Cisco live!

What You Make Possible

TOMORROW starts here.

cisco

Designing Layer 2 Networks - Avoiding Loops, Drops, Flooding

BRKCRS-2661

TOMORROW starts here.

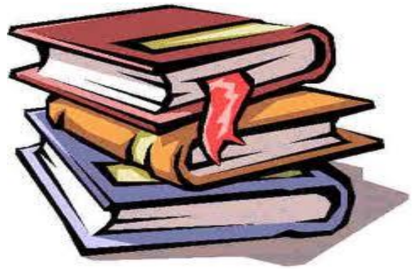
Abstract

Designing Layer 2 networks is easy.

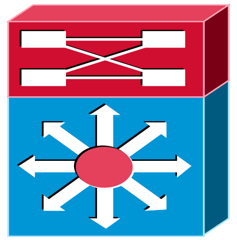
Apparently, in fact there are many traps and dependencies. Three issues of Layer 2 networks - loops, traffic drop and excessive flooding can be demanding. This session is to discuss and present how to avoid them with the standard design techniques or by new mechanisms.

Presentation Legend

Key Points



Reference Material



Standalone Multilayer Switch

Virtual Switching System

Layer 2 Link

Layer 3 Link

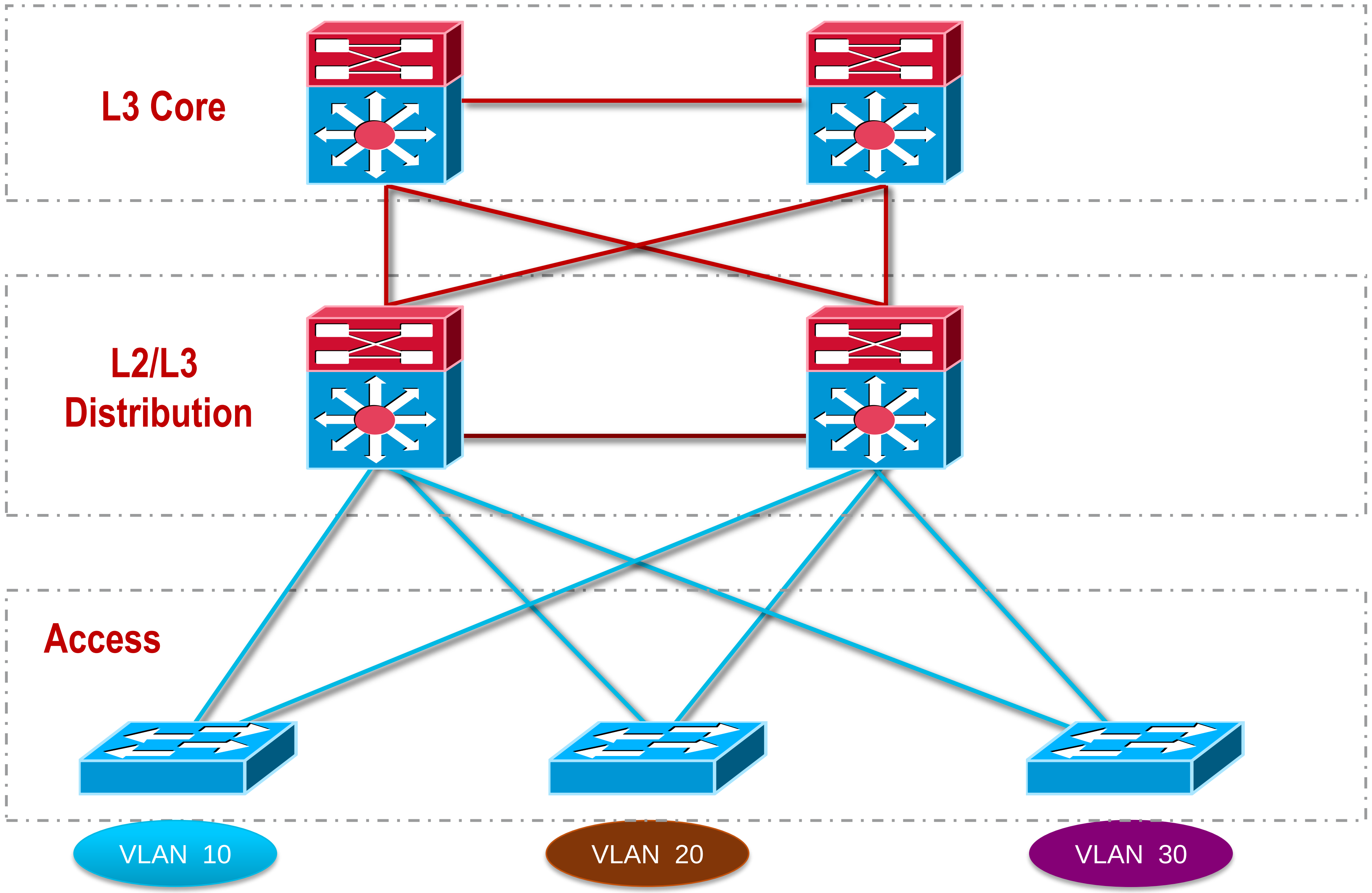
Agenda

- L2 Network Challenges
- Traditional Multilayer Designs
- Virtual Switching Systems (VSS) Designs
- Fabric Path Designs
- Summary

