
TOMORROW starts here.
cisco
Cisco live!

How to Achieve True Active-Active Data Centre Infrastructures

BRKDCT-2615

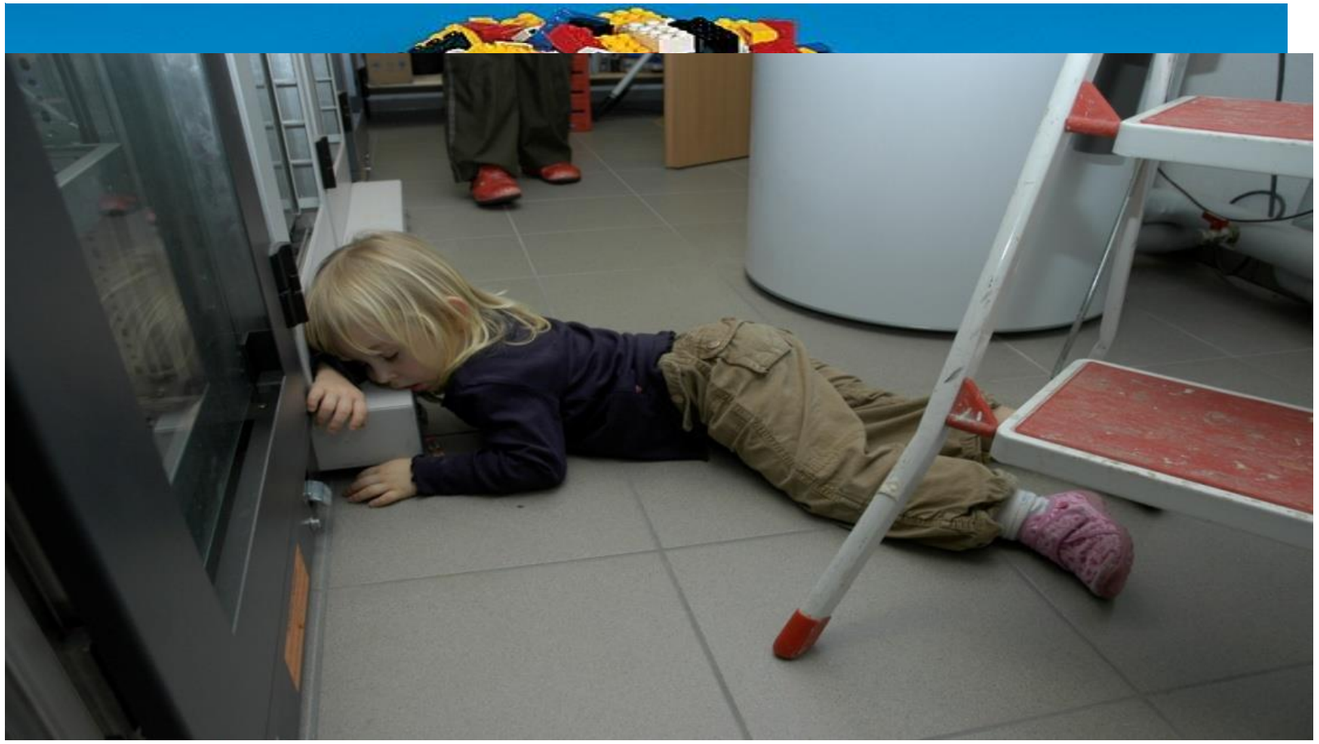
Carlos Pereira
Distinguished Systems Engineer II

Cisco live!

Active / Active Data Centres
Typical Process

Then try to figure that out

… and feel tired (or panic ☺)



Objectives

- Understand the Active/Active Data Centre requirements and considerations
- Provide considerations for Active/Active DC Design, from storage, DCI (including LISP) and network services perspectives
- Share Experiences with State-full Devices placements and their impact within DCI environment
- Briefly discuss about the evolution Active / Active DC with ACI policies, DCI and federation considerations

Reference slides would be
Quickly (if) covered during the session.

Legend

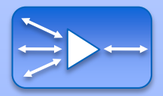

Load Balancer


SSL Offloader

Application Policy Infrastructure Controller

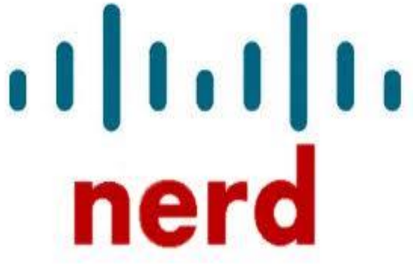

SVI / HSRP Default Gw


IDS / IPS


WAN Accelerator

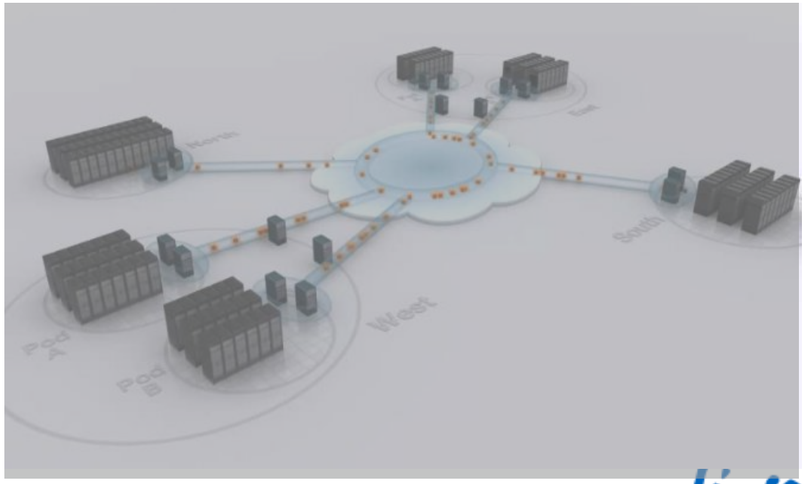

Firewall

Potential Collateral Effect

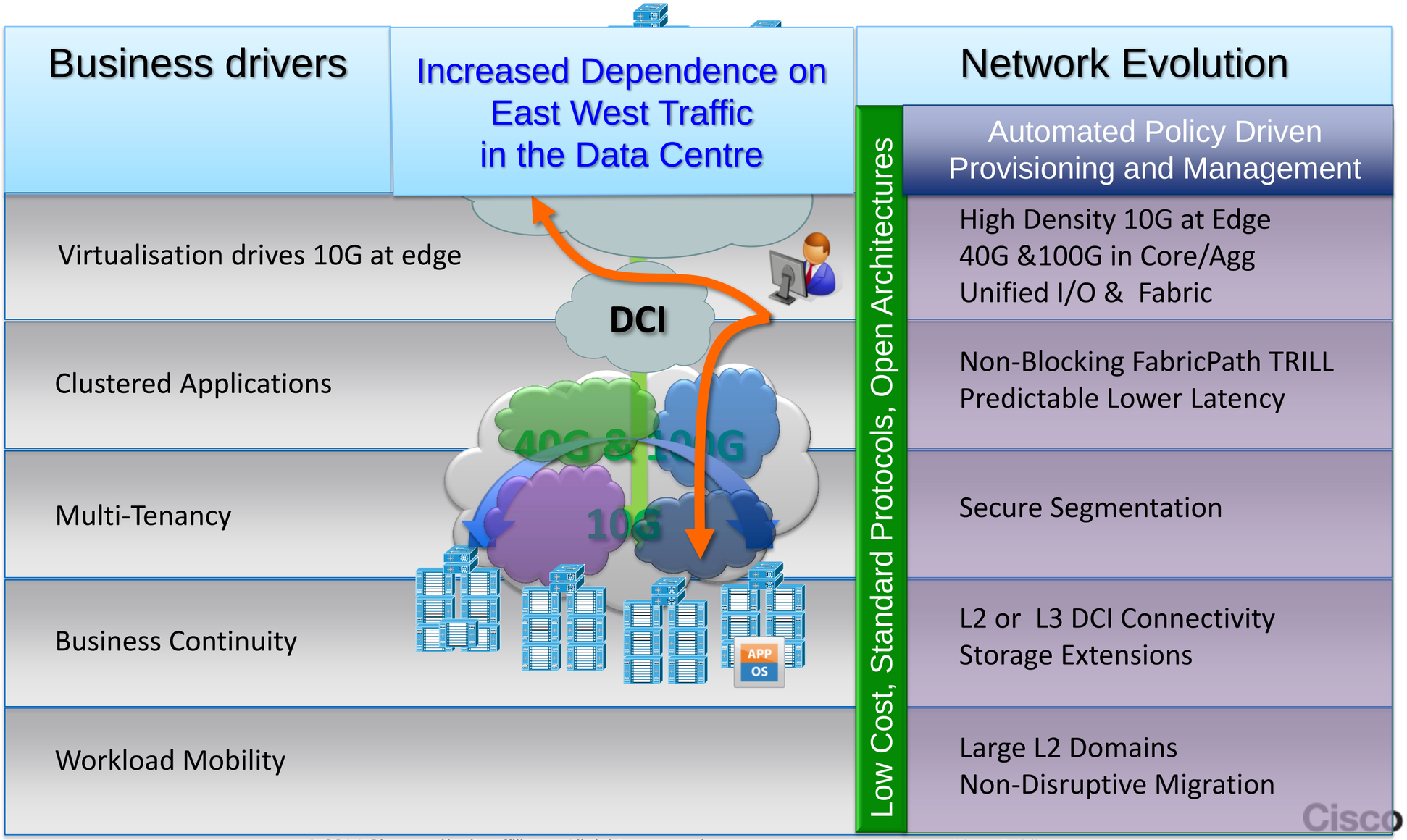


Agenda

- Active-Active Data Centre: Business Drivers and Solutions Overview
- Active / Active Data Centre Design Considerations
  - Storage Extension
  - Data Centre Interconnect (DCI) - LAN Extension Deployment Scenarios
  - Host Mobility using LISP and OTV
  - Network Services and Applications (Path optimisation)
- Cisco ACI and Active / Active Data Centre
- Summary and Conclusions
- Q&A



Data Centre Evolution
Cloud Network Fabric

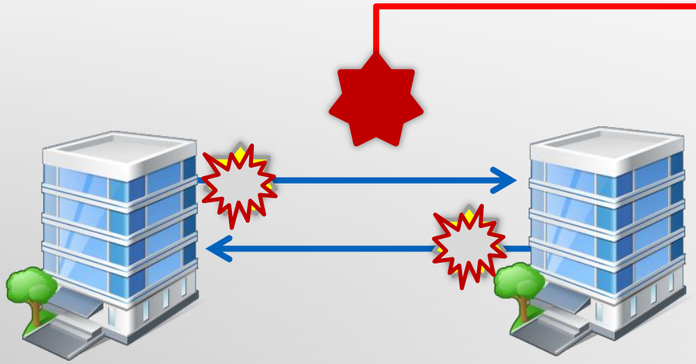


Business Continuity and Disaster Recovery

Ability to Absorb the Impact of a Disaster and Continue to Provide an Acceptable Level of Service.

"Operational Continuity"

Planned / Unplanned Service Continuity within the Metro
(including a single data centre loss)



DC-1 DC-2

Active-Active

Metro Area

"Disaster Recovery (DR)"

Loss of "Regional data centres" leads to recovery in a Remote data centre

DC-3

Pervasive Data Protection
+ **Infrastructure Rebalance**

"Applications and services extended across Metro and Geo distances are a natural "next step" along the virtualisation curve…."

Application Resiliency and Business Criticality Levels

Defining how a Service outage impacts Business will dictate a redundancy strategy

Lowest RTO/RPO

Highest RTO/RPO

Criticality Levels	Term	Impact Description	
C1	Mission Imperative	Any outage results in immediate cessation of a primary function, equivalent to immediate and critical impact to revenue generation, brand name and/or customer satisfaction; **no downtime is acceptable under any circumstances**	~ 10% of Apps
C2	Mission Critical	Any outage results in immediate cessation of a primary function, equivalent to major impact to revenue generation, brand name and/or customer satisfaction;	
C3	Business Critical	Any outage results in cessation over time or an immediate reduction of a primary function, equivalent to minor impact to revenue generation, brand name and/or customer satisfaction	~ 10% of Apps
C4	Business Operational	A sustained outage results in cessation or reduction of a primary function	~ 60% of Apps
C5	Business Administrative	A sustained outage has **little to no impact** on a primary function	~ 20% of Apps

Expected Distribution: (C1 & C2) <10%, C3 ~10%, C4 >60%, C5 <20% of applications

Criticality Classification Matrix

Criticality Classification Matrix v3.0								
Operational Continuity (Planned and Unplanned Downtime)					Disaster Recovery			
Adjusted Availability Ceiling	Planned Downtime Acceptable?	Acceptable Recovery Time (ART, hours)	Acceptable Data Loss (ADL, Hours)	Reduced Performance Acceptable (Single DC Loss)?	Recovery Time Objective (RTO, in Hours)	Recovery Point Objective (RPO, in Hours)	Reduced Performance Acceptable (Large-Scale Disaster)?	Criticality Level
Up to 99.999%	N	~0	~0	N	n/a**	n/a	n/a	C1
Up to 99.995%	N	1	0	N	4	1	N	C2
Up to 99.99%	Y	4	0	N	24	1	Y	C3
Up to 99.9%	Y	24	1	Y	48	24	Y	C4
Up to 99.9%	Y	Best Effort	24	Y	Best Effort	1 wk	Y	C5

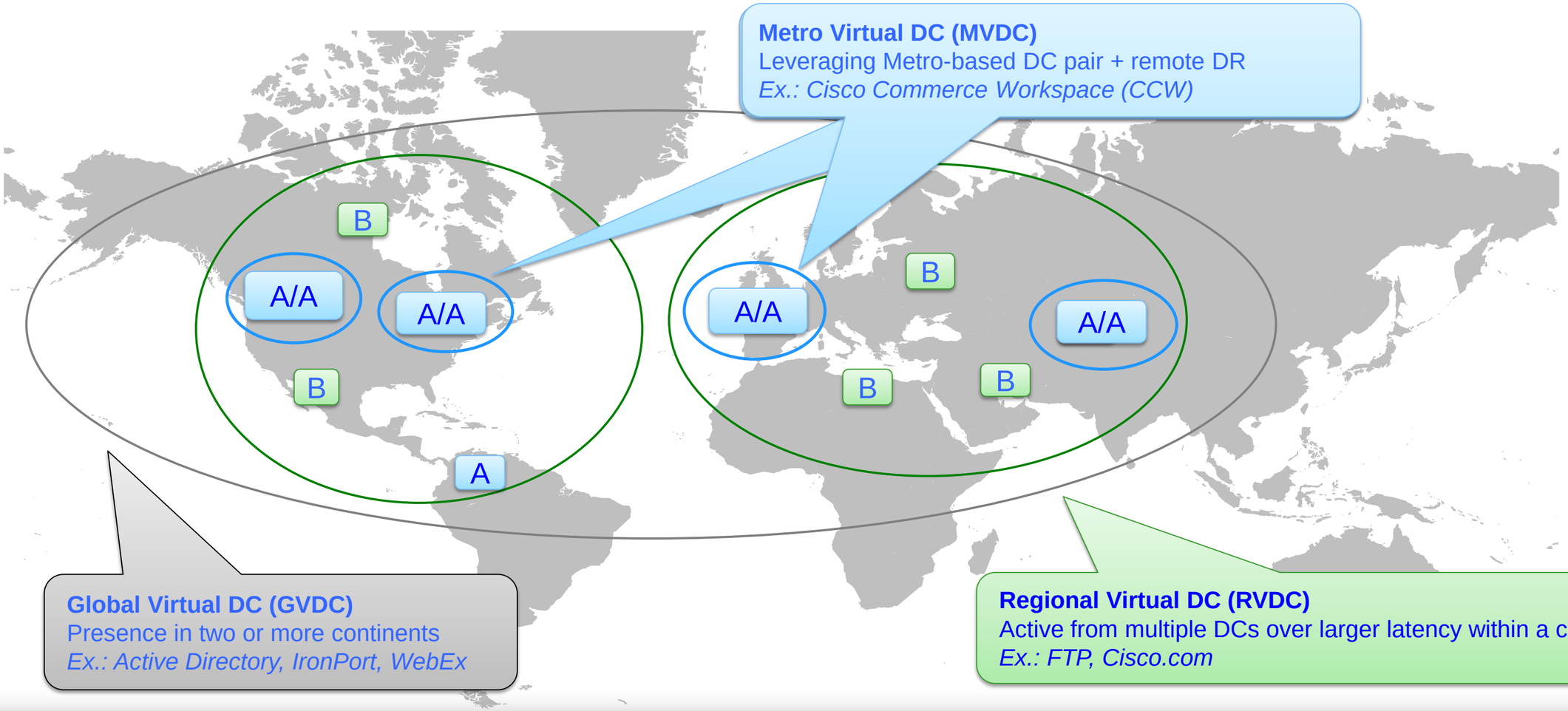
- ART = Maximum downtime following incidents (up to and including one DC in Metro down)
- ADL = Maximum data loss following incidents (up to and including one DC in Metro down)
- RTO = Maximum downtime for applications following large-scale disaster (multiple Tier-III DCs in Metro down, highly unlikely)
- RPO = Maximum data loss following large-scale disaster (multiple Tier-III DCs in Metro down, highly unlikely)

** Targeting distributed architectures (active/active over large distance) to meet service continuity requirements without DR invocation

Expected Distribution: (C1 & C2) <5%, C3 ~10%, C4 >60%, C5 <25% of applications

Global Data Centre Presence—Target State

Shared Resilient Infrastructure Enables Diversified Business Growth

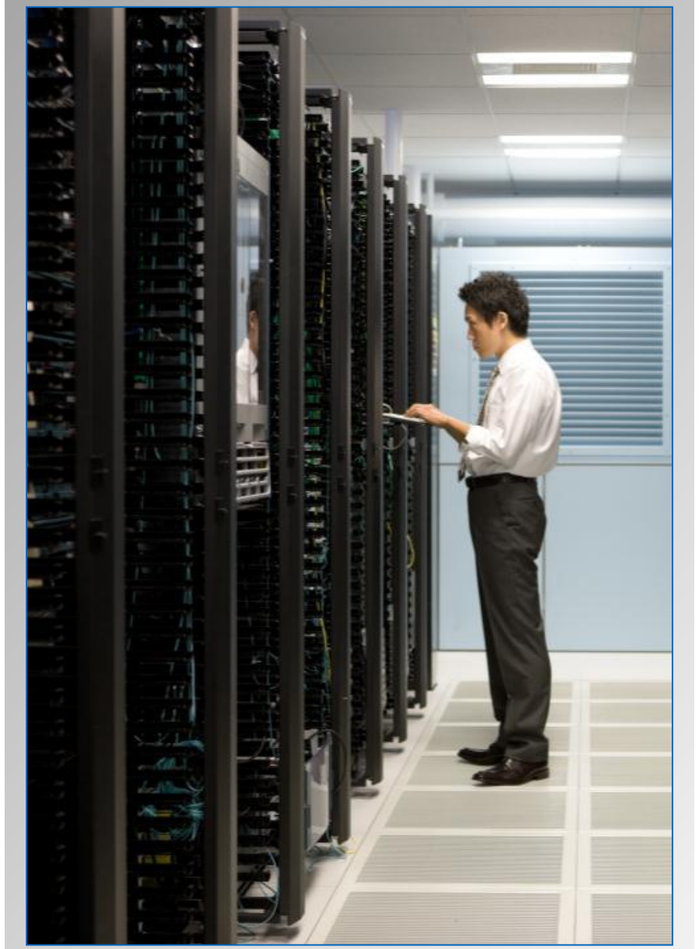


Distributed Virtual Data Centre (DVDC) Architecture

Three Variations Reflecting Varying Latency Constraints and Performance Requirements

Metro Virtual Data Centre (MVDC) Vision

- Generic, high-availability application and data solution architecture which leverages a dual data centre physical infrastructure
- Addresses all levels of the data centre stack
  - ✓ physical layer, network layer, server platform resources, storage resources, application networking services, data tier structure
  - ✓ No physical single points of failure
  - ✓ Optimised use of capacity through virtualisation
- Management and interaction of applications in a paired data centre environment
- Disaster Avoidance and Prevention by Pro-actively migrates seamlessly Virtual Machines with NO interruption
- Support Disaster Recovery capability beyond dual data centre resiliency
- Active-active capability and workload rotation to accelerate incident response time and increase confidence
- Capable of "no data loss" within Metro (synchronous replication, RPO=0)



MVDC Enabling Technologies
Cisco & Strategic Partners

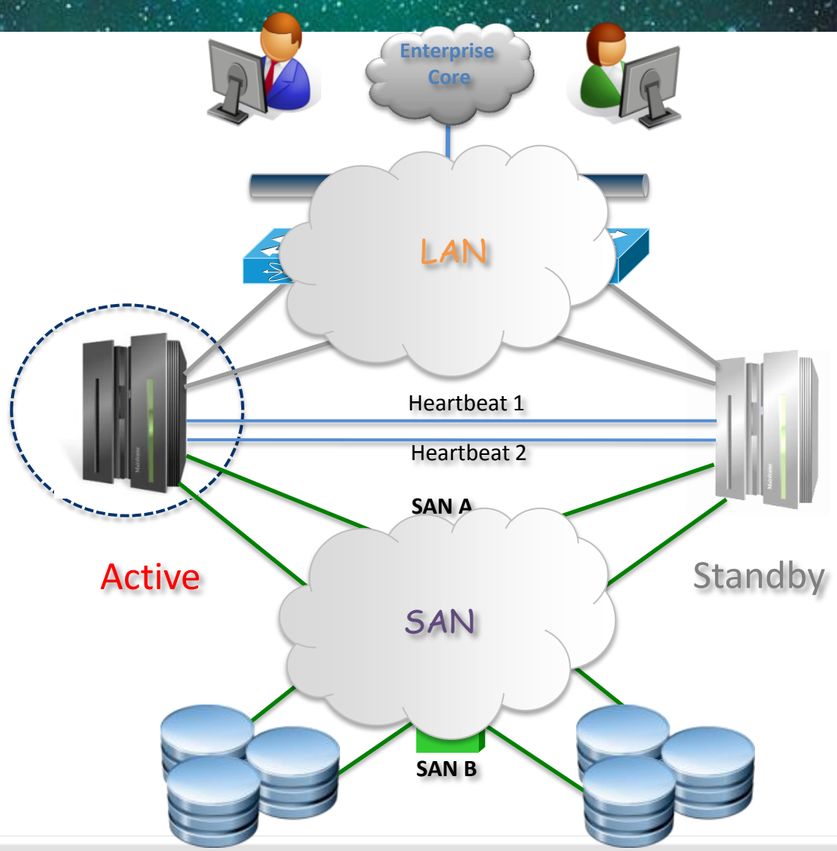
- **Optical Network**
  Create extended Campus using dark fibre or metro distances using DWDM link between DC's
- **Cisco UCS, MDS & Nexus Switches**
  Improved performance, reduce latency, accelerate installation and reduced cabling
- **GSLB or LISP**
  IP localisation Path Optimisation
- **Synchronous Data Replication**
  Applications run Active / Active (from both DC's)
- **System Virtualisation (Hypervisors)**
  Vmware, Microsoft, Citrix, RedHat, to leverage the business continuity with the mobility of Virtual Machines Inter DC.

A/A DC Driver: Business Continuance

High Availability Clusters - Local

Cluster Application such as…

- Microsoft MSCS
- Vmware Cluster (Local)
- Solaris Sun Cluster Enterprise
- Vmware cluster (Local)
- Oracle RAC
- IBM HCMP
- ….

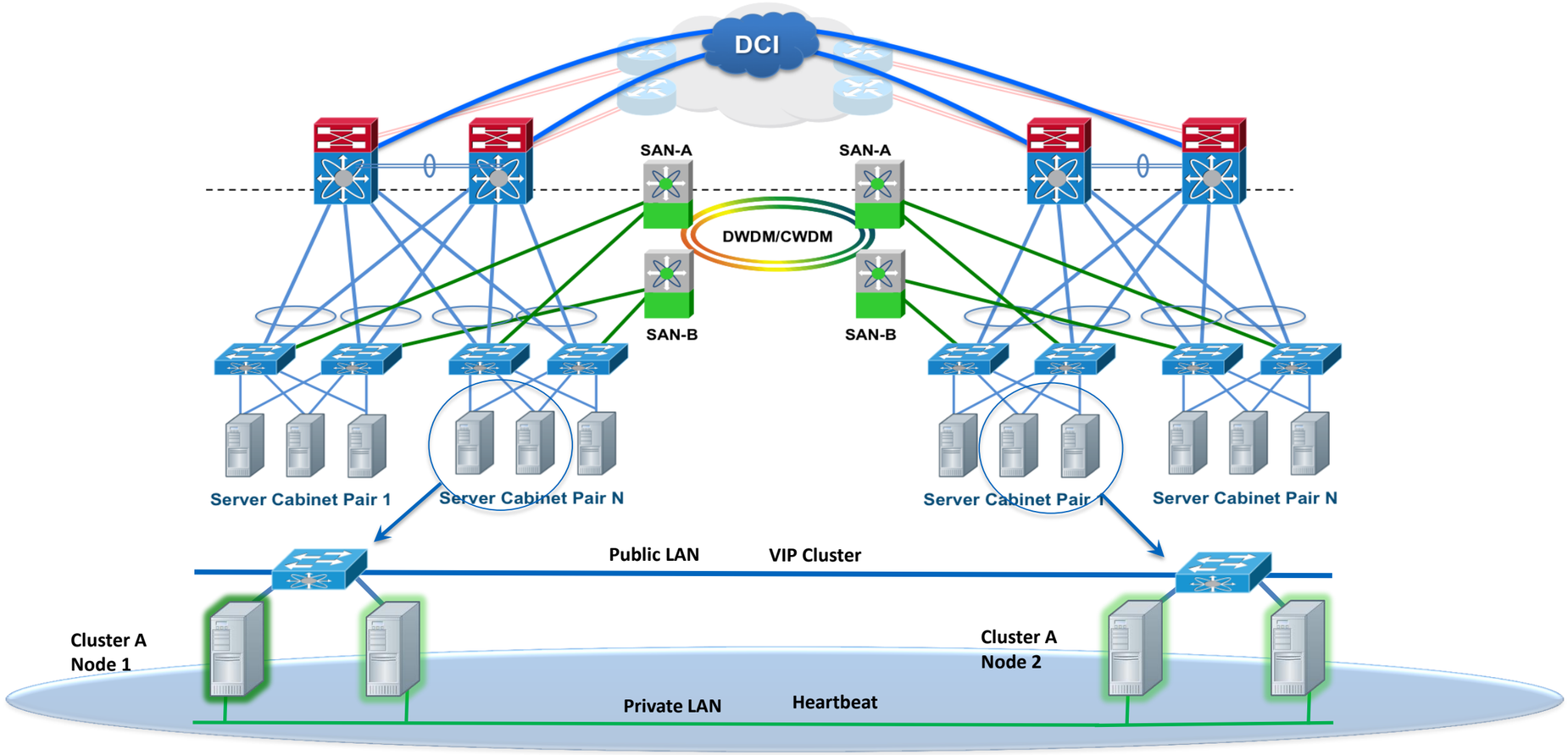


- Typically Active/Standby Cluster failover; Failure transfers Storage ownership
- Inter-server heartbeats, status & control synchronised through private network as well as VIP cluster through the public networks

  Requires Layer 2 path between hosts
- Client reconnection transparent - shared IP address → Layer 2 must be “extended”

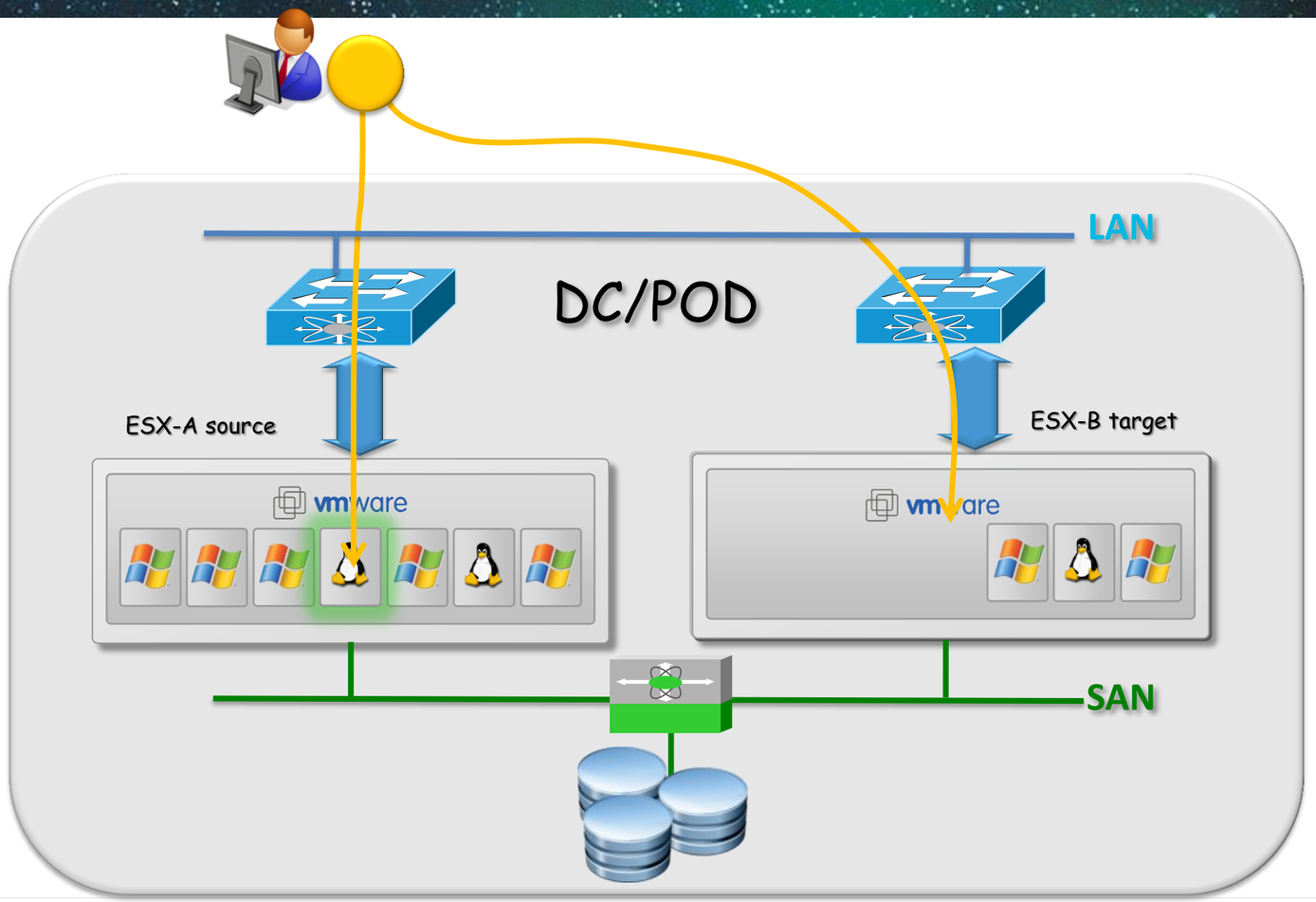
A/A DC Driver: Business Continuance

Multi-Site Geographically Dispersed HA Clusters



- Enhances HA Clusters to protect against Catastrophic site-level failures
- Clustering applications typically require “Stretched” L2 VLANs to peer DC sites
- Some applications support clustering using L3 for Inter-site routing

Virtualised Workload Mobility
Intra-DC Deployment

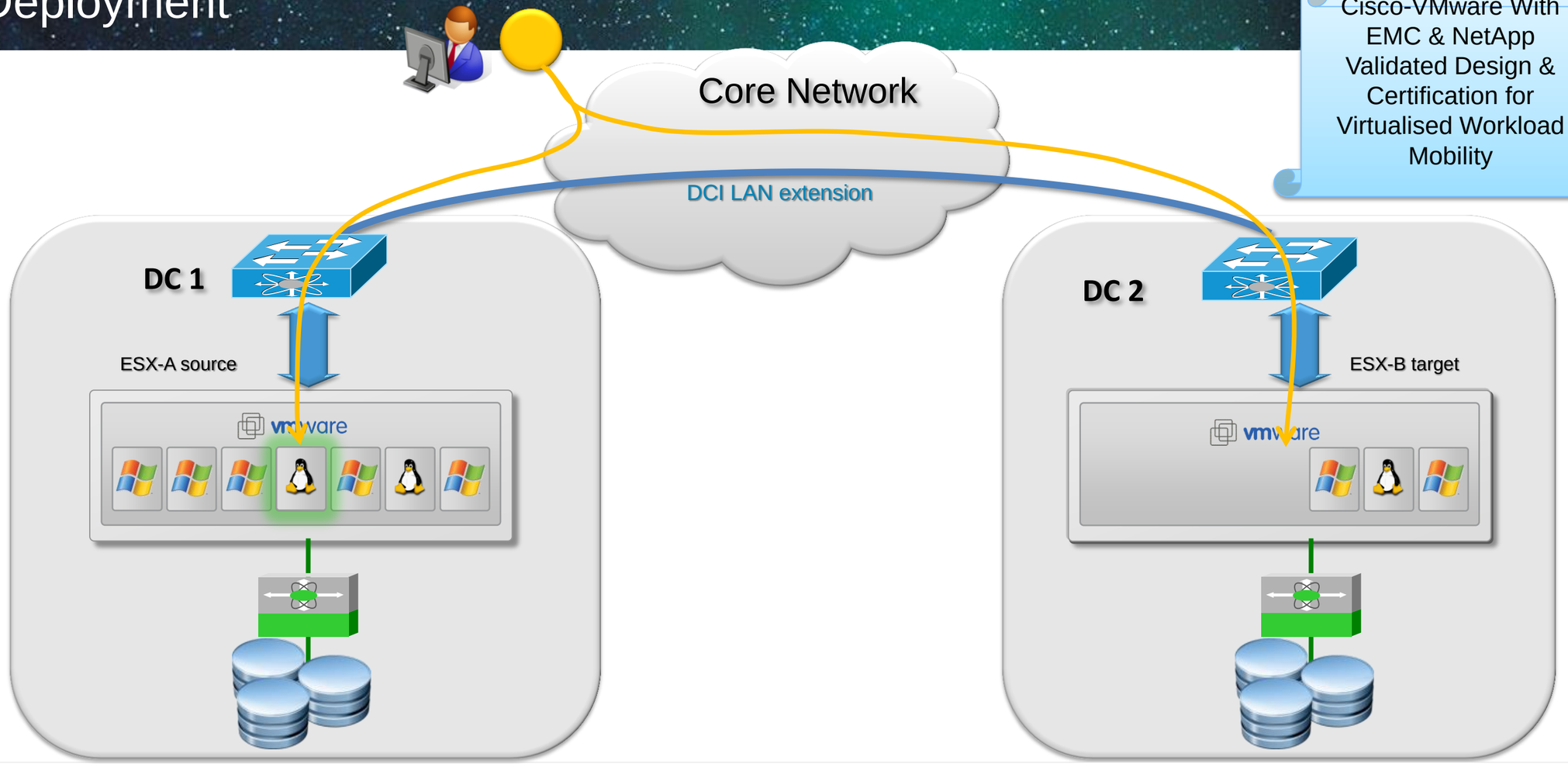


- Virtual Machines migration increases application availability
- L2 network adjacency between ESX hosts is currently required
- Consistent LUN ID must be maintained for stateful migration

A/A DC Driver: Virtualised Workload Mobility

Inter-DCs Deployment

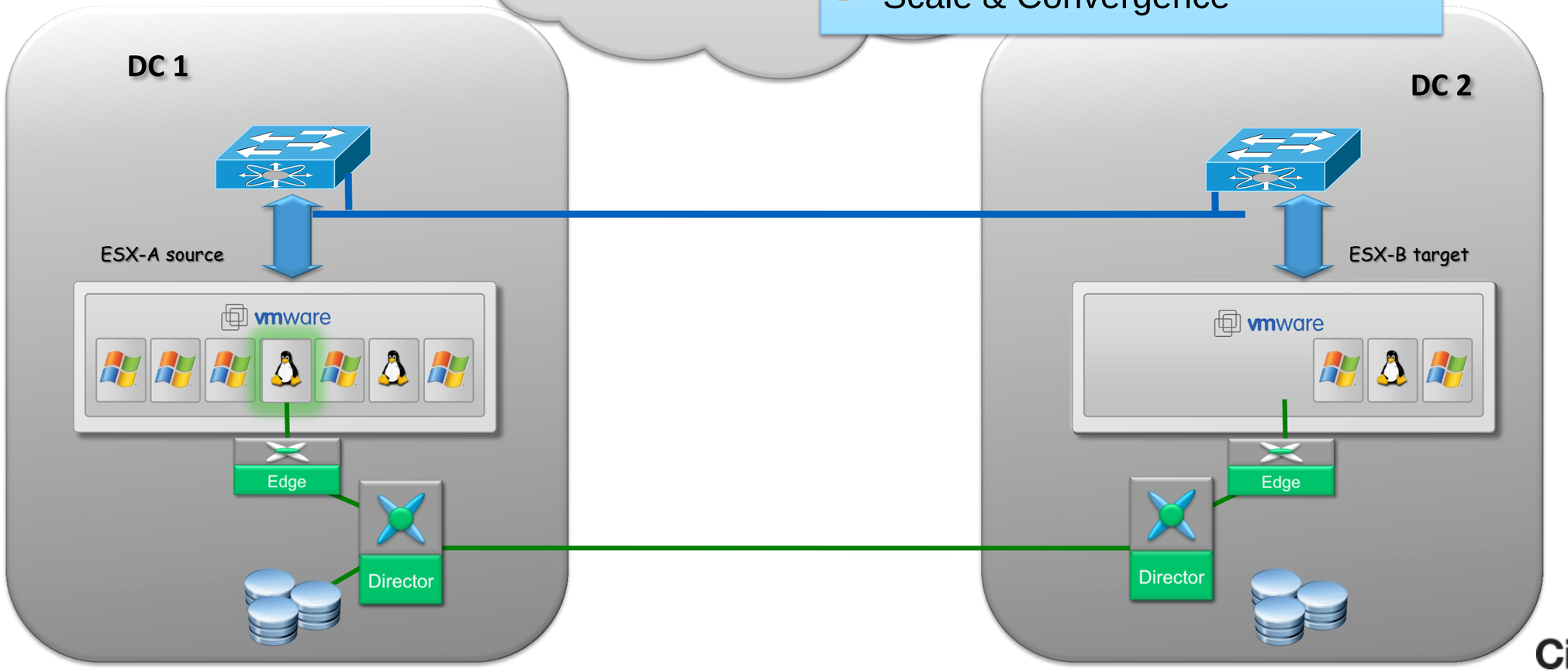
Cisco-VMware With EMC & NetApp Validated Design & Certification for Virtualised Workload Mobility



- Virtualised Workload Mobility across geographically dispersed sites
- Requires to stretch VLANs and to provide consistent LUN ID access
- Disaster Recovery Applications Ex- VMware Site Recovery Manager (SRM)

