Sponzor komunikacijskih tehnologija

Hrvatski Telekom

Sponzori konferencije

SG Equipment Finance

SG Leasing d.o.o.

SOGELEASE

HRVATSKA UDRUGA POSLODAVACA

Komunikacijski partner

spona
Integrirane komunikacije

Partner digitalnog oglašavanja i video nadzora

RECRO net
We know IT

Partneri konferencije

COMPUTECH

computer systems

IBM

KING ICT

RECRO net

s&t

Sedam IT
Primjena informatičkih tehnologija

STORM

verso

Tehnološki sponzor

ComTel Corp.

EMC²

Novi Forum

SAP

Medijski pokrovitelji

vecernji.hr

ADRIA

MREŽA

racunalo.com

Večernji list

infoTrend

PC chip

Cisco Expo 2010

24.–26. ožujka 2010.
Hotel Le Méridien Lav
Split

Kolaboracija -
nova snaga poslovne suradnje

Cisco Expo 2010

Novi mehanizmi redundancije u Ethernet agregacijskim mrezama

Dejan Jaksic – Systems Engineer SP

djaksic@cisco.com

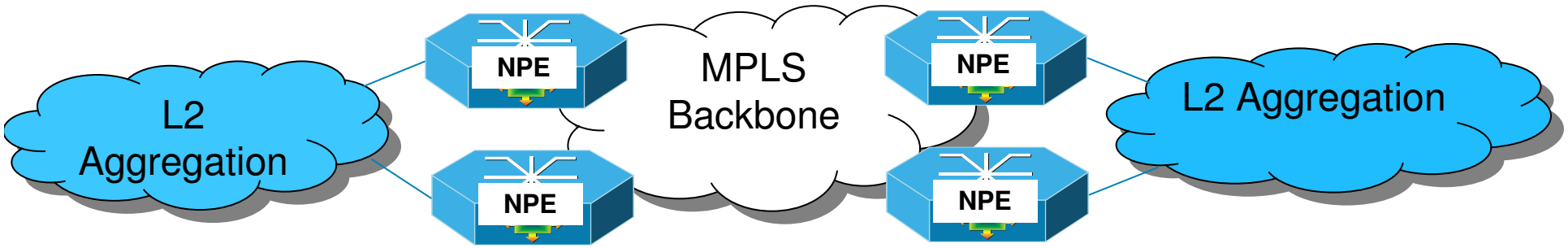
Carrier Ethernet Ring Topologies
Current Network Trends and Challenges

- Large Spanning Tree domains
  - Increasing number of nodes in the ring
- Supporting higher number of subscribers
  - Increasing number of VLANs, MAC addresses per L2 domain
- Carrier Ethernet Trend
  - Fast convergence requirement in the access and aggregation networks
  - Spanning Tree not perceived as Carrier Class
  - Ethernet and IP/MPLS based aggregation
- Complexity of management and troubleshooting as the network grows

L2 Convergence

Motivation

- Large L2 aggregation networks deployed due to cost and operational aspects
- Backbone networks are MPLS/IP based
- Objective: fully redundant end-to-end network. L2 aggregation networks are multi-homed into MPLS/IP backbone to achieve high availability



L2 Convergence

State of Art

- EVC L2 Convergence - SRD and before

  MST

  REP (switchport)

  PW Redundancy (MPLS aggregation)

  Flex-Link

  Etherchannel/LACP (single-homed devices)

  Interface-Backup (single-homed devices)

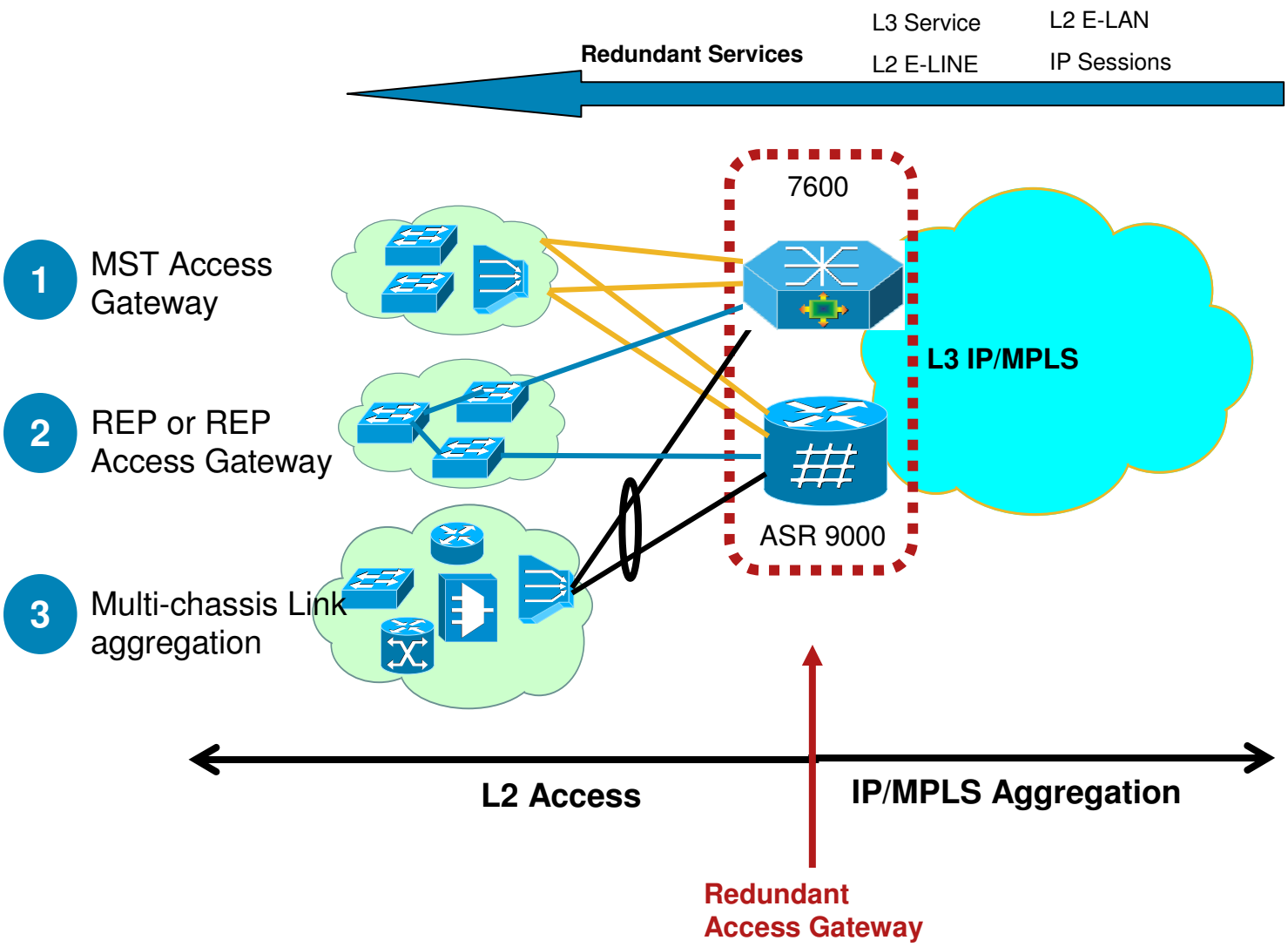
- New features in SRE

  MST AG – simplify MST based deployments

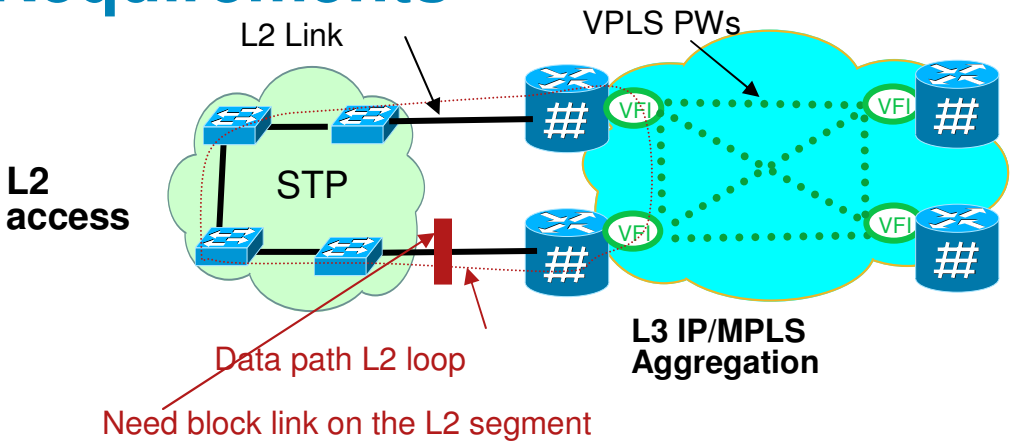
  REP – support for EVC added

  mLACP – dual homed LACP

Redundant Access Gateway Overview



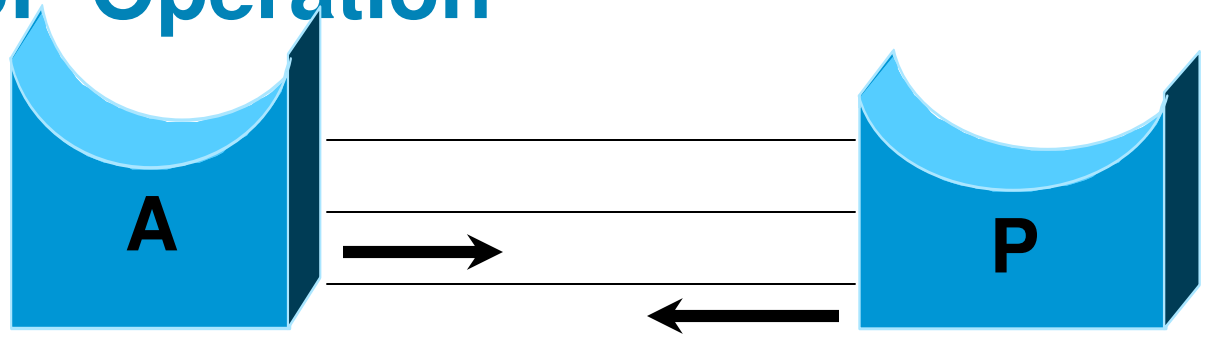
Access Redundancy – The Common Requirements



- **Fundamental Requirement – Block redundant access link – singe exit point from L2!**
  - –STP approach – any vendor and any topology, per VLAN load balancing, slow convergence
  - –REP – ring or spoke-and-hub topology, Cisco specific, per VLAN load balancing, 50 msec convergence
  - –Multi-chassis link aggregation – standard, spoke-and-hub topology only, sub-second convergence, active/standby mode (in current implementation)
- **TCN propagation – MAC flush**
  - –Be able to propagate TCN from access network into MPLS/VPLS
  - –Be able to monitor MPLS network topology change, and feed back into access network. This is enhanced requirement to avoid double failure
- **Access network isolation or separation**
  - –Independent access domain, isolated from each other

mLACP

LACP Operation



- IEEE 802.3ad for link bundling for increasing the bandwidth and also providing redundancy
- LACP BPDUs (01:80:C2:00:00:00) are exchanged on Each Link
- System Attributes
  - Priority
  - MAC Address
- Port Attributes
  - Key
  - Priority
  - Number
  - State

LACP Operation

- **Selection Logic**
  - Each Port is assigned a Key
  - The Actor and the Partner Receive Per-Port Keys in LACP BPDUs.
  - If a Particular Port Key matches the Aggregator Key, the port is part of that Aggregator.
  - In case of Hot standby Links with higher Port-Priorities are bundled as Active links and rest of them are Bundled as Standby Links.
  - “lacp max-links” configuration determines number of active links in a bundle.

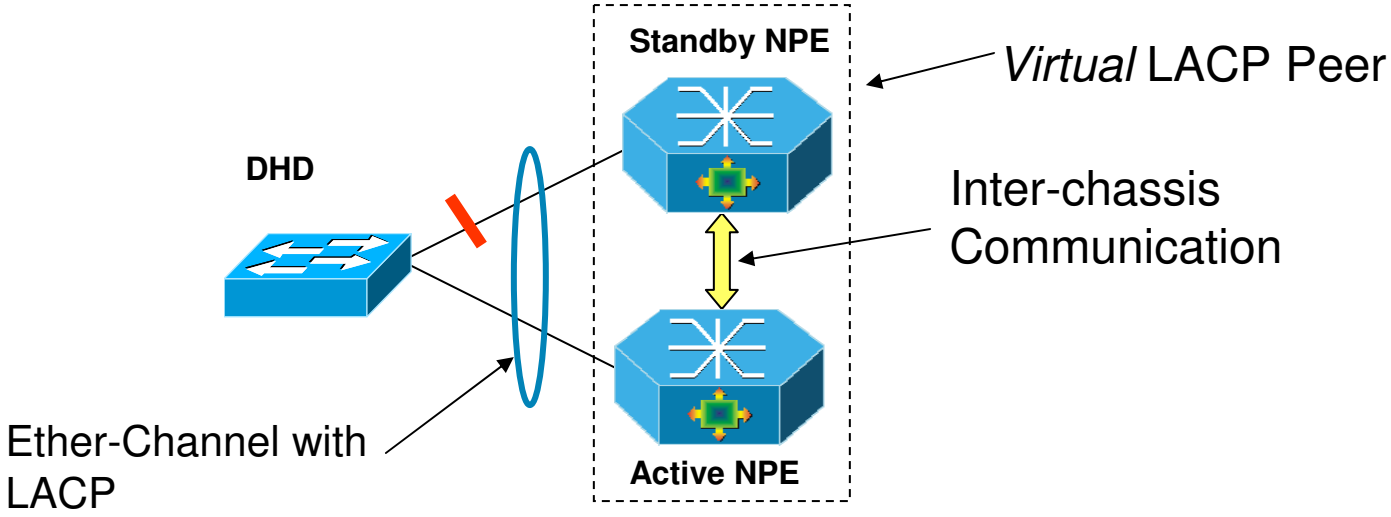
- **LACP State Machine**
  - Detached->Wait->Attached->Collecting_Distribution
  - Attached -> member links are attached to the Aggregator
  - Wait -> During Addition/Deletion of member links or Hot Standby
  - Collecting_Distribution -> Actor and Partner are in Sync and ready for packet forwarding

- **Steady-state Operation (Hot-Standby):**
  - Actor : L1 → COLLECTING_DISTRIBUTING (L1 SELECTED, send LACP IN_SYNC)
    L2 → WAITING (L2 STANDBY, send LACP OUT_OF_SYNC)
  - Partner : L1 → COLLECTING_DISTRIBUTING (L1 SELECTED, send LACP IN_SYNC)
    L2 → WAITING (L2 STANDBY, send LACP OUT_OF_SYNC)

