

Propojování datových center (Data Center Interconnect)

ARCH3 / L3

Miroslav Brzek

Systems Engineer, Cisco

mibrzek@cisco.com

Prosíme, ptejte se nás

- Twitter www.twitter.com/CiscoCZ
- Talk2cisco www.talk2cisco.cz/dotazy
- SMS 721 994 600



Agenda

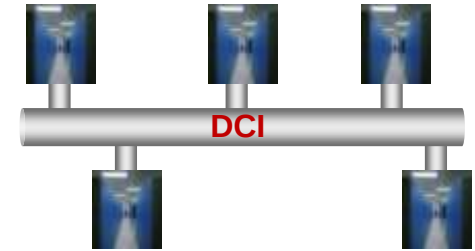
- DCI Business Drivers and Solutions Overview
- SAN Extension Solutions
- LAN Extension Deployment Scenarios
 - Ethernet Based Solutions
 - MPLS Based Solutions
 - IP Based Solutions
- Path optimization
- Conclusions and Q&A



Data Center Interconnect

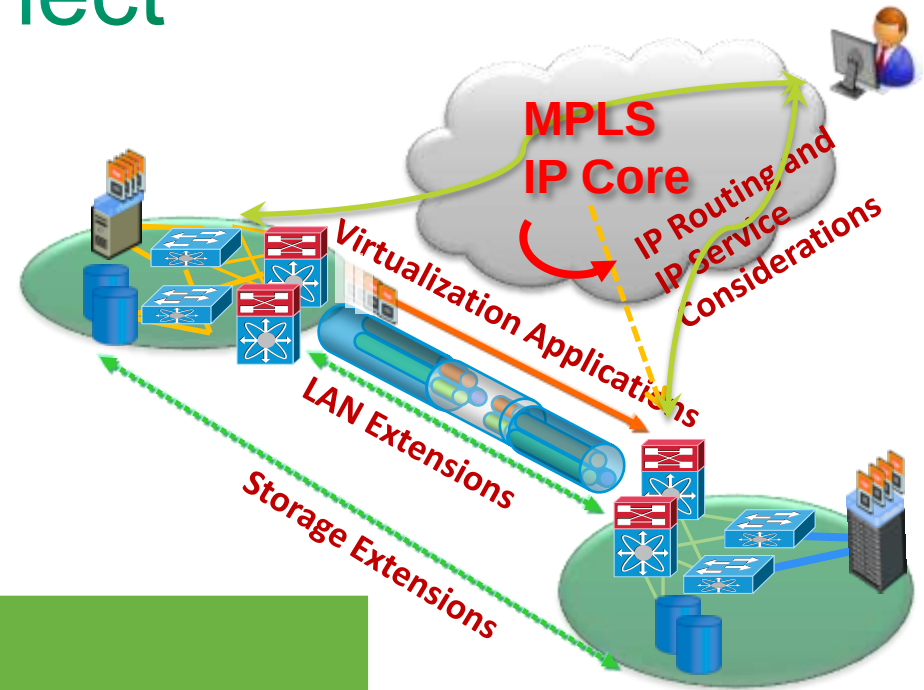
Business Drivers

- Data Centers are extending beyond traditional boundaries
- Virtualization applications are driving DCI across PODs (aggregation blocks) and Data Centers



Drivers	Business Solution	IT Technology
Business Continuity	✓ Disaster Recovery ✓ HA Framework	✓ GSLB ✓ Geo-clusters ✓ HA Cluster
Operation Cost Containment	✓ Data Center Maintenance / Migration / Consolidation	✓ Distributed Virtual Data Center
Business Resource Optimization	✓ Disaster Avoidance ✓ Workload Mobility	✓ VM Mobility
Cloud Services	✓ Inter-Cloud Networking ✓ XaaS	✓ VM Mobility ✓ Automation

Data Center Interconnect Solution Components



DCI Function	Purpose
Storage Extensions	Providing applications access to storage locally, as well as remotely with desirable storage attributes
LAN Extensions	Extend same VLAN across Data Centers, to virtualize servers and applications
Path Optimization	Routing users to the data center where the application resides while keeping symmetrical routing in consideration for IP services (e.g. Firewall)
Inter-DC Routing	Provide routed connectivity between data centers (used for L3 segmentation/virtualization, etc.)

Agenda

- DCI Business Drivers and Solutions Overview
- **SAN Extension Solutions**
- LAN Extension Deployment Scenarios
 - Ethernet Based Solutions
 - MPLS Based Solutions
 - IP Based Solutions
- Path optimization
- Conclusions and Q&A



SAN Extension

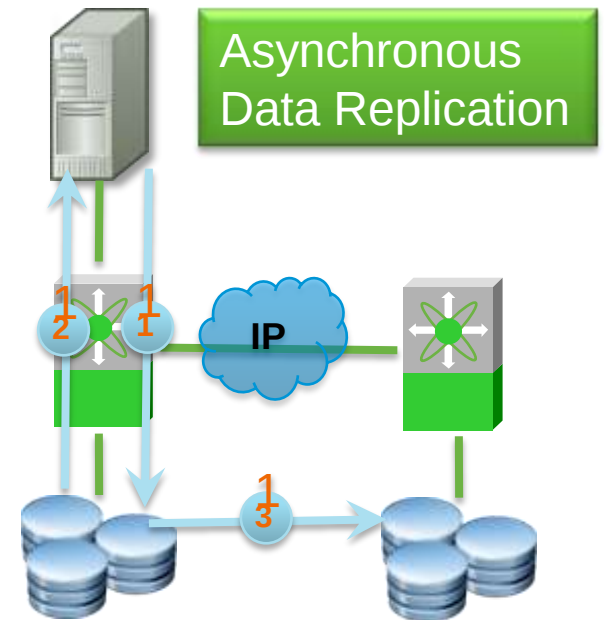
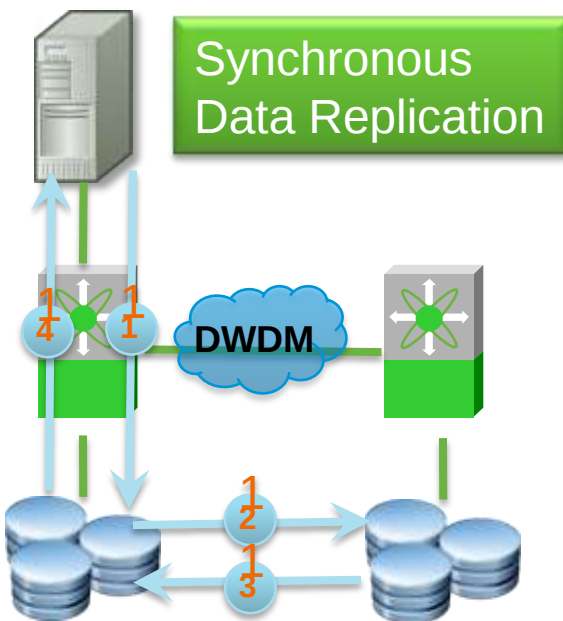
Synchronous vs. Asynchronous Data Replication

- **Synchronous Data replication:** The Application receives the acknowledgement for I/O complete when both primary and remote disks are updated. This is also known as Zero data loss data replication method

Metro Distances (depending on the Application can be 50-300kms max)

- **Asynchronous Data replication:** The Application receives the acknowledgement for I/O complete as soon as the primary disk is updated while the copy continues to the remote disk.

Unlimited distances



Agenda

- DCI Business Drivers and Solutions Overview
- SAN Extension Solutions
- LAN Extension Deployment Scenarios
 - Ethernet Based Solutions
 - MPLS Based Solutions
 - IP Based Solutions
- Path optimization
- Conclusions and Q&A



LAN Extension

Key Technical Challenges

Extending Layer 2 domains across data centers present challenges including, but not limited to:

- STP fault domain isolation
- Achieving High Availability (L2 dual-homing)
- Network loop avoidance, given redundant links and devices without STP
- Bridging data-plane flooding & broadcasting storm control
- Full utilization of cross sectional bandwidth across the Layer 2 domain
- Long distance link protection with fast convergence
- Path diversity
- Multicast optimization
- QoS
- Encryption

LAN Extension

Technology Selection Criteria

Technology Nature

Selection Criteria



Ethernet

➤ *VSS & vPC or FabricPath*

- Multi-Chassis EtherChannel for dual site interconnection
- FabricPath for multi-site deployments
- Over dark fiber or protected D-WDM
- Easy crypto using end-to-end 802.1AE



MPLS

➤ *EoMPLS & A-VPLS*

- L2oL3 for link protection (Fast detection & convergence / Dampening)
- Point-to-Point service model
- Large scale & Multi-tenants
- Works over GRE
- Most deployed today



IP

➤ *OTV*

- L2oL3 for link protection (Fast detection & convergence / Dampening)
- Point-to-cloud service model
- Enterprise focus
- Easy integration over Core, works over any transport
- Innovative MAC routing

Agenda

- DCI Business Drivers and Solutions Overview
- SAN Extension Solutions
- LAN Extension Deployment Scenarios
 - Ethernet Based Solutions
 - VSS/vPC
 - Fabric-Path
 - MPLS Based Solutions
 - IP Based Solutions
- Path optimization
- Conclusions and Q&A



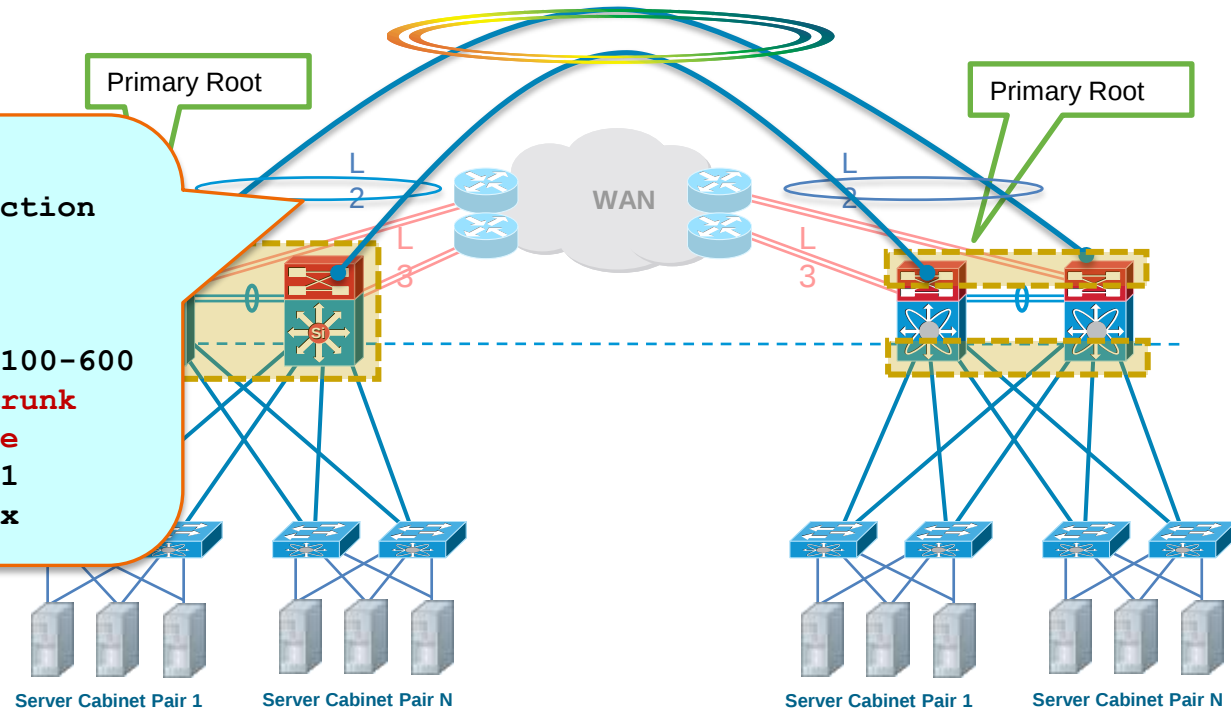
LAN Extension - Dual Sites Interconnection

Leveraging MECs Between Sites

On DCI Etherchannel:

- STP Isolation (BPDU Filtering)
- Broadcast Storm Control
- FHRP Isolation

```
interface port-channel10
desc DCI point to point connection
switchport
switchport mode trunk
vpc 10
switchport trunk allowed vlan 100-600
spanning-tree port type edge trunk
spanning-tree bpdupfilter enable
storm-control broadcast level 1
storm-control multicast level x
```



- VSS or vPC for redundancy/multi-homing
- MEC for loop-prevention/multipathing and STP isolation
 - DCI port-channel: 2 or 4 links
- Requires protected DWDM or Direct fibers

LAN Extension - Cisco FabricPath



Switching

- Easy Configuration
- Plug & Play
- Provisioning Flexibility



Routing

- Multi-pathing (ECMP)
- Fast Convergence
- Highly Scalable



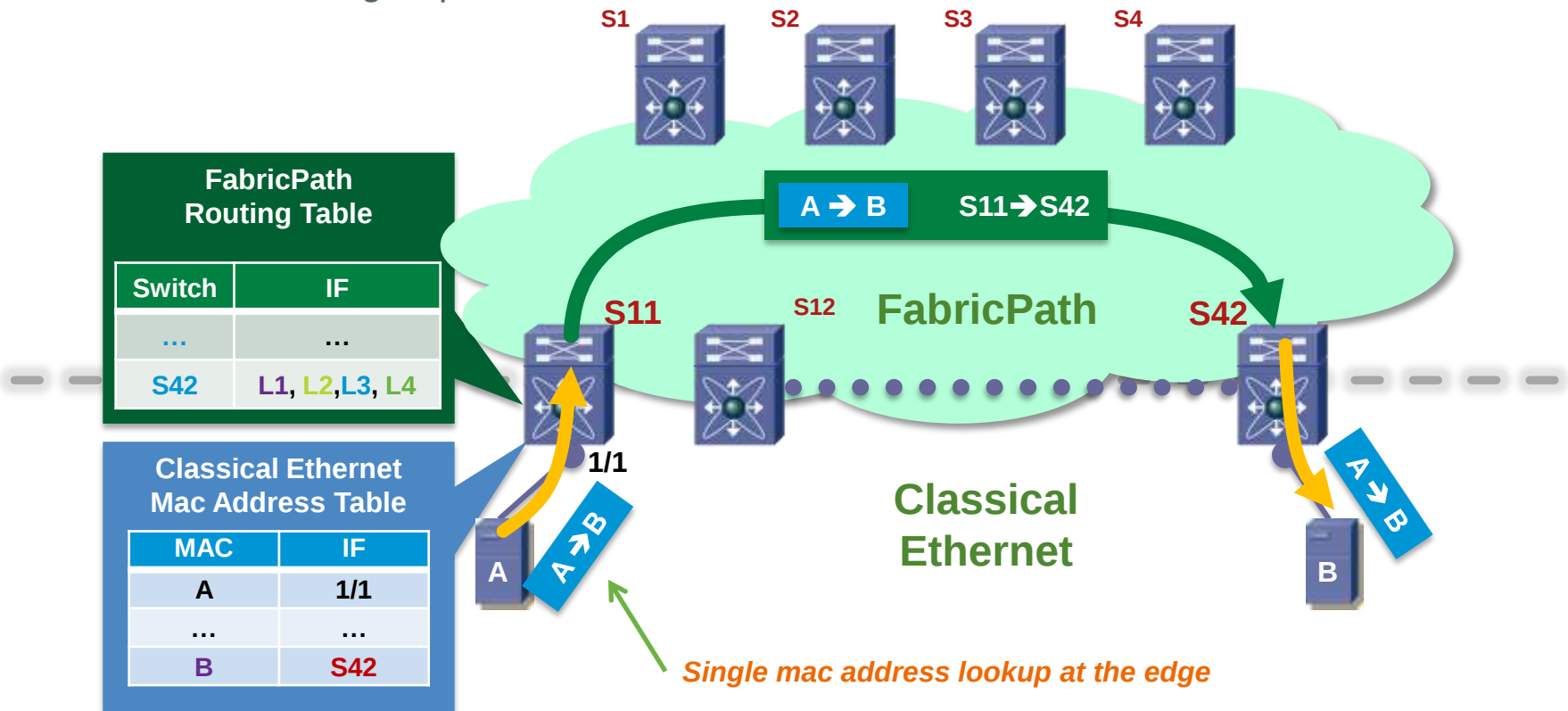
FabricPath

“FabricPath brings Layer 3 routing benefits to flexible Layer 2 bridged Ethernet networks”