L2 Network Design Challenges

Traditional Multi-Layer Design

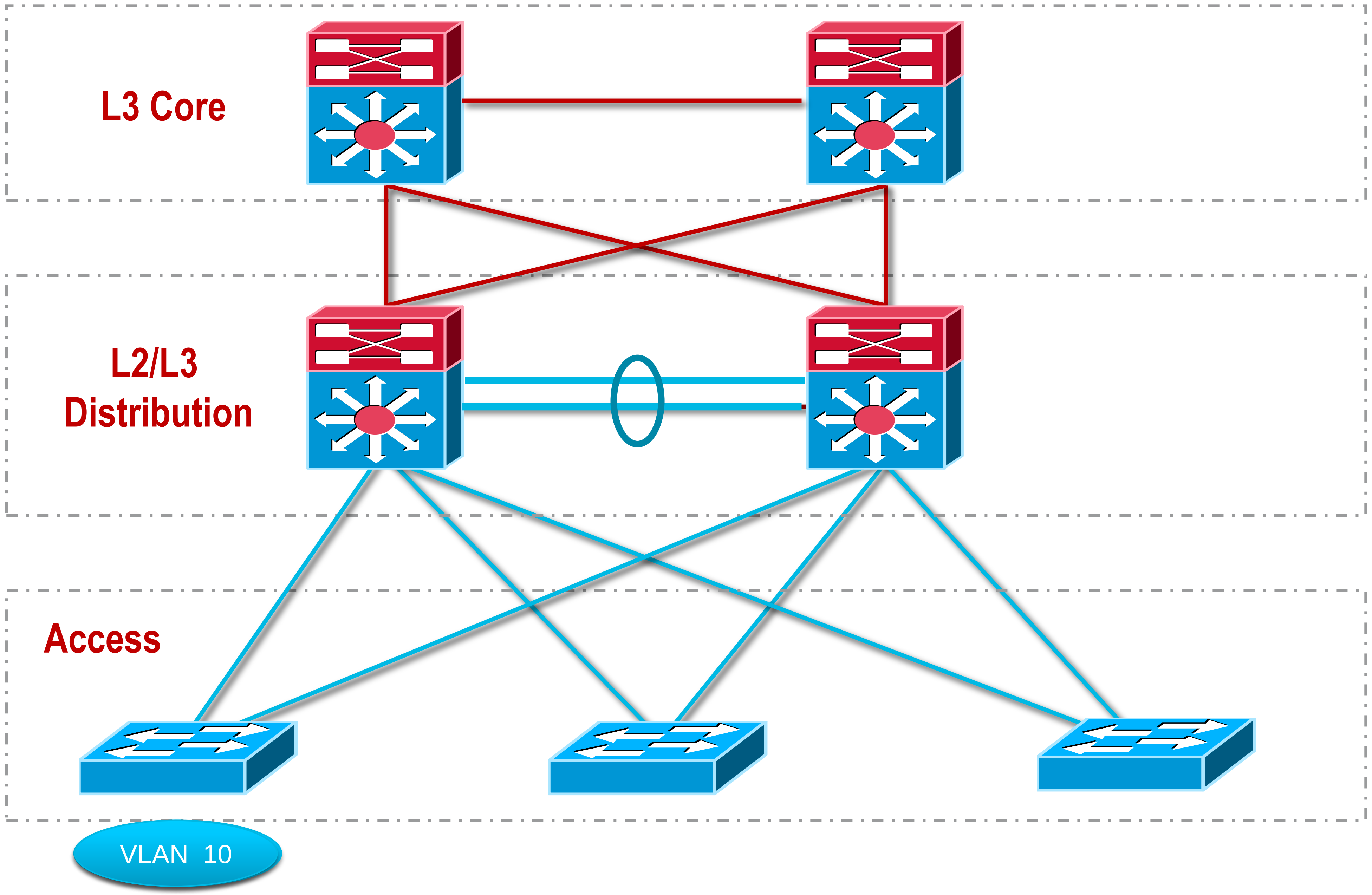
No L2 Loops



- One switch per subnet per vlan
- Simple design
- Limits L2 domain size to port density to size of the access switch