LACP Extensions to support Multi-Chassis

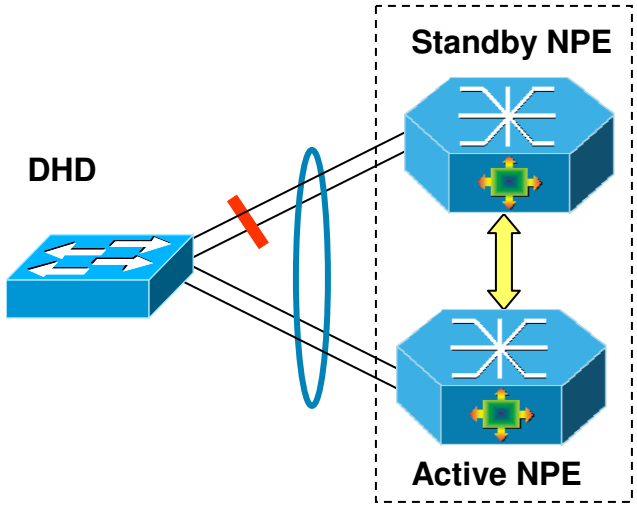
Attribute	Single Chassis	Multi-Chassis
System Mac-address*	ROM Backplane Mac-Address by default. Not User Configurable	Need to synchronize across systems so that DHD is presented with the same system-ID. Make it user configurable
System Priority	User Configurable. Default -32768	Need to synchronize across systems. Pick the lowest priority configured as system priority
Port Key	Operational or Admin Keys per port	No checks. Operator needs to make sure that keys are same on both the systems.
Port Number	Ifindex local to the router	Ifindex. Need to synchronize between systems
Port Priority	User Configurable Default - 32768	Operator needs to make sure configurations are the same across the chassis

Multi-chassis LACP concept



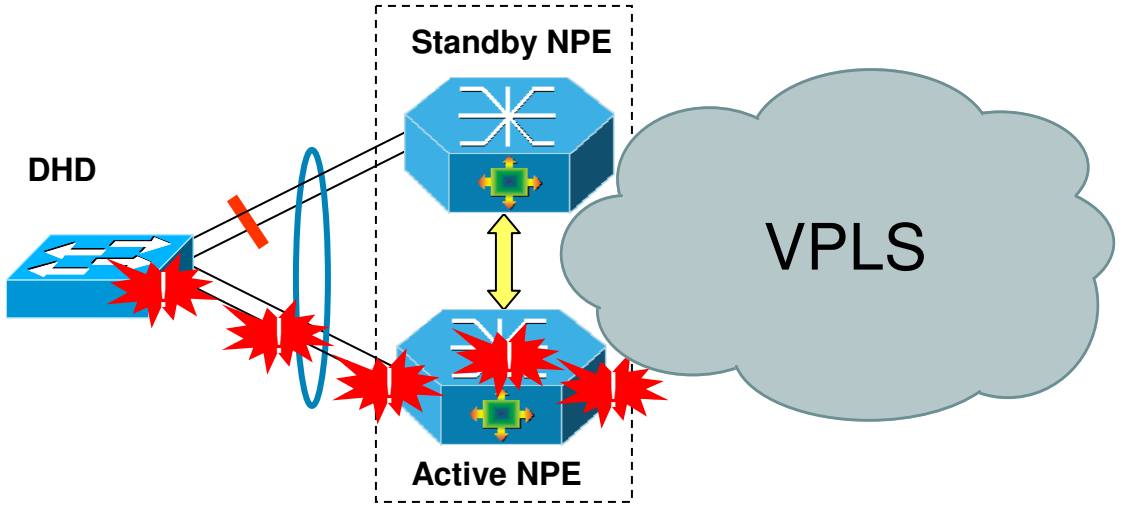
- Dual-home Access Node (DSLAM, CPE) to two N-PEs using Etherchannel LACP
  - Active/Standby operation
  - CPE/DSLAM runs standard 802.3ad LAG with LACP.
- Dual-homed Device (DHD) should 'think' it is connected to single device and run LACP.

Multi-chassis LACP function



- Between DHD and N-PE - Regular LACP function under NPE, DHD or shared control
- Between N-PE's - Redundancy Protocol for active/standby device selection, inter-chassis LACP state communication
- Operation modes – revertive, non-revertive
- mLACP can bundle up to 8 links to each N-PE

mLACP failure scenarios



- A: DHD Uplink Port Failure
- B: DHD Uplink Failure
- C: Active NPE Downlink Port Failure
- D: Active NPE Node Failure
- E: Active NPE Uplink Failure (Core Isolation)

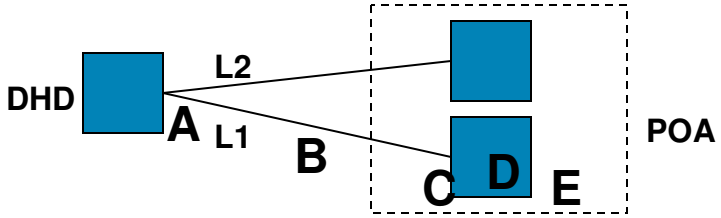
DHD/POA Control Comparison

Control	PROS	CONS
DHD	-No split brain problem	-Control not with SP. -Mis-configuration at the customer side can cause Active-Active. - lacp fast-switchover may not be supported on customer device.
POA	- Control with SP - Faster convergence - No chance for mis-configuration	-Split brain can be a problem
DHD and POA	- Independent control - No split brain problem - No active-active problem	-Potential slower convergence because of LACP state machine

Failover Operations

Operation	Description
Revertive	- Active PoA changes once the fault is cleared For A,B,C,D failures - Enabled by default -Dynamic Port Management for Fault E (180 secs by default Is the minimum)
Non Revertive	- Enabled by configuring “lacp failover non-revertive” under Port-channel. - Reduces churn with mLACP with Pseudowires.
Brute Force	- For Maintenance/Operational purpose. Brings down active PoA links upon failover explicitly.

Revertive Mode



Priority L1* Higher than *Priority L2

- For failures A, B, C and D: revertive operation is guaranteed by virtue of standard LACP operation.

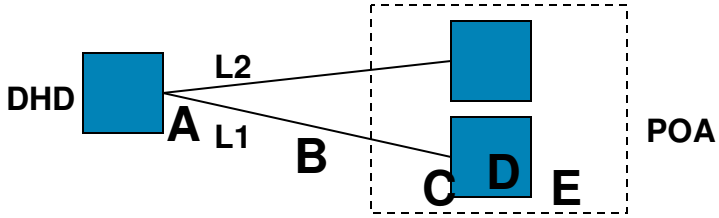
  No dynamic Port Priority management is required.

- For failure E – port-channel up:

  On failure, active PoA changes L1's priority lower than L2

  On clear, active PoA changes L2's priority lower than L1

Non-Revertive Mode



Priority L1* Higher than *Priority L2

- For failure E: non-revertive operation is guaranteed by virtue of dynamic Port Priority Management

  On failure, active PoA changes L1's priority lower than L2

  On clear, L1 continues to be STANDBY because it has lower priority

- For failures A, B, C:

  On failure, previously active PoA must change L1's priority lower than L2 to prevent L1 from resuming primary status on fault clear.

- For failure D:

  Option 1: Standby PoA (now taking over) should latch state to inform the other PoA (when it recovers) over ICC that it needs to bring down the priority on L1. Alternatively:

  Option 2: Standby PoA (now taking over) could increase priority of L2. ← This will slow down failover as L2 has to get unselected and then reselected. Hence Option 1 is preferred.

ICCP Overview

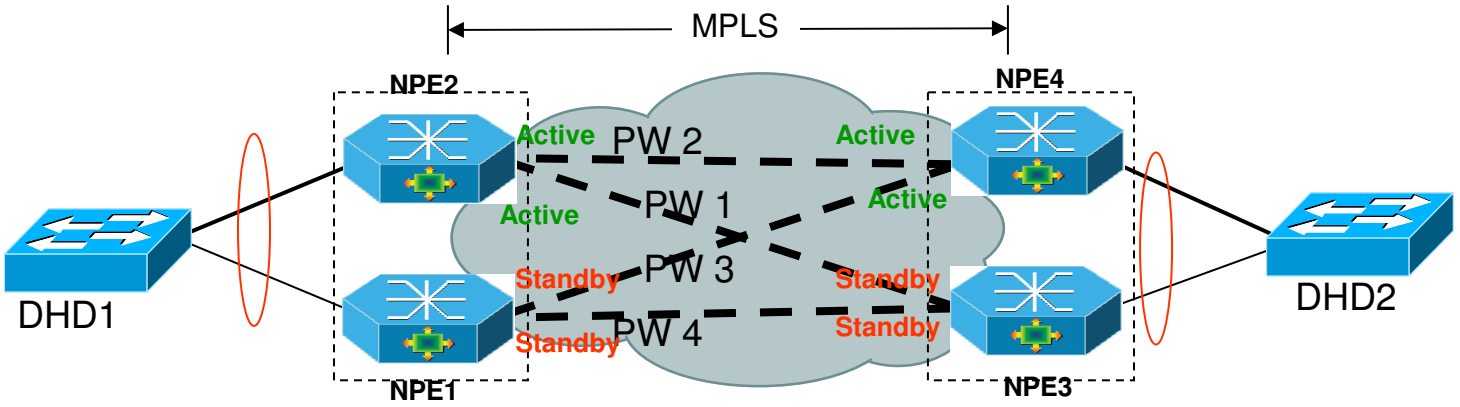
- ICCP is implemented according to the standard draft-martini-pwe3-iccp-00.txt
- ICCP is an extensible Protocol to synchronize event/states between multiple chassis which are part of the redundant group.
- ICCP is a reliable protocol which runs over TCP
- ICCP PDUs are exchanged between Peers to keep the application state consistent across Routers.
- Control Messages to setup, notify and exchange heartbeats.
- Data Messages to exchange the application state consistent across the chassis. Ex: LACP Parameters
- **ICCP failure detection**

  ICC Heartbeat ← Slow (~ 30 sec)

  /32 Next-hop Tracking ← Depends on IGP timers

  BFD ← ~50 – 150 msec

mLACP with VPWS/VPLS Integration



- Attachment Circuit "Active/Standby" state drives & is reflected in advertised PW state

When AC changes state to Active, both PWs will advertise Active

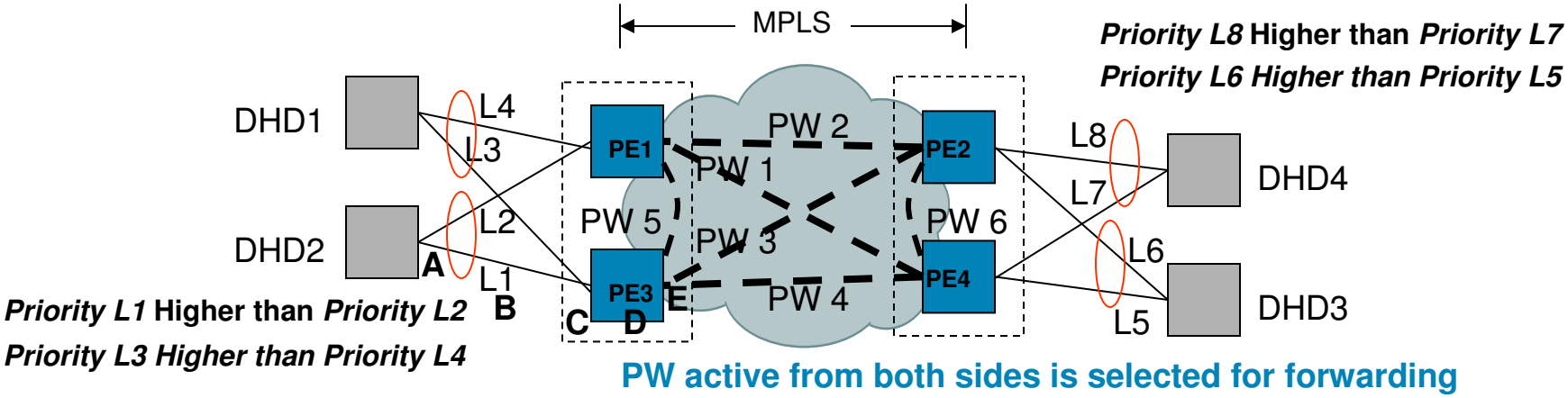
When AC changes state to Standby, both PWs will advertise Standby

- Two modes of operations are possible

Dual-homed/Coupled – AC state reflected into PW state

Decoupled – AC and PW state are independent

VPLS with two-sided PW Redundancy



PW active from both sides is selected for forwarding

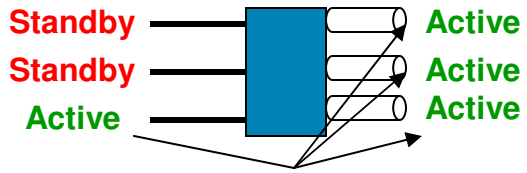
Operation (L1,L3 and L5,L7 are active links)	Active PWs	Standby PWs
coupled Mode (Default Operation)	PE3 : PW3,PW4,PW5 PE4: PW4,PW1,PW6	PE2: PW2,PW6,PW3 PE1:PW1,PW2,PW5
De-coupled Mode (Enabled by CLI) (platform vfi provision vlan <vlan-id")	PE3:PW3,PW5,PW4 PE4: PW4,PW1,PW6 PE2: PW2,PW6,PW3 PE1:PW1,PW2,PW5 (Faster Convergence but with flooding)	None
A,B,C faults (Coupled Mode) : Triggers mac-withdrawl with mac-flush	PE2: PW2,PW6,PW3 PE1:PW1,PW2,PW5	-
D,E (Coupled Mode) : Triggers mac-withdrawl with mac-flush	New active PE signals active on all PWs.	-

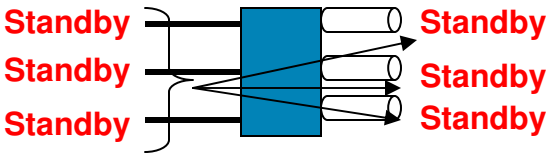
VPLS with mLACP

- VPLS Coupled Mode:

  When at least 1 AC in VFI changes state to Active, all PWs in VFI will advertise Active

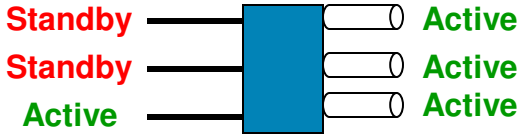
  When all ACs in VFI change state to Standby, all PWs in VFI will advertise Standby mode

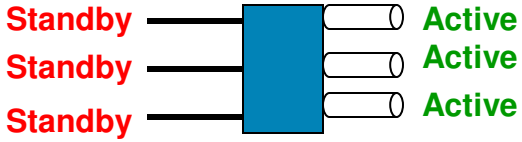




- VPLS Decoupled Mode:

  Have all the PWs in the VFI always active, and have the AC state independent of the PW state. This has the advantage of providing lower switchover time for the scenario where the platform doesn't support Hot Standby PW functionality, at the cost of extra flooding/multicast that will be dropped on the PE with standby ACs.





