ADVANCED VPLS

Emil Gągała
PLNOG, Kraków, 21.10.2010

Agenda

Ingress replication with P2MP LSP

VPLS scaling

- H-VPLS and
- Full Mesh concept

BGP-LDP VPLS interworking

VPLS Multi-homing

Interworking with native Ethernet xSTP access networks

VPLS Intro Quiz

1. How is MAC signalled in VPLS?
2. What are two protocols commonly used for VPLS signalling?
3. Why loops cannot happen in VPLS core?

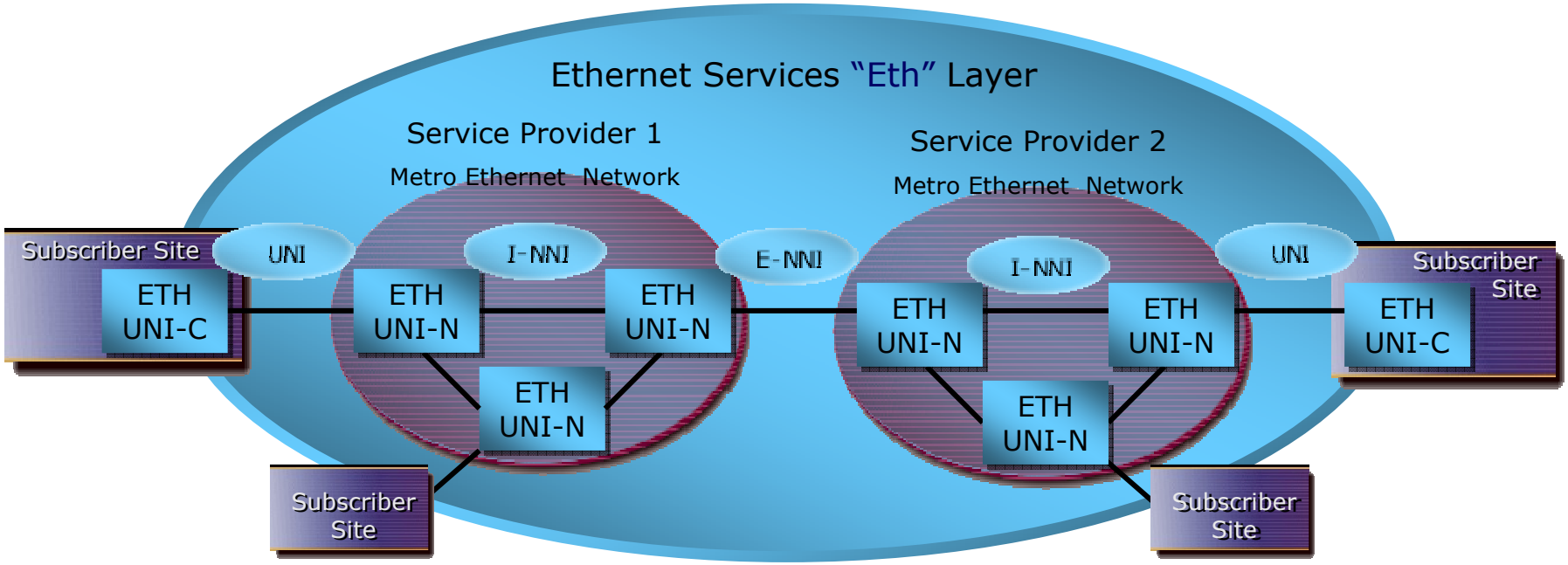
Bonus question:

What is content of VPLS BGP NLRI?

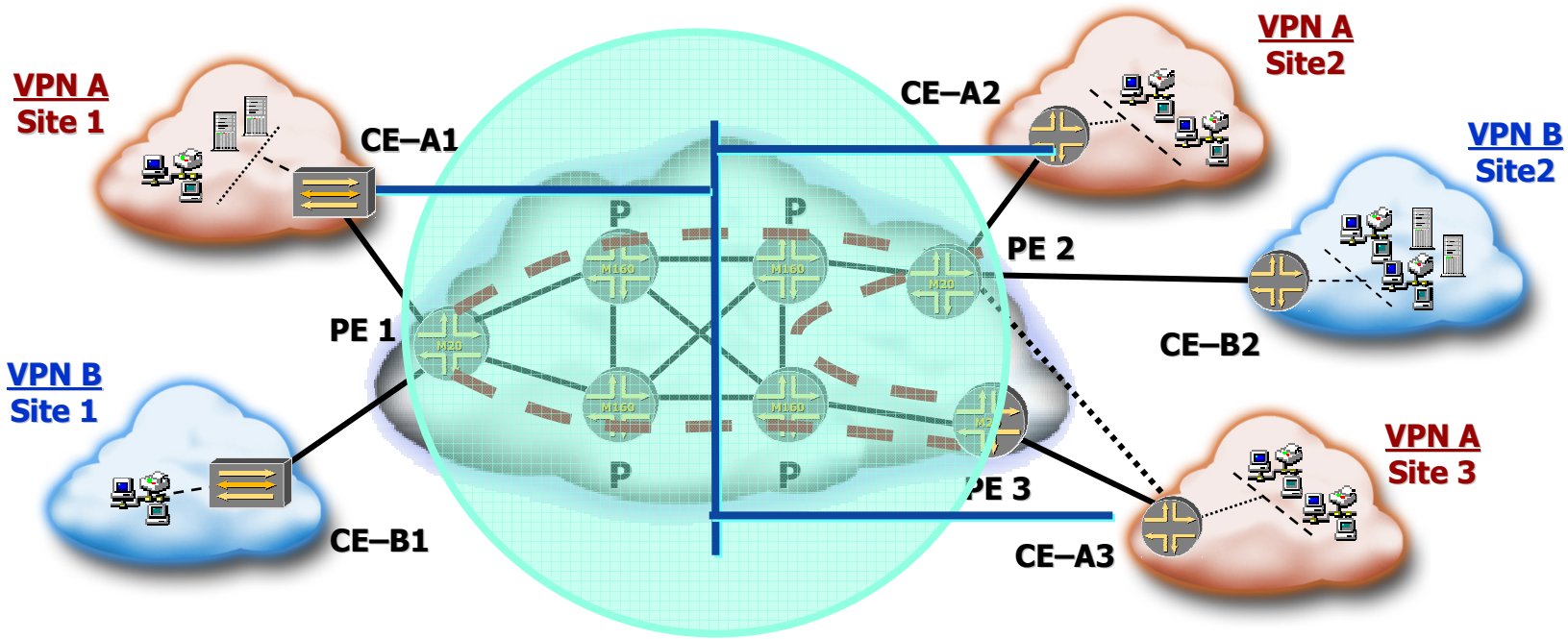
ETHERNET SERVICES

MEF describes Ethernet Services

- E-Line - Point-to-Point
- E-LAN - Multipoint-to-Multipoint
- E-Tree - Point-to-Multipoint



Virtual Private LAN Service



A private Ethernet network constructed over a ‘shared’ infrastructure which may span several metro networks

