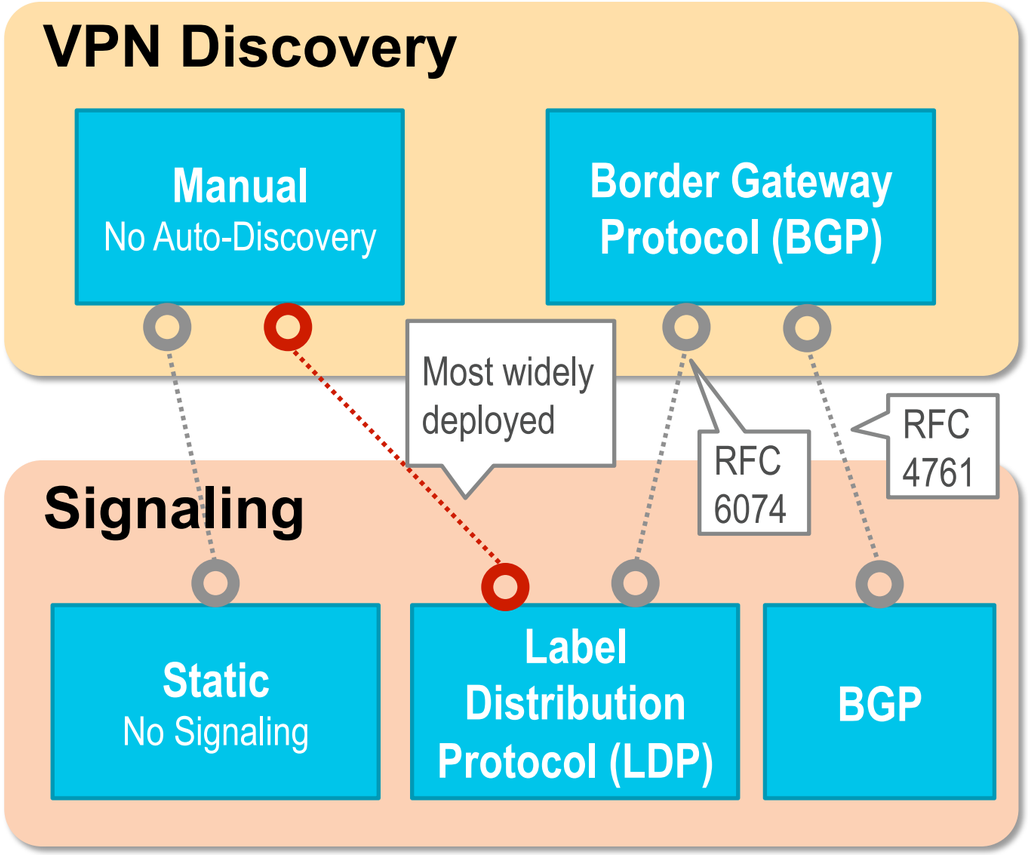
    RFC6624
- VPWS with LDP-signaling and No auto-discovery
  - Most widely deployed solution
- Auto-discovery for point-to-point services not as relevant as for multipoint



VPLS

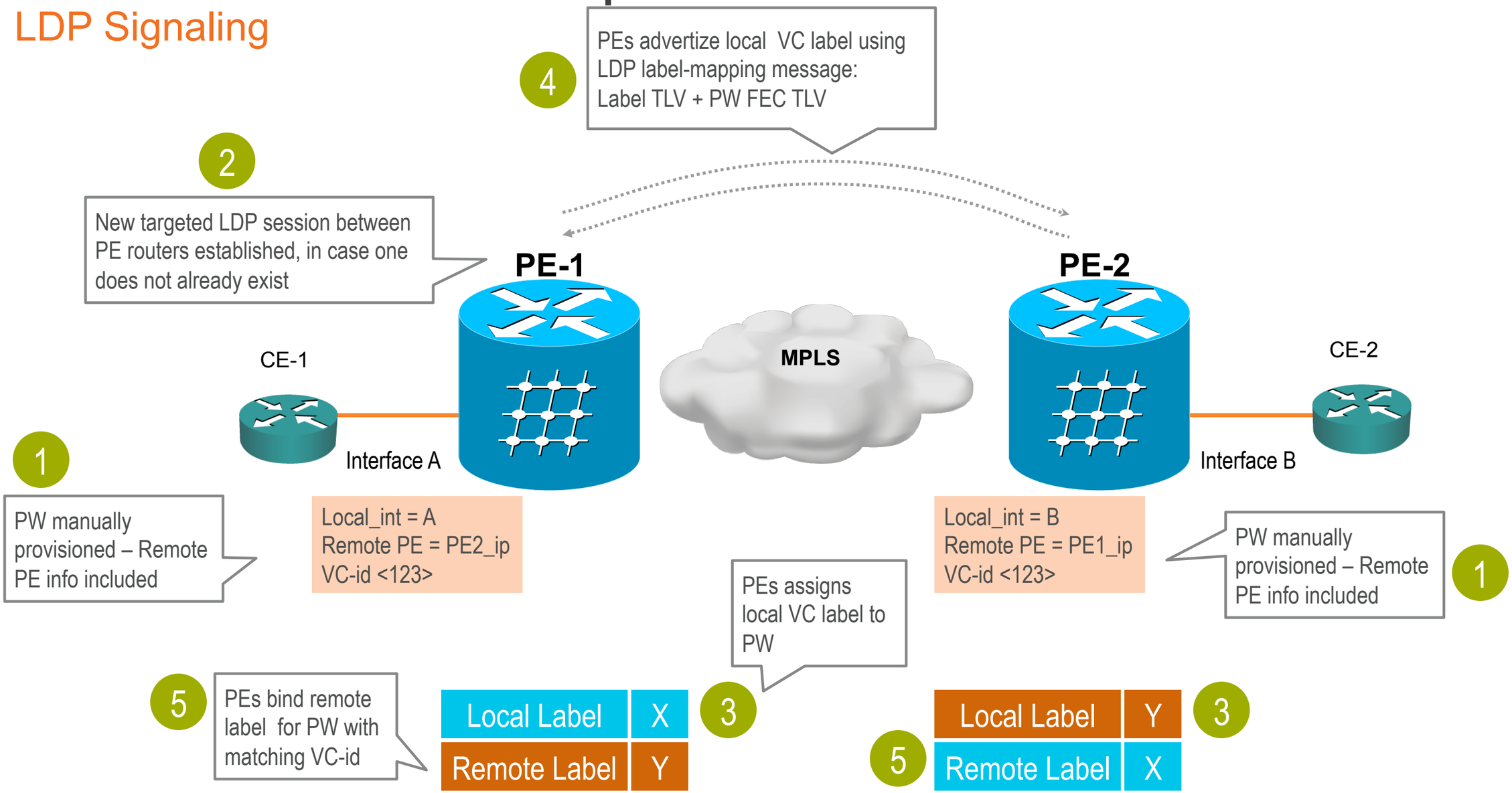
Discovery and Signaling Alternatives

- VPLS Signaling
  - LDP-based (RFC 4762)
  - BGP-based (RFC 4761)
- VPLS with LDP-signaling and No auto-discovery
  - Most widely deployed solution
  - Operational complexity for larger deployments
- BGP-based Auto-Discovery (BGP-AD) (RFC 6074)
  - Enables discovery of PE devices in a VPLS instance



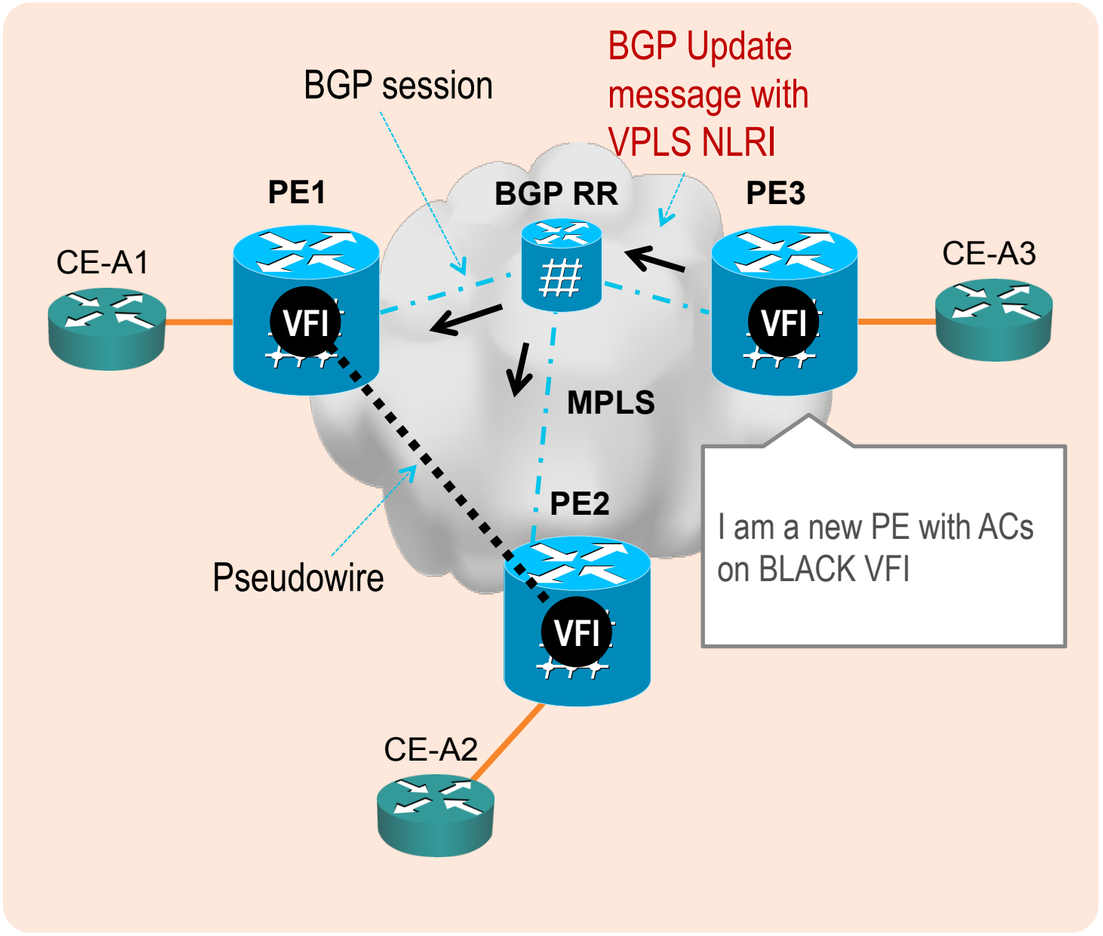
PW Control Plane Operation

LDP Signaling



BGP Auto-Discovery (BGP-AD)

- Eliminates need to manually provision VPLS neighbors
- Automatically detects when new PEs are added / removed from the VPLS domain
- Uses BGP Update messages to advertize PE/VFI mapping (VPLS NLRI)
- Typically used in conjunction with BGP Route Reflectors to minimize iBGP full-mesh peering requirements
- Two (2) RFCs define use of BGP for VPLS AD[1]
  - RFC 6074 – when LDP used for PW signaling
  - RFC 4761 – when BGP used for PW signaling



(1) VPLS BGP NLRIs from RFC 6074 and 4761 are different in format and thus not compatible, even though they share same AFI / SAFI values

Pseudowire Connectivity Verification

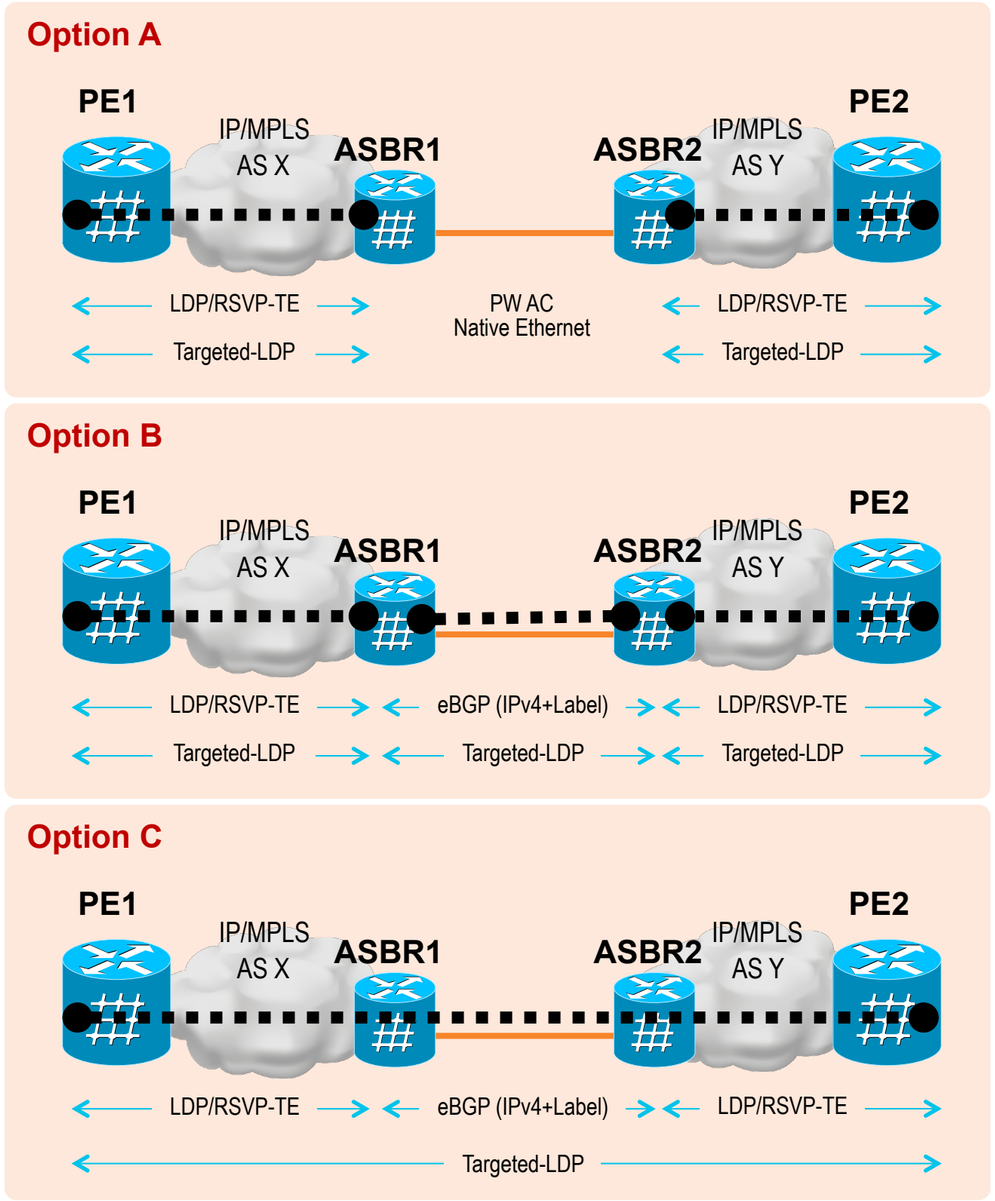
Cisco IOS

```
7604-2#ping mpls pseudowire 104.104.104.104 111000 ?
  destination  Destination address or address range
  exp          EXP bits in mpls header
  interval     Send interval between requests in msec
  pad          Pad TLV pattern
  repeat       Repeat count
  reply        Reply mode
  revision     Echo Packet TLV versioning
  segment      Segment of the MS-PW
  size         Packet size
  source       Source specified as an IP address
  sweep        Sweep range of sizes
  timeout      Timeout in seconds
  verbose      verbose output mode
```

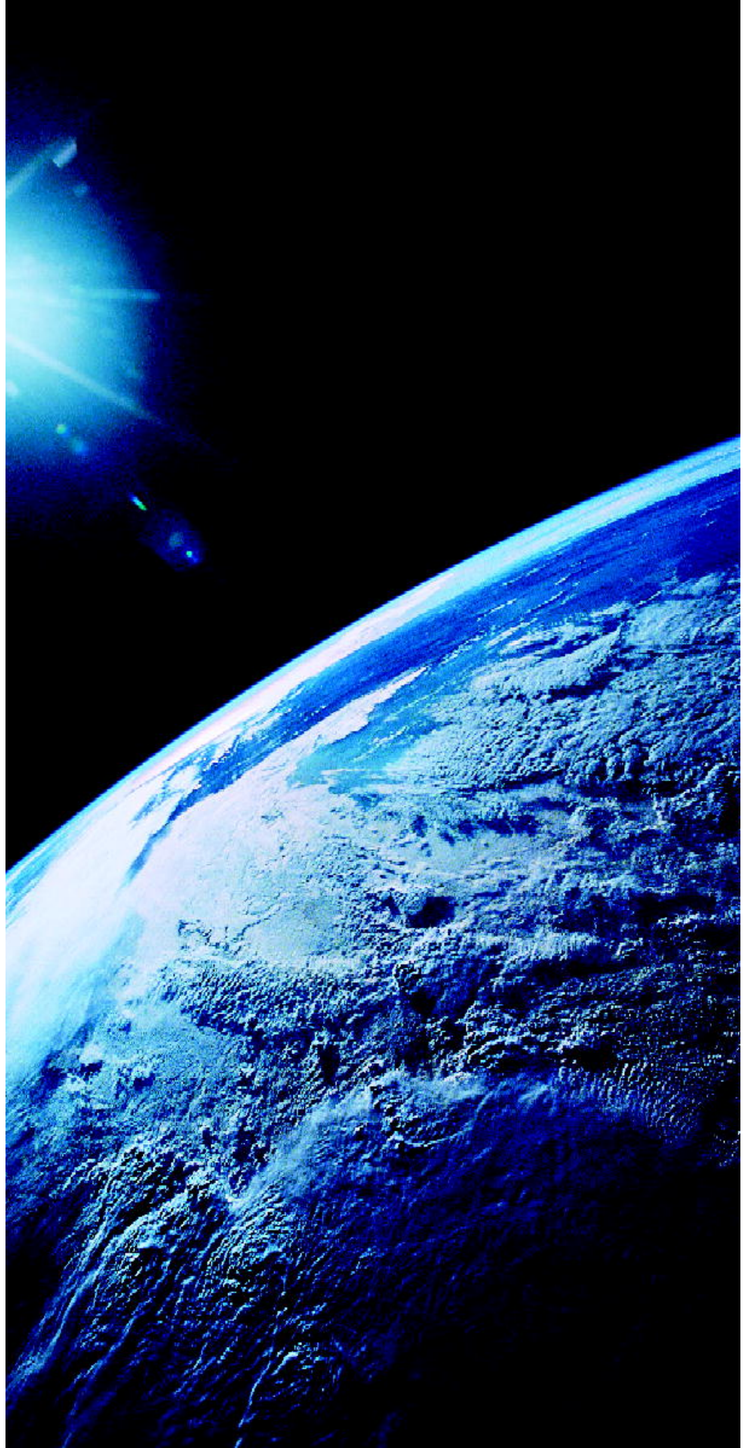
```
7604-2#ping mpls pseudowire 104.104.104.104 111000
Sending 5, 100-byte MPLS Echos to 104.104.104.104,
     timeout is 2 seconds, send interval is 0 msec:
Codes: '!' - success, 'Q' - request not sent, '.' - timeout,
  'L' - labeled output interface, 'B' - unlabeled output interface,
  'D' - DS Map mismatch, 'F' - no FEC mapping, 'f' - FEC mismatch,
  'M' - malformed request, 'm' - unsupported tlvs, 'N' - no label entry,
  'P' - no rx intf label prot, 'p' - premature termination of LSP,
  'R' - transit router, 'I' - unknown upstream index,
  'l' - Label switched with FEC change, 'd' - see DDMAP for return code,
  'X' - unknown return code, 'x' - return code 0
Type escape sequence to abort.
!!!!!
Success rate is 100 percent (5/5), round-trip min/avg/max = 1/1/1 ms
```

L2VPN Inter-AS

- Three (3) deployment models
- Option A
  - No reachability information shared between AS
- Option B
  - Minimal reachability information shared between AS
  - ASBR configured as S-PEs (multi-segment PWs)
  - eBGP (IPv4 prefix + label) used to build PSN tunnel between AS
- Option C
  - Significant reachability information shared between AS
  - Single-segment PW signaled across AS boundary



Next Generation L2VPN Solutions



Overview

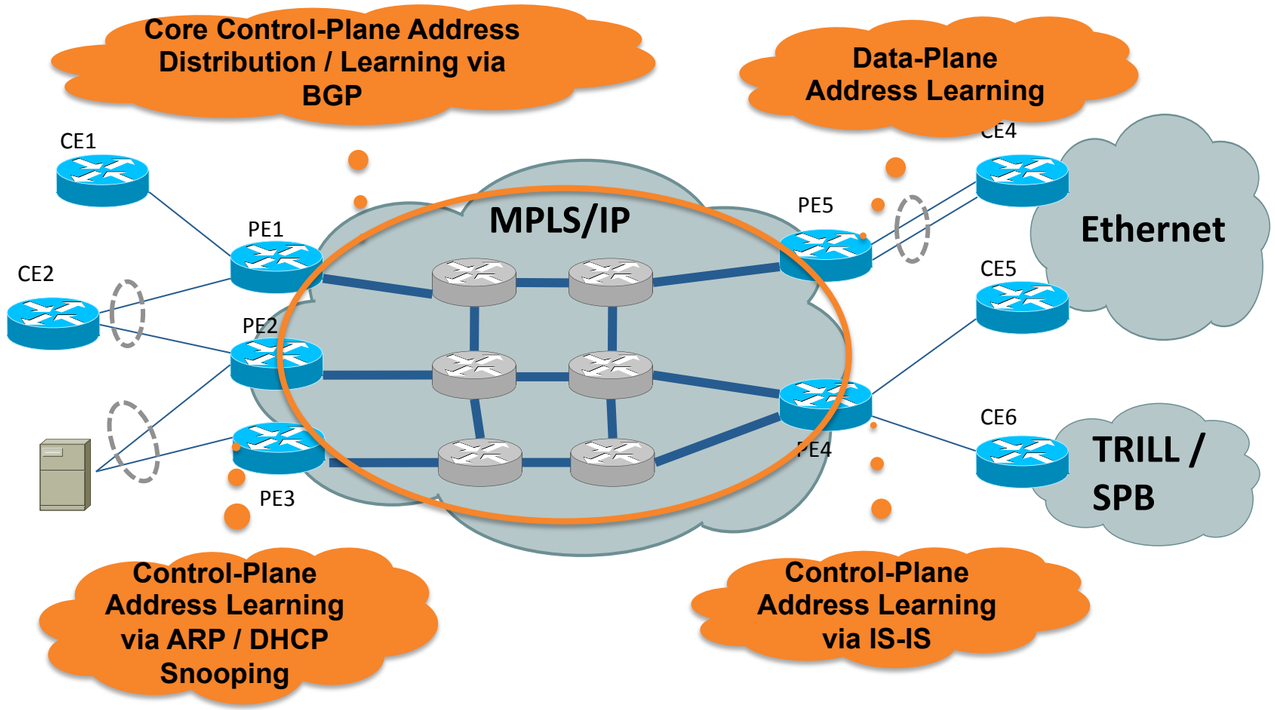
PEs run Multi-Protocol BGP to advertise & learn L2 address information over Core.

