Data Center Interconnect mit Layer 2

- Solutions Overview -

Gerd Pflueger – CSE R&S Germany

gerd@cisco.com

V4

Abstract

Data Center Interconnect mit Layer 2 – eine kritische Bewertung

Neue Technologien halten Einzug in das Data Center von heute. Dabei bilden VMware mit Vmotion, Data Center-Konsolidierungen und Cluster-Lösungen die Treiber für die Nutzung von Layer 2 und Spanning-Tree im Data Center. Doch wegen ihrer Nachteile wurden mehrere alternative Design-Methoden für das Data Center Interconnect (DCI) entwickelt.

Das besondere Augenmerk richten wir dabei auf zwei neue DCI-Techniken: CEE/TRILL (Covergenced Enhanced Ethernet/Transparent Interconnection of Lots of Links) und OTV (Overlay Transport Virtualization). Welche der Design-Methoden ist die beste für Data Center Interconnect (DCI)? Und welche Vor- und Nachteile haben diese Lösungsansätze der modernen Data Center-Architektur? Diskutieren Sie mit uns die richtige Architektur - denn sie allein entscheidet über den Erfolg oder Misserfolg im Data Center.

Agenda (V4)

- Motivation for L2 DC Interconnect
- Solutions Overview
- Solutions Details
- CEE/TRILL – more details
- OTV – deeper dive
- Key Take Aways

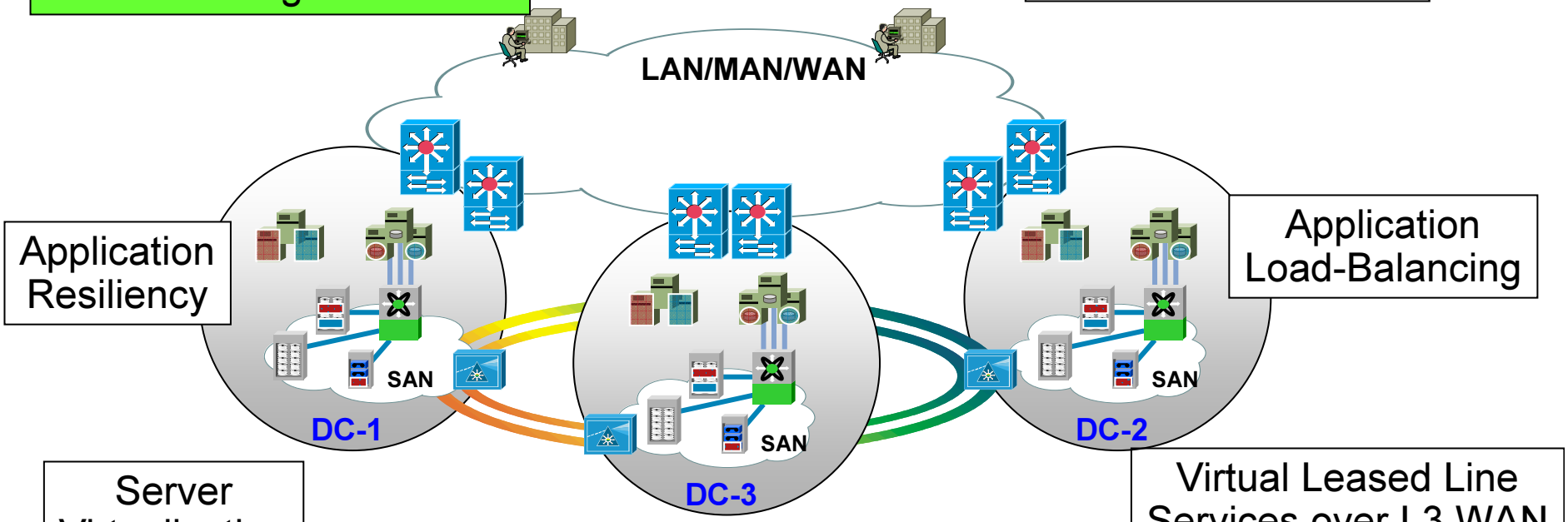
Motivation for L2 DC Interconnect

Motivation for L2 DC Interconnect (cont.)

Data Center Consolidation And Migration

Non-Routable Applications

Distributed Clustered Applications



Application Resiliency

Application Load-Balancing

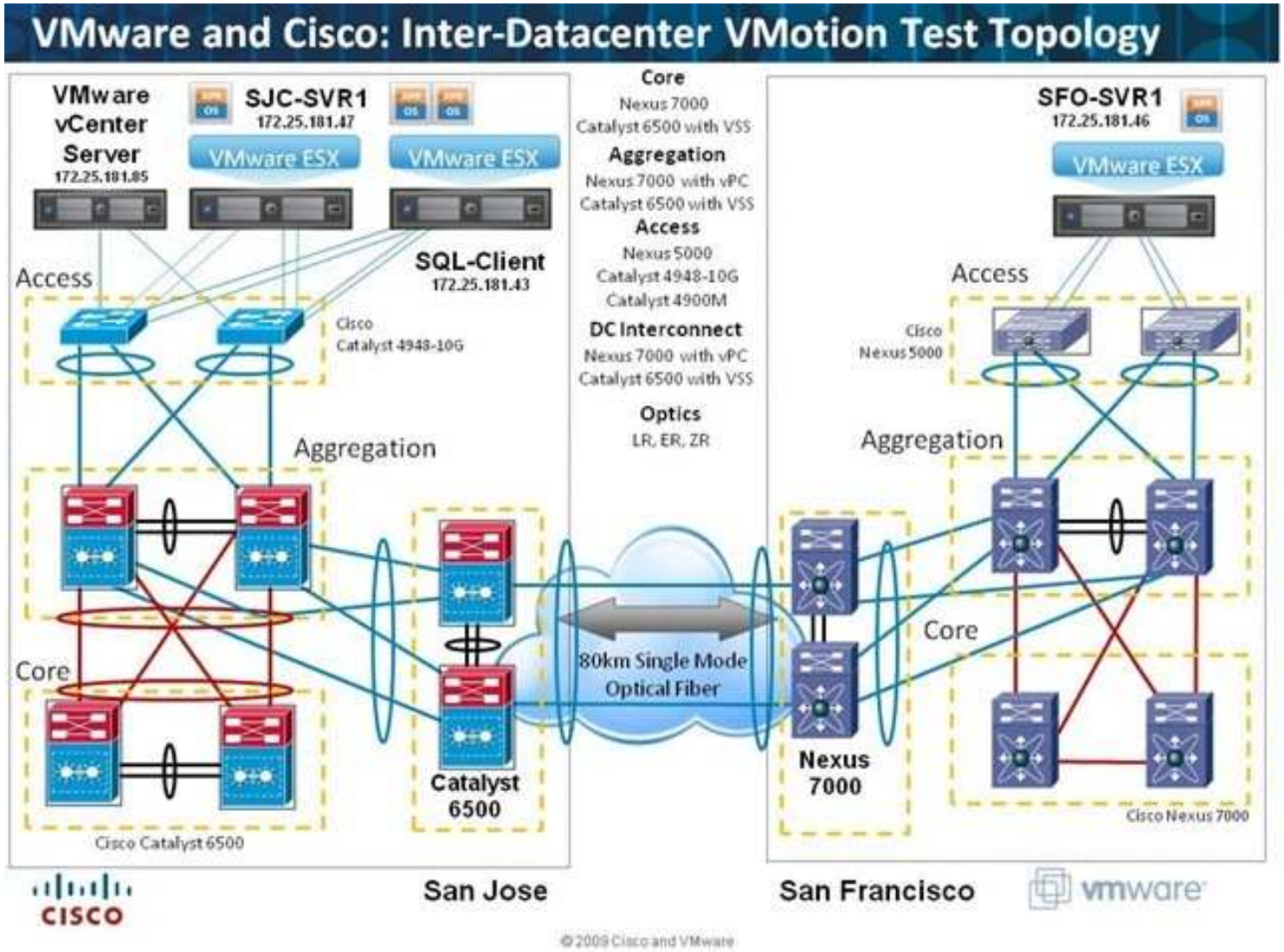
Server Virtualization

Virtual Leased Line Services over L3 WAN

L2 Domain Scaling

Server/Site Mobility (vMotion)

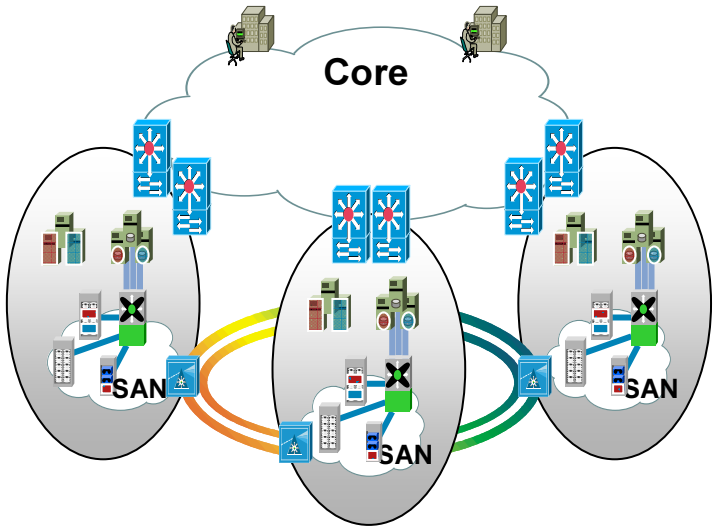
Motivation for L2 DC Interconnect



http://blogs.vmware.com/networking/2009/06/vmotion-between-data-centersa-vmware-and-cisco-proof-of-concept.html

Solutions Overview

Solutions Overview

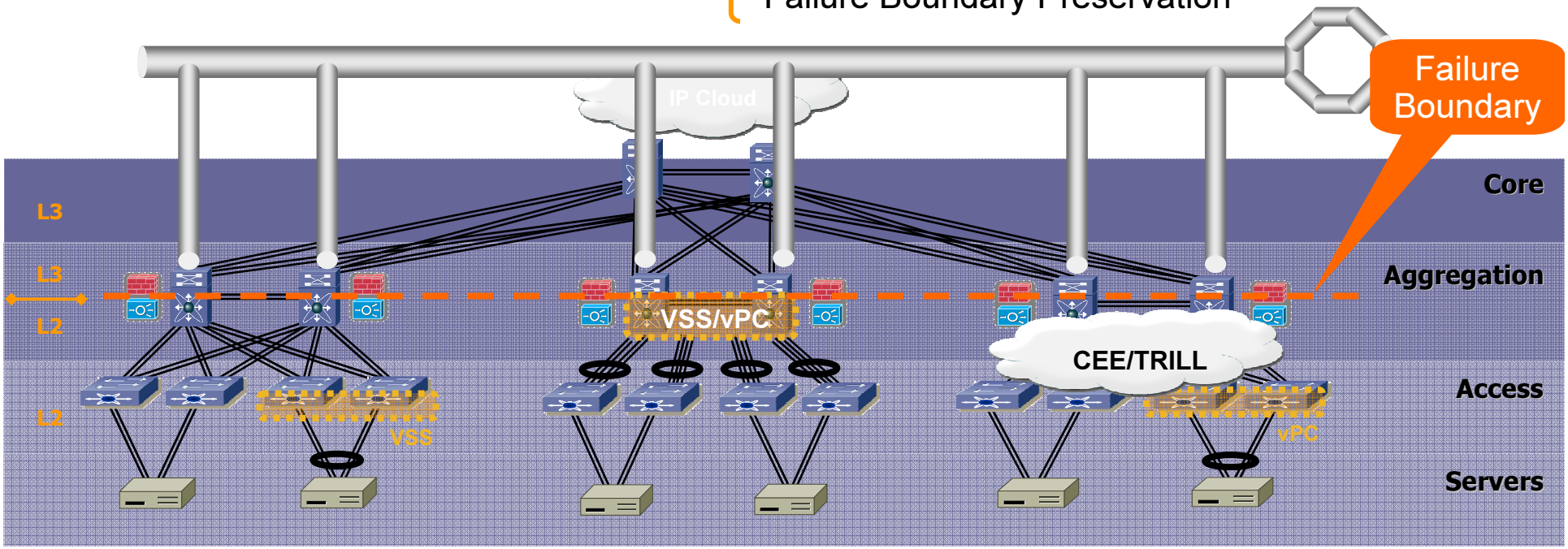


- Spanning Tree
- QinQ & MACinMAC (802.1ad & 802.1ah)
- L2TPv3
- MPLS (EoMPLS & VPLS)
- VSS/vPC
- CEE (aka DCE)/TRILL
- OTV