- H-VPLS (MPLS Access)

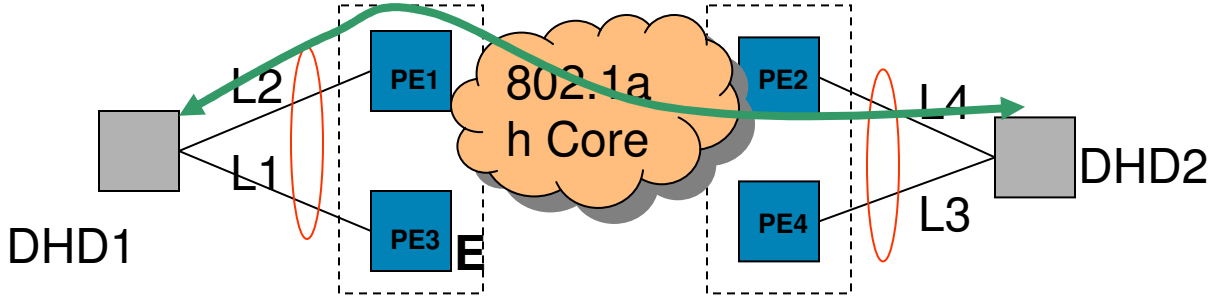
  uPE operation similar to mLACP with VPWS, or alternatively similar to today's 1-way PW redundancy.

  On nPE, all PWs in VFI are Active simultaneously, for both access & core PWs.

Pseudowire Redundancy in SRE

- VCCV over Primary and Backup PW
- Preferential Forwarding status bit according to draft-ietf-pwe3-redundancy-bit
- Upon Receipt of PW switchover status request, The receiver should clear the preferential status forwarding bit and activate the PW.
- Back up Pseudowire will be preprovisioned in the data plane.But forwarding is disabled.
- Supported with Scale EoMPLS configuration only.
- VPLS Redundancy is supported only with mLACP configuration.
- Supported on ES+,ES20 only.

mLACP with 802.1ah (MAC-in-MAC)



- Active PoA maintains a B-MAC to C-MAC mapping and the traffic flows through PE1 and PE2 in the above diagram.
- For A,B,C PE1 flushes its own mac-address table for particular I-SID.
- PE1 Also sends MIRP message to flush the mac-table mapping C-MAC to the B-MAC
- Traffic will be flooded until the C-MAC to B-MAC learning happens on the BEB.
- Configure “mac-flush notification mirp” under MTP Service Instance for Generating mac-flush in case of mLACP switchover.

mLACP CLI

- Interchassis Redundancy Group (2 PoA)

```
redundancy
 interchassis group <rg id>
  member ip <peer ip address>
  monitor peer [bfd | route-watch]
  mlacp node-id <node id>
  backbone interface <backbone if>
```

  - Pseudowire Class to reflect or decouple AC and PW status

```
pseudowire-class <class name>
 encapsulation mpls
 status peer topology dual-homed
```

mLACP CLI (cont)

- Port-Channel Definition

```
interface Port-channel1
 no ip address
 lacp fast-switchover
 lacp max-bundle 1
 mlacp interchassis group <rg id>
 service instance 1 ethernet
  encapsulation dot1q 100
  xconnect <peer ip> <vcid> pw-class <class name>
   backup peer <backup peer ip> <vcid> pw-class <class-name>
```

Solutions Comparison

	Highlights
REP (7600 only, EVC support in SRE release) REP AG (ASR9K only 3.9.1)	Sub 200msec convergence Good access ring isolation Not standard based Spoke-and-hub and ring topology, not works well for mesh network ASR9K interoperability with MST AG
Multi-chassis LAG (7600 support in 12.2SRE, ASR 9000 support target for 4.0)	Simple solution for spoke-and-hub topology, works for both bridging and non-bridging access device Standard based solution by using 802.3ad Sub-second convergence Phase 1 implement is active/standby mode with no per VLAN load balancing. Pseudo-mlacp supports active/active scheme with VLAN load balancing. Ring topologies under investigation.

Switchover Times

- with 4K VCs the switchover time is around 10-12 sec.
- Hot-Standby PW works in conjunction to mLACP for End-End VPWS.
- Per-VC Group PW signalling will reduce the switchover time. Coming soon

E-OAM Considerations

- CFM Port MEPs on DHD can be configured with mLACP and are transparent to the mLACP operation
- 802.3ah RFI when received PoA removes link from the Port-channel
- 802.3ah Error notifications exceeds the configurable limit PoA removes link from the port-channel
- E-LMI is not supported on LACP nor mLACP

Configuration : Best Practices

- Lacp fast switchover needs to be configured for better convergence.
- system-priority on DHD is lesser priority compared to POAs for PoA based control.
- mLACP with VPWS convergence can be improved by configuring LDP GR..
- Must have equal number of links going to each PoA for DHD control
- Max-bundle value must equal to the number of links connected locally to the PoA (no local intra PoA Act/Stby protection).
- Maximum number of mem-links per LAG is restricted to 8 just as in case of single chassis LAG.
- The command " lacp max-bundle <max-links> " should be configured all PoAs. Each PoA may be connected to DHD with varied number of member-links.
- The command " lacp min-bundle <min-links>" should be configured on PoAs.This is local to the box and brings down the port-channel if the number of active links is less than min-links.

Platform Specifics for mLACP

- mLACP is SSO Aware
- mLACP is only supported with EVC configuration.
  - No support for Subinterfaces, Access subinterfaces, Switchport configurations.
- mLACP is only supported with ES20,ES+
- mLACP is not supported with EVC Routed Pseudowire configuration
  - Exception is inter-POA Routed PW use case for VRRP/HSRP
- Number of chassis part of redundancy group is 2
- 802.1ah supports only 1 member link on the Port-Channel. mLACP configuration should be active-standby with 1 member link.
- ASR9K target for mLACP is in release 4.0

Resilient Ethernet Protocol - REP

Resilient Ethernet Protocol (REP)
What is REP?

- **A new protocol designed to provide a solution for fast and predicable layer 2 convergence for Carrier Ethernet networks**

- Fast and predictable convergence

  Convergence time: 50 to 250ms

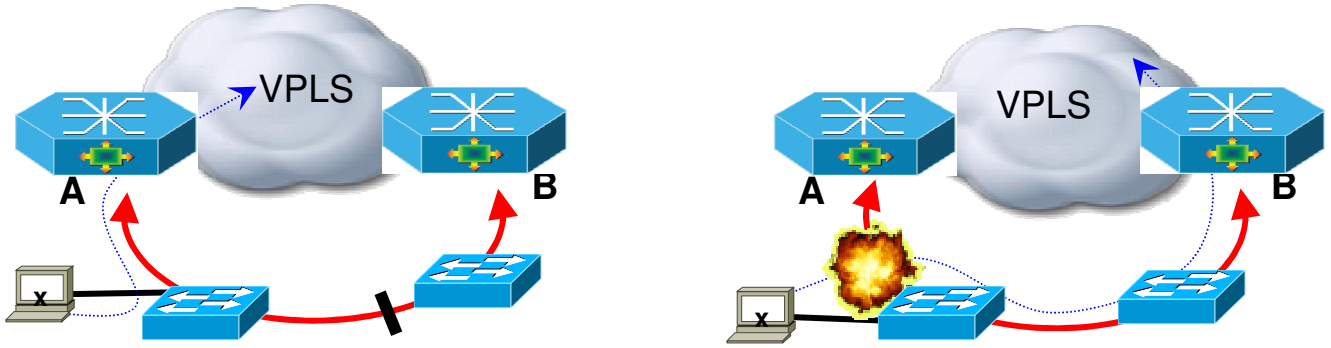
  Fast failure notification even in large rings with high number of node

  Manual configuration for predictable failover behavior

- Limit the scope of Spanning-tree

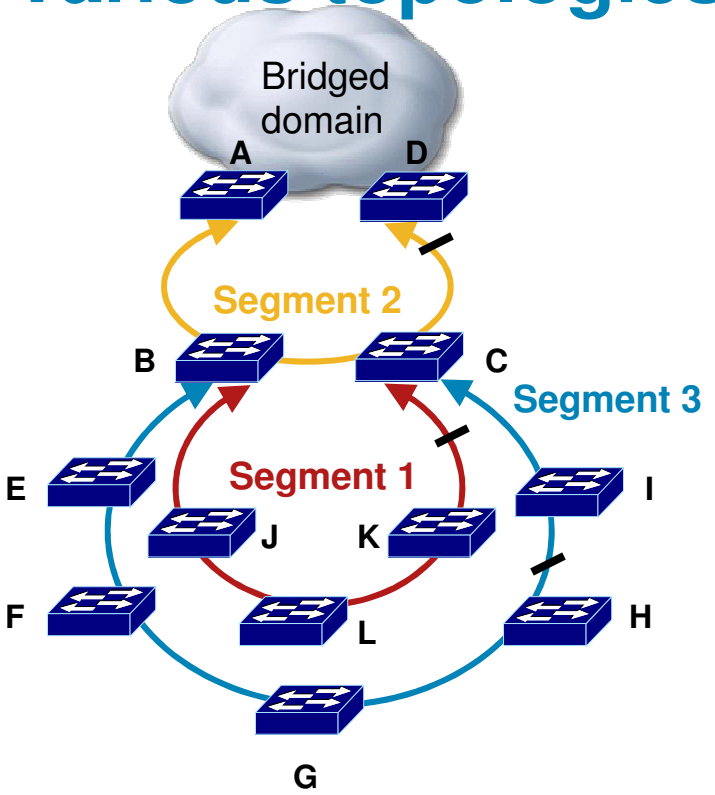
  STP is deactivated on REP interfaces

Resilient Ethernet Protocol (REP)
Its a Ring/Segment Protocol

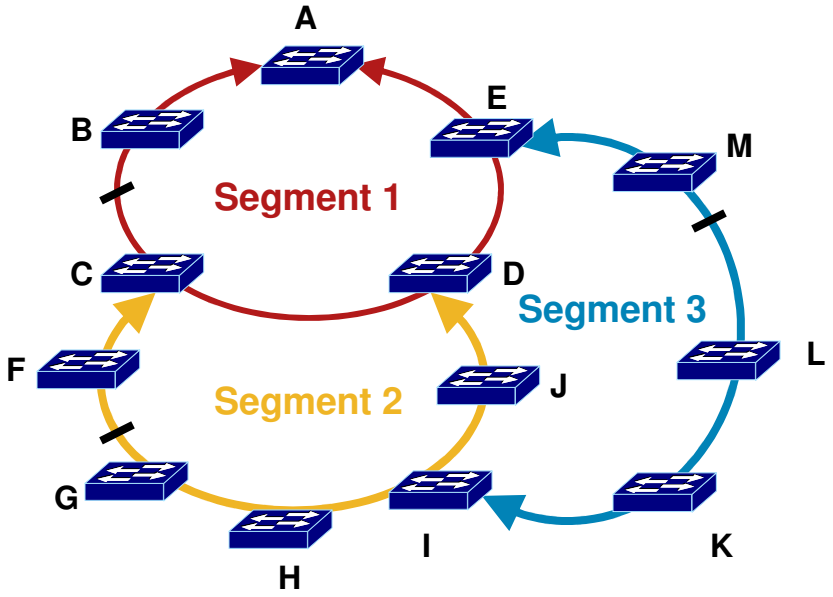


- REP is loop free L2 control protocol
- Fast predictable convergence (~200ms)
- REP segment comprises from Port A to Port B
  - Ports A and Port B are called as REP Edge Ports
  - Ports are explicitly configured to be part of a segment
- When all the links in the segment are operational, a blocked port is determined so that there is no connectivity between the edges A,B through the segment
- If a failure occurs within the segment, the blocked port goes forwarding
- **REP is supported on switchports in SRC and over EVCs in SRE**

REP Flexibility in supporting various topologies REP Flexibility in supporting various topologies



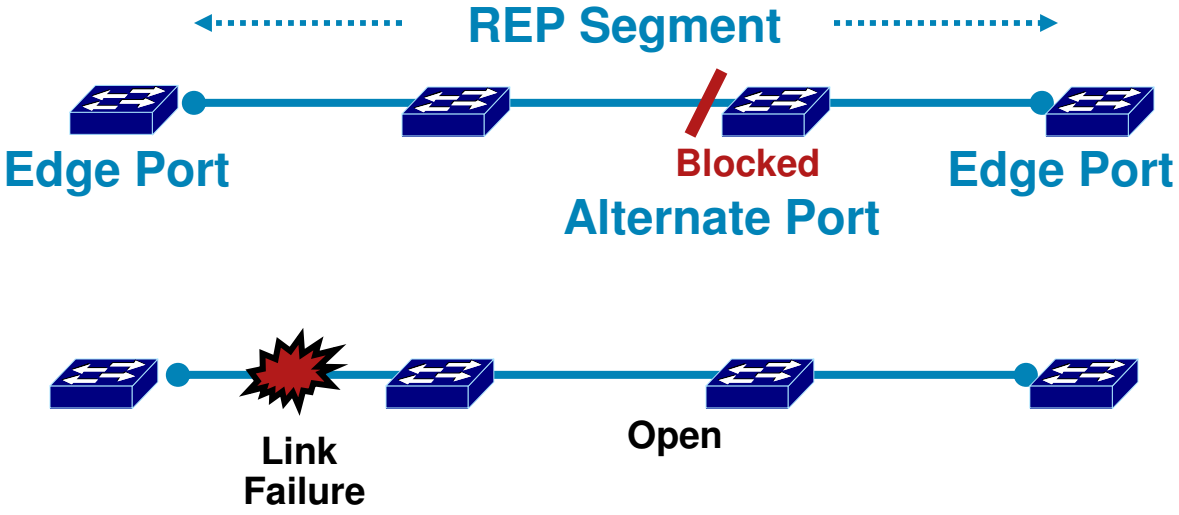
Access and Aggregation rings



Sub-tending rings terminated on different rings

Resilient Ethernet Protocol
A segment Protocol

- REP guarantees there is no connectivity between two edge ports on a segment.
- A REP segment is a chain of ports connected to each other and configured with a segment ID.
- When all interfaces in the segment are UP, the alternate port is blocking
- When a link or switch failure occurs on the segment, then blocked port goes forwarding



Resilient Ethernet Protocol
Fast link failure Notification

Fast Notification results in faster layer 2 MAC entries flushing on all nodes in the segment which results in faster network convergence times.