Traditional Mu

With L2 Loops

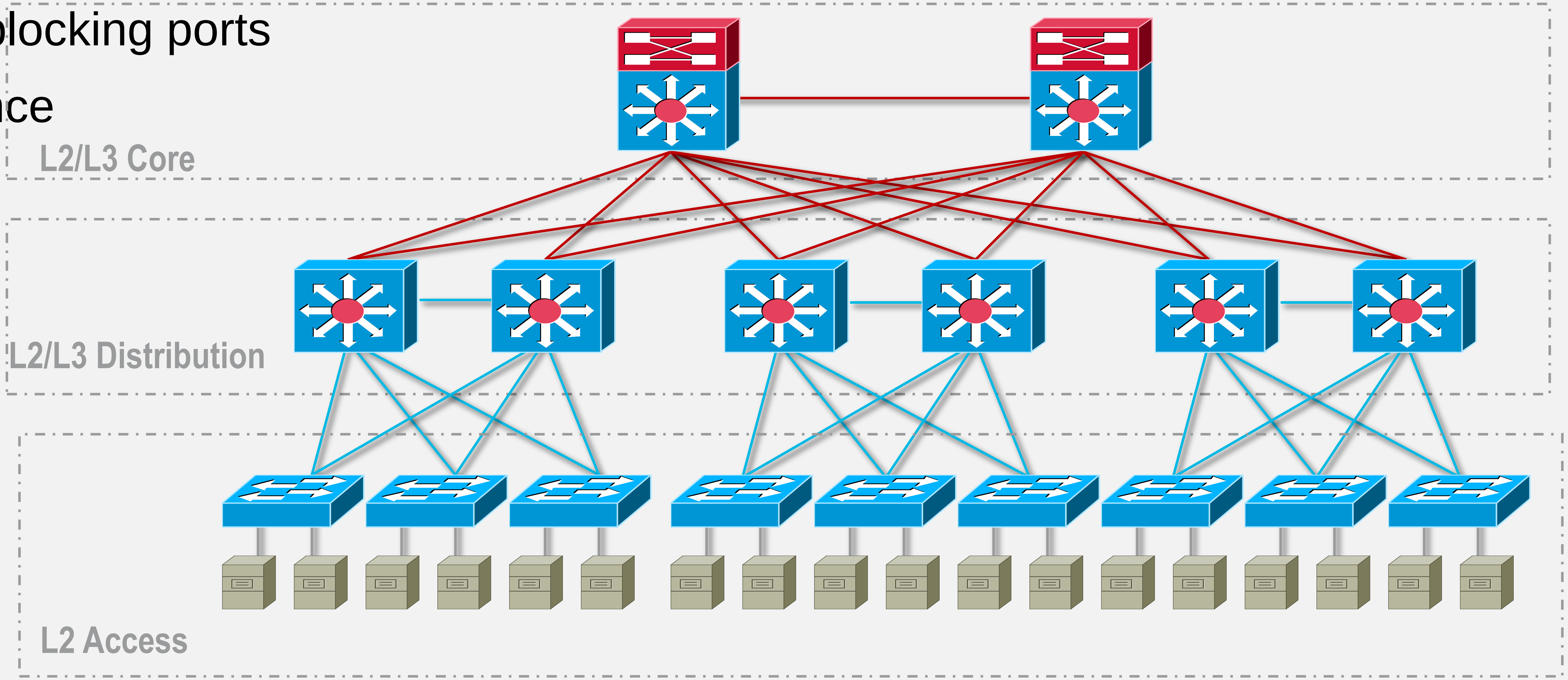


- Extending the L2 domain beyond the single switch
- Best practice says
  - Distribution link must be an L2 link
  - Redundant Links
- Now we have the loop