Solutions Overview – Roadmap/Strategy

Q1CY10 **OTV**

- Inter-POD Connectivity across **L3**
- Failure Boundary Preservation



STP

STP Enhancements

Bridge Assurance

FCS

vPC/VSS

NIC Teaming

Simplified loop-free trees

2x Multi-Pathing

Q4CY08

CEE/TRILL

16x ECMP (**L2MP**)

Low Latency / Lossless

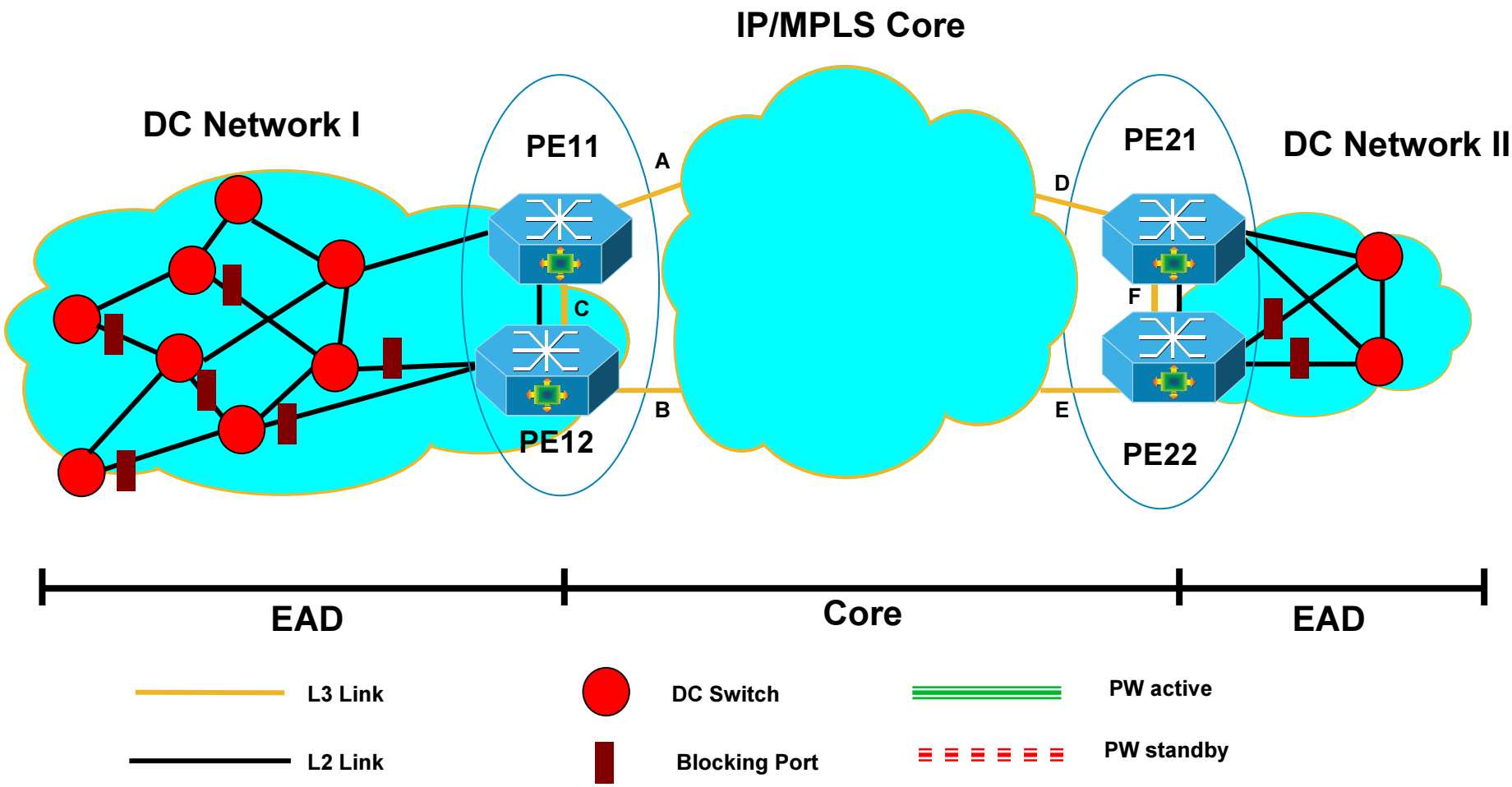
MAC Scaling

Operational Flexibility

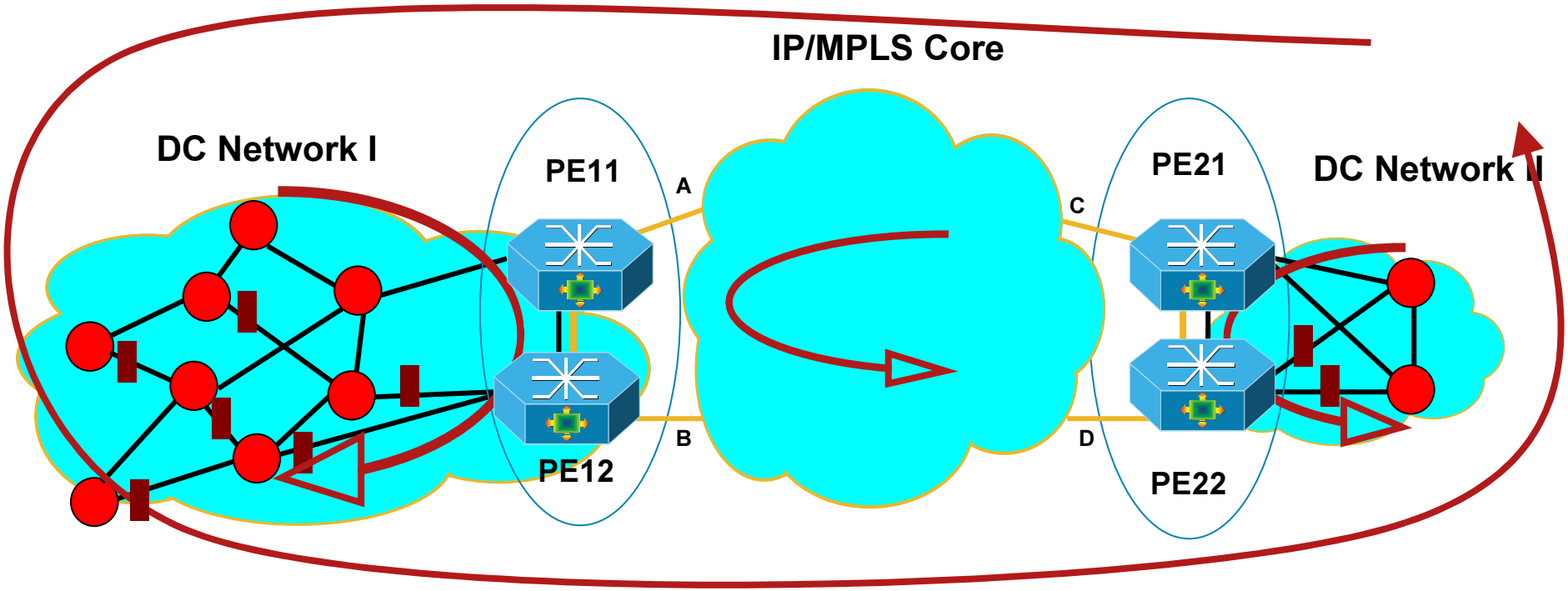
Q1CY10

Solutions Details

Reference Network Design



Network Design – Basic L2 Problem



EAD

STP/PVSTP

RSTP, MST

CEE/TRILL

Core

STP

EoMPLS/VPLS

CEE/TRILL

OTV

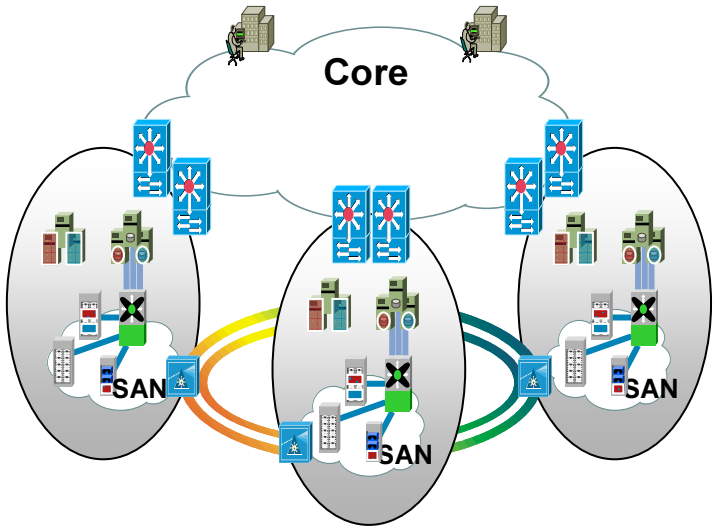
EAD

STP/PVSTP

RSTP, MST

CEE/TRILL

Solutions Overview - Agenda

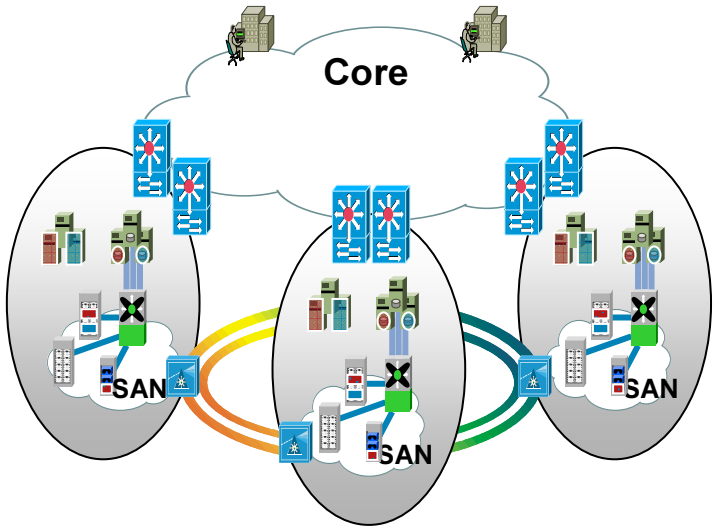


- **Spanning Tree**
- **QinQ & MACinMAC (802.1ad & 802.1ah)**
- **L2TPv3**
- **MPLS (EoMPLS & VPLS)**
- **VSS/vPC**
- **CEE (aka DCE)/TRILL**
- **OTV**