Fast Notification:

Flooded in hardware using Cisco Multicast address - No CPU involved

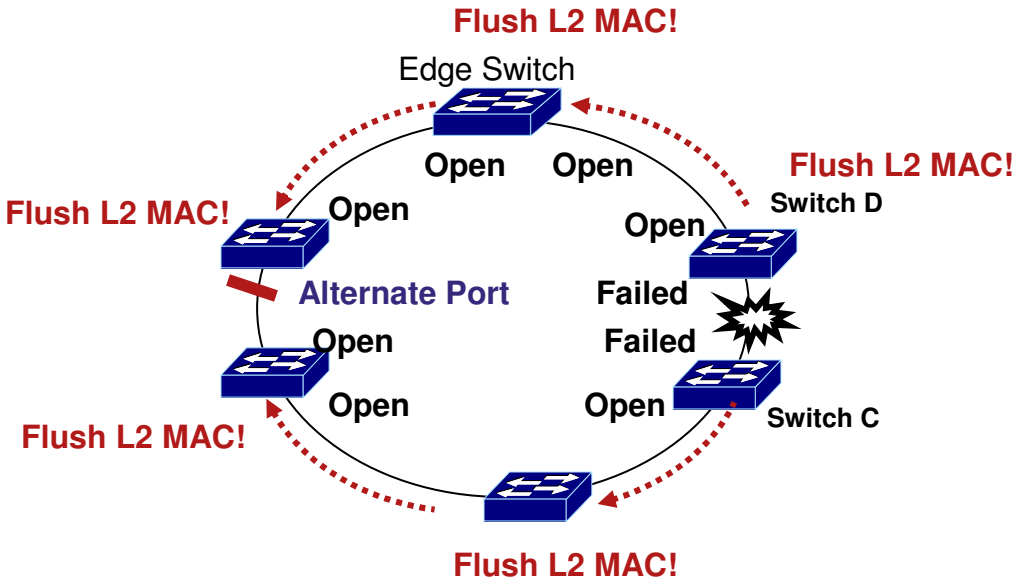
Results in fast Layer MAC address table flushing on each node of the network.

Configurable option to generate Spanning Tree TCN notifications (STCN).

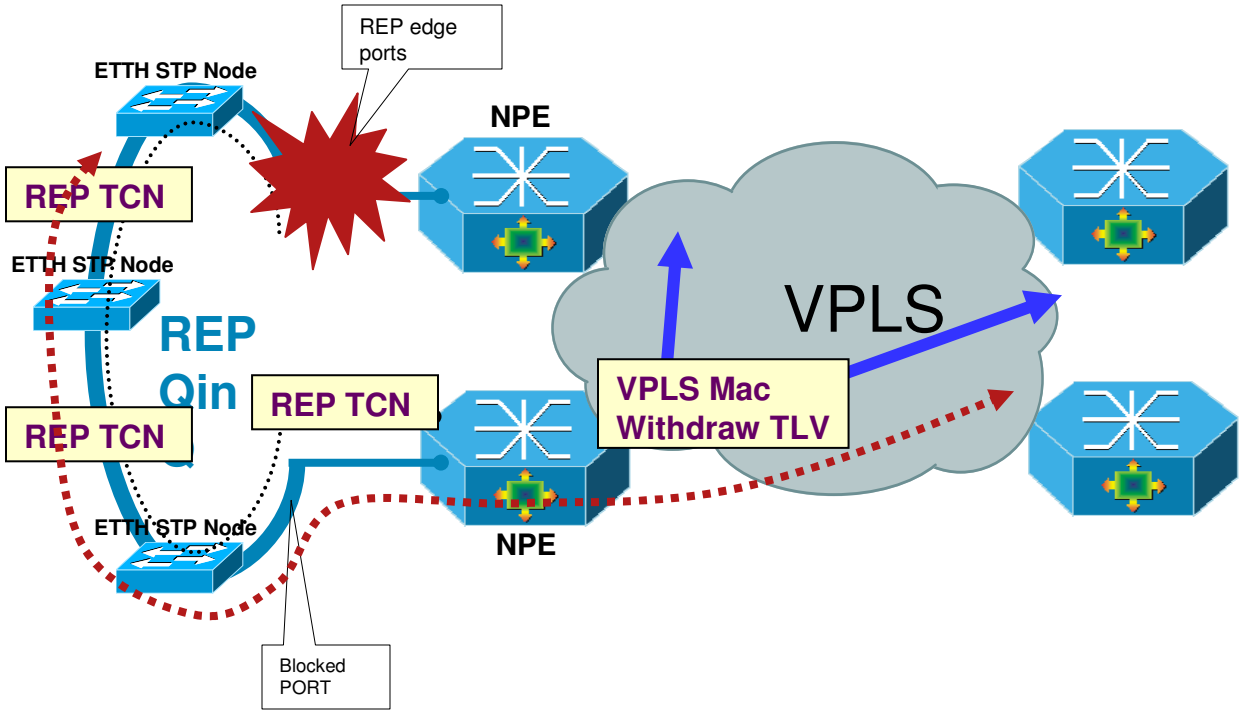
Configurable option to generate STCN on other REP segments

Configurable option to flush Mac addresses on the common link

On 7600, REP TCN triggers VPLS MAC withdrawal

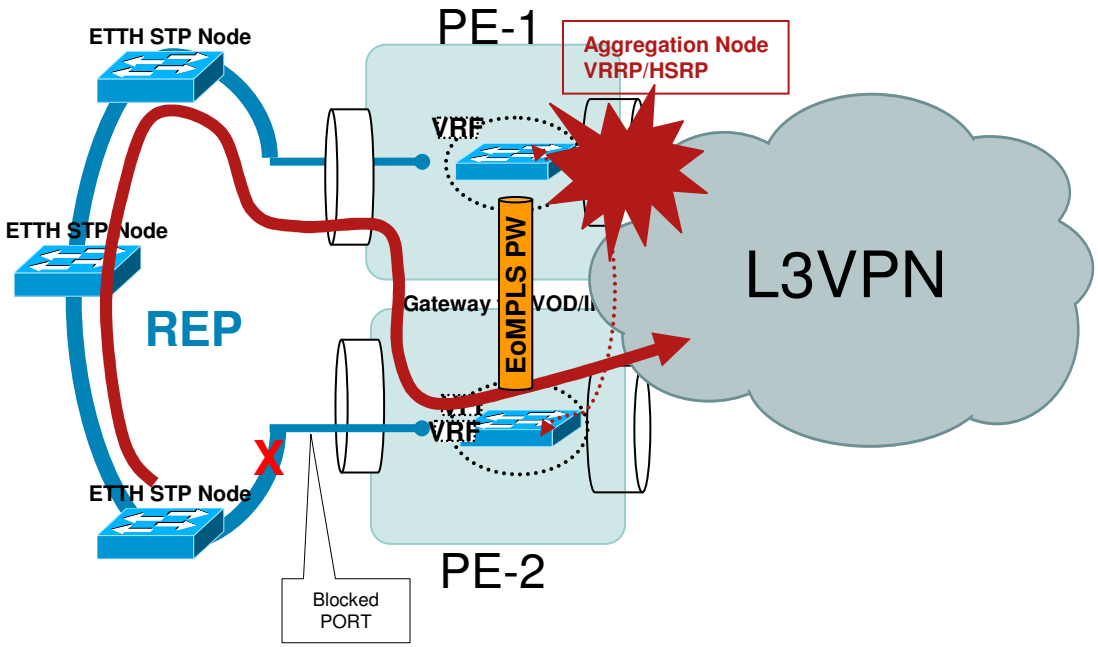


REP with VPLS



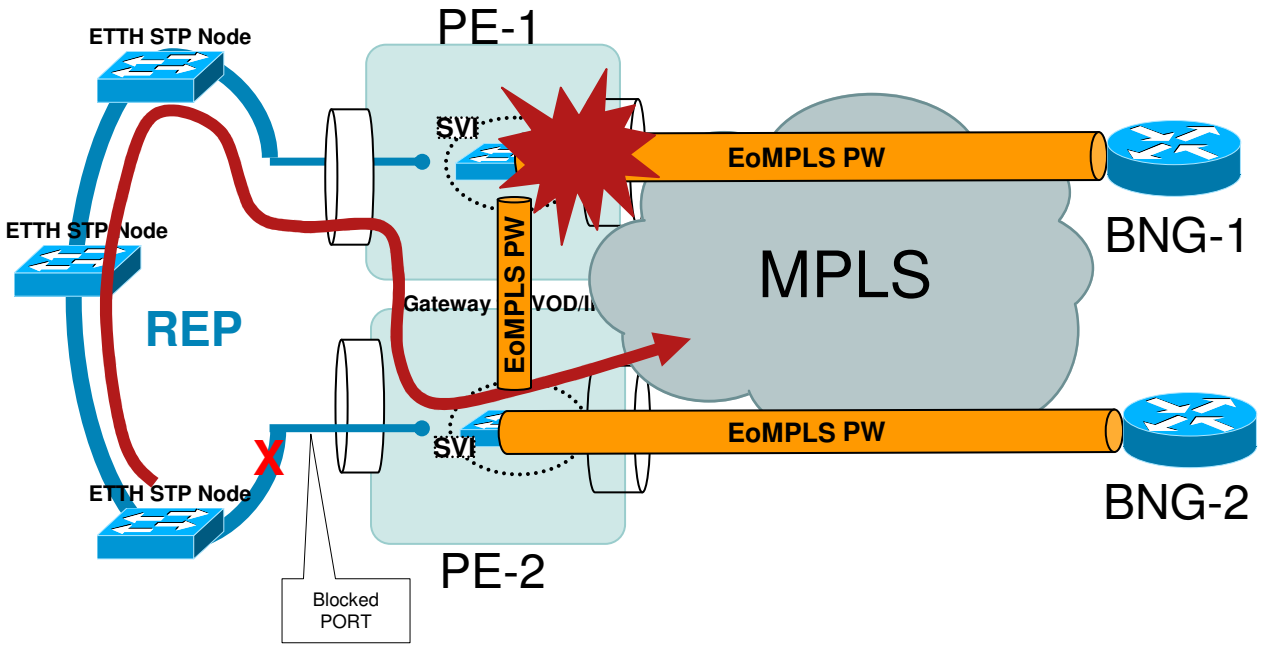
- REP ports are unblocked and access portion restored in sub-second time
- REP Access failure is propagated via REP TCN across the ring
- REP TCN triggers MAC withdrawal and the traffic can be quickly restored over VPLS domain

REP with L3/L3 VPN



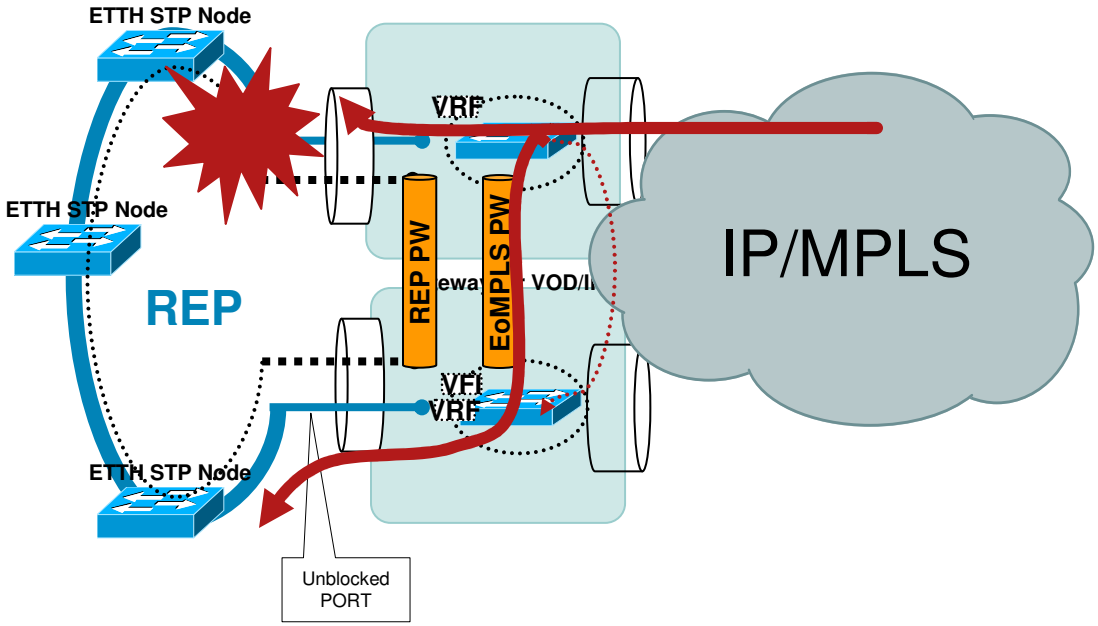
- Ring must be closed to avoid traffic black-hole in case of PE isolation from MPLS network or failure within the ring
- During the failure scenario, traffic flows on the data PW between the PE's
- VRRP/HSRP is required to provide GW resiliency
- Refer to Tahoma-Lippi for topology recommendations

REP with xconnect
Active/Active BNG example



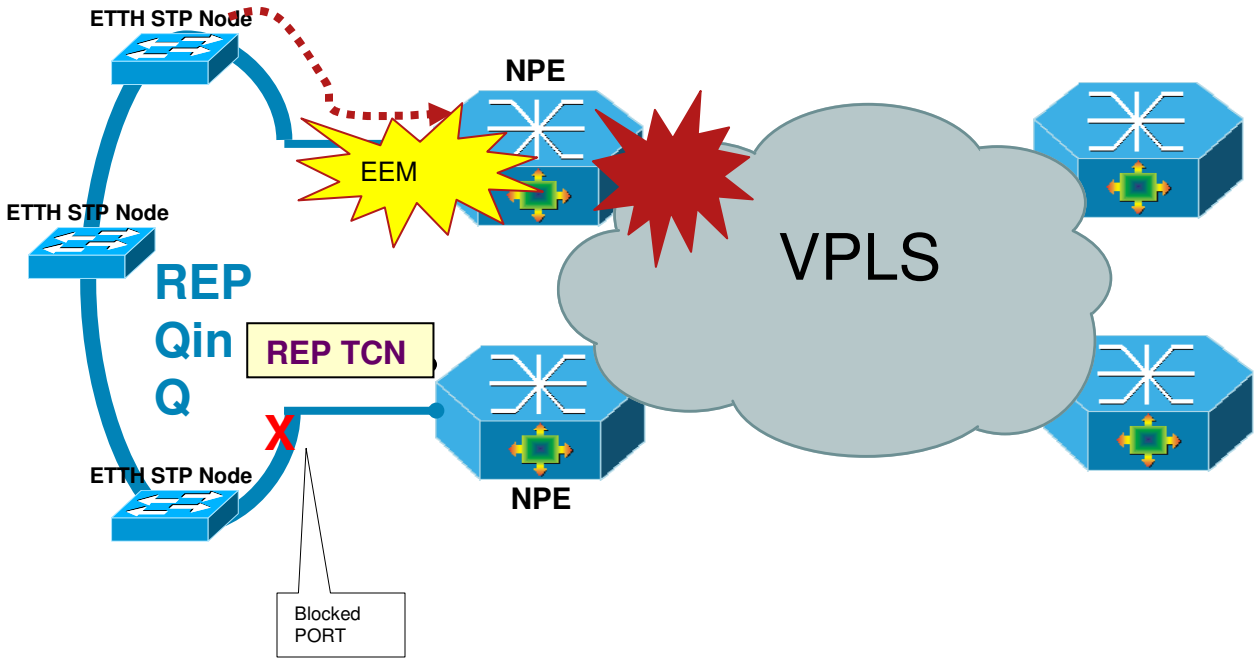
- Closed REP rings will be also required in case of EoMPLS when no loop exist. Special PW relies REP notifications while regular PW relies data traffic.
- Upon PW, Uplink or remote PE failure, the connectivity is restored over the second PE and PW.