Data Centre Interconnect
LAN Extension

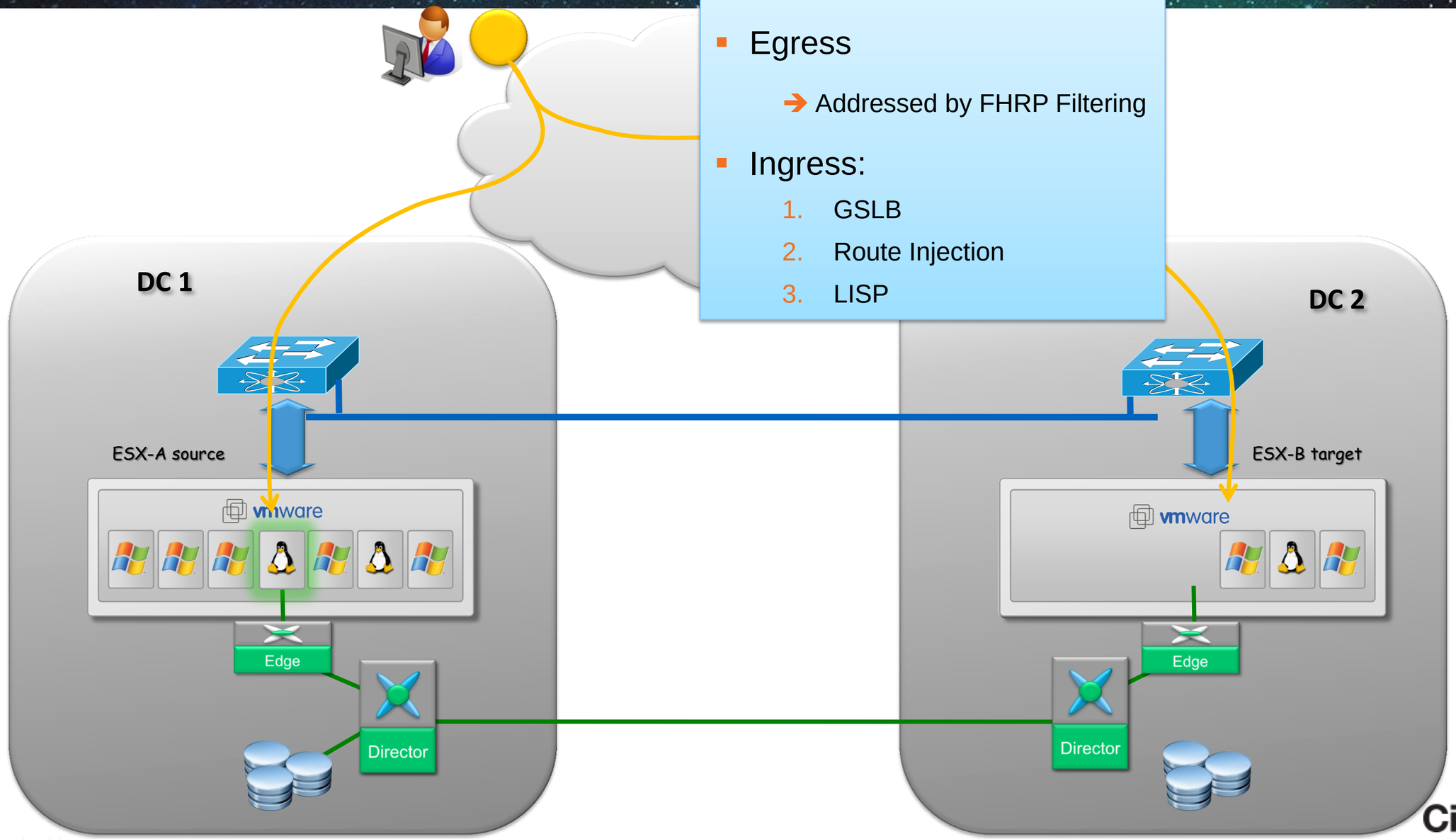
- STP Isolation is the key element
- Multipoint
- Loop avoidance + Storm-Control
  Unknown Unicast & Broadcast control
- Link sturdiness
- Scale & Convergence



Data Centre Interconnect
Path Optimisation

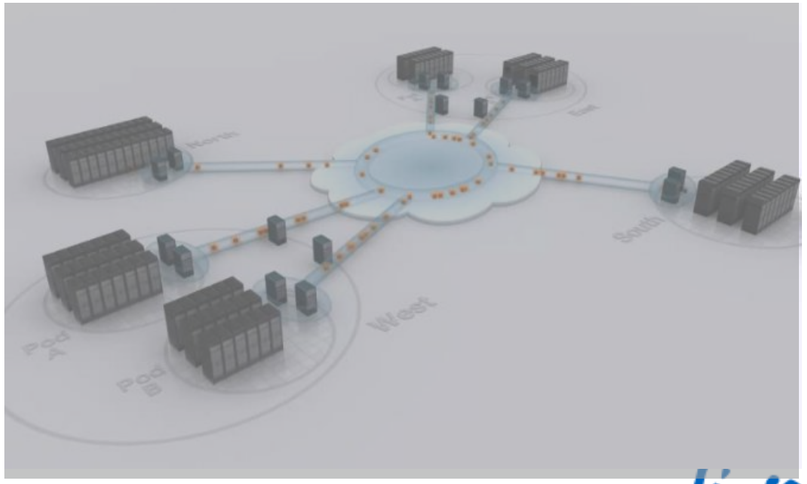
Options

- Egress
  - ➔ Addressed by FHRP Filtering
- Ingress:
  1. GSLB
  2. Route Injection
  3. LISP

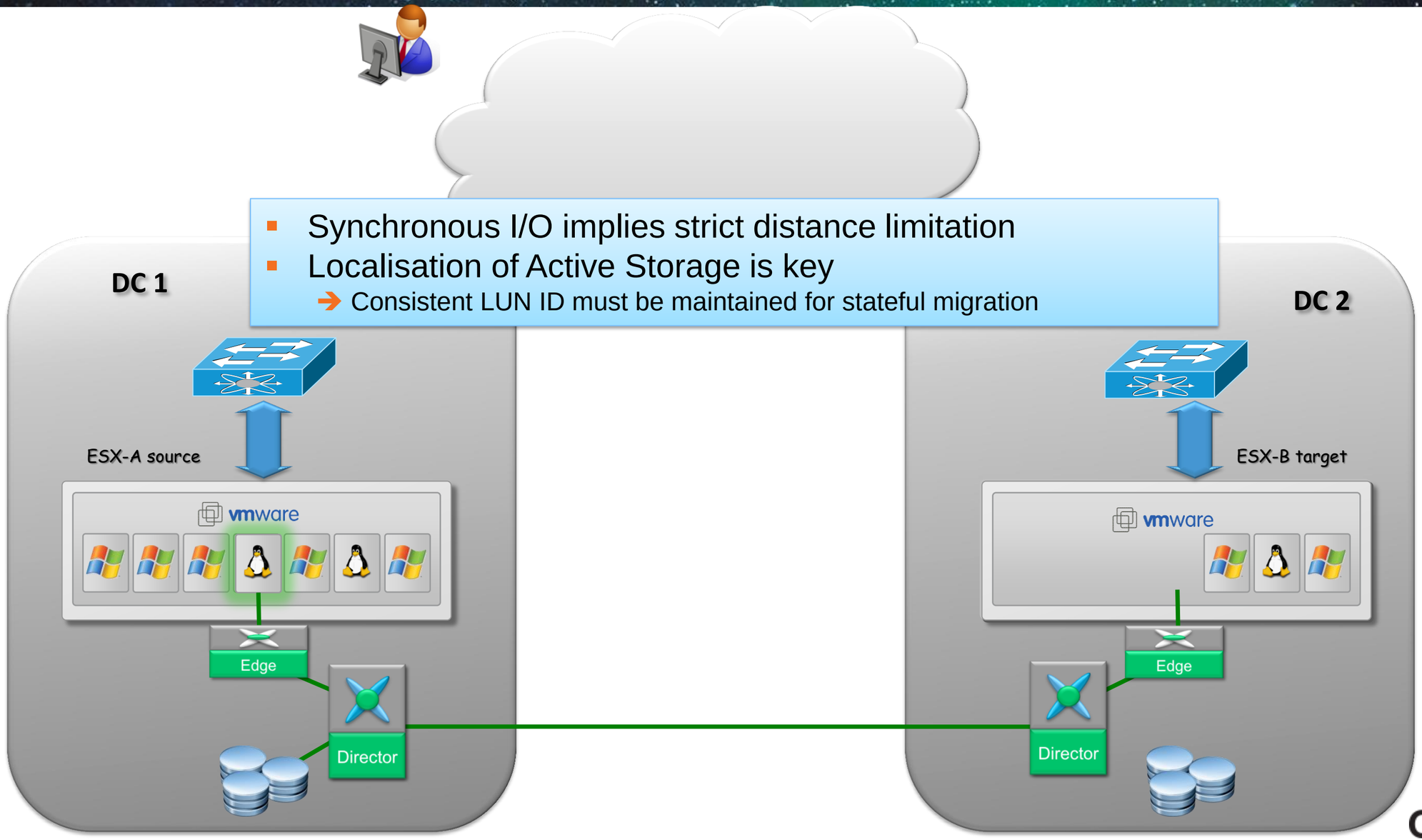


Agenda

- Active-Active Data Centre: Business Drivers and Solutions Overview
- Active / Active Data Centre Design Considerations
  - Storage Extension
  - Data Centre Interconnect (DCI) - LAN Extension Deployment Scenarios
  - Host Mobility using LISP and OTV
  - Network Services and Applications (Path optimisation)
- Cisco ACI and Active / Active Data Centre
- Summary and Conclusions
- Q&A

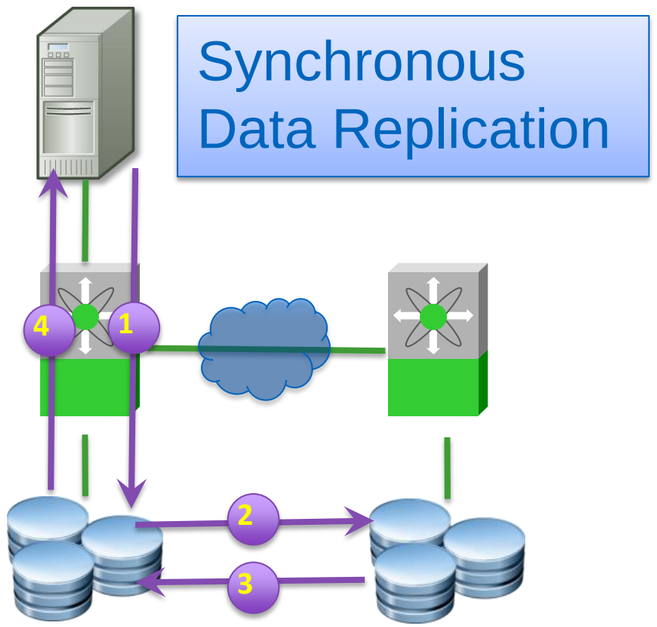


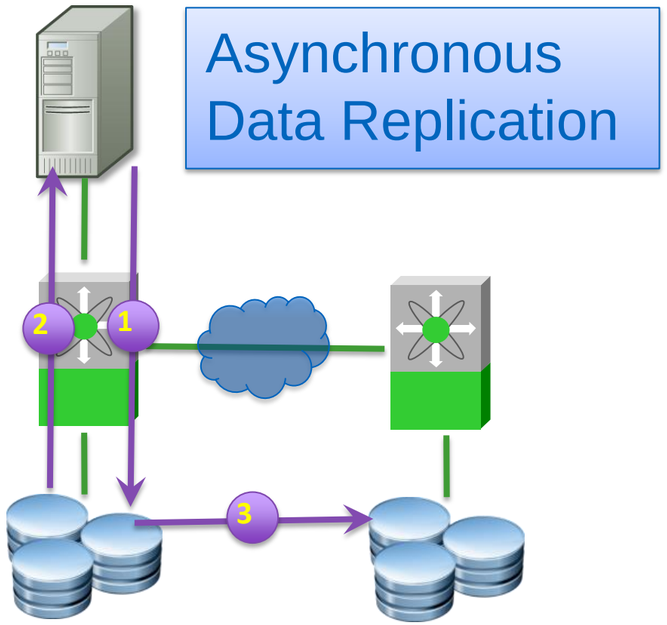
Data Centre Interconnect
SAN Extension



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 (or Zero RPO)
  - 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

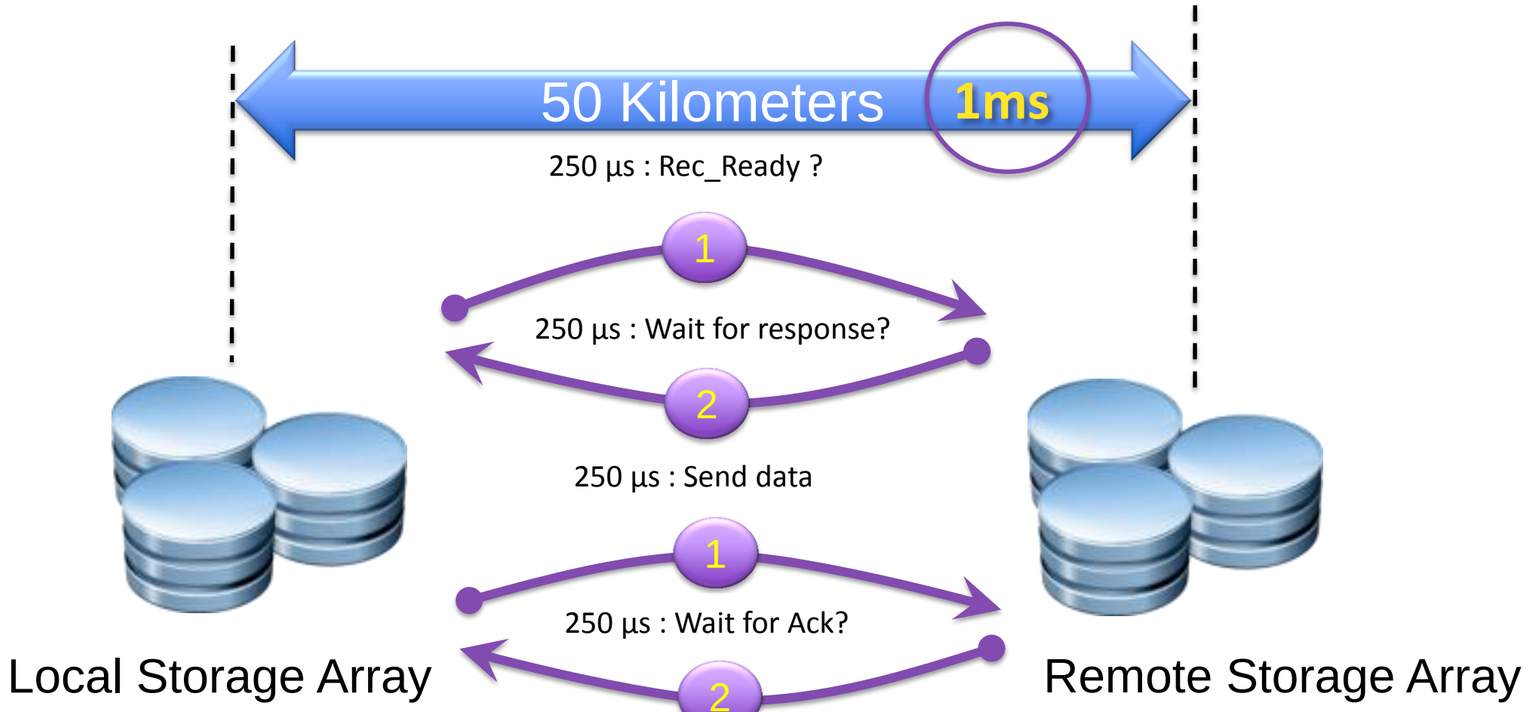




Synchronous Data Replication
Network Latency

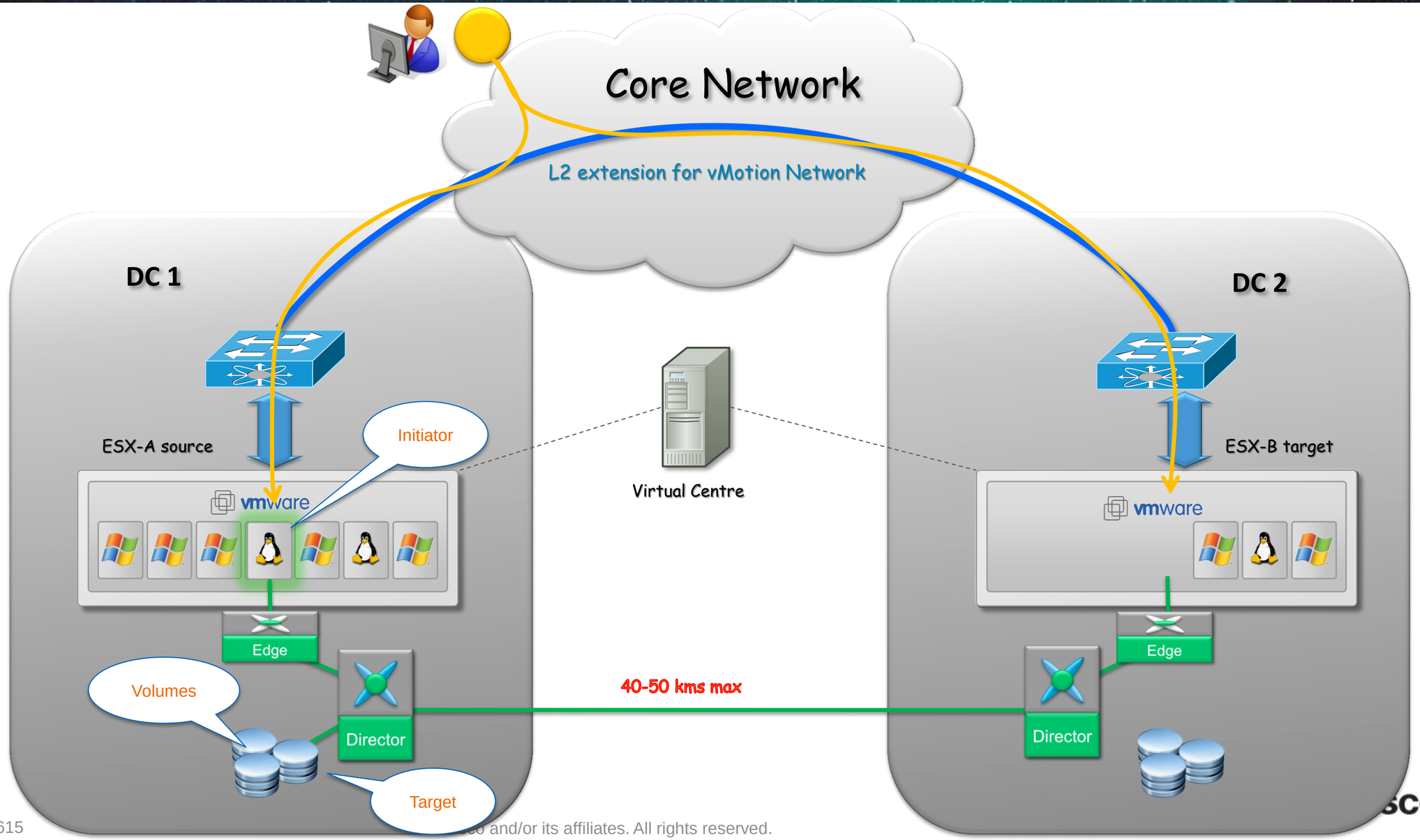
- Speed of Light is about 300000 Km/s
- Speed is reduced to 200000 Km/s ➔ 5 µs per Km (8 µs per Mile)
- That gives us an average of **1ms** for the light to cross **200 Kms** of fibre

- Synchronous Replication: SCSI protocol (FC) takes two round trips
- For each Write cmd each round trips is about 10 µs per kilometre
  ➔ 20µs/km for 2 round trips for Synch data replication



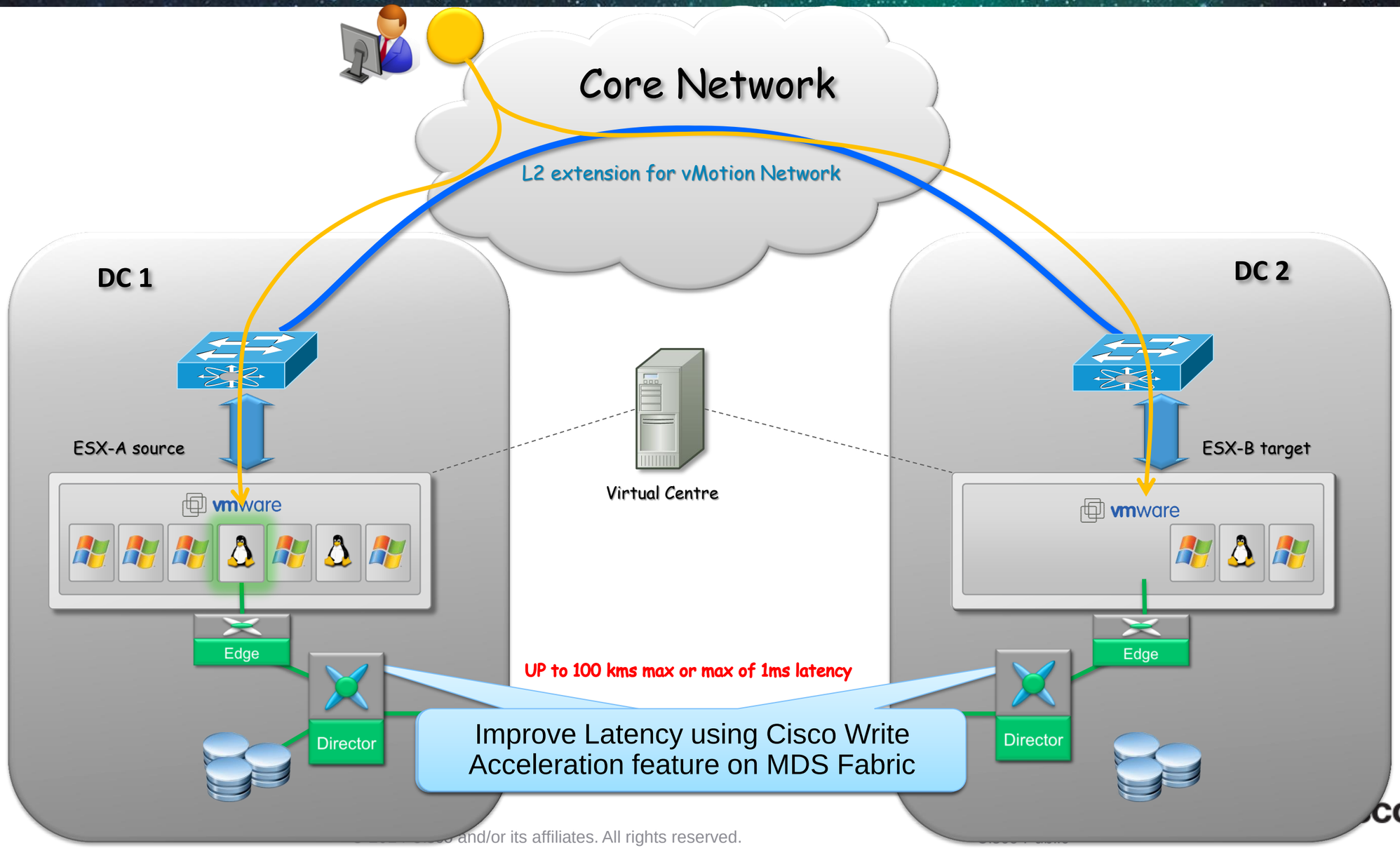
Storage Deployment in DCI

Option 1 - Shared Storage



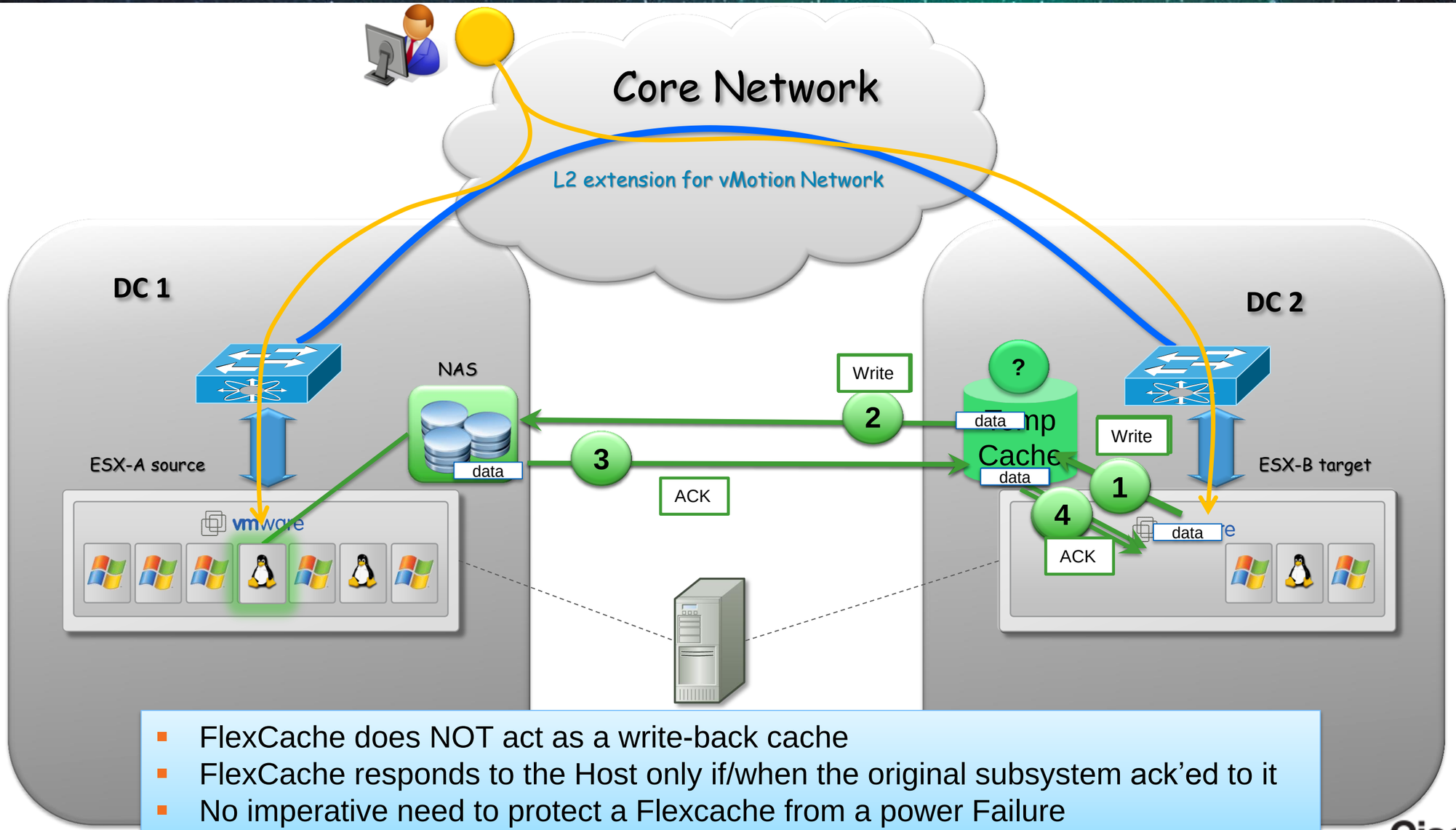
Storage Deployment in DCI

Shared Storage Improvement Using Cisco IOA



Storage Deployment in DCI

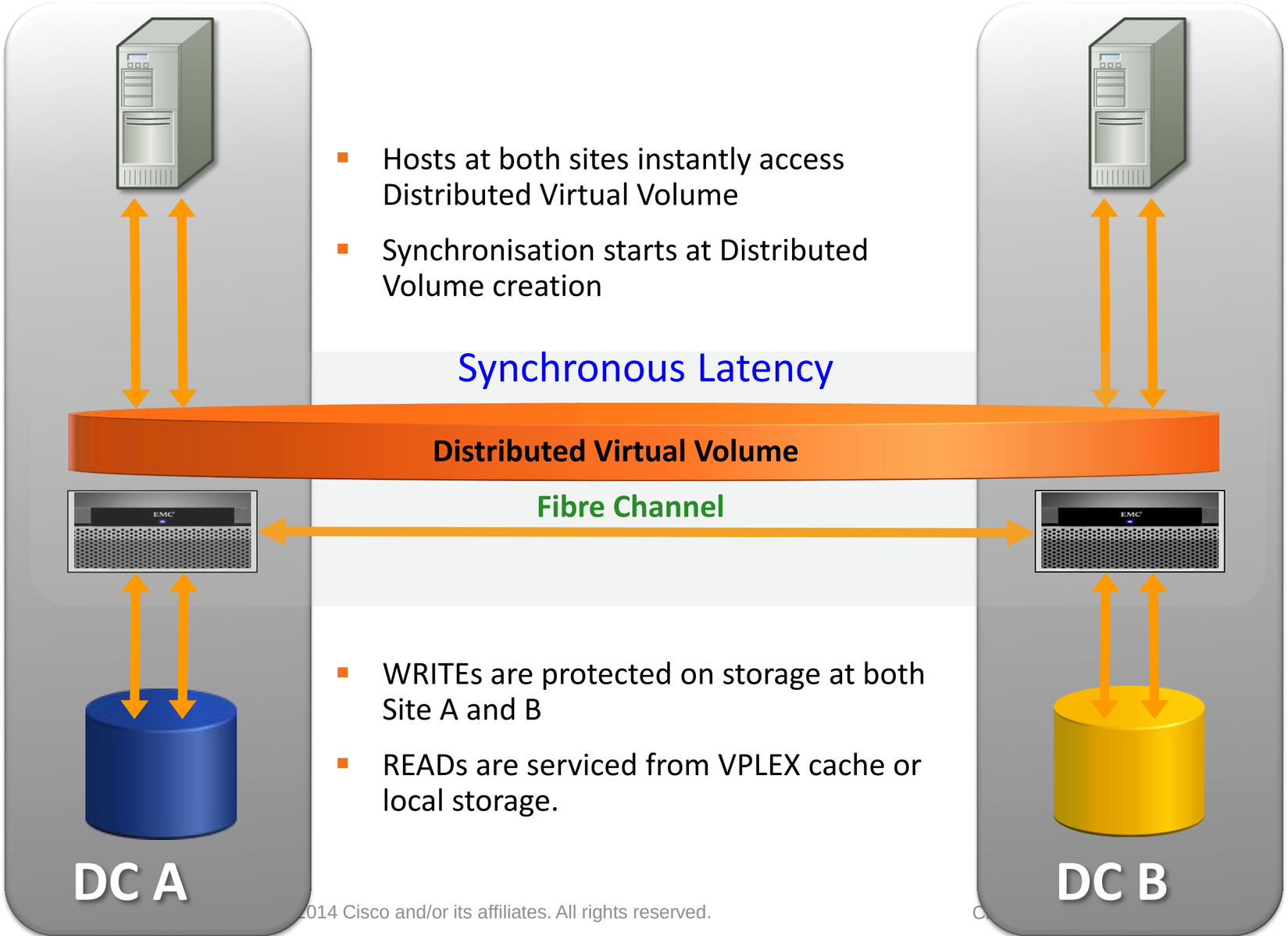
Option 2 - NetApp FlexCache (Active/Cache)



- FlexCache does NOT act as a write-back cache
- FlexCache responds to the Host only if/when the original subsystem ack'ed to it
- No imperative need to protect a Flexcache from a power Failure

http://www.cisco.com/en/US/docs/solutions/Enterprise/Data_Center/DCI/4.0/Netapp/dciNetapp.html

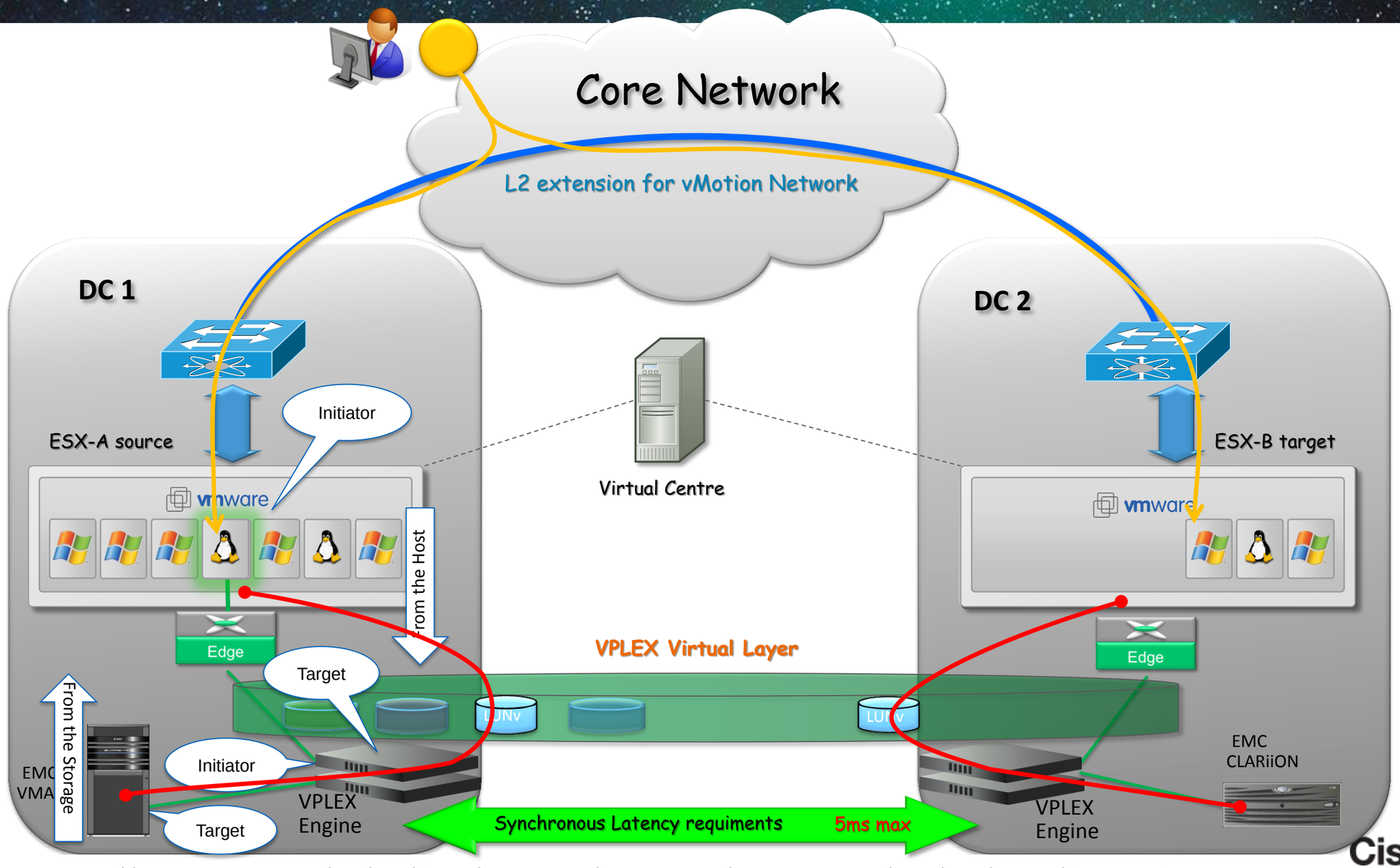
Storage Deployment in DCI
Option 3 - EMC VPLEX Metro (Active/Active)



- Hosts at both sites instantly access Distributed Virtual Volume
- Synchronisation starts at Distributed Volume creation

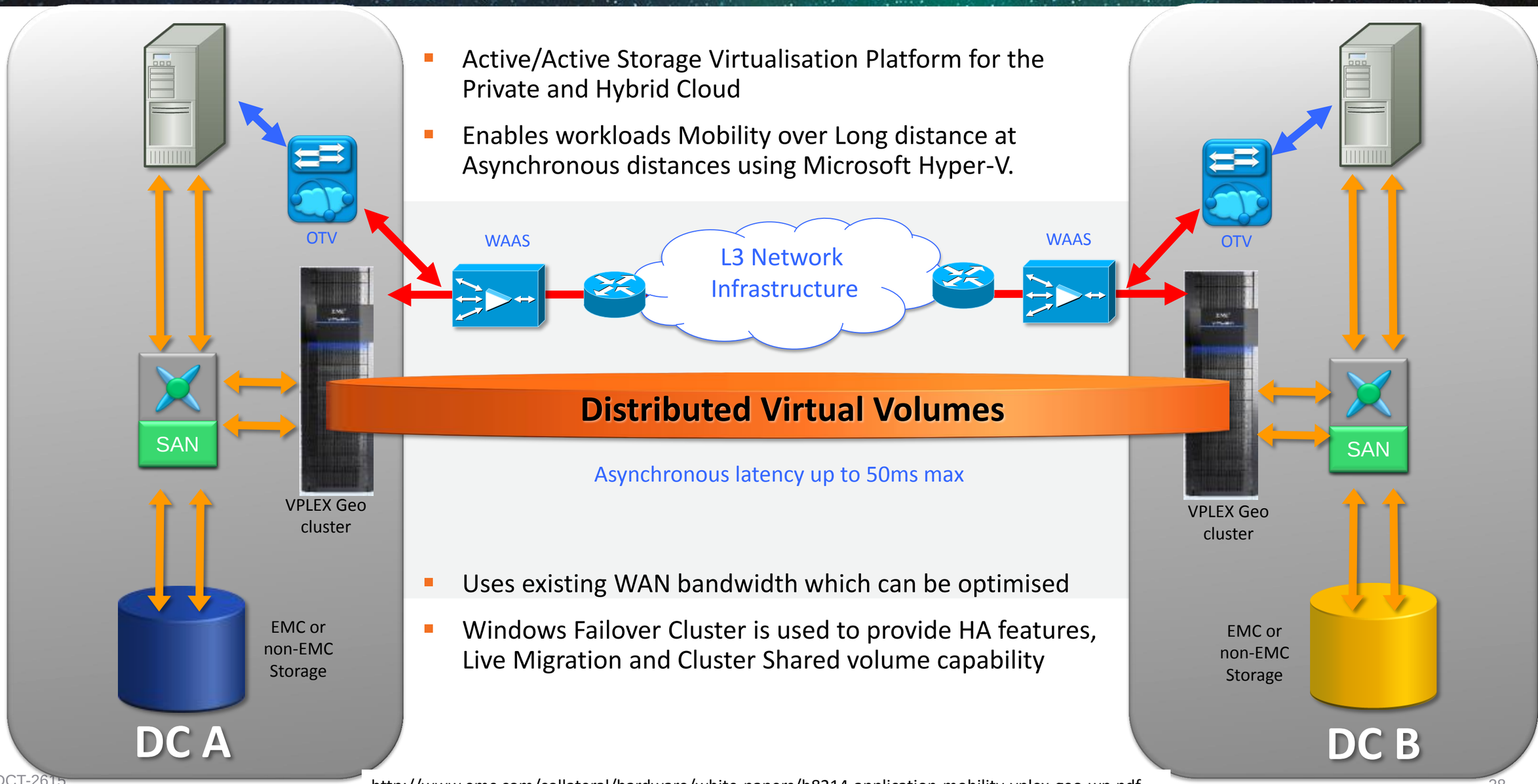
- WRITEs are protected on storage at both Site A and B
- READs are serviced from VPLEX cache or local storage.