Layer 2 Risks

- Flooding of packets between data centers
- Rapid Spanning Tree (RSTP) is not easily scalable and risk grows as diameter grows
- RSTP has no domain isolation – issue in single DC can propagate
- First hop resolution and inbound service selection can cause verbose inter-data center traffic
- In general Cisco recommends L3 routing for geographically diverse locations

Solutions Overview - Agenda

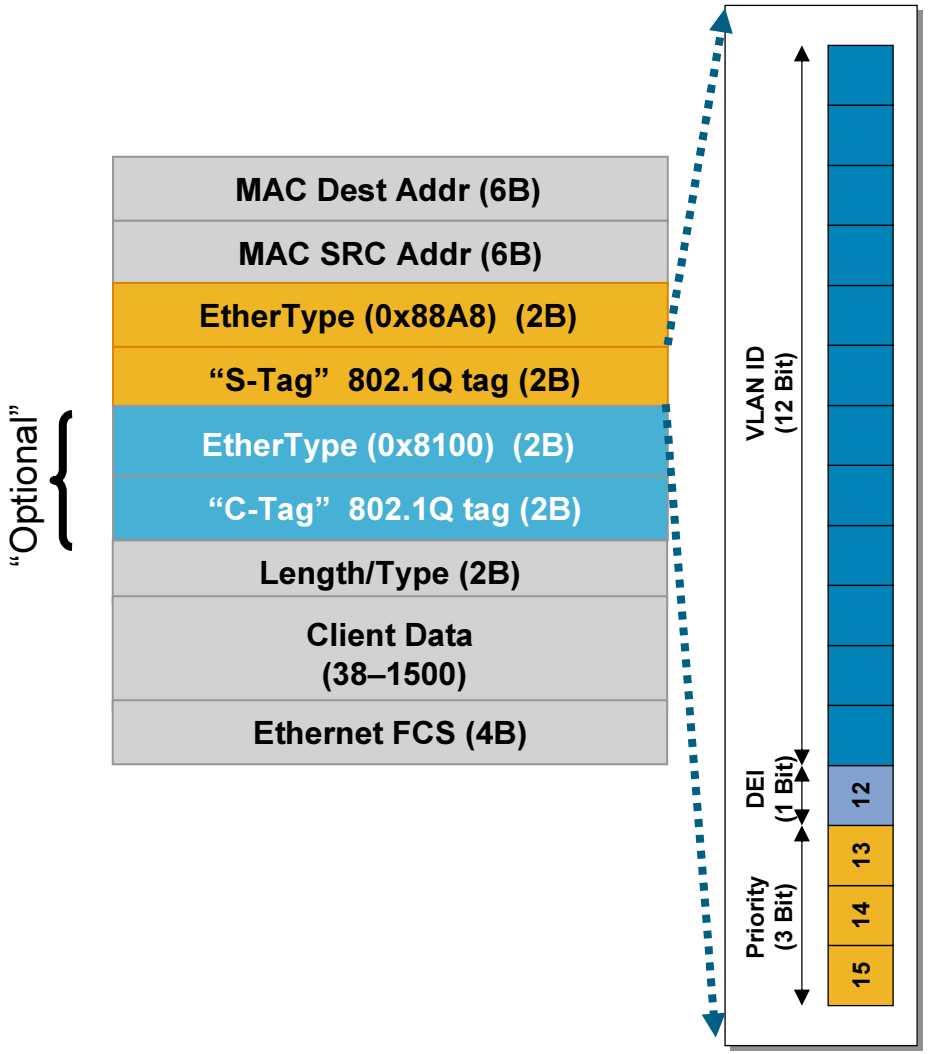


- Spanning Tree
- QinQ & MACinMAC (802.1ad & 802.1ah)
- L2TPv3
- MPLS (EoMPLS & VPLS)
- VSS/vPC
- CEE (aka DCE)/TRILL
- OTV

Building Provider Ethernet Access Networks

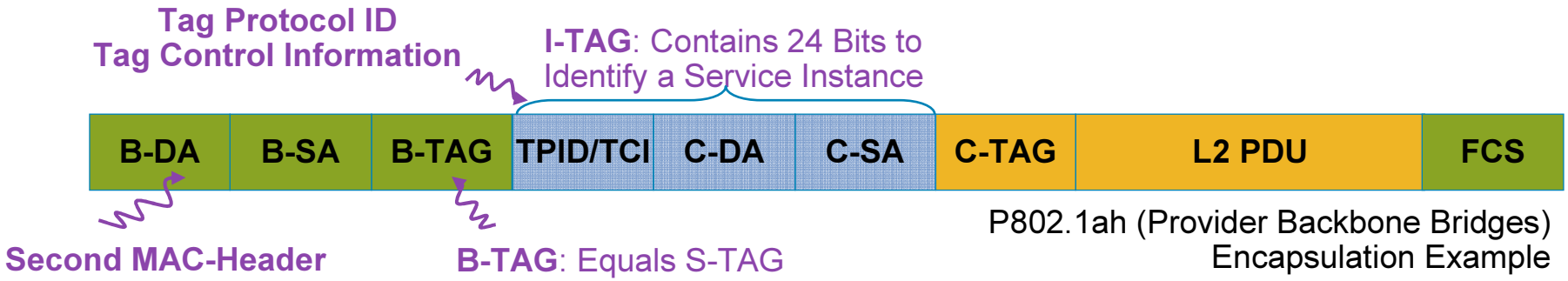
IEEE 802.1ad Provider Bridges

- Customer VLAN Transparency
  - IEEE 802.1ad Provider Bridges will provide a standardized version of “QinQ” (Note: Inner .1Q tag is optional)
  - Standard will include additional enhancements
- Frame Format same “QinQ”
  - New EtherType: 0x88A8
- Technically complete
  - Standard approved 8th Dec 2005 (Draft 6)
- Protocol Tunneling differs from Cisco’s L2PT
- See also: http://www.ieee802.org/1/pages/802.1ad.html



Scaling Provider Bridges

Provider Backbone Bridges (802.1ah): Main Ideas/Concepts



P802.1ah (Provider Backbone Bridges) Encapsulation Example

- Service Scalability

  Define a new "Service Instance Identifier"—24 Bits wide (taking the place of the former "VLAN"): I-SID

- Domain Isolation, MAC-Address Scalability

  Encapsulate Customer MAC-frames at the edge of the network into a "Provider MAC-Frame": New MAC-Header with B-TAG

- "Backward Compatibility" to 802.1ad

  Outer header = normal 802.1ad header

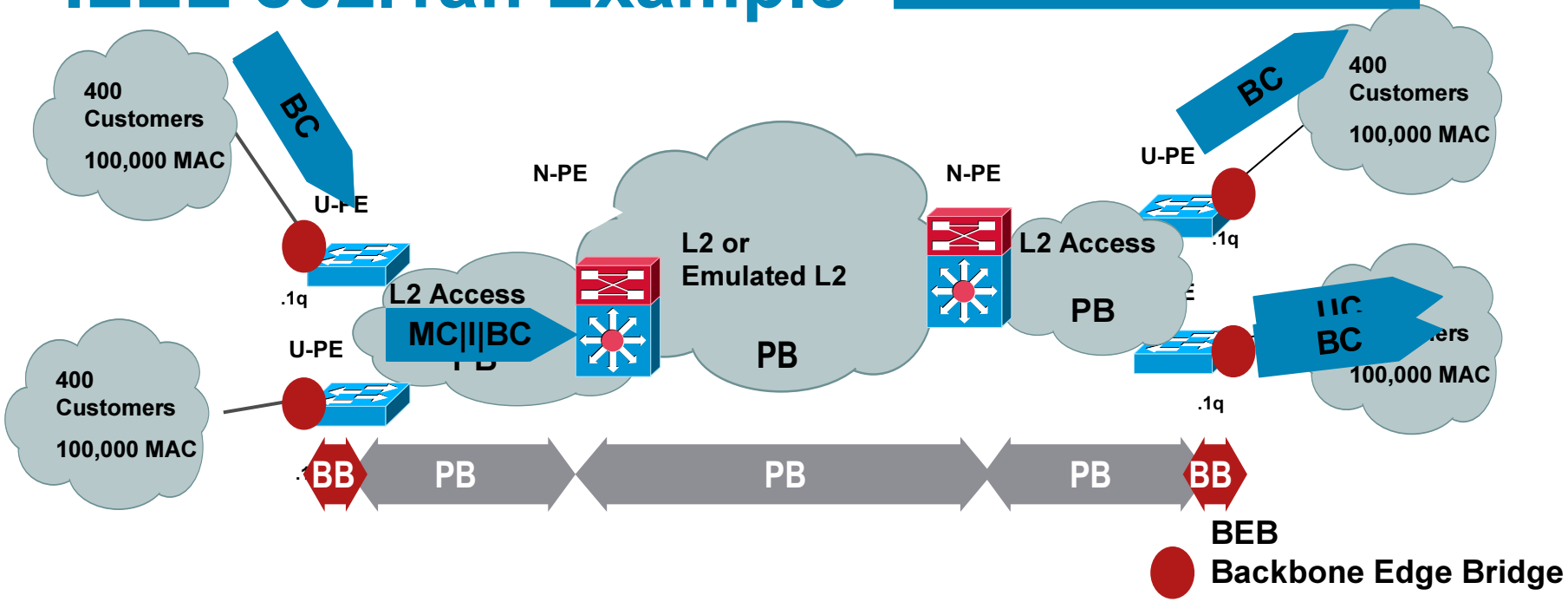
  802.1ah assumes existing L2 control plane mechanisms such as spanning tree and learning/flooding

  Other paradigms are valid (802.1aq, 802.1Qay, topology hiding with VPLS/MPLS and PW redundancy)

- 802.1ah Standard Approved June 12th, 2008

IEEE 802.1ah Example

UC = Unicast (Known)
I = I-SID
BC = Broadcast or Unknown UC
MC = Multicast

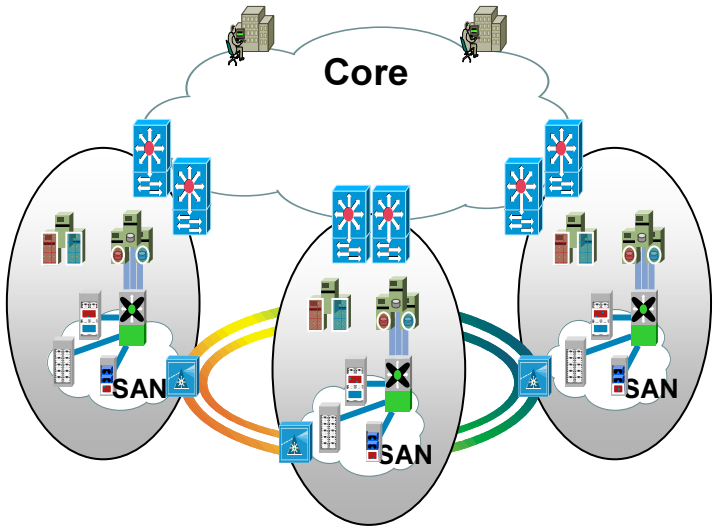


MAC Scalability: Core only needs to learn NNI/UNI_MAC addresses, rather than all customer MAC addresses