Service: Multipoint to Multipoint Ethernet connectivity

- For the CE perspective, the SP network looks like a private Ethernet broadcast domain

Complements Layer 3 2547bis and Layer 2 Services

Service Provider Needs, As We See Them

Scalability – number of MACs and Pseudowires

Resilience - need to multi-home customers

- Don't want to rely on **customer** Spanning Tree

Optimizing multicast

- Use a substrate of RSVP-TE P2MP LSPs

Concerns about "Juniper-proprietary" solution

- Use LDP VPLS if needed, and interwork with BGP

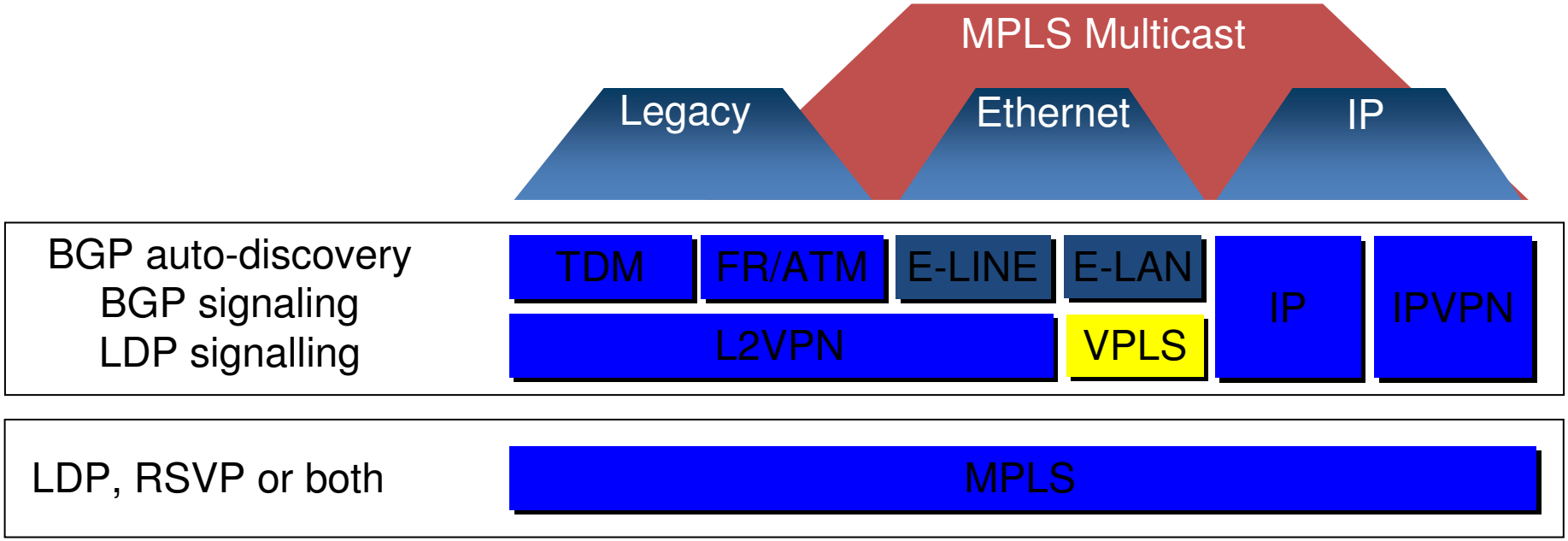
Simplifying configuration

- Automatic assignment of site IDs
- The use of macros

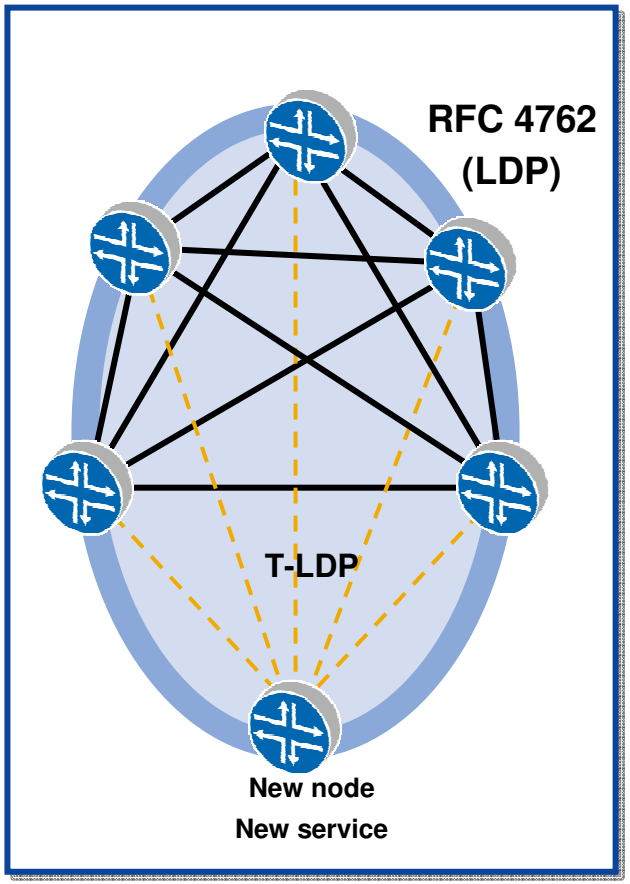
VPLS CONTROL-PLANE

Same operational procedures

- Legacy, Ethernet & IP
- Re-use existing skills: trained personnel advantage



VIRTUAL PRIVATE LAN SERVICES
TWO DEPLOYED STANDARDS

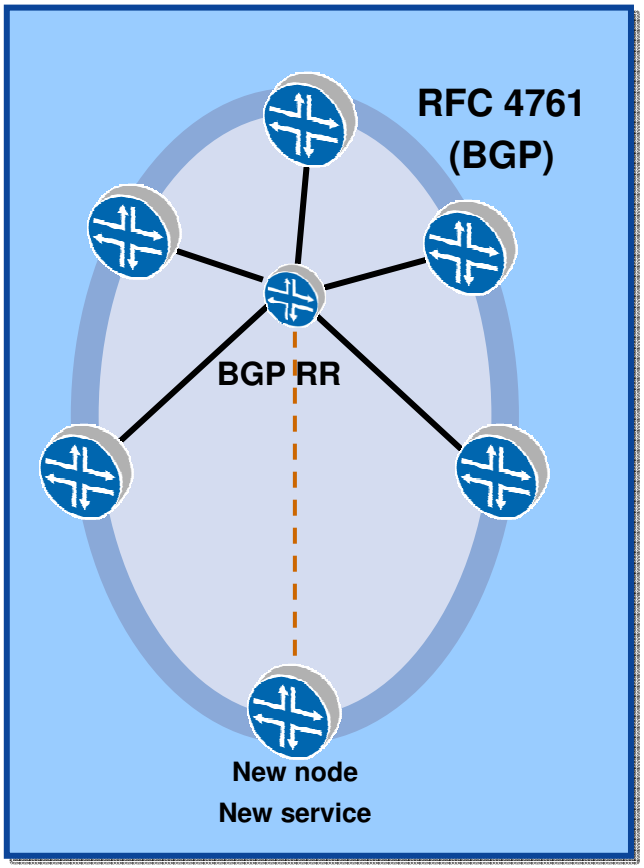


LDP-based

- Signaling only, no auto-discovery
- High-touch provisioning

BGP-based

- Signaling & Auto-discovery
- Inter-area/ metro/provider
- Multicast optimization