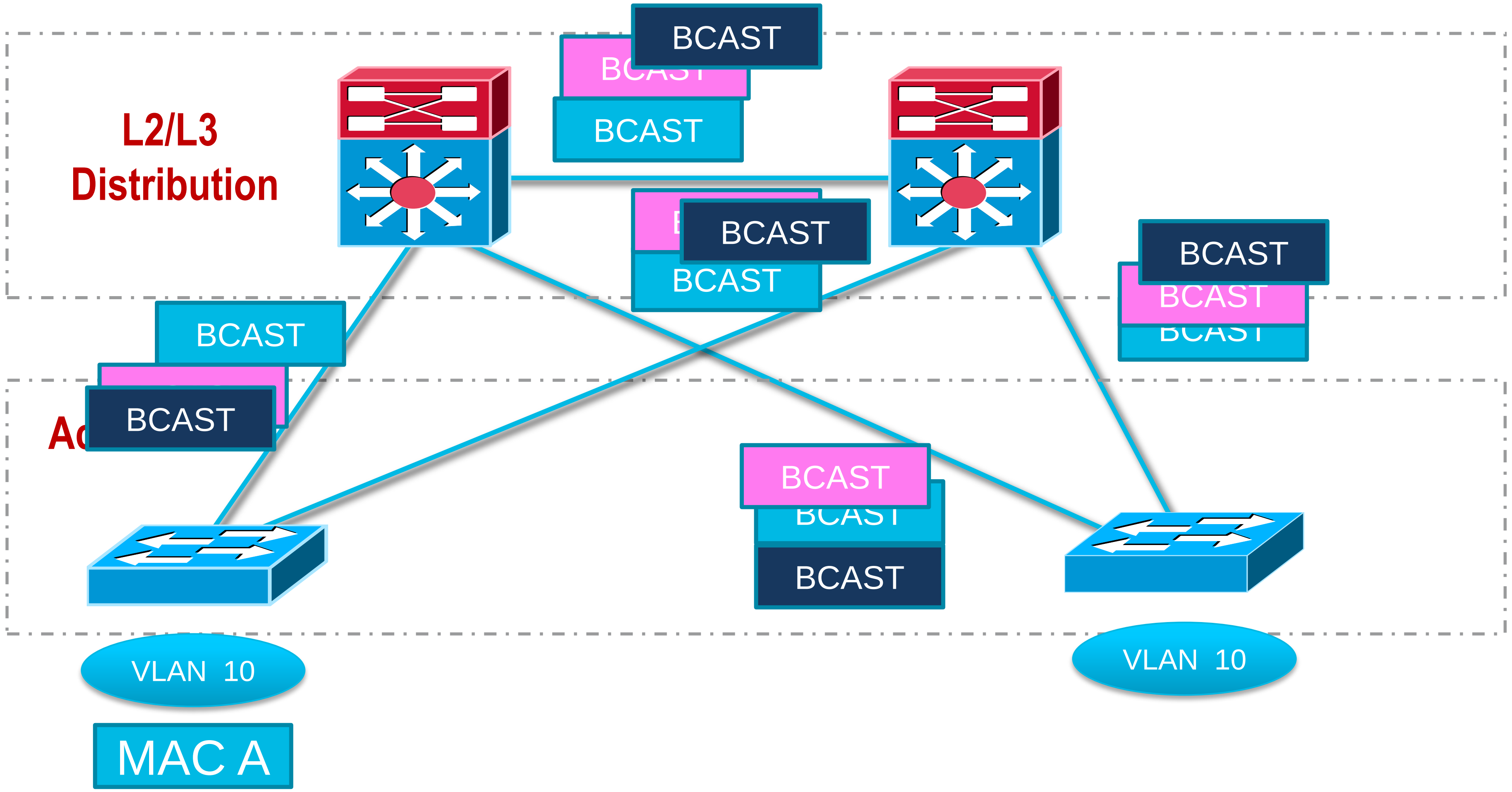
Current Network Challenges

Traditional Data Centre Multi-layer Design

- Extend L2 domains across distribution blocks
- Eliminate STP blocking ports
- Fast Convergence