VLAN Scalability: 24-Bit I-SID allows Millions of Service Instances

No fork lift upgrade: Additional encapsulation is just done at the UNI/NNI, core bridges can be standard Provider Bridges

Solutions Overview - Agenda

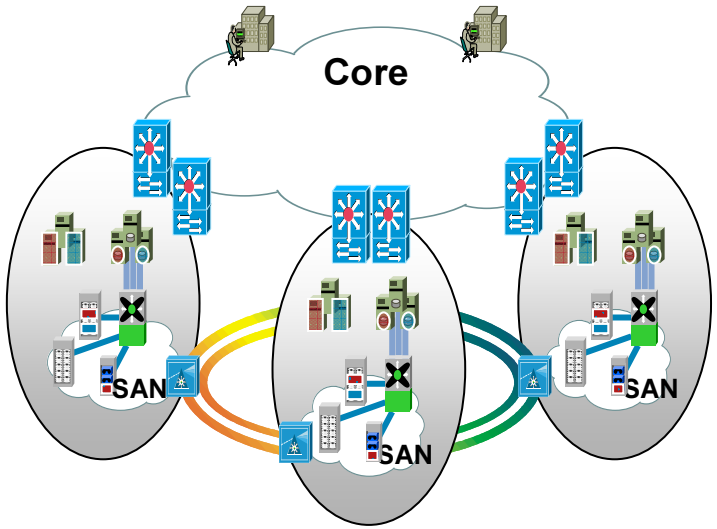


- Spanning Tree
- QinQ & MACinMAC (802.1ad & 802.1ah)
- L2TPv3
- MPLS (EoMPLS & VPLS)
- VSS/vPC
- CEE (aka DCE)/TRILL
- OTV

L2TPv3

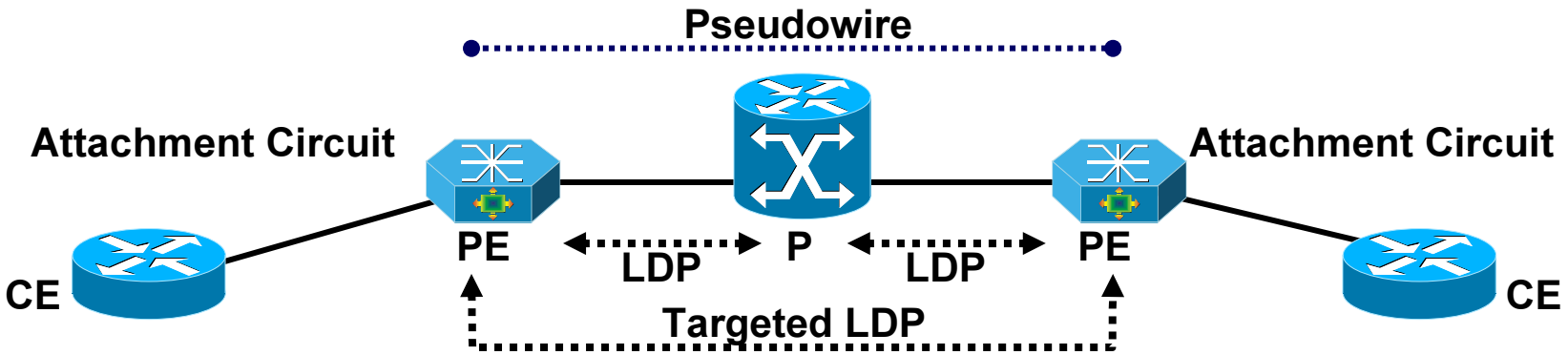
- P2P
- Not HW-supported on Catalyst
- Same design issues like EoMPLS

Solutions Overview - Agenda



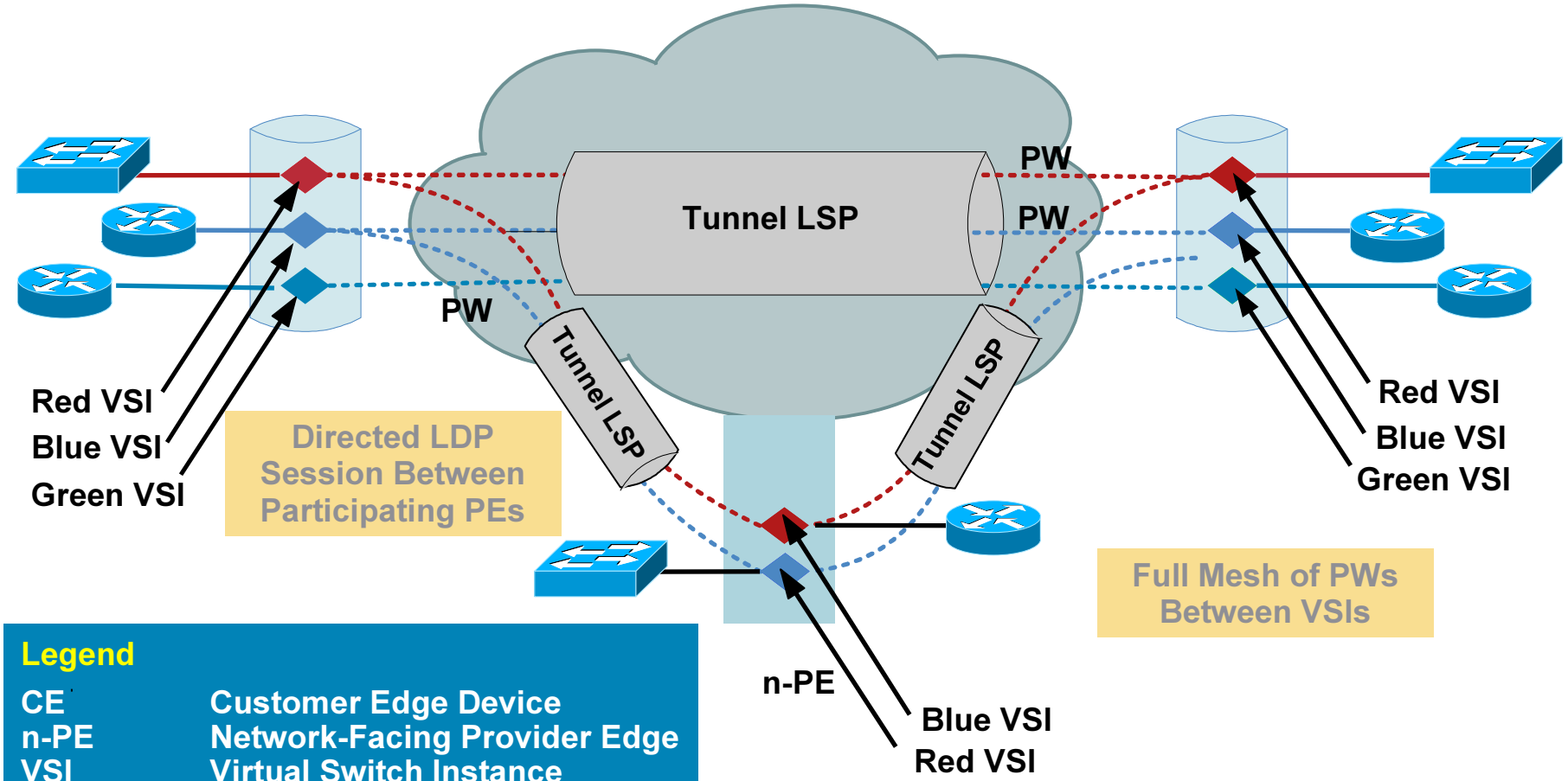
- Spanning Tree
- QinQ & MACinMAC (802.1ad & 802.1ah)
- L2TPv3
- MPLS (EoMPLS & VPLS)
- VSS/vPC
- CEE (aka DCE)/TRILL
- OTV

EoMPLS – Basic Functions



- MPLS transport in the core (needs only LDP, no MP-BGP)
- Targeted (aka directed) LDP session between PEs
- Targeted LDP session distributes pseudowire (PW aka VC) labels
- Port, Sub-Interface Modus - transparent for BPDU - flooding
- SVI Modus - participate in STP - MAC learning – local switching

VPLS Components



Legend

CE	Customer Edge Device
n-PE	Network-Facing Provider Edge
VSI	Virtual Switch Instance
PW	Pseudowire
Tunnel LSP	Tunnel Label Switch Path That Provides PW Transport

VPLS: Layer 2 Forwarding Instance *"VFI"*

A Virtual Switch MUST operate like a conventional L2 switch!

Flooding / Forwarding:

- MAC table instances per customer and per customer VLAN (L2-VRF idea) for each PE
- VSI will participate in learning, forwarding process
- Uses Ethernet VC-Type defined in pwe3-control-protocol-xx

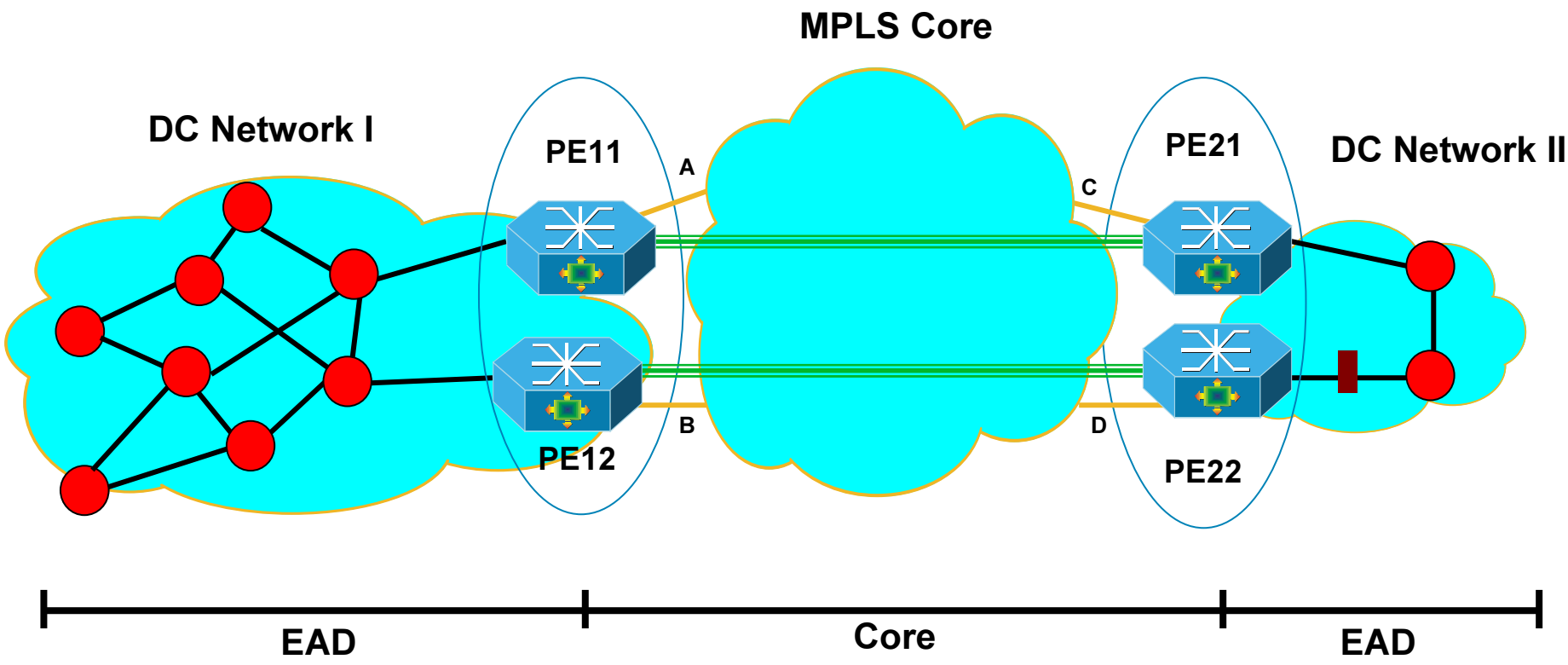
Address Learning / Aging:

- Self Learn Source MAC to port associations
- Refresh MAC timers with incoming frames
- New additional MAC TLV to LDP

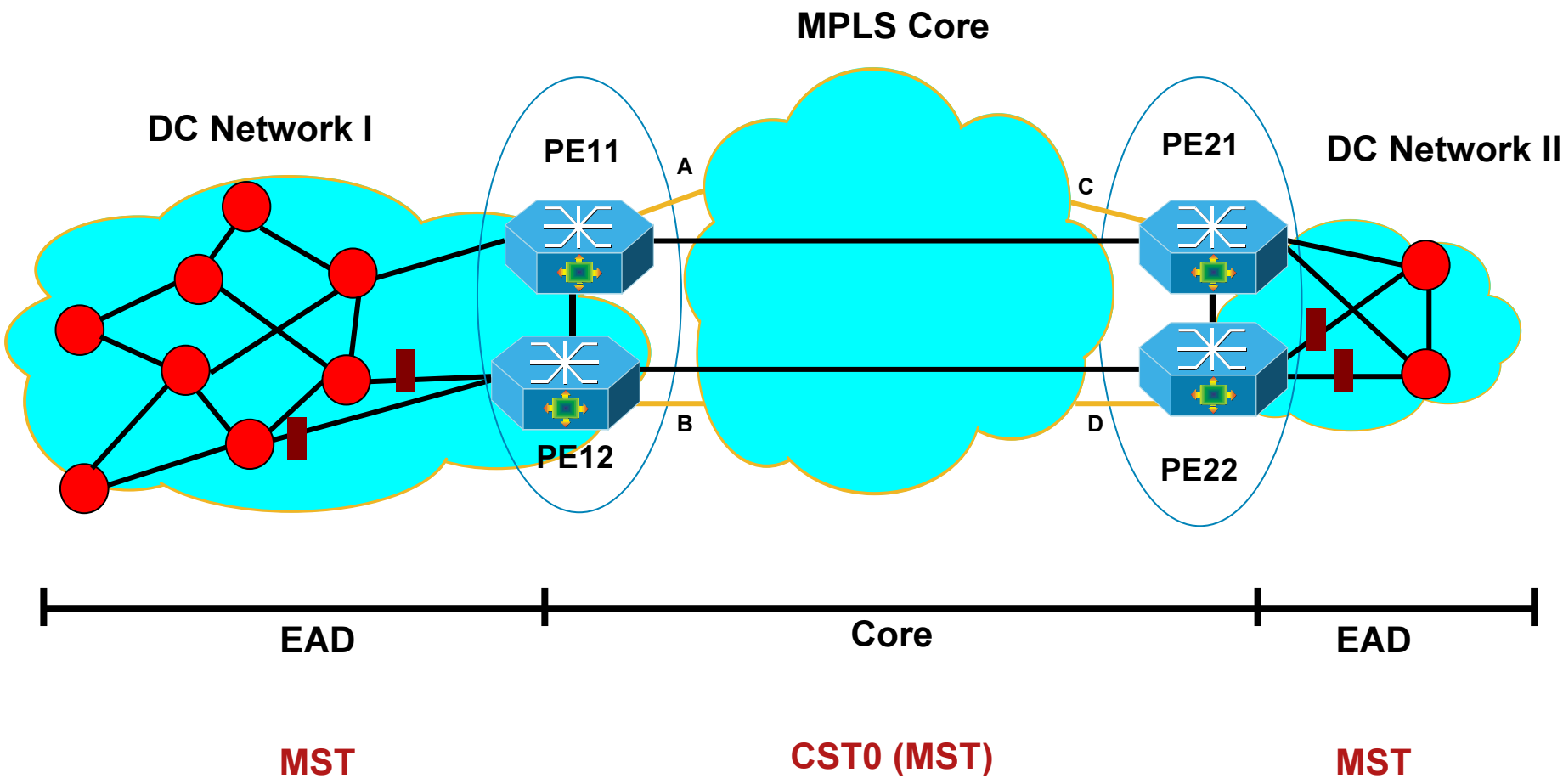
Loop Prevention:

- Create partial or full-mesh of EoMPLS VCs per VPLS
- Use "split horizon" concepts to prevent loops
- Announce EoMPLS VPLS VC tunnels

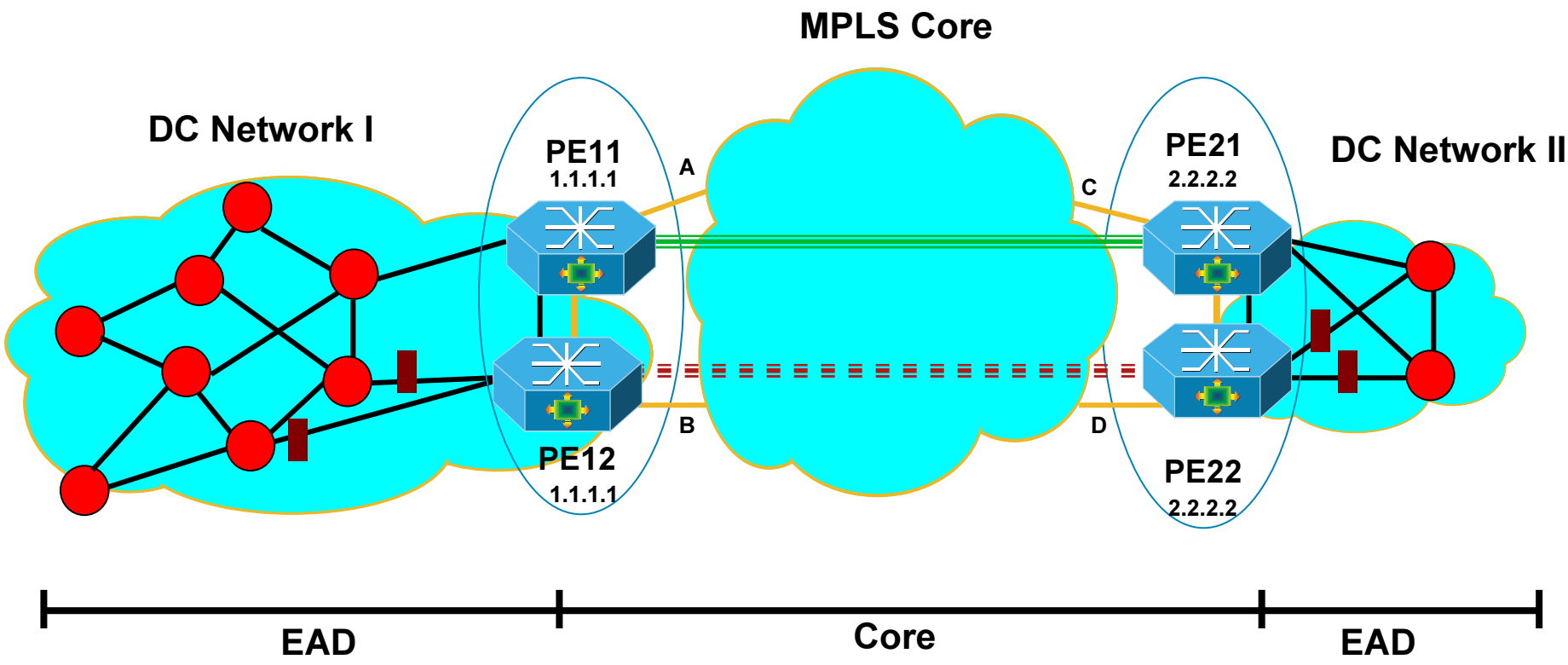
Design 1: STP w/ EoMPLS



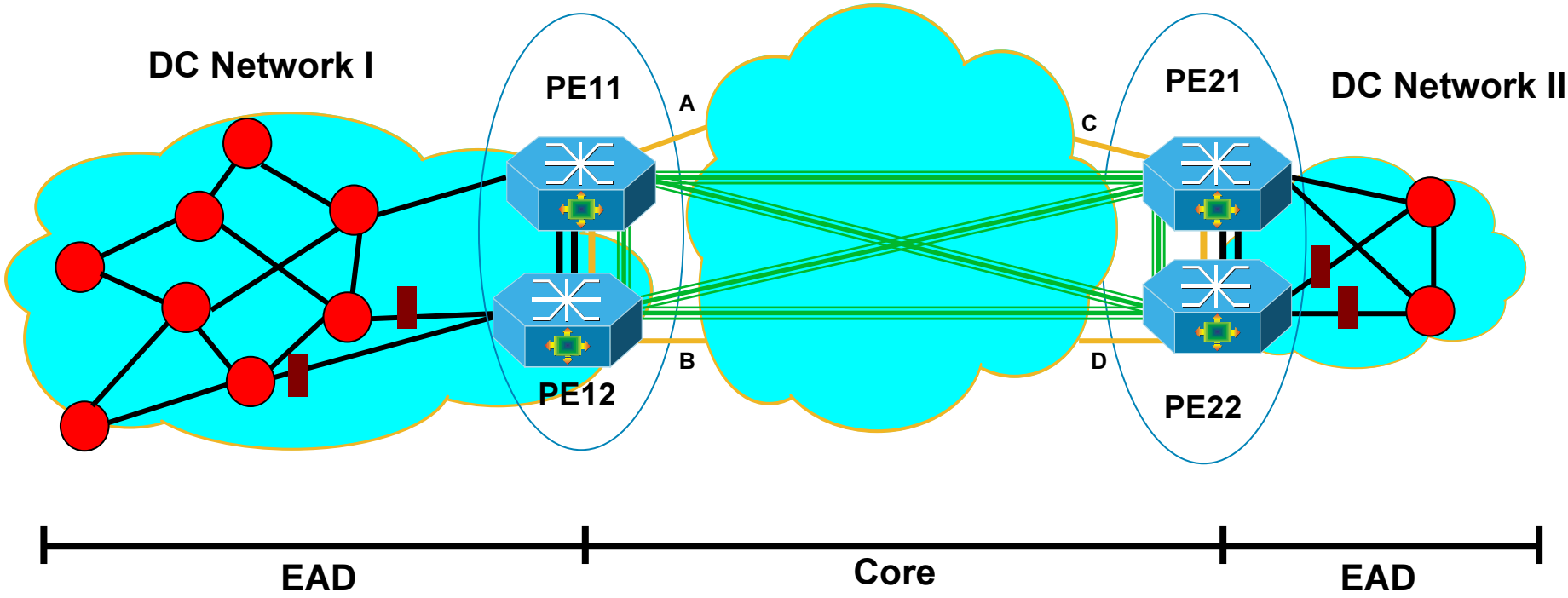
Design 2: (MST Core – STP Design)