Storage Deployment in DCI
Option 3 - EMC VPLEX Metro (Active/Active)



http://www.cisco.com/en/US/docs/solutions/Enterprise/Data_Center/DCI/4.0/ EMC/dciEmc.html

Storage Deployment in DCI
Option 4 - EMC VPLEX Geo (Active/Active)



- Active/Active Storage Virtualisation Platform for the Private and Hybrid Cloud
- Enables workloads Mobility over Long distance at Asynchronous distances using Microsoft Hyper-V.
- Uses existing WAN bandwidth which can be optimised
- Windows Failover Cluster is used to provide HA features, Live Migration and Cluster Shared volume capability

http://www.emc.com/collateral/hardware/white-papers/h8214-application-mobility-vplex-geo-wp.pdf

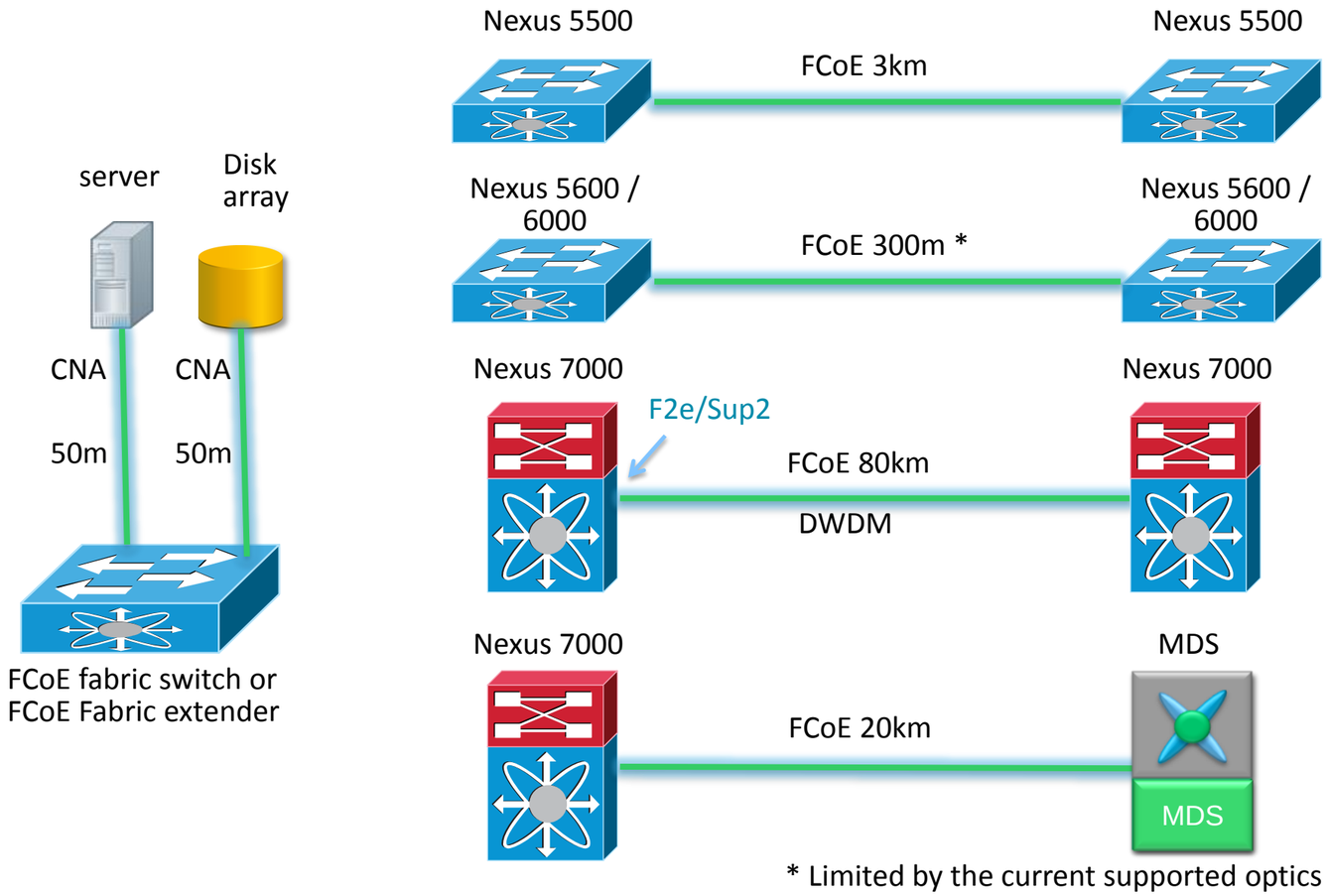
FC Extension Options

FC, FCoE, FCIP: Max Distances

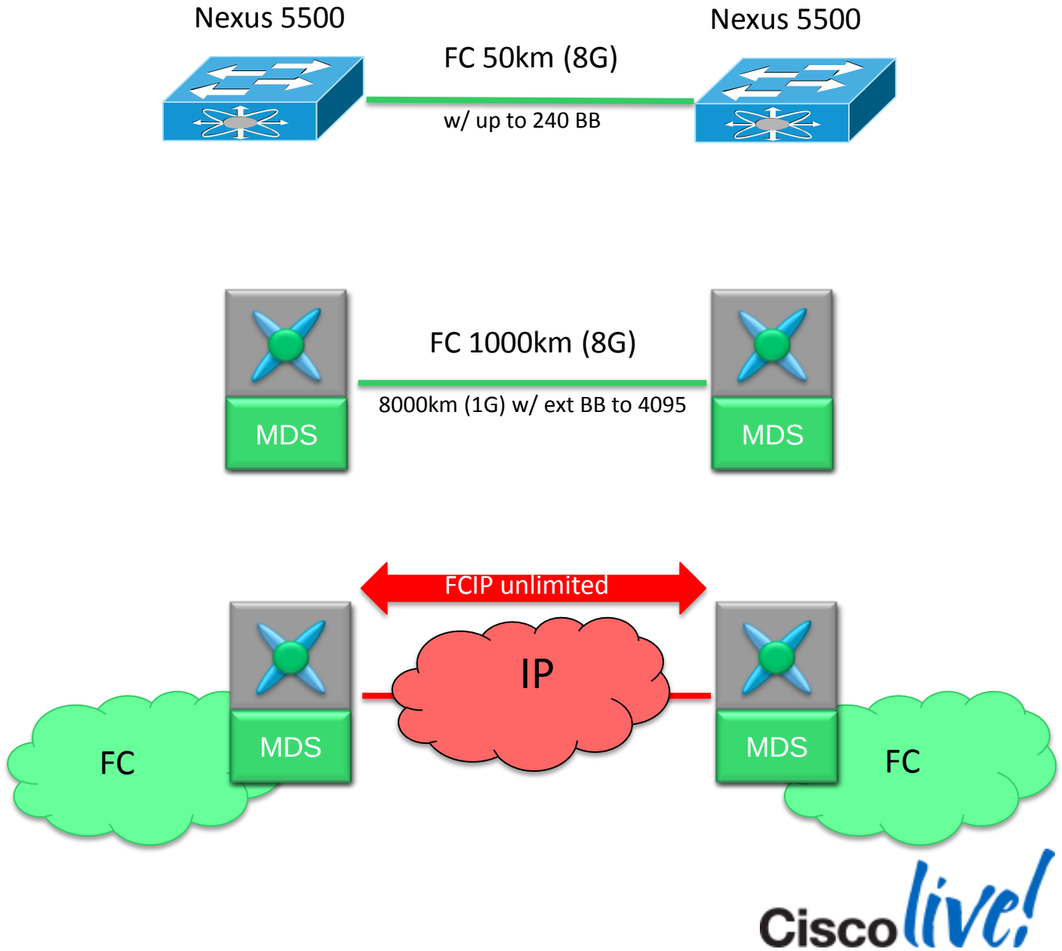
Native FC
FCoE

FCoE options

- ✓ **Requirement:** Maintain loss-less behaviour between the Point-to-Point link
- ✓ Supported distance is governed by the egress buffer size on the switch as well as the qualified optics

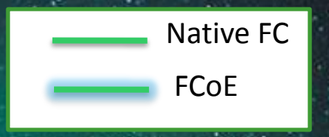


FC options



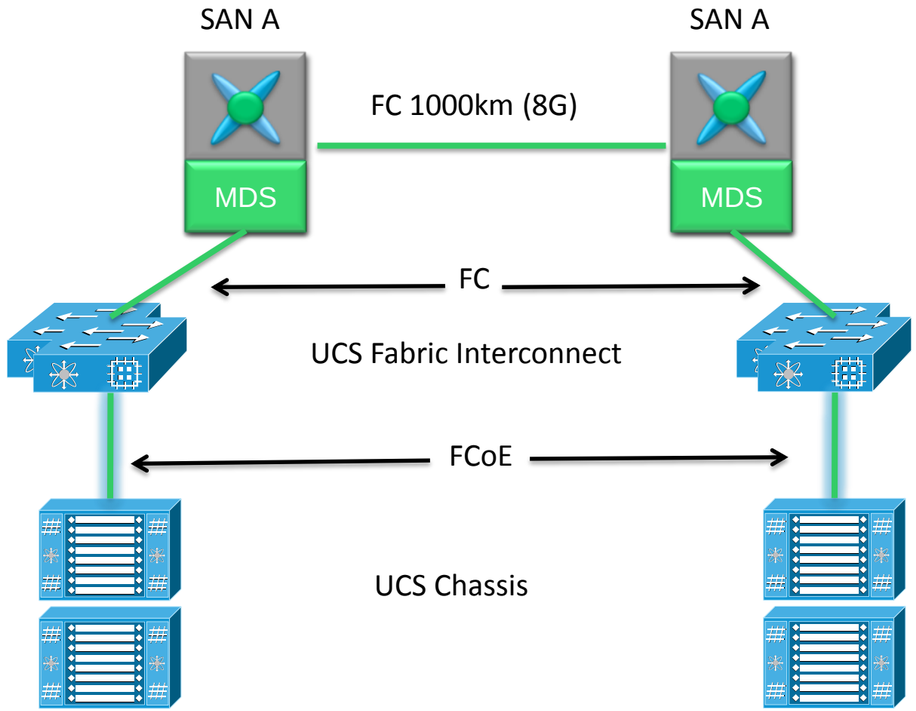
FC Extension Options

Max Distance with UCS



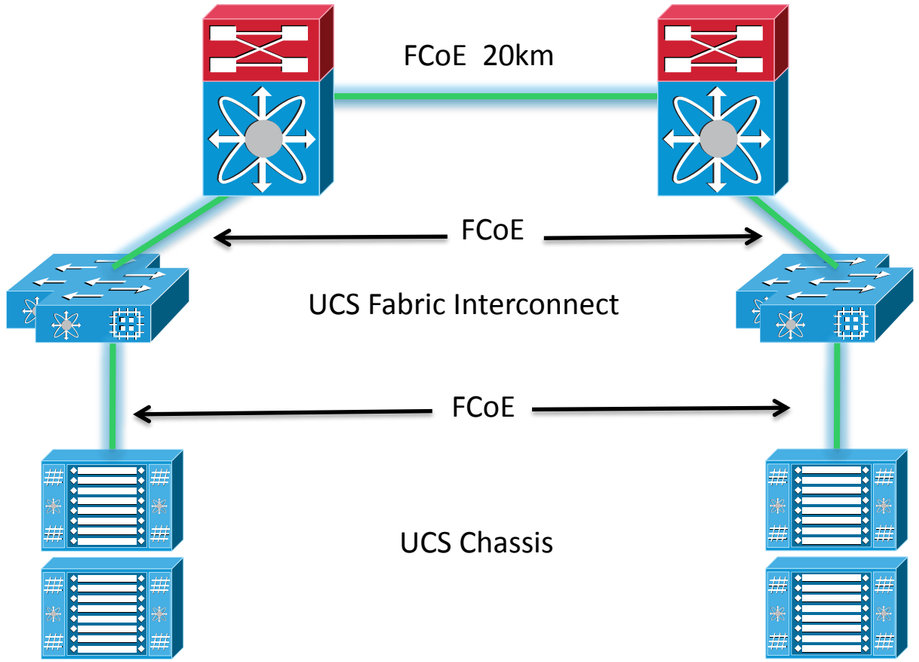
Previous options

- ✓ Currently the Fabric Interconnect does not support Multi-hop FCoE



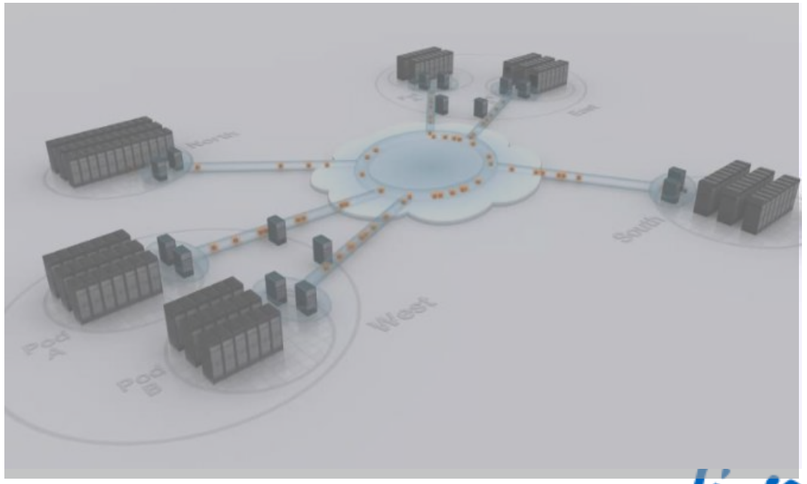
Current options rel. 2.1

- ✓ FCoE NPV / Multi-hop FCOE support on fabric Interconnect 6100 & 6200



Agenda

- Active-Active Data Centre: Business Drivers and Solutions Overview
- Active / Active Data Centre Design Considerations
  - Storage Extension
  - Data Centre Interconnect (DCI) - LAN Extension Deployment Scenarios
  - Host Mobility using LISP and OTV
  - Network Services and Applications (Path optimisation)
- Cisco ACI and Active / Active Data Centre
- Summary and Conclusions
- Q&A



LAN Extension with DCI
VLAN Types

✓ Type T0

Limited to a single access layer device

✓ Type T1

Extended inside an aggregation block (POD)

✓ Type T2

Extended between PODs part of the same DC site

✓ Type T3

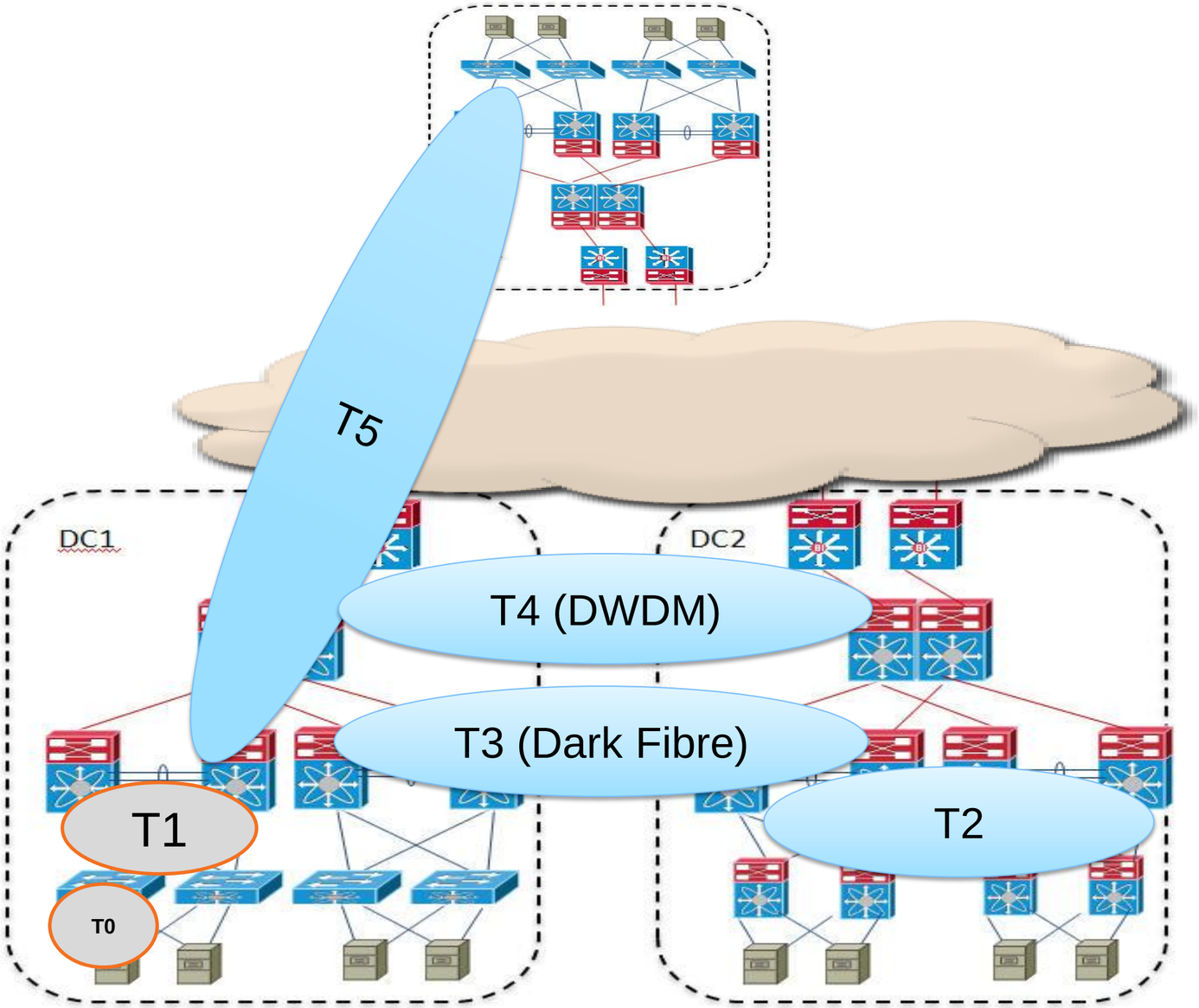
Extended between PODs part of twin DC sites connected via dedicated dark fibre links

✓ Type T4

Extended between PODs part of twin DC sites connected via xWDM links

✓ Type T5

Extended between PODs part of distant remote DC sites



LAN Extension with DCI

Usage and Technology Mapping

VLAN Type	Usage
Type 0	Routed only (or isolated)
Type 1	Clustering intra-POD, VM provisioning flexibility
Type 2	Clustering intra-DC (inter PODs), VM provisioning flexibility
Type 3/4	Disaster Avoidance, Active/Active sites, VM provisioning flexibility (live motion)
Type 5	Disaster Recovery, migration (cold motion)

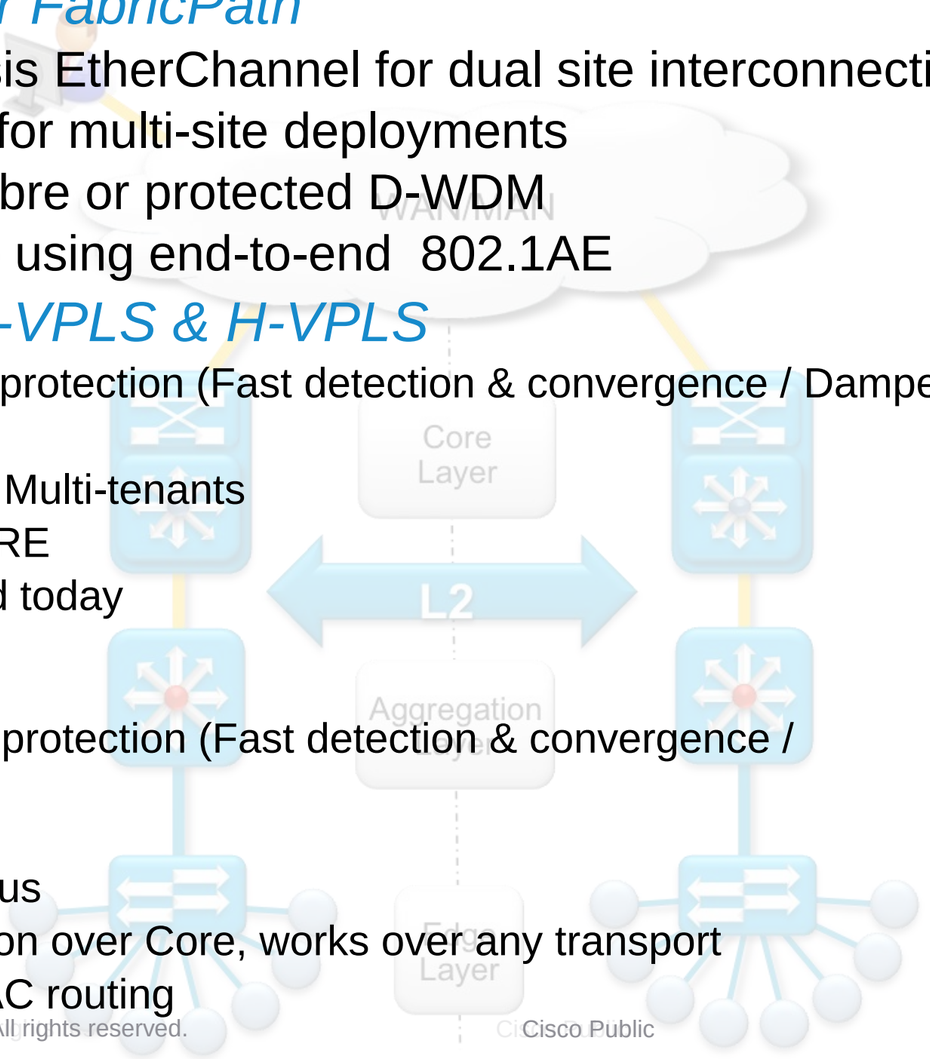
LAN Extension
Key Technical Challenges / Points of attention

- L2 control-plane
  - ✓ STP domain scalability
  - ✓ STP fault domain isolation
  - ✓ L2 Multi-Homing
- L2 data-plane
  - ✓ Bridging data-plane flooding & broadcasting storm control
  - ✓ Outbound MAC learning
- Inter-site transport
  - ✓ Long distance link protection with fast convergence
  - ✓ Point to Point & Multi-points bridging
  - ✓ Path diversity
  - ✓ L2 based Load repartition
  - ✓ Optimised routing egress & ingress
  - ✓ LAN Extension over Layer 3 cloud (IP, MPLS)
  - ✓ Multicast optimisation

Technology challenge:
- ✓ L2 is weak
- ✓ IP is not mobile

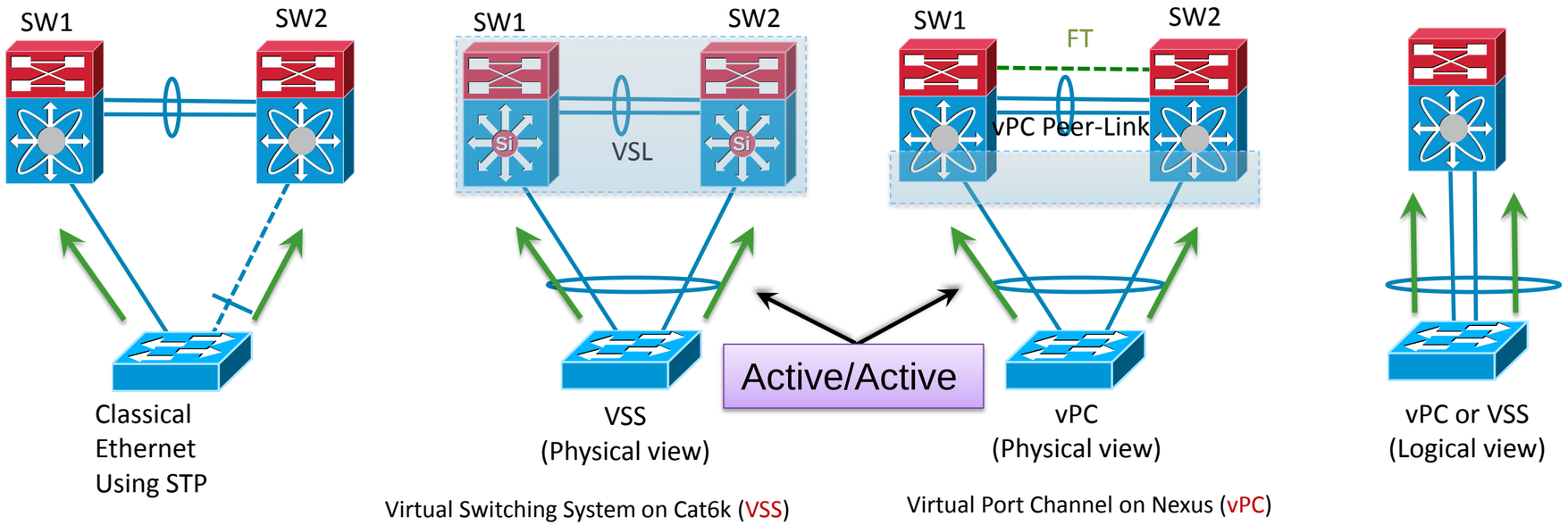
LAN Extension
Technology Selection Criteria

Technology Nature	Selection Criteria
Ethernet	➢ *VSS & vPC or FabricPath* ▪ Multi-Chassis EtherChannel for dual site interconnection ▪ FabricPath for multi-site deployments ▪ Over dark fibre or protected D-WDM ▪ Easy crypto using end-to-end 802.1AE
MPLS	➢ *EoMPLS & A-VPLS & H-VPLS* L2oL3 for link protection (Fast detection & convergence / Dampening) PE style Large scale & Multi-tenants Works over GRE Most deployed today
IP	➢ *OTV* L2oL3 for link protection (Fast detection & convergence / Dampening) CE style Enterprise focus Easy integration over Core, works over any transport Innovative MAC routing



Multi-Chassis EtherChannel (MEC)

Using Multi-Chassis Link Aggregation Control Protocol (mLACP)



- Both VSS-MEC and vPC are a Port-channeling concept extending link aggregation to two separate physical switches
- Allows the creation of resilient L2 topologies based on Link Aggregation.
- Eliminates the dependence on STP in the L2 access-distribution Layer
- Scale Available Layer 2 Bandwidth
- Simplify Network Design

Dual Sites Interconnection
Leveraging MECs Between Sites

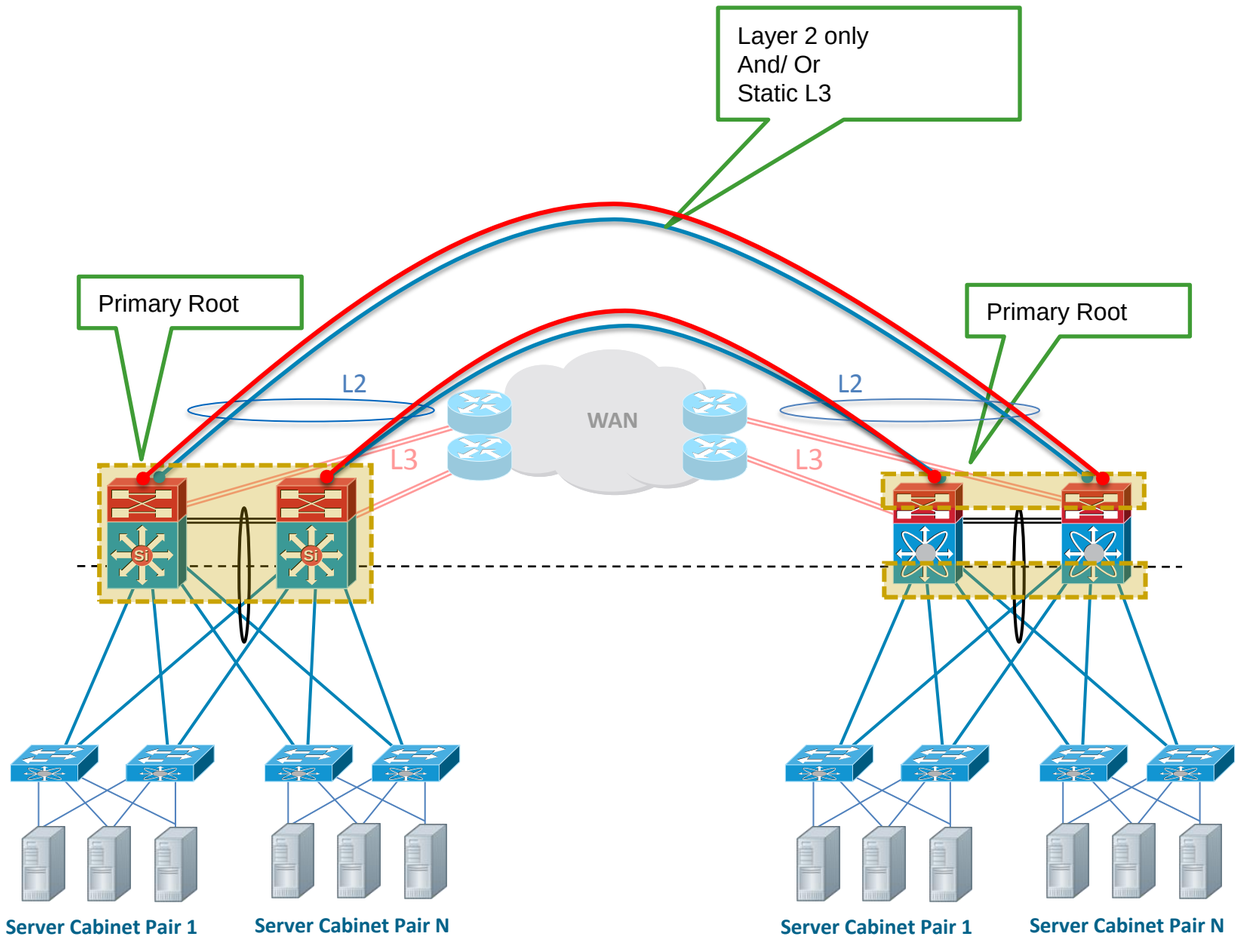
At DCI point:
- STP Isolation (BPDU Filtering)
- Broadcast Storm Control
- FHRP Isolation

- Link utilisation with Multi-Chassis EtherChannel
- Enable Fast LACP with DWDM
- DCI port-channel
  - 2 or 4 links
- Requires protected DWDM or Direct fibres

- Validated design:
  - *200 Layer 2 VLANs + 100 VLAN SVIs*
  - *1000 VLAN + 1000 SVI (static routing)*

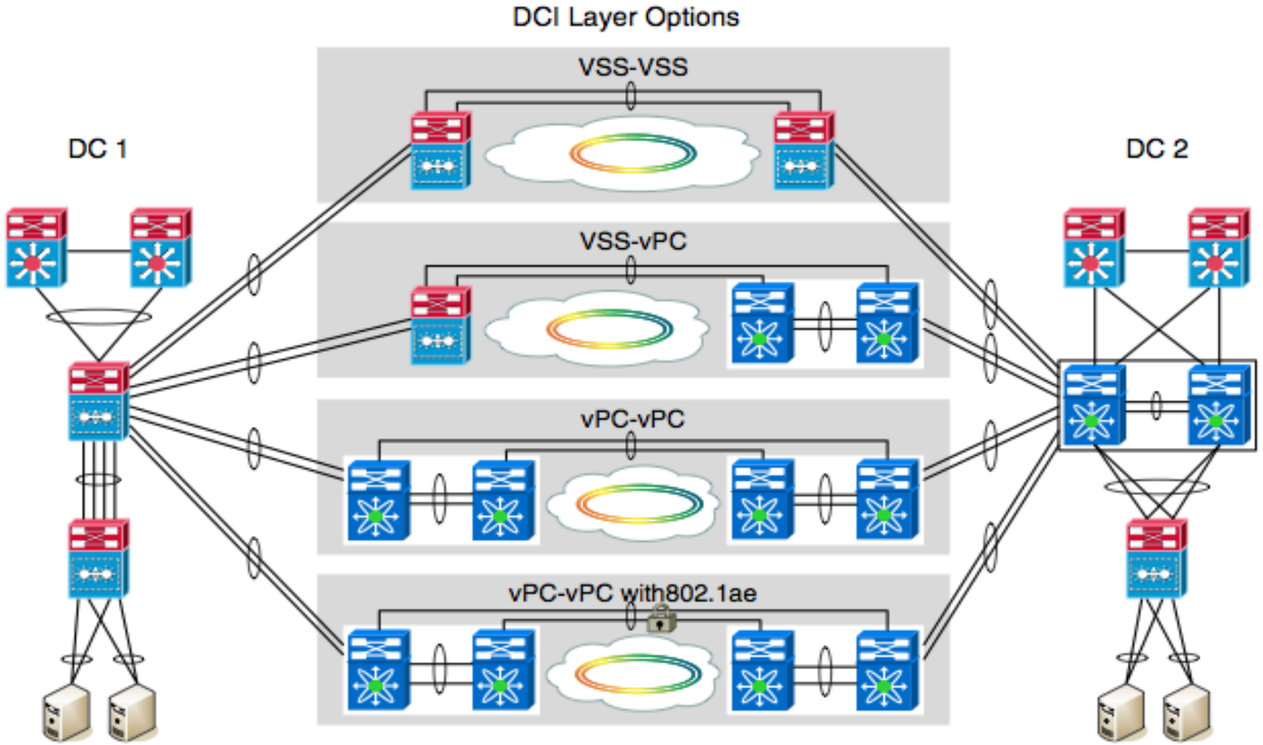
- Currently vPC does not support L3 peering:
  - Use dedicated L3 Links for Inter-DC routing!

- Support for L3 peering by CY2014
  - Requires F2 / F3 Line Card
  - Latest NX-OS release



Dual Sites Use Case Summary

Cisco Validated Design on CCO



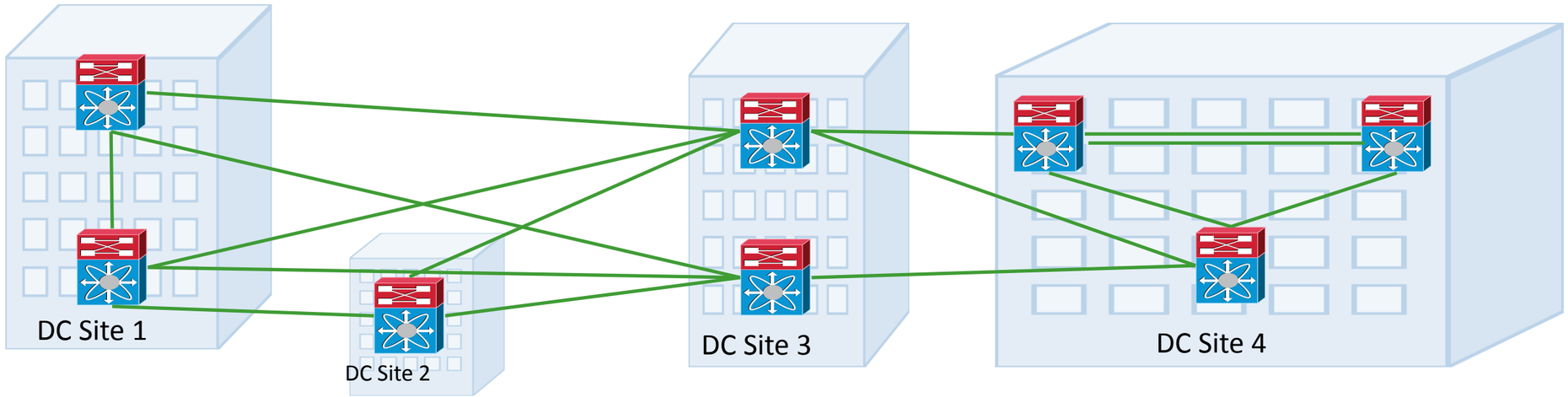
Test Case	Hardware failure Ucast	Hardware failure Mcast	Hardware restore Ucast	Hardware restore Mcast	Link Failure Ucast	Link failure Mcast	Link Restore Ucast	Link Restore Mcast
VSS-VSS	<1.7	<2.3	<1.1	<2.8	<1.3	<1.2	<1.7	<1.2
VSS-vPC	<1.3	<1.7	<2.0	<2.6	<1.2	<1.6	<1.5	<1.4
vPC-vPC	<1.5	<1.6	<2.8	<2.5	<1.2	<0.2	<0.2	<0.2

FabricPath
Problem Statement

Problem/Challenge	Fabric-Path Solution
Desire to deliver a 'workload anywhere' model	L2 Flexibility for allowing VLANs anywhere helps to reduce physical constraints on server location
Too much unused Bandwidth at Layer 2	Up to 16 active paths at L2, each path a 16 member port channel for Unicast and Multicast
Undesirable Failure Handling at Layer 2	Alternative to Spanning Tree 'drawbacks' Leveraging ISIS L3 routing concepts
Scaling MAC address tables in larger L2 domains	Hierarchical addressing + Conversational learning allow more efficient use of the available MAC table space.

FabricPath is primarily positioned for Clos-based architectures

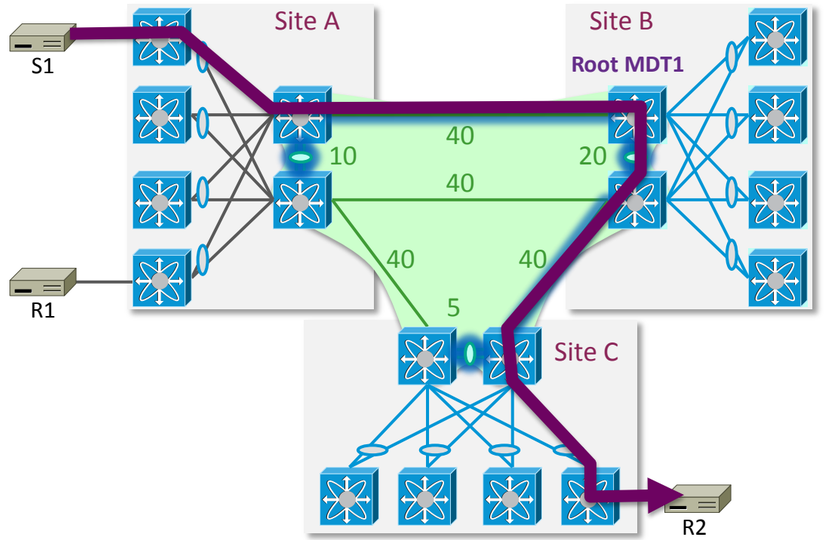
- L2 DCi is **NOT** LAN Switching! Is FabricPath a valid solution for DCI ?