Cisco FabricPath is “L2 Routing”

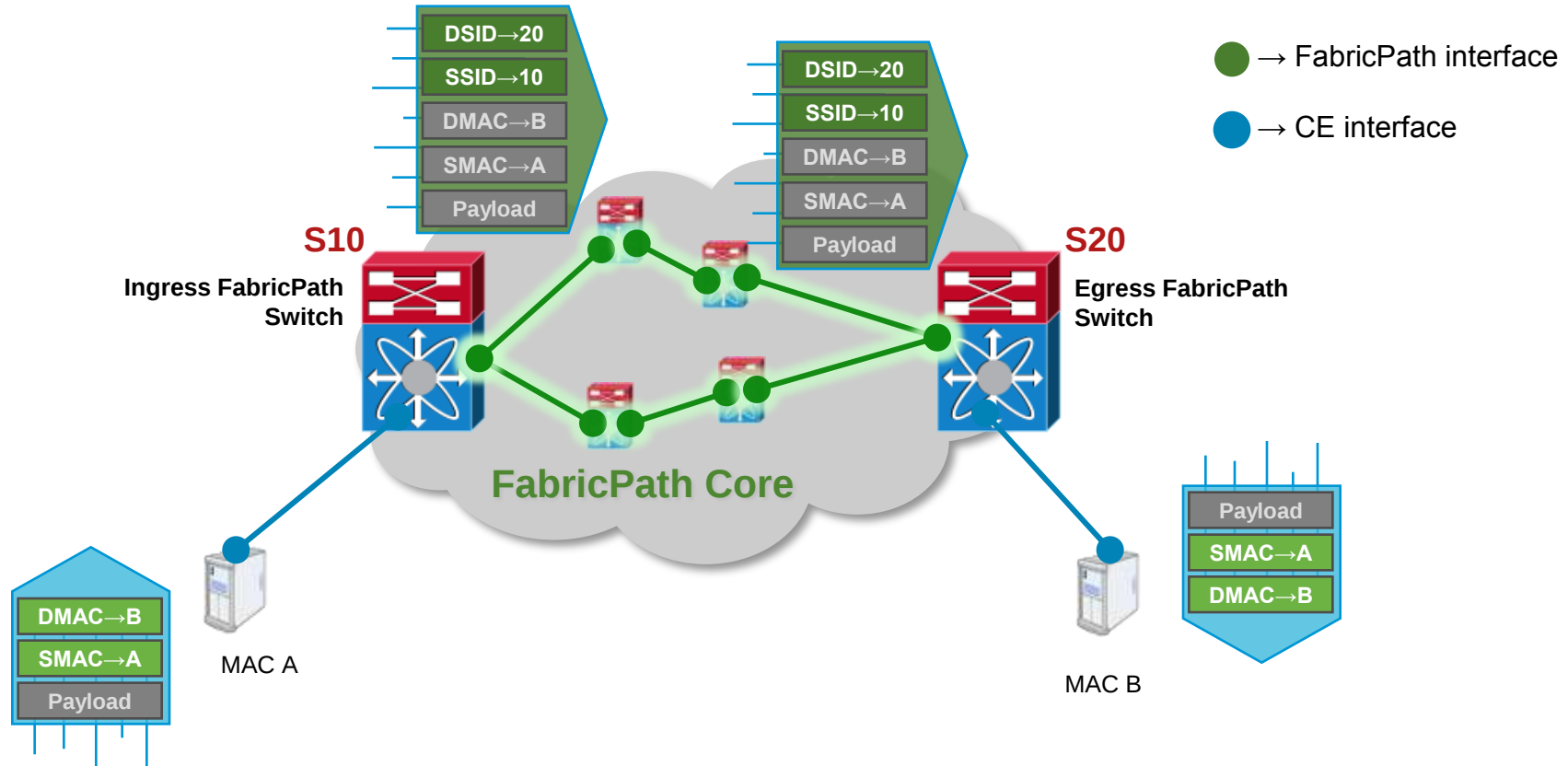
Forwarding decision based on ‘FabricPath Routing Table’

- FabricPath header is imposed by ingress switch
- Only switch addresses are used to make “routing” decisions
- TTL and RPF check the data plane protect against loops
 - L2 can be extended in/accross the data centers (while STP is segmented)
- No MAC learning required inside the L2 Fabric



Cisco FabricPath for LAN Extension

Basic Data Plane Operation



- Ingress FabricPath switch determines destination Switch ID and imposes FabricPath header
- Destination Switch ID used to make routing decisions through FabricPath core
- No MAC learning or lookups required inside core
- Egress FabricPath switch removes FabricPath header and forwards the Ethernet frame

FabricPath

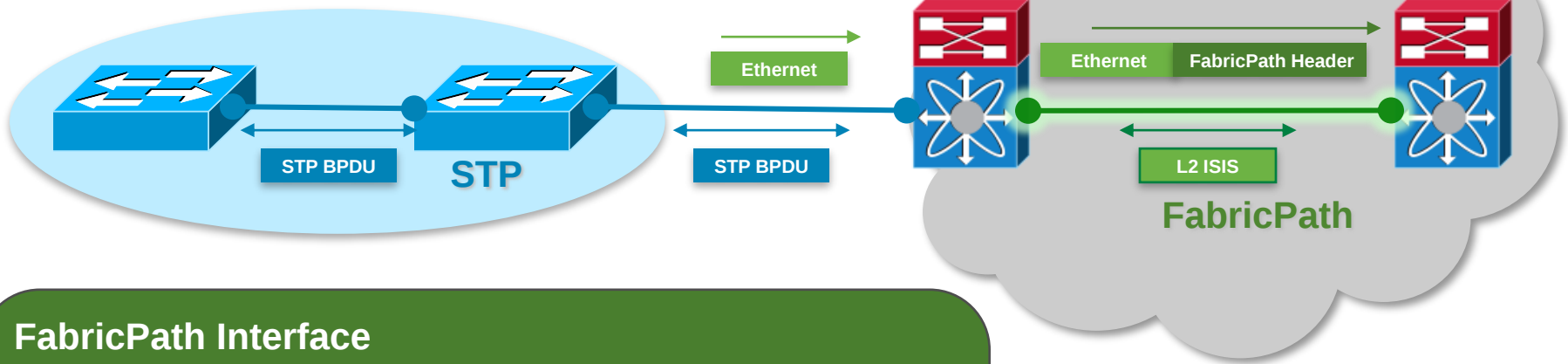
Interaction with Classic Ethernet Interfaces

Classic Ethernet (CE) Interface

- Interfaces connected to existing NICs and traditional network devices
- Send/receive traffic in 802.3 Ethernet frame format
- Participate in STP domain
- Forwarding based on MAC table

● → FabricPath interface

● → CE interface

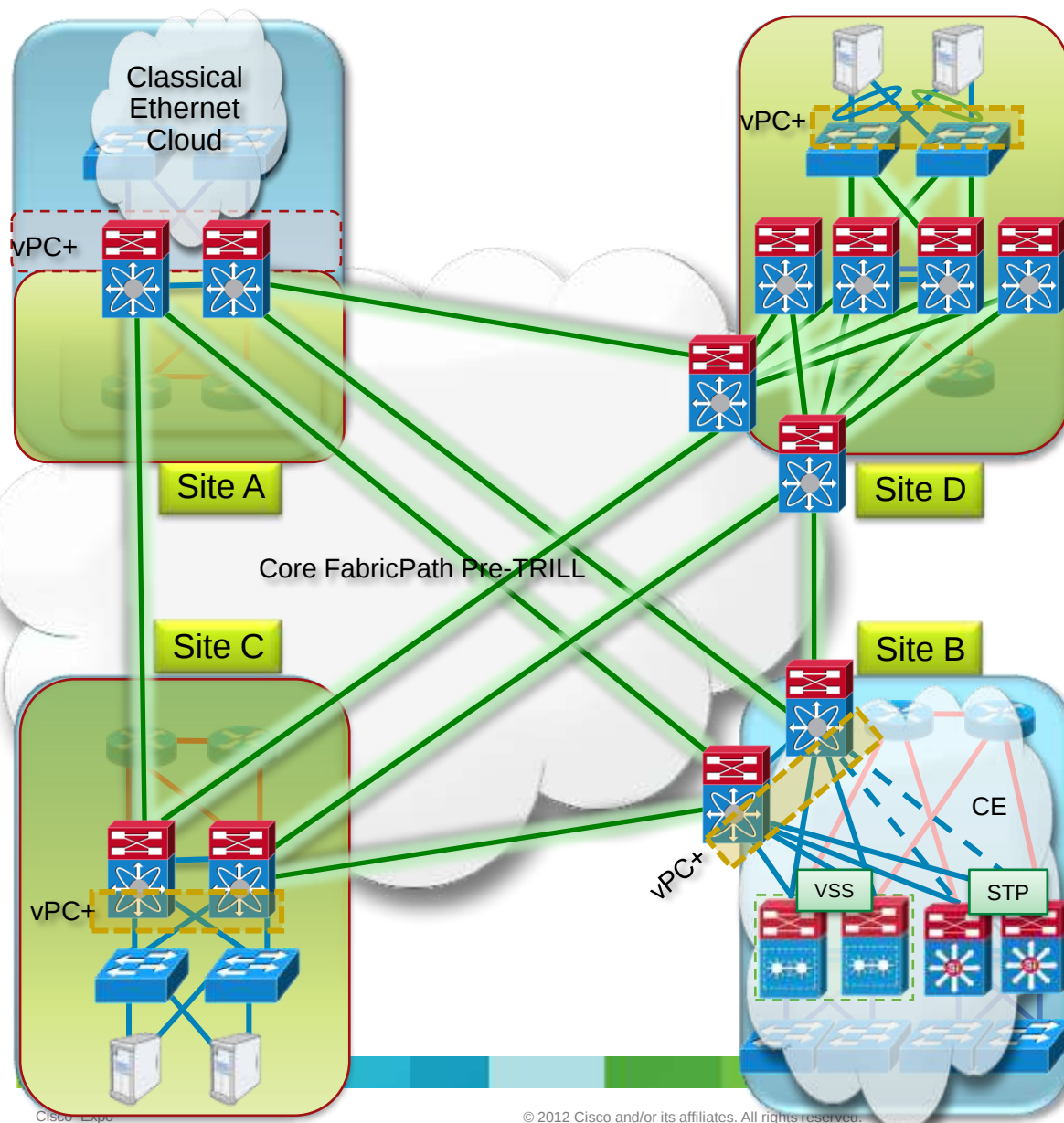


FabricPath Interface

- Interfaces connected to another FabricPath device
- Send/receive traffic with FabricPath header
- No spanning tree!!!
- No MAC learning
- Exchange topology info through L2 ISIS adjacency
- Forwarding based on 'Switch ID Table'

FabricPath for Interconnecting Multiple DC Sites

Partial-Meshed Topology for different models of DC



- Required point to point connections
- Relies on Flooding for Unknown Unicast traffic
- L2 Multipath only for equal cost path can be leveraged (i.e. $A \leftrightarrow B$ or $C \leftrightarrow D$)

- Offer a full HA DCI solution with Native STP Isolation
- Dynamic VLAN pruning
- Provides easy integration with Brownfield DC
- Optimized using vPC+

Cisco FabricPath for LAN Extension

Design Considerations

Benefits

- FabricPath Offers an easy way solution to interconnect multiple DC with STP Isolation
- Optimize bandwidth using Layer 2 Multipath (up to 16 equal cost paths) in full-mesh deployment
- Native Loop free
- Reduce the number of Macs

The FabricPath edge devices learn only based on conversation

The FabricPath Core devices keep only Layer 2 information about the FP edge devices

- FabricPath is transparent to L3 protocols (anything can be carried over the fabric)

Constraints

- Maturity
- Dark Fiber (metro distances)
- Rely on flooding for Unknown Unicast traffic
- Point to point (versus point to cloud) – every device in the path must be FP enabled

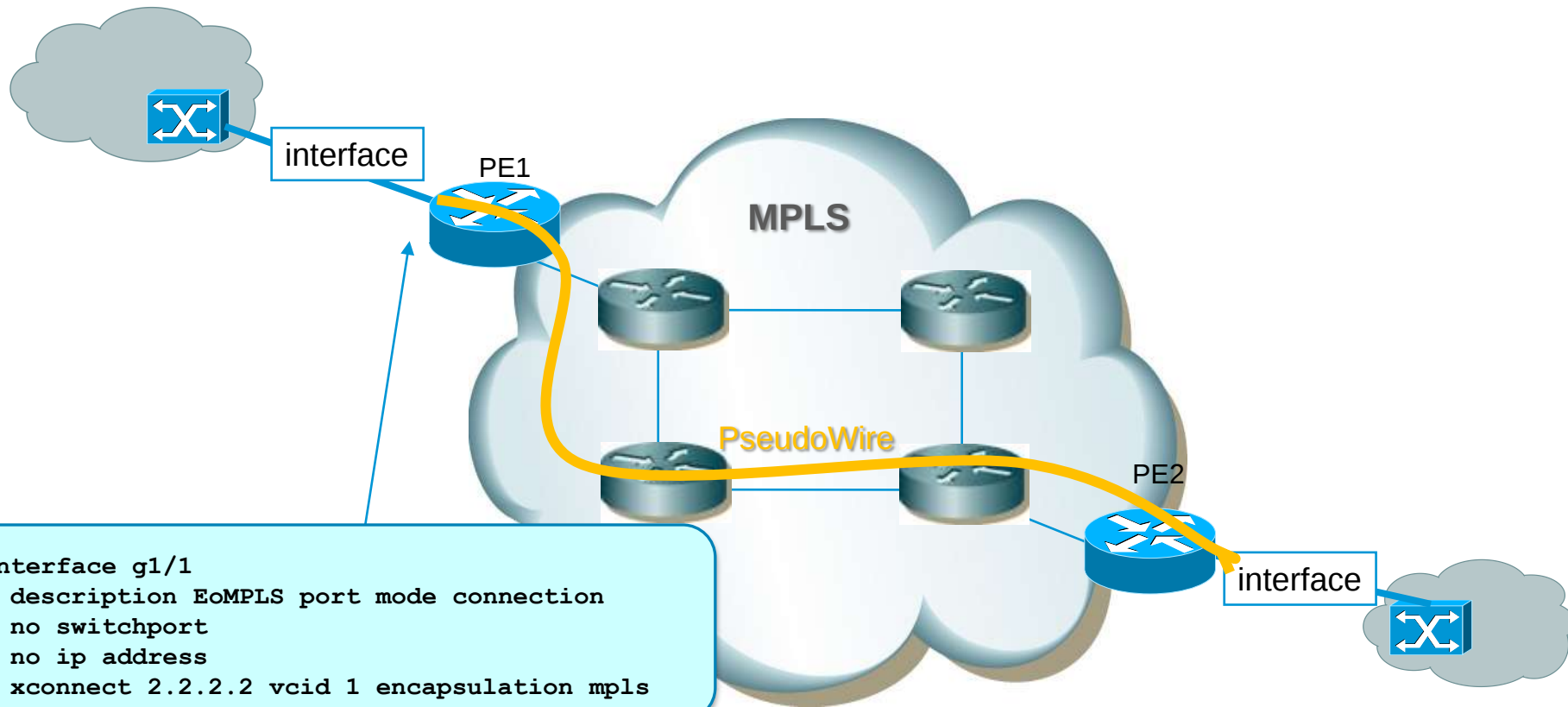
Agenda

- DCI Business Drivers and Solutions Overview
- SAN Extension Solutions
- LAN Extension Deployment Scenarios
 - Ethernet Based Solutions
 - MPLS Based Solutions
 - EoMPLS
 - A-VPLS: VPLS using VSS
 - IP Based Solutions
- Path optimization
- Conclusions and Q&A