Design 3: AnyCast PW (Patrice)



Design 4: Khalil



RST

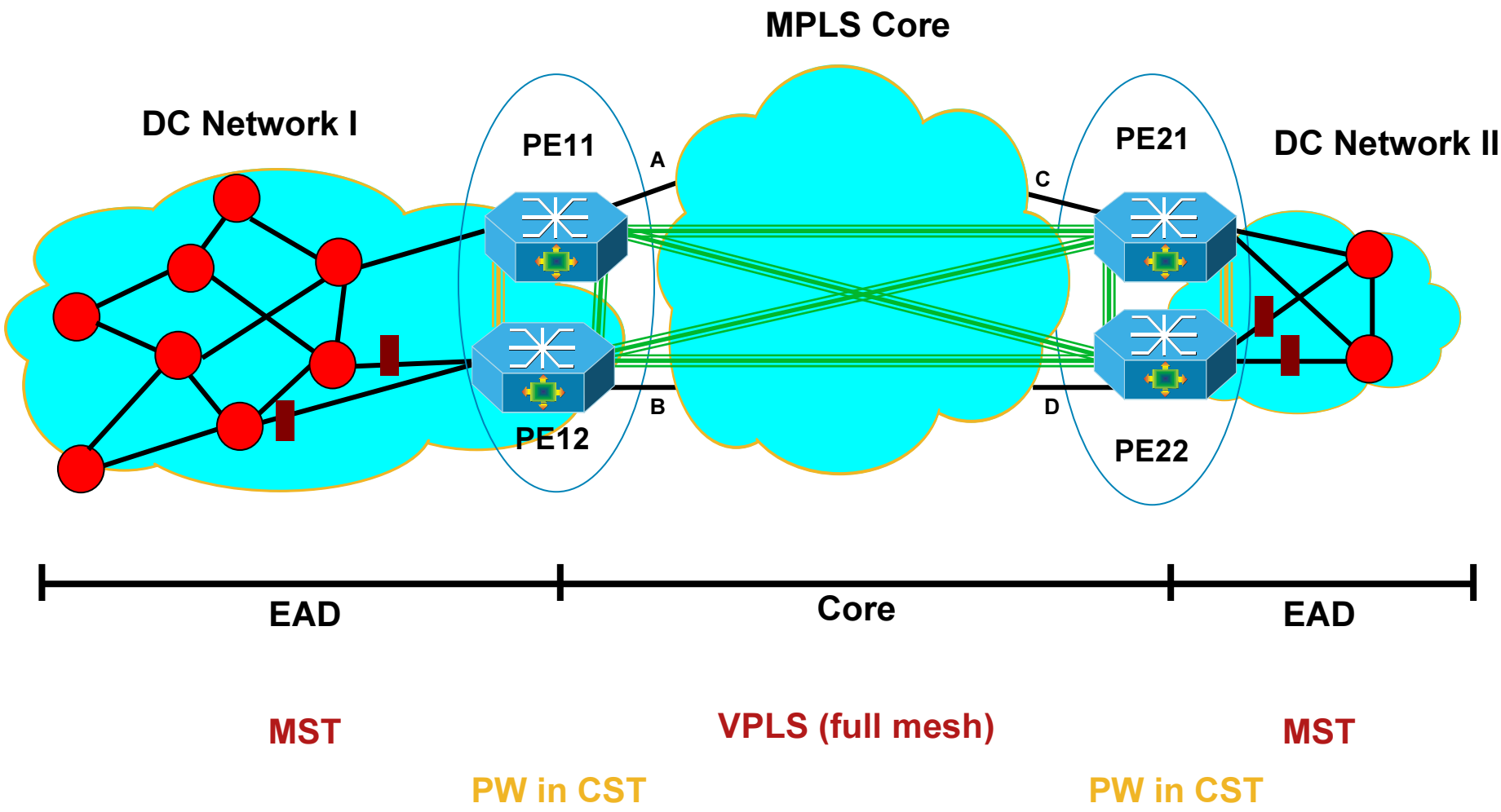
VPLS (full mesh)

RST

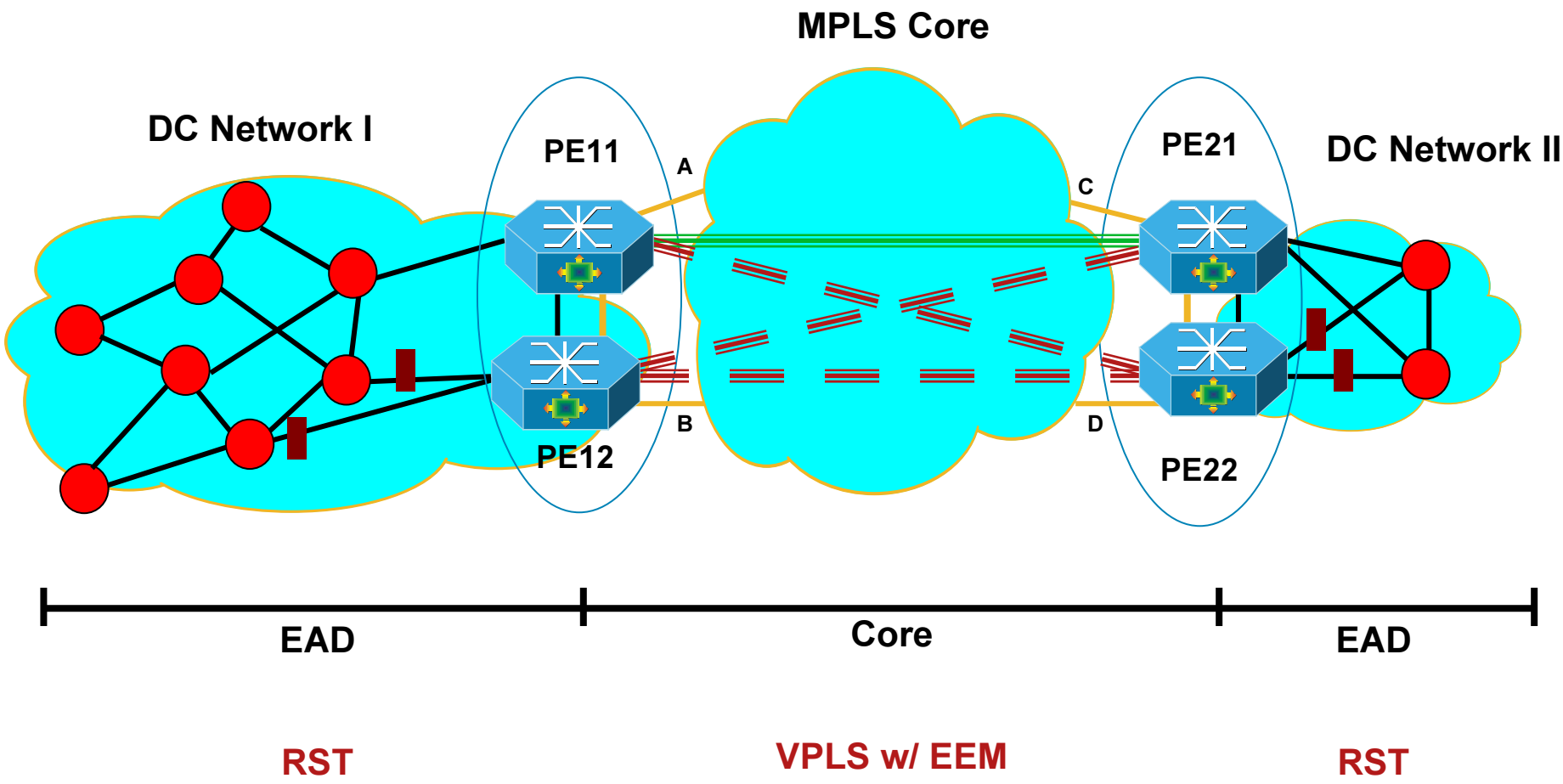
EC w/ native VLAN
in CST (MST0)

EC w/ native VLAN
in CST (MST0)