REP with multicast



- REP TCN triggers multicast membership to be reprogrammed immediately on the alternate port. That allows immediate L2 multicast restoration on intermediate ports.
- **Caveat:** when the failure occurs on the TCN does not cause the multicast entries reprogramming on the **edge ports**. To restore multicast traffic, the join will need to be reissued. This may cause multi-second failure.
- Temporary workaround may consist of static IGMP group statements for most critical channels.
- Multicast enhancements are planned for Q4'2010

REP design considerations – Uplink Tracking



- Isolated PE case may lead to traffic blackhole even if VPLS core connectivity is available to PE-2
  - PE-1 REP SVI remains up therefore REP segment continues to forward to PE-1. It is unaware of the VC failures.
- EEM can be used to track PE isolation case and shut down redundant access ports
- Uplink tracking enhancement to detect uplink failures to avoid traffic blackhole and unblock REP alternate port are planned for Q4'2010

REP CLI

- Configure port as REP edge

```
interface GigabitEthernet1/31
 rep segment 100 edge <preferred>
 rep preempt delay 60
 service instance 3600 ethernet
  encapsulation dot1q 3600
  rewrite ingress tag pop 1 symmetric
  bridge-domain 3600
```

- Display REP topology

```
NPE-1#sh rep topology
REP Segment 100
BridgeName      PortName   Edge Role
--------------- ---------- ---- ----
NPE2            Gi3/15     Pri  Alt
UPE-2           Gi1/1           Open
UPE-2           Gi1/5           Open
UPE-1           Gi1/5           Open
UPE-1           Gi1/1           Open
NPE-1           Gi1/31     Sec  Open
```

Resilient Ethernet Protocol (REP)
What REP is not meant to do:

- REP Does not
  - Replace IEEE 802.17b RPR
    - Provide ring based QoS and Fairness across the ring
  - Replace Spanning Tree for complex layer 2 networks
  - Does not protect against dual failure in the ring

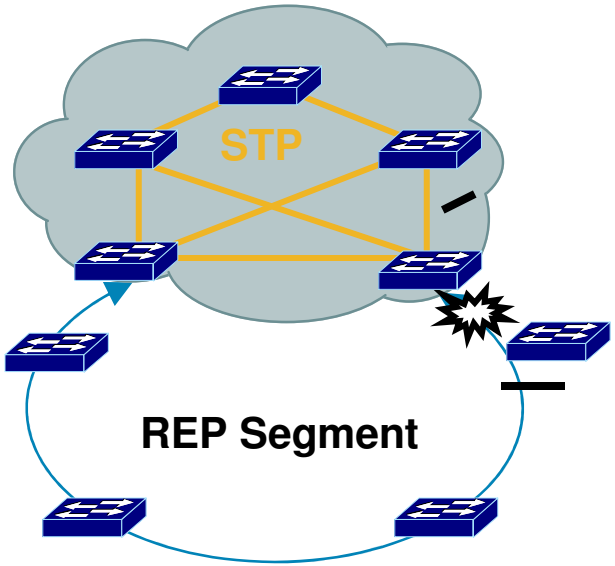
7600 REP on EVC Implementation Specifics

- REP over EVC supports Bridge-Domain, Xconnect and Local connect
  - Xconnect and local connect support must be enabled via CLI
  - R1(config-if)#ethernet vlan color-block all
- REP is not supported over service instances with encap untagged or defaultREP Load Balanced vlans are mapped to outer tags or range of outer tags in case of service instance with QinQ.
  - No CLI to block if load balancing vlans does not match the outer vlan range..
- EFP state can be decoupled with the SVI,VC state by configuring the following command
  - R1(config)#platform vfi provision vlan <vlan#>
- REP over EVC is supported on Ether Channels
- Maximum of 128 REP segments per switch, 2 ports per switch can belong to the same segment
- Recommended number of switches on each REP Segment is 32
- Supported with ES+/ES20
- Supports SSO/ISSU
- **Enhancements:**
  - REP fast hello (Q3'2010) – for uWave links
  - Uplink tracking
  - Multicast Convergence
  - REP edge no-neighbor

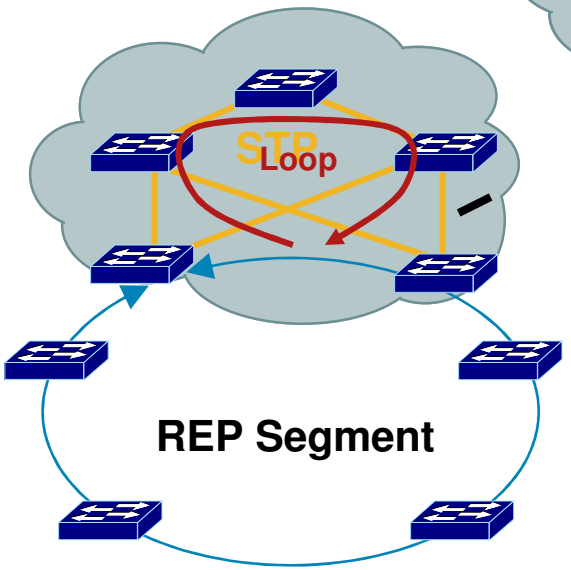
REP and Spanning Tree

- REP and STP are mutually exclusive on the same interface.
- An interface configured with REP does not participate in Spanning Tree.
- REP ports does not forward Spanning Tree BPDUs.
- → Proper design to ensure loop avoidance
- → The common link to REP and STP to run Spanning Tree **not** REP.

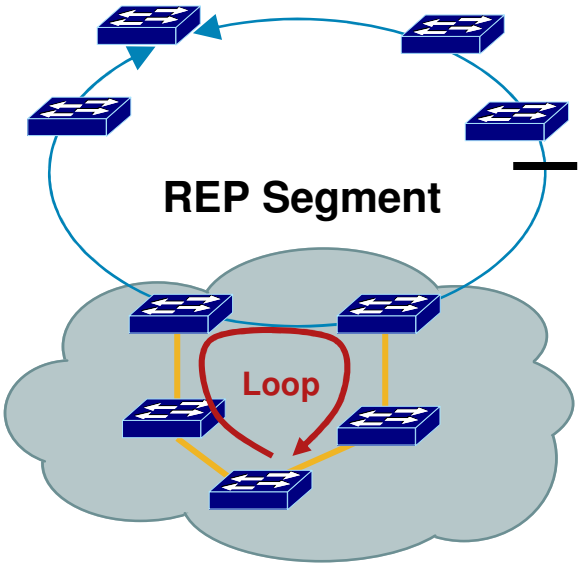
Proper Design



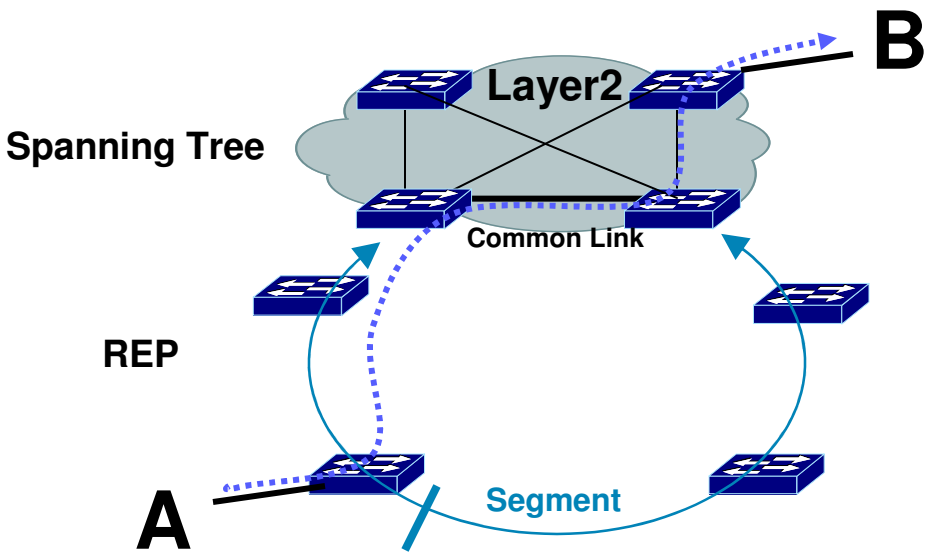
Not recommended

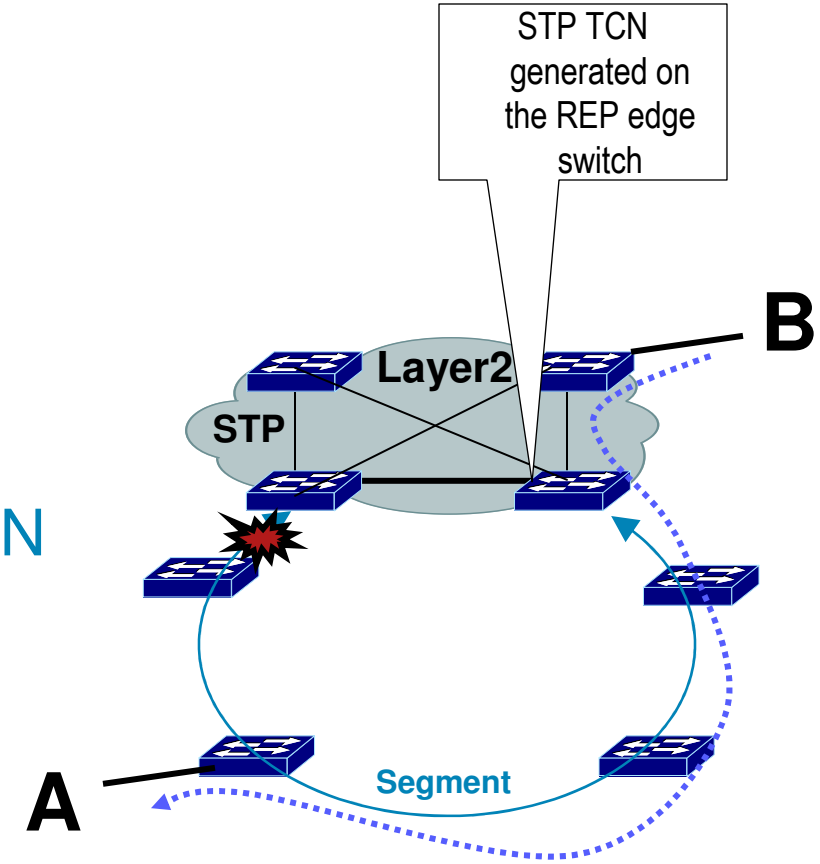


Not recommended



REP to STP TCN propagation

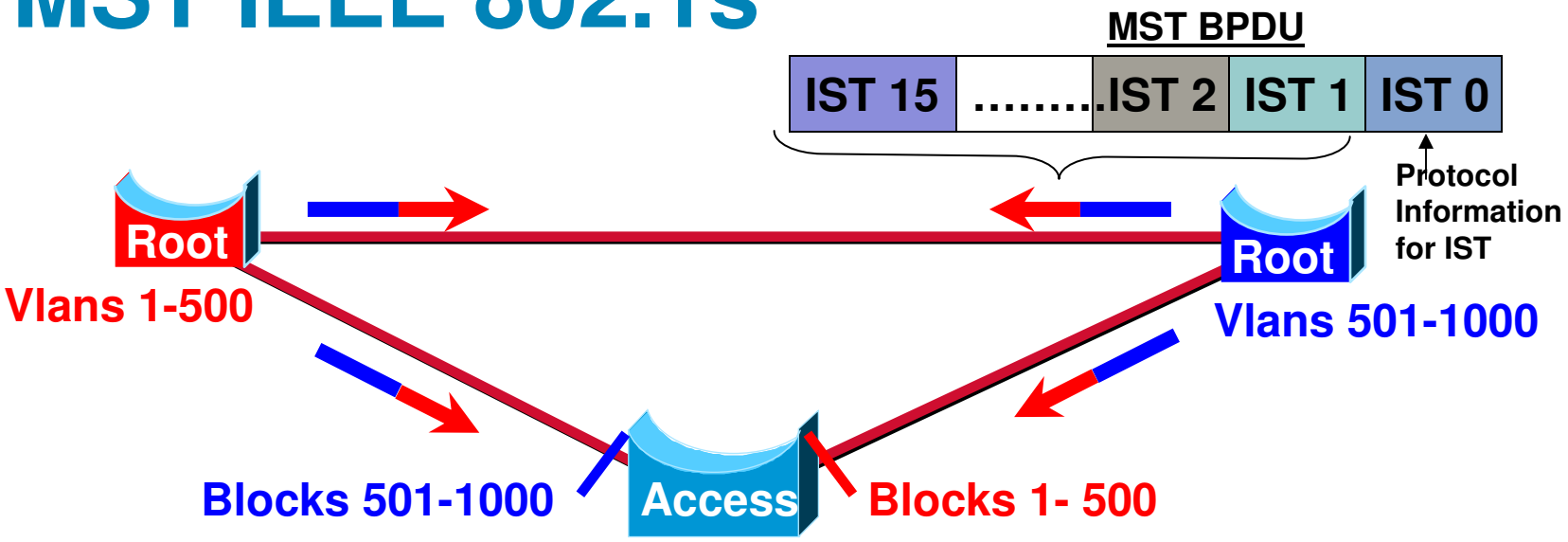




- REP segment propagates TCN notification to Spanning Tree.