LAN Extensions: P2P Topologies

Ethernet over MPLS (EoMPLS)

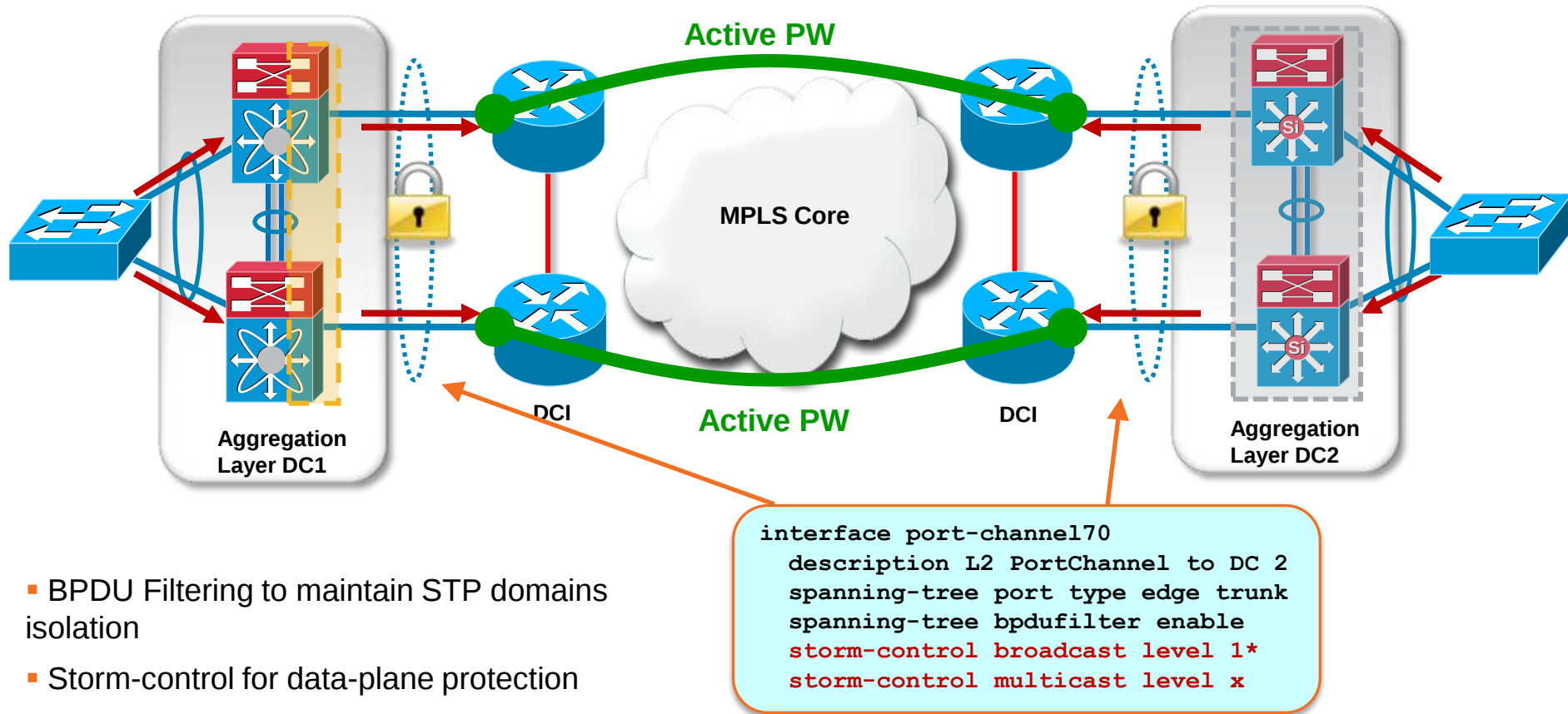


8

1518

EoMPLS for Dual Sites Interconnection

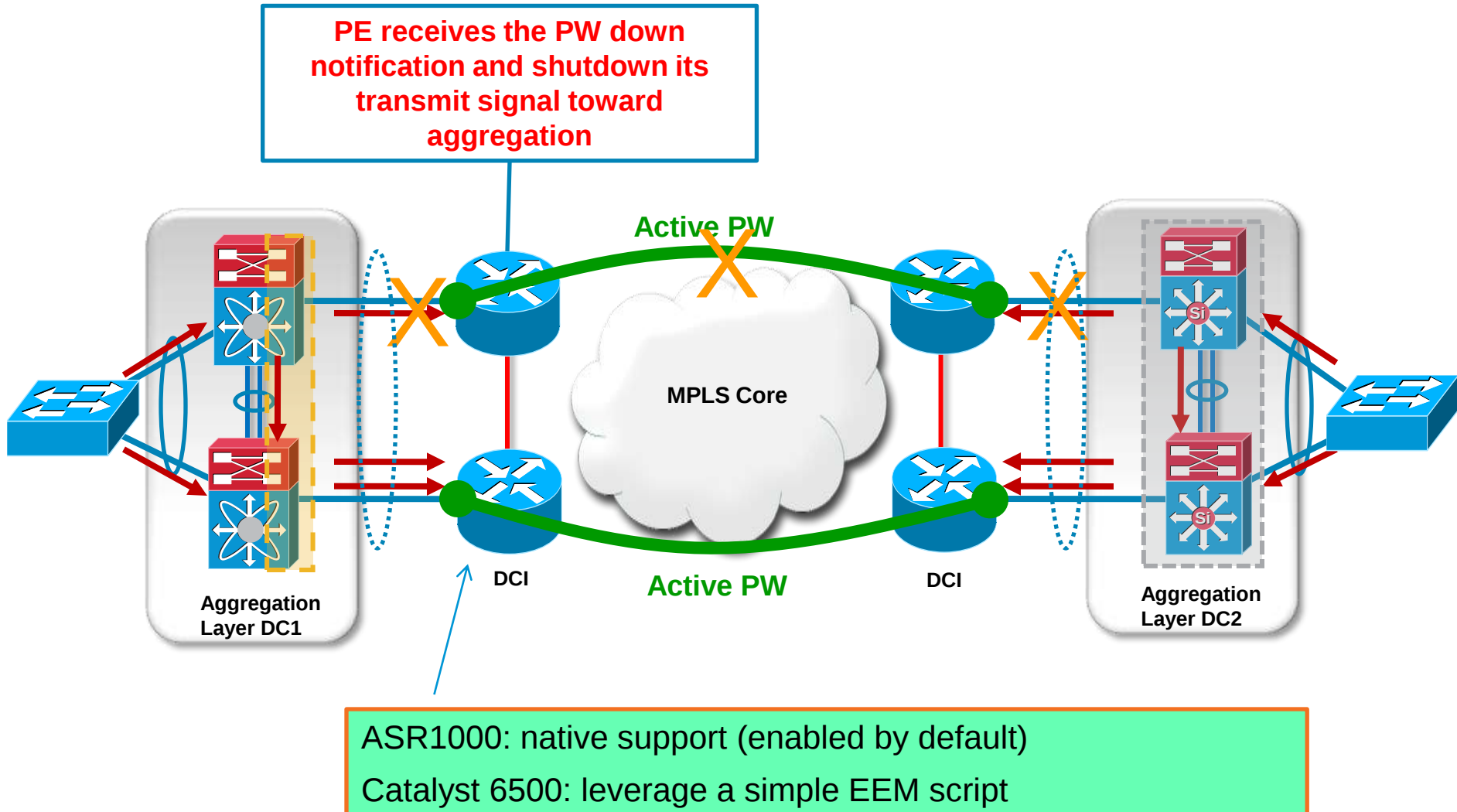
End-to-End Loop Avoidance Using MCEC



- BPDU Filtering to maintain STP domains isolation
- Storm-control for data-plane protection
- Configuration applied at aggregation layer on the logical port-channel interface
- Manual 802.1AE configuration on a physical interface level

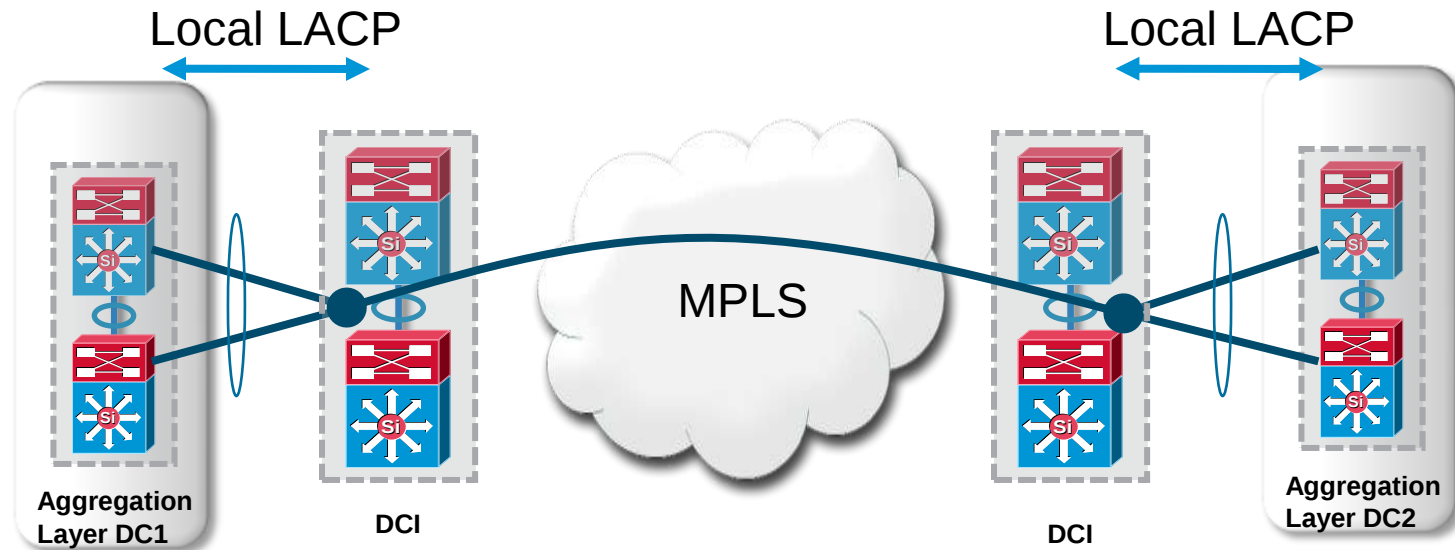
EoMPLS: Dealing with PseudoWire (PW) Failures

Remote Ethernet Port Shutdown



EoMPLS for Dual Sites Interconnection

With Port-Channel xconnect



- Instead of xconnecting physical port, xconnect port-channel
- LACP is kept local, no more extended over EoMPLS
- PW is virtual on both VSS members
- Requires VSS or Nexus as DC device

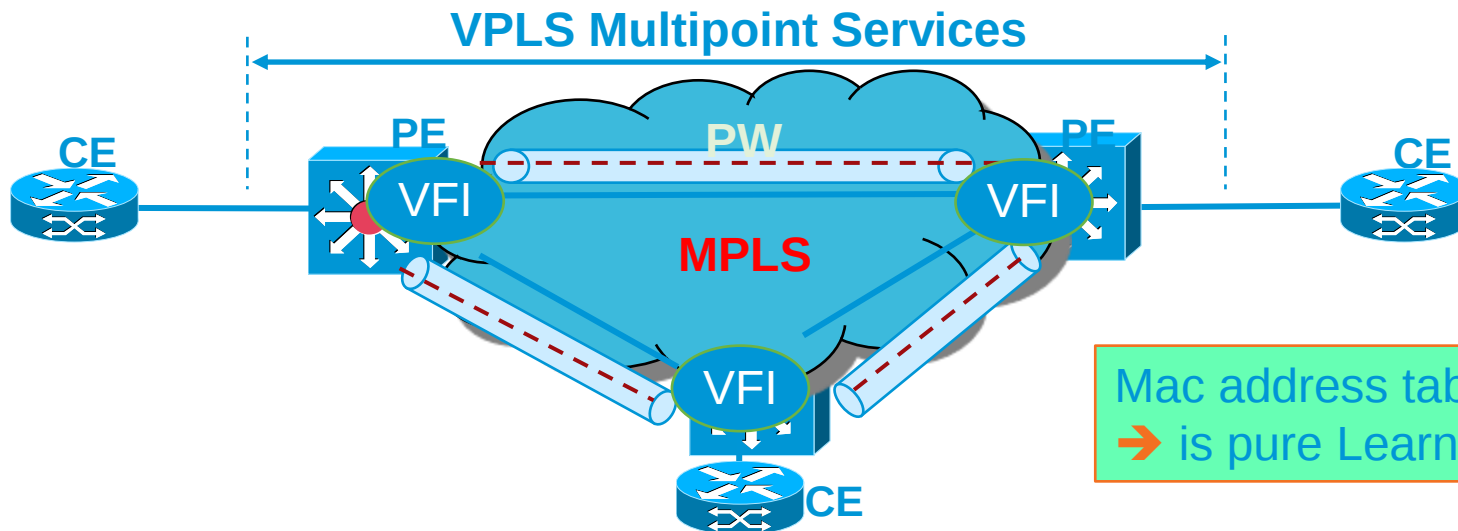
LAN Extensions: Multipoint Topologies

Virtual Private LAN Service (VPLS)

- MPLS Core emulates an IEEE Ethernet bridge (virtual)
- Virtual Bridges (VFI) linked with Pseudowires (PWs)

Assuming PW full-mesh in a VFI

- Split-Horizon for Loop Avoidance in MPLS core (but only for single-homed deployments)
- BPDU are not transmitted by default
- L2 dual-homing is one of the big challenges of VPLS deployment



Mac address table population
→ is pure Learning-Bridge

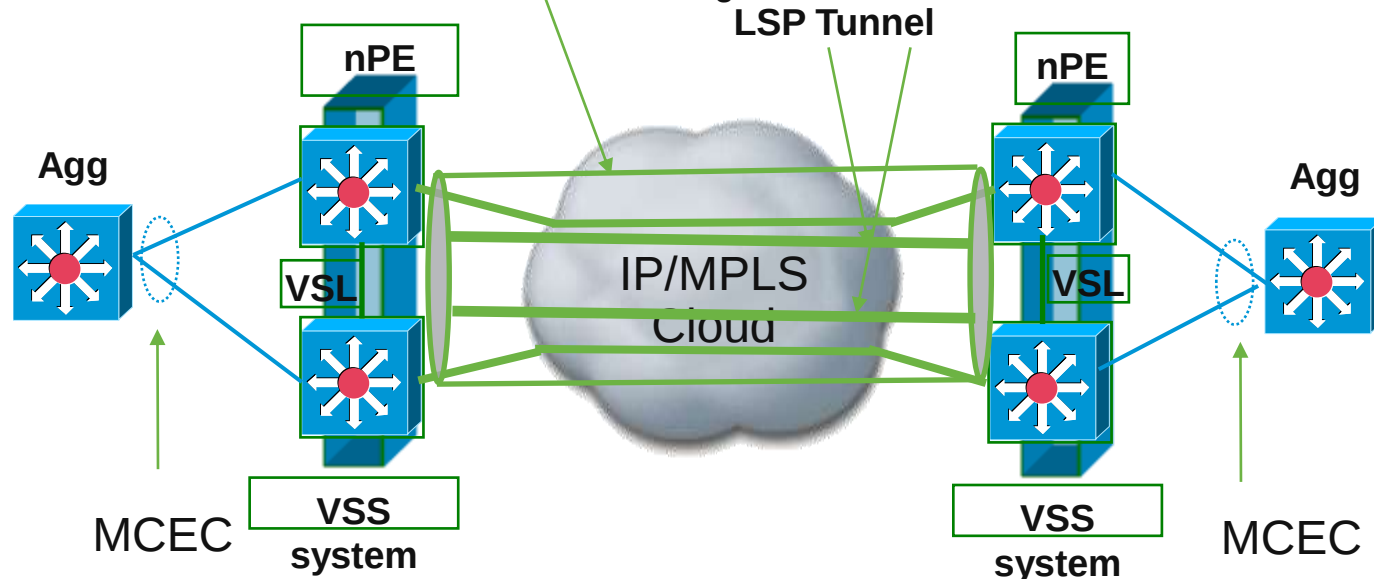
Advanced Virtual Private LAN Service

A-VPLS

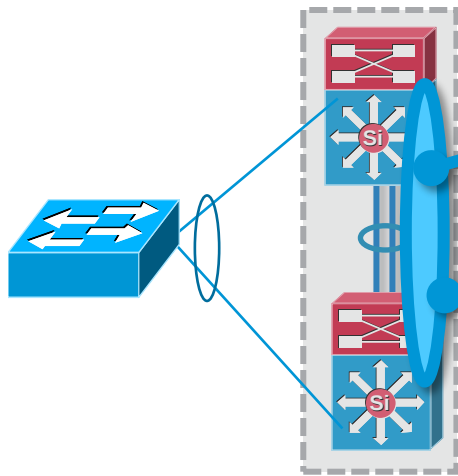
A-VPLS leverages traditional VPLS while adding additional benefits making it a superior solution for Data Center Interconnect deployments