Learning on PE Access Circuits via:

- Data-plane transparent learning, or
- Control-plane (DHCP, ARP, IS-IS)

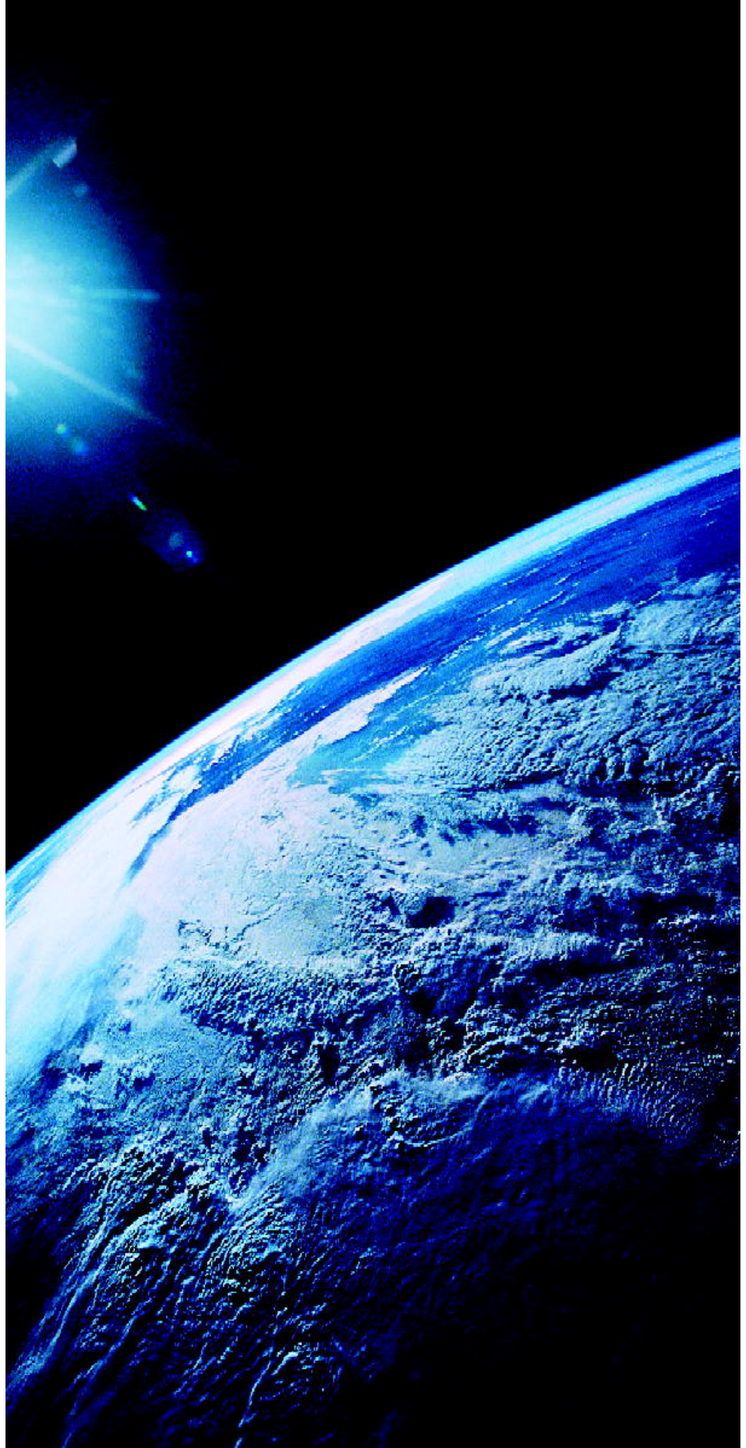
No pseudowires

- Unicast: use MP2P tunnels
- or use LSM



Next Generation L2VPN Solutions

E-VPN/PBB-EVPN



Requirements for Next Generation L2VPNs

- Multi-homing / All-active Redundancy
  - Flow Based Load Balancing on PEs
  - Flow Based Multi-pathing in PSN
  - Geo-redundancy and Flexible Redundancy Grouping
- Simplified Provisioning and Operation
  - Core Auto-Discovery
  - Access Multi-homing Auto-Discovery
  - New Service Interfaces
- Optimal Multicast with LSM
  - P2MP Trees
  - MP2MP Trees
- Fast Convergence
  - Link/Port/Node Failure
  - MAC Mobility
- Support Flexible Forwarding Policies and Topologies

VPLS

E-VPN

- Scalable for SP Virtual Private Cloud services:
  - Support O(1Million) MAC Addresses per DC
  - Confinement of C-MAC Learning
- Support C-MAC (VM) Mobility with MAC Summarization
- Seamless interworking between TRILL / 802.1Qaq / 802.1Qbp and Legacy DC
  - Guarantee C-MAC Transparency on WAN Edge PE.
- Fast Convergence
  - Avoiding C-MAC address Flushing

PBB-EVPN

Optimal Forwarding

Optimal forwarding for unicast and multicast.

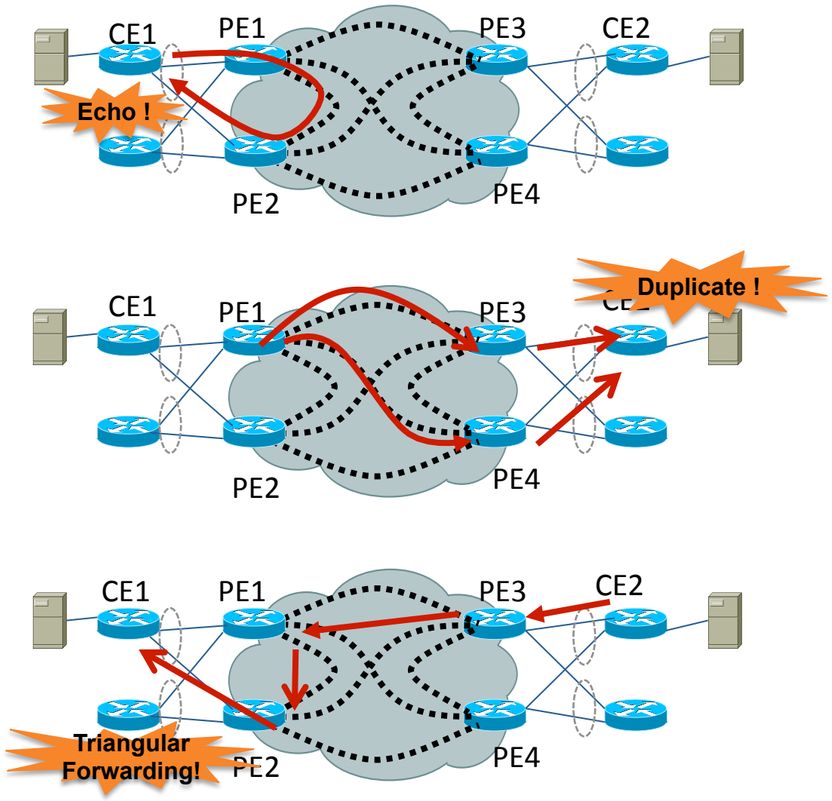
Shortest path – no triangular forwarding at steady-state.

Loop-Free & Echo-Free Forwarding.

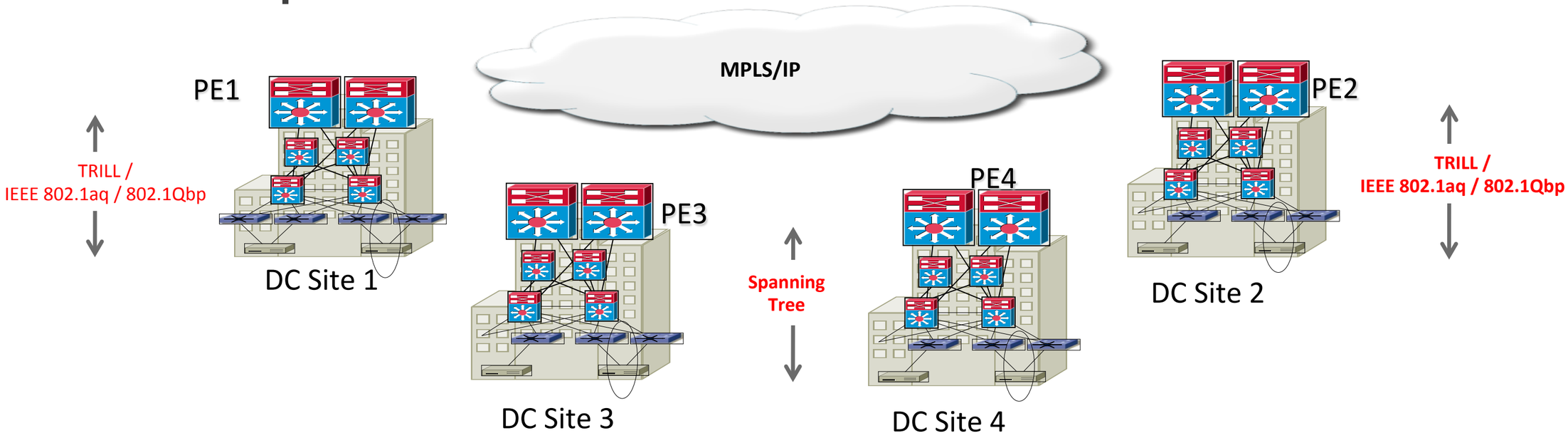
Avoid duplicate delivery of flooded traffic.

Multiple multicast tunneling options:

- Ingress Replication
- P2MP LSM tunnels
- MP2MP



Seamless Interworking with TRILL, SPB & IEEE 802.1Qbp



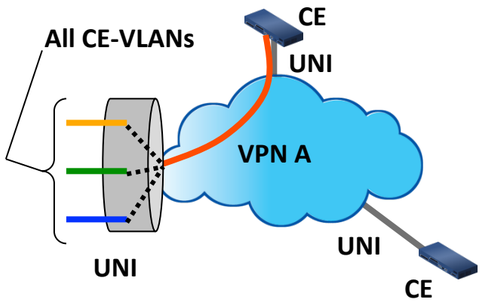
- Support seamless inter-working of classical Ethernet & next-generation data center solutions (i.e. TRILL, IEEE 802.1Qaq or 802.1Qbp).
- Requirements:
  - Control Plane separation between DC sites
  - C-MAC address transparency on the PEs
  - Resilient DC to PE connectivity

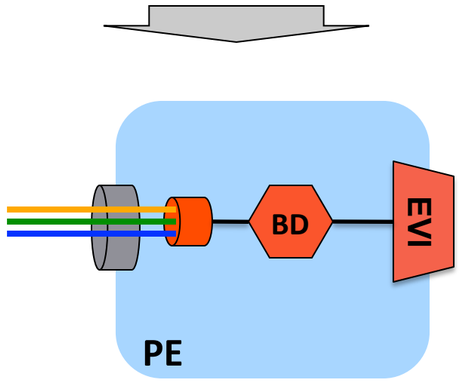
E-VPN Instance (EVI) & Service Interfaces

- E-VPN Instance (EVI) identifies a VPN in the MPLS/IP network.
- EVI may encompass one or more bridge-domains, depending on PE's service interface type:

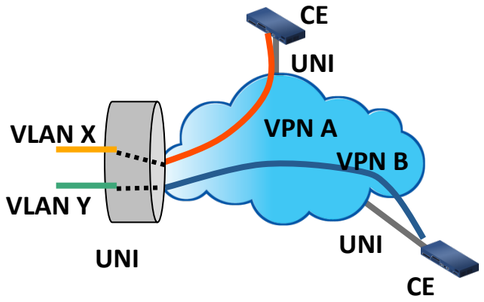
New!

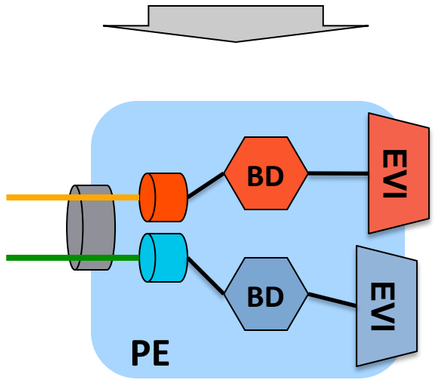
Port Based Service Interface



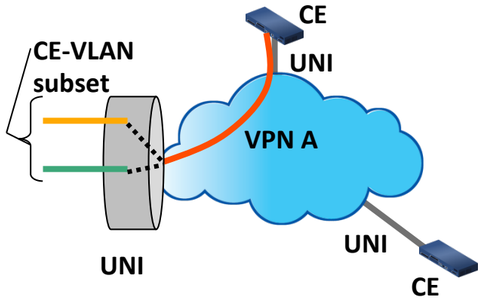


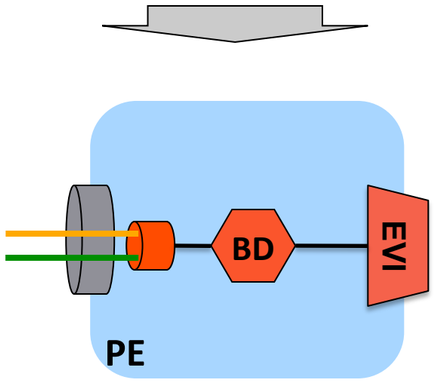
VLAN Based Service Interface



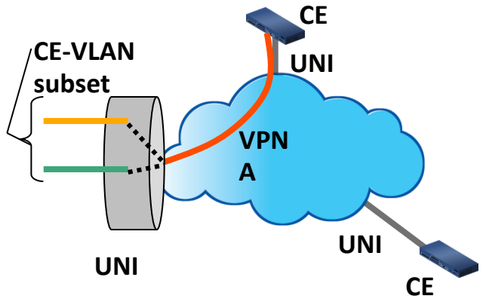


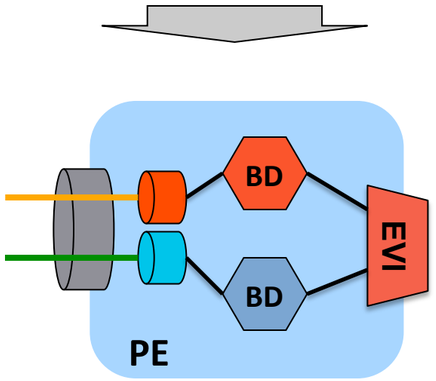
VLAN Bundling Service Interface





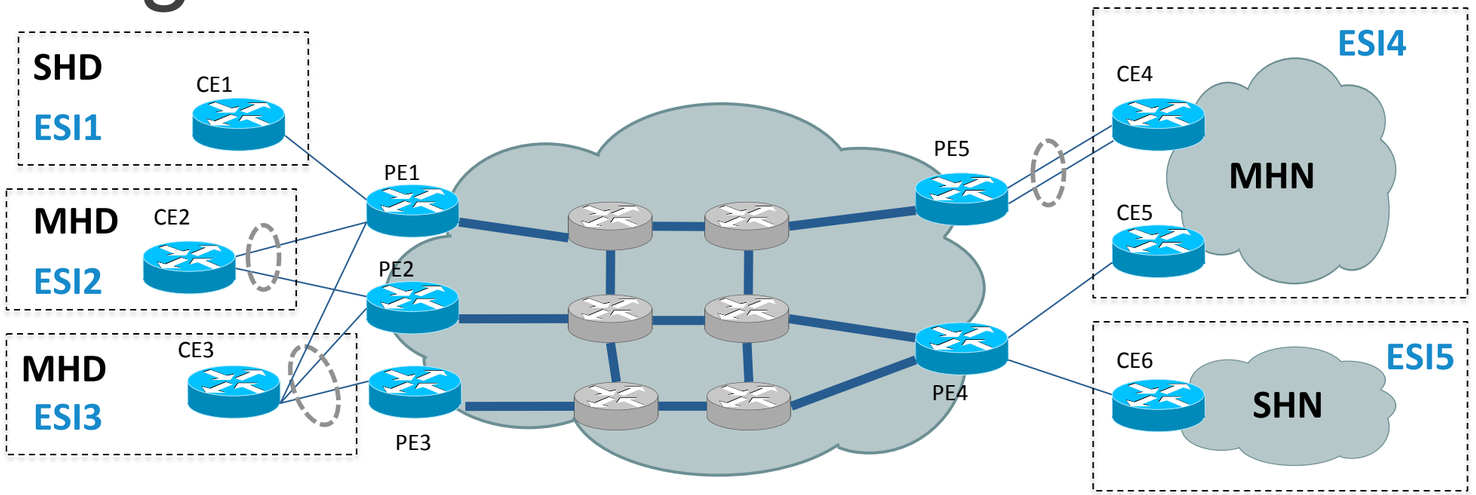
VLAN Aware Bundling Service Interface





Ethernet Segment

Definition

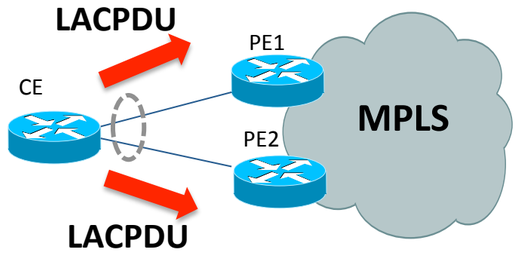


- Ethernet Segment is a ‘site’ connected to one or more PEs.
- Ethernet Segment could be a single **device** (i.e. CE) or an entire **network**.
  - Single-Homed Device (SHD)
  - Multi-Homed Device* (MHD) using Ethernet Multi-chassis Link Aggregation Group
  - Single-Homed Network (SHN)
  - Multi-Homed Network* (MHN)
- Uniquely identified by a 10-byte global Ethernet Segment Identifier (**ESI**).