Existing control-plane session
New control-plane session

INGRESS REPLICATION

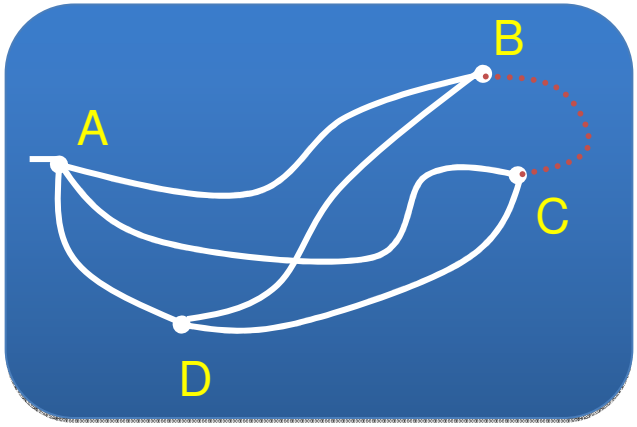
REGULAR VPLS (RFC 4761/4762)

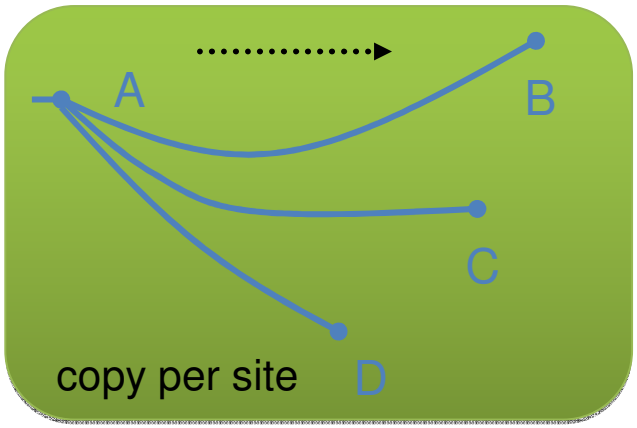
VPLS requires full-mesh

- Split-horizon rule
- Prevents loops

Ingress replication is A Good Thing

- But it's inefficient, real pain for IPTV
- Two solutions:
  - Move replication down one level
  - Distribute all replication into the network





RECIPE FOR ANY VPLS SERVICE

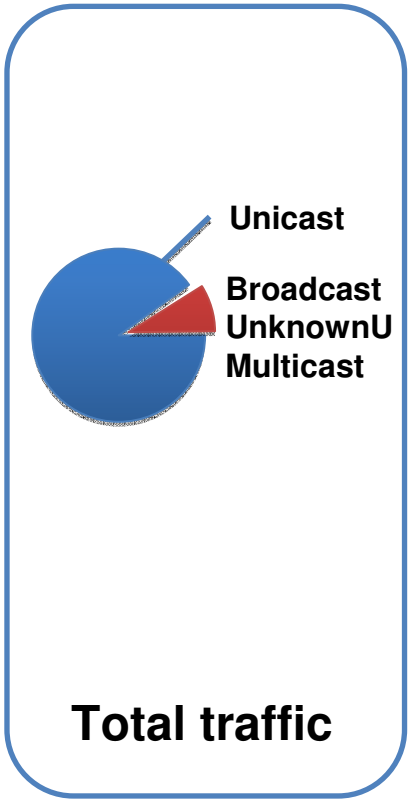
Unicast traffic
- Limit total unicast traffic

Unknown unicast traffic
- Someone intentionally tried to emulate multicast w/o using multicast addressing
- A decent conversation became one-way
- => All cases: police down to minimum amount / disable

Multicast traffic
- Limit % of multicast traffic – filter/police on Mcast MAC range

Broadcast traffic
- Limit % of bcast traffic – filter/police on DA MAC
- Limit # of MACs per logical port/interface

Flood traffic
- All Broadcast, Multicast & unknown unicast traffic
- Usually policer enforceable per VSI/VPLS Instance