- Perception on FabricPath DCi
  - Plug and play
  - No Spanning Tree events shared between DC sites
  - Can do IP routing over FP DCi
  - One single protocol to manage end to end
  - One single Fabric end to end
  - Works also with N5K only scenarios

FabricPath DCI
Lessons learned

- Dependencies with L1 WAN links
  - Requires point to point high quality connections
  - Golden rule : WAN links must support Remote Port Shutdown and micro flapping protection
- Multidestination traffic impacts
  - Must tune multicast tree to avoid local traffic to fly over root tree site
  - Cannot avoid multicast to fly over root tree site for DCI multicast
- IP routing over FabricPath
  - Ship in the night effect
  - OSPF hellos are multicast and will fly over root site



- STP interactions with FabricPath DCI
  - The Fabric becomes STP root for all propagated VLAN, means that twin site vPC will be blocking
- FabricPath & HSRP Localisation
  - HSRP Control-plane can be isolated with mismatching authentication key
  - But HSRP data-plane cannot be isolated when DC is also FP, leading to flapping vMAC
- High Availability
  - L2 ISIS fine tuning is required: allocate-delay timer, transition-delay, linkup-delay, spf-interval, lsp-gen-interval
  - Sub second convergence, except node recovery in 3s

FabricPath DCI - Key Takeaways

	Customer references	Operations simplicity	Domino effect prevention	DCI link quality mgmt	3+ Sites optimisation	High Availability	L2 functions	L3 Unicast functions	Multicast functions	Scalability
FabricPath	★★	★★★★★	★★	★★	★★★	★★★★	★★★	★★★	★★	★★★

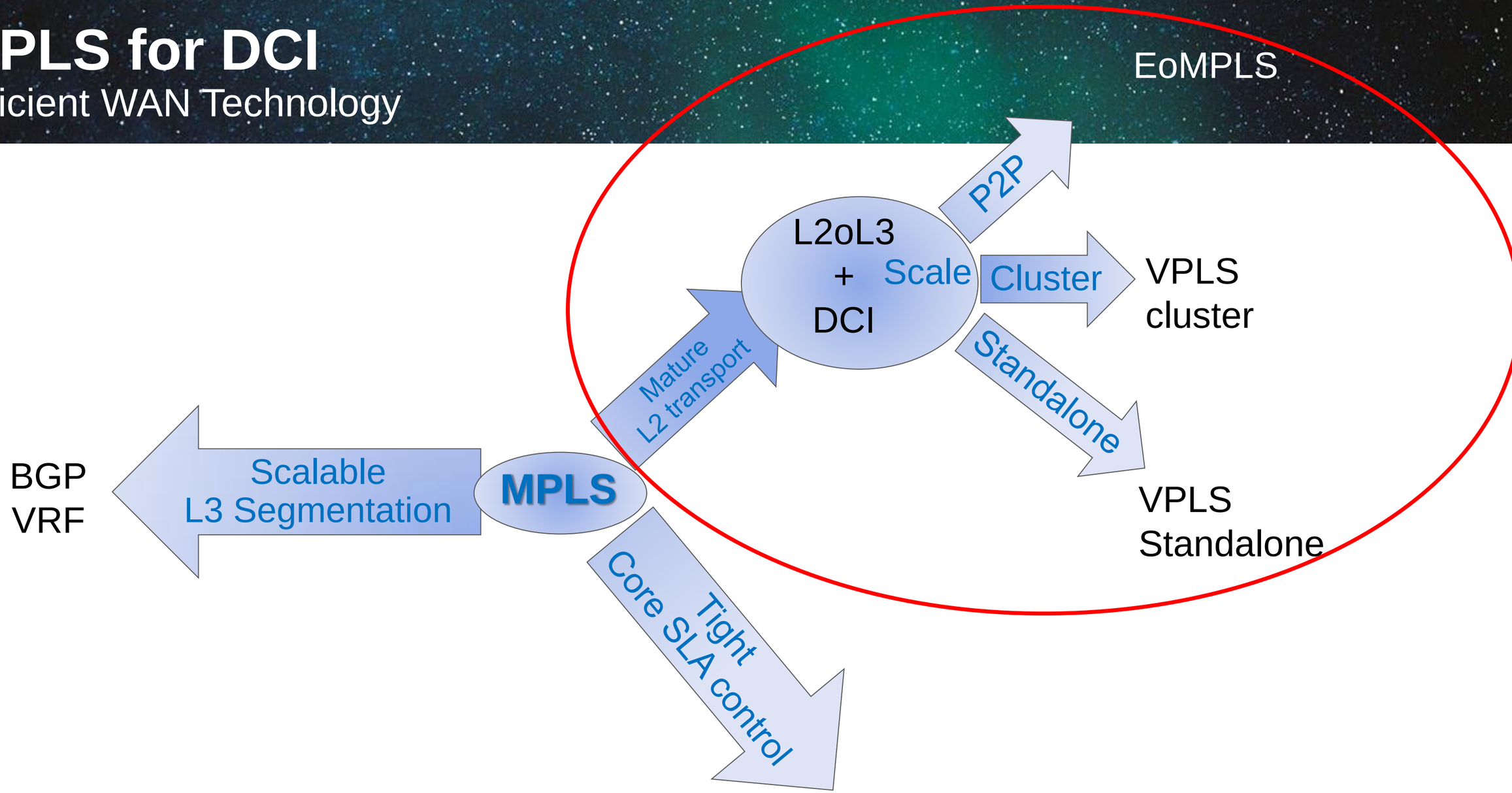
- On DCI, FabricPath is not so Plug and Play actually
  - No specific DCI functions compared to OTV, VPLS
  - Several designs gotchas but do not impact all customers
  - Multidestination Trees capacity planning may be complex
  - Multiple Topologies do enhance the overall solution
- Still a common control-plane between sites, even if hardened
- By default, OTV/VPLS should be the first solutions to be considered
  - Cisco Validated Designs (CVDs)
  - Specific DCI features
  - Offer an efficient independence between DC
- FabricPath is a valid DCI solution when :
  - Short distances between DCs via optical links (tromboning is not a issue)
  - Multicast is not massively used




MPLS-Based Solutions

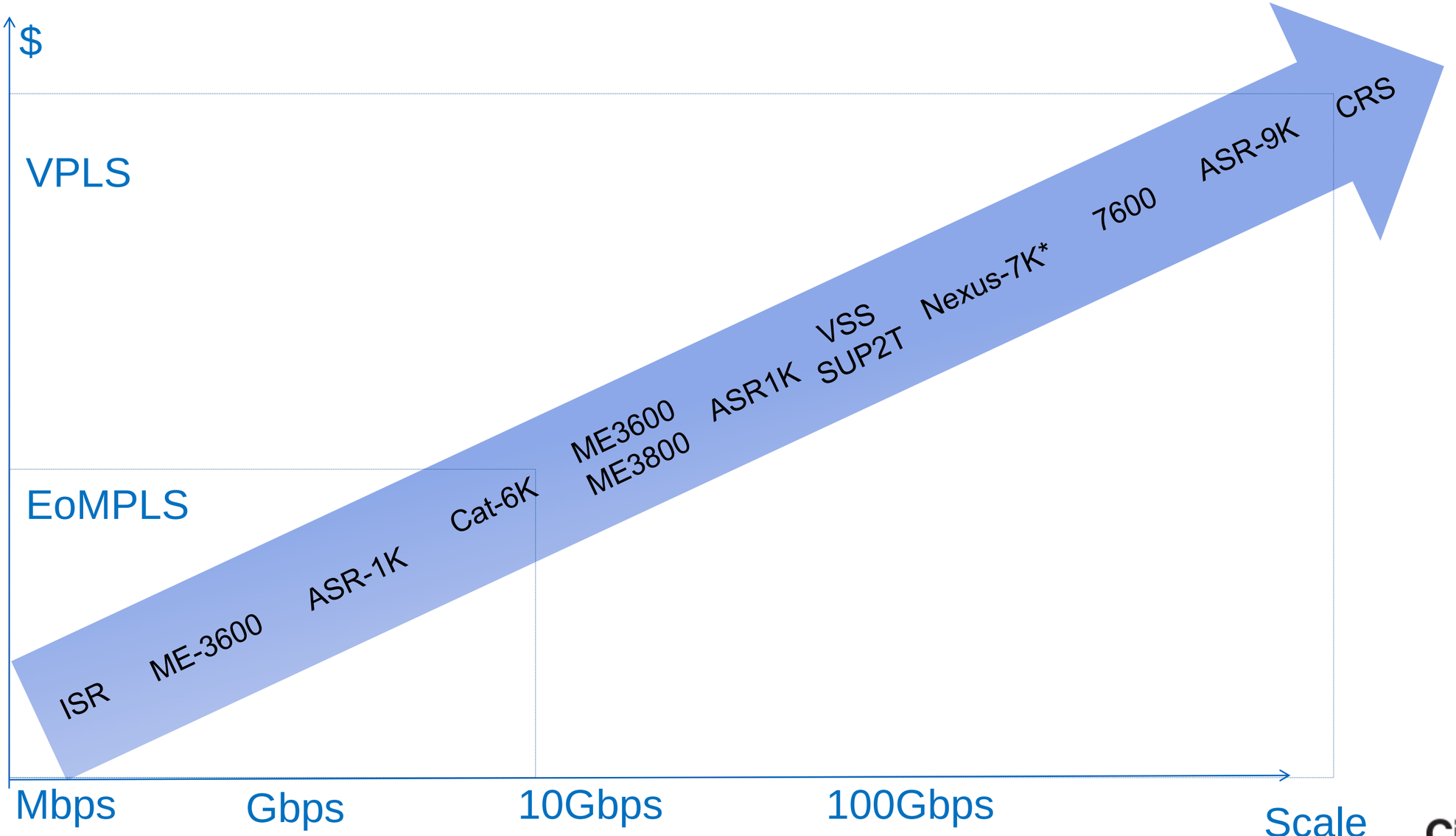
EoMPLS, VPLS

MPLS for DCI
Efficient WAN Technology



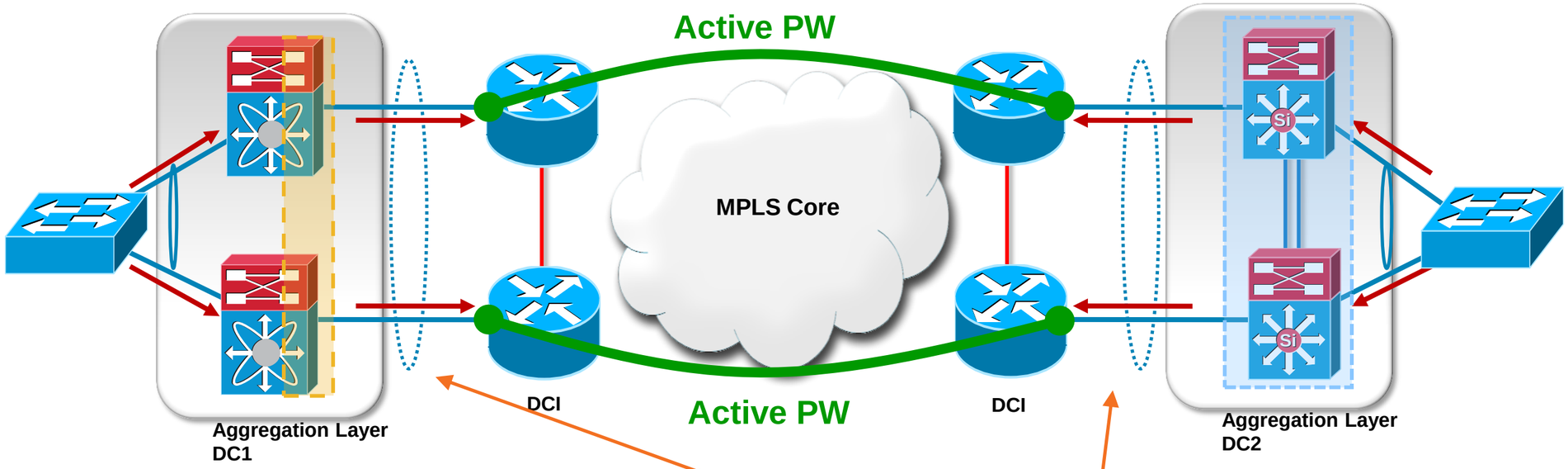
Fast ReRoute for sub-50ms convergence
Traffic-Engineering for SLA control and path diversity

MPLS for DCI
Large Choice of Devices



EoMPLS Usage with DCI

End-to-End Loop Avoidance Using Edge to Edge LACP



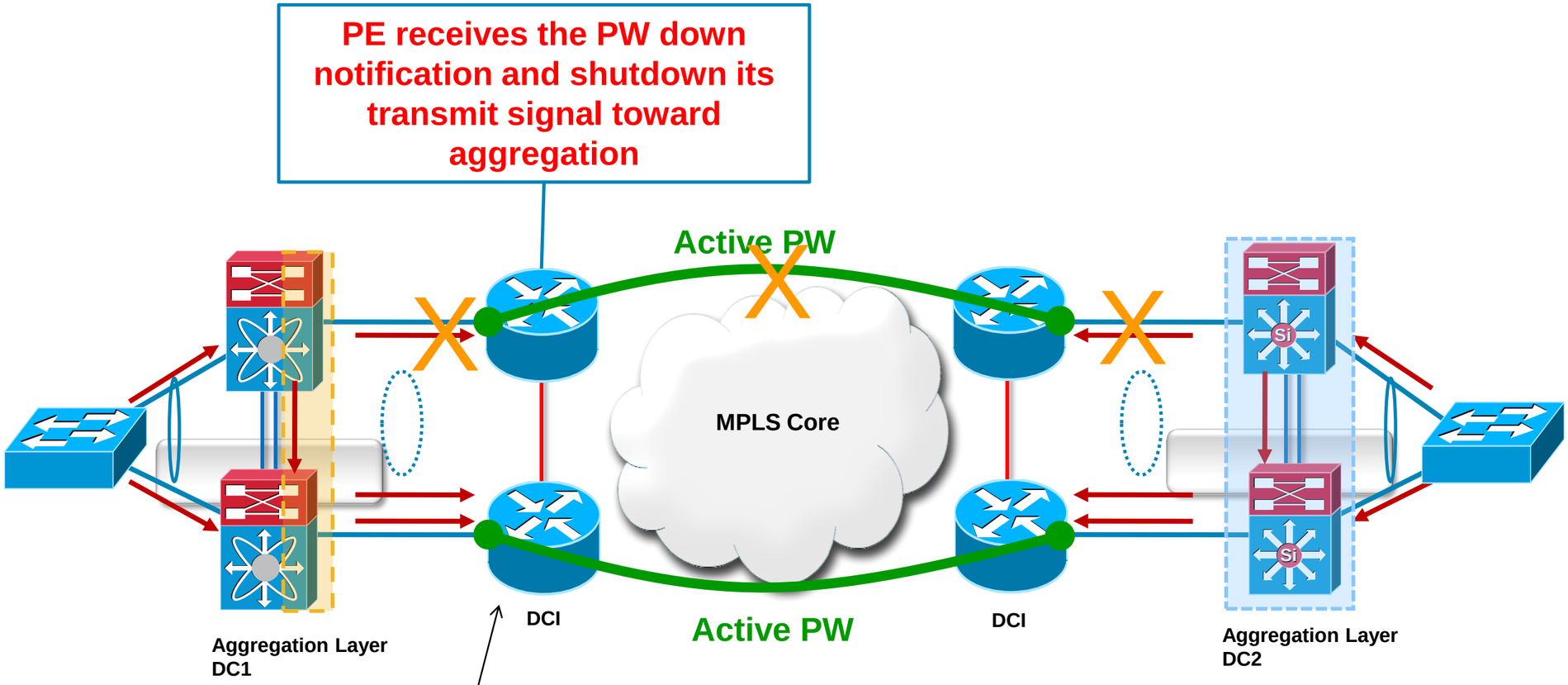
- BPDU Filtering to maintain STP domains isolation
- Storm-control for data-plane protection
- Configuration applied at aggregation layer on the logical port-channel interface

```
interface port-channel70
  description L2 PortChannel to DC 2
  spanning-tree port type edge trunk
  spanning-tree bpdufilter enable
  storm-control broadcast level 1*
  storm-control multicast level x
```

*Value to be tuned, min is 0.3

Dealing with PseudoWire (PW) Failures

Remote Ethernet Port Shutdown



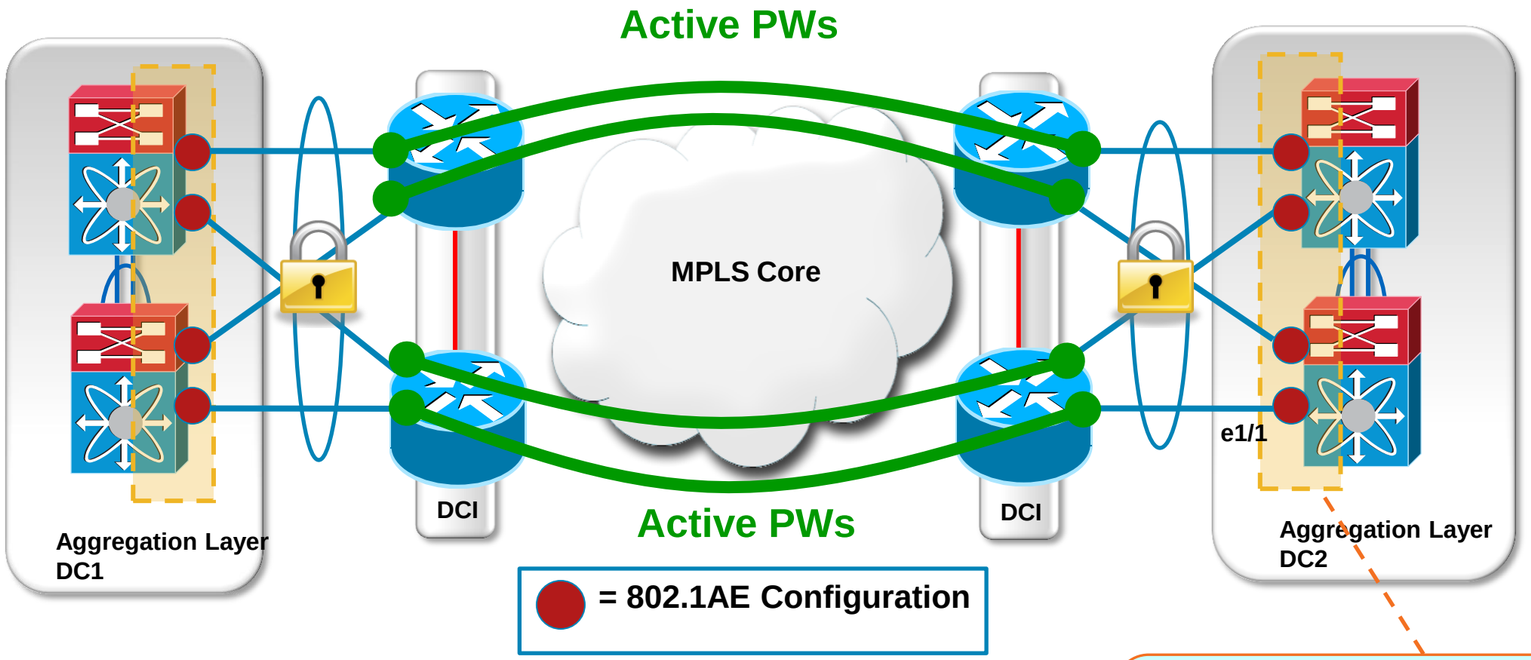
ASR1000 feature configuration:

```
interface GigabitEthernet1/0/0
  xconnect 1.1.1.1 1 pw-class eompls
  remote link failure notification          ! (default)
```

		Failover (msec)	Fallback (msec)
Bridged traffic	→	281	54
	←	453	300

EoMPLS Port Mode

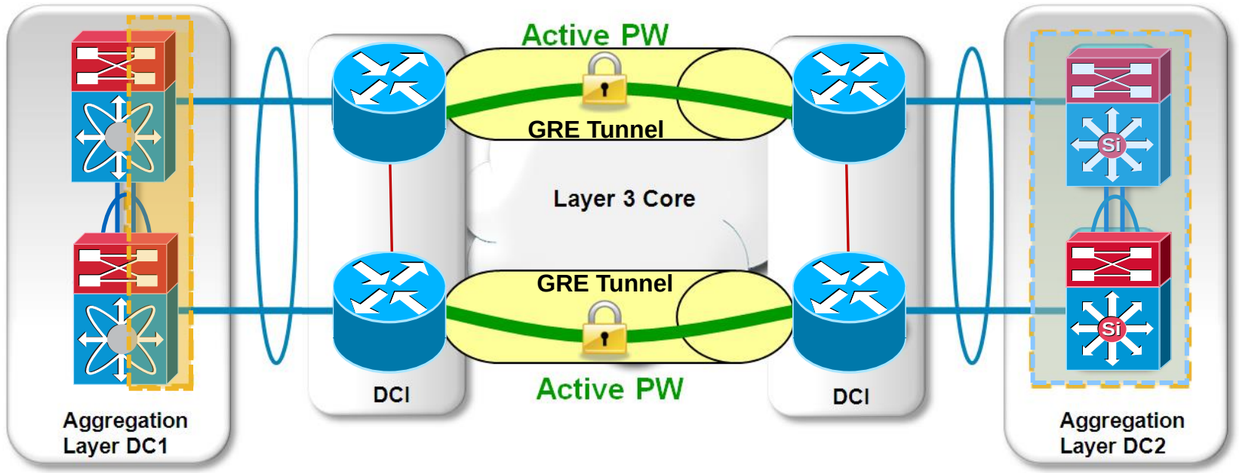
Encryption Services with 802.1AE




```
interface Ethernet1/1
  description PortChannel Link
  cts manual
    no propagate-sgt
    sap pmk 1234000000000000…
```

- “Manual” 802.1AE configuration on a physical interface level
- Traffic encryption end-to-end (intra- and inter-data centre)
- Requires the deployment on Nexus 7000 in the aggregation layer for both sites
- Note the link full-mesh to ensure vPC fast convergence

EoMPLSoGRE
IPSec-Based Encryption Services



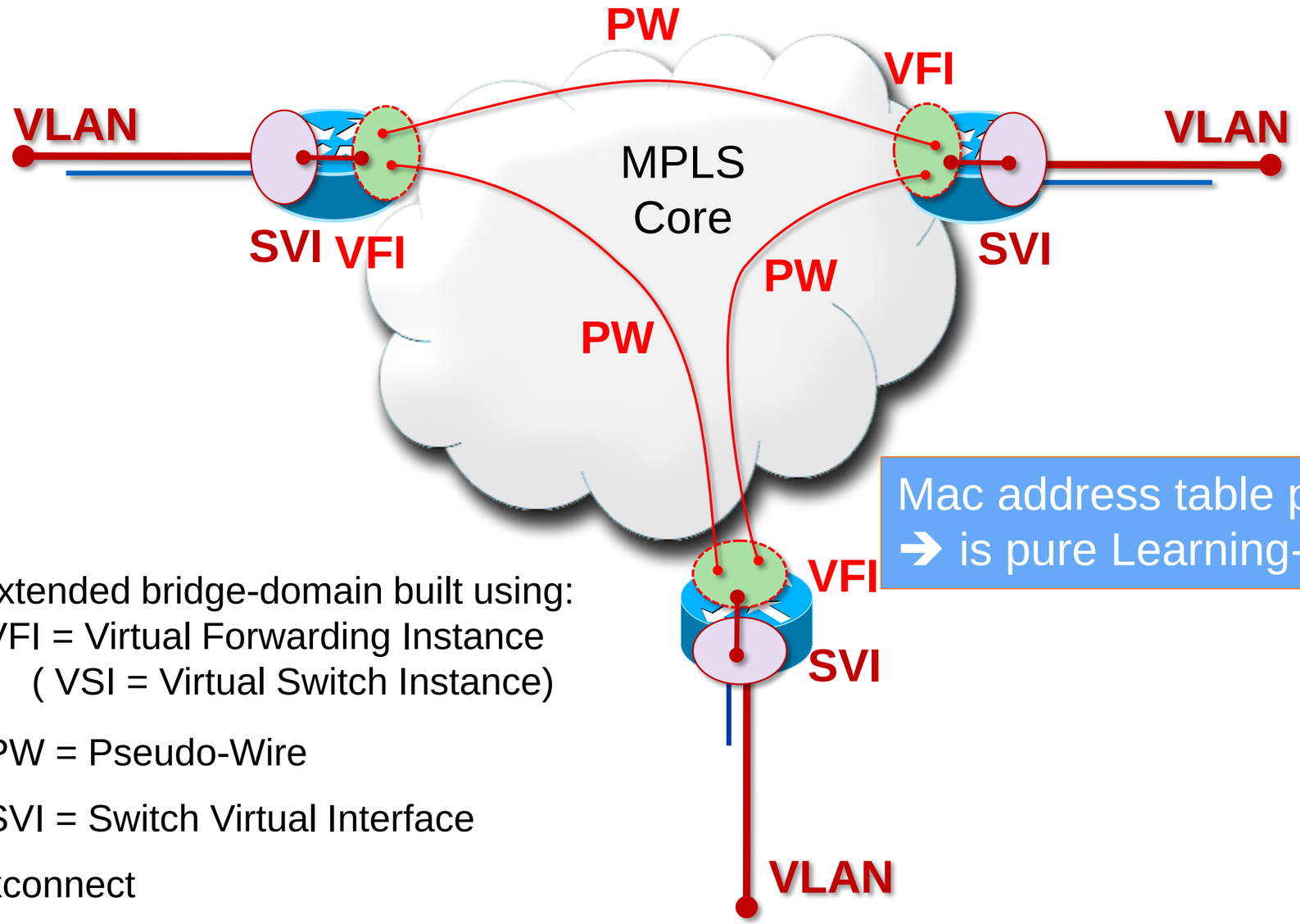
- Native with ASR1000 / ISR
- Requires SIP-400 with Catalyst 6500
  - With loopback cable for crypto
- Tunnel protection is the recommended approach
  - Applied directly to the GRE interface

```
crypto isakmp policy 10
 authentication pre-share
crypto isakmp key CISCO address 0.0.0.0 0.0.0.0
crypto ipsec transform-set MyTransSet esp-3des esp-sha-hmac
crypto ipsec fragmentation after-encryption

crypto ipsec profile MyProfile
 set transform-set MyTransSet

interface Tunnel100
 ip address 100.11.11.11 255.255.255.0
 ip mtu 9216
 mpls ip
 tunnel source Loopback100
 tunnel destination 12.11.11.21
 tunnel protection ipsec profile MyProfile
```

Multi-Point Topologies
What Is VPLS?

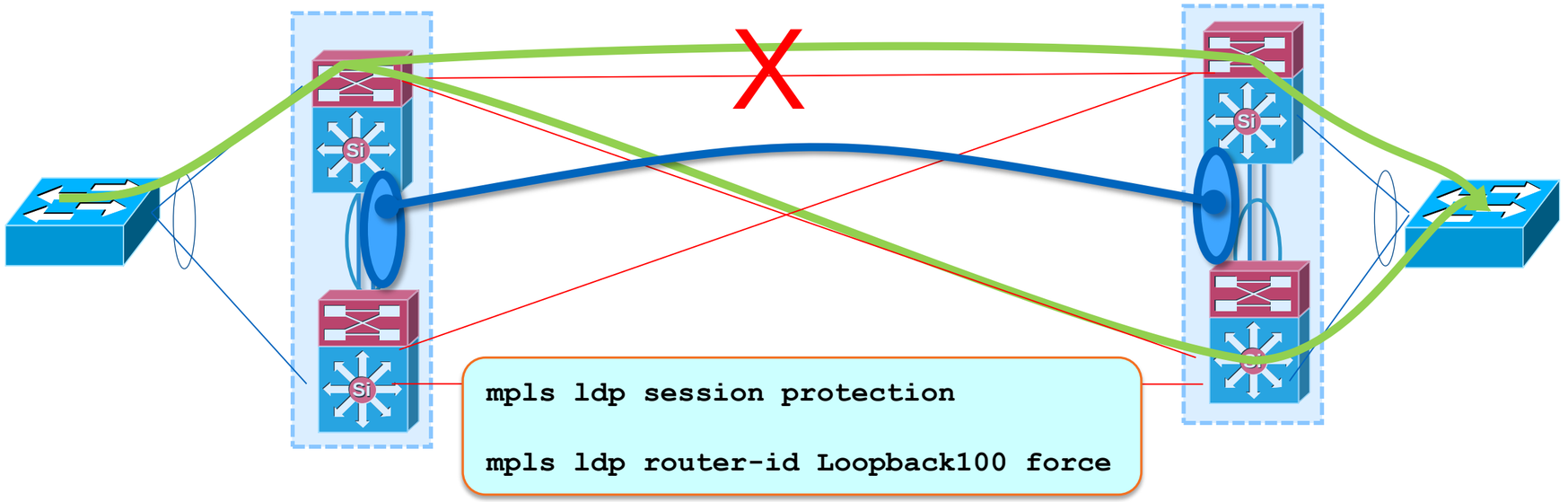


Mac address table population
➔ is pure Learning-Bridge

One extended bridge-domain built using:

- VFI = Virtual Forwarding Instance
  (VSI = Virtual Switch Instance)
- PW = Pseudo-Wire
- SVI = Switch Virtual Interface
- xconnect

Cluster VPLS – Redundancy
Making Usage of Clustering

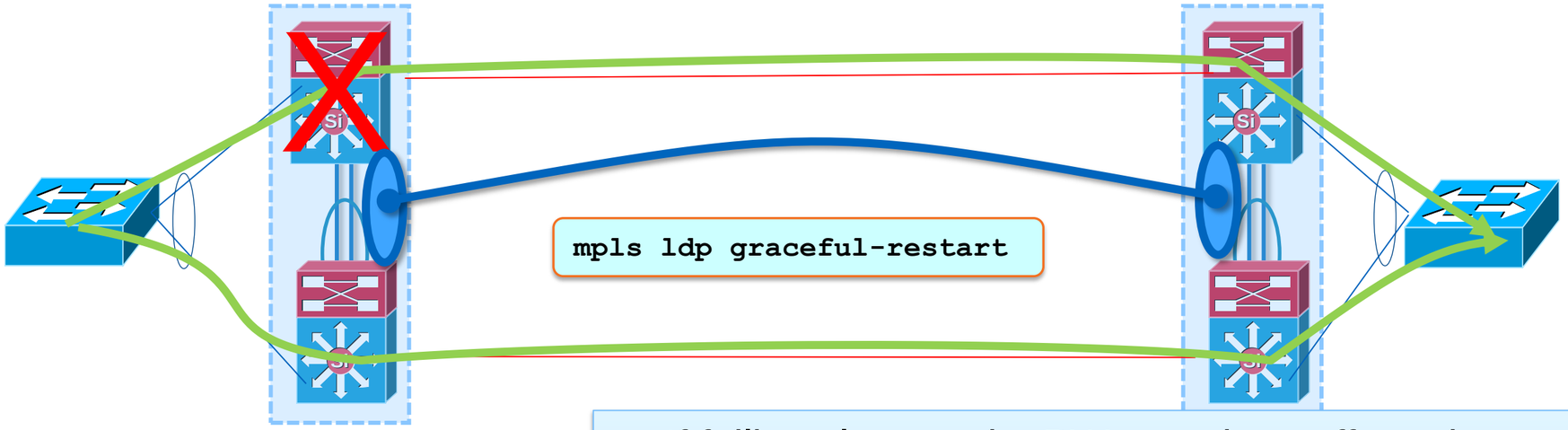


VSS		Failover (msec)	Fallback (msec)
Bridged traffic	➔	258	218
	➔	162	174

- LDP session protection & Loopback usage allows PW state to be unaffected
- LDP + IGP convergence in sub-second
  Fast failure detection on Carrier-delay / BFD
- Immediate local fast protection
  Traffic exit directly from egress VSS node

Cluster VPLS – Redundancy

Making Usage of Clustering

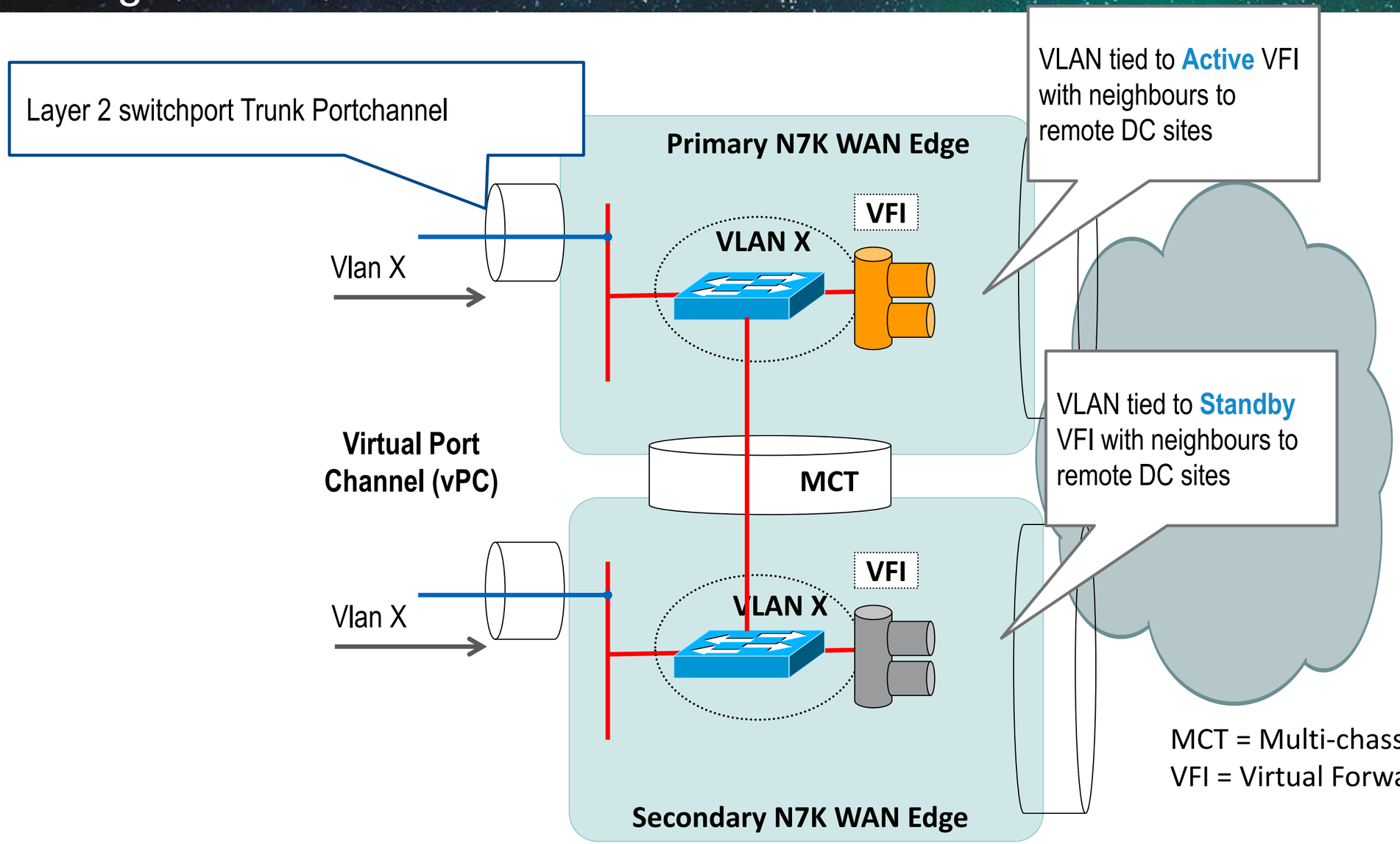


VSS		Failover (msec)	Fallback (msec)
Bridged traffic	➔	224	412
	➔	326	316

- If failing slave node: PW state is unaffected
- If failing master node:
  - PW forwarding is ensured via SSO
  - PW state is maintained on the other side using Graceful restart
- Edge Ether-channel convergence in sub-second
- Traffic is directly going to working VSS node
- Traffic exits directly from egress VSS node
- Quad sup SSO for SUP2T in 1QCY13

Nexus 7000 - Data Centre Interconnect with VPLS

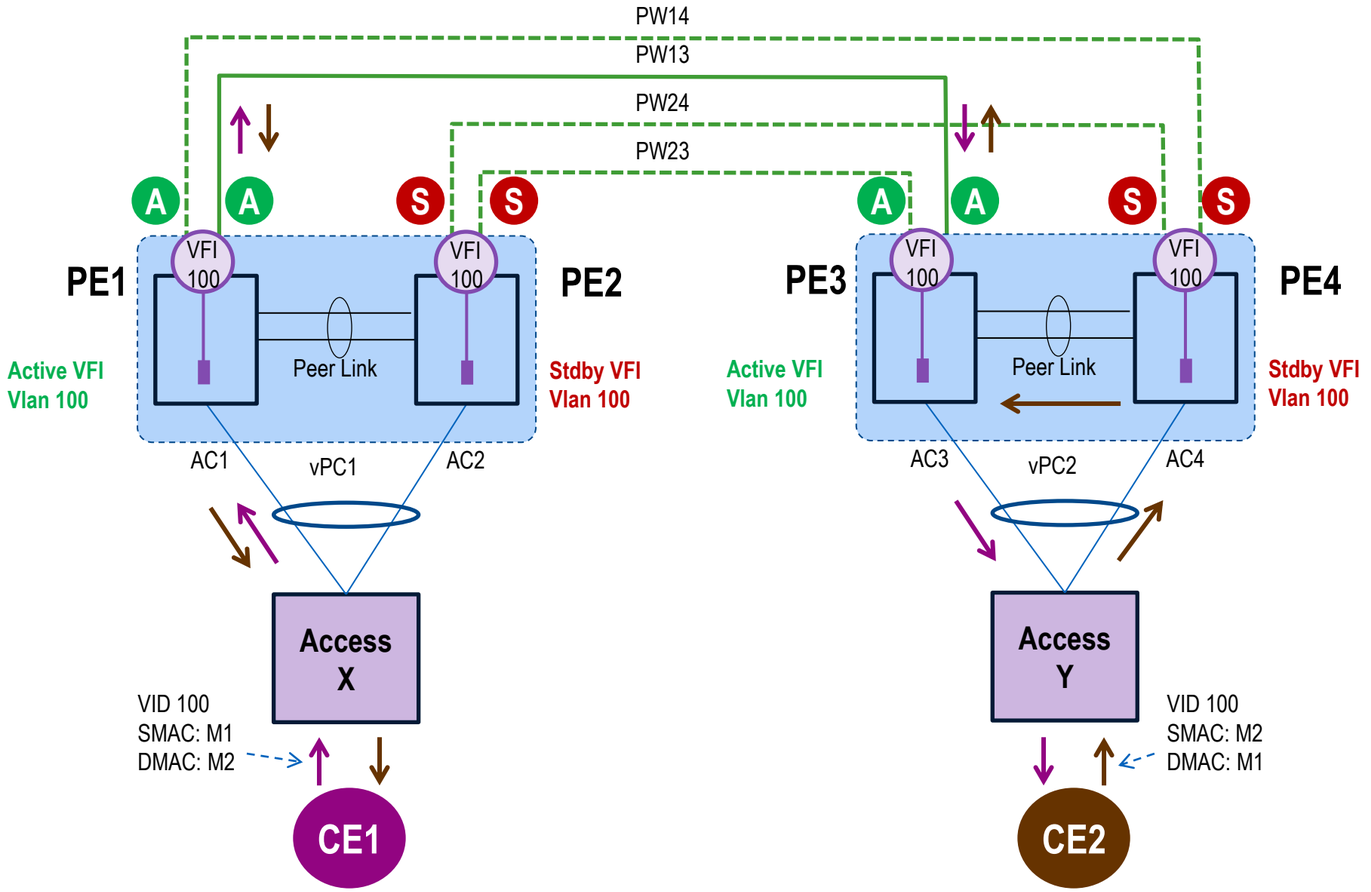
Logical View



MCT = Multi-chassis Trunk Interface
VFI = Virtual Forwarding Instance

Nexus 7000 - Data Centre Interconnect with VPLS

- VPLS multi-homing solution relies on regular vPC procedures / mechanisms as well as extended messaging (ICRM) using CFSoE over Peer link
- vPC validates the VPLS config between vPC Operational Primary and Secondary devices and Slave
- vPC Peer Link used in steady state



PE1

MAC	Port
M1	vPC1 (ac1)
M2	PE3 (pw13)

PE2

MAC	Port
M1	vPC1 (ac2)
M2	PE1(mct)

PE3

MAC	Port
M1	PE1 (pw13)
M2	vPC2 (ac3)

PE4

MAC	Port
M1	PE3 (mct)
M2	vPC2 (ac4)

Nexus 7000 - Data Centre Interconnect with VPLS

Sample Configuration – Nexus 7000

PE 1

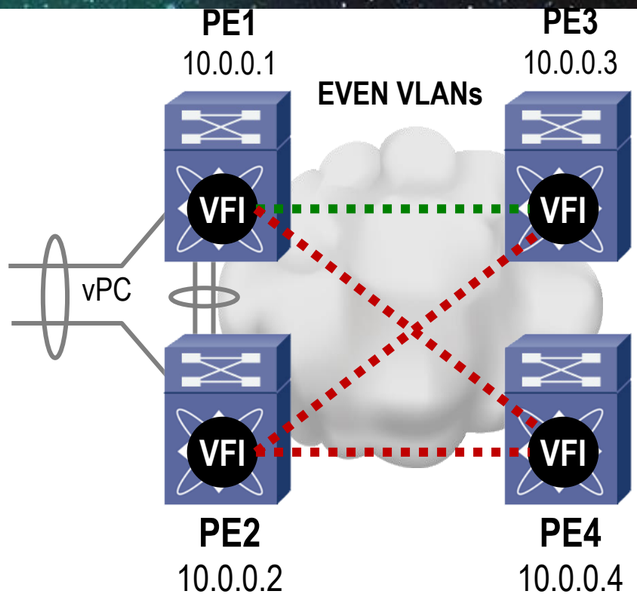
- Primary VFI owner for EVEN vlans
- Secondary owner for ODD vlans