- Simplified redundancy with VSS (dual-homed deployments)
- Native STP isolation, Multipoint loop-free connectivity with VSS
- Enhanced VPLS traffic load-balancing capabilities with FAT
- VPLS configuration simplifications

A-VPLS Pseudowire — Single Virtual Ethernet Interconnect across Multiple Interfaces



A-VPLS - Configuration



```
#sh mpls 12 vc
```

Local intf	Local circuit	Dest address	VC ID	Status
VFI VFI_610_	VFI	10.100.2.2	610	UP
VFI VFI_610_	VFI	10.100.3.3	610	UP
VFI VFI_611_	VFI	10.100.2.2	611	UP
VFI VFI_611_	VFI	10.100.3.3	611	UP

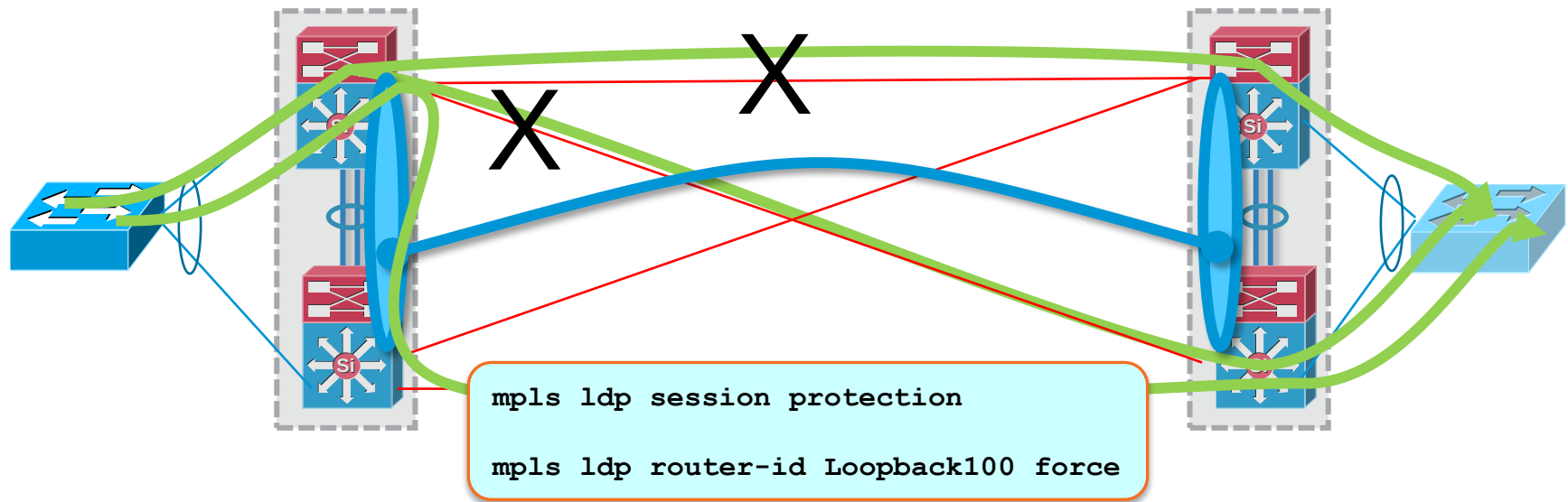
Rem: One PW per VLAN per destination

```
interface Virtual-Ethernet1
  switchport
  switchport mode trunk
  switchport trunk allowed vlan
610-611
  transport vpls mesh
  neighbor 10.100.2.2 pw-class Core
  neighbor 10.100.3.3 pw-class Core

pseudowire-class Core
  encapsulation mpls
```



A-VPLS - Redundancy/Dual-Homing Using VSS



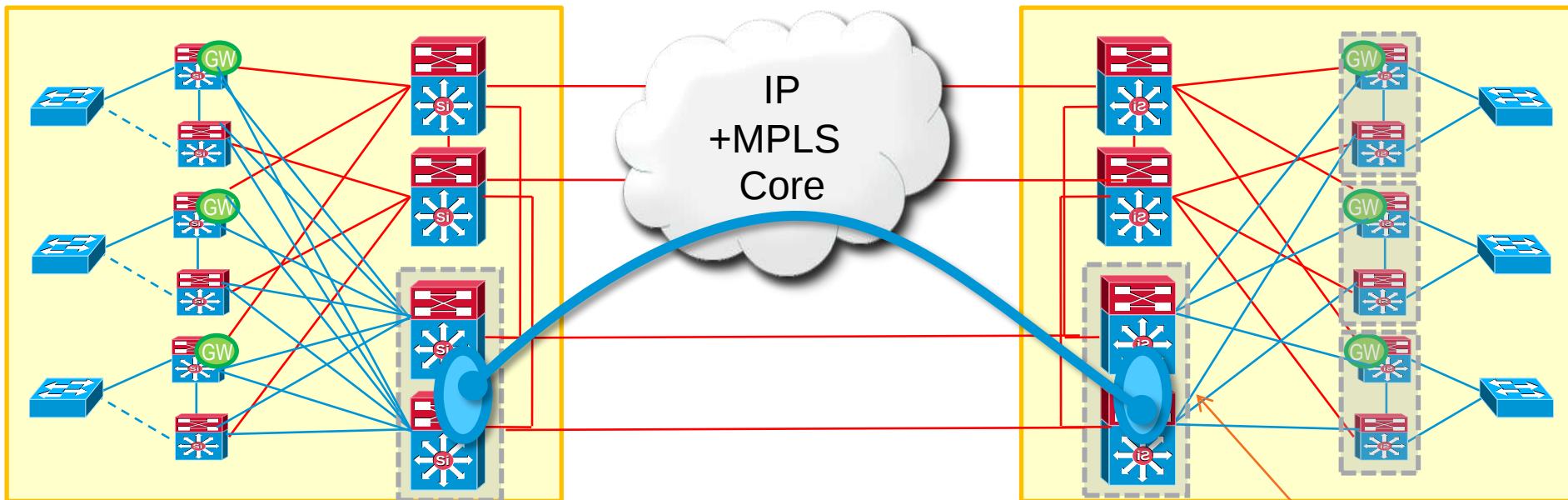
- LDP session protection & Loopback usage allows PW state to be unaffected

- PW state is unaffected
- LDP + IGP convergence in sub-second
Fast failure detection on Carrier-delay / BFD
- Traffic flows through the VSL link
Traffic exits directly from egress VSS node

- convergence in sub-second
- detection on Carrier-delay / BFD
- local fast protection
- directly from egress VSS node

A-VPLS - Deployment Consideration

Dedicated VSS for DCI

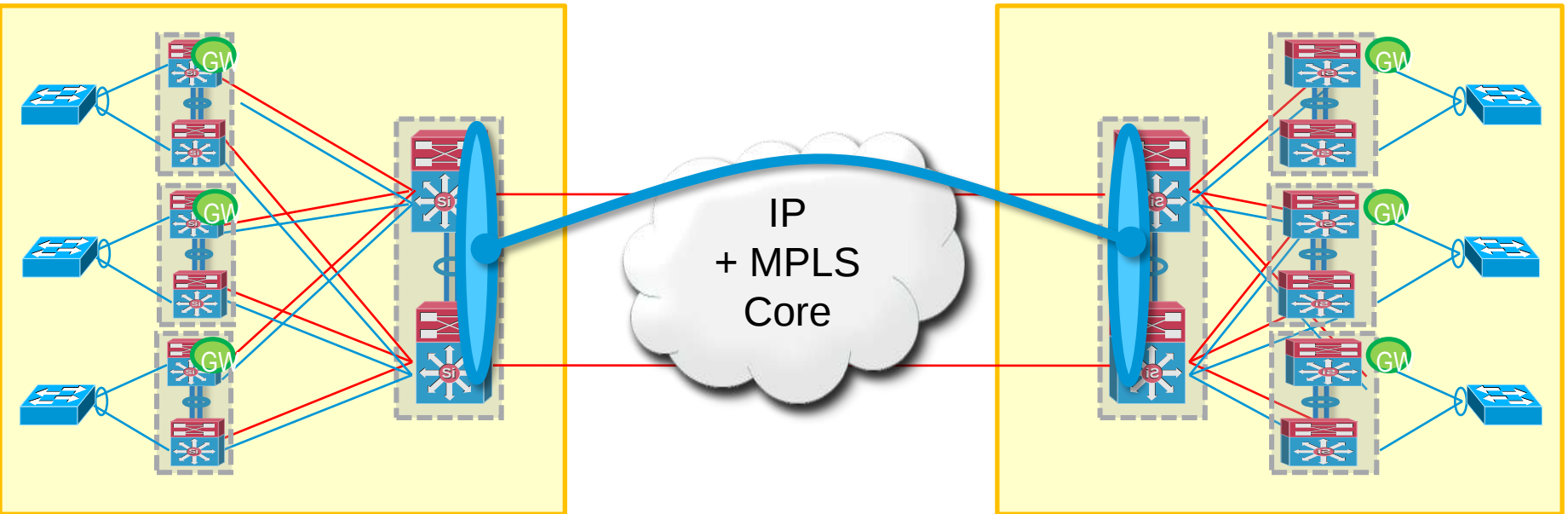


- ✓ Extend VLAN from aggreg to core using physical octopus
 - ✓ Full mesh octopus is required when STP connection
- ✓ Use A-VPLS to extend them in multi-point over MPLS
- 40Gbps with ES+
- ✓ SVI routing is still in aggregation

STP Isolation (default)
Storm control
FHRP Isolation

A-VPLS -Deployment Consideration

Fusion DCI Layer into DC Core with L3 Aggregation



- ✓ Extend VLAN from aggreg to core using dot1Q
- ✓ Use A-VPLS to extend them
- ✓ SVI routing is still in aggregation

MPLS for LAN Extensions Conclusion

Design Considerations

Benefits

- EoMPLS is an easy point to point solution
- A-VPLS based on node clustering (VSS)
 - Dual-homing support without additional protocols
 - Configuration simplicity
- Complementary with full MPLS featuring
 - VRF MP-BGP
 - RSVP for Traffic-Engineering & Fast-ReRoute

Constraints

- Rely on flooding for Unknown Unicast traffic
- Point to point (versus point to cloud)
- Full mesh of pseudo-wires/tunnels must be in place.
- Head-end replication for multicast and broadcast. Sub-optimal BW utilization

Agenda

- DCI Business Drivers and Solutions Overview
- SAN Extension Solutions
- LAN Extension Deployment Scenarios
 - Ethernet Based Solutions
 - MPLS Based Solutions
 - IP Based Solutions
 - Overlay Transport Virtualization
- Path optimization
- Conclusions and Q&A