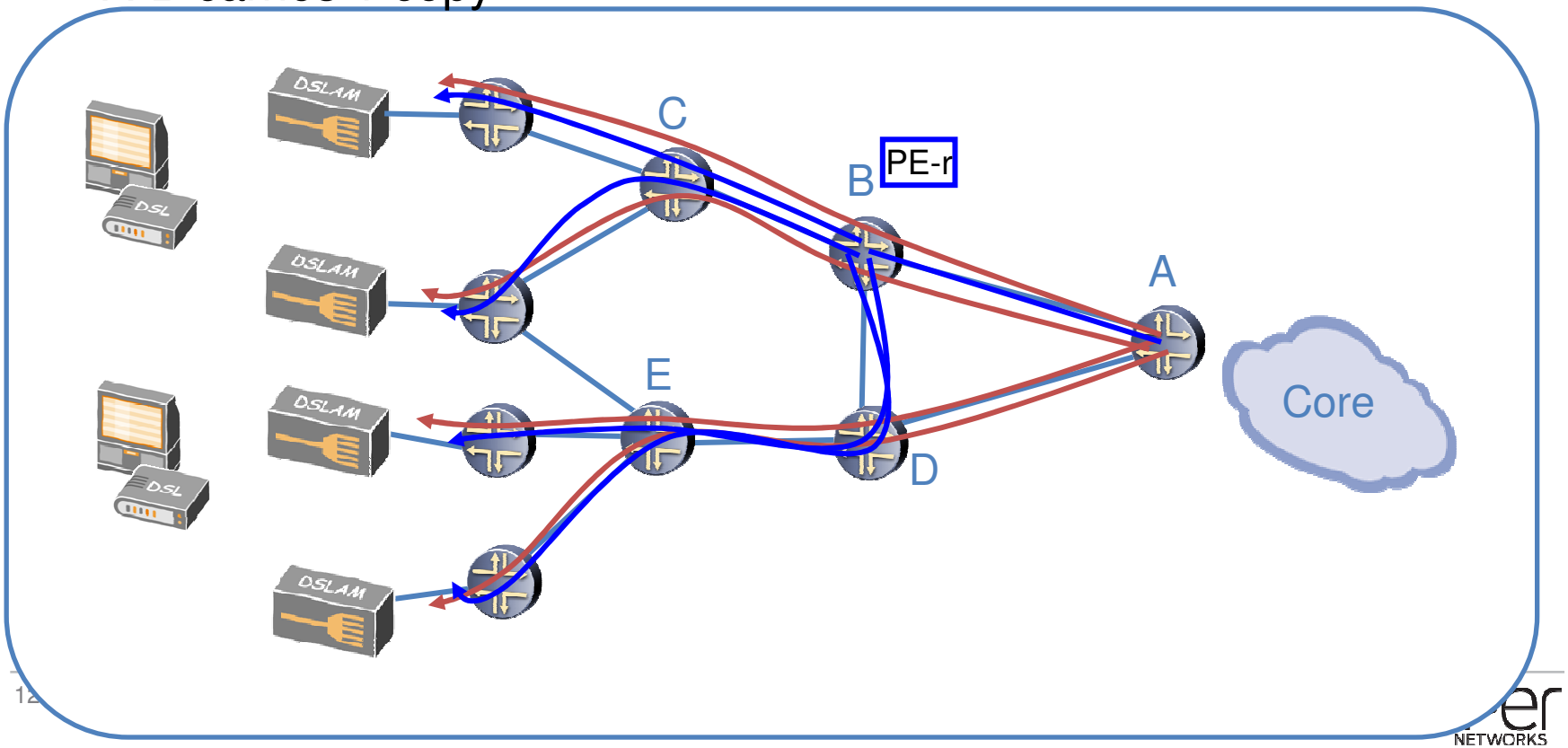
REPLICATION IN A REAL NETWORK

Regular VPLS

- A-B & A-D carry 2 copies (4x 250Mbps = 1Gbps)

H-VPLS

- A-B carries 1 copy



DRAWBACKS

VPLS has ingress replication

H-VPLS solves part of the problem

- First hop/link no longer abused
- But that comes at a cost:
  - More nodes need to learn MAC addresses
  - PE-r single-point of failure
- More layers are possible
  - Increasing MAC spread & complexity of operation

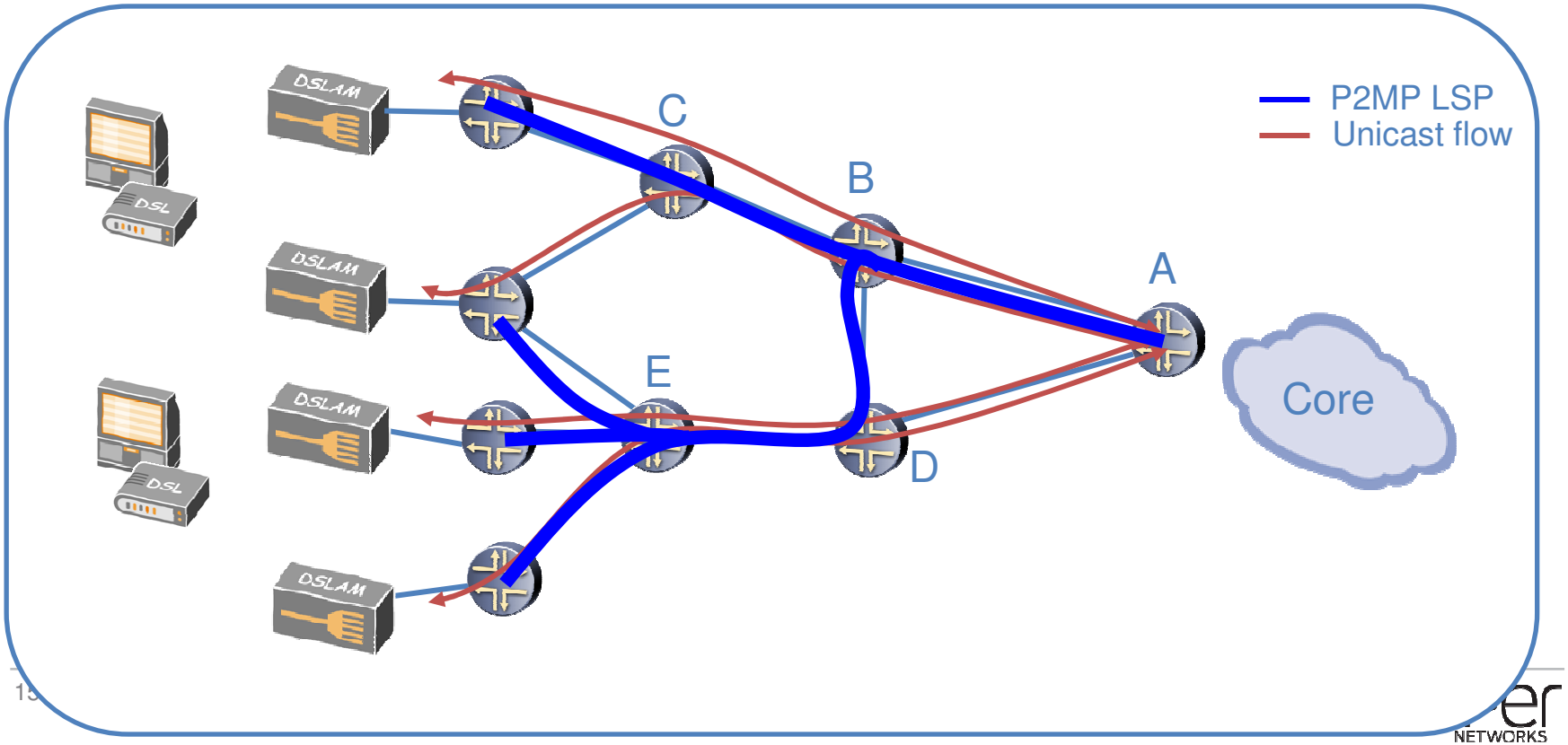
H-VPLS does not solve all

- Manual configuration
- Further replication down-stream possible

EXAMPLE SOLUTION
RFC 4761 VPLS & P2MP LSPS

Unicast traffic uses normal labels & LSPs

Broadcast traffic uses a P2MP LSP



VPLS + P2MP : ADVANTAGES

In general: solve a forwarding plane problem in the forwarding plane

Replication options, technically feasible:

- MPLS RSVP-TE P2MP LSP
- MPLS mLDP

Benefits of RSVP-TE:

- Selectable per ingress PE
- One tree per VPLS per ingress PE
- Automatically setup & update of LSP structure (leafs)
- Strict traffic engineering
- P2MP uses MPLS operations: does not require learning MACs

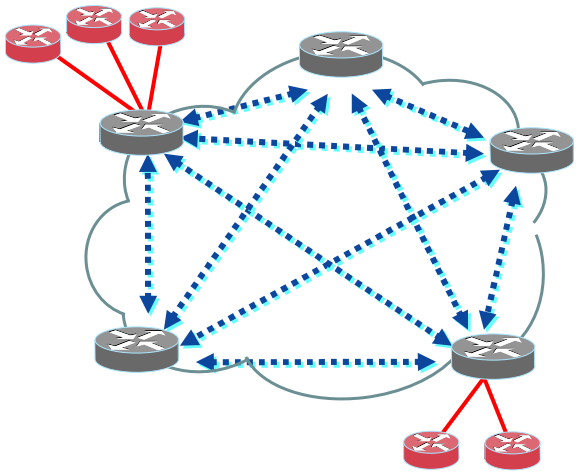
SCALABILITY & INTERWORKING LDP & BGP VPLS