***: Includes Dual-Homed**

Ethernet Segment

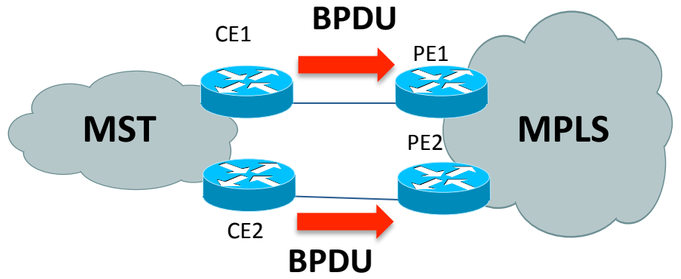
ESI Auto-Sensing



MHD with Multi-chassis LAG

- ESI is auto-discovered via LACP.
- ESI is encoded using the CE's LACP parameters:

System Priority	System MAC Address	Port Key
2 bytes	6 bytes	2 bytes



MHN with MST

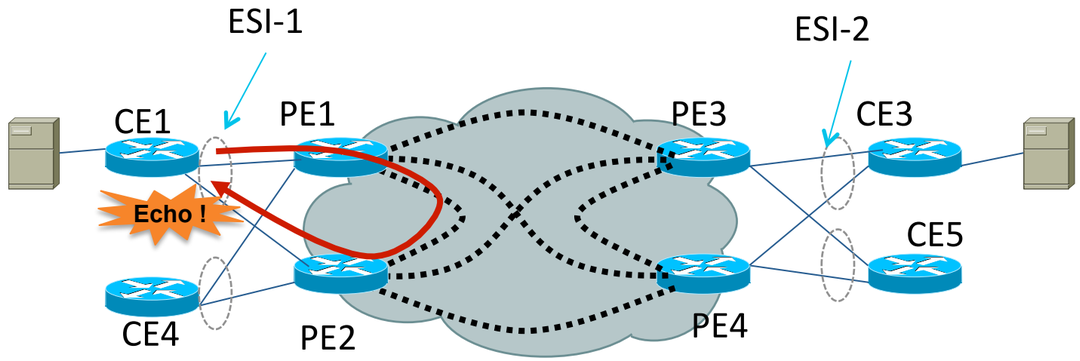
- ESI is auto-discovered via MST BPDU snooping.
- ESI is encoded using the IST's root parameters:

Bridge Priority	Root Bridge MAC	0x0000
2 bytes	6 bytes	2 bytes

Split Horizon

For Ethernet Segments – E-VPN

Challenge:
How to prevent flooded traffic from echoing back to a multi-homed Ethernet Segment?

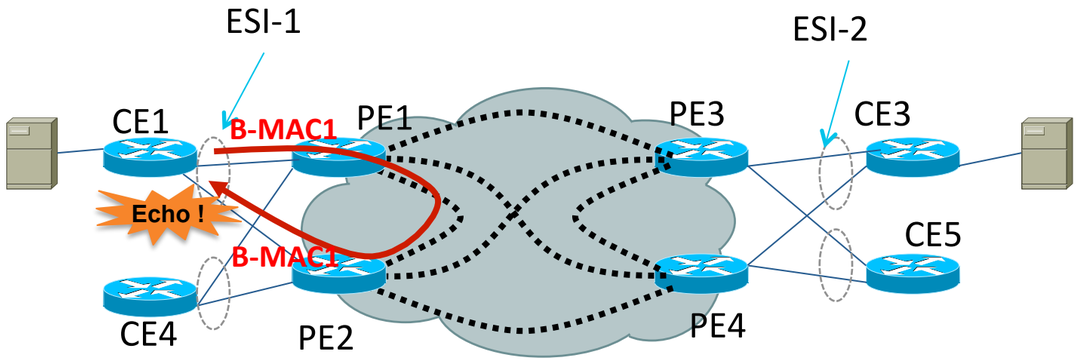


- PE advertises in BGP a split-horizon label (ESI MPLS Label) associated with each multi-homed Ethernet Segment.
- Split-horizon label is only used for multi-destination frames (Unknown Unicast, Multicast & Broadcast).
- When an ingress PE floods multi-destination traffic, it encodes the Split-Horizon label identifying the source Ethernet Segment in the packet.
- Egress PEs use this label to perform selective split-horizon filtering over the attachment circuit.

Split Horizon

For Ethernet Segments – PBB-EVPN

Challenge:
How to prevent flooded traffic from echoing back to a multi-homed Ethernet Segment?

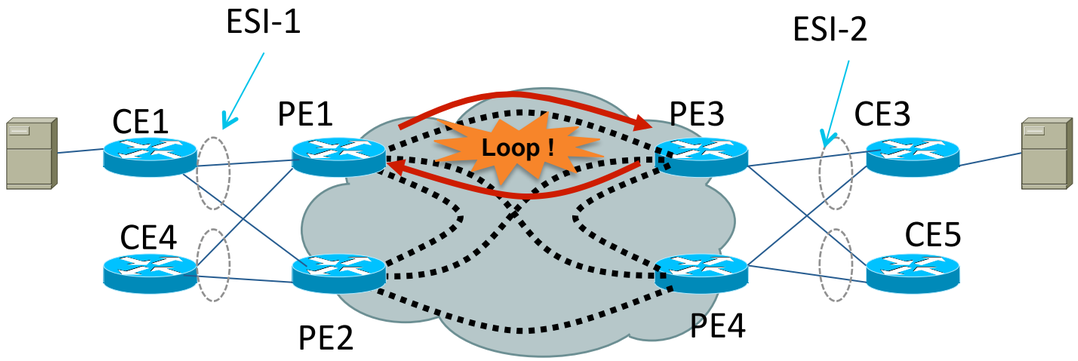


- PEs connected to the same MHD use the same B-MAC address for the Ethernet Segment
  - 1:1 mapping between B-MAC and ESI (for All-Active Redundancy with flow-based LB)
- Disposition PEs check the B-MAC source address for Split-Horizon filtering
  - Frame not allowed to egress on an Ethernet Segment whose B-MAC matches the B-MAC source address in the PBB header.

Split Horizon

For Core Tunnels

Challenge:
How to prevent flooded traffic from looping back over the core?

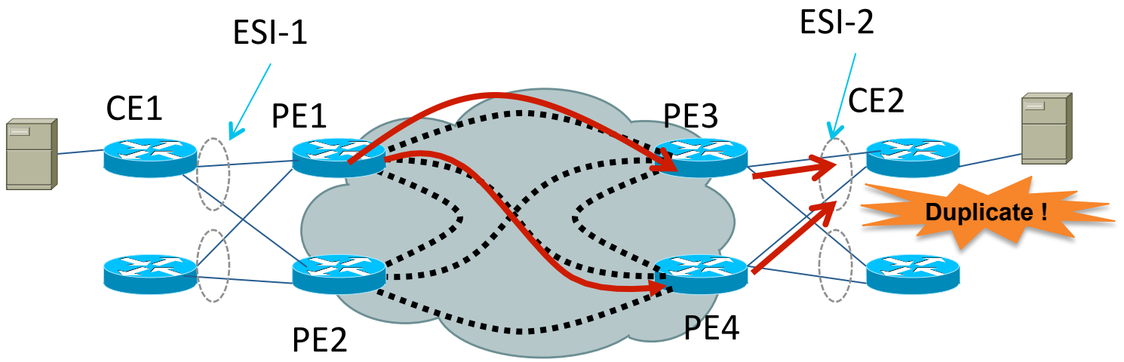


- Traffic received from an MPLS tunnel over the core is never forwarded back to the MPLS core.
- This is similar to the VPLS split-horizon filtering rule.

Designated Forwarder (DF)

DF Election

Challenge:
How to prevent duplicate copies of flooded traffic from being delivered to a multi-homed Ethernet Segment?

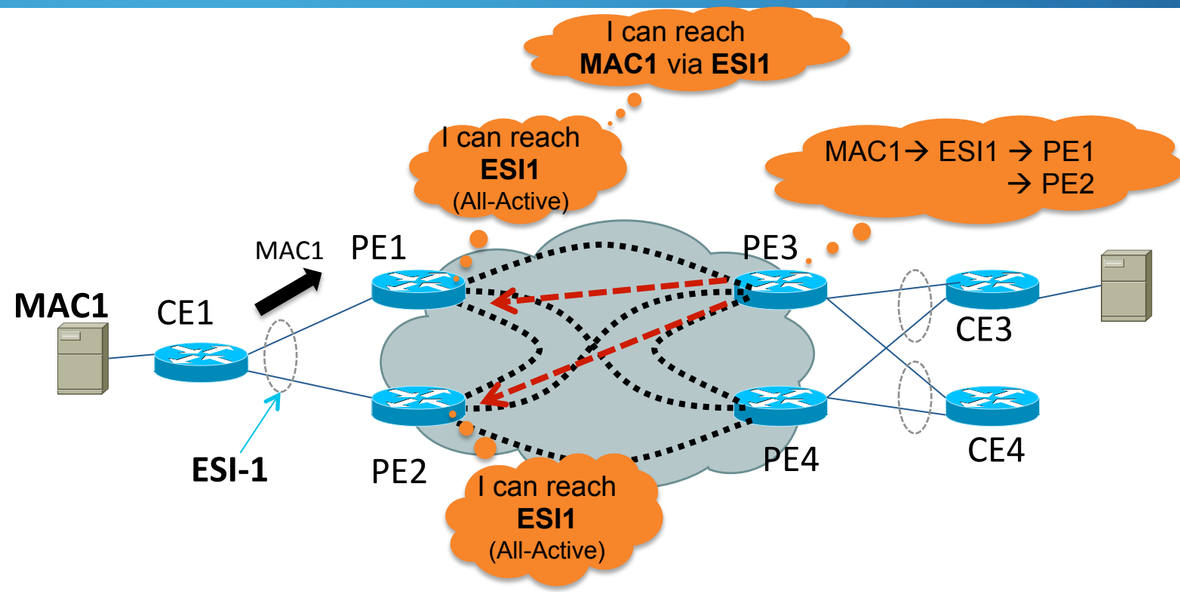


- PEs connected to a multi-homed Ethernet Segment discover each other via BGP.
- These PEs then elect among them a Designated Forwarder responsible for forwarding flooded multi-destination frames to the multi-homed Segment.
- DF Election granularity can be:
  - Per Ethernet Segment (Single PE is the DF)
  - Per EVI (E-VPN) or I-SID (PBB-EVPN) on Ethernet Segment (Multiple DFs for load-balancing)

Aliasing

E-VPN

Challenge:
How to load-balance traffic towards a multi-homed device across multiple PEs when MAC addresses are learnt by only a single PE?

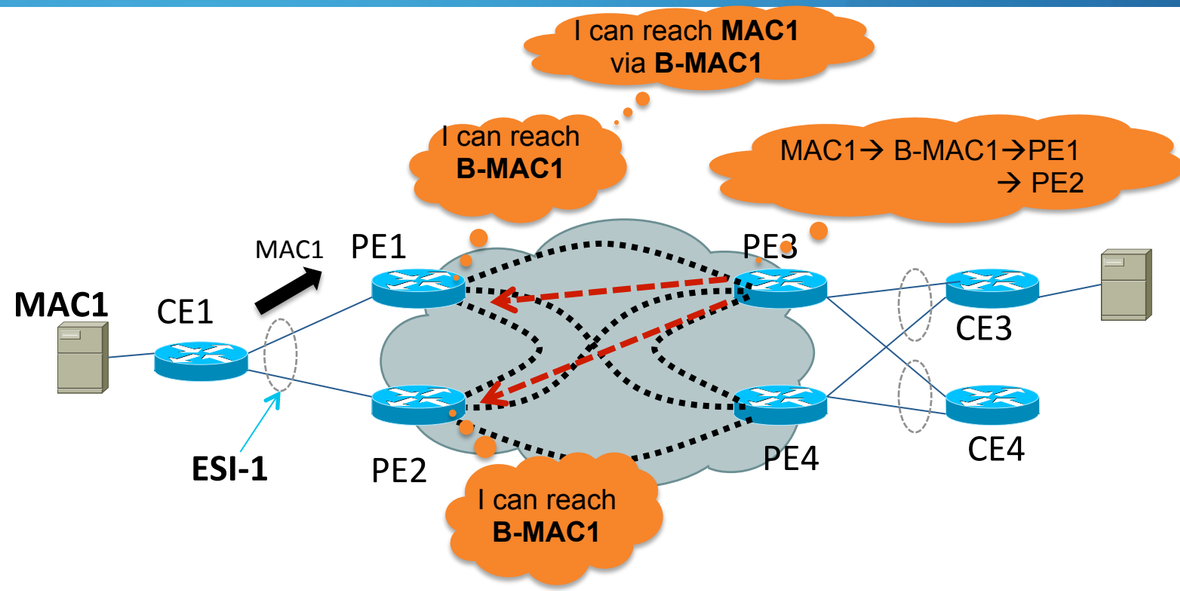


- PEs advertise in BGP the ESIs of local multi-homed Ethernet Segments.
  - All-Active Redundancy Mode indicated
- When PE learns MAC address on its AC, it advertises the MAC in BGP along with the ESI of the Ethernet Segment from which the MAC was learnt.
- Remote PEs can load-balance traffic to a given MAC address across all PEs advertising the same ESI.

Aliasing

PBB-EVPN

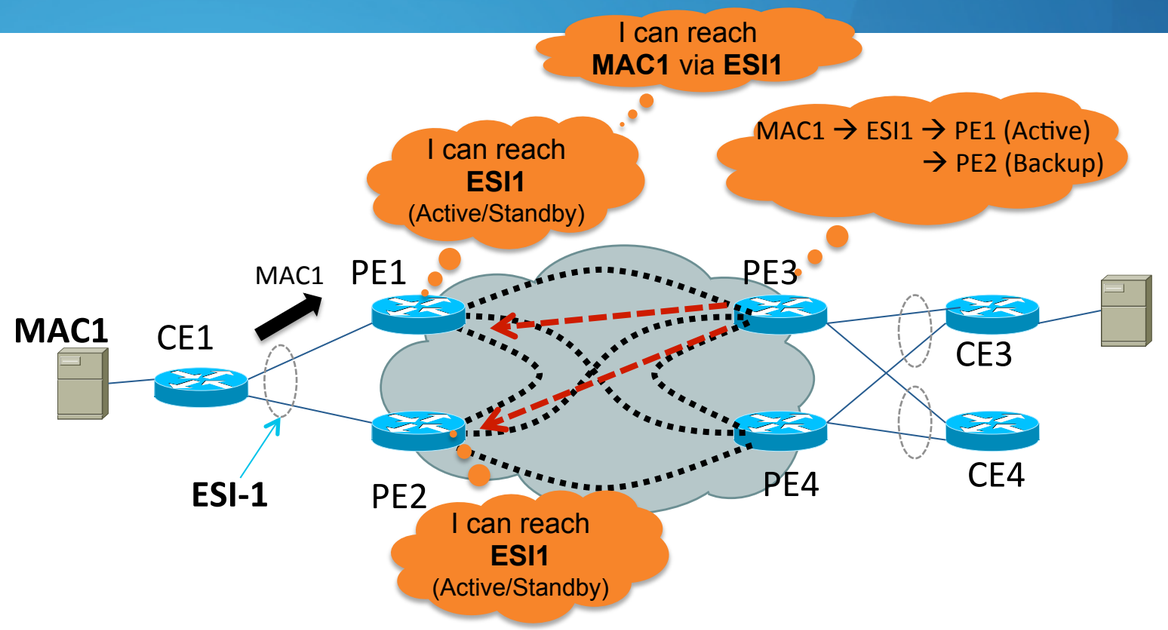
Challenge:
How to load-balance traffic towards a multi-homed device across multiple PEs when MAC addresses are learnt by only a single PE?



- PEs connected to the same MHD use the same B-MAC address for the Ethernet Segment
  - 1:1 mapping between B-MAC and ESI (for All-Active Redundancy with flow-based LB)
- PEs advertise their B-MAC addresses independent of the C-MAC learning state.
- Remote PEs can load-balance traffic to a given C-MAC across all PEs advertising the same associated B-MAC.

Backup Path

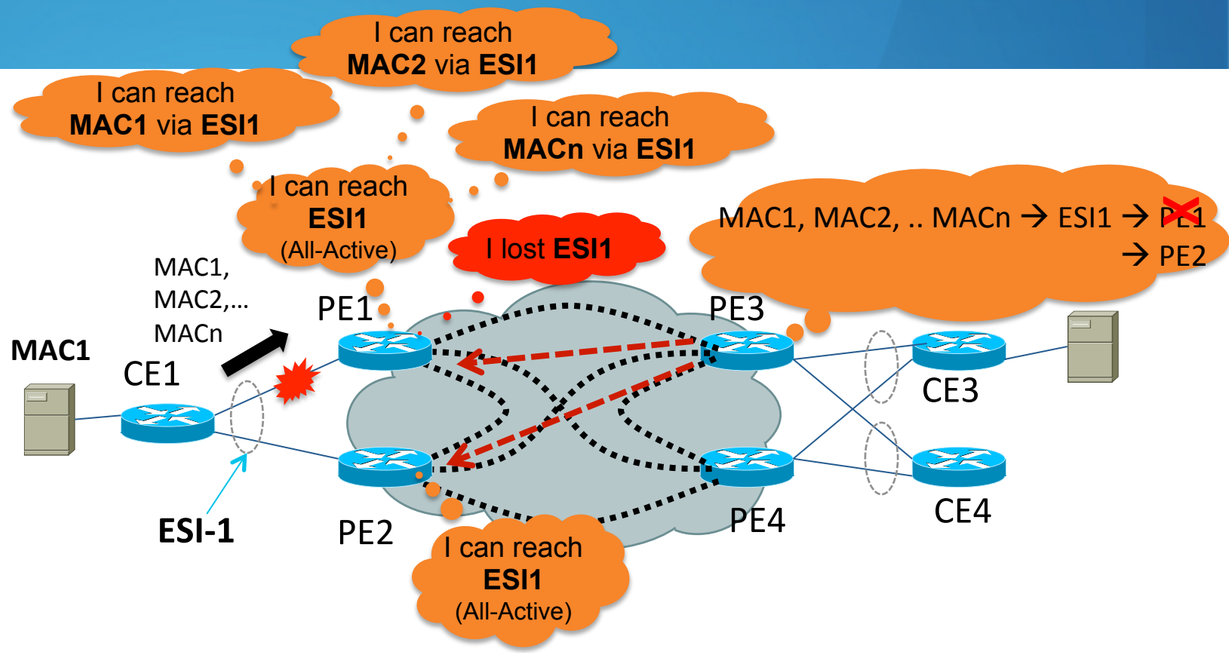
Challenge:
How to identify PEs that have a backup path to a multi-homed Ethernet Segment?



- PEs advertise in BGP connectivity to ESIs associated with local multi-homed Ethernet Segments.
  - Active/Standby Redundancy Mode is indicated
- When PE learns a MAC address on its AC, it advertises the MAC in BGP along with the ESI of the Ethernet Segment from which the MAC was learnt.
- Remote PEs will install:
  - **active path** to the PE that advertised **both MAC Address & ESI**
  - **backup path** to the PE that advertised **ESI only**

MAC Mass-Withdraw

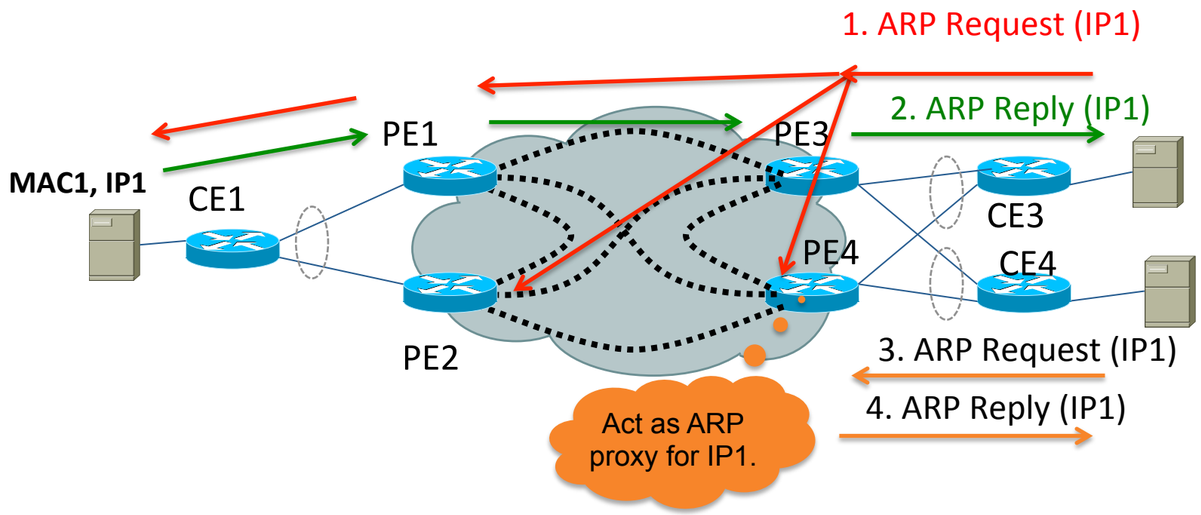
Challenge:
How to inform remote PEs of a failure affecting many MAC addresses quickly while the control-plane re-converges?