Overlay Transport Virtualization

Technology Pillars



OTV is a “MAC in IP” technique to extend Layer 2 domains
OVER ANY TRANSPORT



Dynamic Encapsulation

No Pseudo-Wire State Maintenance

Optimal Multicast Replication

Multipoint Connectivity

Point-to-Cloud Model



Nexus 7000

*First platform to support OTV
(from release 5.0)*

Protocol Learning

Preserve Failure Boundary

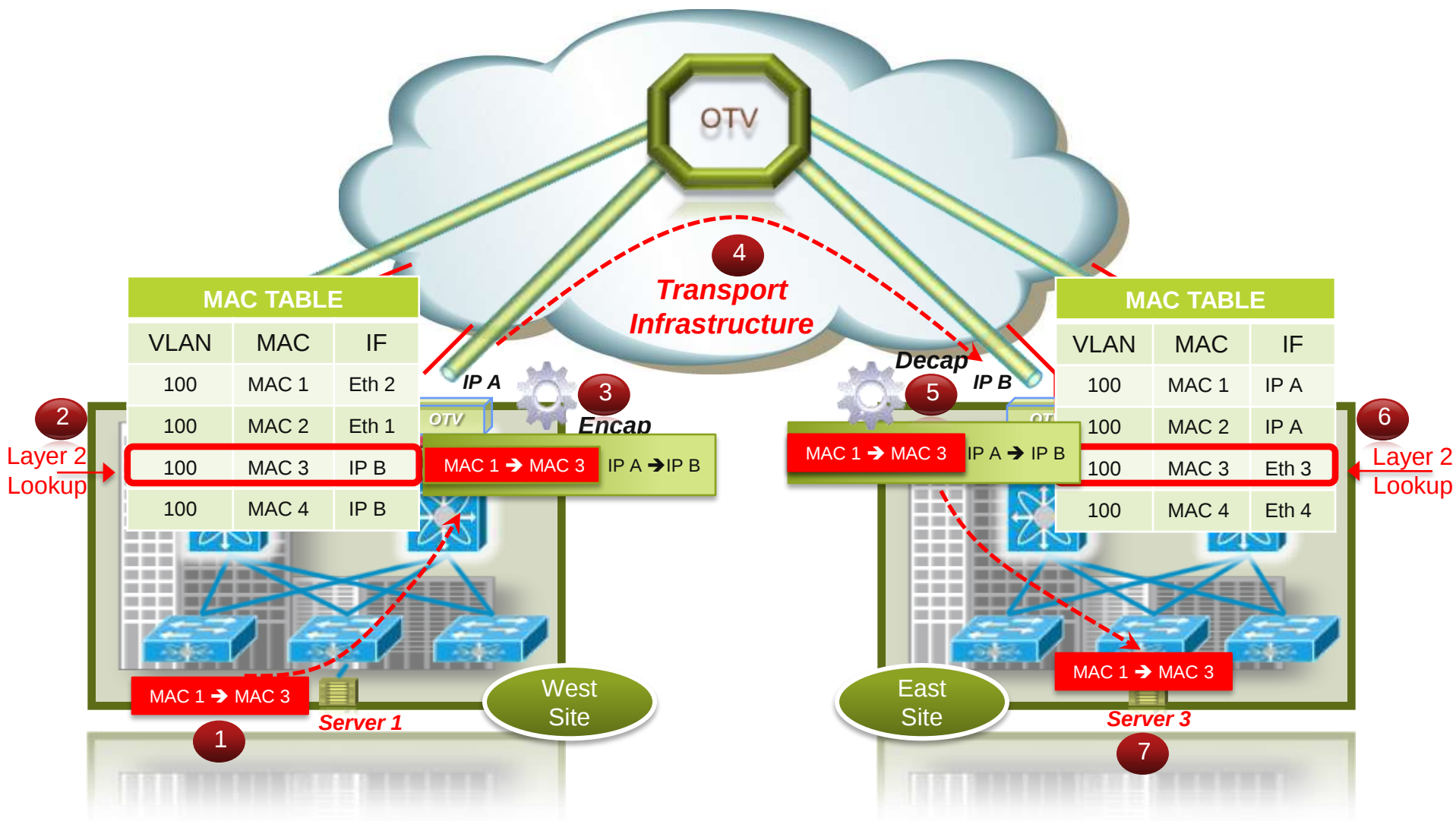
Built-in Loop Prevention

Automated Multi-homing

Site Independence

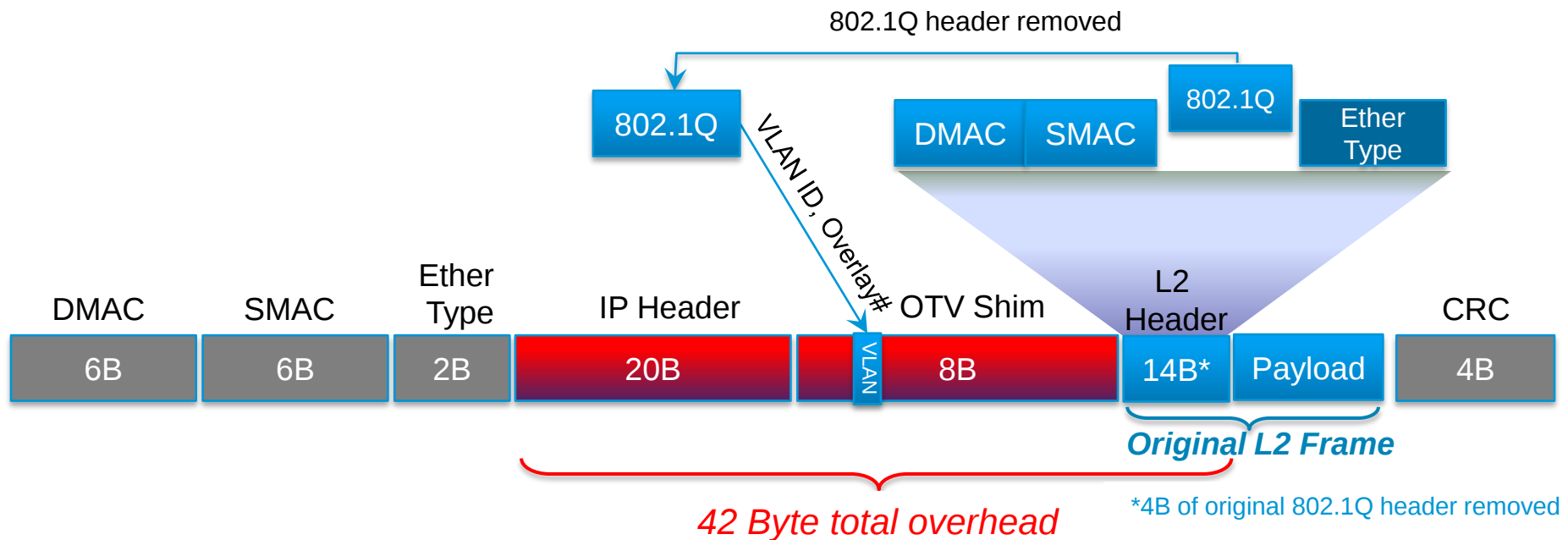
OTV Data Plane

Inter-Sites Packet Flow



OTV Data Plane Encapsulation

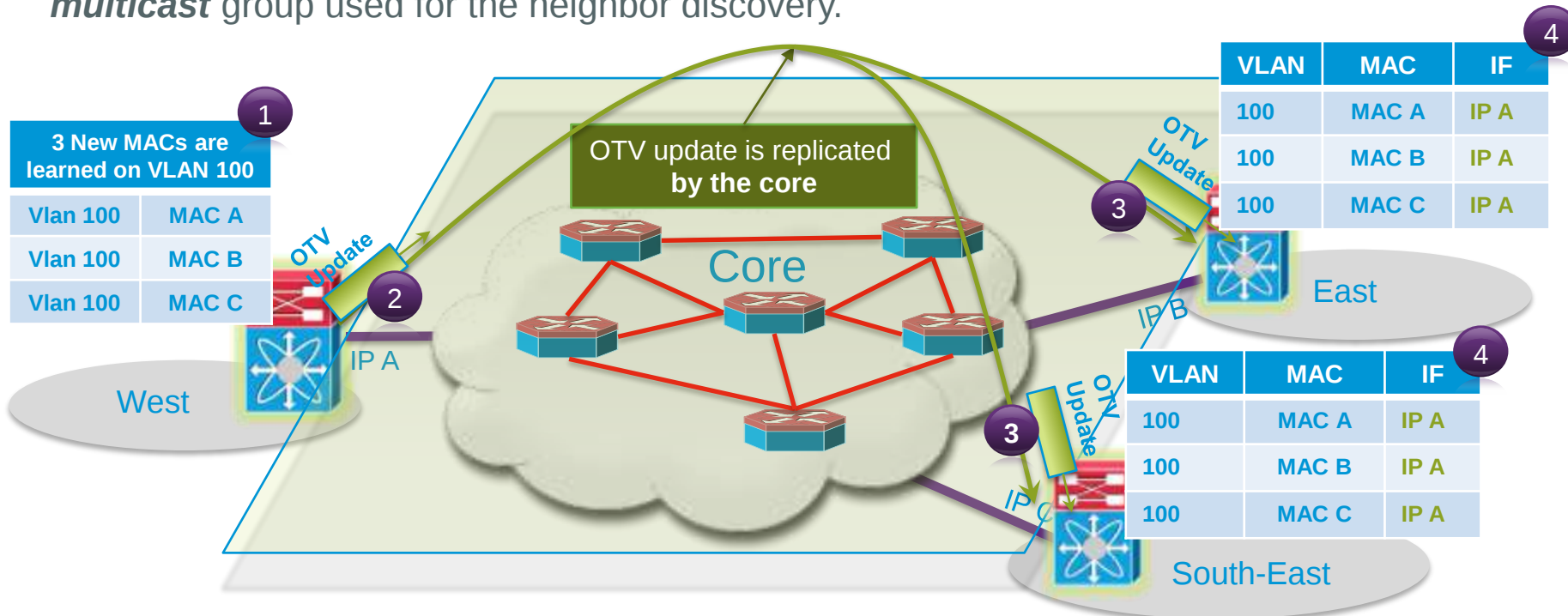
- OTV encapsulation adds 42 Bytes to the packet IP MTU size
 - Outer IP Header and OTV Shim Header in addition to original L2 Header
- The outer OTV shim header contains information about the overlay (VLAN, overlay number)
- The 802.1Q header is removed from the original frame and the VLAN field copied over into the OTV shim header



OTV Control Plane

MAC Address Advertisements (Multicast-Enabled Transport)

- Every time an Edge Device learns a new MAC address, the OTV control plane will advertise it together with its associated VLAN IDs and IP next hop.
- The IP next hops are the addresses of the Edge Devices through which these MACs addresses are reachable in the core.
- A single OTV update can contain multiple MAC addresses for different VLANs.
- A single update reaches all neighbors, as it is encapsulated in the same **ASM multicast** group used for the neighbor discovery.



OTV Configuration

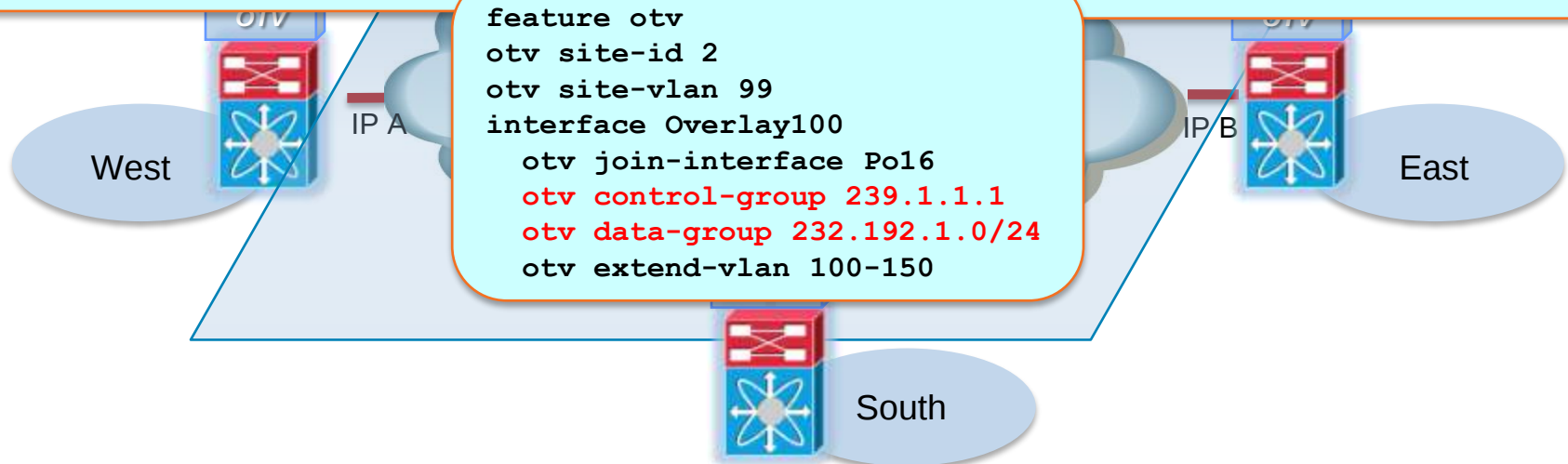
OTV over a Multicast Transport

- Minimal configuration required to get OTV up and running

```
feature otv
otv site-id 1
otv site-vlan 99
interface Overlay100
  otv join-interface e1/1
  otv control-group 239.1.1.1
  otv data-group 232.192.1.0/24
  otv extend-vlan 100-150
```

```
feature otv
otv site-id 3
otv site-vlan 99
interface Overlay100
  otv join-interface e1/1.10
  otv control-group 239.1.1.1
  otv data-group 232.192.1.0/24
  otv extend-vlan 100-150
```

```
feature otv
otv site-id 2
otv site-vlan 99
interface Overlay100
  otv join-interface Po16
  otv control-group 239.1.1.1
  otv data-group 232.192.1.0/24
  otv extend-vlan 100-150
```



OTV solves Layer 2 Fault Propagation

- **STP isolation – No configuration required**

No BPDUs forwarded across the overlay
STP remains local to each site

- **Unknown unicast isolation – No configuration required**

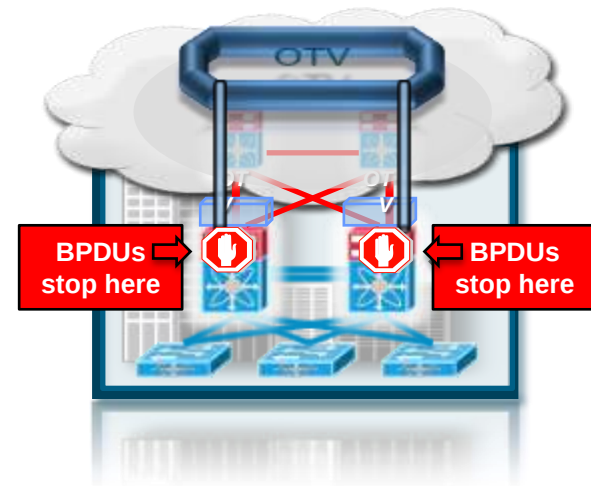
No unknown unicast frames flooded onto the overlay
Assumption is that end stations are not silent
Option for selective unknown unicast flooding (for certain applications)

- **Proxy ARP cache for remote-site hosts – On by default**

ARP cache maintained in Edge Device by snooping ARP replies

First ARP request is broadcasted to all sites. Subsequent ARP requests are replied by local Edge Device

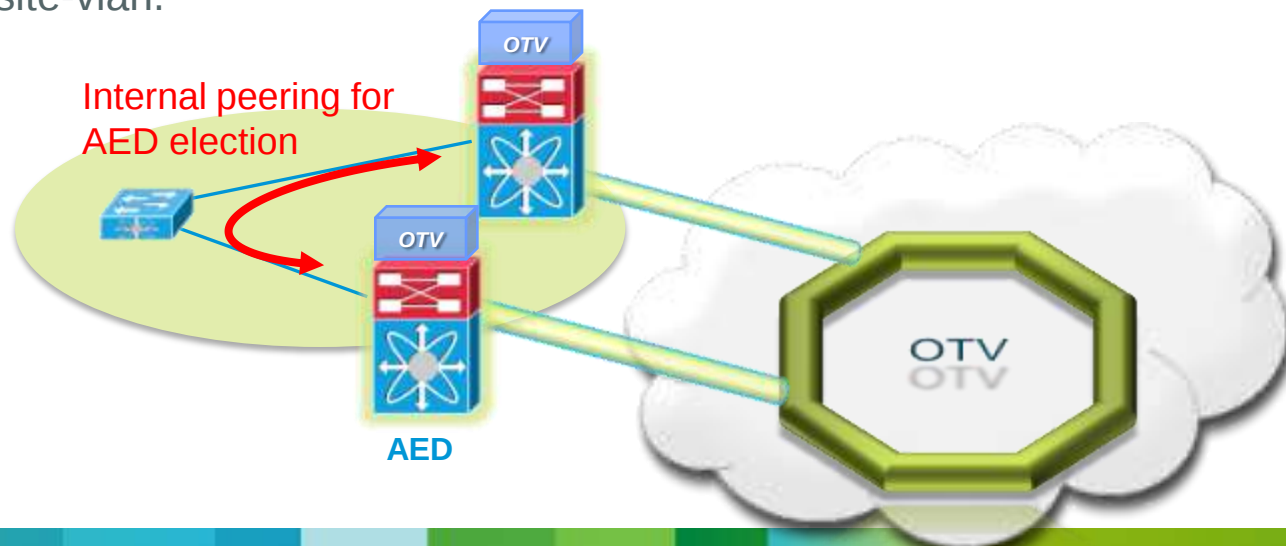
- Broadcast can be controlled based on a white list as well as a rate limiting profile



OTV Automated Multi-homing

Per VLAN Authoritative Edge Device

- The detection of the multi-homing is **fully automated** and **it does not require additional protocols and configuration**
- OTV provides loop-free multihoming by electing a designated forwarding device **per site for each VLAN**
- The designated forwarder is referred to as the **Authoritative Edge Device (AED)**.
 - forwards traffic to and from the overlay
 - advertises MAC addresses for any given site/VLAN
- The Edge Devices at the site peer with each other on the internal interfaces to elect the AED. The peering takes place over the OTV “**site-vlan**”. It’s recommended to use a dedicated VLAN as site-vlan.