Control Plane

Characteristic	LDP VPLS	BGP VPLS
Full-mesh requirement	Alleviated only somewhat by H-VPLS, though at the expense of introducing changes and additional overhead in the dataplane	Solved by the use of BGP Route Reflector hierarchy
Provisioning task of adding or removing PE router	Only somewhat simpllified by H-VPLS	Highly simplified by use of Route Reflector
Provisioning task of adding or changing VPLS cutomer sites	Manual or through provisioning	Automated by BGP'd autodiscovery
VPLS with P2MP LSP integration to scale forwarding and data planes	Currently not supported in any commercial implementation	Supported in standards and currently implemented
Signaling overhead	Increases in proportion to the total number of PWs in the network	Minimal because each signaling update can be used to establish multiple PWs

Single Domain VPLS Scalability

- Suitable for simple/small implementations
- Full mesh of PWs required (for both LDP and BGP based)
  - N*(N-1)/2 Pseudo Wires
  - For LDP based full mesh of directed LDP sessions required
  - For LDP based manual configuration & provisioning issues
- No hierarchical scalability
- Potential packet replication overhead
  - PE-to-PE flooding used to use ingress replication of unknowns, broadcast and multicast
  - CPU overhead for replication



INTER-AS VPLS OPTIONS

Re-cap on inter-connection options:

- Option A : plain Ethernet
- Option B : PWs from all PEs to all PEs, ASBRs have a control-plane info for each VPLS
- Option C : PWs from all PEs to al PEs, RRs distribute reachability between Ases

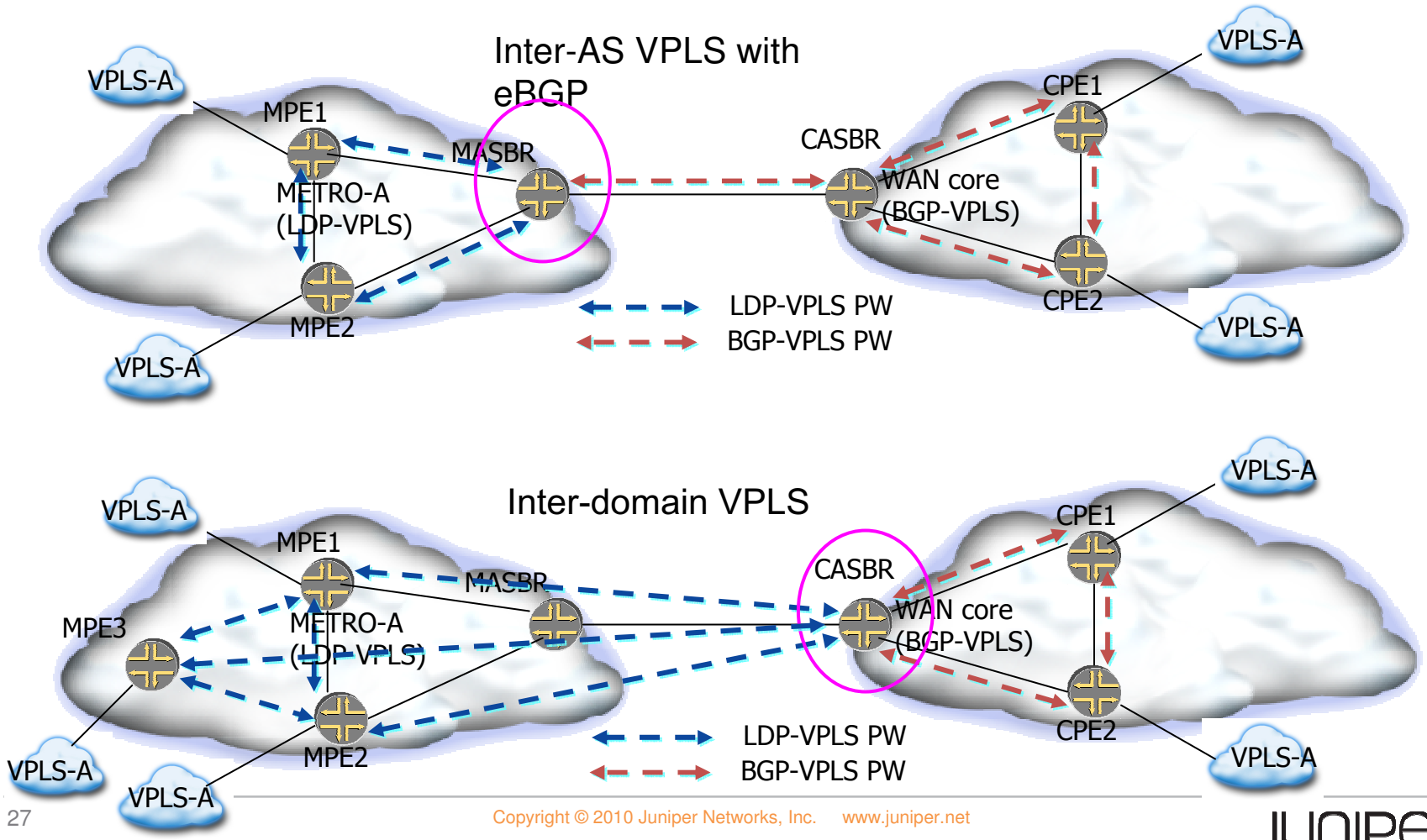
Option E

- Similar to Option B
- Uses MAC table instead of stitching
- Reduces the amount of PWs between two ASes to 1