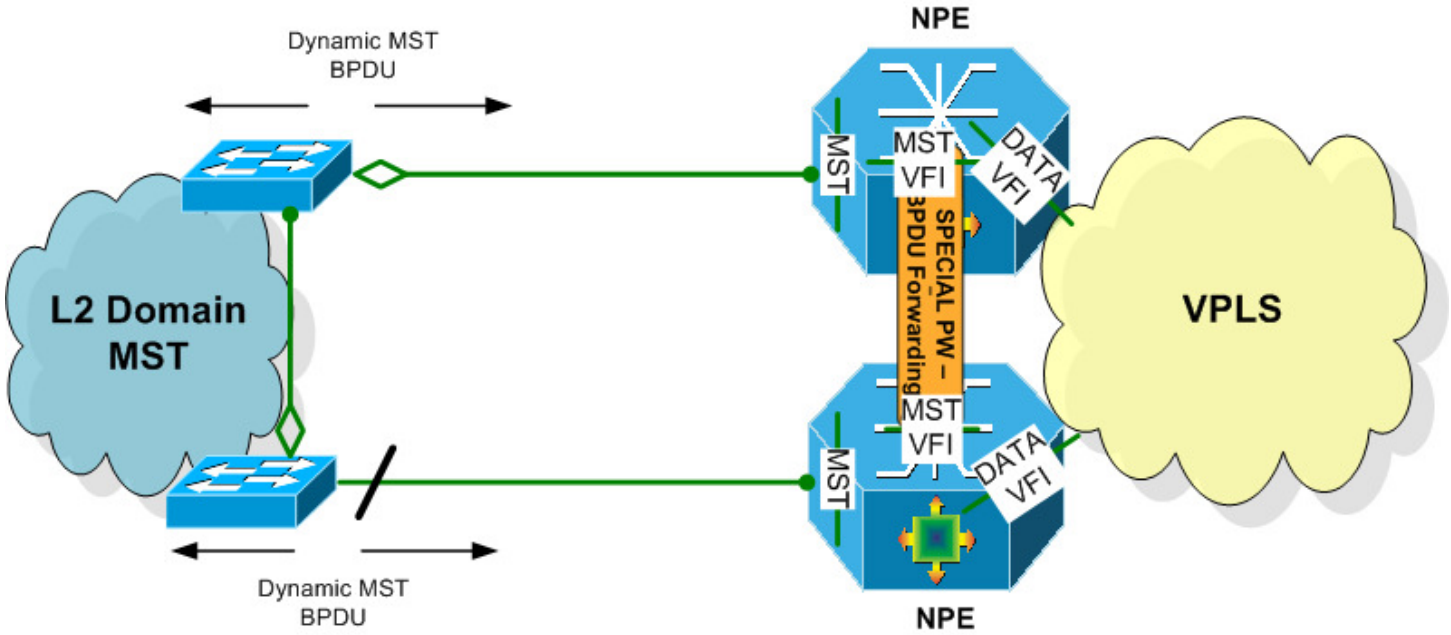
MST Access GW

MST IEEE 802.1s



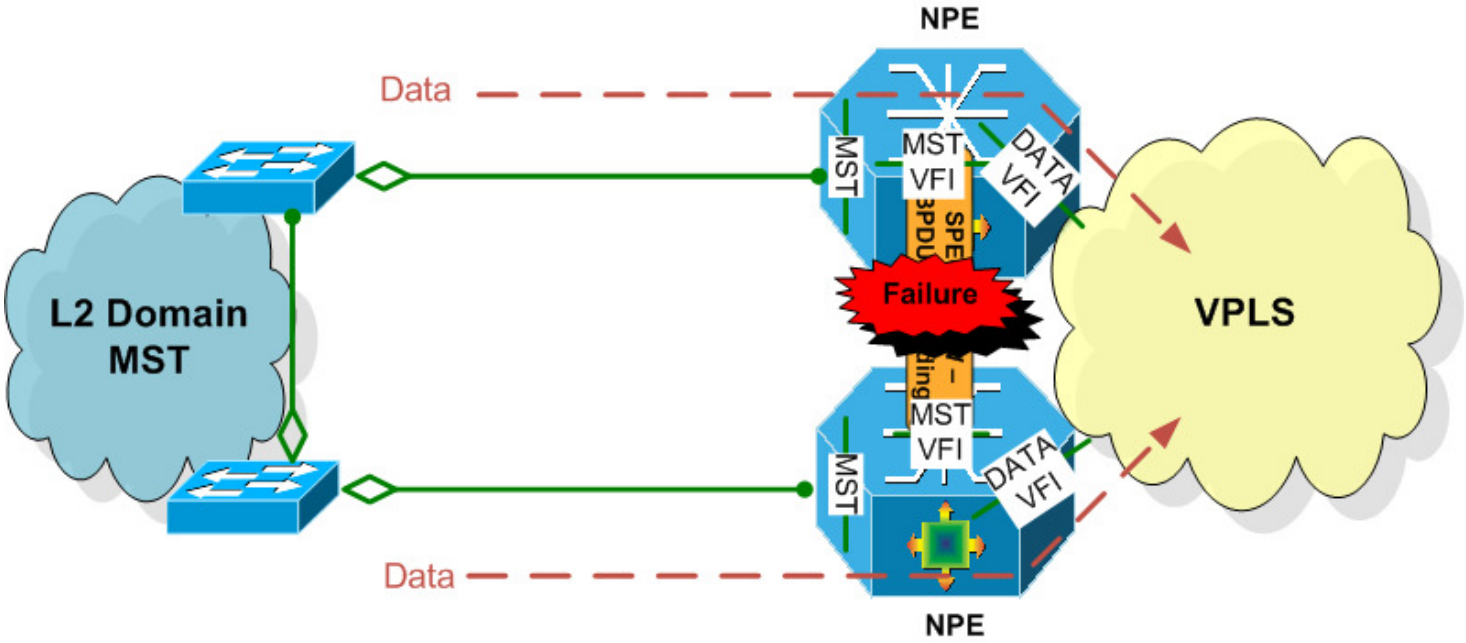
- MST optimizes STP solution building multiple logical STP on a single physical network
  - VLANs are placed into groups and assigned to Spanning Tree Instances (IST)
  - Each Spanning Tree Instance builds each own topology
- An MST region is a group of switches having the same configuration
  - Name (32 bytes), Revision number (16 bits), Vlan to Instance Table must be the same on all switches
- MST only **sends** 1 BPDU (whether on a Trunk port or Access Port) that contains data for all the instances.
- MST BPDU is sent on Native VLAN using well known MAC
  - Unlike PVST or PVRST which are tagged

MST on NPE with special PW



- STP runs on N-PE. L2 links participate STP, thus all L2 links are **point-to-point.**
- Dedicated PW will be treated as **"virtual link"** to participate STP as well. It simulates BPDU connectivity to match data connectivity provided by VPLS.
- PE must be STP root or tune STP parameter to make sure dedicated PW is not blocked by STP.
- MAC withdrawal is triggered by STP TCN
- Convergence time is determined by STP. 1-2 seconds with MST

MST on NPE – Special PW Failure



- Dedicated PW failure, it could be native SVI interface down, PW down, etc. This failure can cause redundant L2 link unblocked which will cause duplicated frames sent to other site, and cause L2 loop
- **if VPLS PW come up first before the dedicated PW, then both L2 links are in forwarding state, this will cause temporarily loop**
- **Single point of failure, vulnerable for human mistake**
- In normal case, if dedicated PW goes down, it means MPLS uplinks on one of the PE has problem. Thus VPLS data PW will go down as well on that PE. As result, there is no L2 loop since only one PE has active PWs.

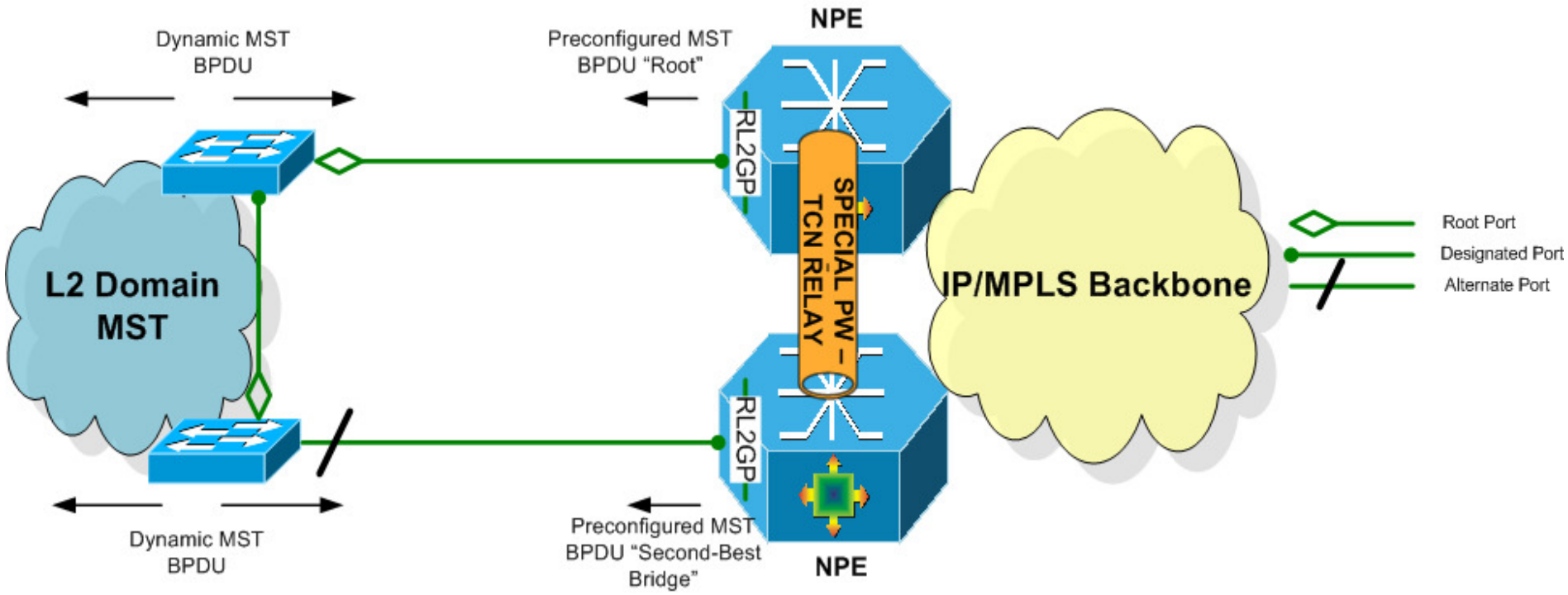
MST on EVC implementation on Cisco 7600 – 12.2SRD

- MST is supported for EVC with Bridge Domains
  - MST is based on the outer VLAN, EVC must have unique outer VLAN encapsulation
  - VLAN must be assigned to an IST
- Special PW between two Aggregation Nodes is a part of MST topology
  - Special PW is never blocked (low port cost assigned internally)
  - Special PW is tied to native VLAN
- Data packet goes through regular VPLS VC
- STP TCN triggers VPLS MAC withdrawal in aggregation network
- Limitations
  - No boundary functionality - peer must be running MST
  - Bridged Services blocking
    - Only bridge-domain EVCs are blockedVPLS or switchport in the core
    - Service instances with xconnect/connect are not blocked
  - No support for the following service instances: Default, Untagged, Range or list on outer tag
    - If xconnect/connect configured on default or untagged EFP, BPDUs are tunneled (i.e. MST on the port does not work)
  - EVCs on physical ports – no support for port-channel yet. Target for SRF.
- Recovery times: 1-2s

Why MST Access Gateway?

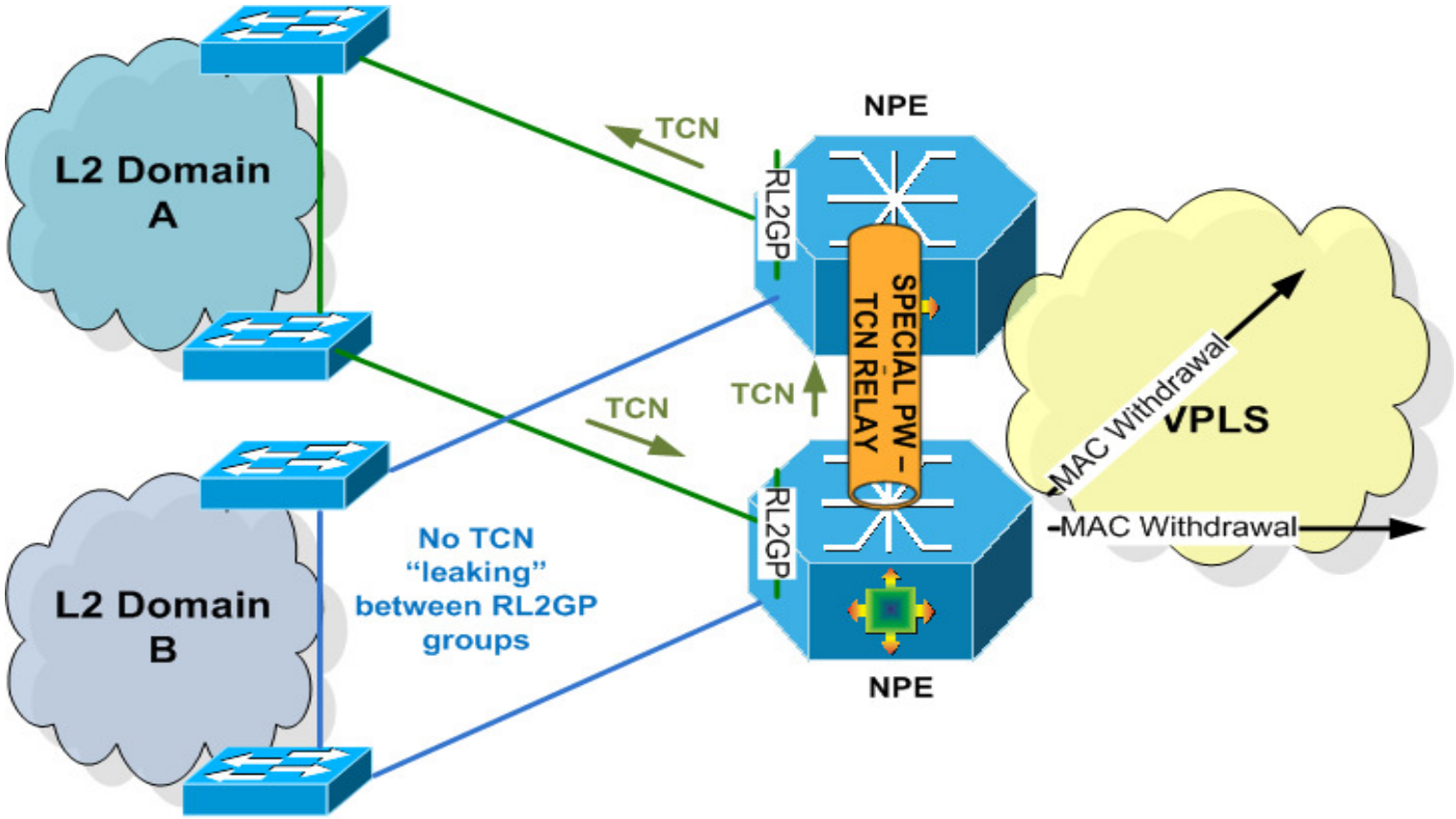
- Known also as Reverse L2 Gateway Protocol – R-L2GP
- Avoid running full Spanning Tree protocol on NPE – STP is challenging to troubleshoot
  - Terminate multiple Ethernet access rings running MST on NPE without running full STP
  - Each ring can have its own independent topology
  - Isolate topology changes/MAC flushes localized to each ring
- Improve scalability
  - No full STP processes on NPE routers
- Maintain existing STP topologies on the access networks
  - Access nodes just speak regular MSTP/RSTP
- Platform Support
  - ASR9K – since FCS
  - 7600 – 12.2SRE