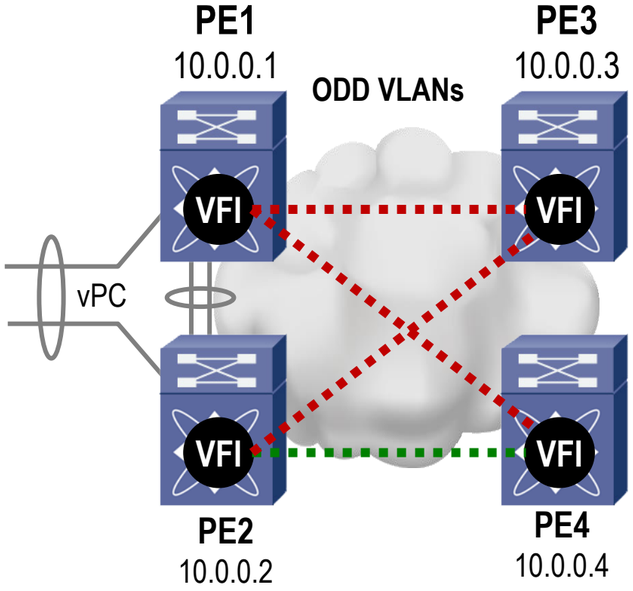
```
vlan 80-81
!
vlan configuration 80
 member vfi vpls-80
!
vlan configuration 81
 member vfi vpls-81
!
l2vpn vfi context vpls-80
  vpn id 80
  redundancy primary
  member 10.0.0.3 encapsulation mpls
  member 10.0.0.4 encapsulation mpls
!
l2vpn vfi context vpls-81
  vpn id 81
  redundancy secondary
  member 10.0.0.3 encapsulation mpls
  member 10.0.0.4 encapsulation mpls
!
interface port-channel50
 switchport mode trunk
 switchport trunk allowed vlan 80,81
```

PE 2

- Primary VFI owner for ODD vlans
- Secondary owner for EVEN vlans

```
vlan 80-81
!
vlan configuration 80
 member vfi vpls-80
!
vlan configuration 81
 member vfi vpls-81
!
l2vpn vfi context vpls-80
  vpn id 80
  redundancy secondary
  member 10.0.0.3 encapsulation mpls
  member 10.0.0.4 encapsulation mpls
!
l2vpn vfi context vpls-81
  vpn id 81
  redundancy primary
  member 10.0.0.3 encapsulation mpls
  member 10.0.0.4 encapsulation mpls
!
interface port-channel50
 switchport mode trunk
 switchport trunk allowed vlan 80,81
```





Note: Virtual Port Channel (vPC) configuration not shown




Overlay Transport Virtualisation (OTV)

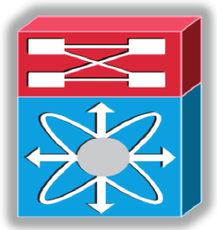
IP Based Solutions

Overlay Transport Virtualisation
Technology Pillars

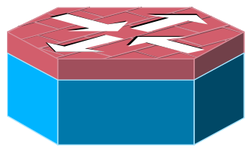
OTV is a “MAC in IP” technique to extend Layer 2 domains
OVER ANY TRANSPORT

Dynamic Encapsulation

- No Pseudo-Wire State Maintenance
- Optimal Multicast Replication
- Multipoint Connectivity
- Point-to-Cloud Model



Nexus 7000
First platform to support OTV (since 5.0 NXOS Release)



ASR 1000
Now also supporting OTV (since 3.5 XE Release)

Protocol Learning

- Preserve Failure Boundary
- Built-in Loop Prevention
- Automated Multi-homing
- Site Independence

OTV Enhancements on ASR 1000

Internal Interfaces - Deployment Guidelines

ASR 1000

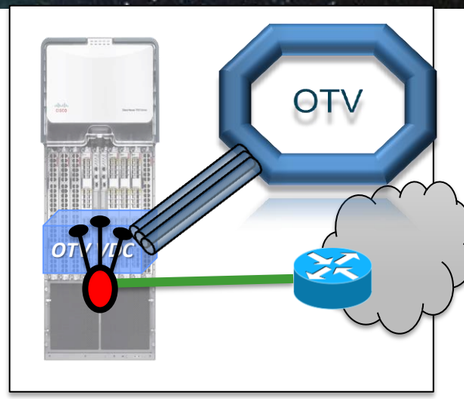
- 1GE and TenGE Ethernet interfaces are supported
- Adjacency Server support (XE 3.9 Mar 2013)
- Multiple Internal Interfaces running Spanning tree and extending the same set of VLANs are not supported
- RPVST Support
- Improvement (XE 3.10 Jul 2013)
  - Port channel interfaces for Join Interface
  - VRF Aware
  - Sub-interfaces as Join interface
  - Layer 2 Port channel

New Features for OTV @ Nexus 7000

New Release 6.2

Tunnel Depolarisation & Secondary IP

- Secondary IP command introduced
  - Configured within interface, not OTV interface
- Introduction of multiple IPs results in tunnel depolarisation




```
OTV-a(config-if)# ip address 2.100.11.1/24 secondary
Disabling IP Redirects on port-channel11 :secondary address
configured.
OTV-a(config-if)# sh run int po11

!Command: show running-config interface port-channel11
!Time: Wed Mar 27 23:05:21 2013

version 6.2(2)

interface port-channel11
  no ip redirects
  ip address 2.100.11.100/24
  ip address 2.100.11.1/24 secondary
  ip ospf network point-to-point
  ip router ospf 1 area 0.0.0.0
  ip igmp version 3
```

```
OTV-a (config-if)# sh otv

OTV Overlay Information
Site Identifier 0000.0000.0011

Overlay interface Overlay1

 VPN name              : Overlay1
 VPN state             : UP
 Extended vlans        : 25-50 72-227 (Total:182)
 Control group         : 224.1.1.0
 Data group range(s)   : 232.1.0.0/24
 Broadcast group       : 224.1.1.0
 Join interface(s)     : Po11 (2.100.11.100)
    Secondary IP Addresses: : 2.100.11.1
 Site vlan             : 1 (up)
 AED-Capable           : Yes1
 Capability            : Multicast-Reachable
```

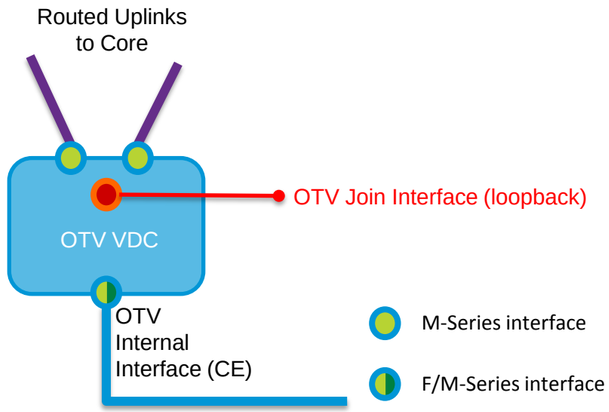
New Feature for OTV @ Nexus 7000

Source Interface & Loopback Addresses

Release 6.2 Maint.

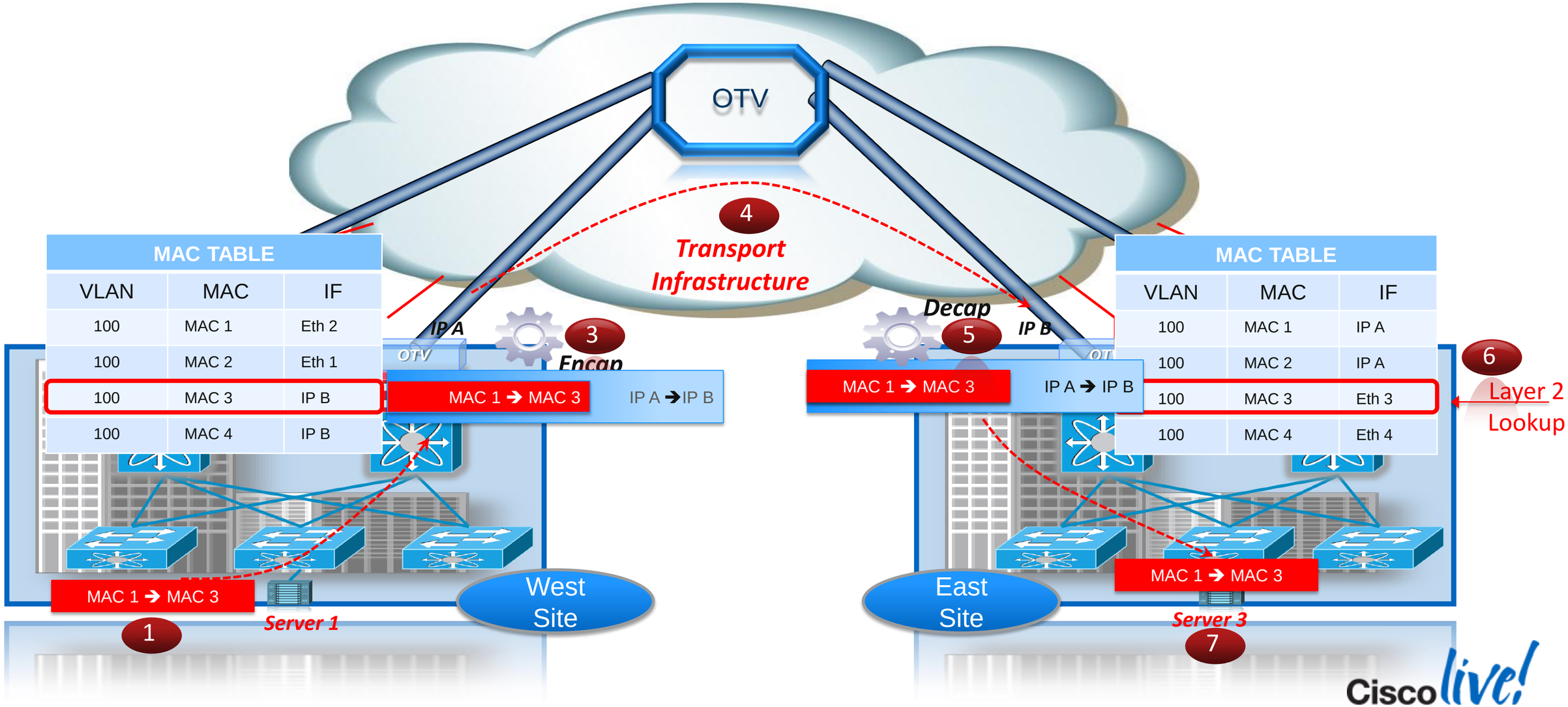
- New command introduced
  - "otv source-interface <interface>"
- Source-interface should be a loopback to guarantee interface is up/up
- One source-interface configurable per overlay
  - Can be shared with multiple overlays
- N7K will now join as PIM router
- All available uplinks to core can be used to reach destination DC
  - Hash will be calculated with the source-interface as the source address
- Source-interface will take priority over join-interface if both are configured
- Need to configure "ip pim sparse-mode" on core L3 interfaces

```
interface Overlay3
  otv source-interface loopback11
  otv control-group 224.3.3.0
  otv data-group 232.3.0.0/24
  otv extend-vlan 198-227
  no shutdown
```



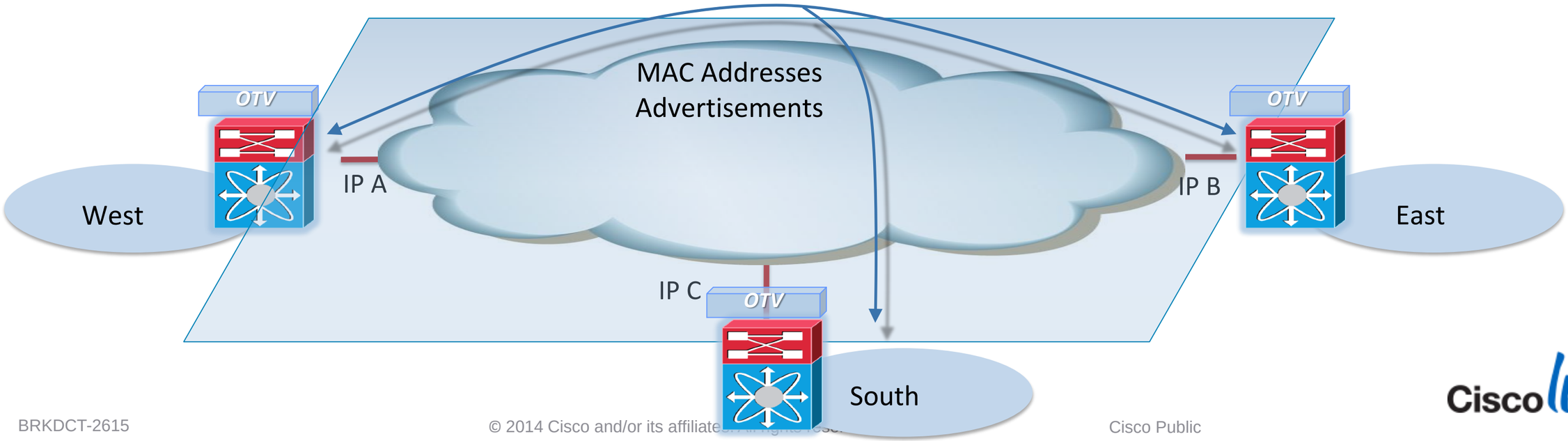
Logical Join Interface

OTV Data Plane
Inter-Sites Packet Flow



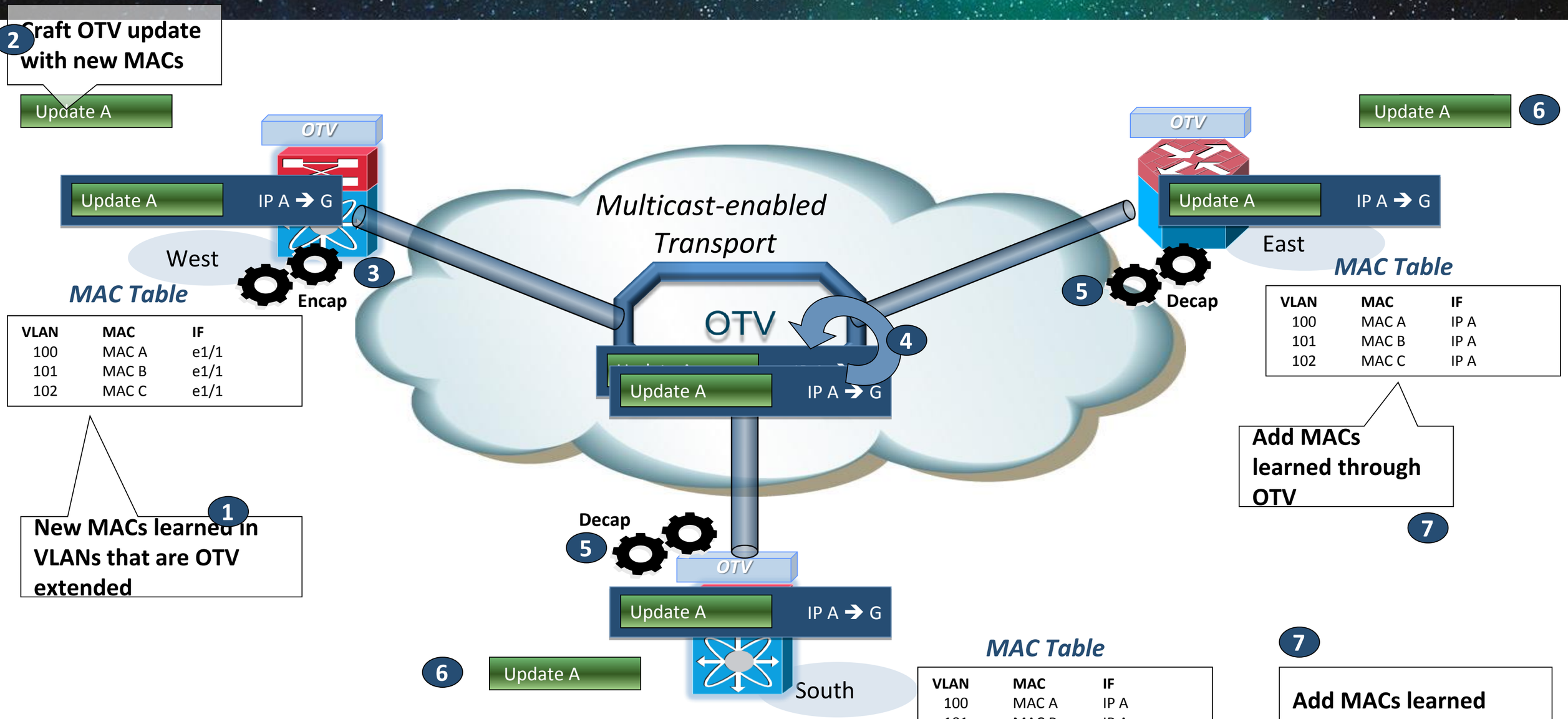
OTV Control Plane
Building the MAC Tables

- **No Unknown Unicast flooding by default**
  - **Selective Unknown Unicast flooding with 6.2**
- **Control Plane Learning with proactive MAC advertisement**
- Background process with no specific configuration
- IS-IS used between OTV Edge Devices



OTV Control Plane

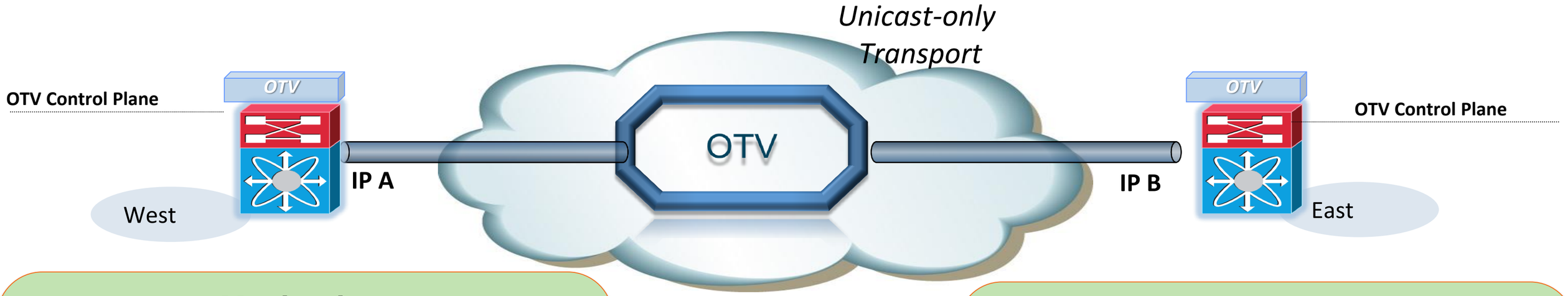
Route (MAC) Advertisements (over Multicast Transport)



OTV Control Plane
Neighbour Discovery (Unicast-Only Transport)

Release 5.2 and above

- Ideal for connecting a small number of sites
- With a higher number of sites a multicast transport is the best choice



Mechanism

- Edge Devices (EDs) register with an *"Adjacency Server"* ED
- EDs receive a full list of Neighbours (oNL) from the *AS*
- OTV hellos and updates are encapsulated in IP and ***unicast*** to each neighbour

End Result

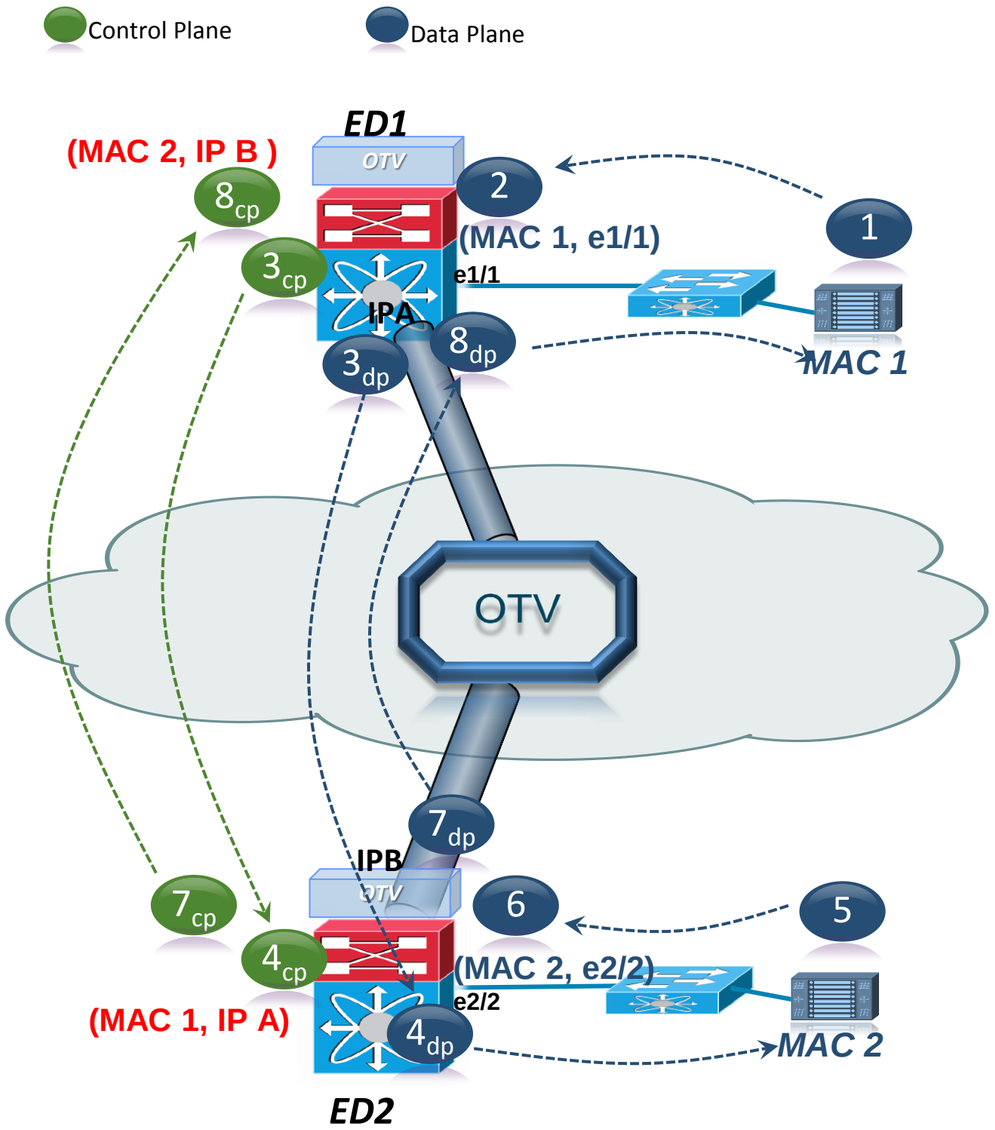
Neighbour Discovery is automated by the "*Adjacency Server*"

All signalling must be replicated for each neighbour

Data traffic must also be replicated at the head-end

OTV Packet Walk

Establishing Inter-Site Unicast Communication



1 – Broadcast ARP for MAC 2

2 – Broadcast ARP received by ED1. MAC 1 is learnt on its internal interface

3_{cp} – ED1 advertises MAC 1 in an OTV Update sent to the other EDs part of the Overlay

4_{cp} – ED2 receives the update and stores MAC1 in MAC table, next-hop is ED1

3_{dp} – ED1 forward the broadcast frame to the Overlay. All EDs receive it

4_{dp} – ED2 decapsulates the frame and forwards the ARP broadcast request into the site

5 – Server 2 receives the ARP and replies with a unicast ARP to MAC 1

6 – ED2 learns MAC 2 on its internal interface

7_{cp} – ED2 advertises MAC 2 with an update sent to the other EDs

8_{cp} – ED1 receives the update and stores MAC2 in MAC table, next-hop is ED2

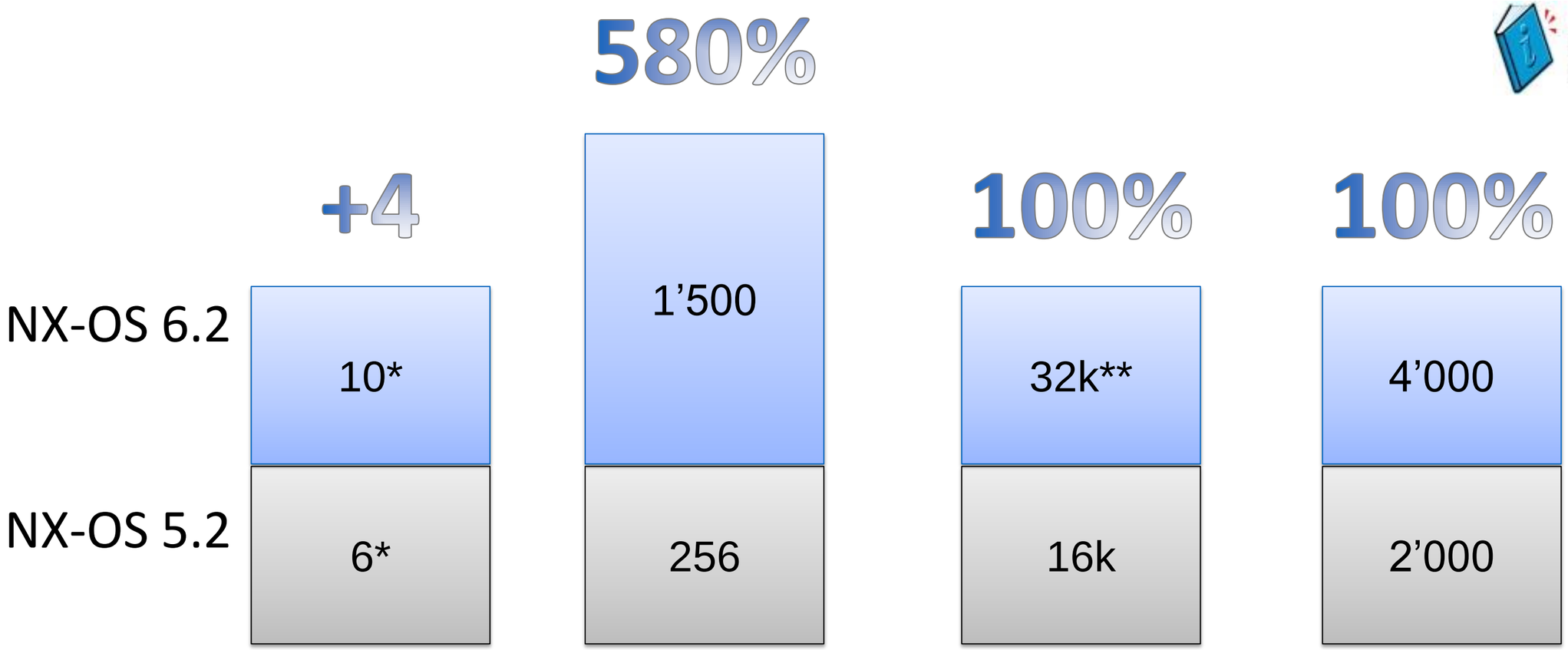
7_{dp} – ED2 knows that MAC 1 is reachable via IP A. It encapsulates the packet (IP A is dest IP) and sends it unicast to ED1

8_{dp} – Core delivers packet to ED1. ED1 decapsulates and forwards it into the site to MAC 1

OTV Scalability
Current Supported Values – Nexus 7000

Release 6.2

For Your Reference



** two ED per Site*
*** with Supervisor 2E*

DCI Convergence Summary

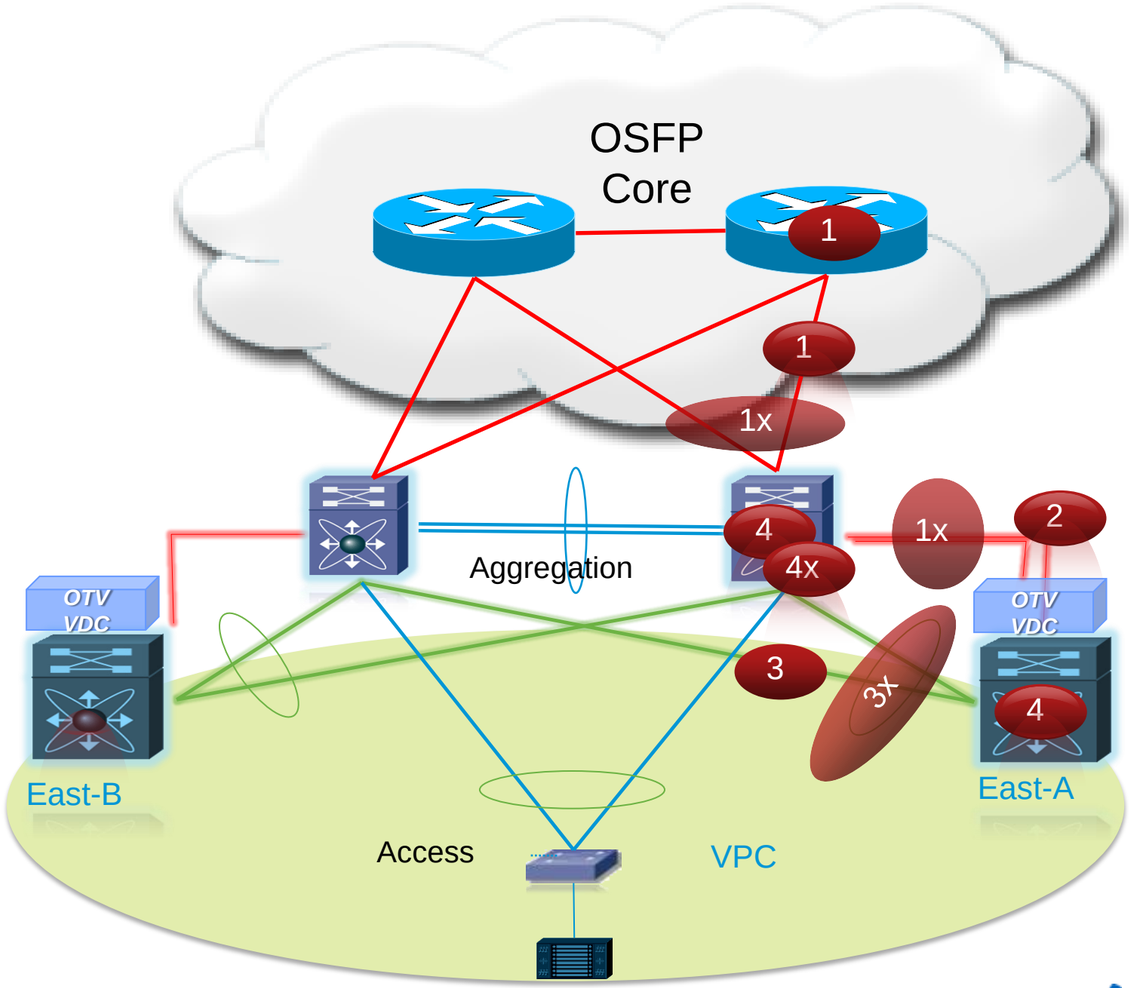
Robust HA is the guiding principle

Common Failures:

1. Core failures
   Multipath routing (or TE FRR) ➔ **sub-sec** ✓
2. Join interface failures
   Link Aggregates across line-cards ➔ **sub-sec** ✓
3. Internal Interfaces failures
   Multipath topology (vPC) & LAGs ➔ **sub-sec** ✓
4. ED component failures
   HW/SW resiliency ➔ **sub-sec** ✓

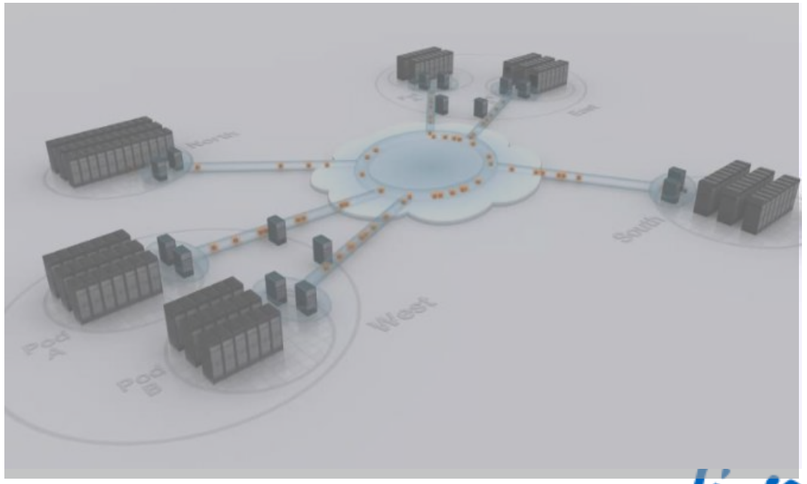
Extreme failures (unlikely):

1x. Core partition
3x. Site partition
4x. Device down
   Require OTV reconvergence
   < 10s (5.2(1))
   6.2 ➔ < 5s ✓



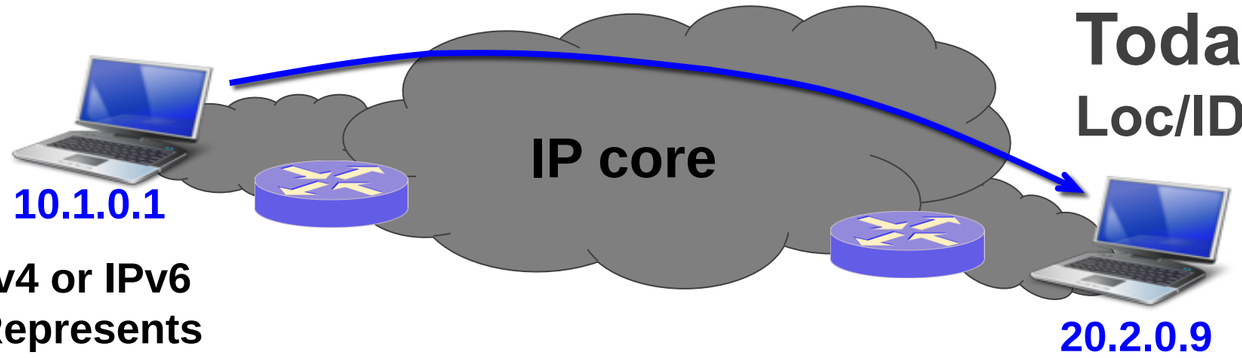
Agenda

- Active-Active Data Centre: Business Drivers and Solutions Overview
- Active / Active Data Centre Design Considerations
  - Storage Extension
  - Data Centre Interconnect (DCI) - LAN Extension Deployment Scenarios
  - Host Mobility using LISP and OTV
  - Network Services and Applications (Path optimisation)
- Cisco ACI and Active / Active Data Centre
- Summary and Conclusions
- Q&A



Location Identity Separation Protocol (LISP)

- What do we mean by “Location” and “Identity”

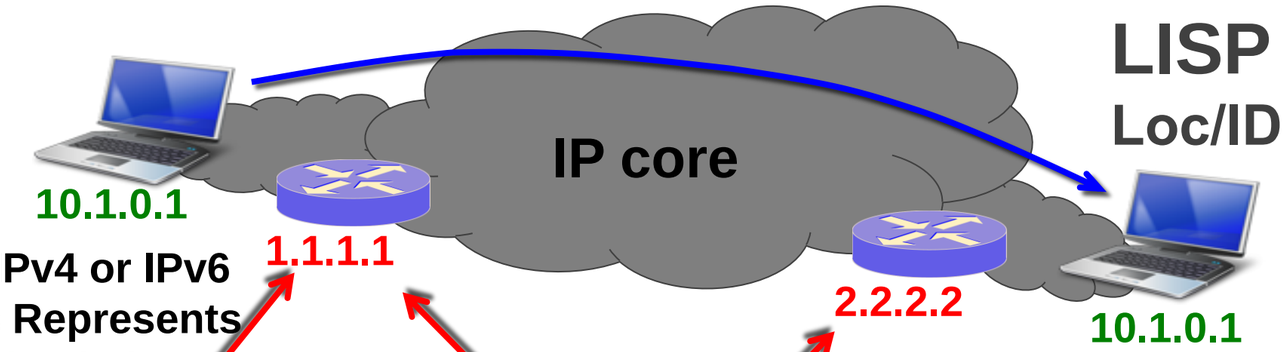


Today’s IP Behaviour

Loc/ID “Overloaded” Semantic

Device IPv4 or IPv6 Address Represents Identity and Location

When the Device Moves, It Gets a New IPv4 or IPv6 Address for Its New Identity and Location



LISP Behaviour

Loc/ID “Split”

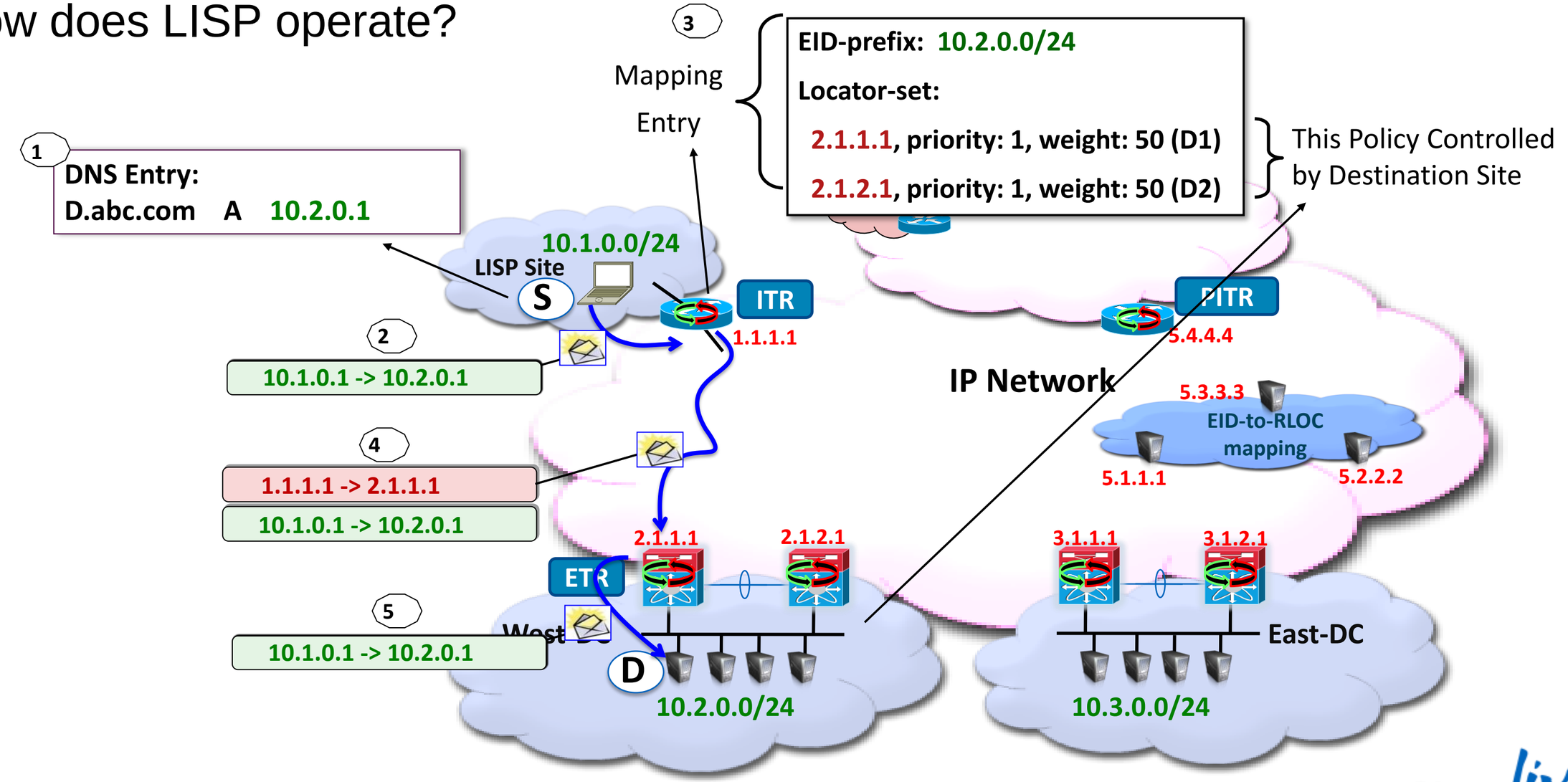
Device IPv4 or IPv6 Address Represents Identity Only.
Its Location Is Here!