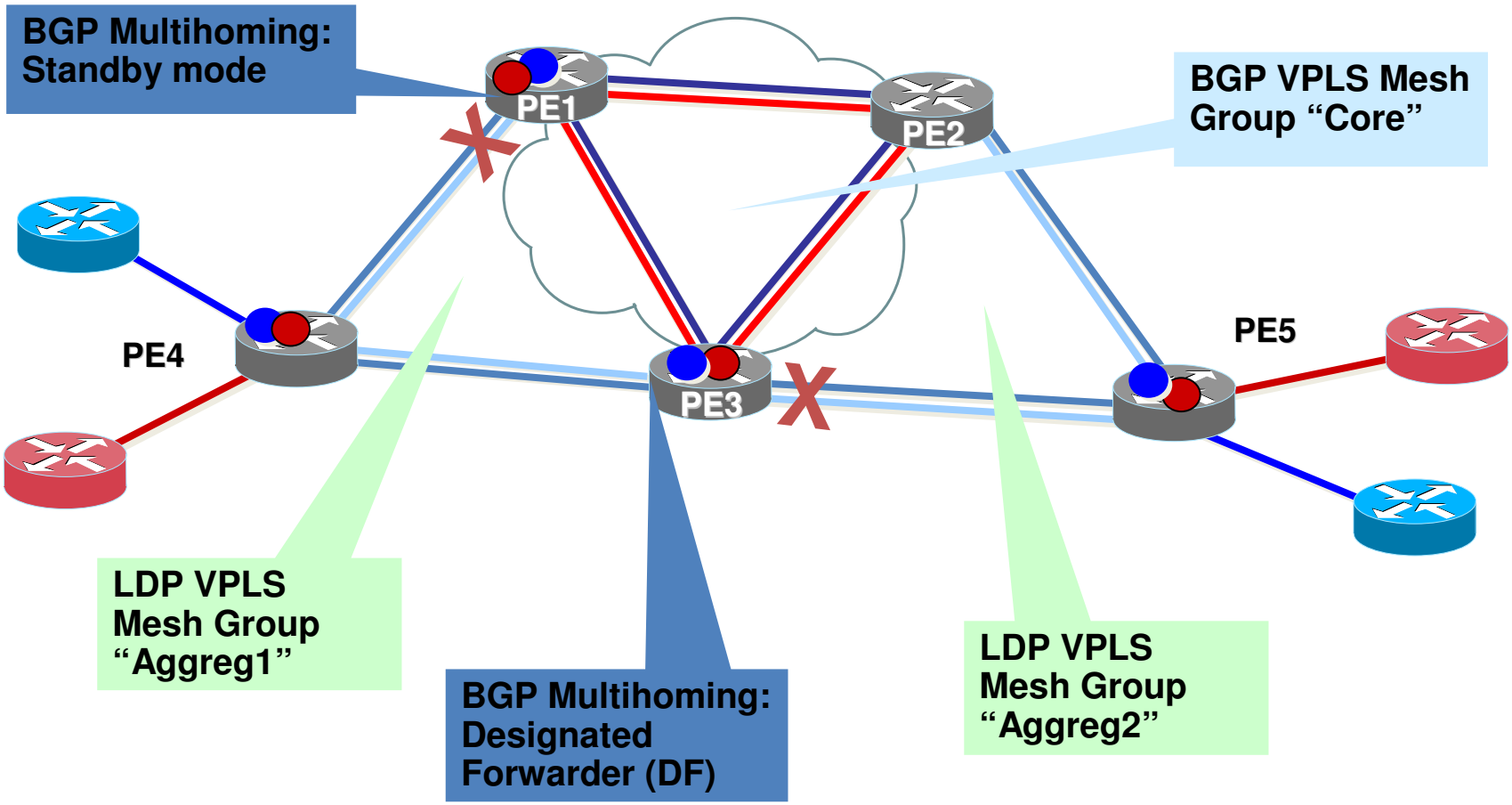
Building block: Mesh-groups

- Control flooding in a flexible way

INTER-DOMAIN LDP-BGP VPLS: 2 SUB-CASES



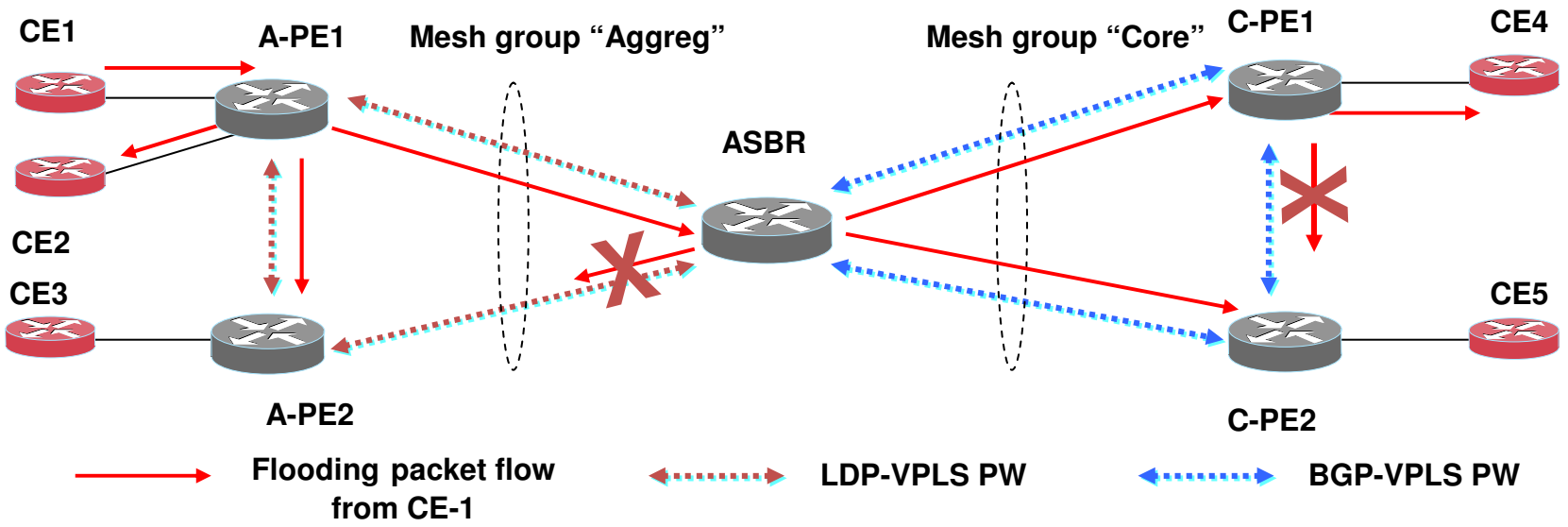
Core BGP VPLS interconnected to Aggregation LDP VPLS



Multi Domain Split Horizont

- Split-horizon is much more complex now:
  - PE1 must **not** forward packets received from metro 1 back to metro 1, or from metro 2 to metro 2, or from core to core
  - But forwards in case of different mesh groups
- ASBR (autonomous-system-boundary) PE router hosts more than one mesh groups (any combination of LDP and BGP)
- For a multi-homed site, a PE can be a designated forwarder only if the site hosted on it is operationally UP

Mesh Groups: Flooding example



- Assume CE-1 sends a broadcast ARP request packet. It will be flooded by A-PE1, to all PEs in "Aggreg" domain (including ASBR) as it's fully-meshed.
- ASBR receives this packet from A-PE1 and forwards it to all the mesh-groups (PWs) except the one in which packet is received, as a result packet is forwarded on all PWs part of mesh-group "Core".
- Upon receiving this packet from A-PE1, ASBR learns the CE-1 MAC address via A-PE1 PW.