- PEs advertise two sets of information:
  - MAC addresses along with the ESI from the address was learnt
  - Connectivity to ESI(s)
- If a PE detects a failure impacting an Ethernet Segment, it withdraws the route for the associated ESI.
  - Remote PEs remove failed PE from the path-list for **all MAC addresses associated with an ESI**.
  - This effectively is a MAC ‘mass-withdraw’ function.

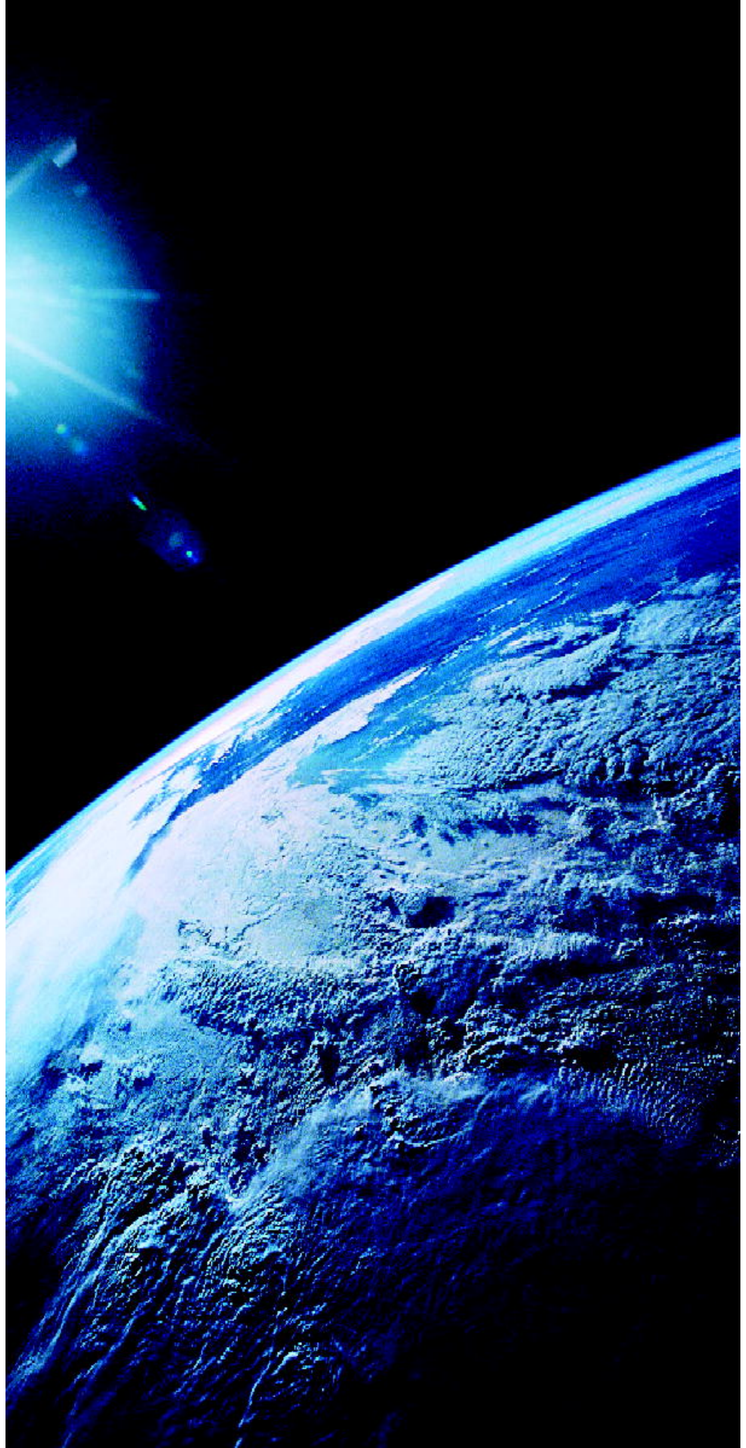
ARP Broadcast Suppression

Challenge:
How to reduce ARP broadcasts over the MPLS/IP network, especially in large scale virtualized server deployments?

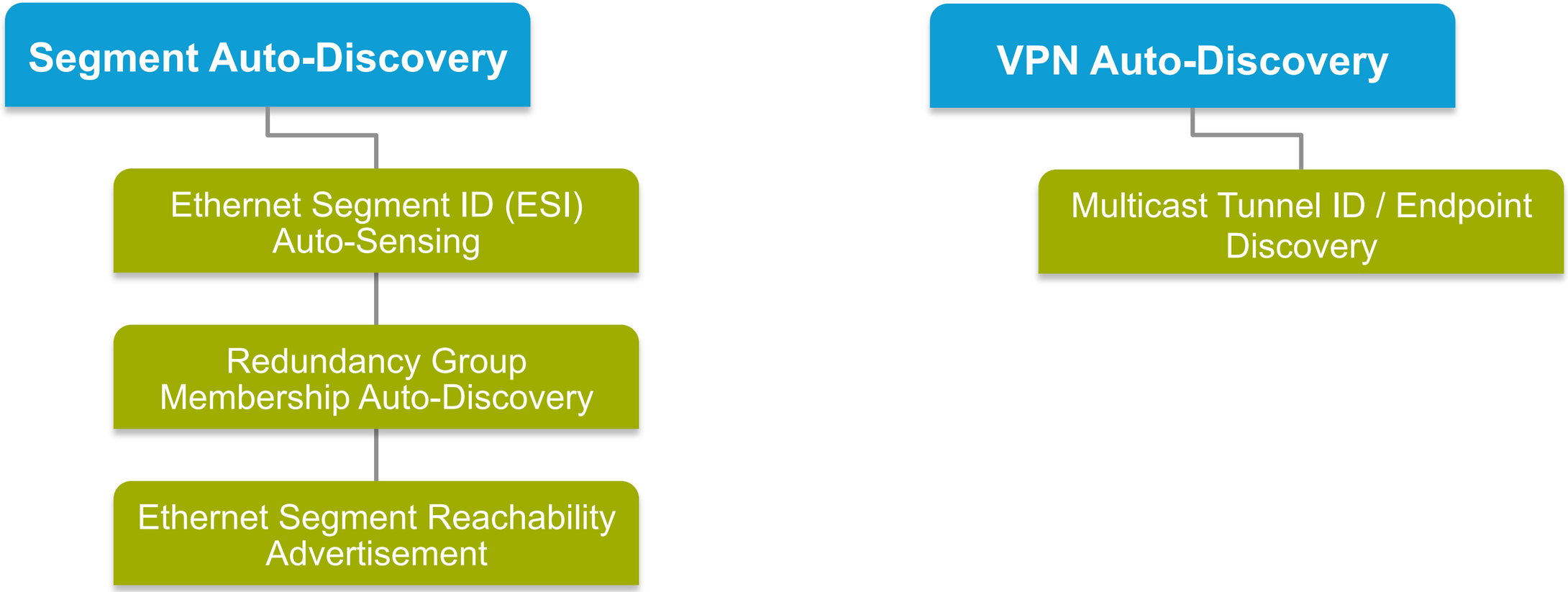


- Construct ARP caches on the E-VPN PEs and synchronize them either via BGP or data-plane snooping.
- PEs act as ARP proxies for locally attached hosts, thereby preventing repeated ARP broadcast over the MPLS/IP network.

E-VPN
Startup Sequences



E-VPN Startup Sequence



E-VPN Startup Sequence (cont.)

ESI Auto-Sensing

Segment Auto-Discovery

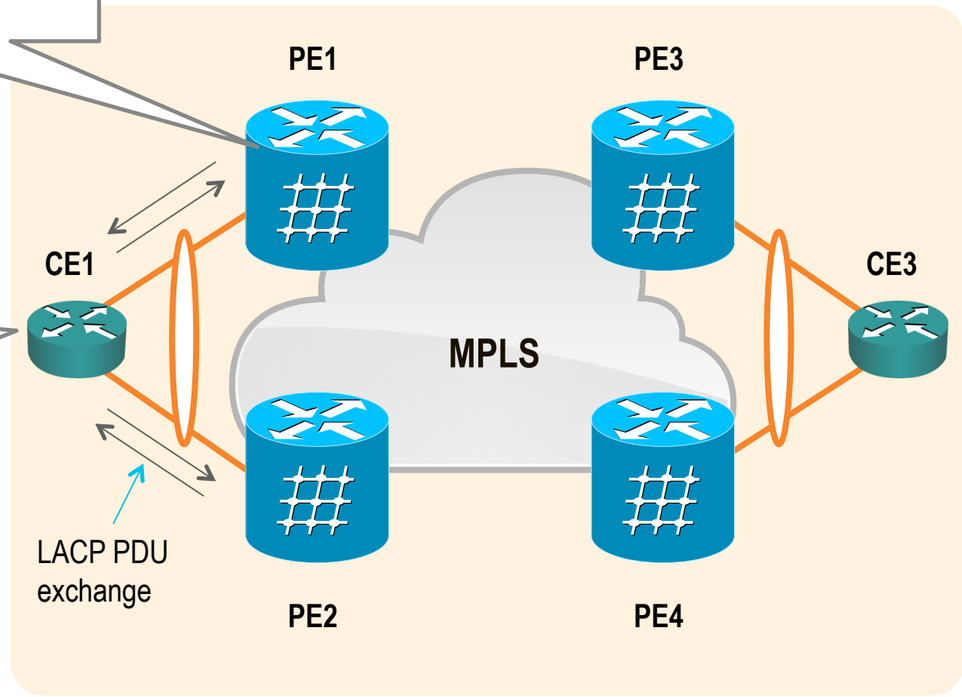
Ethernet Segment ID (ESI) Auto-Sensing

ESI (10B) can be auto-generated* from CE's LACP information → Concatenation of CE's LACP **System Priority** + **System ID** + **Port Key**

Example:
0000. 0011.0022.0033.0018

System Priority	System MAC Address	Port Key
2 bytes	6 bytes	2 bytes

CE LACP info:
LACP System Priority (2B)
e.g. 0000
LACP System ID (MAC) (6B)
e.g. 0011.0022.0033
LACP Port Key (2B)
e.g. 0018



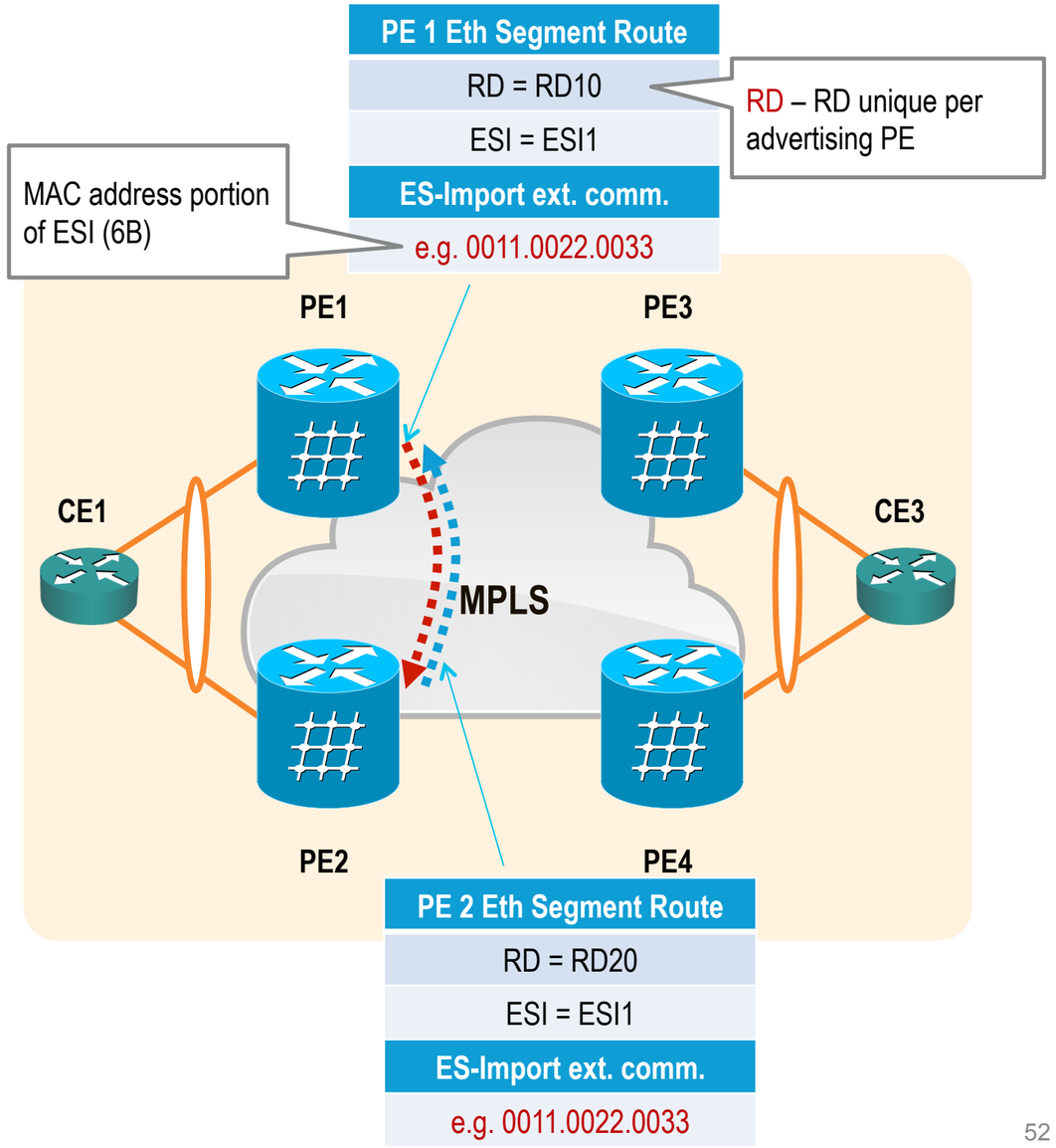
(*) ESI can also be manually configured

E-VPN Startup Sequence (cont.)

BGP Ethernet Segment Route

Segment Auto-Discovery

- Ethernet Segment ID (ESI) Auto-Sensing
- Redundancy Group Membership Auto-Discovery



E-VPN Startup Sequence (cont.)

BGP Ethernet AD Routes – Per-ESI

Segment Auto-Discovery

Ethernet Segment ID (ESI) Auto-Sensing

Redundancy Group Membership Auto-Discovery

Ethernet Segment Reachability Advertisement

Per-ESI adv.

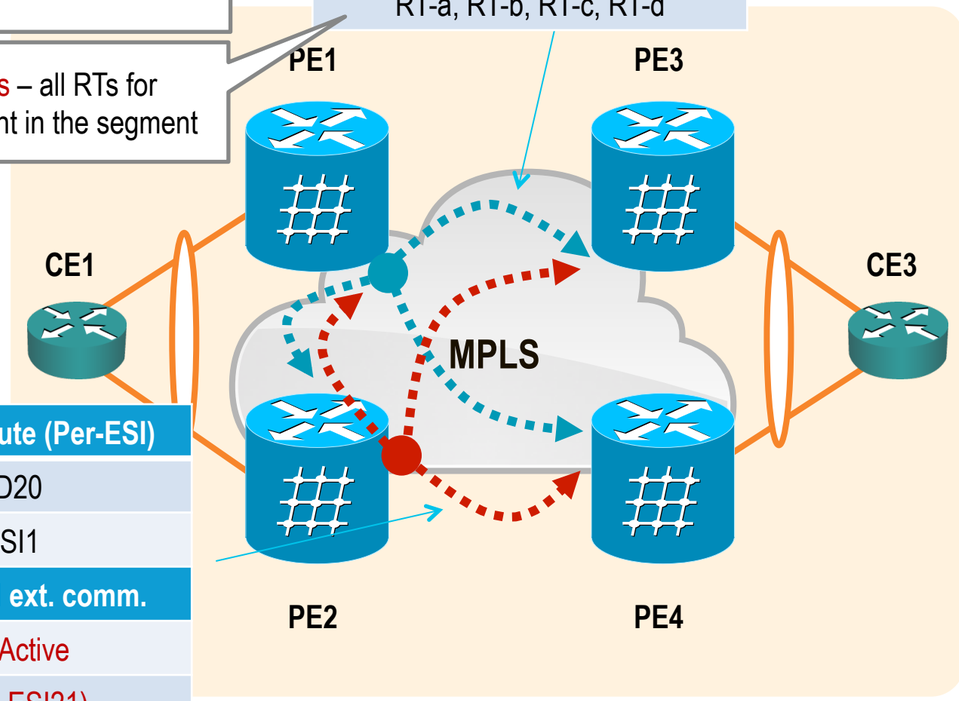
Flags– Redundancy mode - All-Active or Active-Standby

ESI MPLS Label – used by local PEs for split-horizon. **Downstream** assigned (**ingress replication**) or **Upstream** assigned (**P2MP LSP**)

Multiple RTs – all RTs for EVIs present in the segment

PE 1 Eth A-D Route (Per-ESI)
RD = RD10
ESI = ESI1
ESI MPLS Label ext. comm.
Flag (e.g. All-Active)
Label (e.g. LESI11)
RT ext. community
RT-a, RT-b, RT-c, RT-d

RD – RD unique per advertising PE



PE 2 Eth A-D Route (Per-ESI)
RD = RD20
ESI = ESI1
ESI MPLS Label ext. comm.
Flag = All-Active
Label (e.g. LESI21)
RT ext. community
RT-a, RT-b, RT-c, RT-d

E-VPN Startup Sequence (cont.)

BGP Ethernet AD Routes – Per-EVI

Segment Auto-Discovery

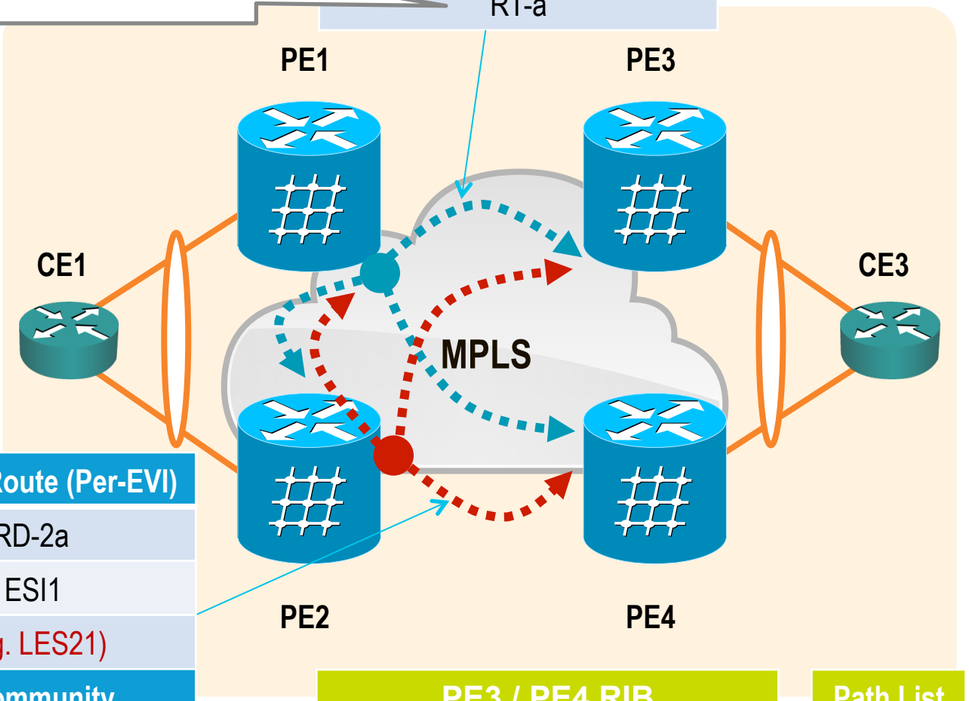
- Ethernet Segment ID (ESI) Auto-Sensing
- Redundancy Group Membership Auto-Discovery
- Ethernet Segment Reachability Advertisement
  - Per-ESI adv.
  - Per-EVI adv.