L2 Loop – Whats the Problem ?



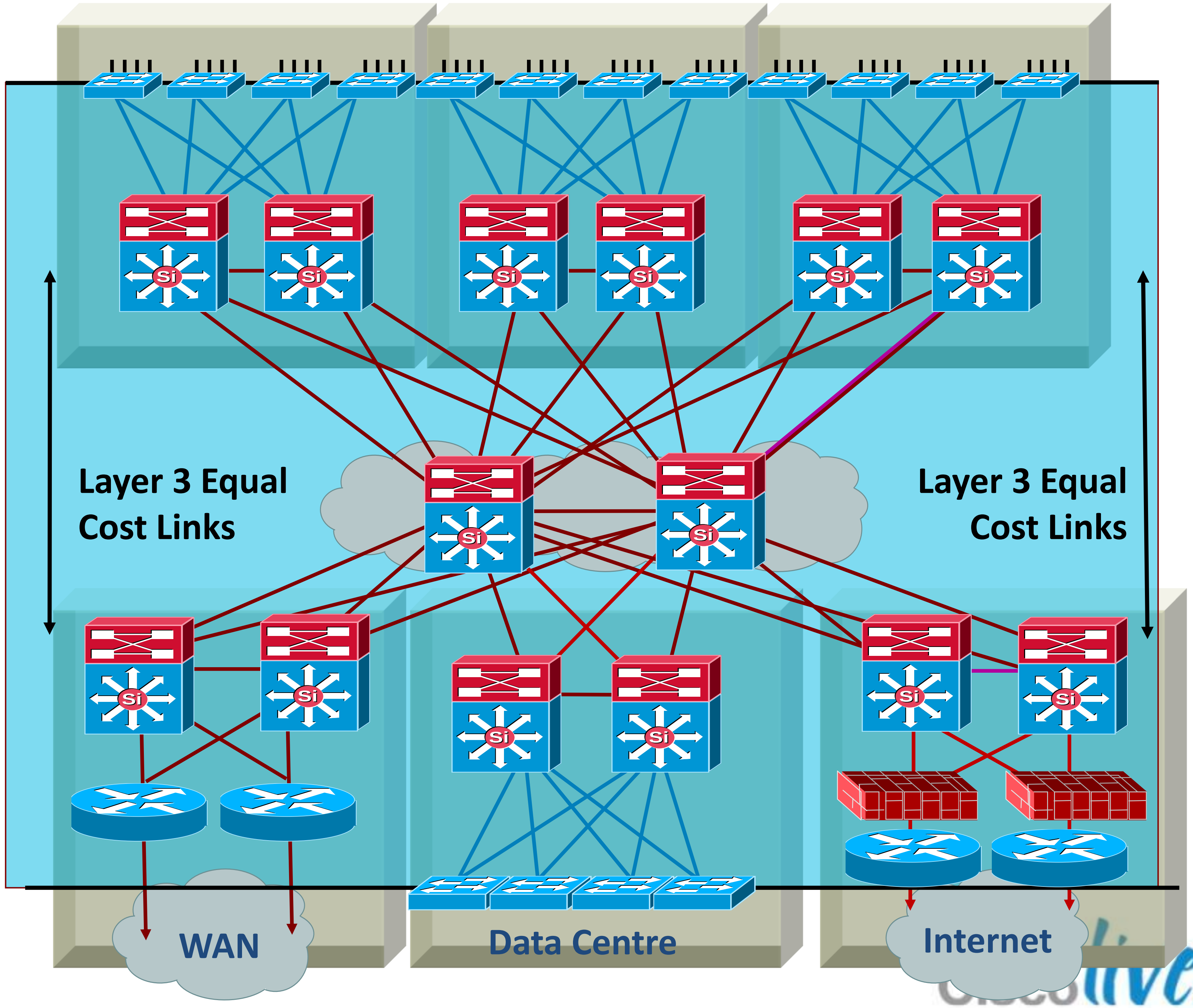
- Broadcast and multicast storm
- Source MAC address appear to be moving around as the MAC gets learned on different ports
- Traffic drops