When the Device Moves, Keeps Its IPv4 or IPv6 Address. It Has the Same Identity

Only the Location Changes

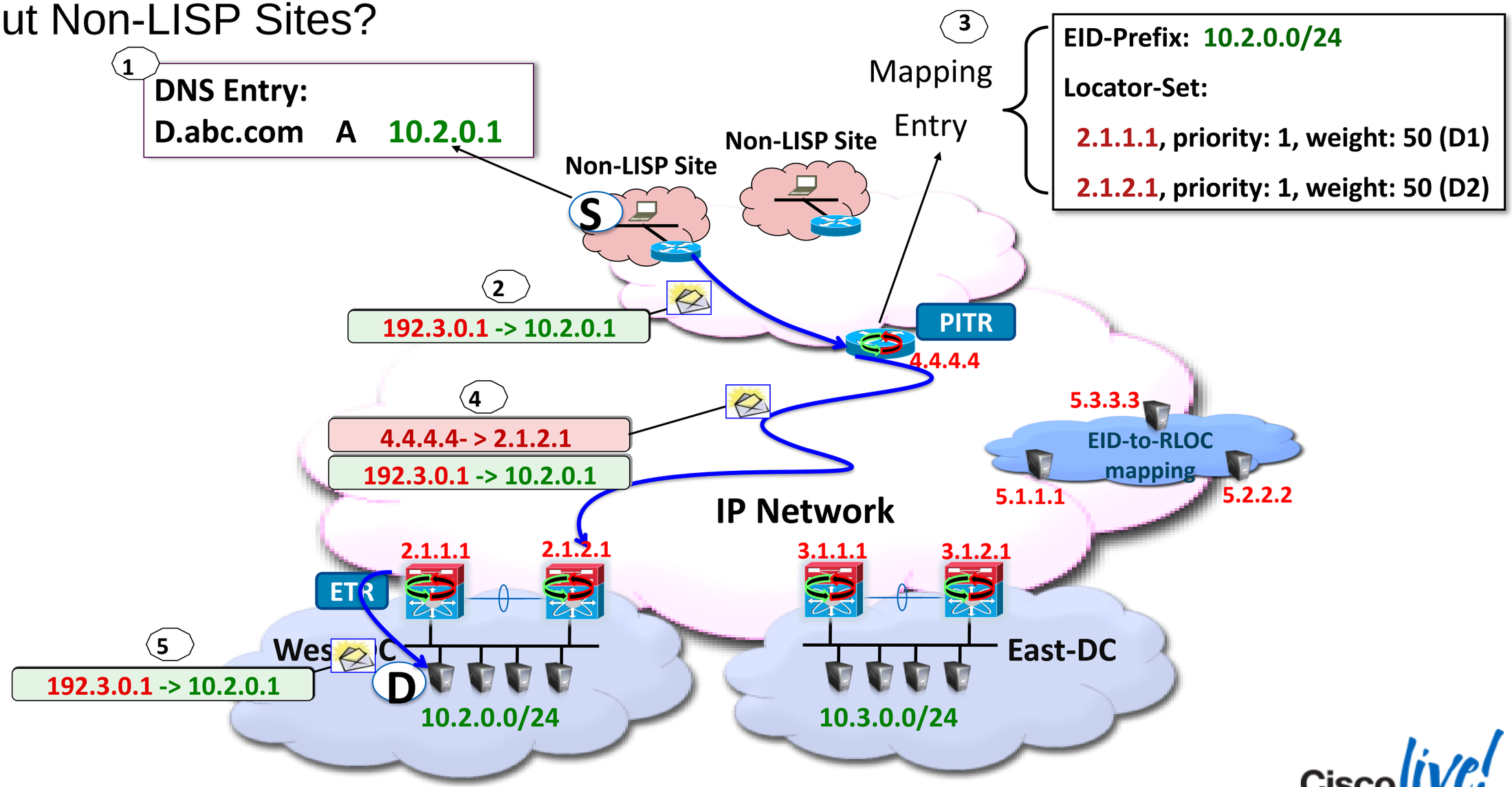
A LISP Packet Walk

- How does LISP operate?



A LISP Packet Walk

- How about Non-LISP Sites?

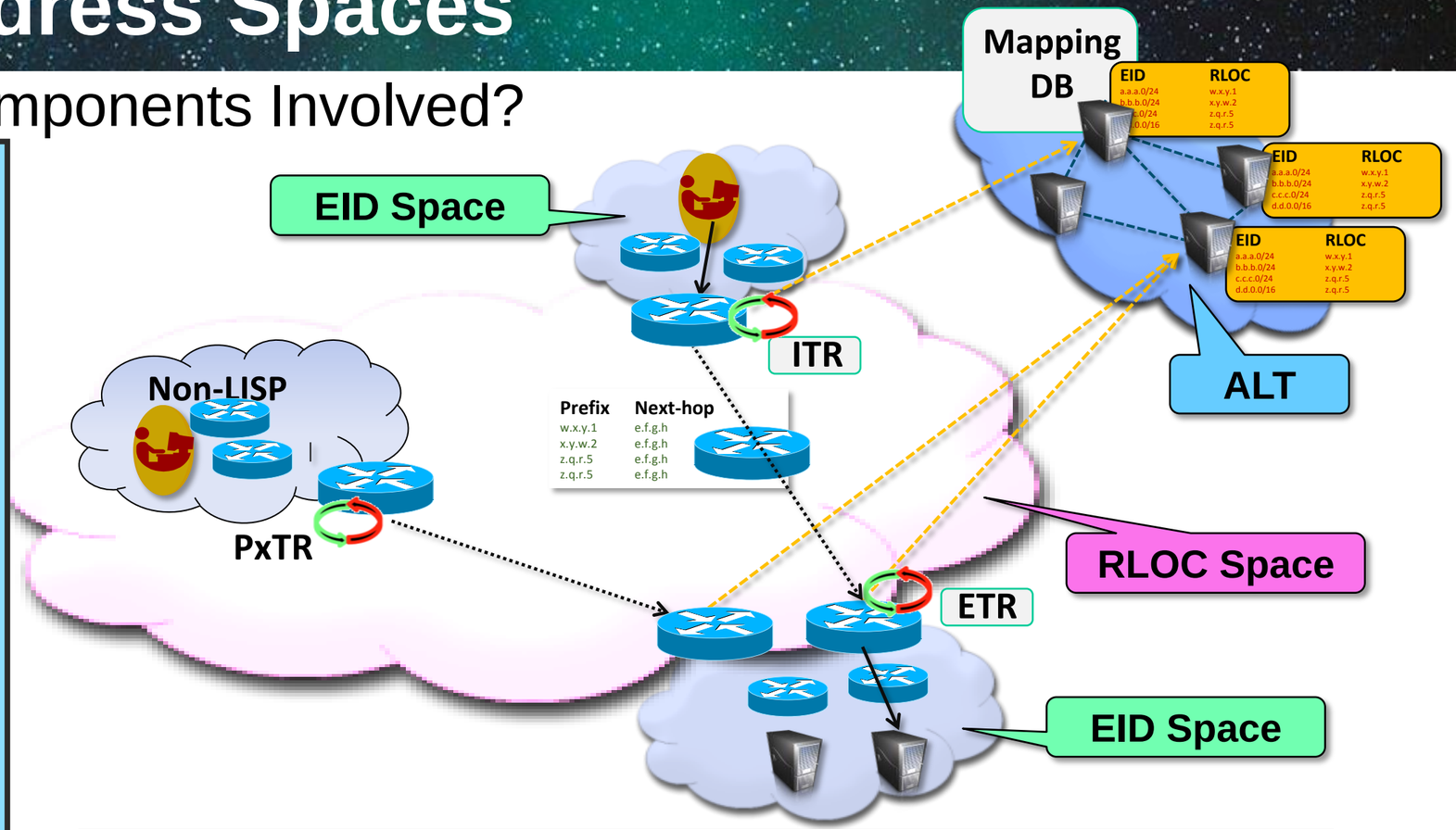


LISP Roles and Address Spaces

- What are the Different Components Involved?

LISP Roles

- **Tunnel Routers - xTRs**
  - Edge devices encap/decap
  - Ingress/Egress Tunnel Routers (ITR/ETR)
- **Proxy Tunnel Routers - PxTR**
  - Coexistence between LISP and non-LISP sites
  - Ingress/Egress: PITR, PETR
- **EID to RLOC Mapping DB**
  - RLOC to EID mappings
  - Distributed across multiple Map Servers (MS)

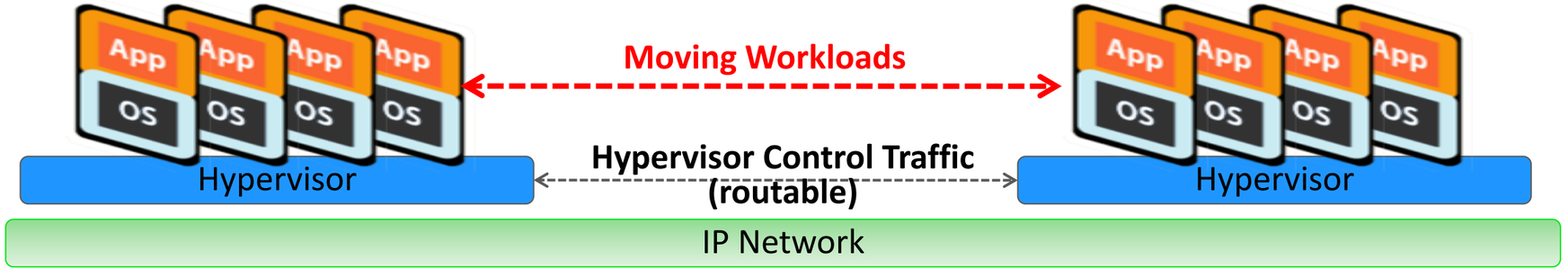


Address Spaces

- **EID = End-point Identifier**
  - Host IP or prefix
- **RLOC = Routing Locator**
  - IP address of routers in the backbone

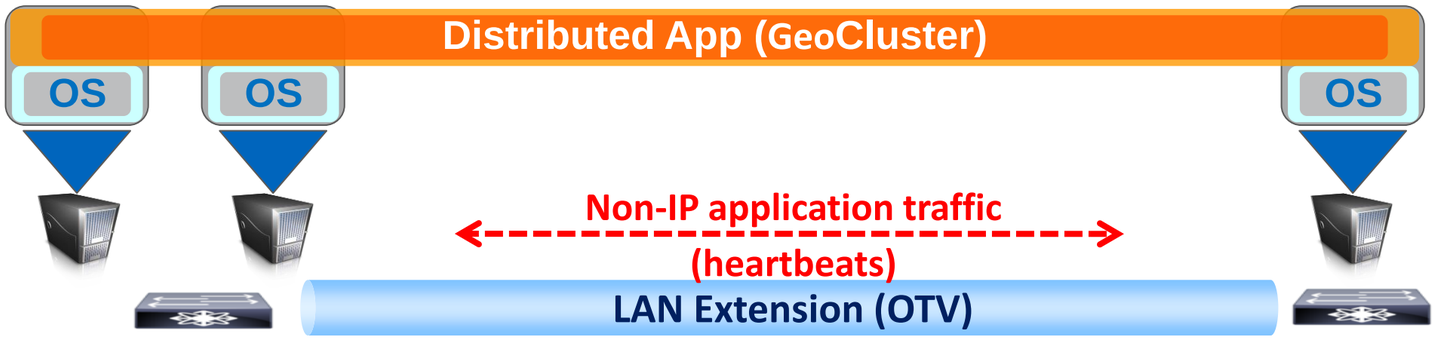
Moving vs. Distributing Workloads

- Why do we really need LAN Extensions?



  - **Move workloads** with IP mobility solutions: LISP Host Mobility
    - IP preservation is the real requirement (LAN extensions not mandatory)

  - **Distribute workloads** with LAN extensions
    - Application High Availability with Distributed Clusters



LISP Host-Mobility

Needs:

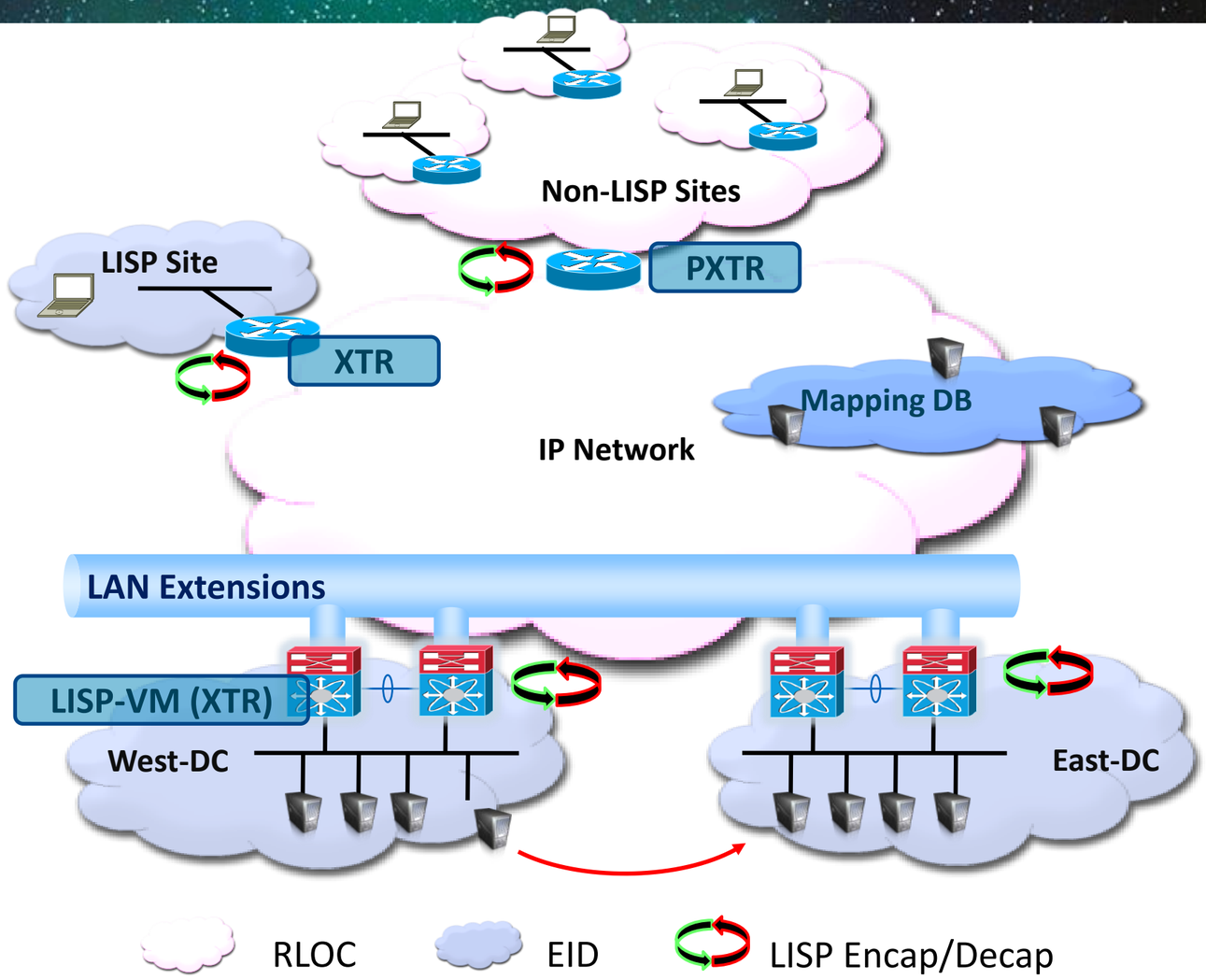
- Global IP-Mobility **across subnets**
- Optimised routing **across extended subnet sites**

LISP Solution:

- Automated **move detection** on XTRs
- Dynamically update EID-to-RLOC mappings
- **Traffic Redirection** on ITRs or PITRs

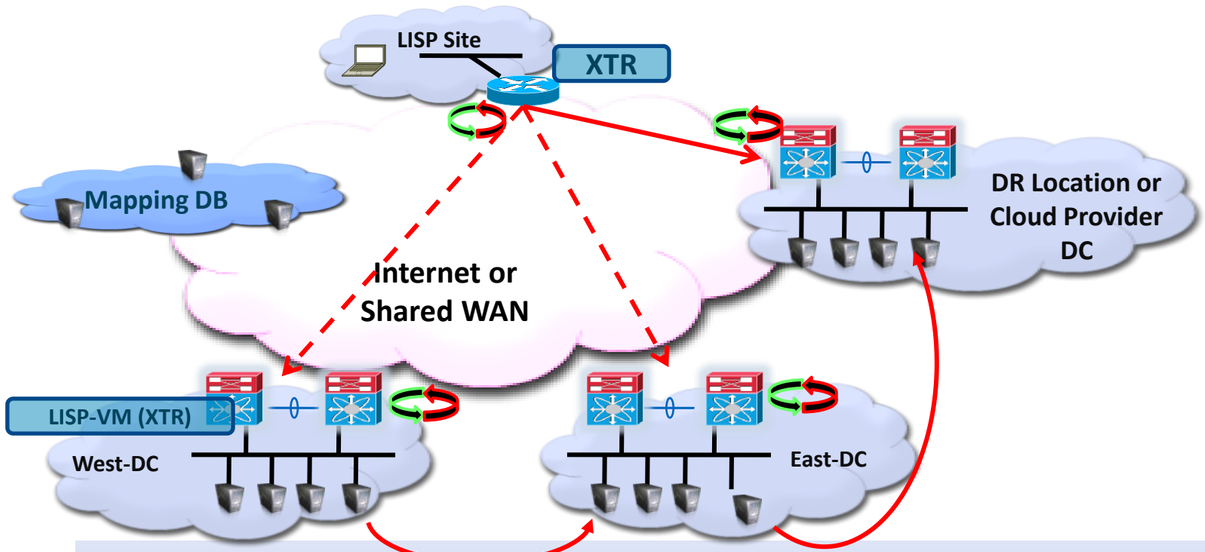
Benefits:

- Direct Path (no triangulation)
- Connections maintained across move
- No routing re-convergence
- No DNS updates required
- Transparent to the hosts
- Global Scalability (cloud bursting)
- IPv4/IPv6 Support



Host-Mobility Scenarios

Moves Without LAN Extension



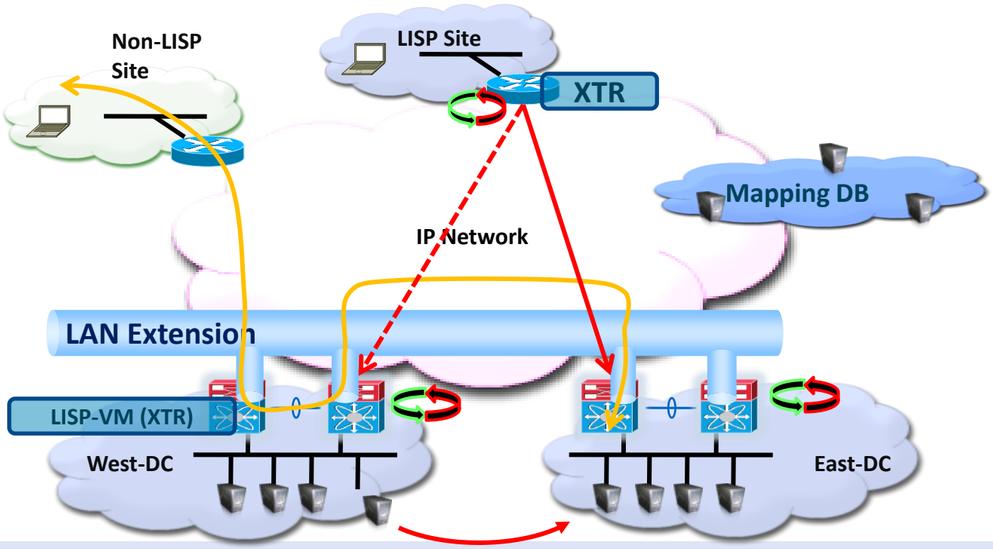
IP Mobility Across Subnets

Disaster Recovery

Cloud Bursting

Application Members in One Location

Moves With LAN Extension



Routing for Extended Subnets

Active-Active Data Centres

Distributed Clusters

Application Members Distributed (Broadcasts across sites)

LISP Host-Mobility – Move Detection

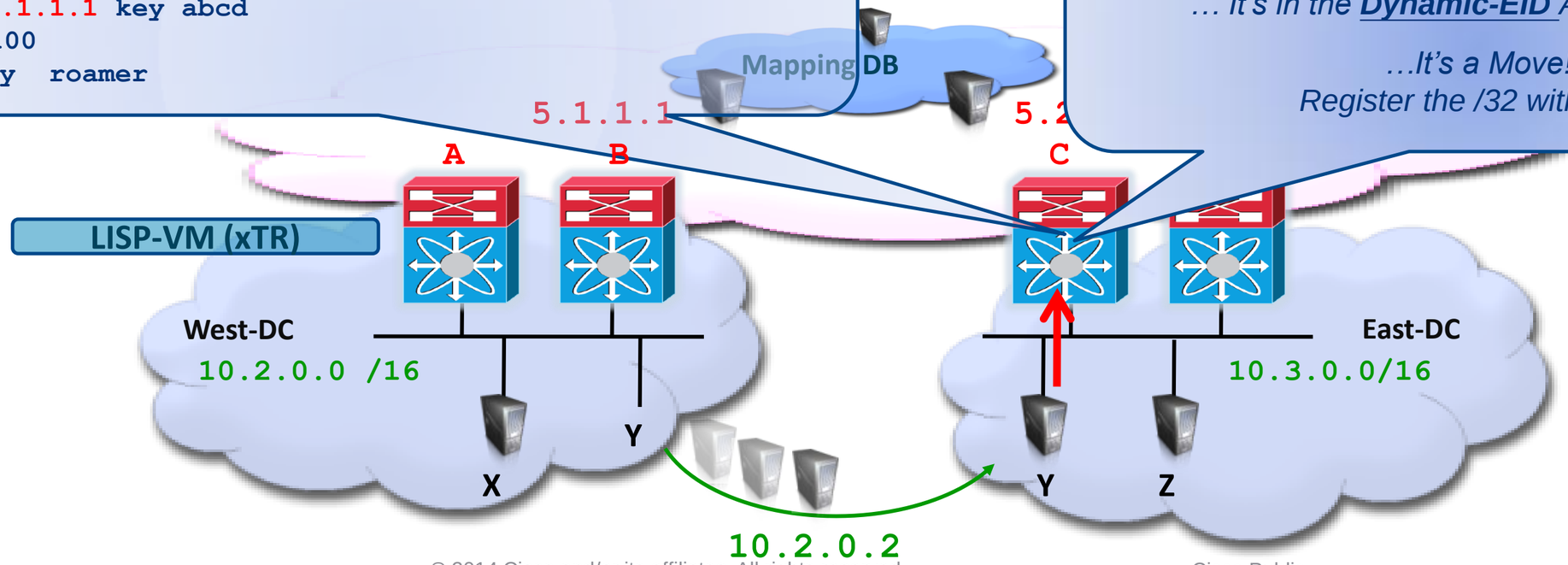
- Monitor the source of Received Traffic
- The new xTR checks the source of received traffic
- Configured dynamic-EIDs define which prefixes may roam

```
lisp dynamic-eid roamer
    database-mapping 10.2.0.0/24 <RLOC-C> p1 w50
    database-mapping 10.2.0.0/24 <RLOC-D> p1 w50
    map-server 5.1.1.1 key abcd
interface vlan 100
    lisp mobility  roamer
```

Received a Packet …

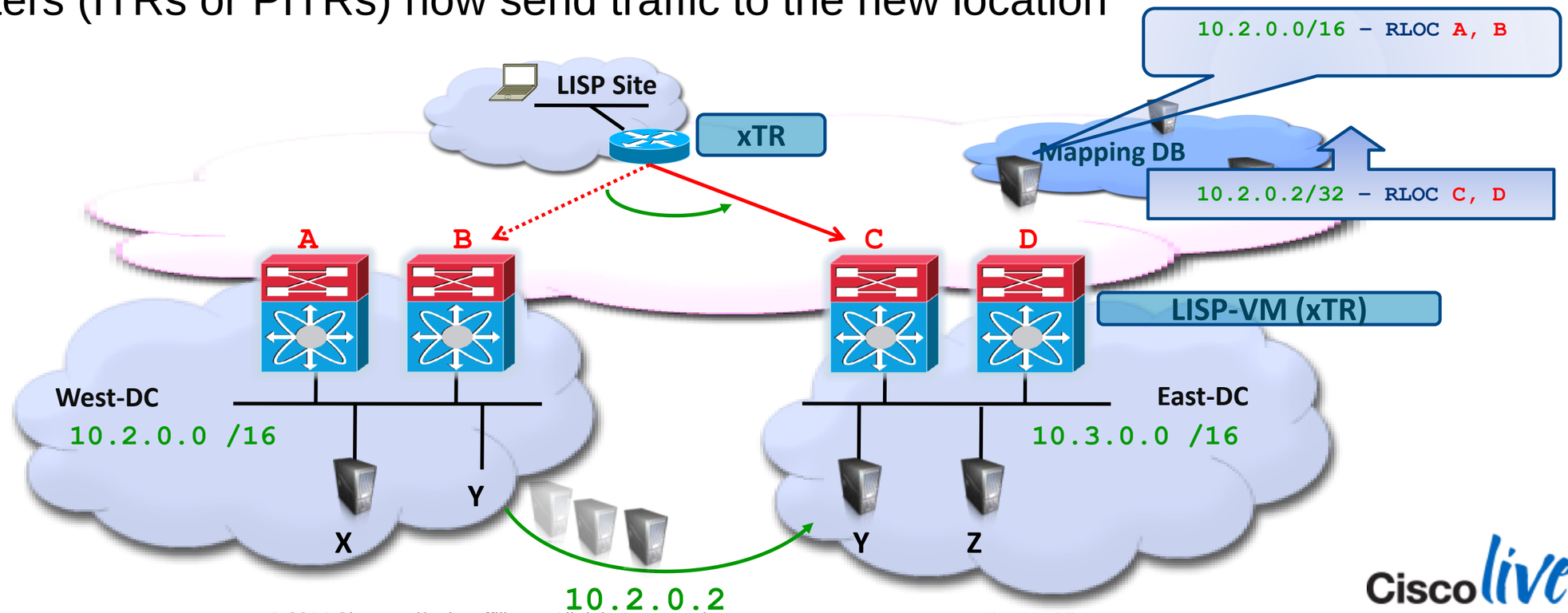
… It's from a "New" Host
*… It's in the **Dynamic-EID** Allowed Range*

…It's a Move!
Register the /32 with LISP



LISP Host-Mobility – Traffic Redirection

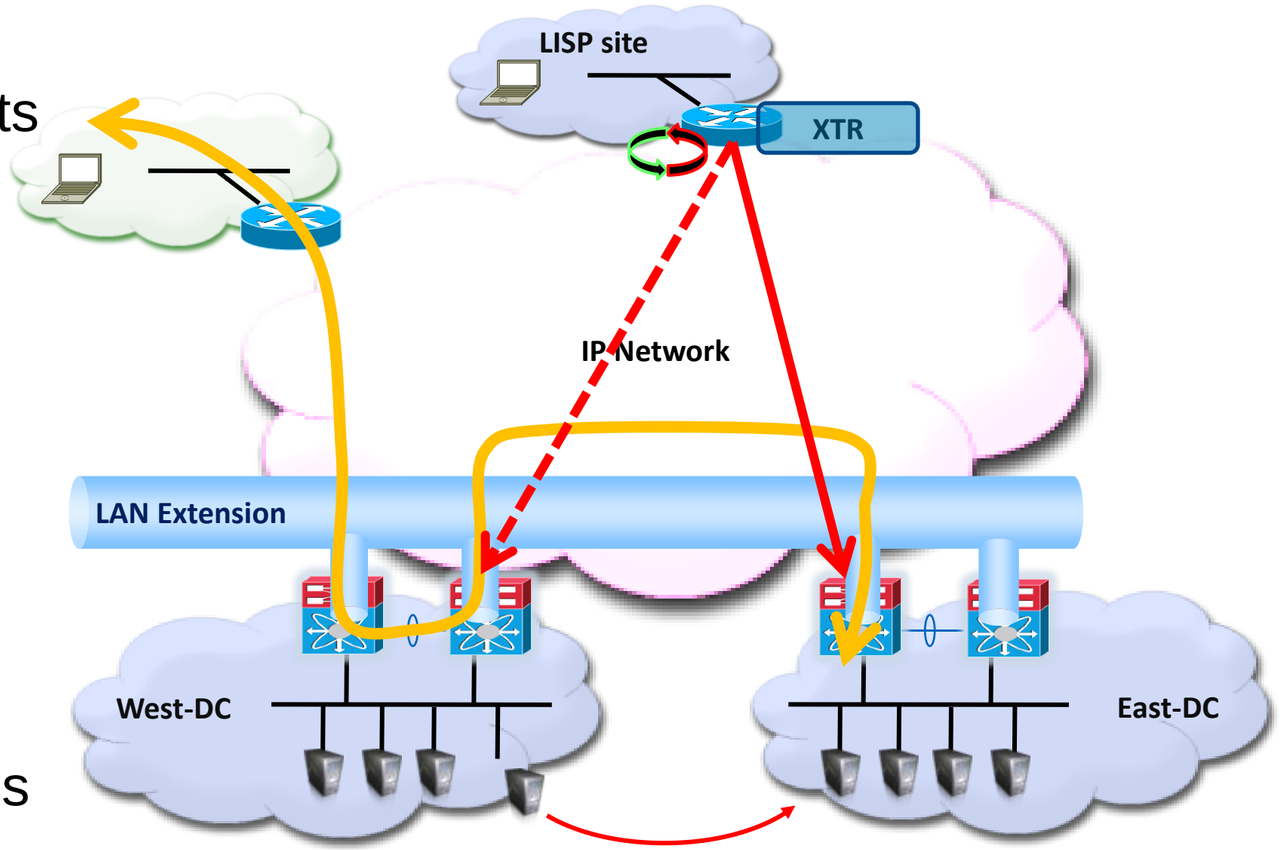
- Update Location Mappings for the Host System Wide
- When a host move is detected, updates are triggered:
  - The host-to-location mapping in the Database is updated to reflect the new location
  - The old ETR is notified of the move
  - ITRs are notified to update their Map-caches
- Ingress routers (ITRs or PITRs) now send traffic to the new location



Ingress Routing Challenge in DCI

Extending Subnets Creates a Routing Challenge

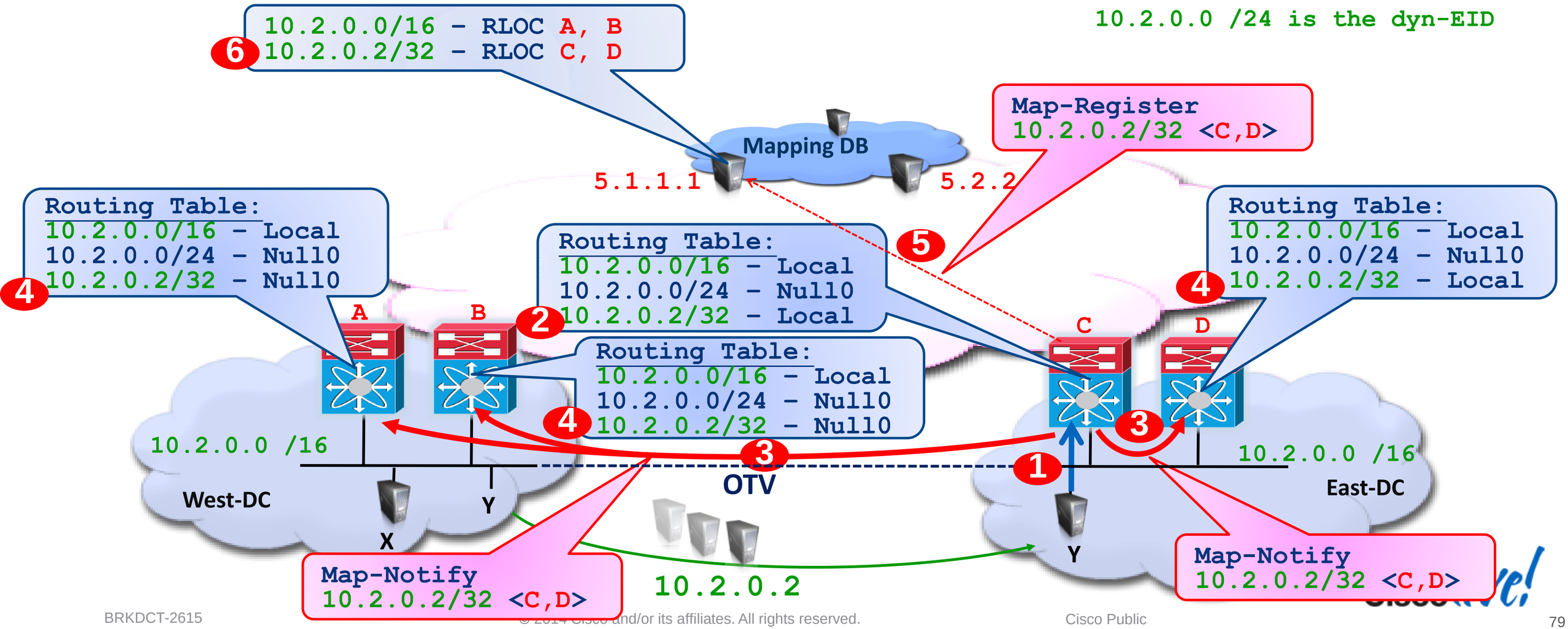
- A subnet traditionally implies location
- Yet we use LAN extensions to stretch subnets across locations
  - Location semantics of subnets are lost
- Traditional routing relies on the location semantics of the subnet
  - Can't tell if a server is at the East or West location of the subnet
- More granular (host level) information is required
  - LISP provides host level location semantics



Host-Mobility and Multi-homing

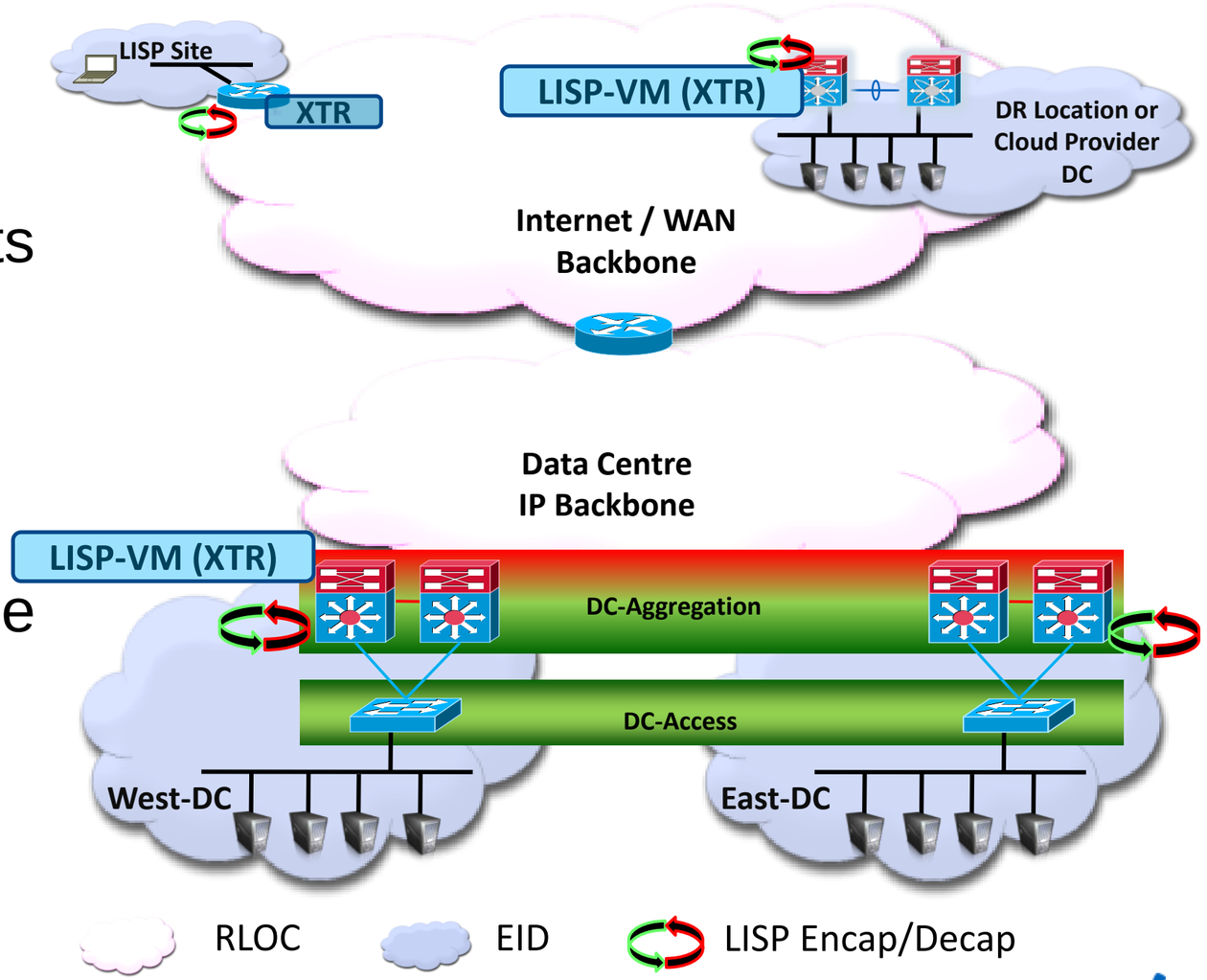
- ETR updates – Extended Subnets

Null0 host routes indicate the host is “away”