Aliasing MPLS Label – used by remote PEs to load-balance among local PEs

RT – RT associated with a given EVI

RD – RD unique per advertising PE per EVI

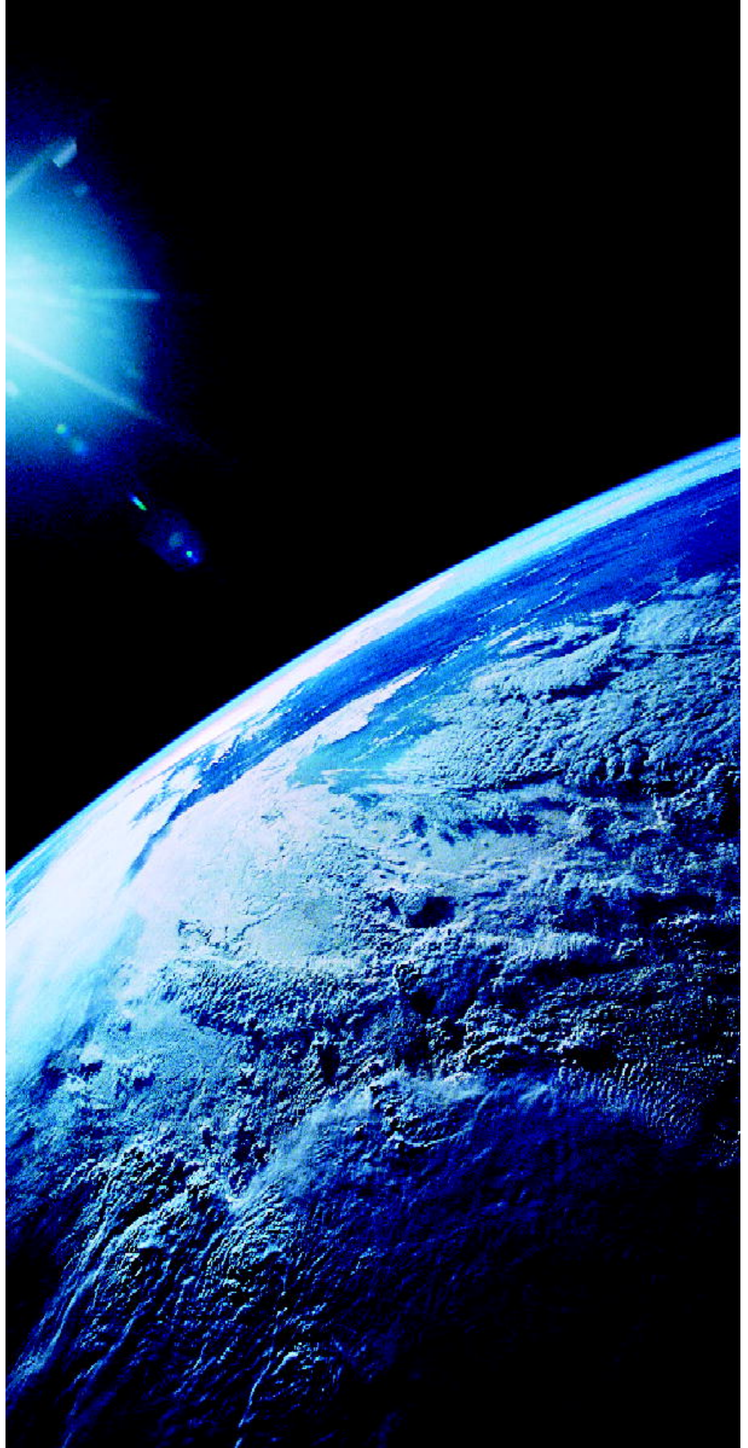
PE 1 Eth A-D Route (Per-EVI)
RD = RD-1a
ESI = ESI1
Label (e.g. LES11)
RT ext. community
RT-a



PE 2 Eth A-D Route (Per-EVI)
RD = RD-2a
ESI = ESI1
Label (e.g. LES21)
RT ext. community
RT-a

PE3 / PE4 RIB			Path List
VPN	**MAC**	**ESI**	**NH**
RT-a	-	ES1	PE1
			PE2

E-VPN
Life of a Packet

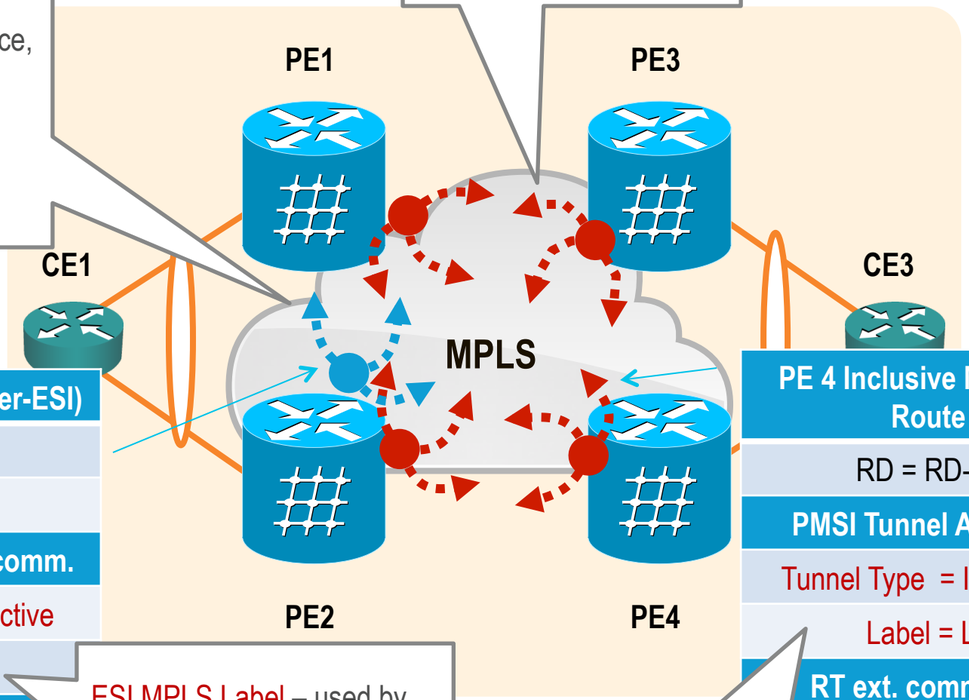


Life of a Packet

Ingress Replication – Multi-destination Traffic Forwarding

During start-up sequence, PE2 sent Per-ESI Ethernet AD route with ESI MPLS label (split-horizon) (see below)

During start-up sequence, PE1, PE2, PE3, PE4 sent Inclusive Multicast route which include Mcast label



PE 2 Eth A-D Route (Per-ESI)
RD = RD20
ESI = ESI1
ESI MPLS Label ext. comm.
Redund. Flag = All-Active
Label = L5
RT ext. community
RT-a, RT-b, RT-c, RT-d

PE 4 Inclusive Multicast Route
RD = RD-4a
PMSI Tunnel Attribute
Tunnel Type = Ing. Repl.
Label = L4
RT ext. community
RT-a

ESI MPLS Label – used by local PEs for split-horizon - downstream assigned (for ingress replication)

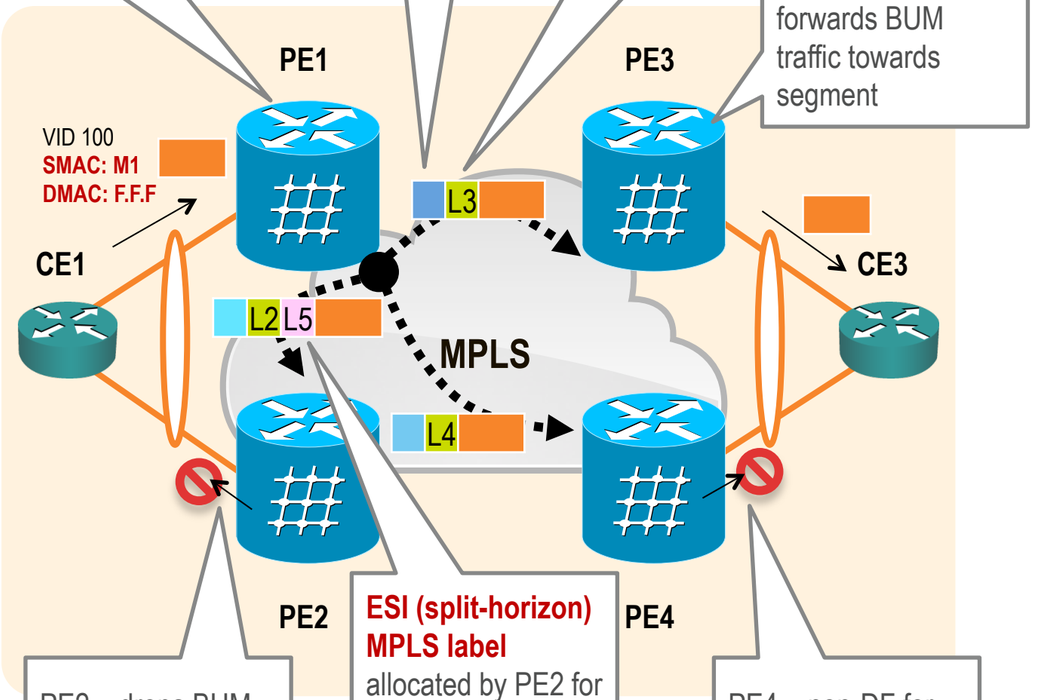
Mcast MPLS Label – used to transmit BUM traffic - downstream assigned (for ingress replication)

PE1 receives broadcast traffic from CE1. PE1 forwards it using ingress replication – 3 copies created

PSN MPLS label to reach PE3

Mcast MPLS Label assigned by PE3 for incoming BUM traffic on a given EVI

PE3 – as DF, it forwards BUM traffic towards segment



PE2 – drops BUM traffic originated on ES1

ESI (split-horizon) MPLS label allocated by PE2 for segment ES1

PE4 – non-DF for given EVI drops BUM traffic

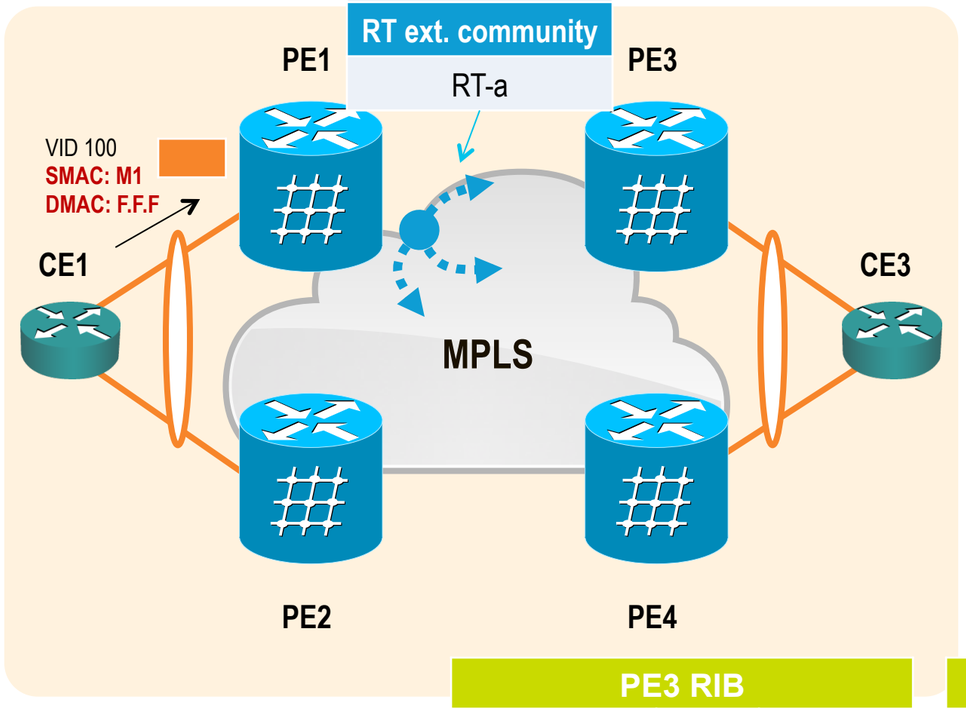
Life of a Packet (cont.)

Unicast Traffic Forwarding

MP2P VPN Label – downstream allocated label used by other PEs to send traffic to advertised MAC

PE1 MAC Route
RD = RD-1a
ESI = ESI1
MAC = M1
Label = L1
RT ext. community
RT-a

MAC advertised by route

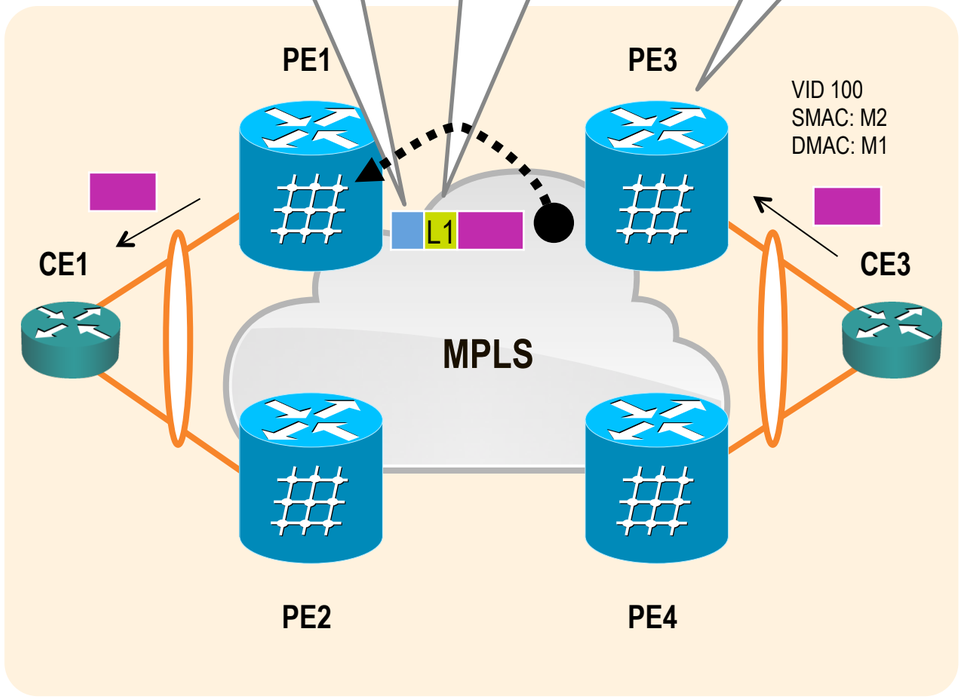


PE3 RIB			Path List
VPN	MAC	ESI	NH
RT-a	M1	ES1	PE1

PSN MPLS label to reach PE1

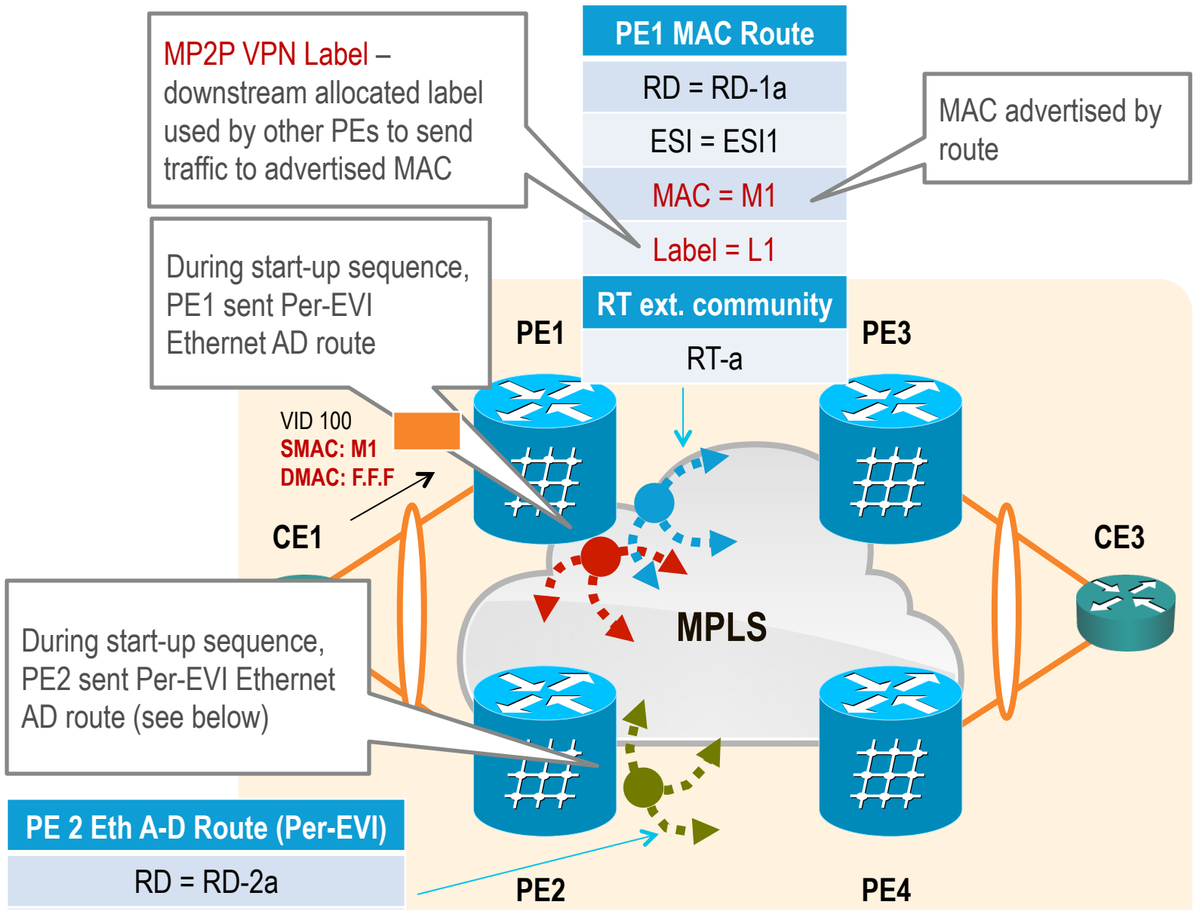
MP2P VPN Label assigned by PE1 for incoming traffic for the target EVI

PE3 forwards traffic destined to M1 based on RIB information (PE1)



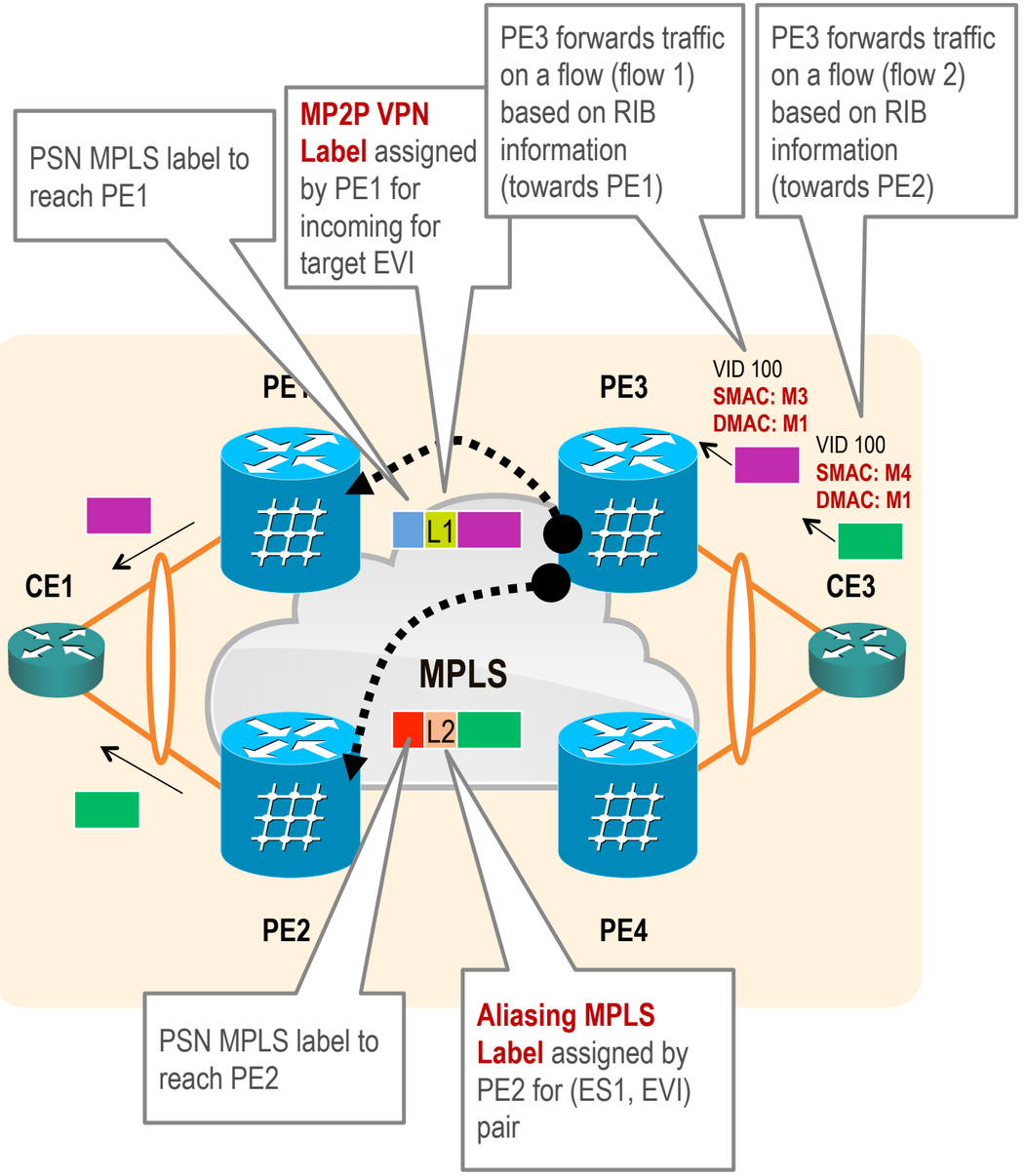
Life of a Packet (cont.)