Traditional Multi-Layer Designs

Best Practices

Layer 1 Physical Things

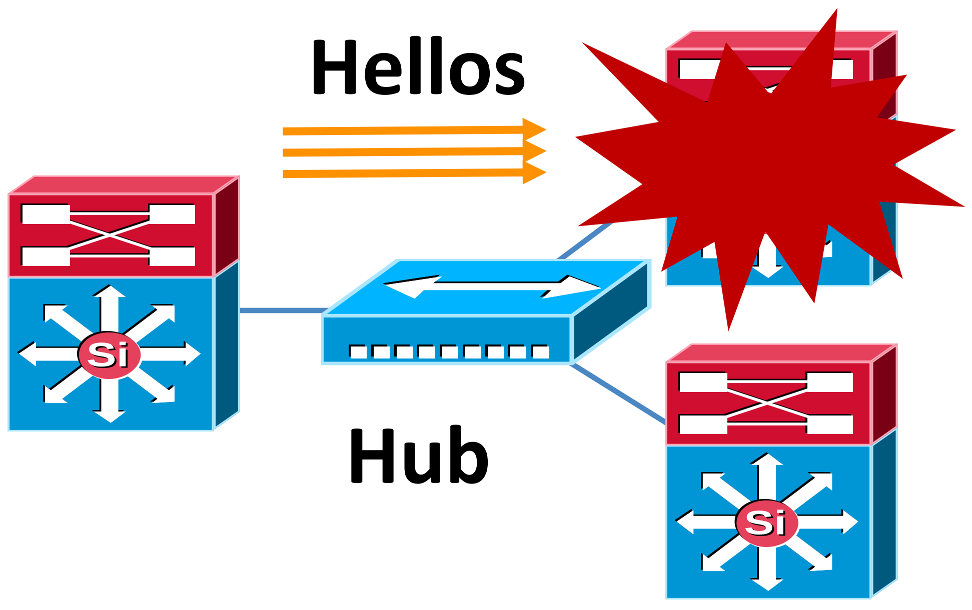
- Use point-to-point interconnections—no L2 aggregation points between nodes
- Use fibre for best convergence (debounce timer)
- Tune carrier delay timer
- Use configuration on the physical interface not VLAN/SVI when possible