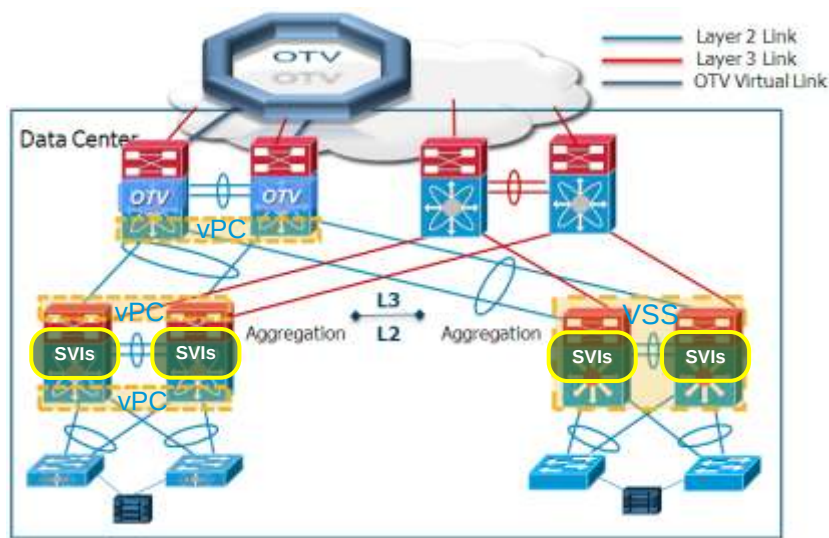
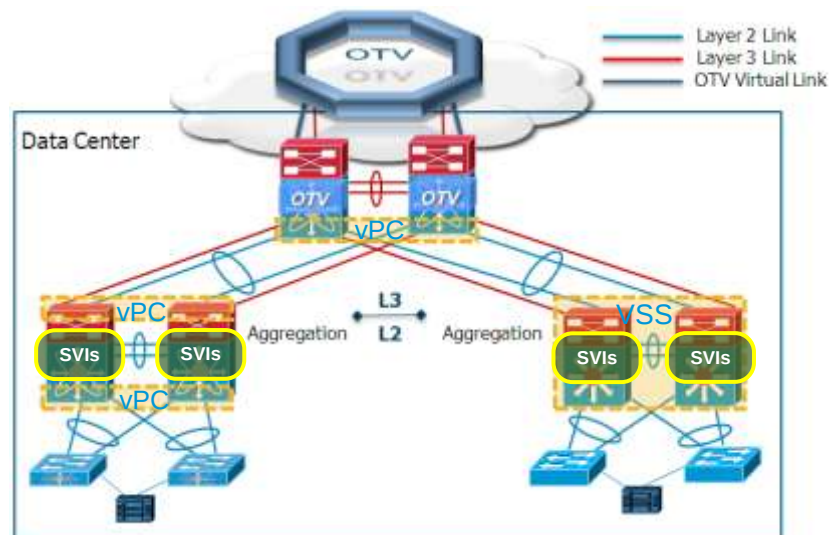
Placement of the OTV Edge Device

Option 1: OTV in the DC Core with L3 Boundary at Aggregation

- Easy deployment for Brownfield
- L2-L3 boundary remains at aggregation
- DC Core devices performs L3 and OTV functionalities
 - May use a pair of dedicated Nexus 7000
- VLANs extended from aggregation layer

L2 “Octopus” design

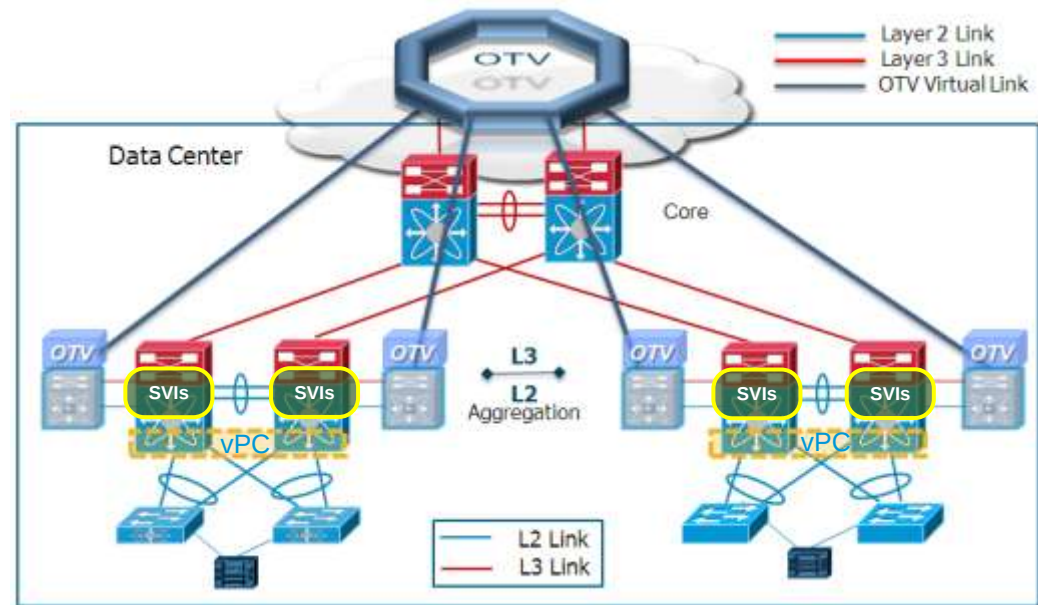
Recommended to use separate physical links for L2 & L3 traffic



Placement of the OTV Edge Device

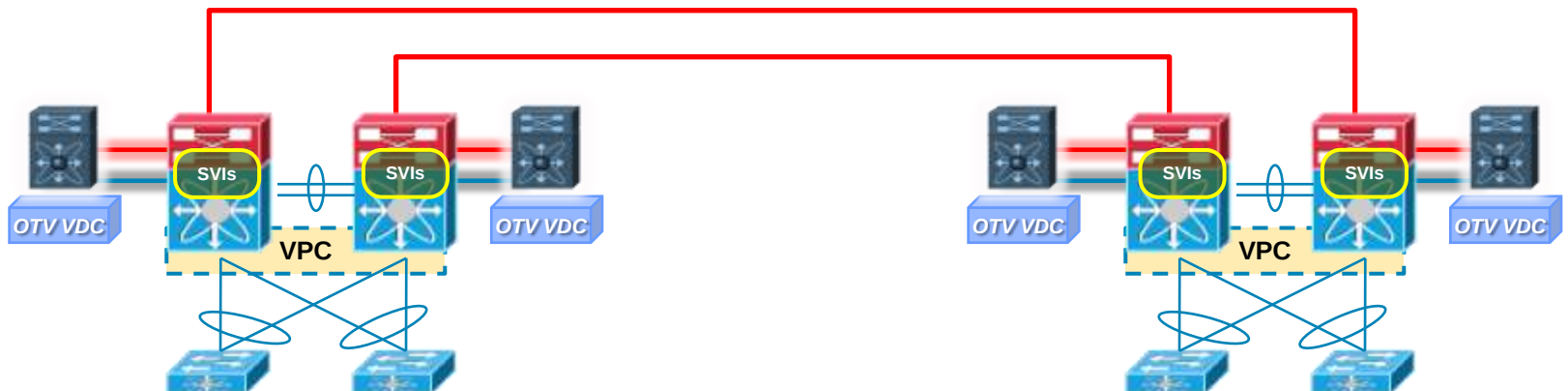
Option 2: OTV in the DC Aggregation

- L2-L3 boundary at aggregation
- DC Core performs only L3 role
- Intra-DC and Inter-DCs LAN extension provided by OTV
 - Requires the deployment of dedicated OTV VDCs
- Ideal for single aggregation block topologies
- Recommended for Green Field deployments
 - Nexus 7000 required in aggregation



Option 3: OTV over Dark Fiber Deployments

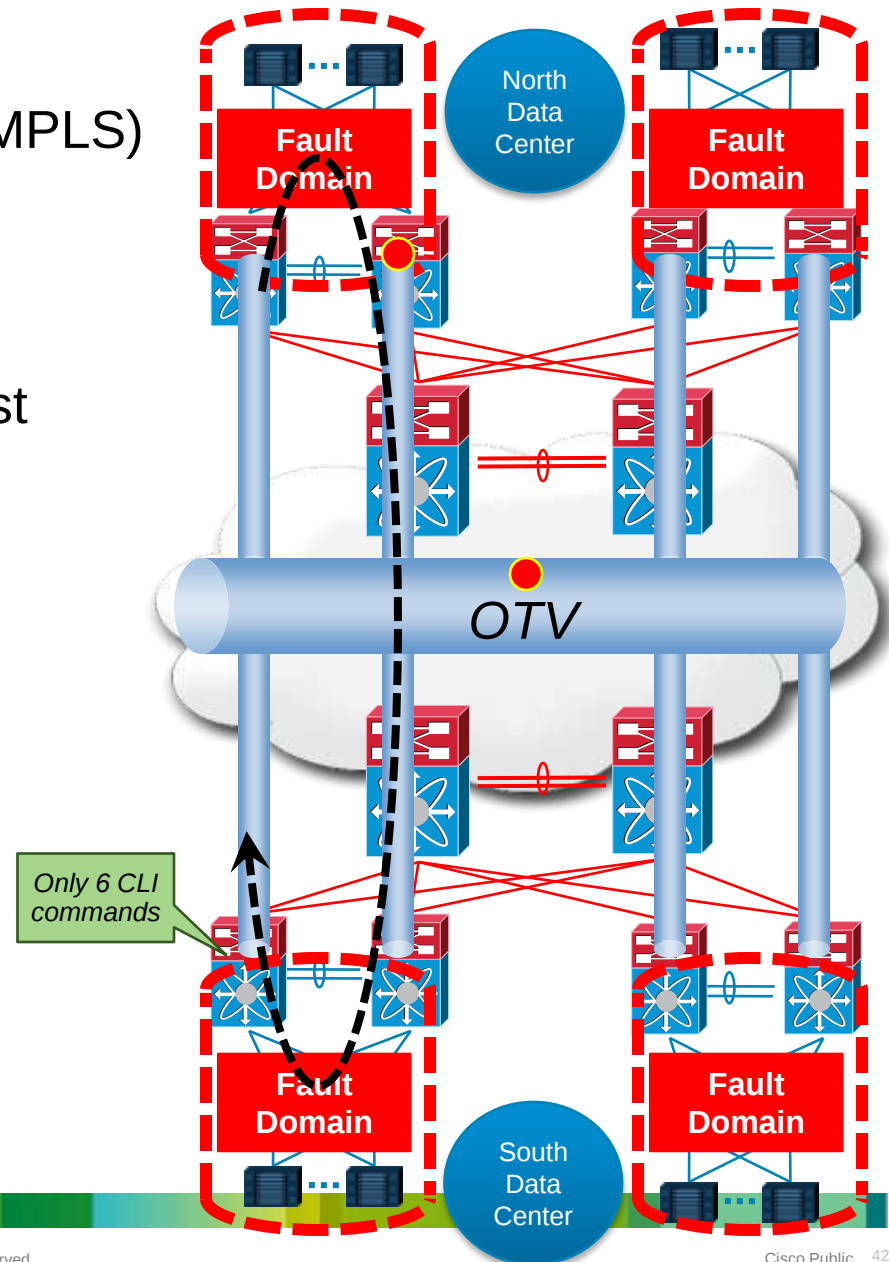
- Data Centers directly connected at the Aggregation
- Currently mandates the deployment of dedicated OTV VDCs
 - OTV Control Plane messages must always be received on the Join Interface
 - Requires IGP/PIM peering between aggregation devices (via peer-link)
- Advantages over VSS-vPC solution:
 - ✓ Provision of Layer 2 and Layer 3 connectivity leveraging the same dark fiber connections
 - ✓ Native STP isolation: no need to explicitly configure BPDU filtering
 - ✓ ARP Optimization with the OTV ARP Cache
 - ✓ Simplified provisioning of FHRP isolation
 - ✓ Easy Addition of Sites



OTV for LAN Extensions Conclusion

Design Considerations

- Extensions over any transport (IP, MPLS)
- Failure boundary preservation
- Site independence
- Optimal BW utilization with multicast enabled transport infrastructure (no head-end replication)
- Automated Built-in Multihoming
- End-to-End loop prevention
- Scalability
 - Sites, VLANs, MACs
- Operations simplicity



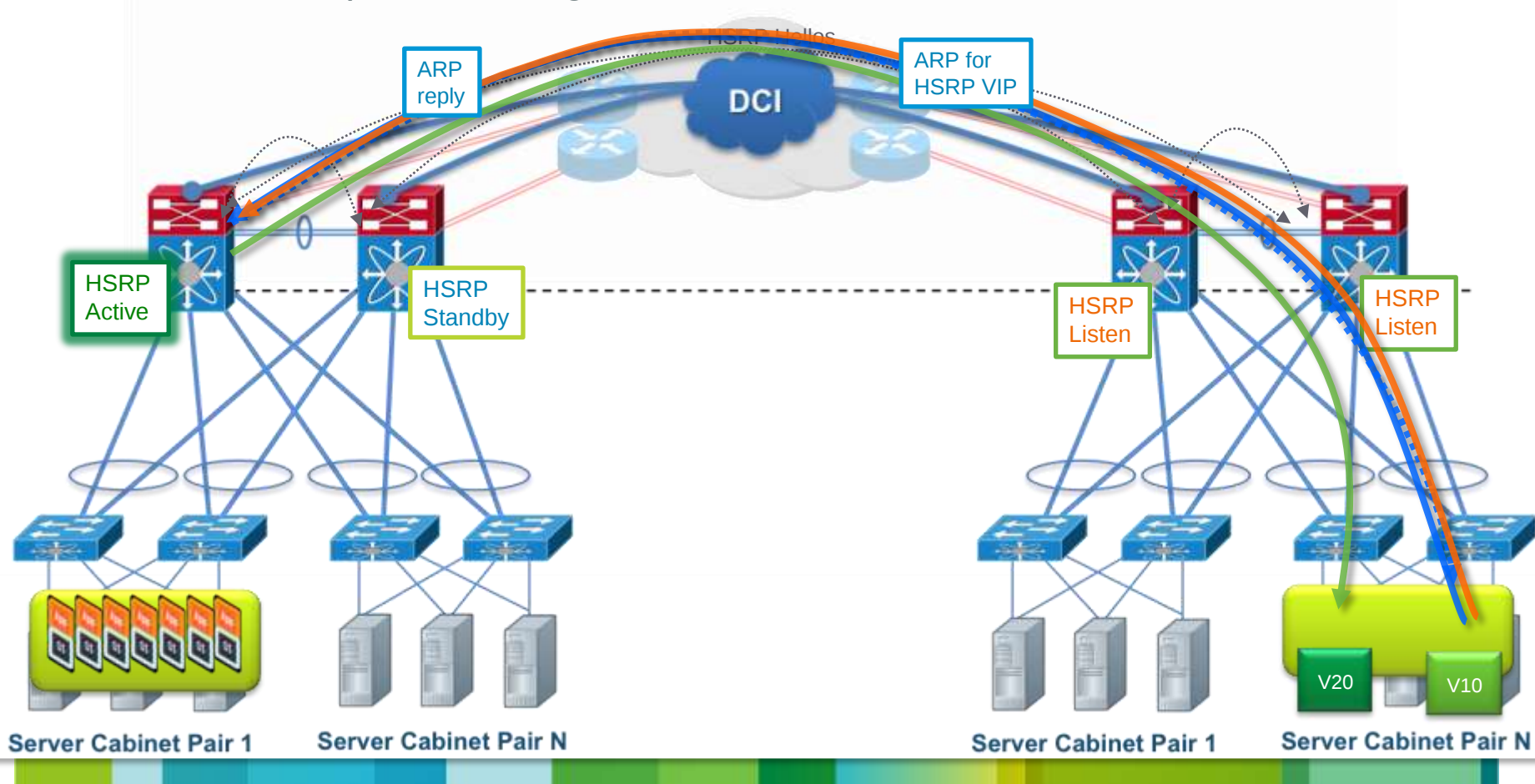
Agenda

- DCI Business Drivers and Solutions Overview
- SAN Extension Solutions
- LAN Extension Deployment Scenarios
 - Ethernet Based Solutions
 - MPLS Based Solutions
 - IP Based Solutions
- Path optimization
- Conclusions and Q&A