Unicast Forwarding and Aliasing

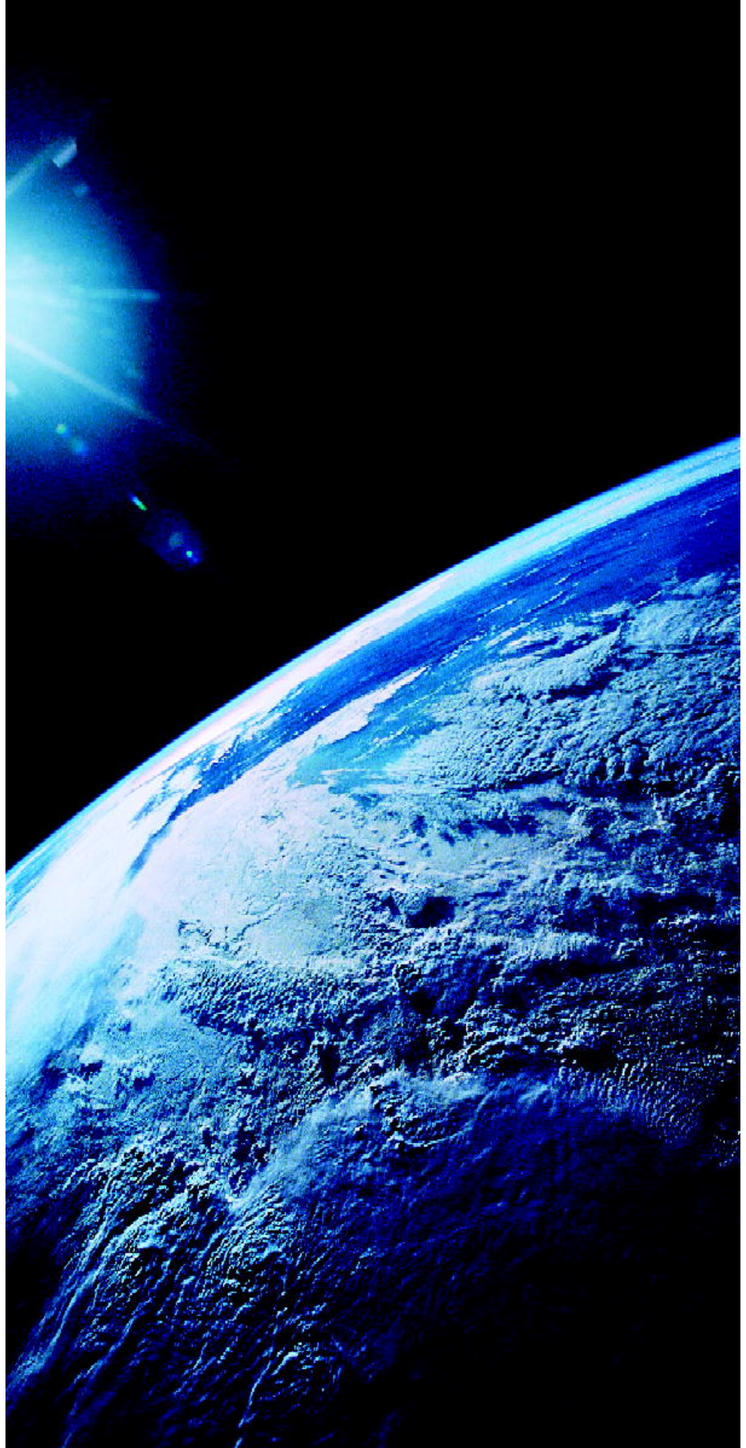


PE3, PE4 RIB		
VPN	MAC	ESI
RT-a	M1	ES1

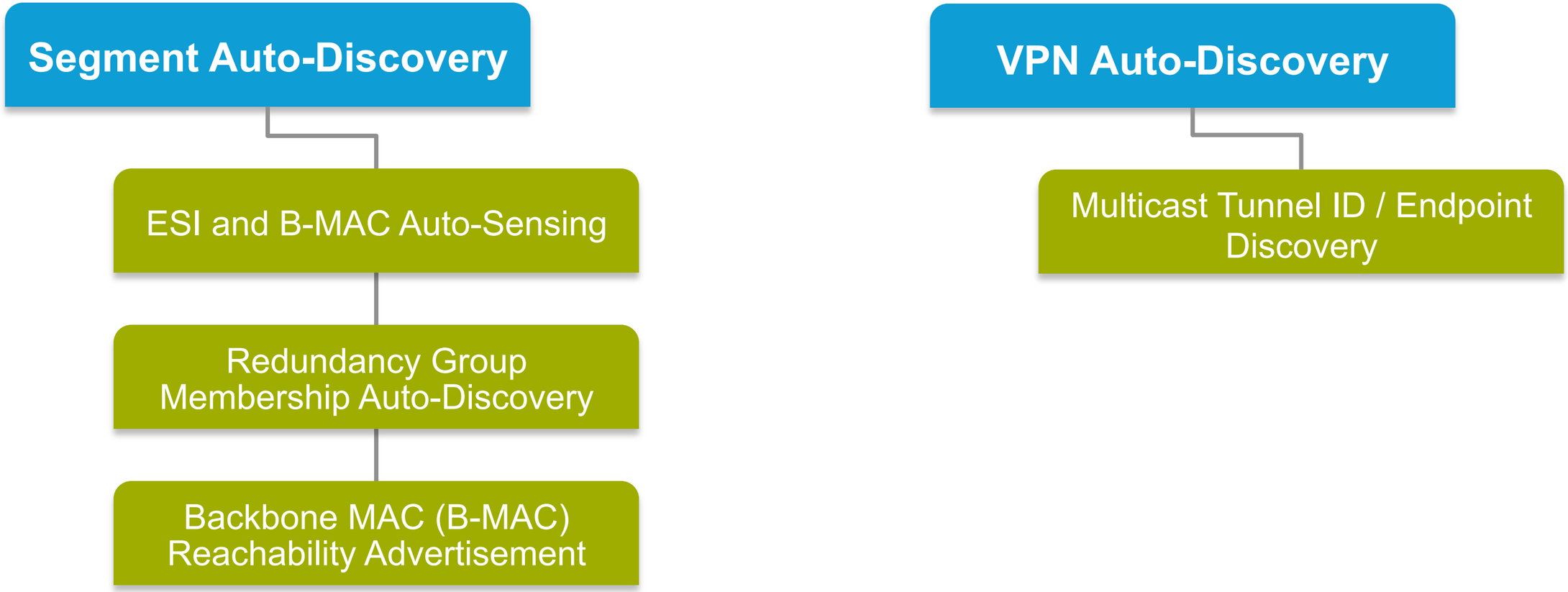
Path List
NH
PE1
PE2



PBB-EVPN Startup Sequences



PBB-EVPN Startup Sequence



PBB-EVPN Startup Sequence (cont.)

ESI and B-MAC Auto-Sensing

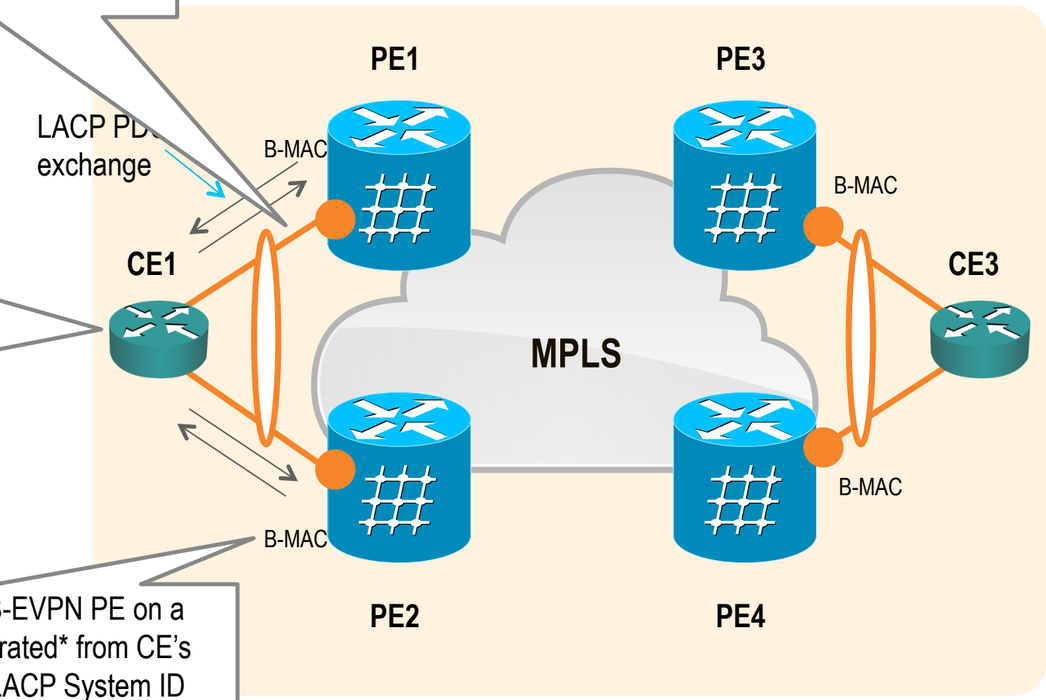
Segment Auto-Discovery

ESI and B-MAC Auto-Sensing

ESI (10B) can be auto-generated* from CE's LACP information -> concatenation of CE's LACP System Priority + Sys ID + Port Key

Example:
0000. 0011.0022.0033.0018

System Priority	System MAC Address	Port Key
2 bytes	6 bytes	2 bytes



CE LACP info:
LACP System ID (MAC) (6B)
e.g. 0011.0022.0033
LACP System Priority (2B)
e.g. 0000
LACP Port Key (2B)
e.g. 0018

Source B-MAC used at PBB-EVPN PE on a given ESI can be auto-generated* from CE's LACP information -> CE's LACP System ID MAC with U/L** (Universal / Locally Administered) bit flipped

Example: 0211.0022.0033

(*) ESI and B-MAC can also be manually configured
(**) U/L is second-least-significant bit of most significant byte

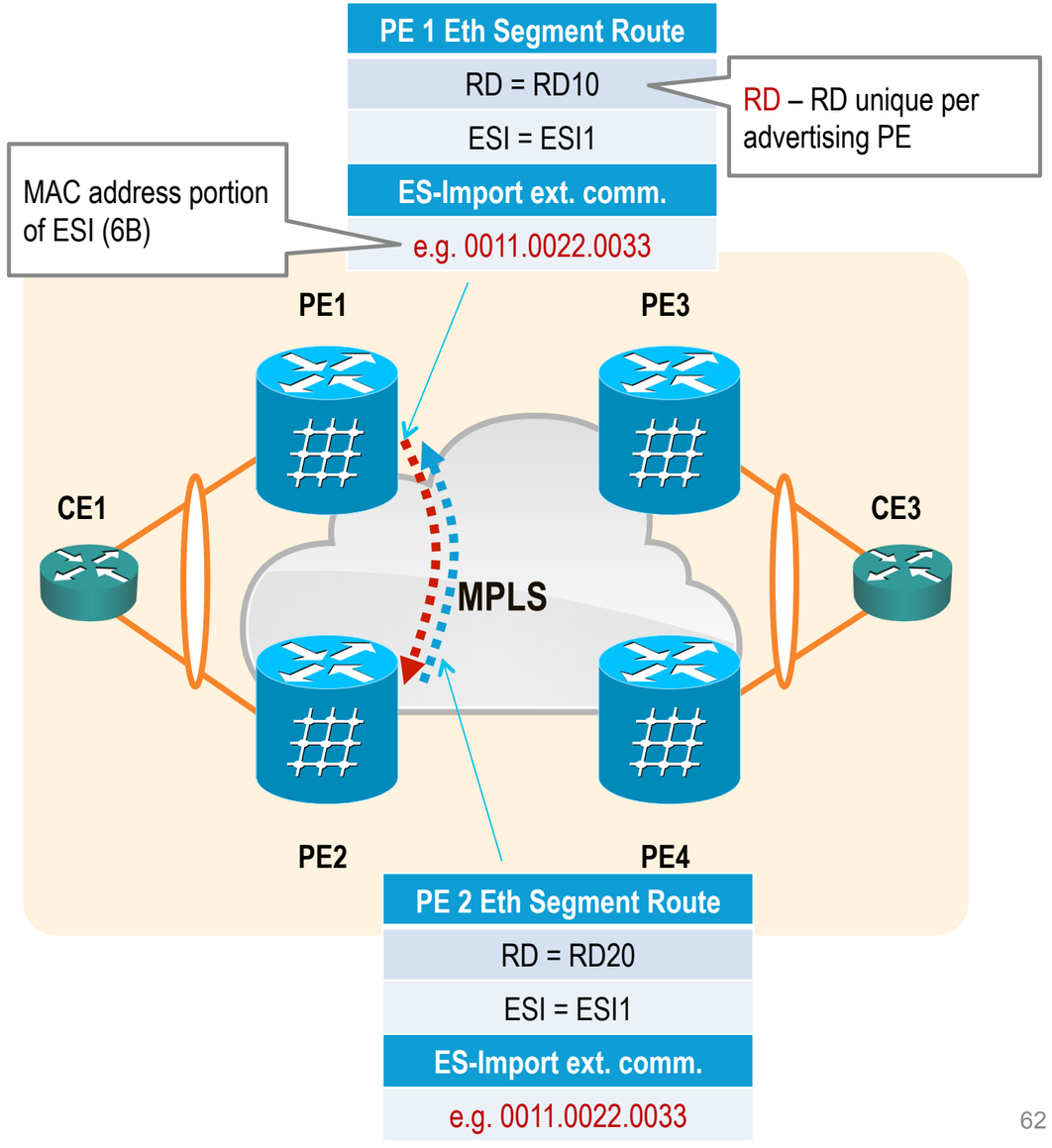
PBB-EVPN Startup Sequence (cont.)

BGP Ethernet Segment Route

Segment Auto-Discovery

ESI and B-MAC Auto-Sensing

Redundancy Group Membership Auto-Discovery



PBB-EVPN Startup Sequence (cont.)

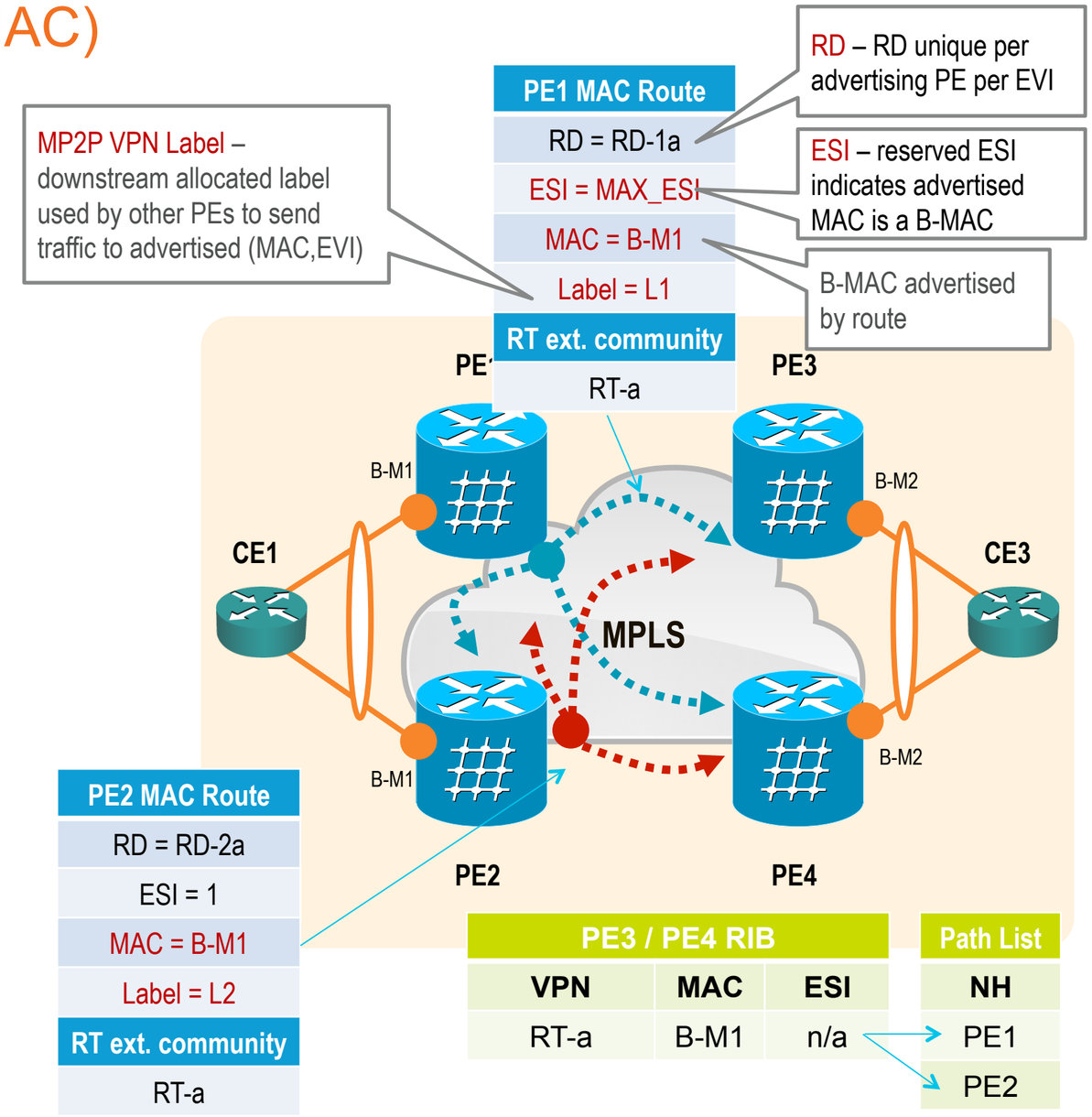
BGP MAC Advertisement Route (B-MAC)