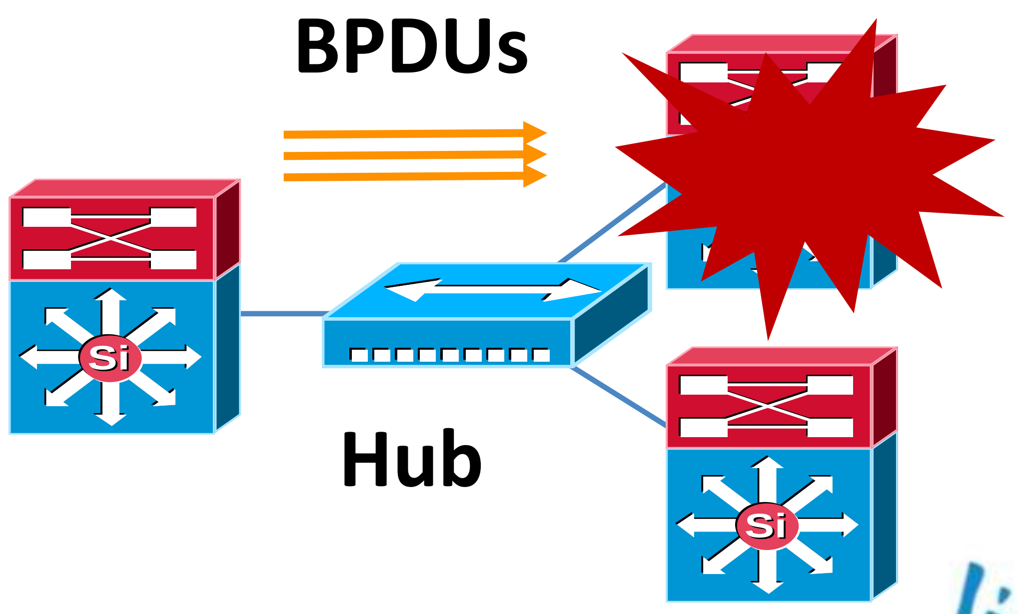


Redundancy and Protocol Interaction

Link Neighbour Failure Detection

- Indirect link failures are harder to detect
- With no direct HW notification of link loss or topology change convergence times are dependent on SW notification
- Indirect failure events in a bridged environment are detected by spanning tree hellos
- In certain topologies the need for TCN updates or dummy multicast flooding (uplink fast) is necessary for convergence
- You should not be using hubs in a high-availability design