MST AG operation



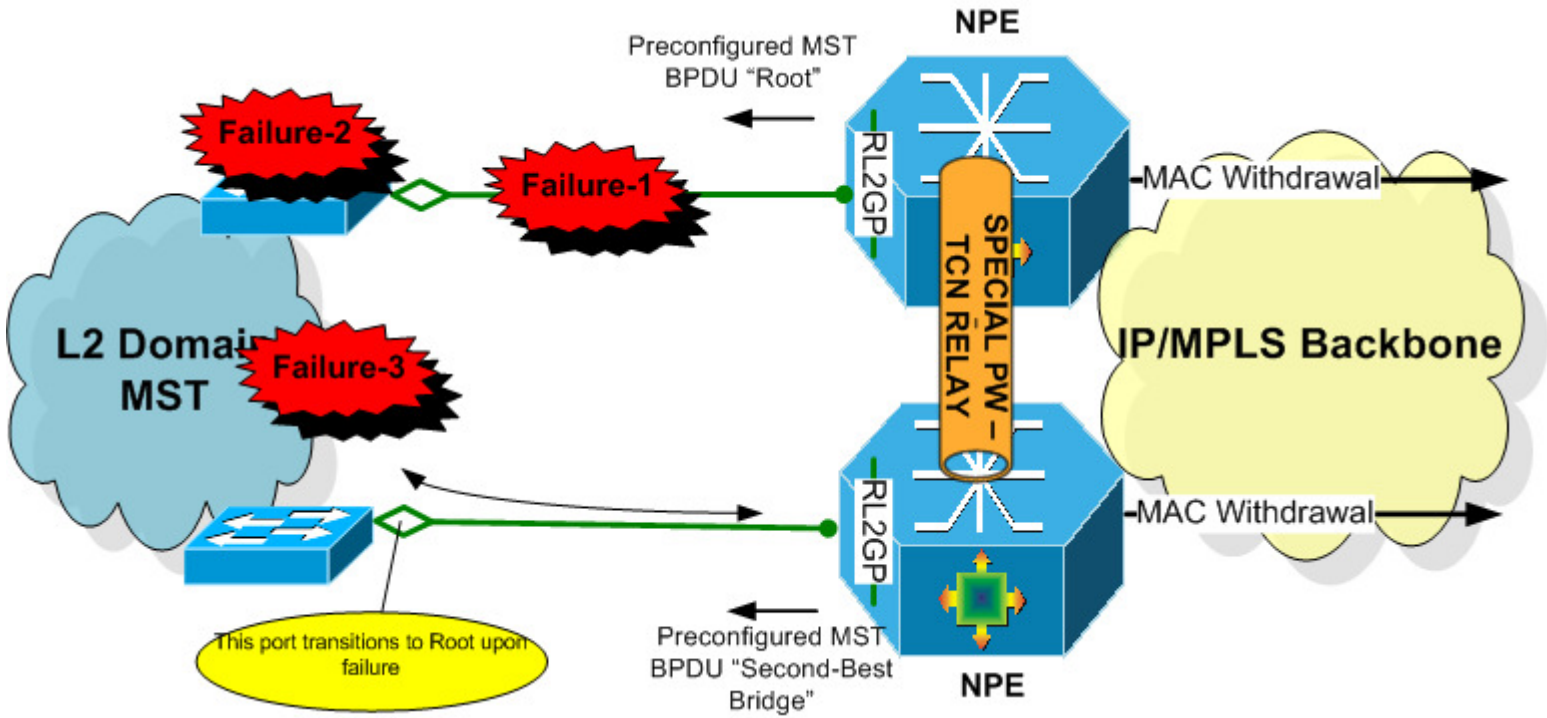
- MST AG ports send **preconfigured BPDU's with root or zero cost** to root information towards access network. Access network sees a loop because of root reachability from both NPE's.
- Both NPE's can send the same information or arbitrarily can be set as best and second best bridge via priority or cost setting for load balancing purposes
- **L2 domain runs regular MST protocol. All convergence operations and port state transitioning happen in the access network.** Ignore BPDU and relay on TCN

MST AG TCN Propagation



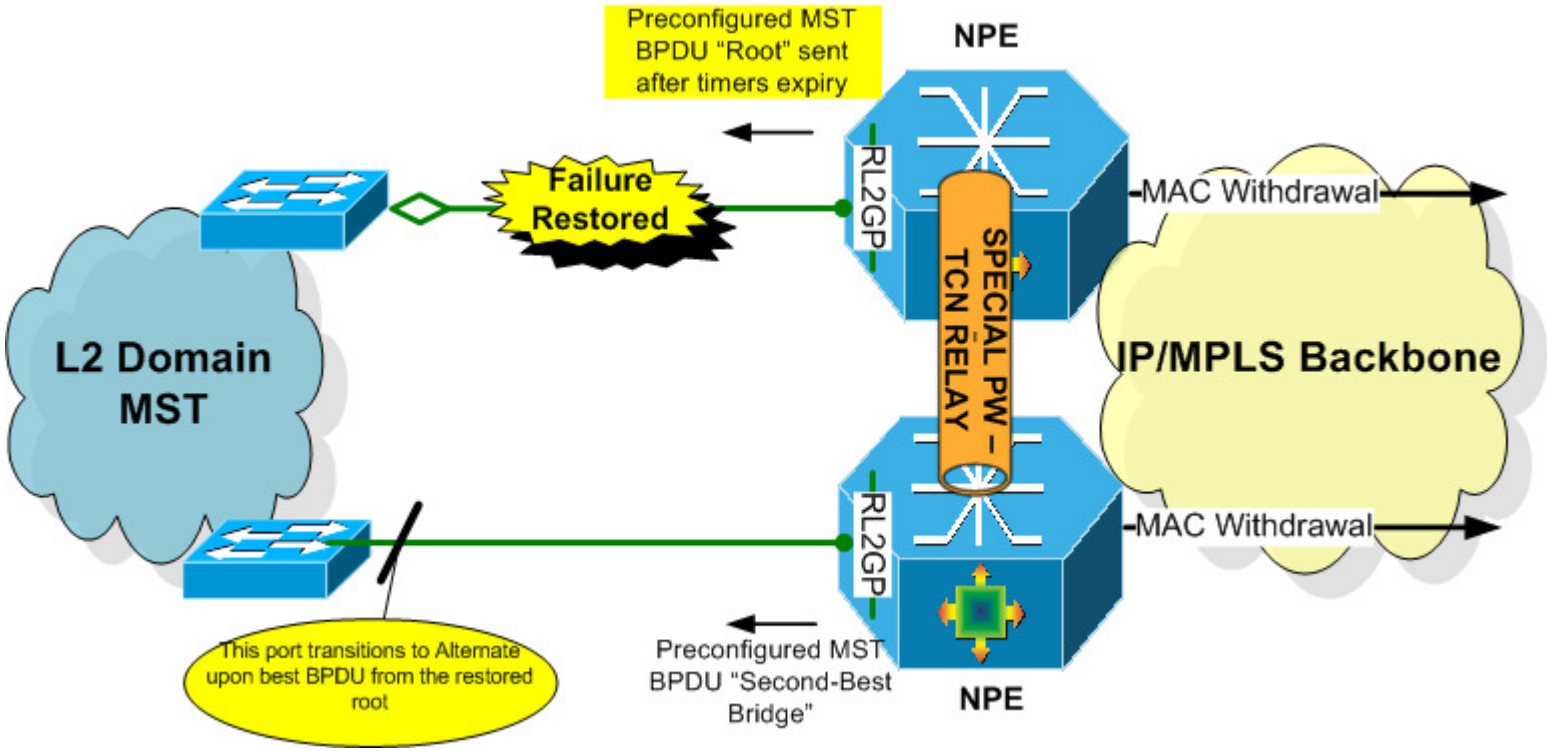
- NPE's snoop and relay TCN from BPDU received from access network
- **TCN stays within MST AG group!!! Reusing outer VLAN tag**
- NPE's trigger MAC withdrawal to neighbors
- TCN is forwarded only to the port within the same MST AG group thus providing L2 domains isolation

Access Network Failure Scenarios



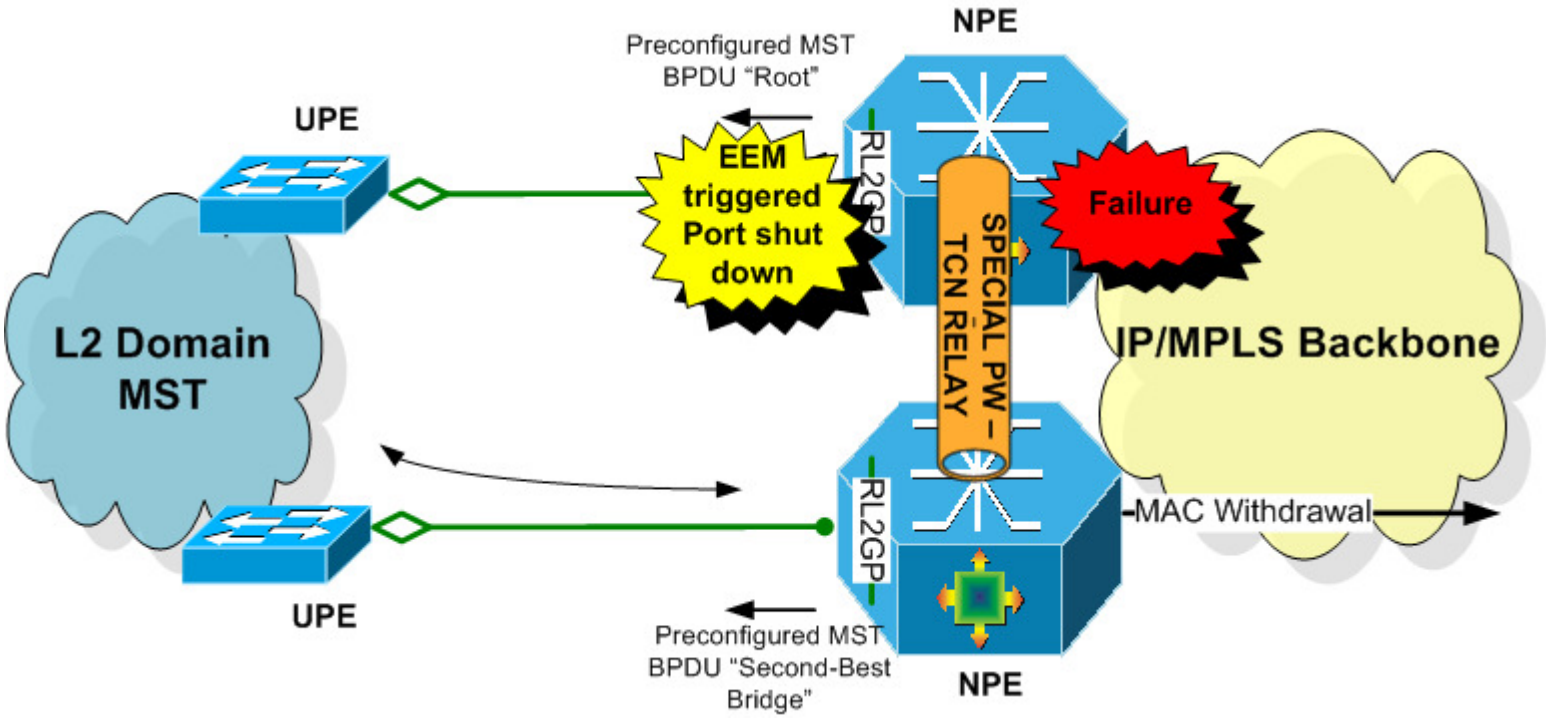
- Failure Scenarios 1,2 and 3 cause a primary data path disruption
- UPE-2 BPDU on Atlernate Port is has now the best BPDU – port transitions to root port role and forwarding state and data path is restored
- TCN propagates across L2 domain and is relayed between NPE's
- NPE's trigger MAC withdrawal (TCN snooping)

Root Port Recovery



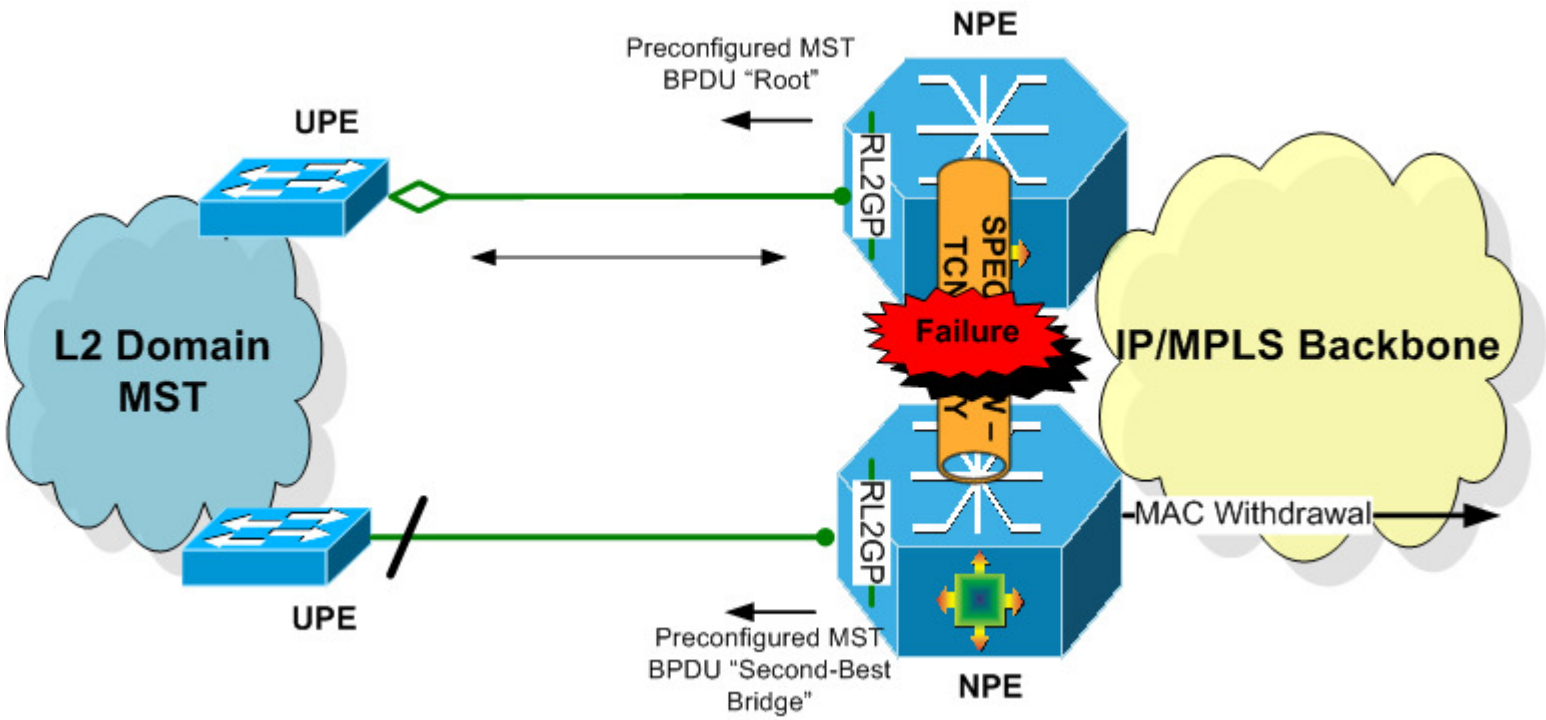
- When root bridge recovers it starts sending best BPDU towards the access network and convergence to the original path occurs
- Sending of the best BPDU has to be delayed to allow core convergence; e.g. if the router was reloaded
- 7600 router runs STP state machine on MST AG when the port is coming from down to up state. The ports is going through LST-LRN-FWD states. To disable this behavior **spanning tree port fast** has to be configured on MST AG ports. **BPDU's are sent immediately upon port recovery which can cause traffic black-hole if core has not converged. EEM can be used to delay port-up event under certain scenarios.**
- ASR 9K does not run spanning tree state machine and defines a **dedicated timer** to delay the best BPDU generation.