Segment Auto-Discovery

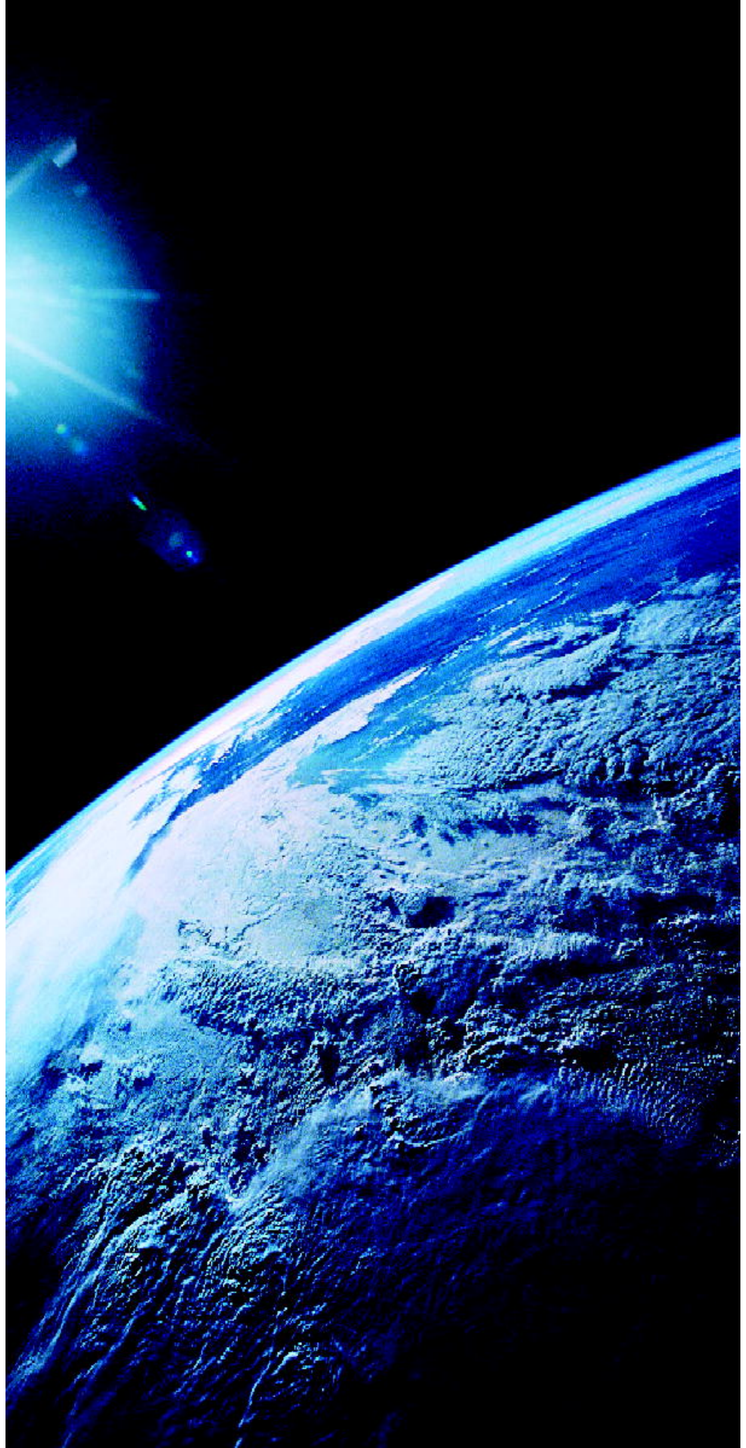
ESI and B-MAC Auto-Sensing

Redundancy Group Membership Auto-Discovery

Backbone MAC (B-MAC) Reachability Advertisement



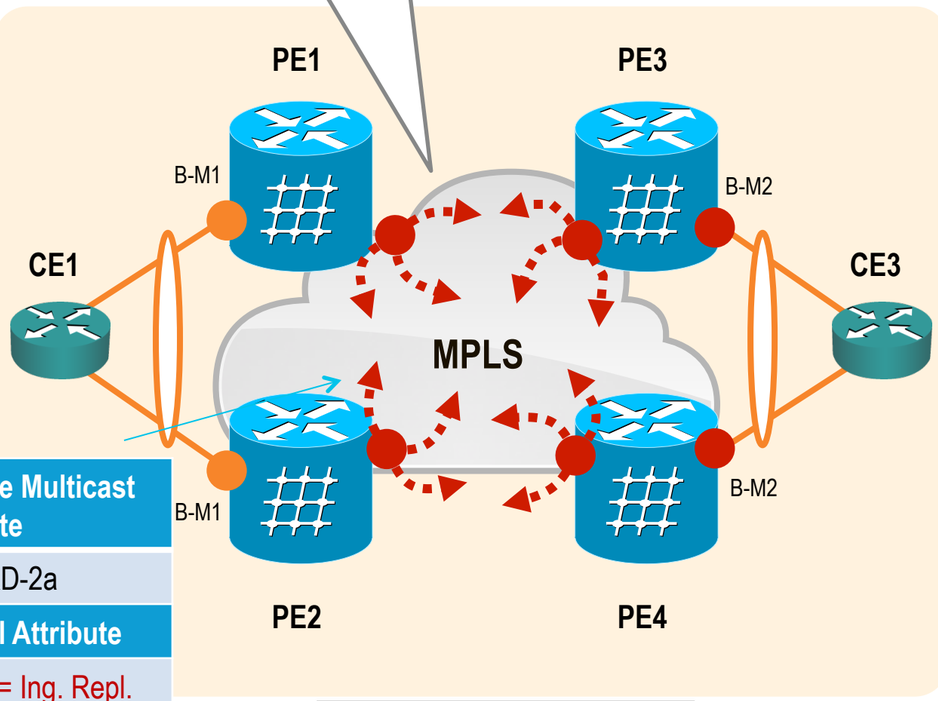
PBB-EVPN
Life of a Packet



Life of a Packet

Ingress Replication – Multi-destination Traffic Forwarding

During start-up sequence, PE1, PE2, PE3, PE4 sent Inclusive Multicast route which include Mcast label



PE 2 Inclusive Multicast Route
RD = RD-2a
PMSI Tunnel Attribute
Tunnel Type = Ing. Repl.
Label = L2
RT ext. community
RT-a

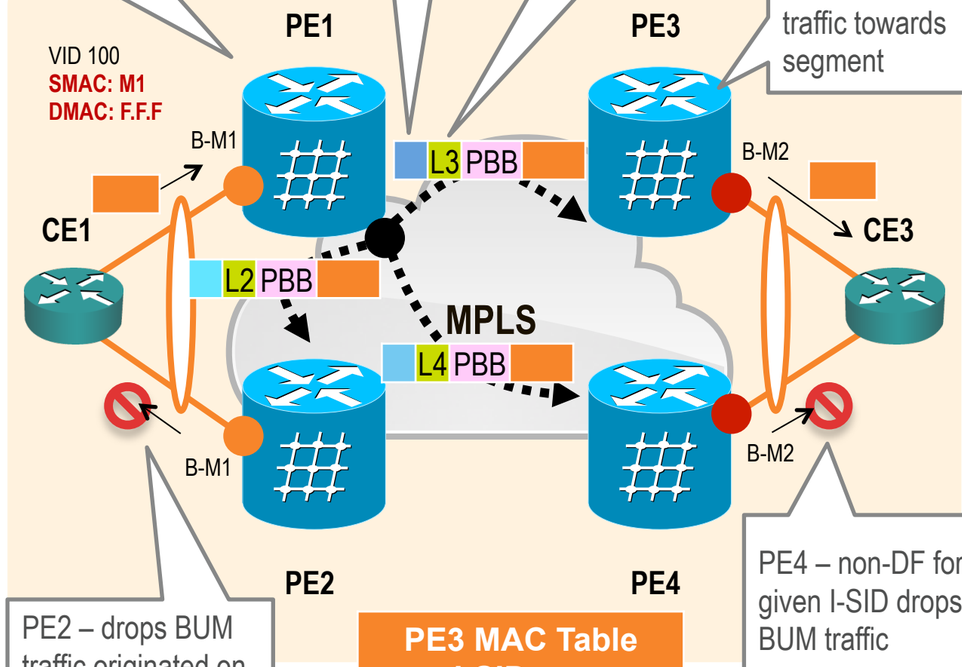
Mcast MPLS Label – used to transmit BUM traffic - downstream assigned (for ingress replication)

PE1 receives broadcast traffic from CE1. PE1 adds PBB encapsulation and forwards it using ingress replication – 3 copies created

PSN MPLS label to reach PE3

Mcast MPLS Label assigned by PE3 for incoming BUM traffic on a given EVI

PE3 – as DF, it forwards BUM traffic towards segment



PE2 – drops BUM traffic originated on same source B-MAC (B-M1)

PE4 – non-DF for given I-SID drops BUM traffic

PE3 MAC Table I-SID xyz

C-MAC	B-MAC
M1	B-M1

Data-plane based MAC learning for C-MAC / B-MAC association

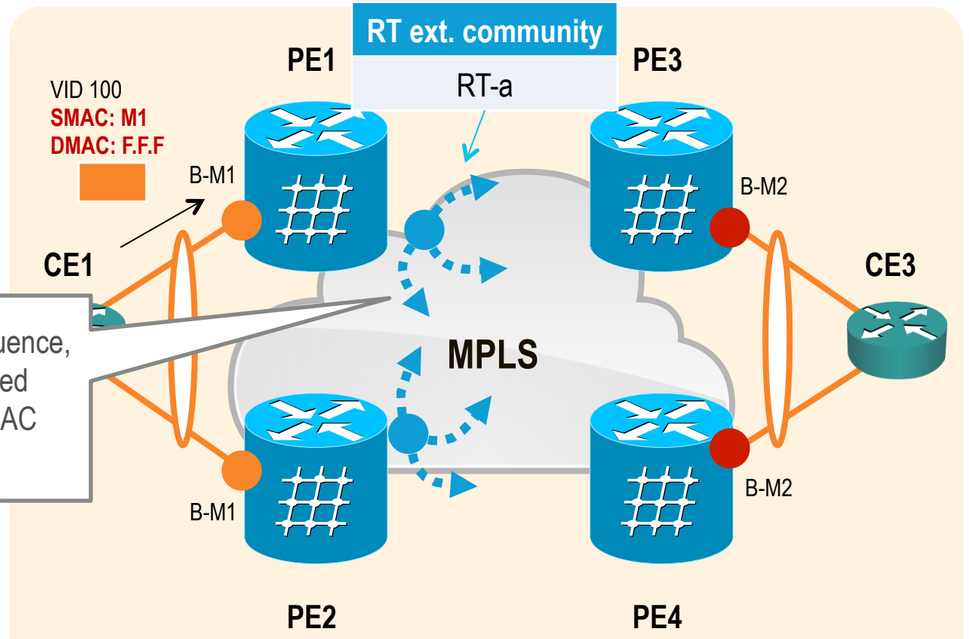
Life of a Packet (cont.)

Unicast Traffic Forwarding

MP2P VPN Label – downstream allocated label used by other PEs to send traffic to advertised MAC

PE1 MAC Route
RD = RD-1a
ESI = 1
MAC = B-M1
Label = L1
RT ext. community
RT-a

MAC advertised by route



During start-up sequence, PE1 & PE2 advertised MAC routes for B-MAC (B-M1)

PE3 RIB		
VPN	**MAC**	**ESI**
RT-a	B-M1	n/a

Path List
NH
PE1
PE2

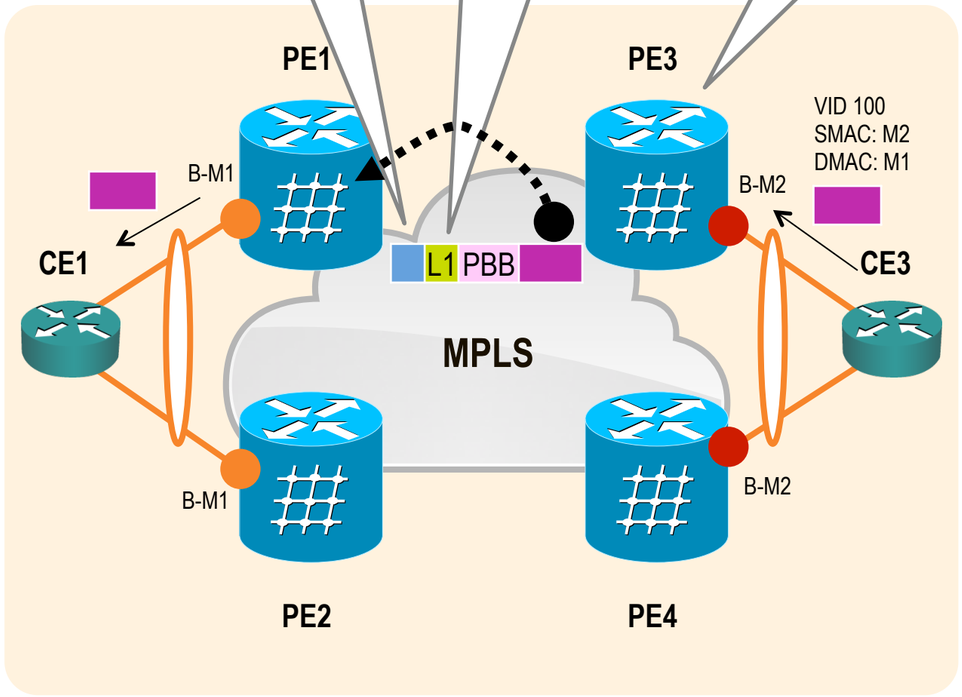
PE3 MAC Table I-SID xyz	
C-MAC	**B-MAC**
M1	B-M1

Data-plane based MAC learning for C-MAC / B-MAC association

PSN MPLS label to reach PE1

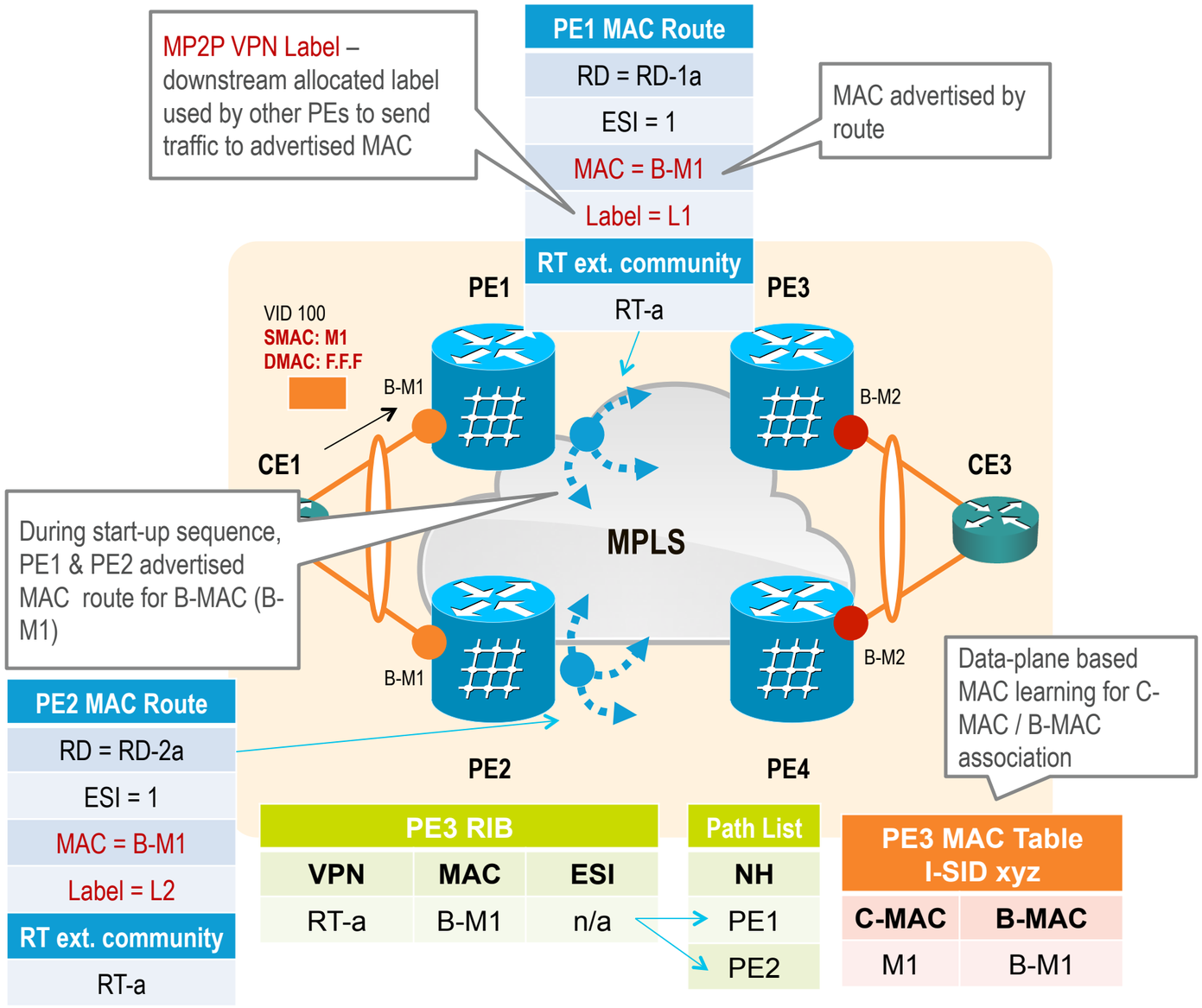
MP2P VPN Label assigned by PE1 for incoming traffic for the target EVI

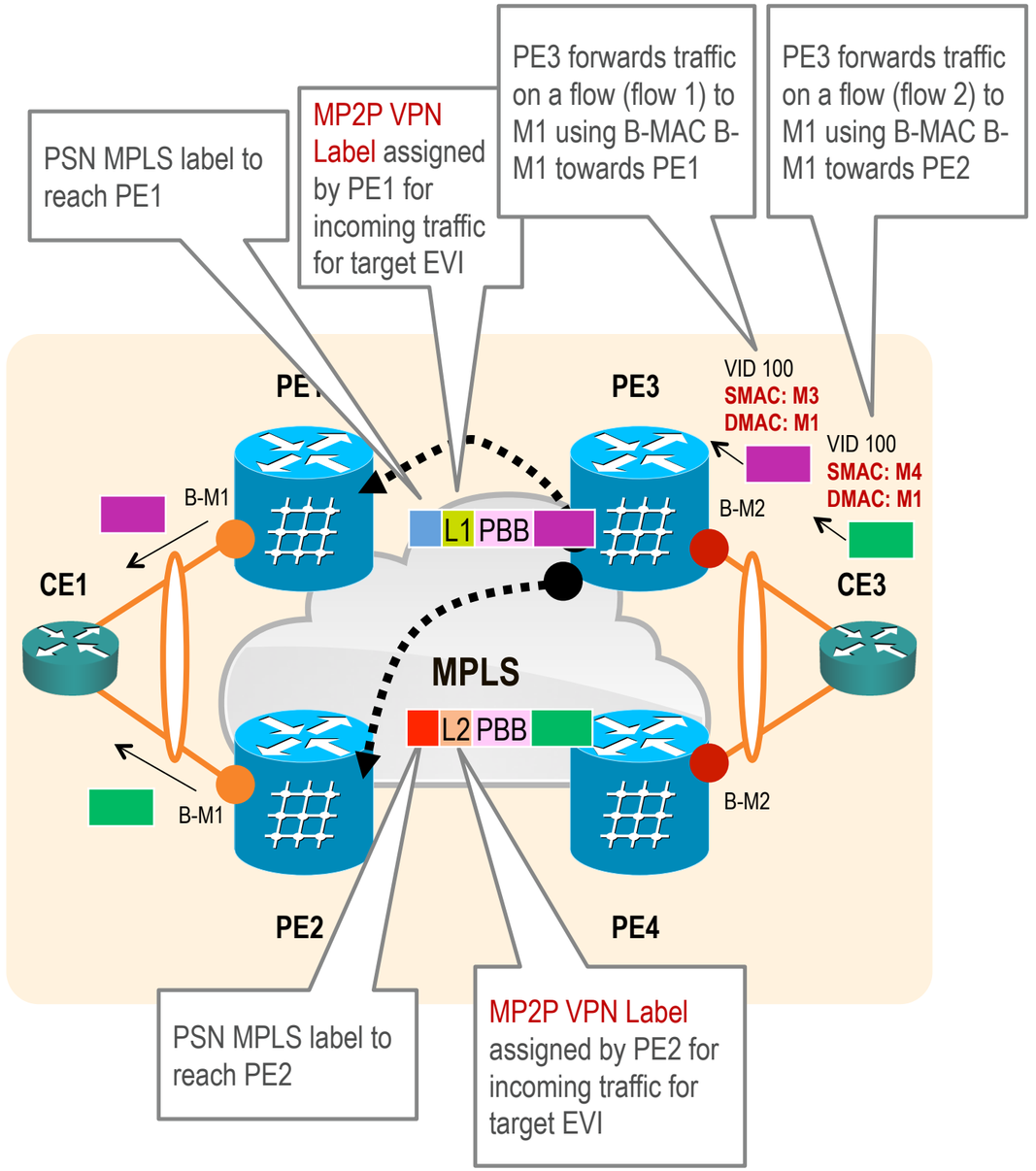
PE3 forwards traffic destined to M1 using B-MAC B-M1 towards PE1



Life of a Packet (cont.)