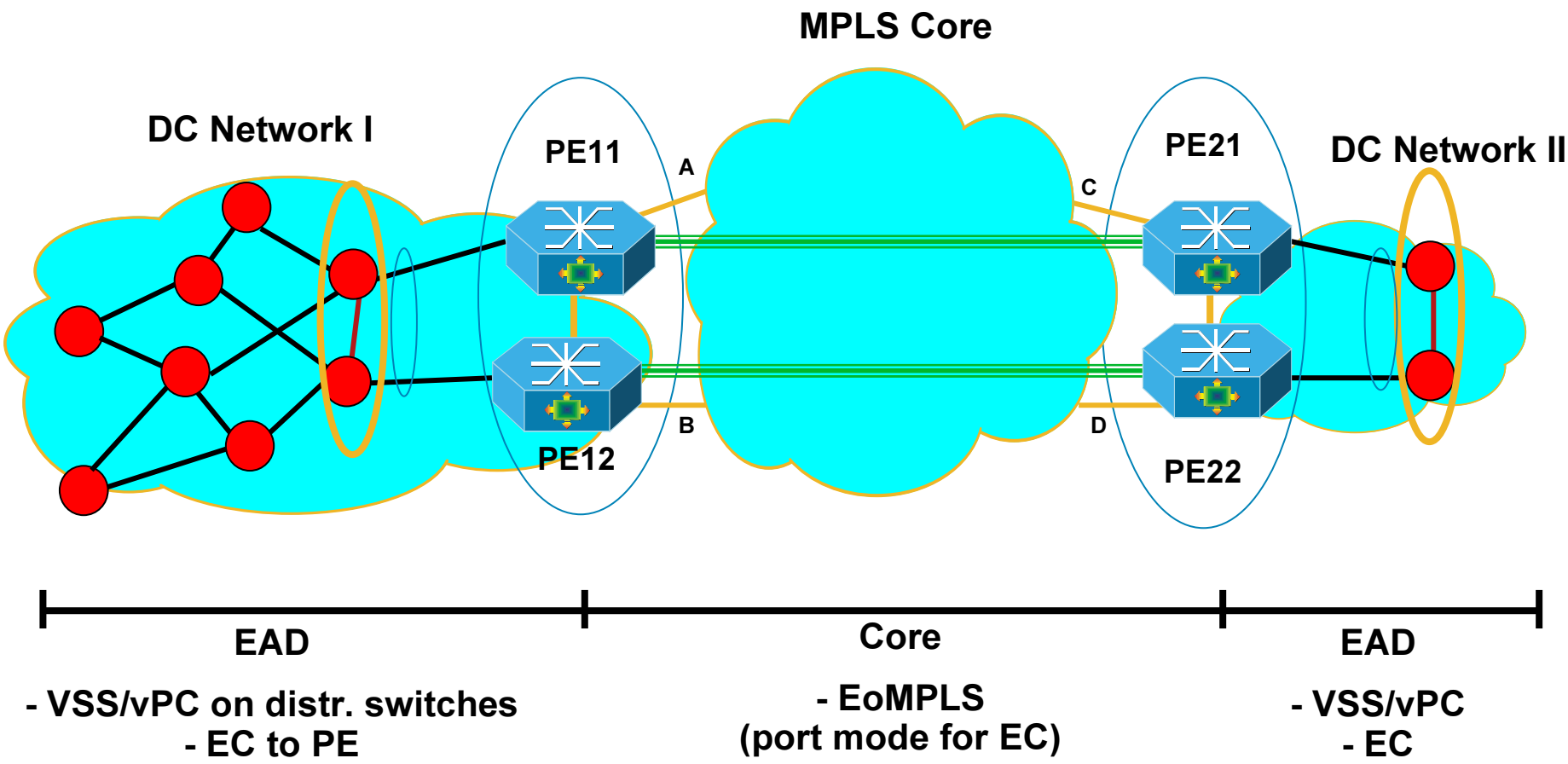
Design 5: Dennis



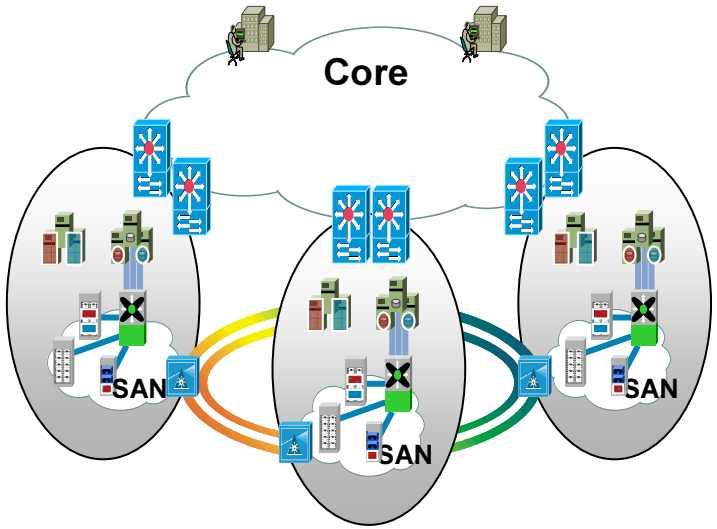
Design 6: VPLS w/ EEM (Patrice)



Design 7: VSS/vPC over EoMPLS

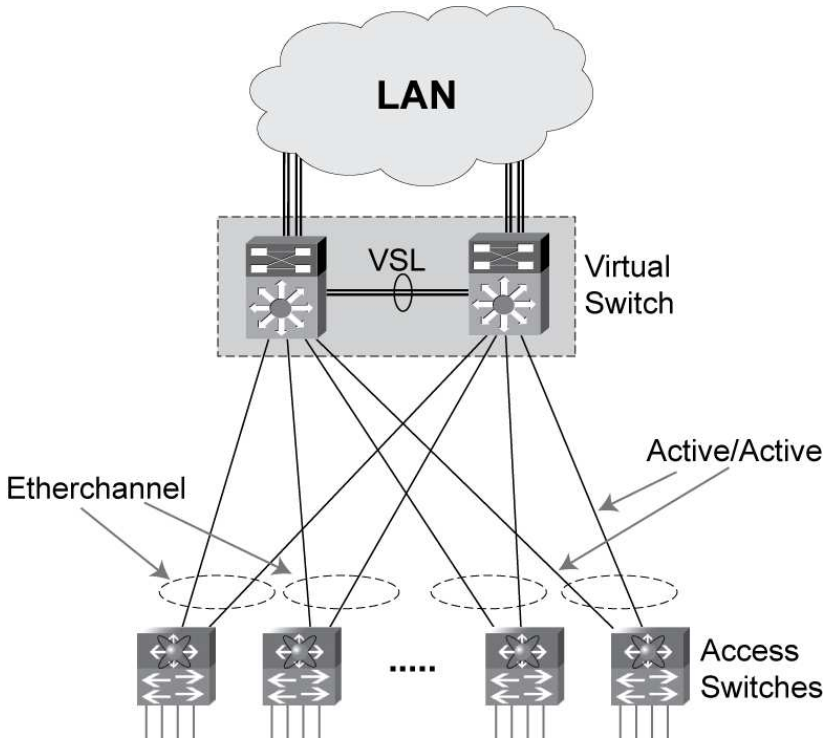


Solutions Overview - Agenda



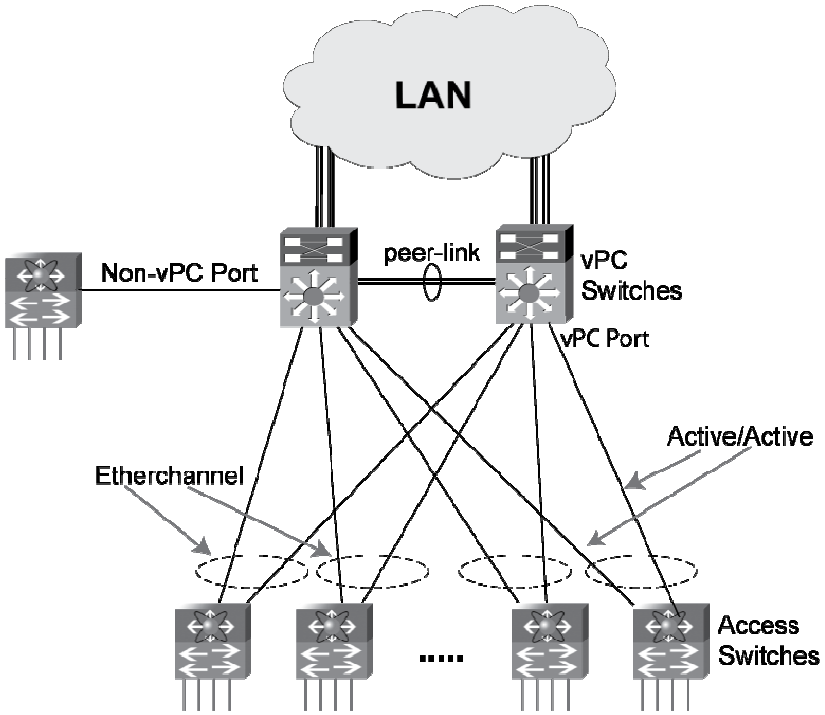
- Spanning Tree
- QinQ & MACinMAC (802.1ad & 802.1ah)
- L2TPv3
- MPLS (EoMPLS & VPLS)
- VSS/vPC
- CEE (aka DCE)/TRILL
- OTV

VSS (Virtual Switching System)



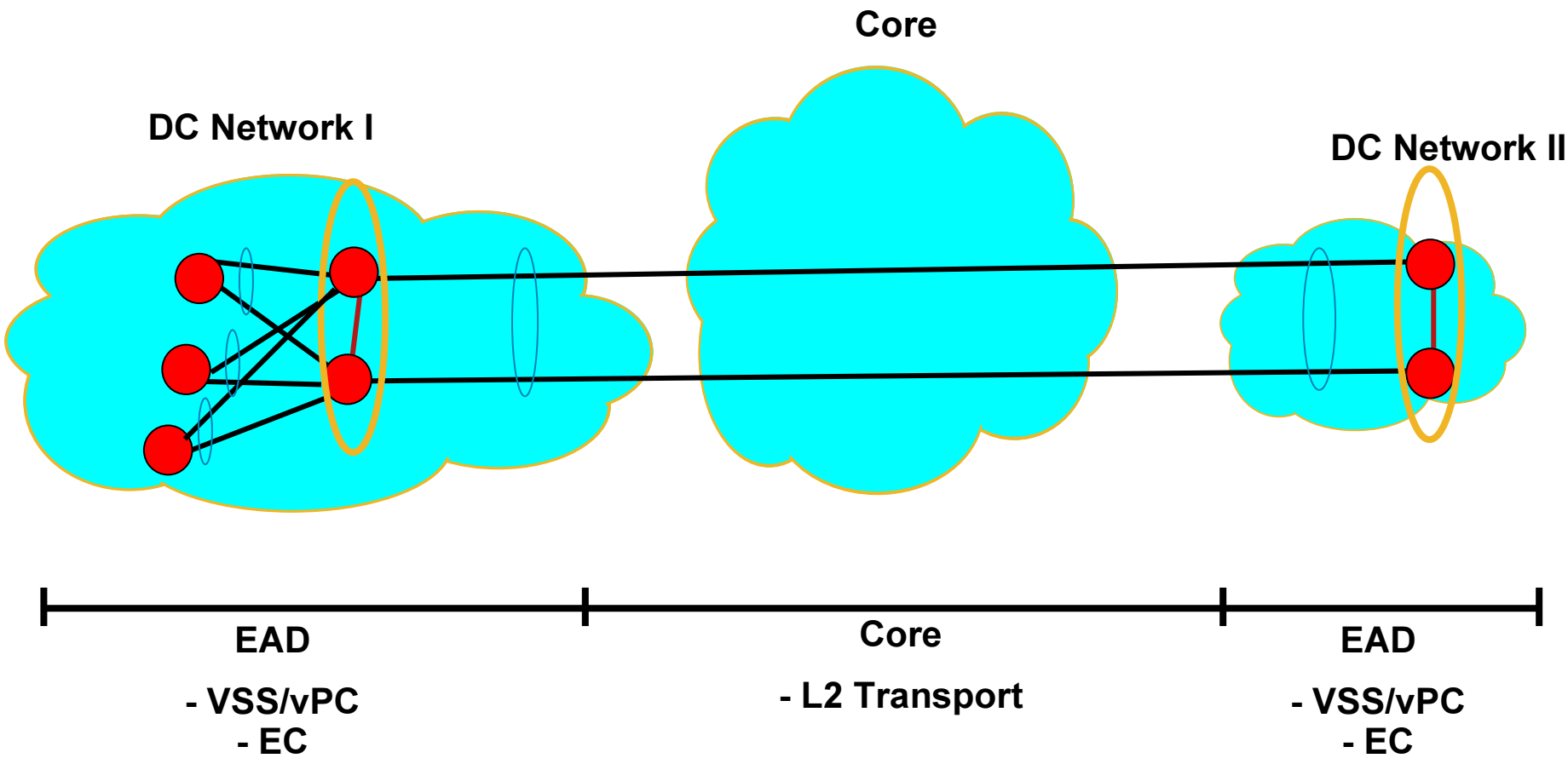
- Cluster two physical chassis
  - Tightly coupled, they become "undistinguishable"
  - Etherchannel from the access switches works unchanged
- Data Plane
  - both switches are active
- Control Plane
  - Active / Stand-by
- Single point of management

vPC (virtual Port Channel)



- Aka MCEC (Multi-Chassis Etherchannel)
  - Etherchannel from the access switches works unchanged
- Loosely Coupled
- Data and Control Plane
  - Both switches are active
- Two points of management
- Peer-Link is a standard link

VSS/vPC (native)