NPE isolation failure scenario



- N-PE isolation occurs if all core facing interfaces are not available resulting in VPLS, Psedowires or L3 connectivity failure
- **N-PE isolation failure is not propagated into access interfaces therefore STP topology remains unchanged, this results in traffic blackhole as access network continues forwarding towards isolated PE**
- An uplink tracking feature is under consideration for future releases
- Current solution is based on EEM – when router isolation is discovered the access interfaces from redundant networks can be shut down which triggers MST convergence. Upon recovery, timer can be set to delay access links recovery and avoid immediate BPDU sending to the access network.

Special PW failure scenario



- Special PW failure can be result of PE isolation or a miss-configuration
- Unlike MST, R-L2GP special PW failure **does not cause a loop and therefore does not cause permanent traffic loss** because BPDU forwarding topology remains unchanged and is not affected by this failure
- TCN will not be relayed between the two NPE's – MAC flush may not happen in a part of L2 domain which may cause temporary traffic loss until MAC aging occurs. Bidirectional traffic will be restored immediately.
- MAC withdrawal will still be generated by the NPE receiving TCN

EVC STP Modes Comparison

STP mode	RPVST/ PVST	MST	MST AG
EVC Support	NO BPDU relay only	SRD	SRE
VLAN-STP mapping	Per VLAN	Single MST Region VLAN to Instance mapping applies to all ports	Single MST Region VLAN to Instance mapping applies to all ports
Dynamic Port State	N/A always FWD	YES	NO Designated ports, always forwarding
Dynamic BPDU	N/A	YES	NO preconfigured BPDU's
TCN isolation	YES	NO	YES TCN forwarded between ports within the same L2GP group
MAC Withdrawal	NO	YES	YES
Complexity	Medium	High	Low Prone to miss-configuration

MST AG configuration steps

1. Configure MST parameters

MST AG reuses global MST configuration template to construct BPDU's. To insure proper MST function, parameters like name, revision and timers should match on other bridges.

Note: due to single domain support the same MST parameters will be used on all MST AG groups. In particular IST to VLAN mapping.

```
spanning-tree mode mst
spanning-tree mst configuration
 name c7600
 revision 1
 instance 1 vlan 3500-3599
spanning-tree mst hello-time 1
spanning-tree mst forward-time 4
spanning-tree mst max-age 6
```

MST AG configuration steps

2. Configure MST AG Pseudo-Information

```
spanning-tree pseudo-information transmit 1
  remote-id 2                                  /*** use the number of pseudo-
                                               information of the peer router ***/
  mst 0-1 root 24576 001e.f7f6.6040
                                               /*** root bridge and priority that
                                               will be send in BPDU on MST AG
                                               ports***/
```

3. Assing MST AG Pseudo-Information to a port

```
interface GigabitEthernet1/32
  spanning-tree pseudo-information transmit 1
                                               /*** the port will send preconfigured
                                               BPDU's as per MST global and
                                               pseudo-information group 1
                                               configuration ***/
```

MST AG configuration steps

- Configure Special PW for TCN relay

  This is the same configuration step as for MST BPDU relay

  Note: to activate MST/MST AG on special PW, VLAN's mapped to the MST must be created in the global VLAN database (at least one)

```
interface Vlan1
 no ip address
 xconnect vfi BPDU
end

NPE-1#sh run | sec BPDU
l2 vfi BPDU manual
 vpn id 1
 forward permit l2protocol all
 neighbor 10.1.1.6 encapsulation mpls
```

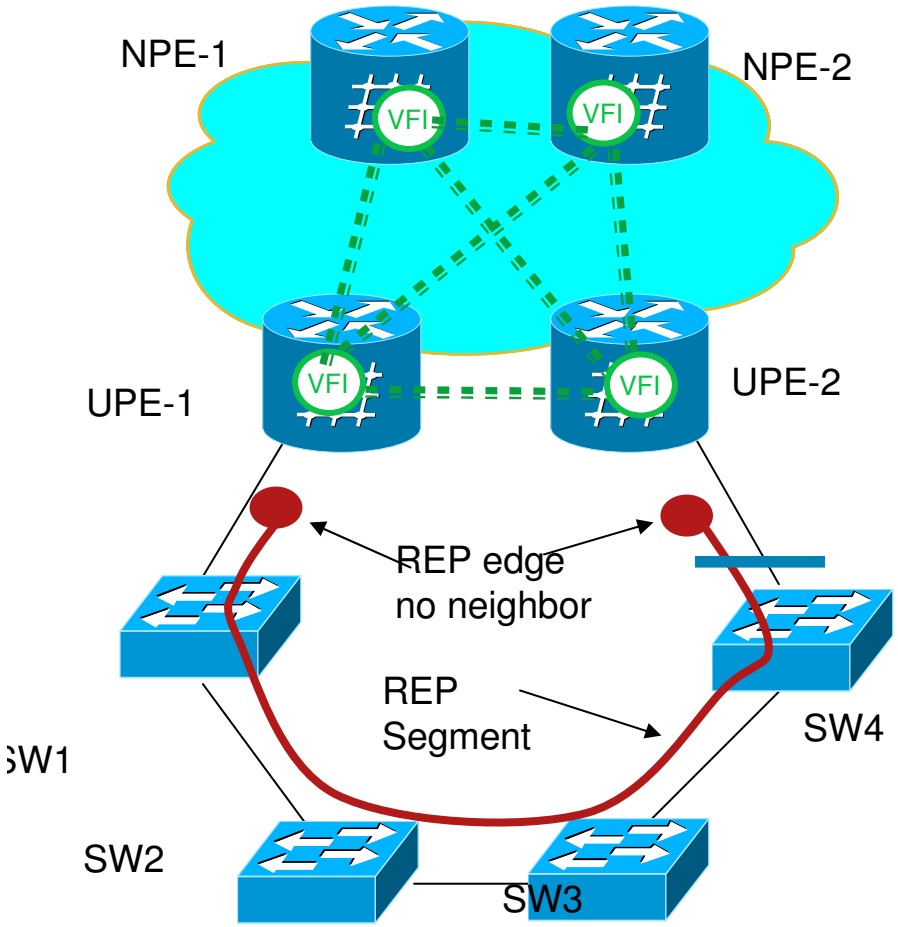
- Configure Service Instances and Bridge Domains
- Configure all Egde Ports explicitly with "portfast" feature

  avoid LRN/LSTN states when bridge is converging

MST AG Restrictions

- Supported on ES+ and ES20
- Applicable to EVC with Bridge Domain only
  - No xconnect, connect or subinterface support
- No EVC untagged, priority tagged or default encapsulation support
  - Native VLAN is used for BPDU forwarding
- Single MST region support
  - All MST AG groups share MST Instance - VLAN mapping, name and revision
- Service instance global outer tag
  - At least one outer tag per instance must be defined as a global VLAN to insert Special PW into MST/R-L2GP topology

ASR9K MST AG - Inter-operate with REP



It works!

- **Access switches are in the REP segment by themselves**
- **Access network topology change will generate TCN. MST gateway router will snoop the TCN and trig MAC withdrawal accordingly. No packet black hole**
- **MST gateway keep sending BPDU to access switches which will be ignored**

Drawback

- **Unnecessary BPDU generation, which is not good, but not a big concern either**
- **Require the STP configuration on the gateway router, which cause operational complexity**

Solution – REP access gateway

- **REP access gateway doesn't run any REP code, but just solve the above two issues**
- **Not generate BPDU, but continue snooping TCN**
- **Have a simple configuration**

Conclusions

MST AG provides an appealing option to operate STP networks to service providers:

- Maintaining access networks without modification
- Lower maintenance complexity on N-PE's – no full spanning tree support
- Lower troubleshooting complexity on the network

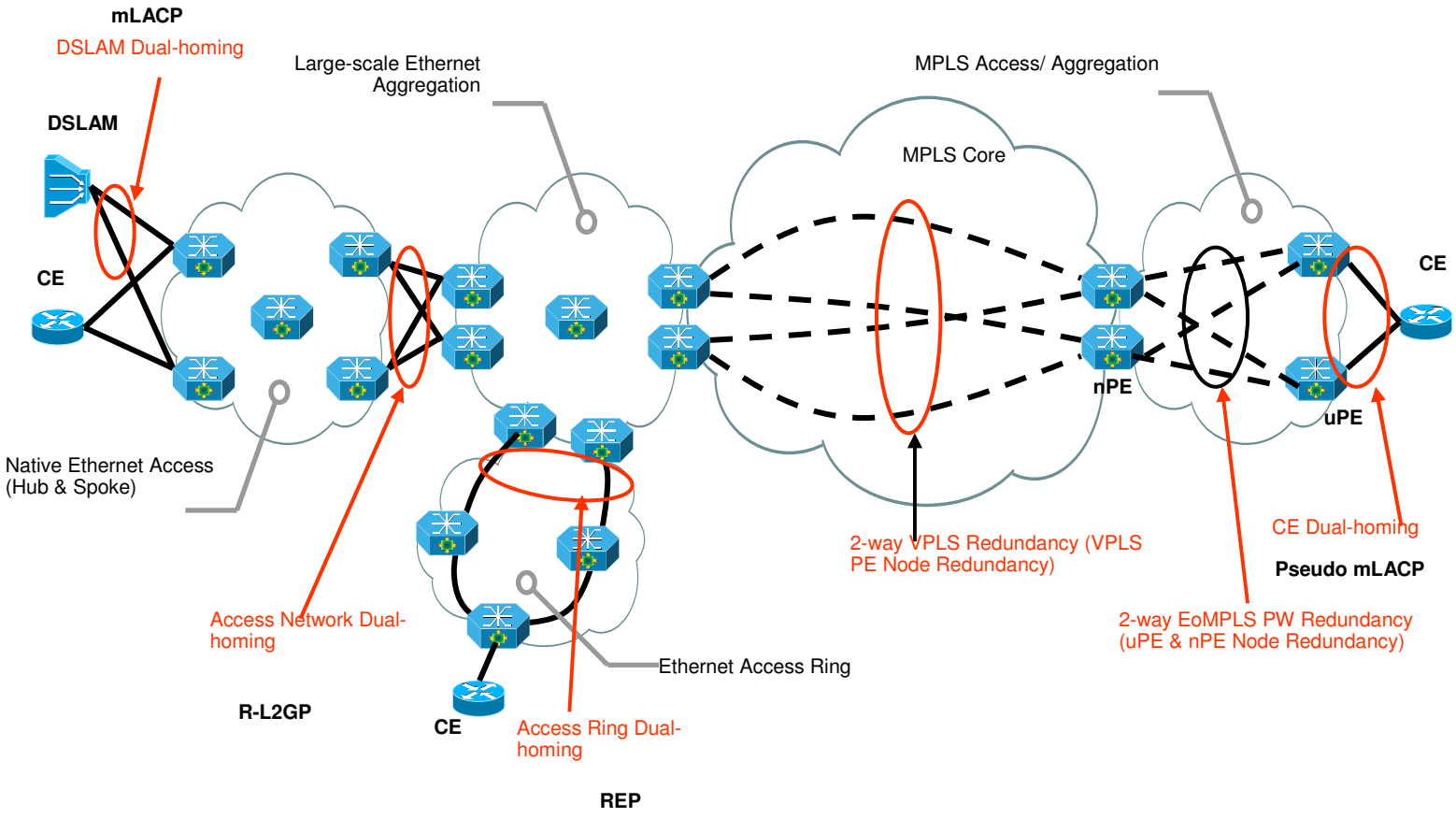
  STP isolation for L2 aggregation domains separated by VPLS core

  Deterministic root location
- Improvements from MST/EVC:

  TCN isolation between access domains

  More robust implementation, special PW failure does not cause traffic black-hole

Layer 2 Redundancy Options



Layer-2 Redundancy Options

	Pros	Cons	Supported Line Cards
802.1D with extensions	Ubiquitous	Slow convergence. Can be as high as 50 secs	LAN cards,ES20,ES40 SIP600
MST	802.1s Fast Convergence	1-2 secs (may be more). Depends on MST Ring size	LAN Cards,SIP600 ES20,ES40 MST/EVC Supported
REP	Easy to Configure <1sec convergence	Cisco Proprietary. Not a Standard	REP/EVC only on ES20/ES+ REP/Switchport on LAN Cards,SIP600
R-L2GP	Static BPDUs for n-PE root election,Improves Convergence	Cisco "Proprietary" Not a Standard	ES20/ES+ only
Hot standby PW	Pre programming of MPLS labels on redundant LC	Slow convergence with High scalability of attachment circuits. Only Scalable EoMPLS supported	ES20/ES+ only. (No SIP-400 Support)
mLACP with HSPW	End-End Fast Convergence solution	Slow Convergence for certain Faults in the network	ES20/ES+ only.

Convergence Times – L2 Redundancy Options

	Scale	Convergence Times (Access)	Convergence Times (End-End VPLS)
MST	64 + 1 Instances	1-2 secs	5-6 secs
REP	6K EVC over 1K BD VLANS (Access) 1K EVC over 1K BD Vlans, 1K Mac (VPLS)	60-80msec	200msec
R-L2GP	16K EVCs	1-2 secs	4-5 secs
Hot standby PW	4K PWs,16k PW,32k PW	5-6 secs	N/A

Q & A


CISCO