VPLS MULTIHOMING

VPLS PE Site redundancy

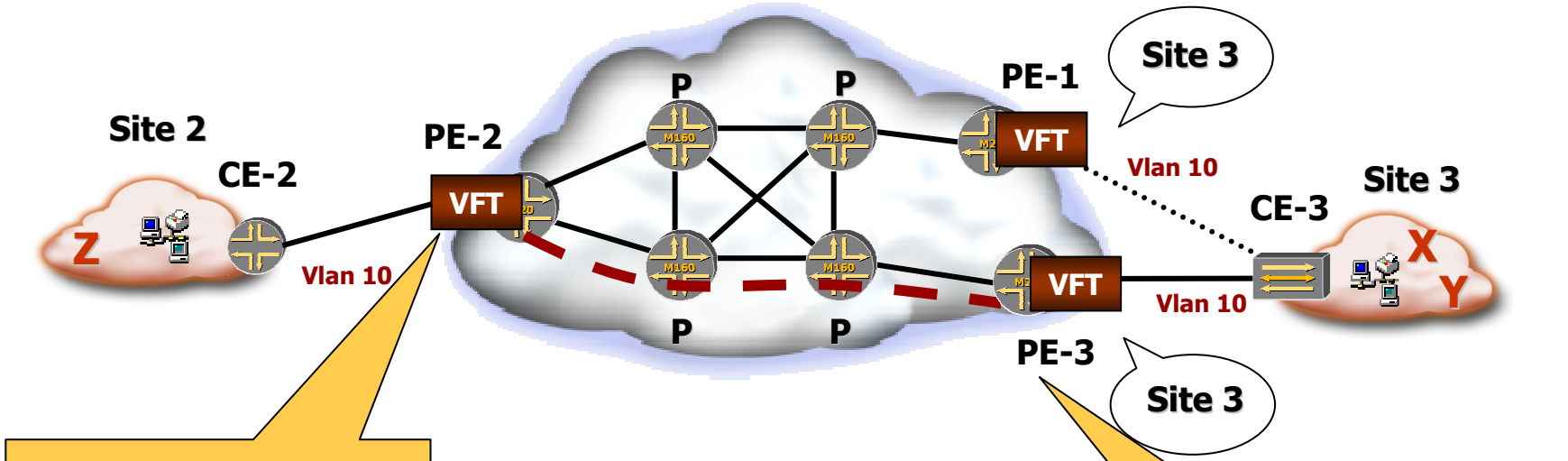
Multiple Options:

- BGP Multi-homing
- Interworking STP with root-protect
- MC-LAG
- PW Redundancy (for H-VPLS topologies in LDP VPLS)
- Using CFM (Ethernet OAM) on access ring

BGP VPLS Multi-homing – Solution Outline

- A CE device that is multi-homed to multiple PEs is given the same site ID on all those PEs
  - If desired, one can set the Local Preference on these PEs to control BGP path selection
- The algorithm essentially selects the VE that originated the "best" advertisement with a particular site ID as the designated forwarder
  - BGP path selection is used
  - IGP metric is not part of the selection process

VPLS Multi-homing - VE Roles

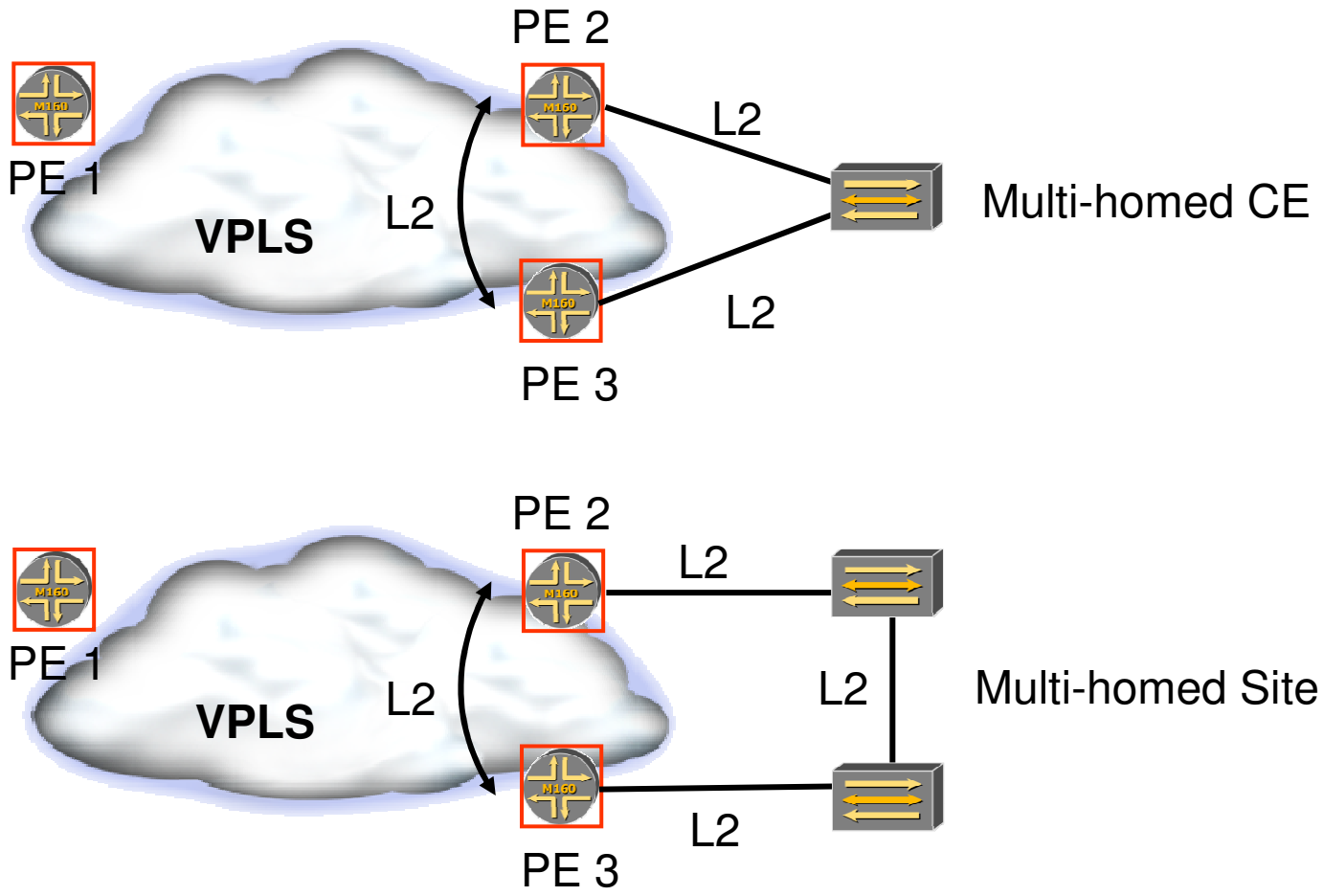


Step 2: All PEs run BGP path selection, and choose PE-3 as D-VE for site 3

Roles:
for site 3, PE-3 is **designated VE**, PE-1 is **non-designated VE**, and PE-2 is the **remote VE**

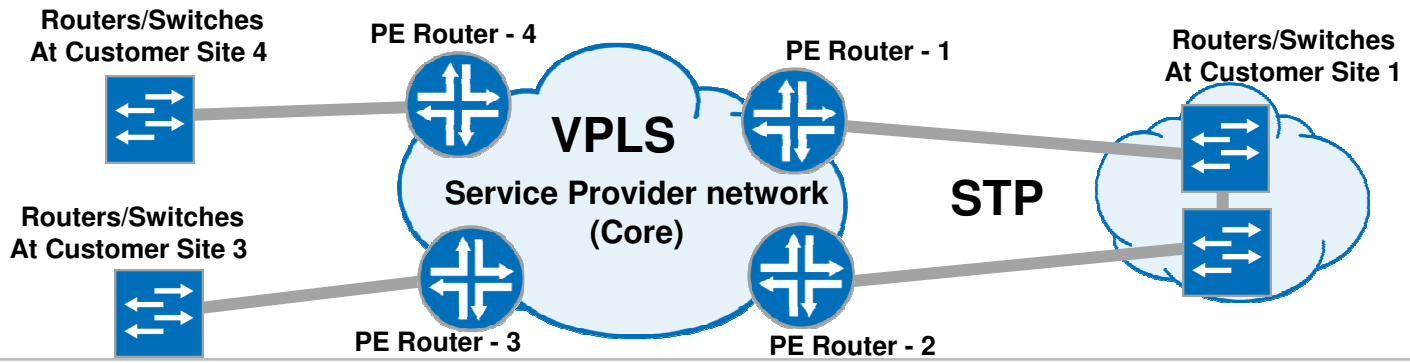
Step 1: Both PE-1 and PE-3 have same site ID (3) for CE-3
PE-3 has better Local Preference

LOOP SCENARIOS



VPLS Multihoming interaction with STP

VPLS and STP domain pass info about topology changes between domains, and update other nodes connected only to VPLS or only to STP domain.