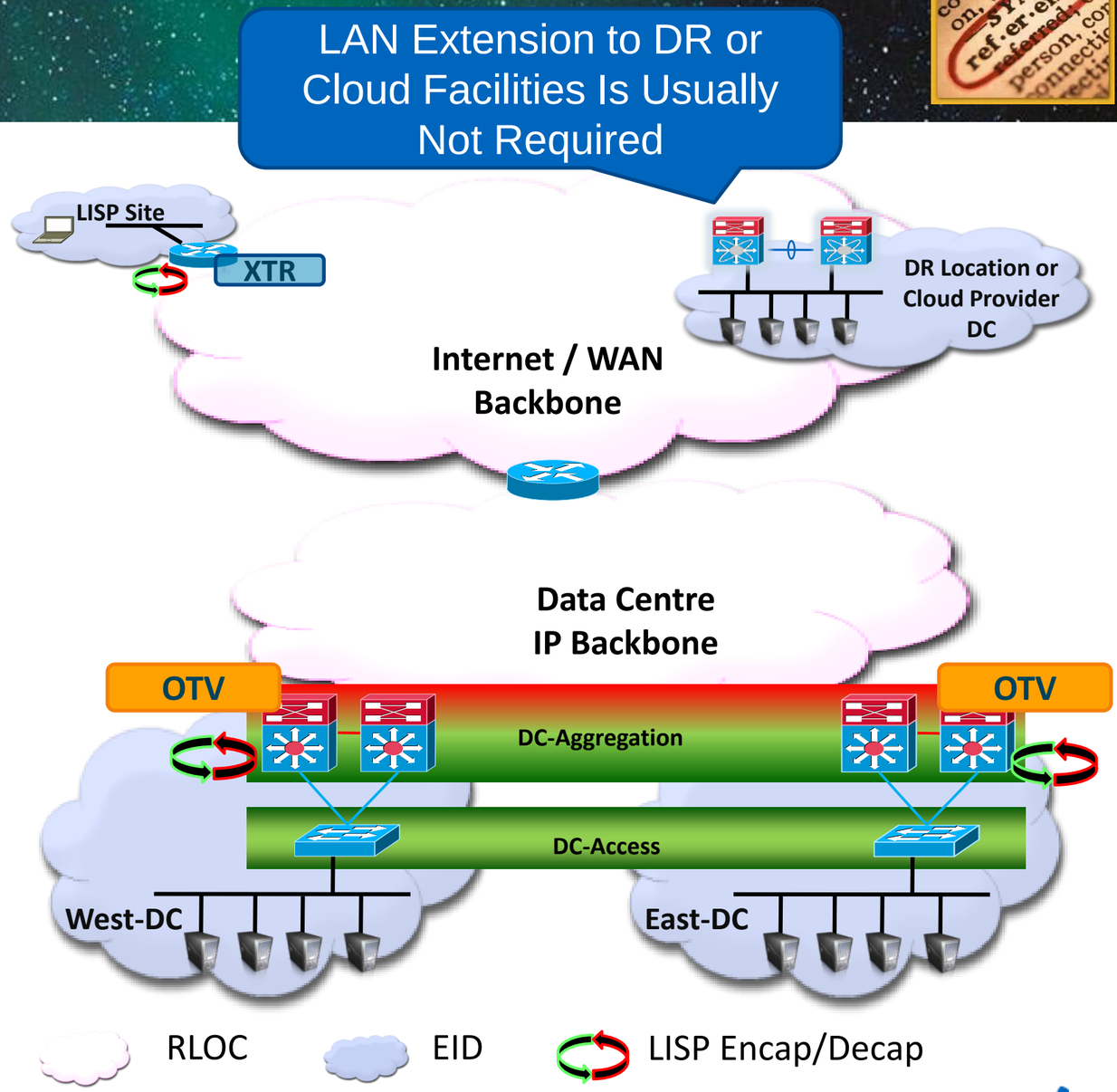
LISP Host-Mobility – Router Placement

- @ Main Data Centres
- @ Disaster Recover facilities
- Ideally: First hop routers for the subnets in which the mobile hosts reside:
  - Detect host moves
  - Provide a consistent first hop presence
  - Could also be the second hop
- Usually the Aggregation Switches in the Data Centre
- Customer Managed



OTV Router Placement

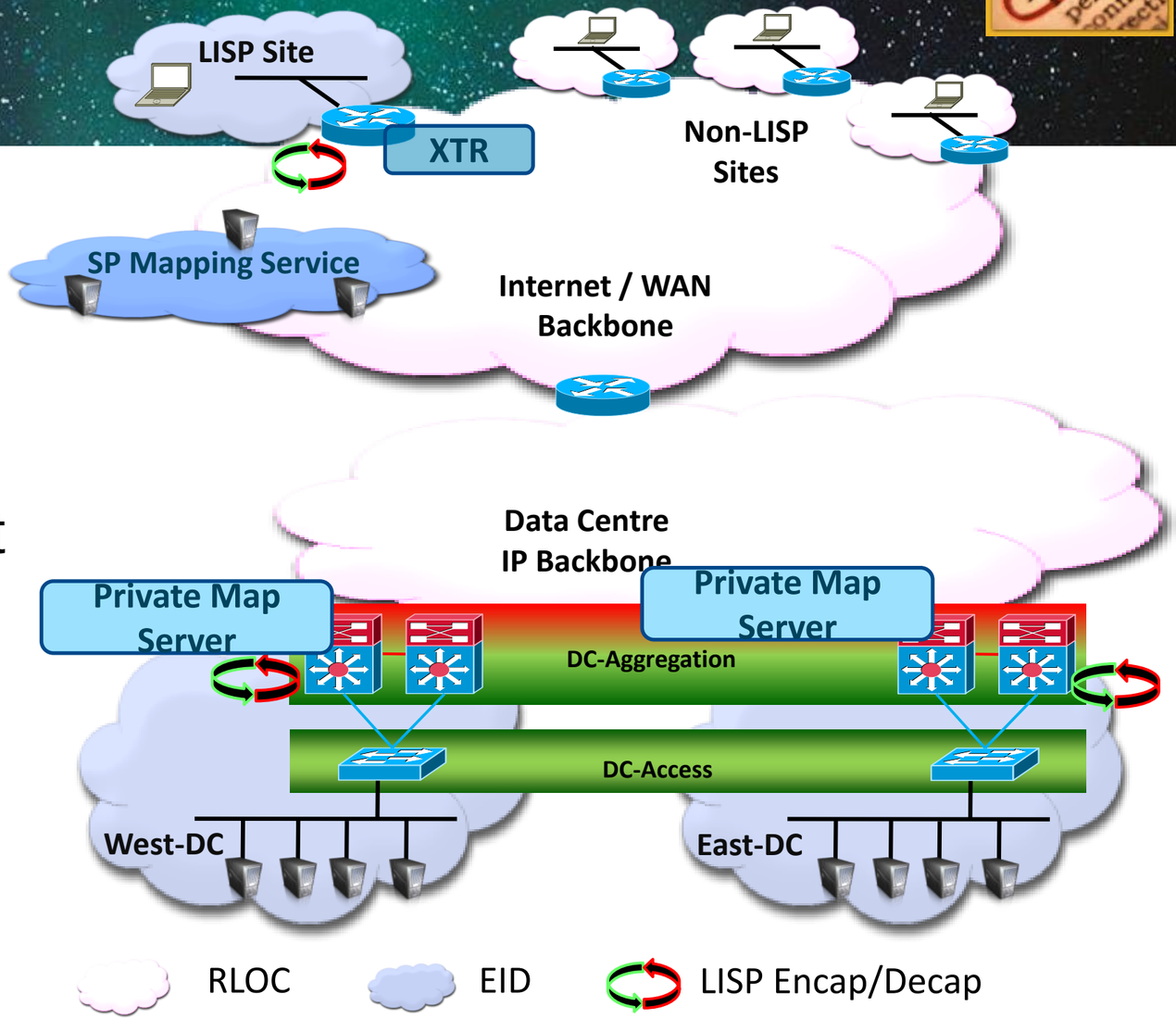
- @ Main Data Centres only
- Typically not required @ Disaster Recover facilities
- First hop routers for the subnets in which the mobile hosts reside:
  - Connect to the VLANs to be extended
  - Connect to the IP core
- Usually the Aggregation Switches in the Data Centre
- Customer Managed



Map Server Placement

A Daemon on a Router

- The Map Server functionality can be enabled on any router
  - BGP route-reflectors are a good analogy
  - Off path is good, but not mandatory
- Distribute Map Servers across different locations
  - Private Data Centres (Self managed)
  - SP Data Centres/Cloud (SP Service)
- Map Server resiliency options:
  - Clustered and distributed
  - Distributed Database (DDT)



Agenda

- Active-Active Data Centre: Business Drivers and Solutions Overview
- Active / Active Data Centre Design Considerations
  - Storage Extension
  - Data Centre Interconnect (DCI) - LAN Extension Deployment Scenarios
  - Host Mobility using LISP and OTV
  - Network Services and Applications (Path optimisation)
- Cisco ACI and Active / Active Data Centre
- Summary and Conclusions
- Q&A

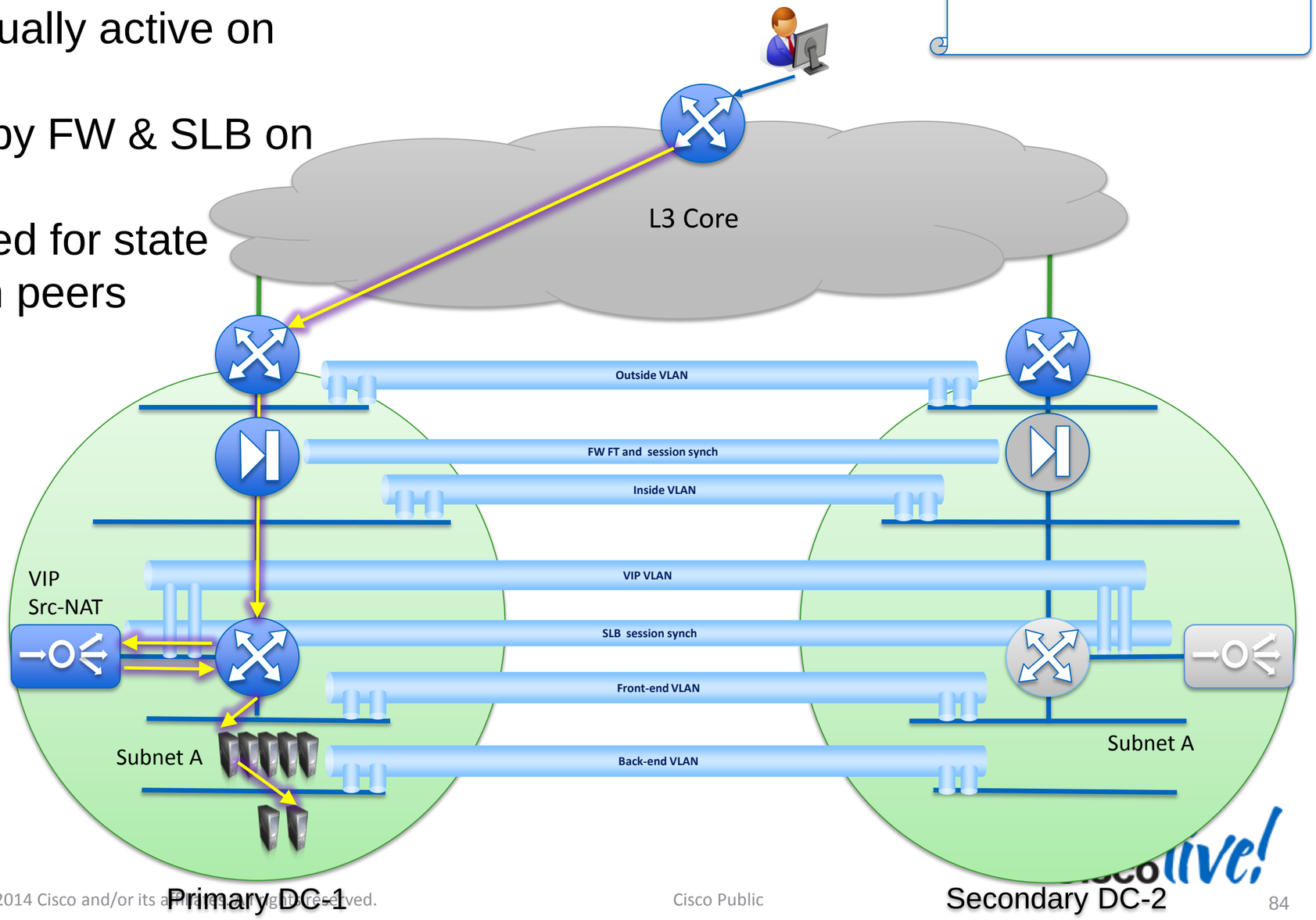
Network Service Placement for Metro Distances

A/S state-full devices stretched across 2 locations – nominal workflow

- This has been deployed for most of Metro Virtual DC (Twin-DC)
- Almost 80% of Twin-DC follows this model

- Network Services are usually active on primary DC
  - Distributed pair of Act/Sby FW & SLB on each location
- Additional VLAN Extended for state synchronisation between peers
- Source NAT for SLB VIP

Nota: With traditional pair cluster this scenario is limited to 2 sites

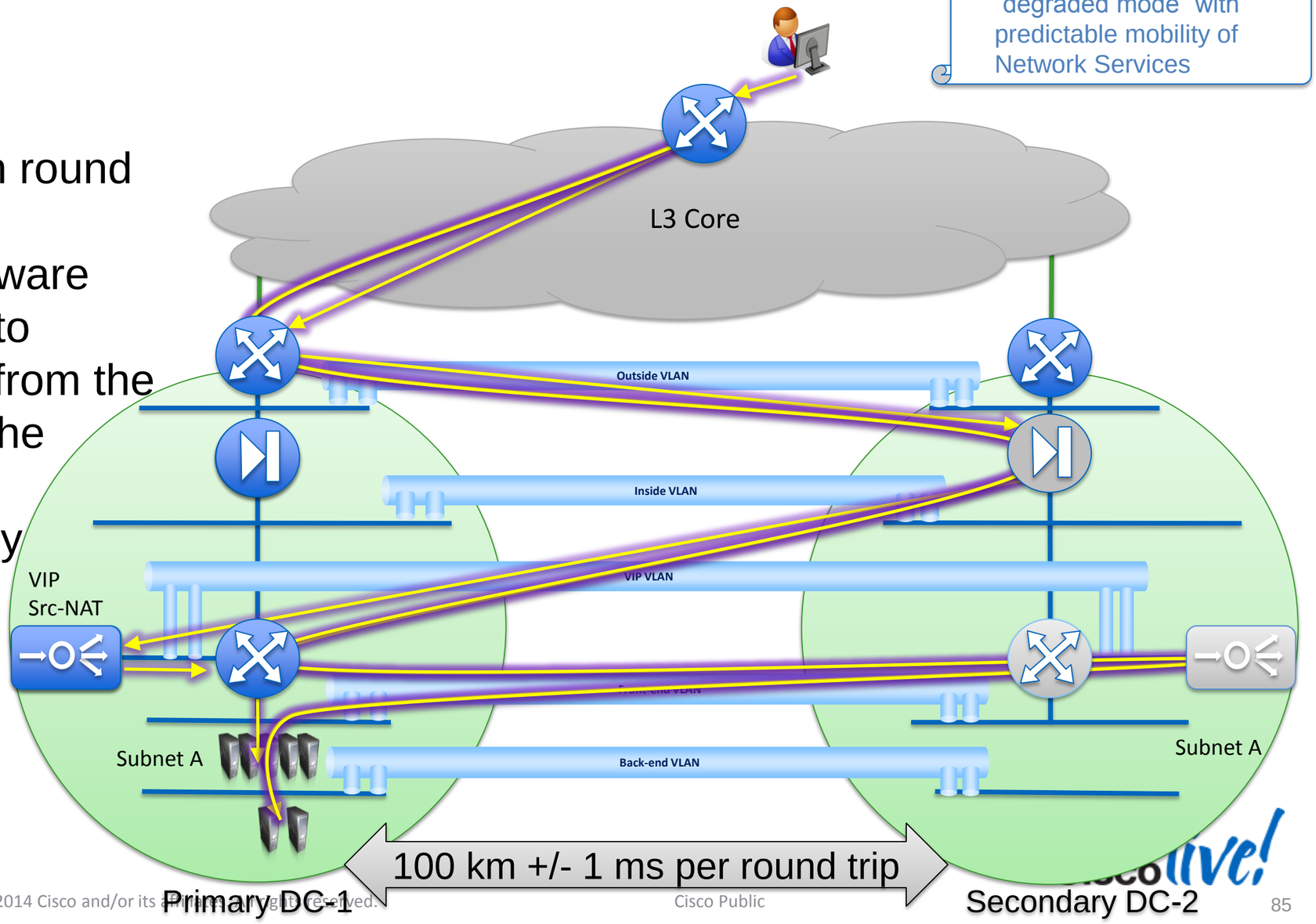


Network Service Placement for Metro Distances

Ping-Pong impact with A/S state-full devices stretched across 2 locations

- Historically limited to Network services and HA clusters offering state-full failover & fast convergences
- It is accepted to work in “degraded mode” with predictable mobility of Network Services

- FW failover to remote site
- Source NAT for SLB VIP
- Consider +/- 1 ms for each round trip for 100 km
- For Secured multi-tier software architecture, it is possible to measure + 10 round-trips from the initial client request up to the result.
- Interface tracking optionally enabled to maintain active security and network services on the same site

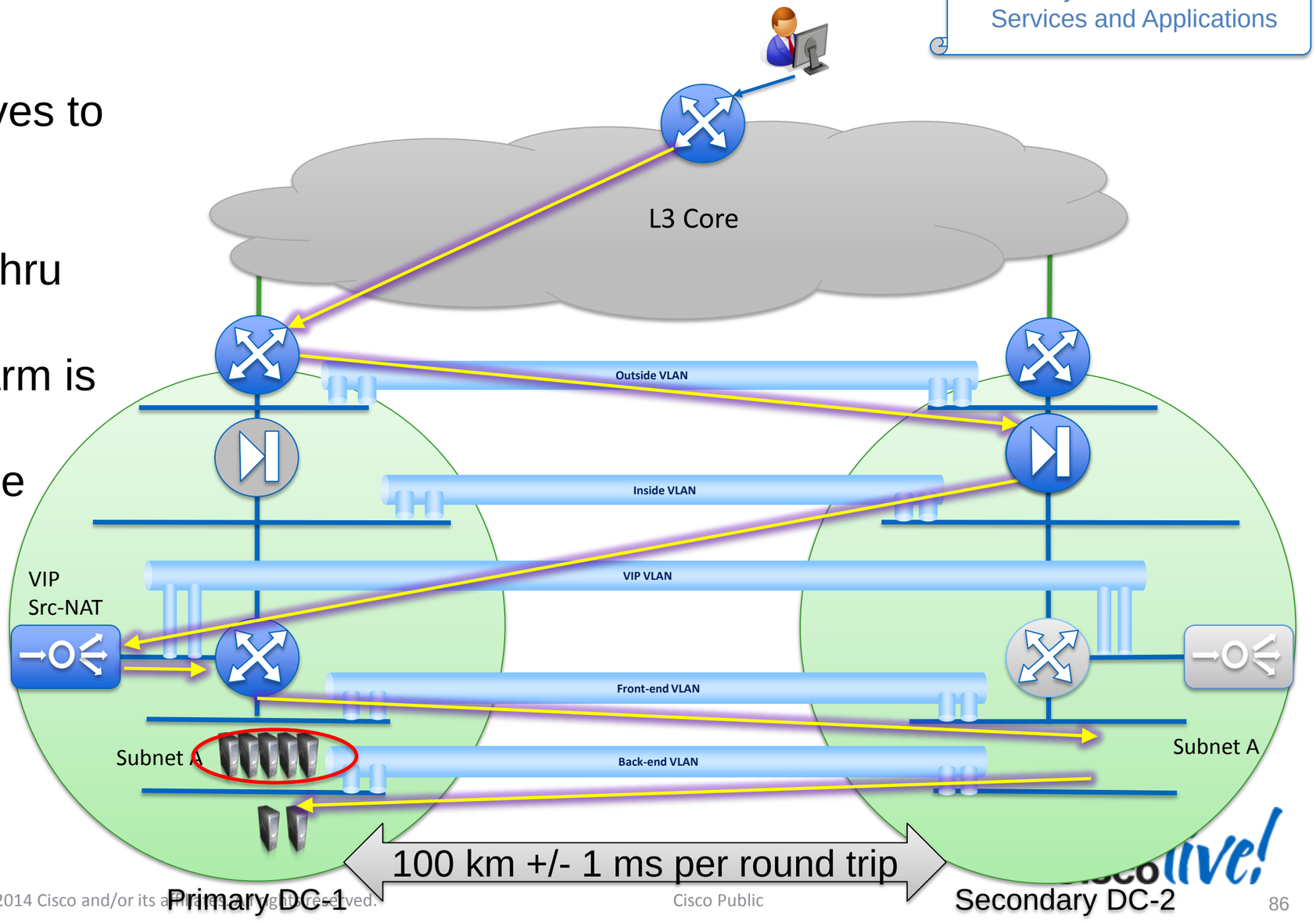


Network Service placement for Metro Distances

Additional Ping-Pong impact with IP mobility between 2 locations

- Network team is not necessarily aware of the Application/VM mobility
- Uncontrolled degraded mode with unpredictable mobility of Network Services and Applications

- FW failover to remote site
- Front-end server farm moves to remote site
- Source NAT for SLB VIP maintains the return path thru the Active SLB
- Partial move of a server farm is not optimised
- Understand and identify the multi-tier frameworks

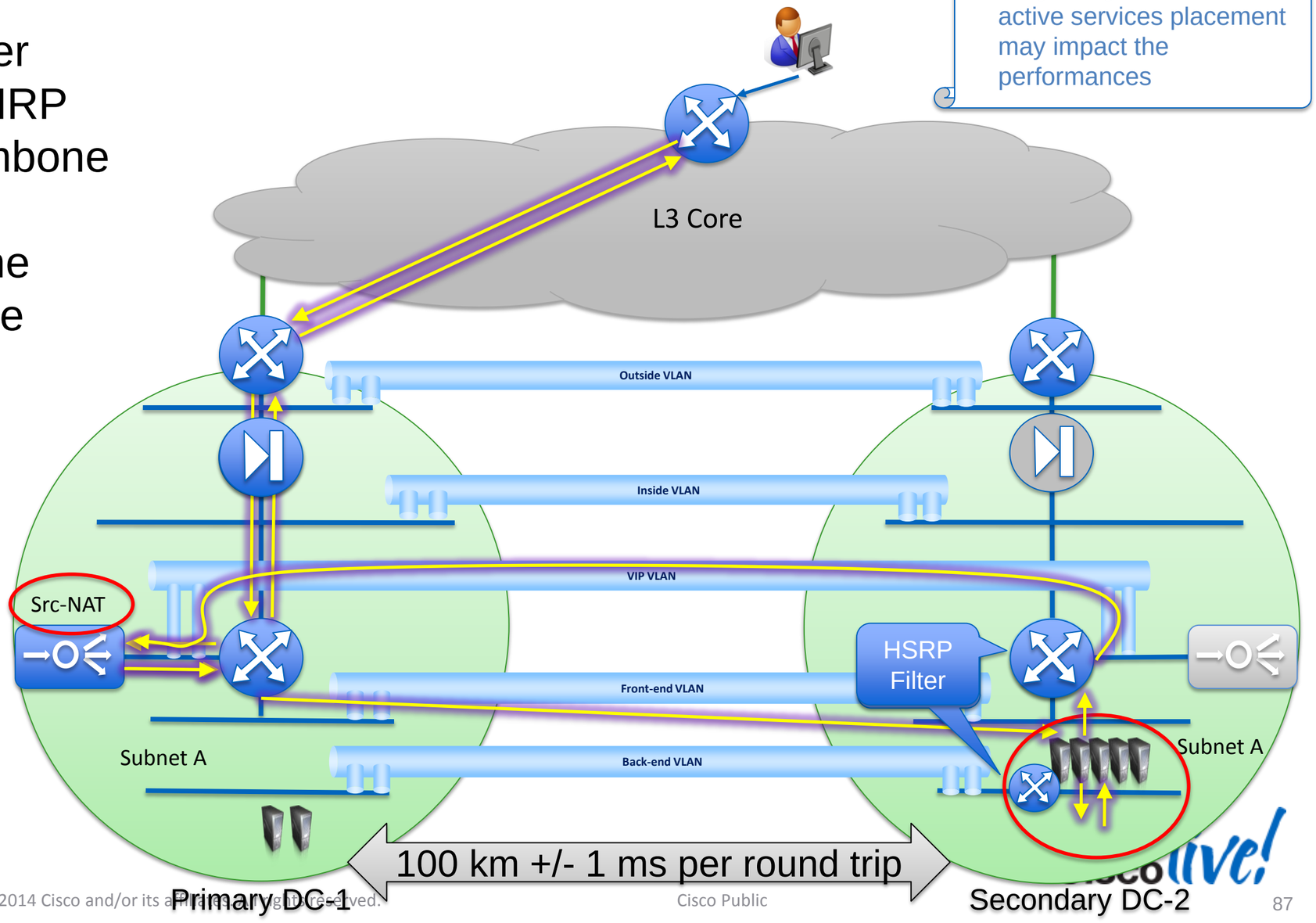


Network Service placement for Metro Distances

State-full Devices and Trombone effect for IP Mobility between 2 locations

- Limited relation between server team (VM mobility) and Network Team (HSRP Filtering) and Service Team (FW, SLB, IPS..)
- Ping-Pong effect with active services placement may impact the performances

- Migrate the whole multi-tier framework and enable FHRP filtering to reduce the trombone effect
- FHRP filtering is ON on the Front-end & Back-end side gateways
- Source NAT for SLB VIP maintains the return path thru the Active SLB
- Understand and identify the multi-tier frameworks
- Therefore it is preferred to move the whole application tiers.



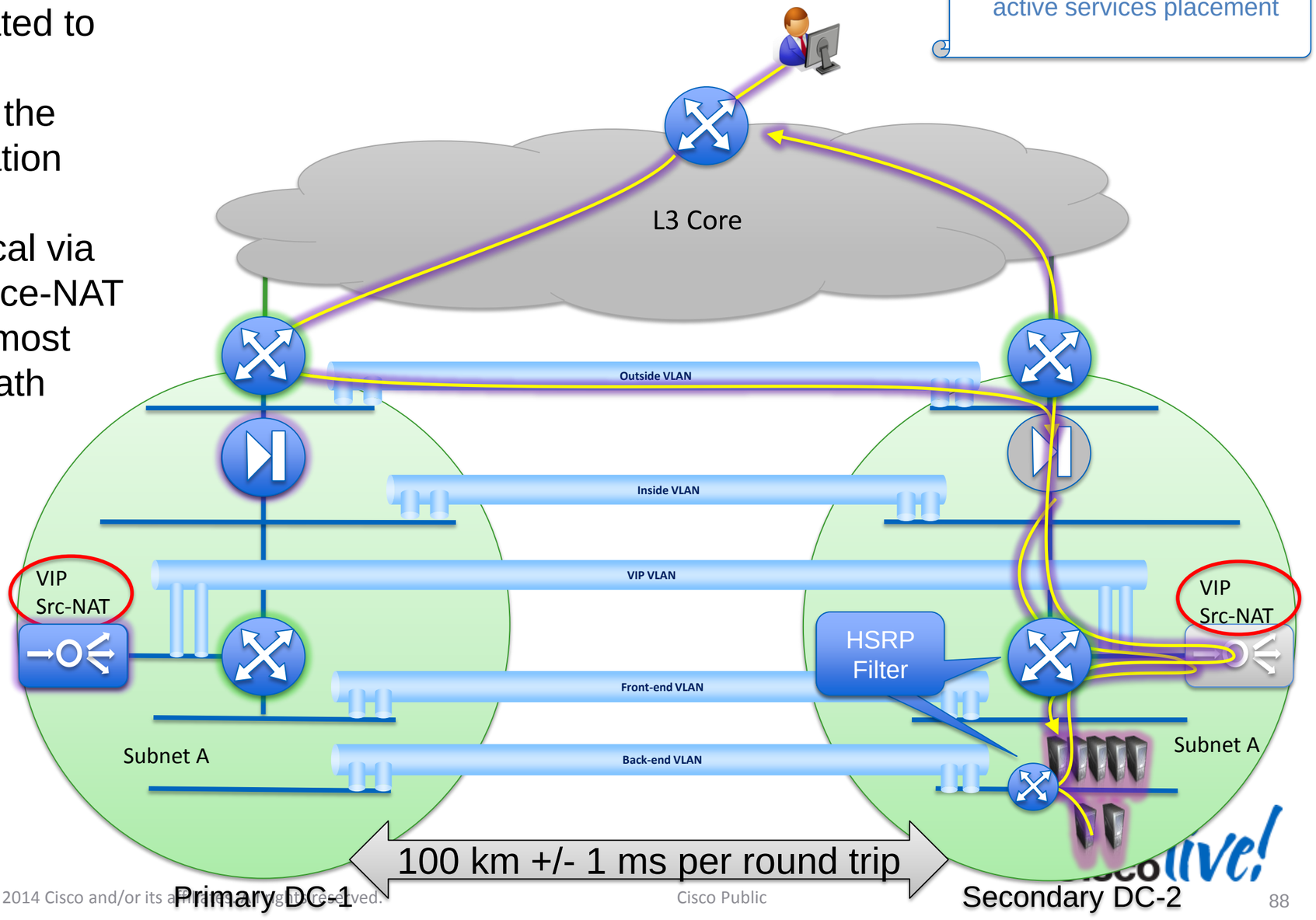
Network Service Placement for Metro Distances

Intelligent placement of Network Services based on IP Mobility localisation

- Improving relations between sillo'ed organisations increases workflow efficiency
- Reduce trombon'ing with active services placement

- Move the FW Context associated to the application of interests
- Interface Tracking to maintain the state-full devices in same location when possible
- Return traffic keeps symmetrical via the state-full devices and source-NAT
- Intra-DC Path Optimisation almost achieved , however Ingress Path Optimisation may be required

- Sillo'ed organisations
  - Server/app
  - Network/HSRP filter service & security
  - Storage

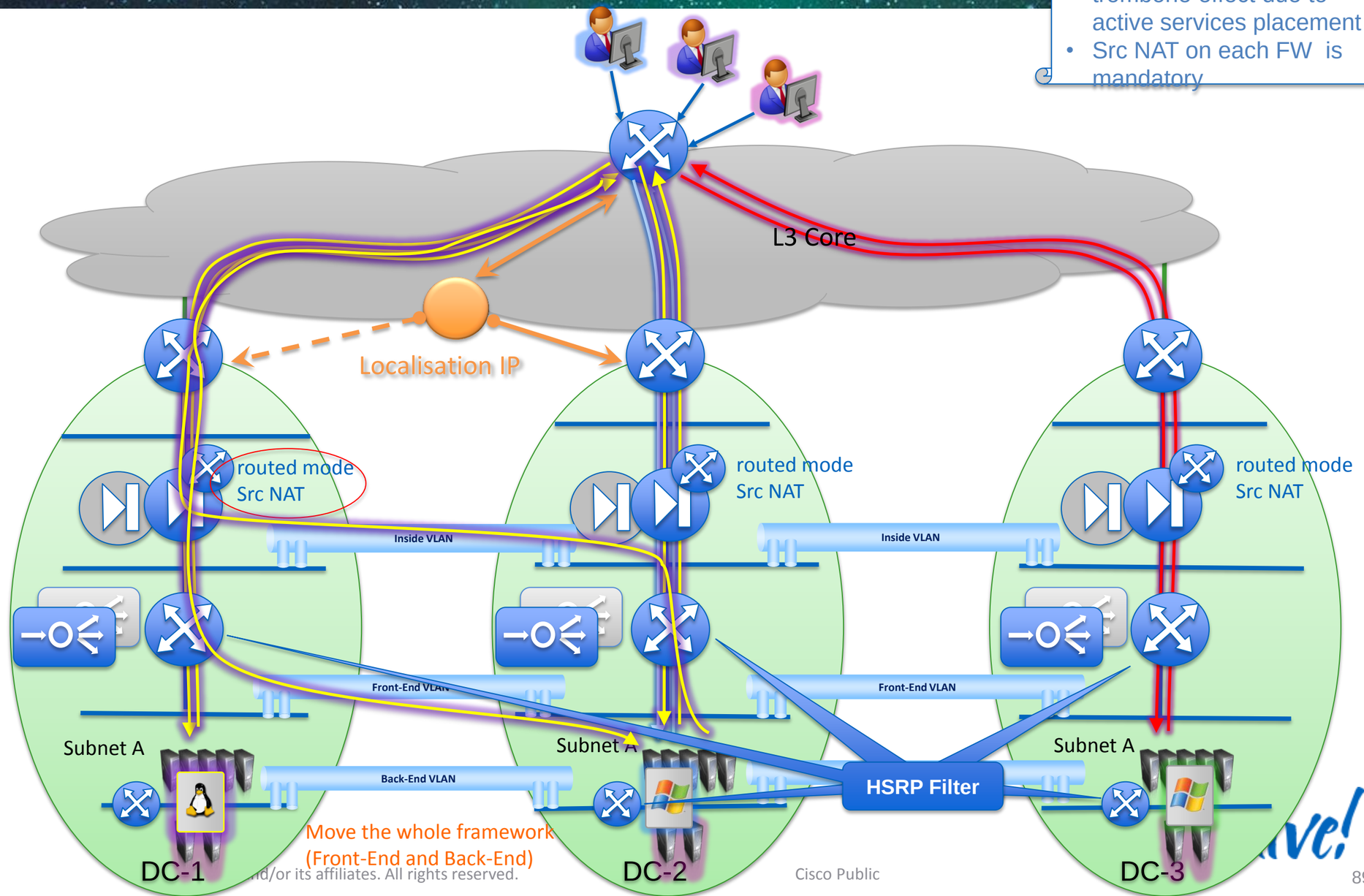


Network Service Placement for Long Distances

Active/Standby Network Services per Site with Extended LAN (State-full Live migration)

- Extend the VLAN of interests
- FW and SLB maintain state-full session per DC.
- No real limit in term of number of DC
- Granular migration is possible only using LISP or RHI (if the Enterprise owns the L3 core)

- ~~Subnet Replication is~~ possible using NAT or LISP
- Ingress Path Optimisation can be initiated to reduce trombone effect due to active services placement
- Src NAT on each FW is mandatory

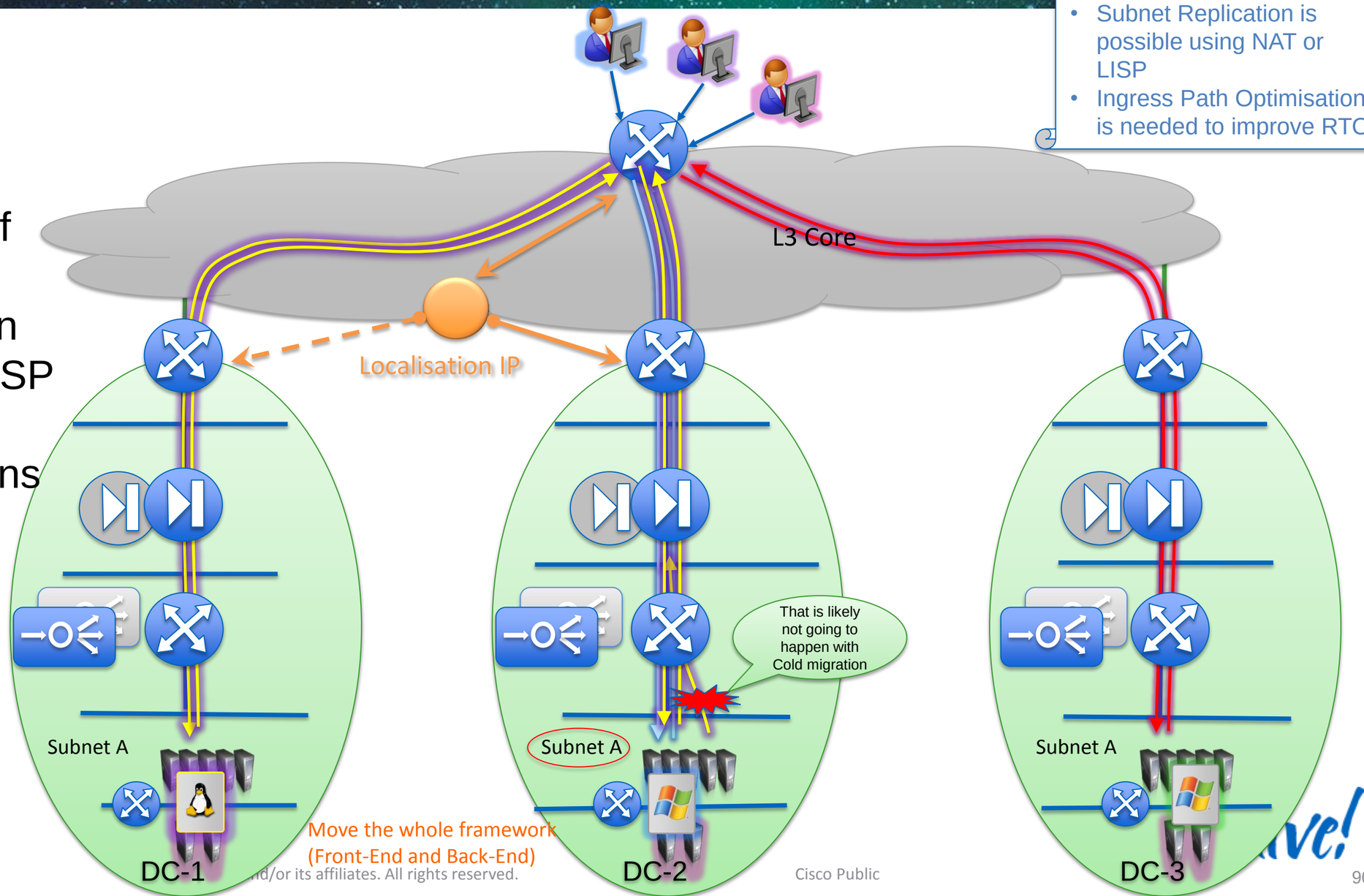


Network Service Placement for Long Distances

Active/Standby Network Services per Site across Subnets (Cold migration)

- Implies “Cold” migration (stateless)
- LAN Extension is not required for “Cold” migration
- Subnet Replication is possible using NAT or LISP
- Ingress Path Optimisation is needed to improve RTO

- FW and SLB maintain state-full session per DC.
- No Limit in term of number of DC
- Granular migration is possible with LISP or RHI (assuming the Enterprise owns the L3 core)
- FW forwarding mode can be Routed or Transparent



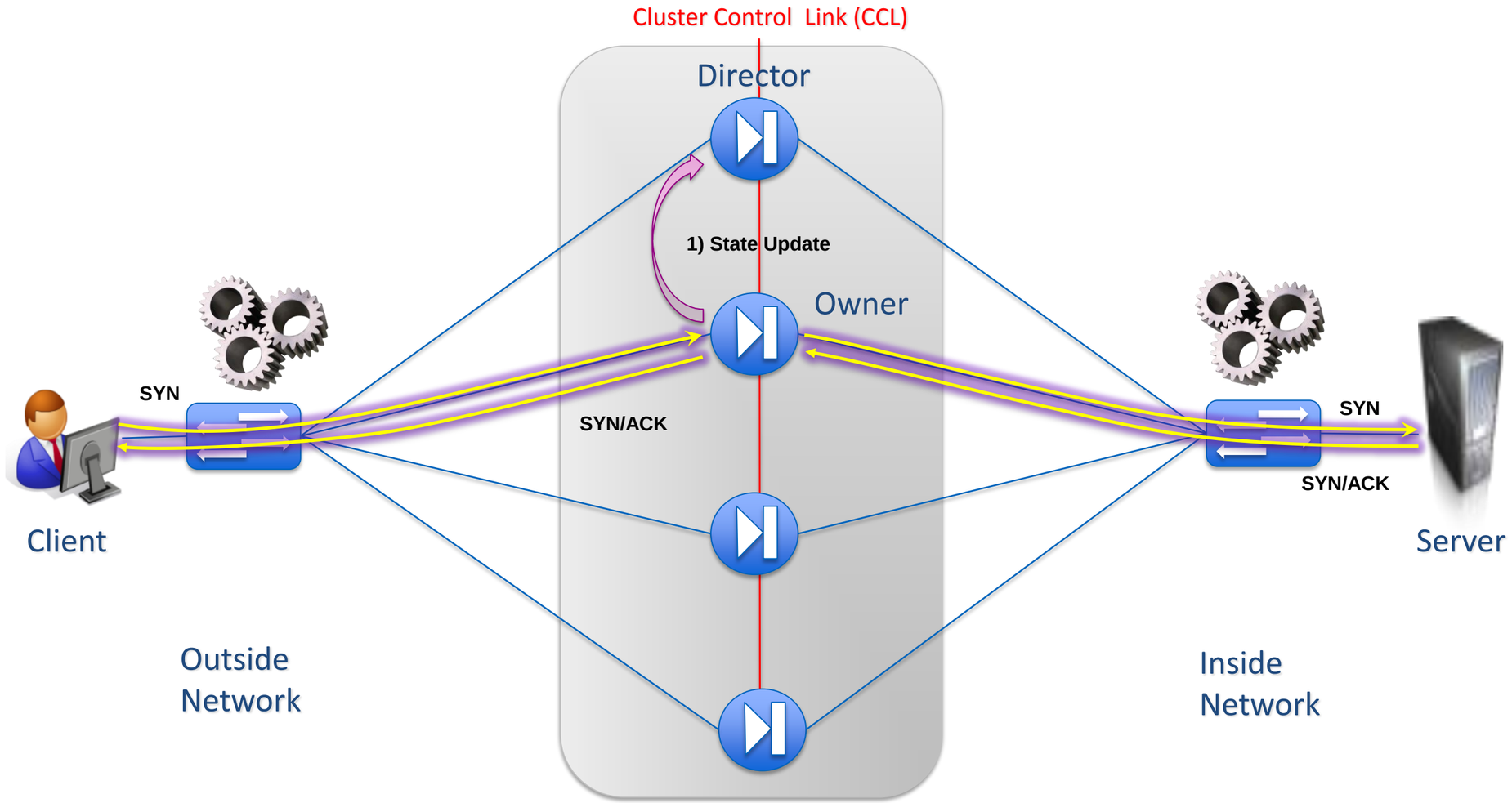



Can ASA Clustering Improve This?