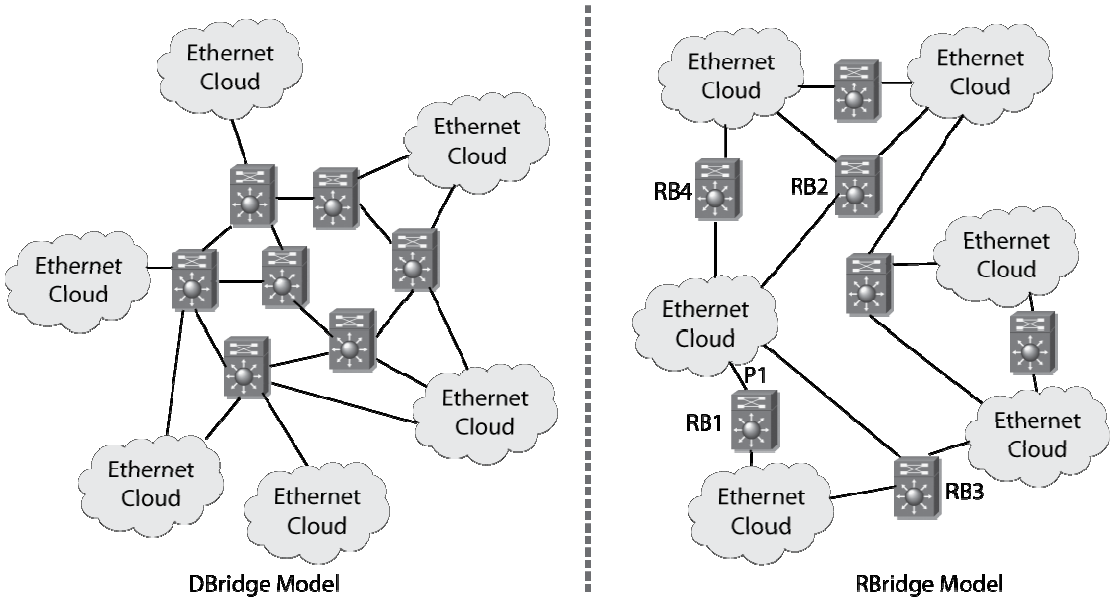


CEE/TRILL – more Details

L2MP Terminology

- Cisco L2MP
  - Aka Cisco DCE or now CEE
  - An internal Cisco effort to provide L2MP
  - Devices that support Cisco L2MP are called “DBridges”
- TRILL (TRansparent Interconnect of Lots of Links)
  - An IETF project
    - http://www.ietf.org/html.charters/trill-charter.html
  - Devices that support Cisco L2MP are called “RBridges”
- Extremely similar – interoperability is possible
  - The term “D/Rbridge” is used when there is commonality

DBridge and RBridge Models

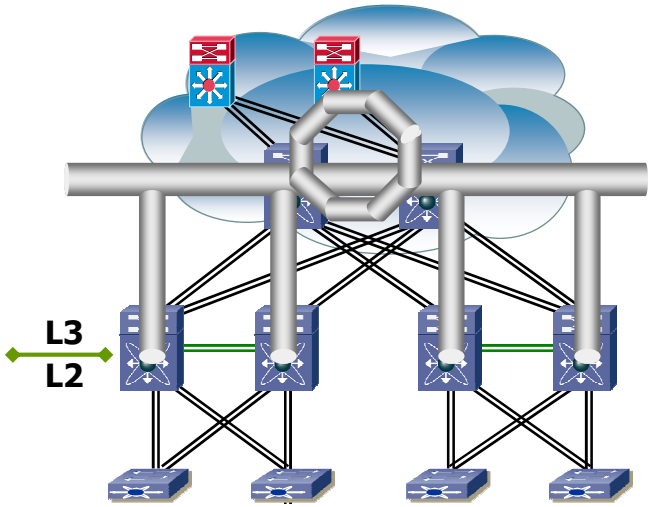


- DBridges are designed to form a L2MP core
  - They are interconnected by point-to-point links
  - They connect the legacy Ethernet cloud at the periphery
- RBridges can be intermixed with legacy Ethernet clouds

OTV – Deeper Dive

Overlay Transport Virtualization

- IP based Ethernet (L2) VPN solution
  - MAC routing
  - IP encapsulated forwarding
- 'Smart' Ethernet Pseudo-node
- Core and Site Transparency



Protocol Learning

Built-in loop prevention (no STP)

Failure domain is bound

Floods/b-casts can be suppressed

Seamless adds/removes & multi-homing

Full cross-sectional BW

- Equal cost multi-pathing
- All-active multi-homing

Packet Switching

Multi-point connectivity

Minimal state

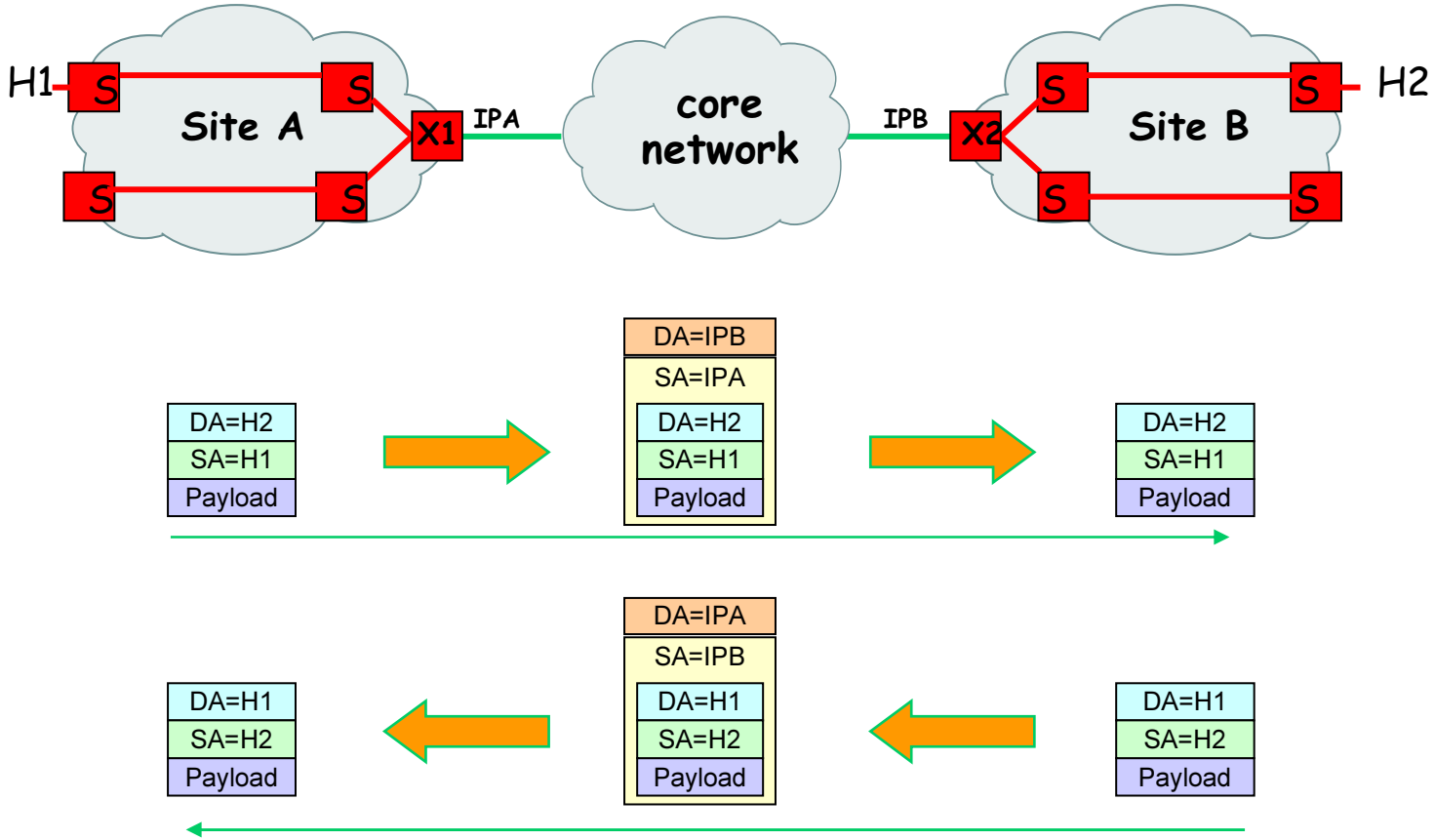
- No PW state preserved

Optimal m-cast replication

Point-to-cloud model for improved manageability

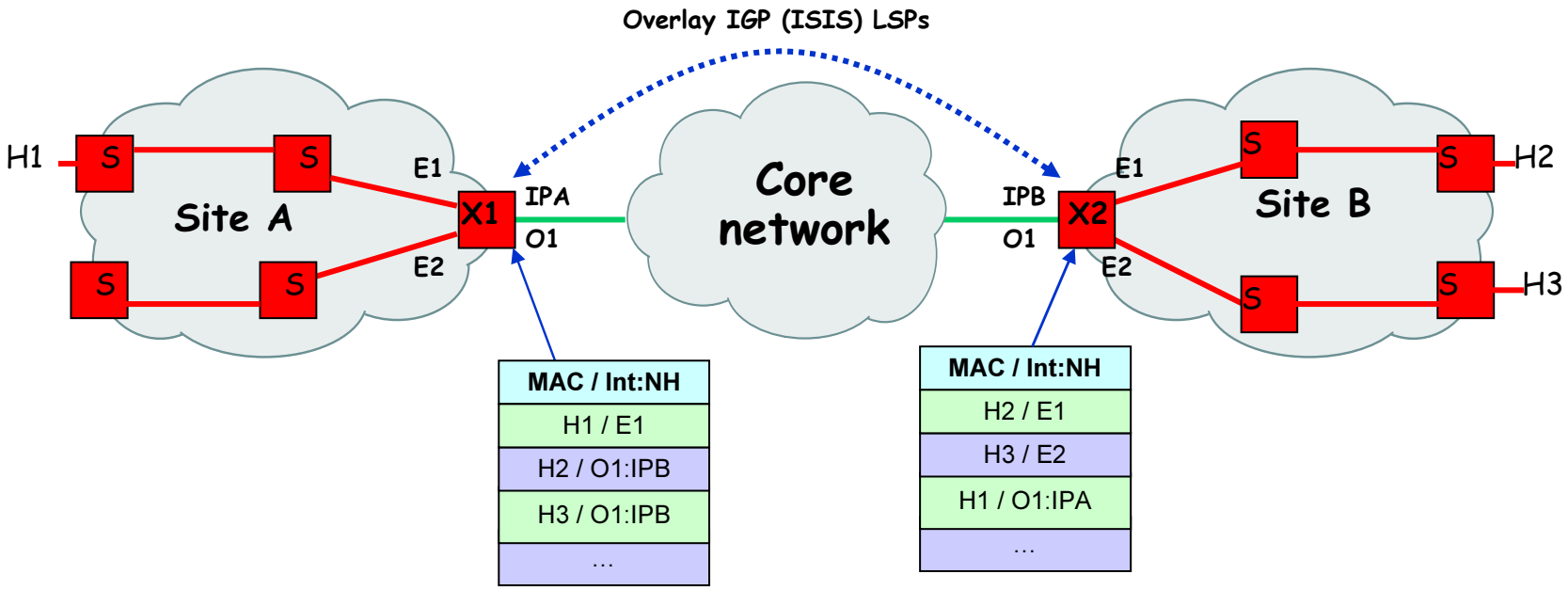
OTV Uses “MAC in IP” Encapsulation

- IP encapsulation → Core Transparency



OTV – MAC Bridging + MAC Routing

- Sites: MAC Bridging → Site Transparency
- Core: “MAC routing” via “overlaid IGP”
- MAC table: “destination ports” and “IP routes”



Key Take Aways

Key Take Aways

- Plan today: VSS/vPC
- Avoid traps w/ VPLS (STP, EoMPLS, …)
- Design Nexus 7000 for DC (and Campus)
- OTV - discuss, test, pilot, install ☺
- Migration to CEE/TRILL later

- Start discussion w/ your account team today!!!

Thank you!!!

Gerd Pflueger – CSE R&S Germany

gerd@cisco.com