Unicast Traffic Forwarding and Aliasing





Life of a Packet (cont.)

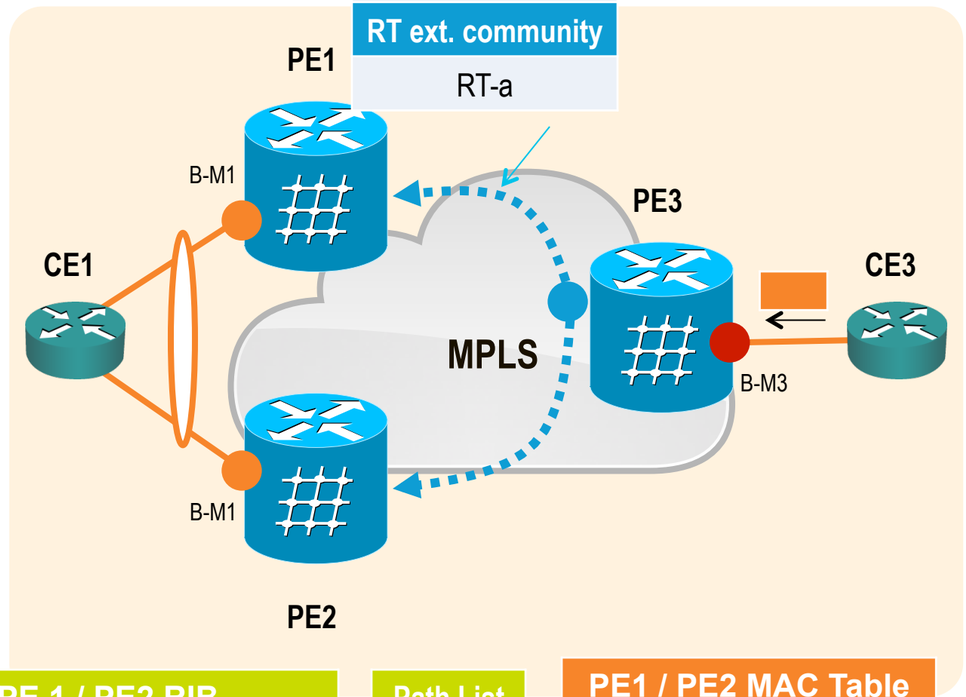
Active / Active Load Balancing from CE

PE3 MAC Route
RD = RD-3a
ESI = 0
MAC = B-M3
Label = L3
RT ext. community
RT-a

MP2P VPN Label – downstream allocated label used by other PEs to send traffic to advertised MAC

ESI == 0 used for Single Home Device

MAC advertised by route



PE 1 / PE2 RIB		
VPN	**MAC**	**ESI**
RT-a	B-M3	0

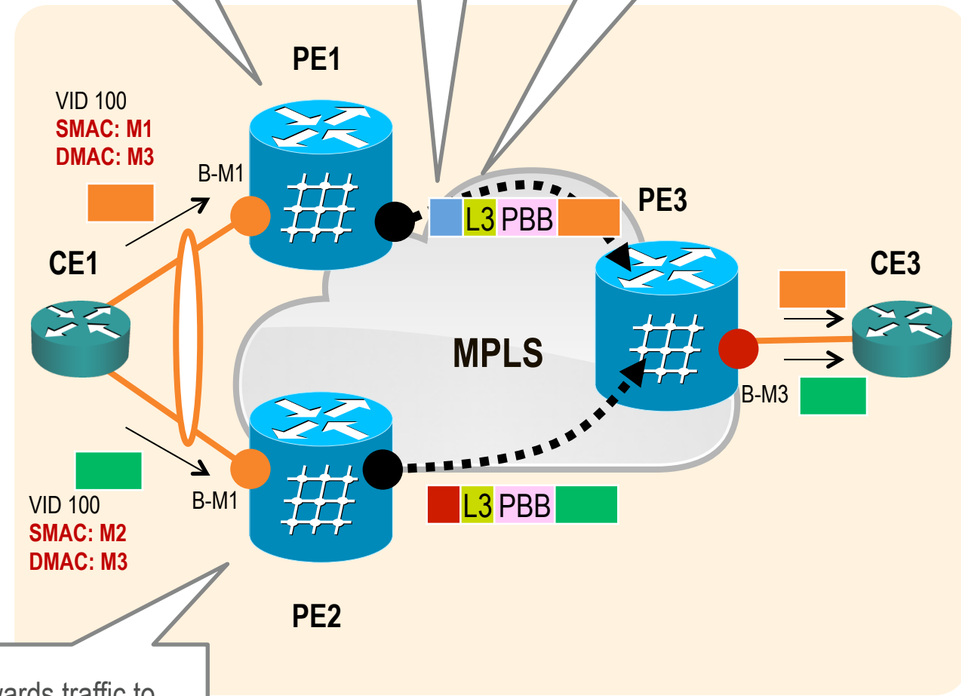
Path List
NH
PE3

PE1 / PE2 MAC Table I-SID xyz	
C-MAC	**B-MAC**
M3	B-M3

PE1 forwards traffic to M3 using B-MAC B-M3 towards PE3

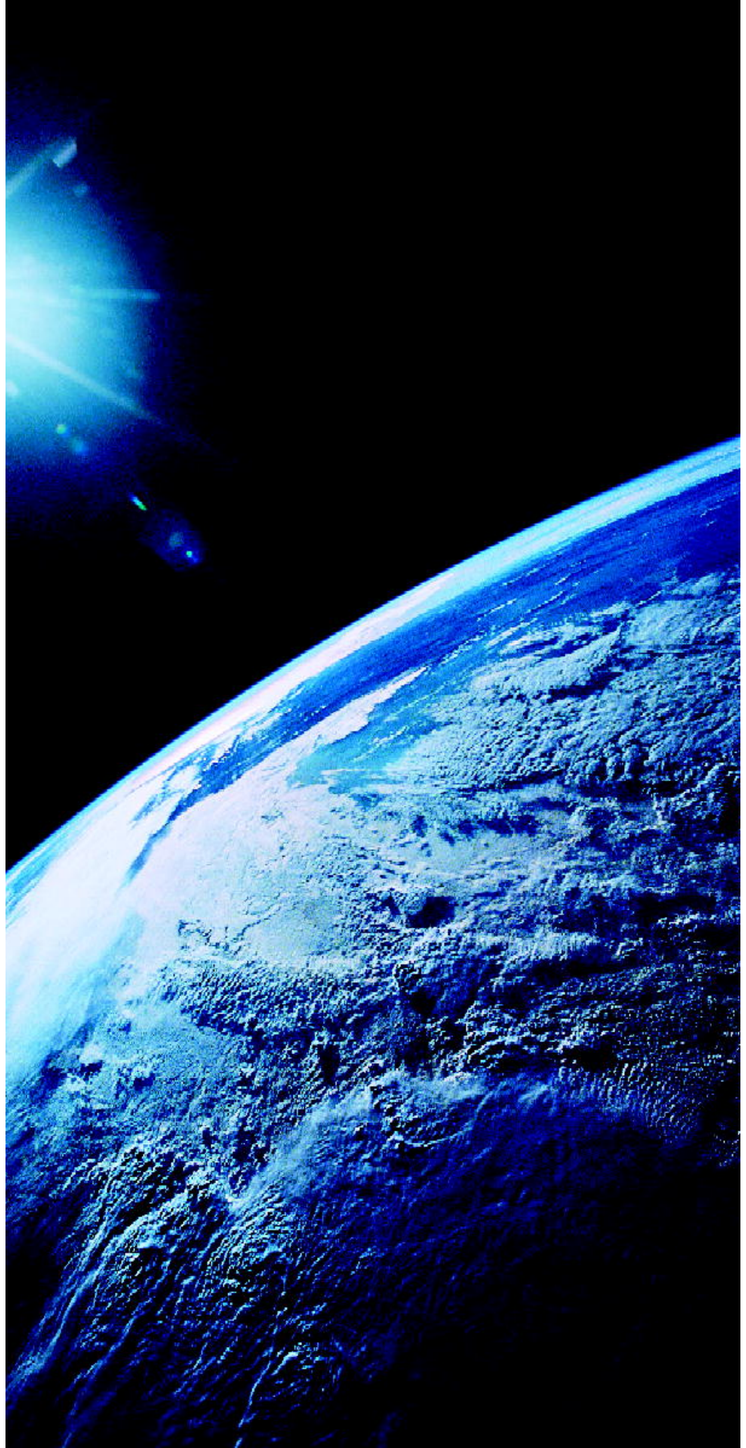
PSN MPLS label to reach PE3

MP2P VPN Label assigned by PE3 for incoming traffic for target EVI



PE2 forwards traffic to M3 using B-MAC B-M3 towards PE3

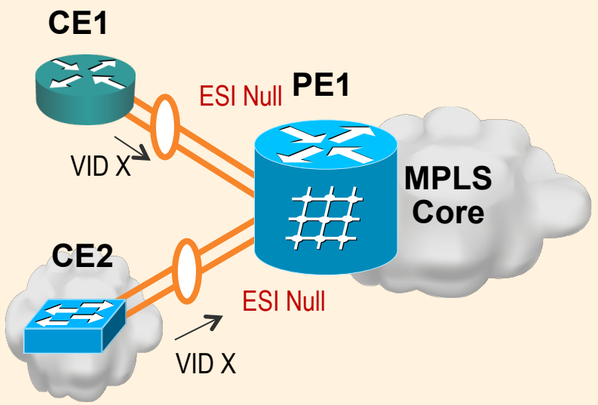
Sample Supported Access Topologies



PBB-EVPN

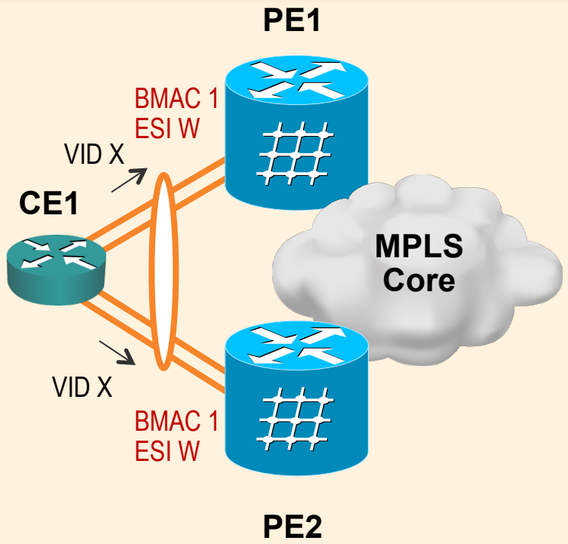
Sample Supported Access Topologies

Single Home Device (SHD) Single Home Network (SHN)



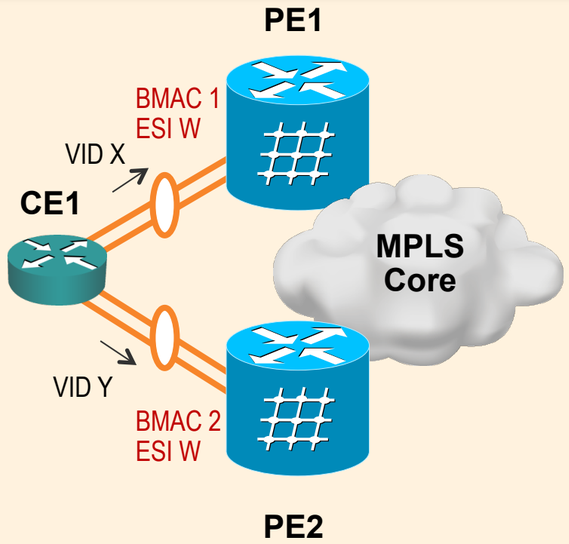
- Null Ethernet Segment Identifier (ESI)

Dual Home Device (DHD) Active / Active Per-Flow LB



- Identical B-MAC on PBB-EVPN PEs (PE1 / PE2)
- Identical ESI on PBB-EVPN PEs

Dual Home Device (DHD) Active / Active Per-Service LB

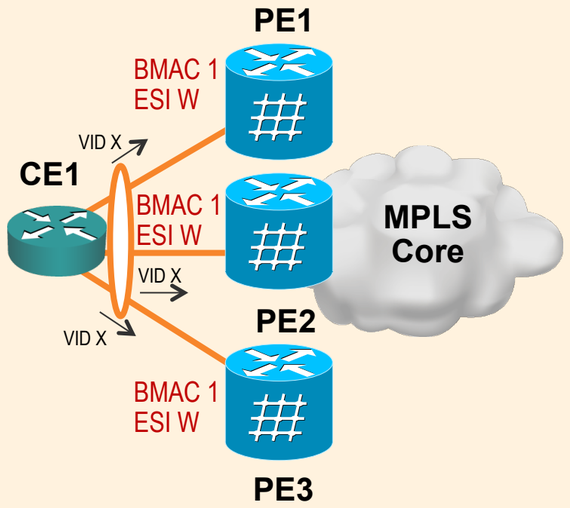


- Different B-MAC on PBB-EVPN PEs (PE1 / PE2)
- Identical ESI on PBB-EVPN PEs
- Per service (I-SID) carving (manual or automatic)

PBB-EVPN

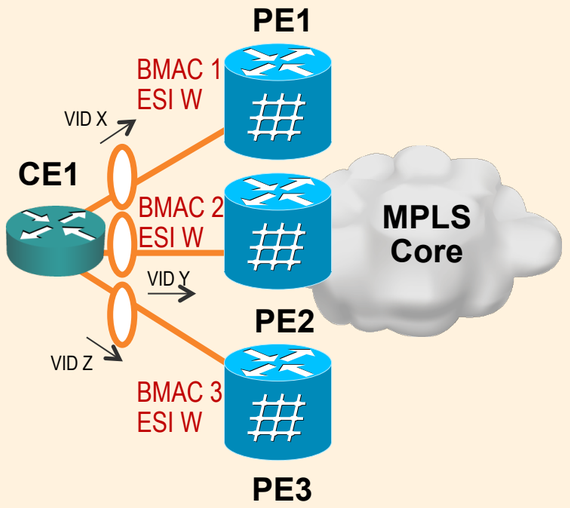
Sample Supported Access Topologies (cont.)

Multi Home Device (MHD) Active / Active Per-Flow LB



- More than two (2) PEs in redundancy group
- Same as DHD Act/Act per-flow LB

Multi Home Device (MHD) Active / Active Per-Service LB

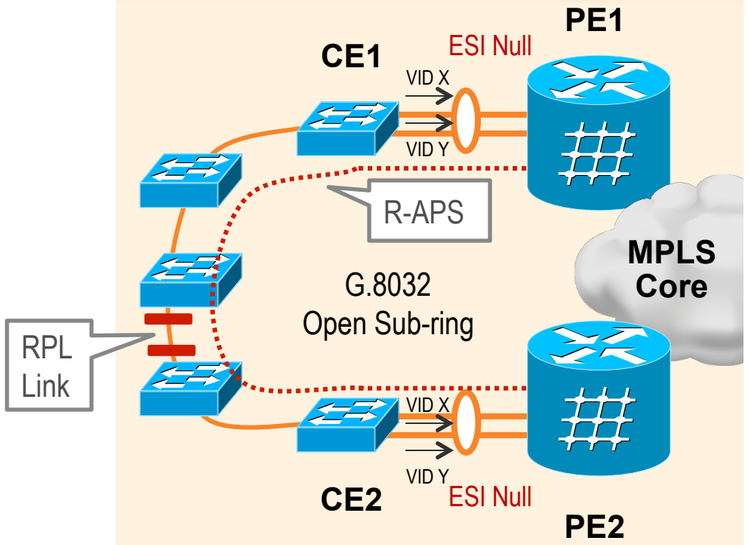


- More than two (2) PEs in redundancy group
- Same as DHD Act/Act per-service LB

PBB-EVPN

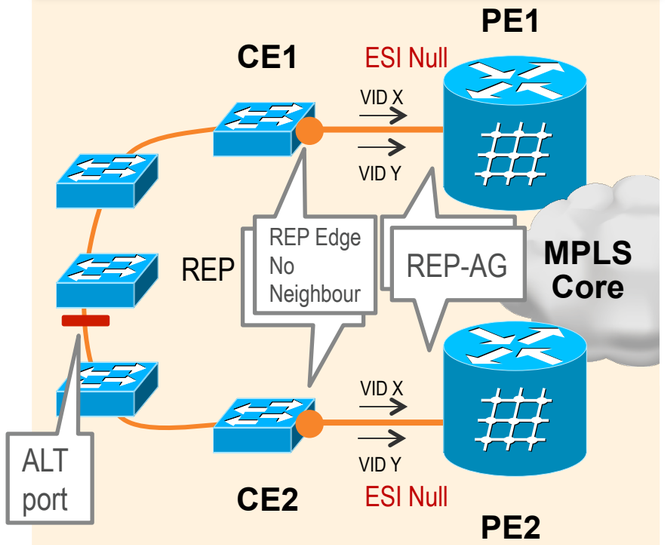
Sample Supported Access Topologies (cont.)

Dual Home Network (DHN) ITU-T G.8032



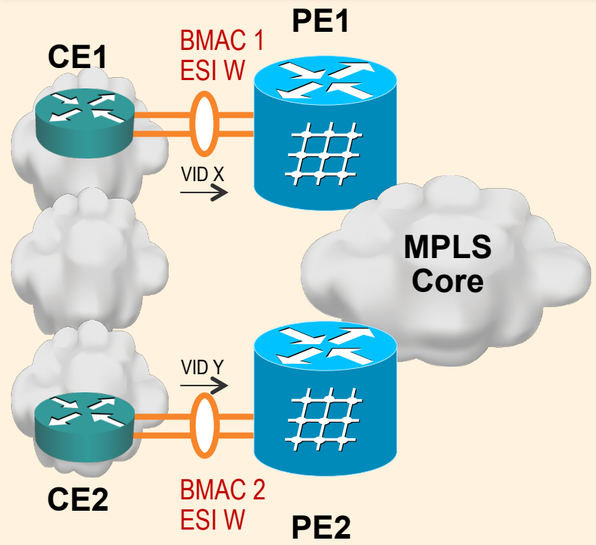
- Treated as SHN by PBB-EVPN PEs (PE1 / PE2)
  - Null ESI; No DF election / No service carving
- Ring operation controlled by R-APS protocol

Dual Home Network (DHN) REP



- Treated as SHN by PBB-EVPN PEs (PE1 / PE2)
  - Null ESI; No DF election / No service carving
- Segment operation controlled by REP protocol

Dual Home Network (DHN) Active / Active Per-Service LB



- Different B-MAC on PBB-EVPN PEs (PE1 / PE2)
- Identical ESI on PBB-EVPN PEs
- Per service (I-SID) carving (manual or automatic)

Summary

- E-VPN & PBB-EVPN are next generation L2VPN Solutions that address resiliency and forwarding policy requirements.
- E-VPN & PBB-EVPN use BGP for MAC distribution/learning over the PSN.
- The following concepts were discussed:
  - Ethernet Segments
  - DF Election and Filtering
  - Split Horizon
  - Aliasing
  - Backup Path
  - MAC Mass Withdrawal
- The operation of E-VPN & PBB-EVPN was discussed.

Comparison of L2VPN Solutions

Requirement	VPLS	E-VPN	PBB-EVPN
All-Active Redundancy with Flow Based Load-balancing		✔	✔
Flow Based Multi-pathing		✔	✔
Geo-redundancy and Flexible Redundancy Grouping		✔	✔
Core Auto-Discovery	✔	✔	✔
Access Multi-homing Auto-Discovery		✔	✔
New Service Interfaces		✔	✔
LSM with P2MP Tree	✔	✔	✔
LSM with MP2MP Tree		✔	✔
Fast Convergence on Failure	✔	✔	✔
Fast Convergence on MAC Mobility	✔	✔	✔
Fast Convergence: Avoiding C-MAC Flushing			✔
Scale to Millions of MAC Addresses			✔
Confinement of C-MAC Learning			✔
Support C-MAC Mobility with MAC Sub-netting			✔
Seamless interworking between TRILL /802.1Qaq/802.1Qbp		✔ (C-MAC Transparency issue)	✔

Otázky a odpovědi

Zodpovíme též v “Ptali jste se” v sále LEO v 17:45 – 18:30

e-mail: connect-cz@cisco.com

Prosíme, ohodnoťte tuto přednášku.

Děkujeme za pozornost.