ASA Clustering Deployment

ASA Clustering (9.0)

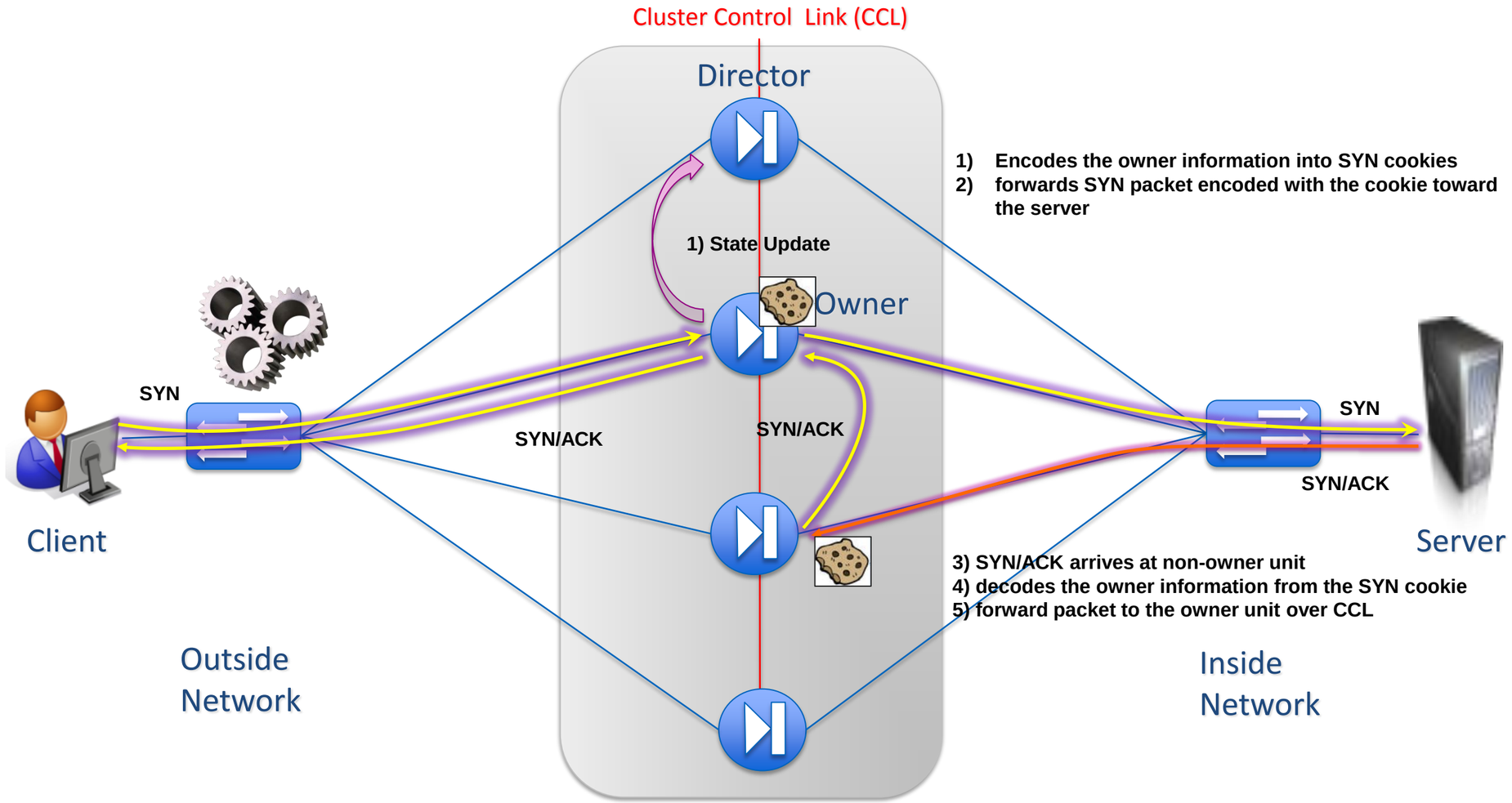
Connection setup when traffic is Symmetric



- State replication from Owner to Director, also serves as failover msg to provide redundancy should owner fail
- Director is selected per connection using consistent hashing algorithm.

ASA Clustering (9.0)

TCP SYN cookies with Asymmetrical Traffic workflows

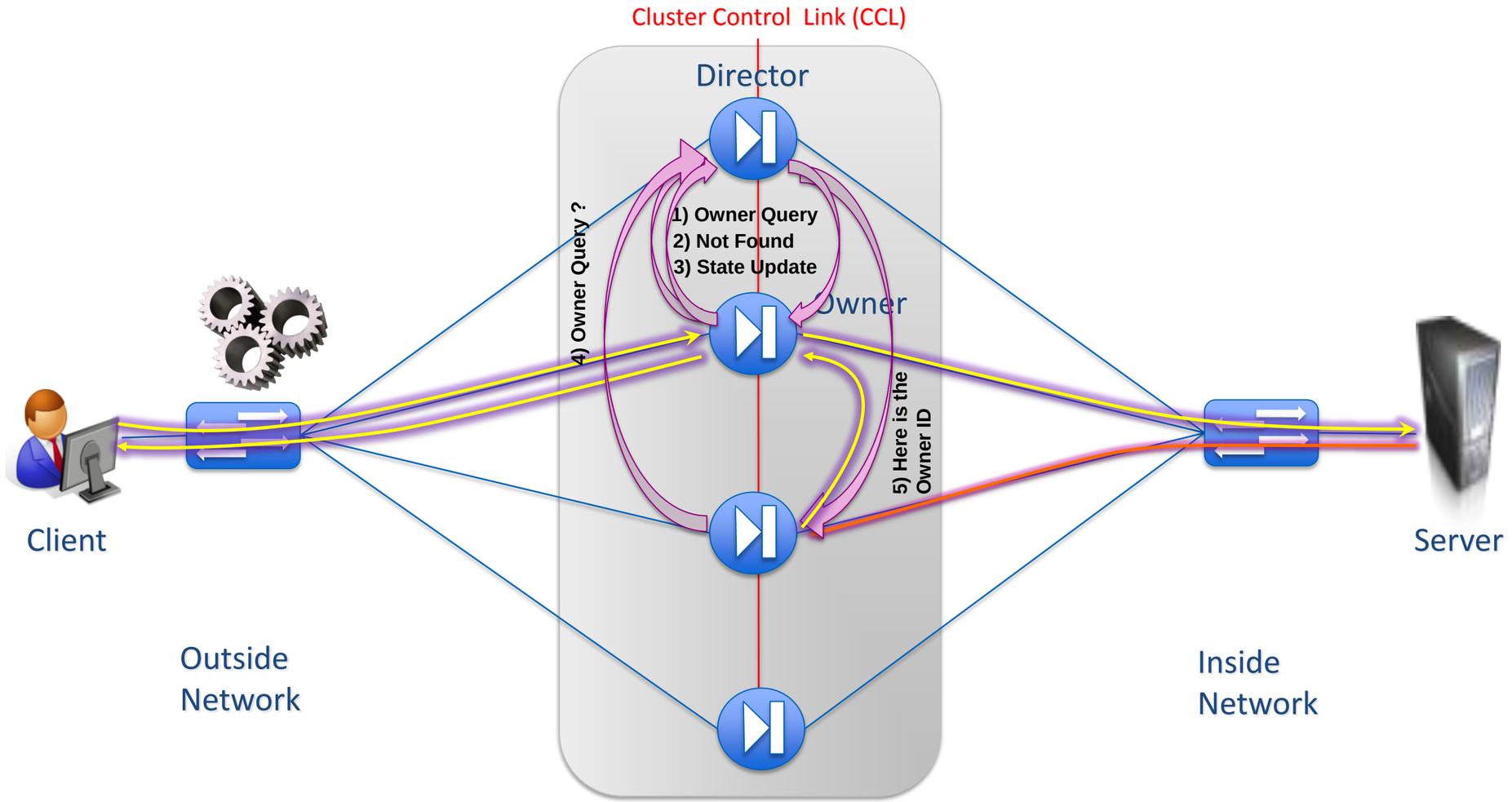


It is possible that the SYN/ACK from the server arrives at a non-owner unit before the connection is built at the director.

- As the owner unit processes the TCP SYN, it encodes within the Sequence # which unit in the cluster is the owner
- Other units can decode that information and forward the SYN/ACK directly to the owner without having to query the director

ASA Clustering (9.0)

UDP sessions with Asymmetric Traffic workflows

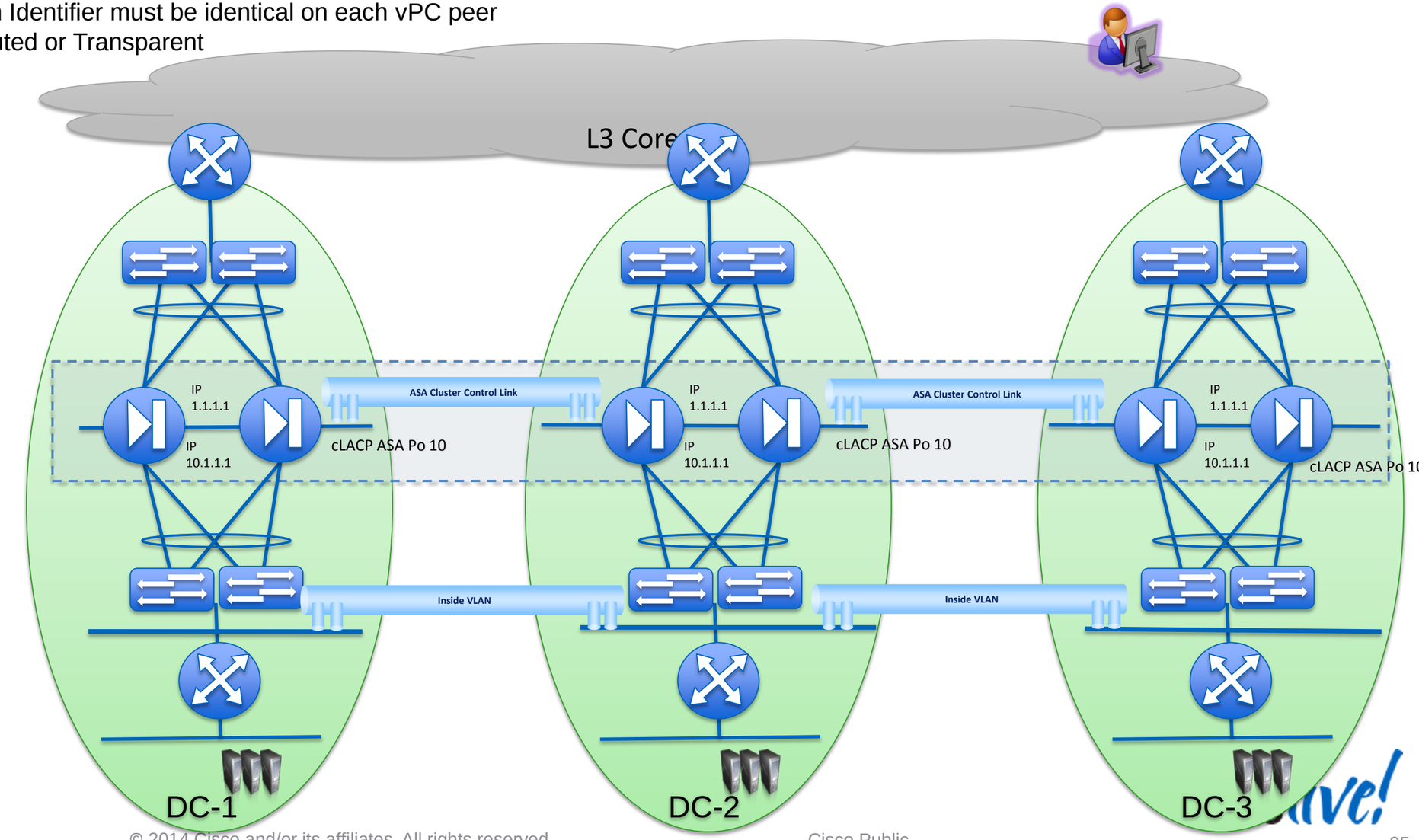


- When a unit receives a UDP packet for a flow that it does not own, it queries the director to find the owner
- Thereafter, it maintains a forwarding flow. It can punt packets directly to the owner, bypassing the query to the director
- Short-lived flows (eg. DNS, ICMP) do not have forwarding flows

Single ASA Cluster stretched across multiple sites

ASA Clustering Data Plane Load Distribution using interface Layer 2 mode

- Only 1 port-channel from the ASA clustering
- cLACP imposes that the same port channel must exist across the same ASA cluster
- Therefore the same vPC Domain Identifier must be identical on each vPC peer
- FW forwarding mode can be Routed or Transparent



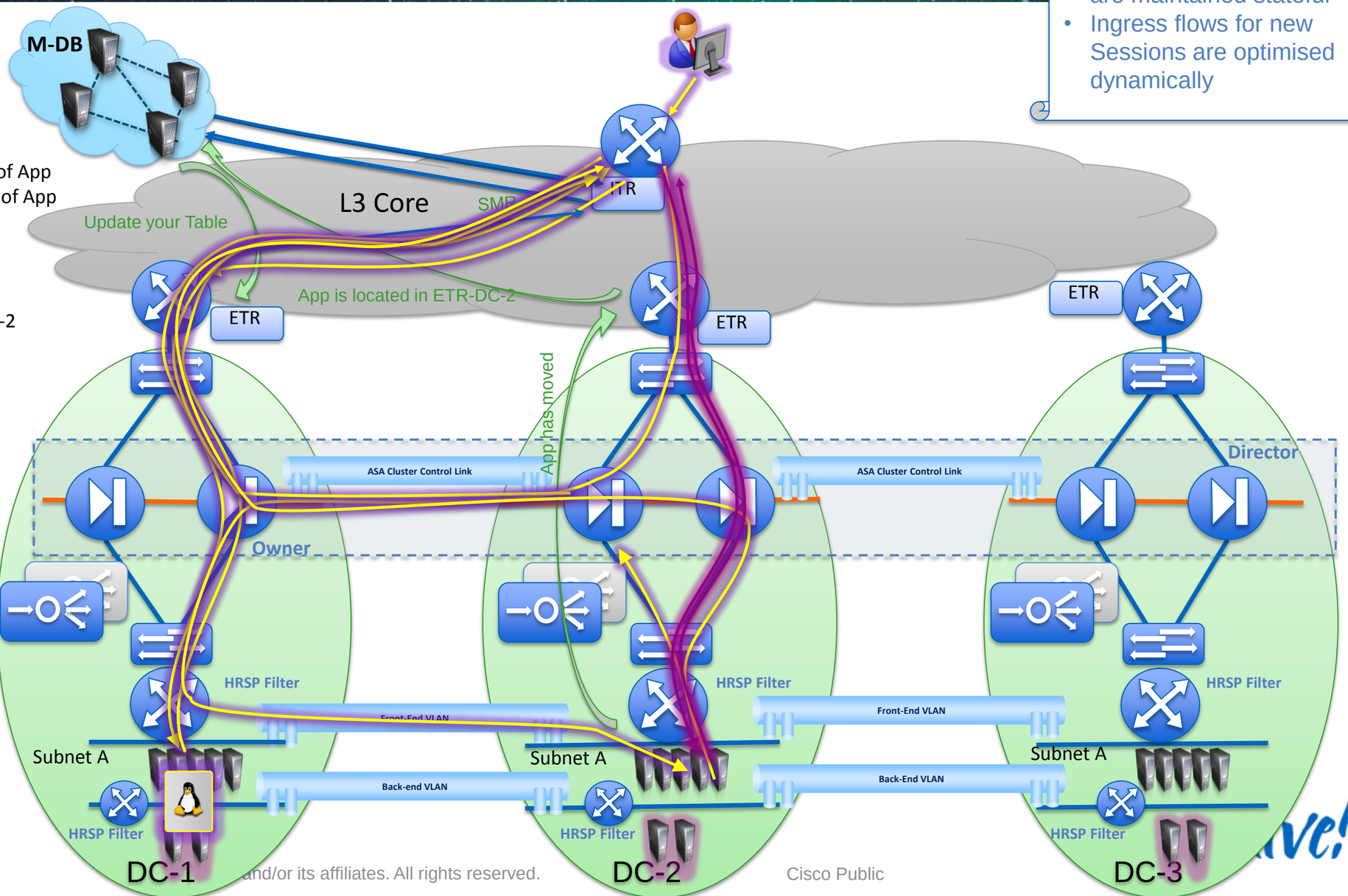
Single ASA Clustering Stretched Across Multiple DC

LISP Extended Subnet Mode with ASA Clustering (Stateful Live migration with LAN extension)

- One Way Symmetric Establishment is achieved via the CCL
- Current active sessions are maintained stateful
- Ingress flows for new Sessions are optimised dynamically

1 - End-user sends Request to App
2 - ITR intercepts the Req and check the localisation
3 - MS replies location for Subnet A being ETR DC-1
3” - ITR encaps the packet and sends it to RLOC ETR-DC-1

4 – LISP Multi-hop informs ETR on DC-2 about the move of App
5 – Meanwhile ETR DC-2 informs MS about new location of App
6 – MR updates ETR DC-1
7 – ETR DC-1 updates its table (App:Null0)
8 – ITR sends traffic to ETR DC-1
9 – ETR DC-1 replies with a Solicit Map Req
8 – ITR sends a Map Req and redirects the Req to ETR DC-2



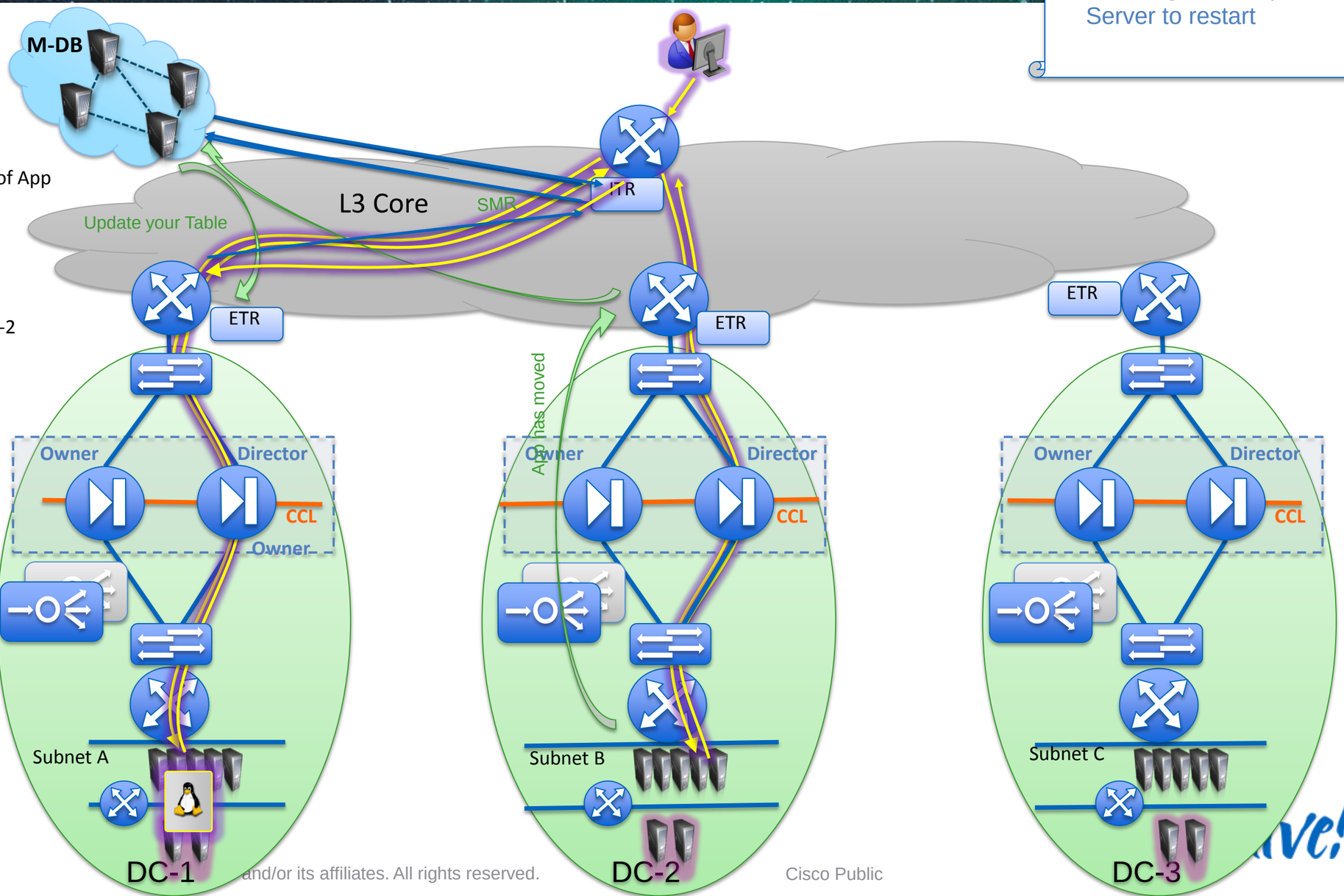
ASA Clustering per DC across Multiple sites

LISP Across Subnet Mode with local ASA Clustering (Cold migration)

- Business continuity assumes the user re-establish a new session
- TCP session is re-initiated
- Cold Migration implies the Server to restart

1 - End-user sends Request to App
2 - ITR intercepts the Req and check the localisation
3 - MS replies location for Subnet A being ETR DC-1
3” - ITR encaps the packet and sends it to RLOC ETR-DC-1

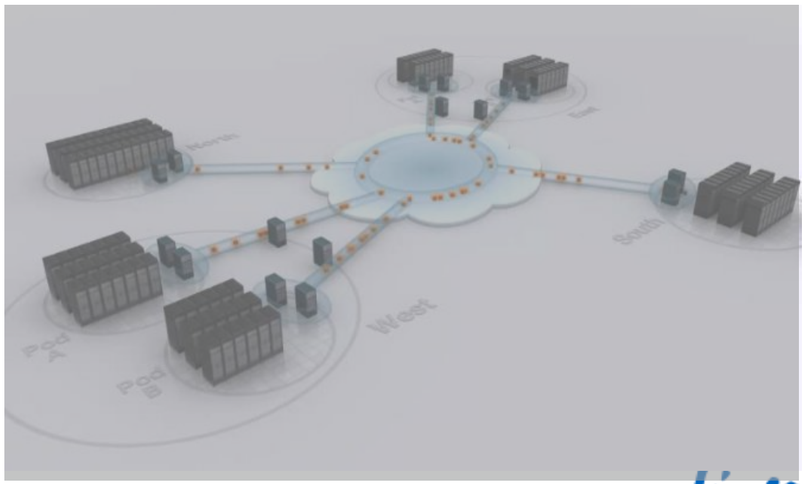
4 – LISP Multi-hop informs ETR on DC-2 about the move of App
5 – ETR DC-2 informs MS about new location of App
6 – MR updates ETR DC-1
7 – ETR DC-1 updates its table (App:Null0)
8 – ITR sends traffic to ETR DC-1
9 – ETR DC-1 replies Solicit Map Req
8 – ITR sends a Map Req and redirects the Req to ETR DC-2



Agenda

- Active-Active Data Centre: Business Drivers and Solutions Overview
- Active / Active Data Centre Design Considerations
  - Storage Extension
  - Data Centre Interconnect (DCI) - LAN Extension Deployment Scenarios
  - Host Mobility using LISP and OTV
  - Network Services and Applications (Path optimisation)

- Cisco ACI and Active / Active Data Centre
- Summary and Conclusions
- Q&A

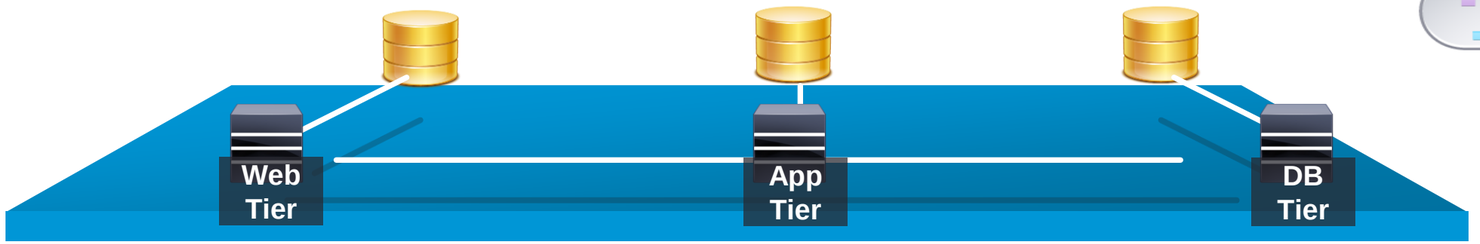


ACI Goal: Common Policy and Operations Framework

Application Admin

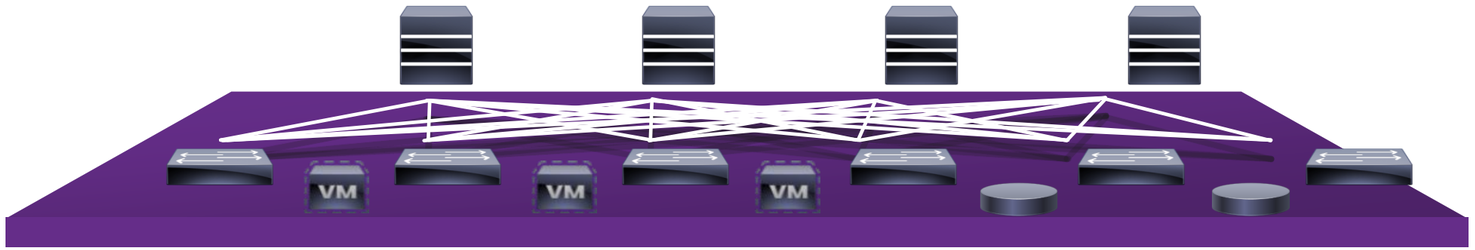
Security Admin

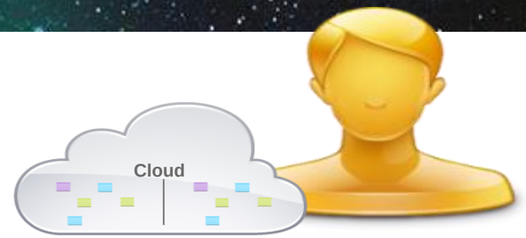
Network Admin



APPLICATION

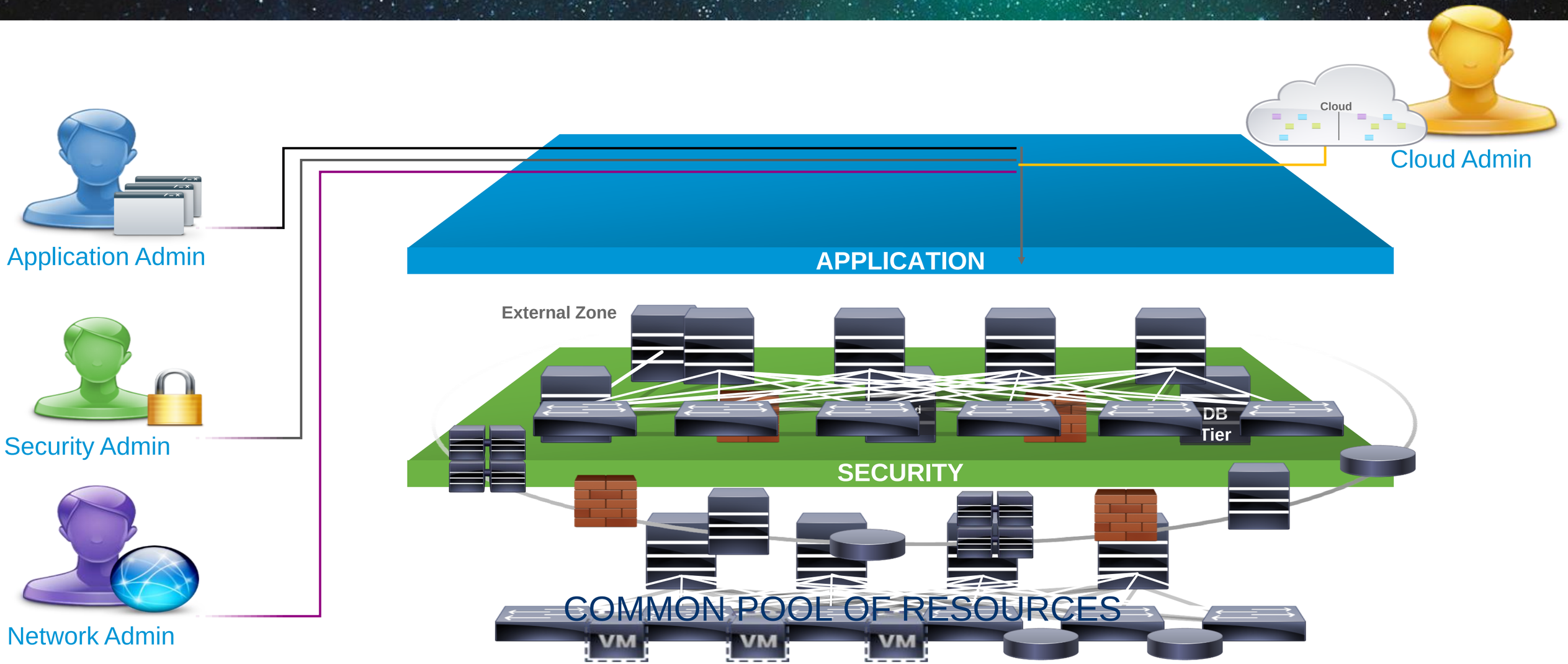
SECURITY





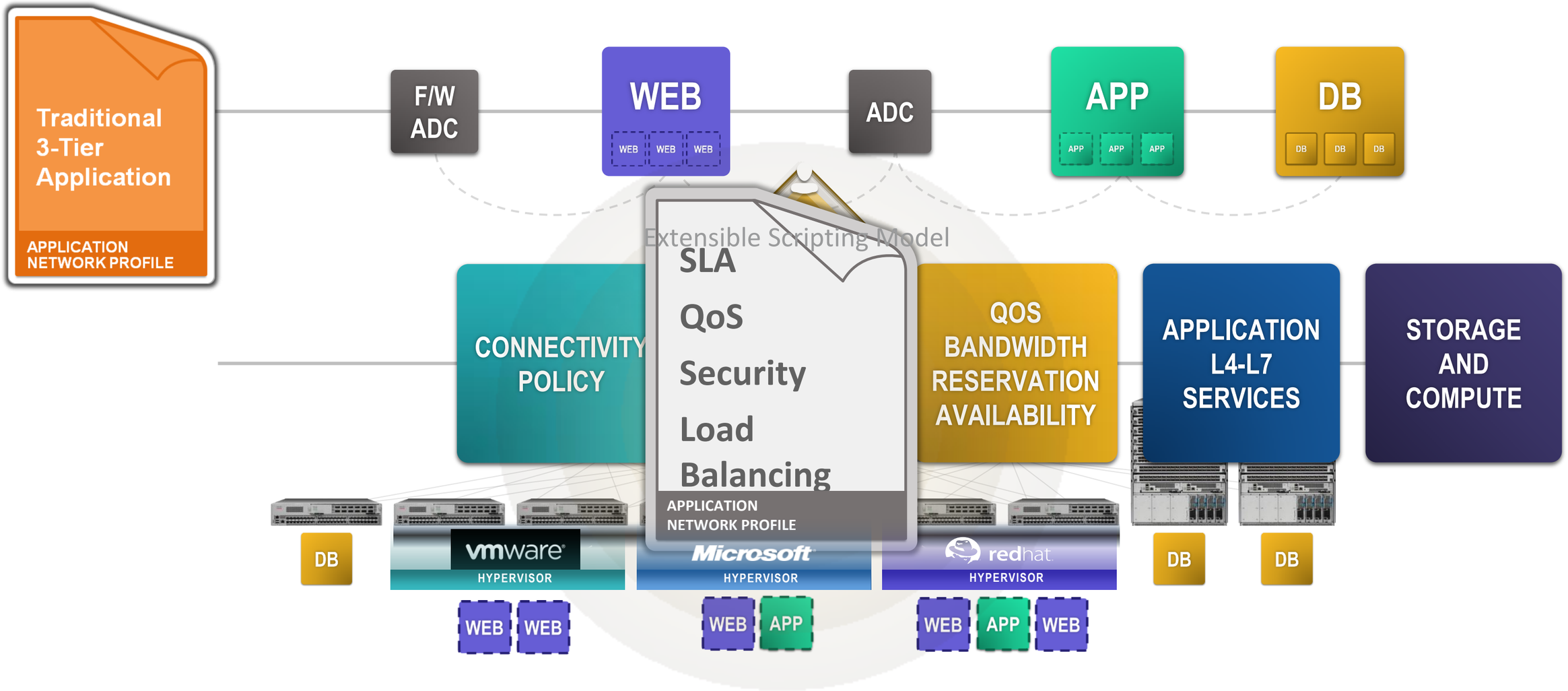
Cloud Admin

ACI Goal: Common Policy and Operations Framework

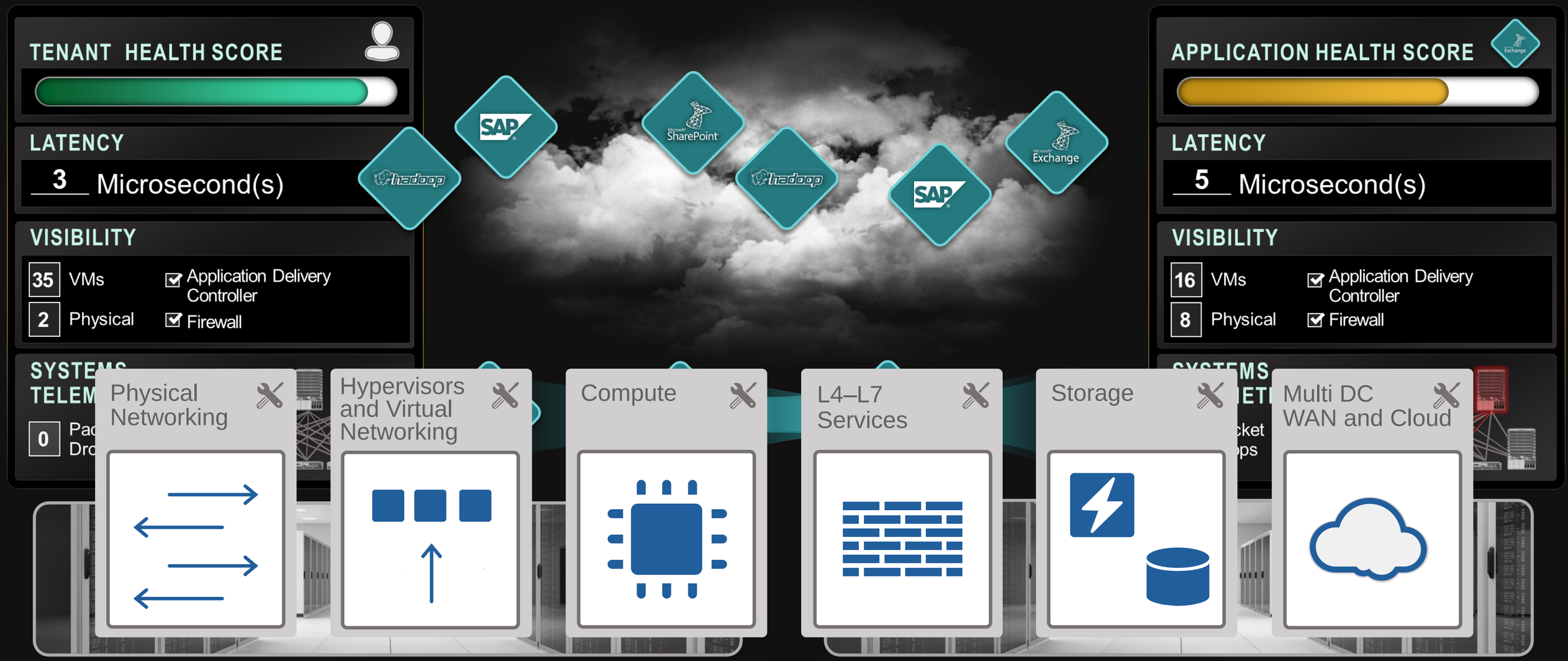


ACI: Any Application, Anywhere

Physical or Virtual Common Application Network Profile



ACI: RAPID DEPLOYMENT OF APPLICATIONS ONTO NETWORKS WITH SCALE, SECURITY AND FULL VISIBILITY



ENABLED BY PHYSICAL AND VIRTUAL INTEGRATION

Agenda

- Active-Active Data Centre: Business Drivers and Solutions Overview
- Active / Active Data Centre Design Considerations
  - Storage Extension
  - Data Centre Interconnect (DCI) - LAN Extension Deployment Scenarios
  - Host Mobility using LISP and OTV
  - Network Services and Applications (Path optimisation)
- Cisco ACI and Active / Active Data Centre
- Summary and Conclusions
- Q&A

Recommendations

1. Layer 2 extensions represent a challenge for optimal routing
2. Consider the implications of stretching the network and security services over multiple DCs
3. For live migration over long distances, when possible enable network path optimisation for traffic :
   - Client to server communication (Ingress Optimisation)
   - Server to Client communication for symmetrical return traffic (Egress optimisation)
   - Server to Server communication (bandwidth and Latency optimisation)
4. Otherwise provision enough bandwidth (2 times the needs) and compute the total latency due to ping-pong workflows
5. When moving a VM /Tier move all the framework
6. Network and security policies must be maintained




Q & A

Complete Your Online Session Evaluation

Give us your feedback and receive a Cisco Live 2014 Polo Shirt!

Complete your Overall Event Survey and 5 Session Evaluations.

- Directly from your mobile device on the Cisco Live Mobile App
- By visiting the Cisco Live Mobile Site www.ciscoliveaustralia.com/mobile
- Visit any Cisco Live Internet Station located throughout the venue

Polo Shirts can be collected in the World of Solutions on Friday 21 March 12:00pm - 2:00pm

Learn online with Cisco Live!

Visit us online after the conference for full access to session videos and presentations.
www.CiscoLiveAPAC.com


CISCO
TM