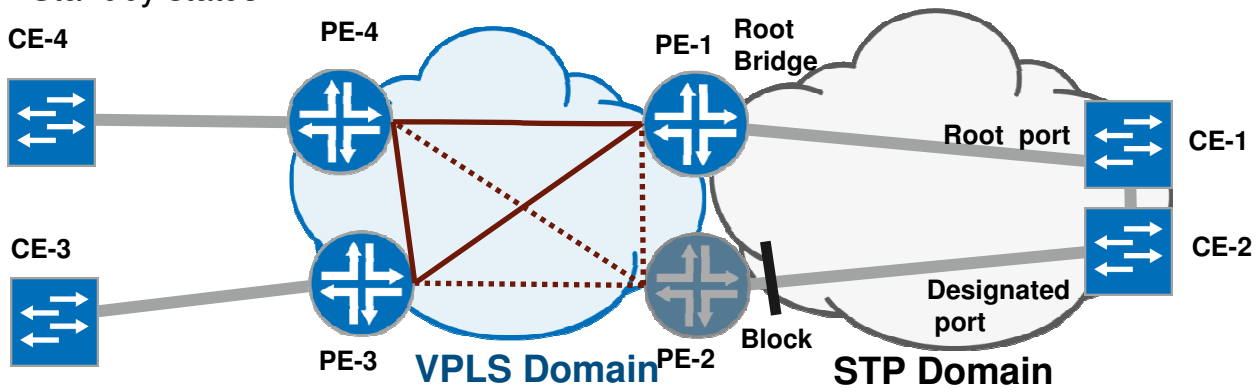


HOW DOES IT WORK UNDER NORMAL CONDITIONS?

Under regular circumstances - all links & nodes are active

1. PE-1 and PE-2 participate in both VPLS domain and STP domain.
2. CE-1 and CE-2 participate only in STP domain.
3. STP works between PE-1, PE-2 and CE-1, CE-2 but it doesn't extend in VPLS domain.
4. STP configured on PEs controls connectivity of CE network to VPLS domain by setting appropriate Bridge Priority.
5. All PE routers connected to STP domain have better Bridge priority than any CE router. If two or more PEs 'sees' each other in STP domain, only one of them will connect STP domain to VPLS domain, other PEs will block.
6. PEs have Pseudo Wires (PW) to all other peers (mesh) and signal them as Up or Standby according to their status in STP domain. If PE is blocking its port in STP domain, it will signal Standby status for PWs (dashed line on diagram).

In example depicted below, PE-1 is the Root Bridge. Root-protect feature on PE-2 blocks traffic on interface toward CE-2 and prevents network loop – as long as PE-2 receives Root Bridge BPDUs on that interface. PWs for PE-2 are in Standby status.



ABOUT SUCCESSFUL RECOVERY

For network to successfully recover from failure and restore events, it is important to:

A. Re-create loop-free L2 network topology – by blocking some interfaces,

B. Communicate updates (mac-flush) for MAC filtering/forwarding tables to avoid traffic black-holing – by sending

   I. STP topology change messages (TC), and

   II. VPLS messages (TLV)

.

HOW DOES IT WORK – AFTER CE-CE LINK FAILURE?

After link between two CE routers fail (CE-1 --- CE-2) :

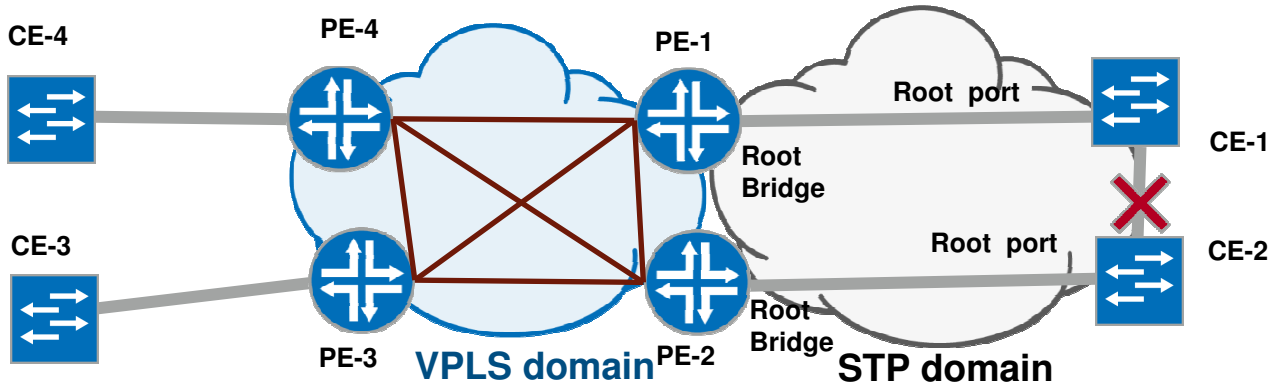
In STP domain:

1. TC (Topology Change) communicated in STP domain.
2. PE-2 becomes Root Bridge in it's network partition – because it has superior bridge priority than CE-2.
3. PE-1 stays Root Bridge in it's network partition – because it has superior bridge priority than CE-1.
4. PE-2 unblocks link to CE-2 – because there are no more superior BPDUs coming from CE-2
5. CE-1, CE-2 flush L2 forwarding tables, starts learning MAC addresses again and flooding traffic for unknown destinations.

In VPLS domain:

1. PE-2 changes VPLS PWs in Up status (from Standby) – because local port is not blocked anymore
2. PE-1 keeps VPLS PWs in Up status
3. PE-2 sends signal to VPLS peers to flush all MACs learned from PE-1 (previous RB).
4. Other VPLS peers (PE-3, PE-4) start using BOTH PE-1 and PE-2 for delivering traffic to (partitioned) Site-1

Note: PW between PE-1 and PE-2 is used to pass traffic between CE-1 and CE-2



SUMMARY

VPLS can be complex ☺

Best practise:

- VPLS BGP core
- VPLS LDP aggregation

VPLS Toolbox:

- Ingress replication with P2MP LSP
- H-VPLS and Full Mesh concept
- BGP-LDP VPLS interworking
- VPLS Multi-homing
- Interworking with native Ethernet xSTP access networks


everywhere