Egress DC Routing with LAN Extension

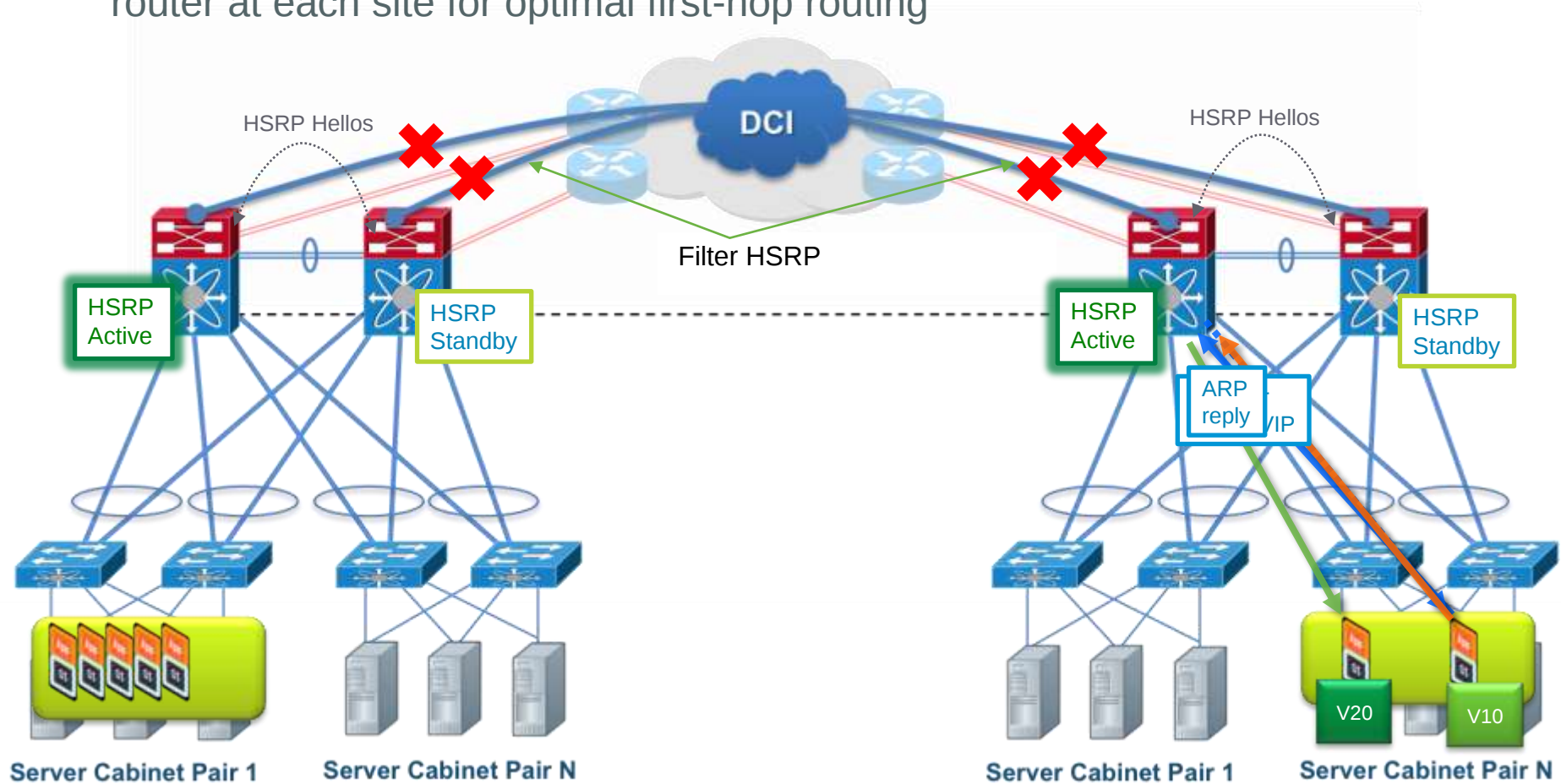
- Extended VLAN typically has associated HSRP group
- Only one HSRP router active, with all servers pointing to HSRP VIP as default gateway
- **Result:** sub-optimal routing



Egress DC Routing Localization

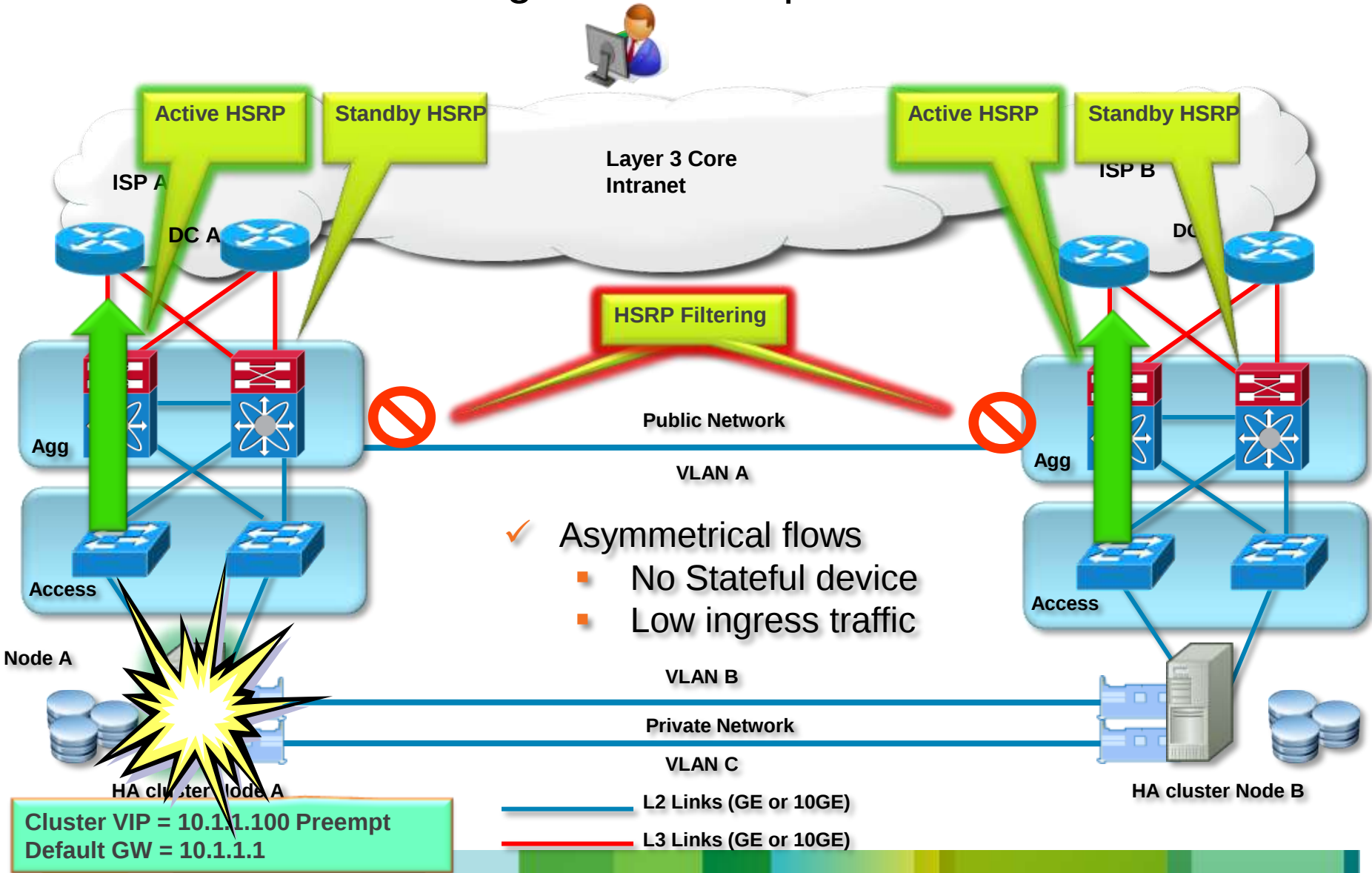
FHRP Filtering Solution

- Filter FHRP with combination of VACL or PACL
- **Result: Still have one HSRP group with one VIP, but now have active router at each site for optimal first-hop routing**

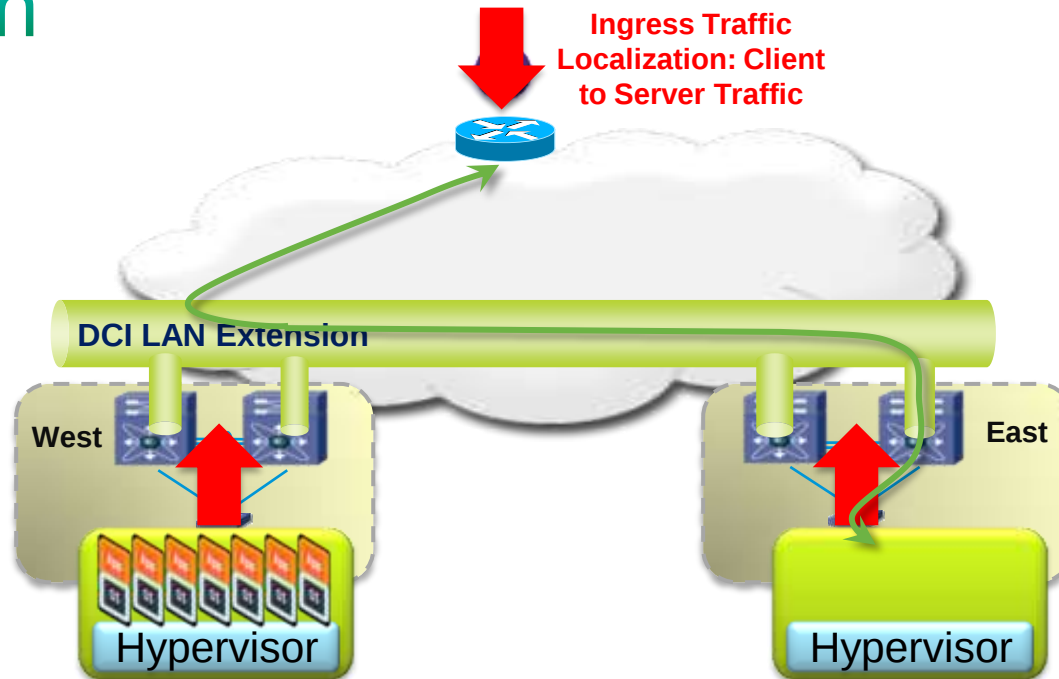


Sample Cluster - Primary Service in Left DC

FHRP Localization – Egress Path Optimization



Ingress DC Routing Localization with LAN Extension



Challenge

- Subnets are spread across locations
- Subnet information in the routing tables is not specific enough
- Routing doesn't know if a server has moved between locations
- Traffic may be sent to the location where the application is not available



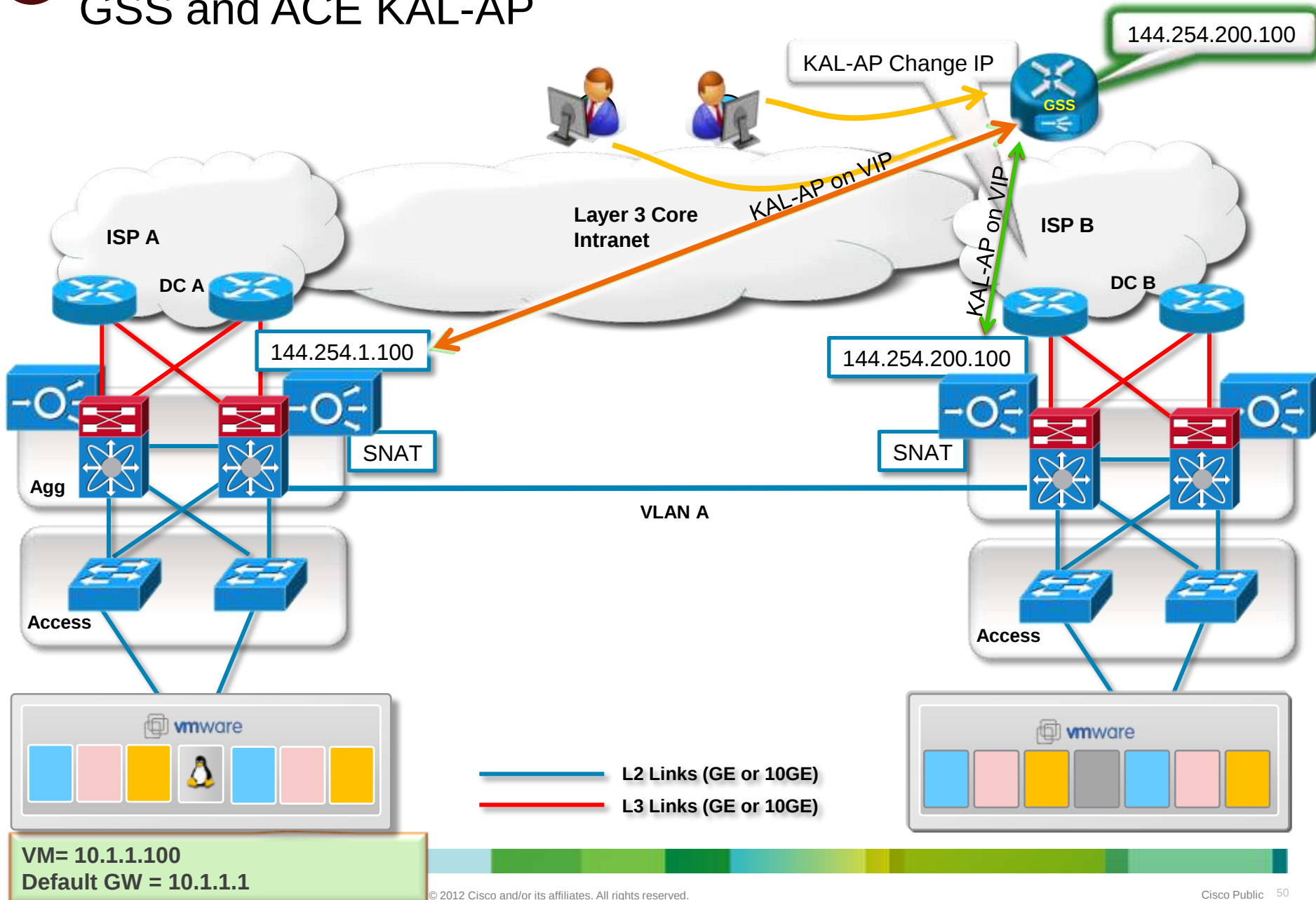
Options

- DNS Based
 1. DNS redirection with ACE/GSS
- Routing Based
 2. Route Injection
 3. LISP

1

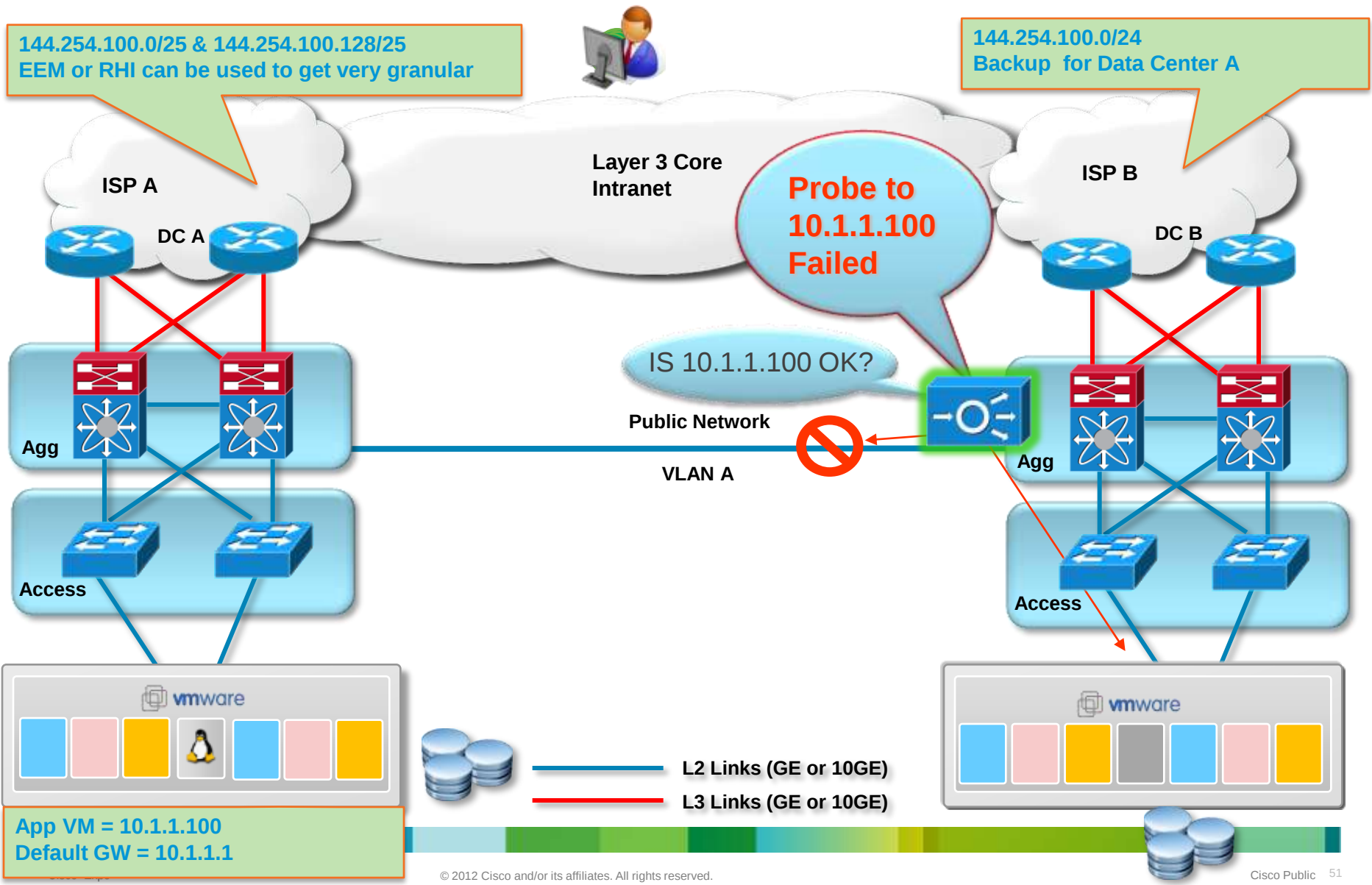
VMotion - Primary Service in Left DC

GSS and ACE KAL-AP



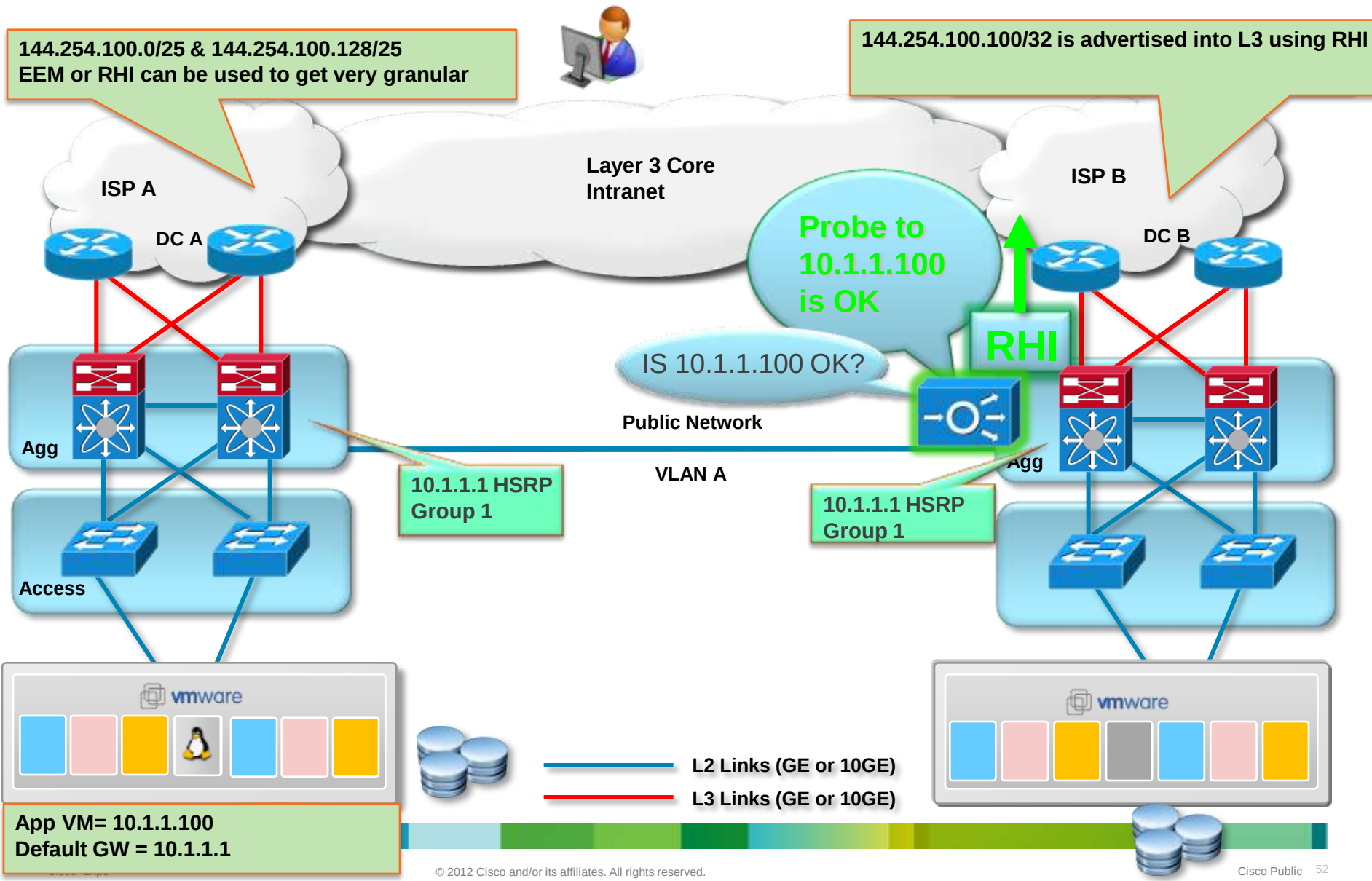
2 VMotion - Primary Service in Left DC

Detection of Movement of VM using ACE Probes – Ingress Path Optimization



2 VMotion - Primary Service in Left DC

Detection of Movement of VM using ACE Probes – Ingress Path Optimization



LISP for Ingress Routing Localization

Needs:

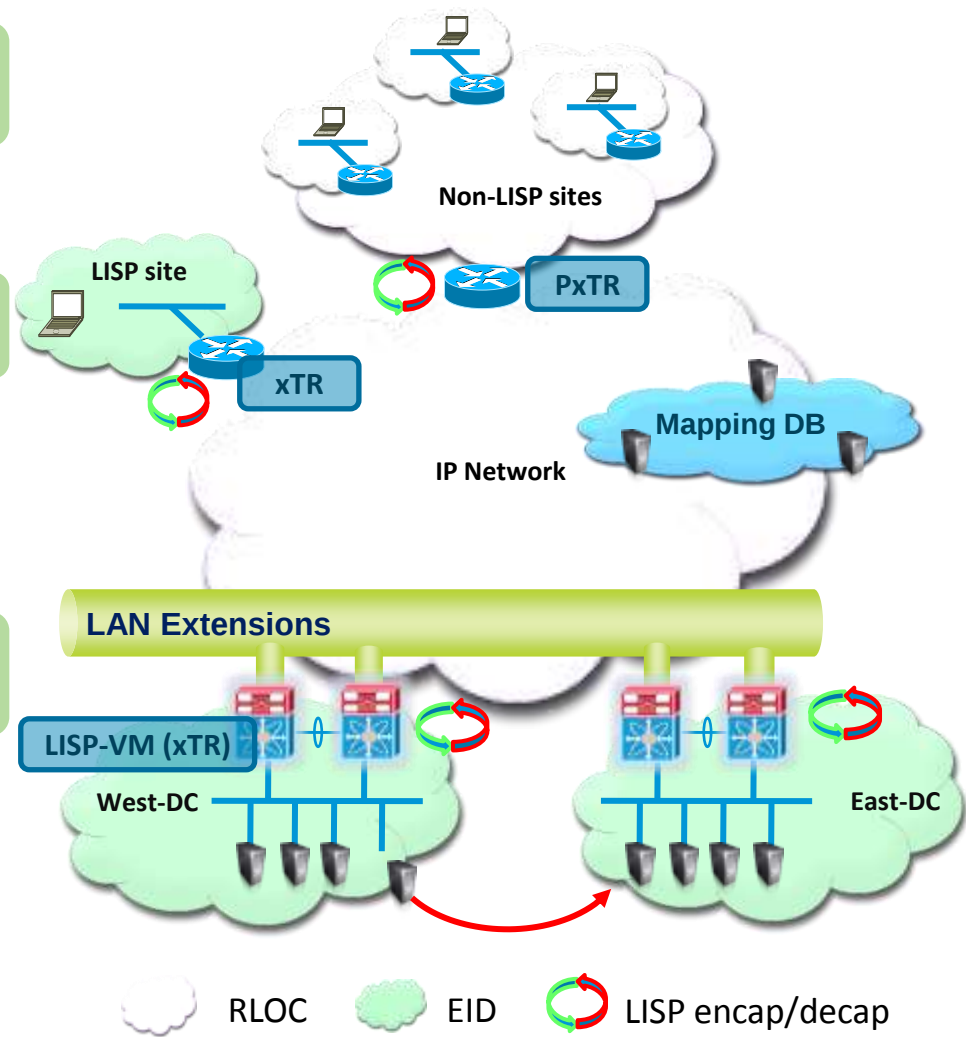
- Optimized routing **across extended subnet sites**

LISP Solution:

- Automated move detection on xTRs
- Dynamically update EID-to-RLOC mappings
- Traffic Redirection on iTRs or PiTRs

Benefits:

- Direct Path (no triangulation)
- Connections maintained across move
- No routing re-convergence
- No DNS updates required
- Transparent to the hosts



Agenda

- DCI Business Drivers and Solutions Overview
- SAN Extension Solutions
- LAN Extension Deployment Scenarios
 - Ethernet Based Solutions
 - MPLS Based Solutions
 - IP Based Solutions
- Path optimization
- Conclusions and Q&A



Data Center Interconnect

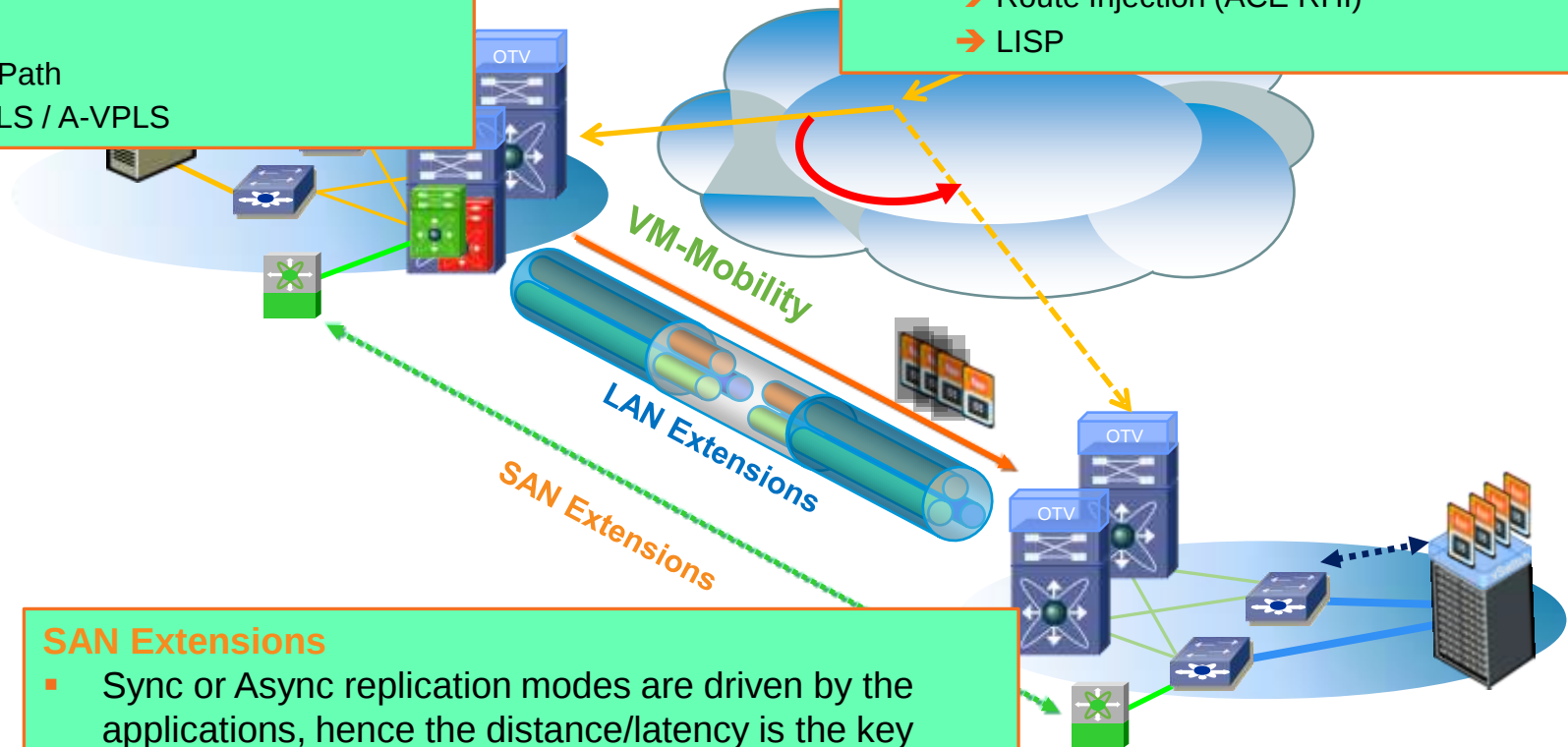
Summary

LAN Extensions

- STP Isolation, End-to-End loop prevention and Multi-homing are the key elements
- DCI Solutions
 - ➔ VSS / vPC
 - ➔ OTV
 - ➔ FabricPath
 - ➔ EoMPLS / A-VPLS

Path Optimization

- Optimal DC Egress/Ingress
- Egress routing optimization
 - ➔ Addressed by FHRP Filtering
- Ingress routing optimization:
 - ➔ DNS redirection with ACE/GSS
 - ➔ Route Injection (ACE RHI)
 - ➔ LISP



SAN Extensions

- Sync or Async replication modes are driven by the applications, hence the distance/latency is the key component to select the choice
 - ➔ Distance can be improved using IO accelerator or caching

Related Cisco Expo 2012 Events

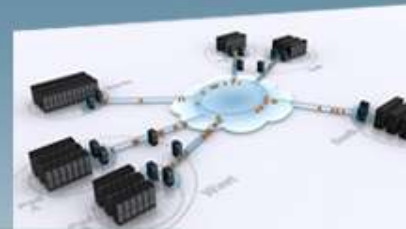
To learn more ...

Kód přednášky	Název přednášky
T-DC1	Technologie Cisco FabricPath a zkušenosti z nasazení
T-DC2	Mobilita ve virtualizovaném datovém centru s OTV a LISP
ARCH5	LISP v příkladech

Odkazy

Extend Layer 2 Applications

Overlay Transport Protocol (OTV) technology supports transparent workload mobility within and across dispersed data centers. (3:54 min)



Watch Video

Overview

In-Depth

Products

Resources

Get Instant Workload Mobility Among Data Centers

Cisco Data Center Interconnect (DCI) solutions can help your IT organization meet business continuity and corporate compliance objectives.

- Reduce the business impact of disaster events and help ensure business continuity
- Improve productivity through enhanced application and data availability
- Meet corporate and regulatory compliance needs and improve data security

These solutions transparently extend LAN and SAN connectivity and provide accelerated, highly secure data replication, server clustering, and workload mobility between geographically dispersed data centers. This enhances business resilience, and helps enable application and data mobility between data centers, while maintaining operational consistency.

Featured Products



Cisco Nexus 7000 Series LAN Extension

Simplify Layer 2 applications across distributed data centers.



Cisco Catalyst 6500 Series LAN Extension

Deliver high-performance, scalable Layer 2 extension with subsecond convergence.



Cisco MDS 9500 Series SAN Extension Over IP

Gain an integrated, cost-effective, reliable business continuance solution.

<http://www.cisco.com/go/dci>

Otázky a odpovědi

- Twitter www.twitter.com/CiscoCZ
 - Talk2Cisco www.talk2cisco.cz/dotazy
 - SMS 721 994 600
-
- Zveme Vás na **Ptali jste se...** v sále **LEO**
1.den 17:45 – 18:30
2.den 16:30 – 17:00

**Prosíme, ohodnot'te
tuto přednášku.**