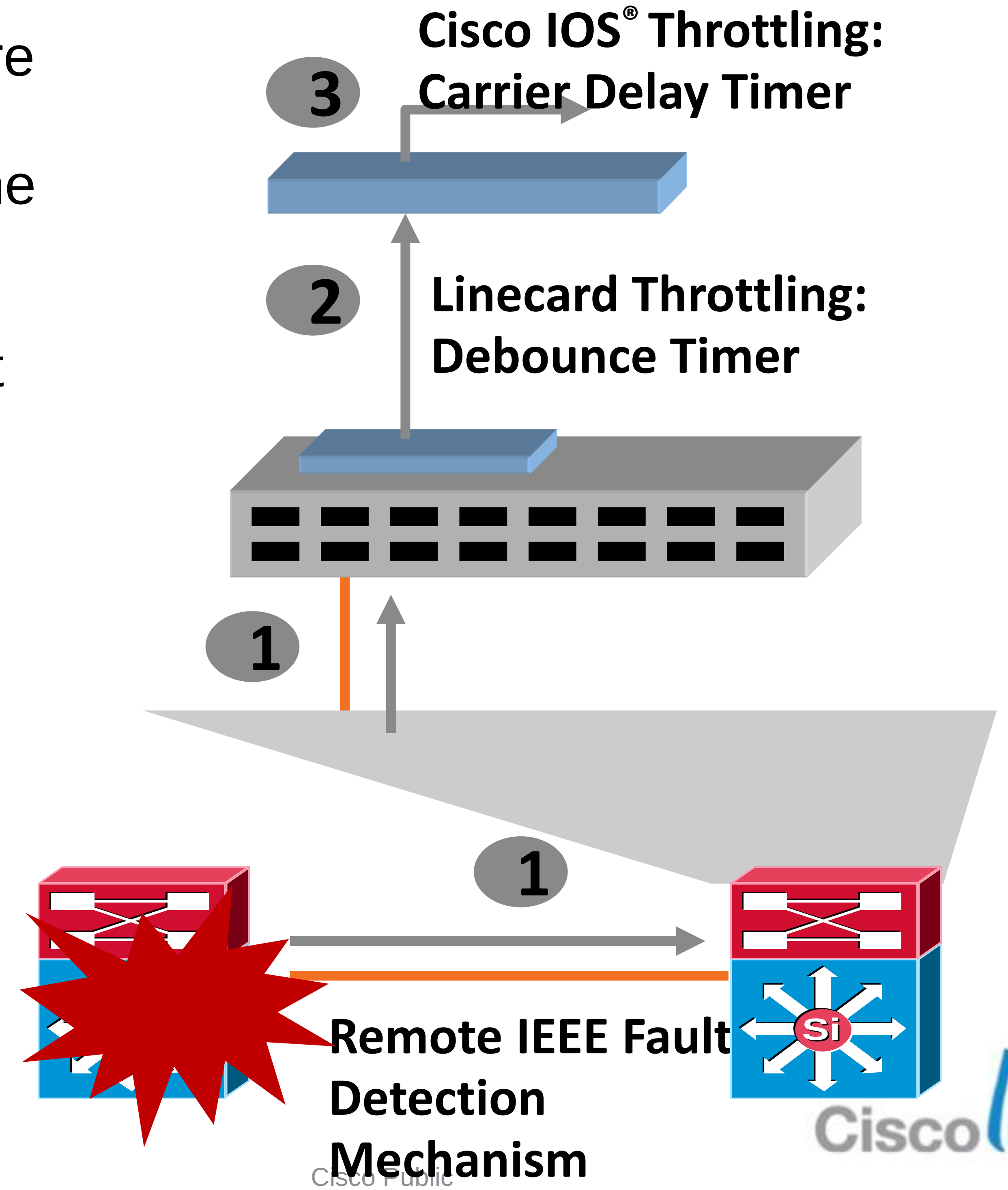




Redundancy and Protocol Interaction

Link Redundancy and Failure Detection

- Direct point-to-point fibre provides for fast failure detection
- IEEE 802.3z and 802.3ae link negotiation define the use of remote fault indicator and link fault signalling mechanisms
- Bit D13 in the Fast Link Pulse (FLP) can be set to indicate a physical fault to the remote side
- Do not disable auto-negotiation on GigE and 10GigE interfaces
- The default debounce timer on GigE and 10GigE fibre linecards is 10 msec
- The minimum debounce for copper is 300 msec
- Carrier-delay
  - 3560, 3750, and 4500—0 msec
  - 6500—leave it set at default



Redundancy and Protocol Interaction

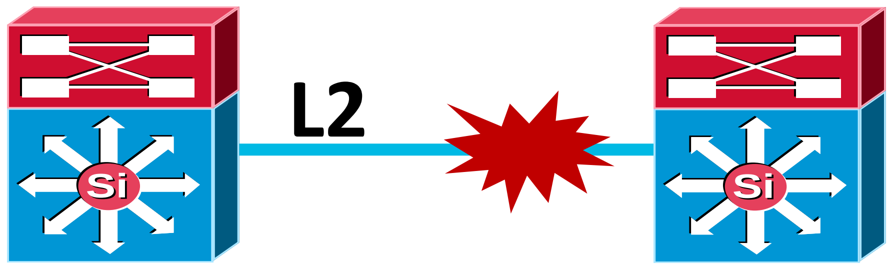
Layer 2 and 3—Why Use Routed Interfaces

Configuring L3 routed interfaces provides for faster convergence than an L2 switch port with an associated L3 SVI

1. Link Down
2. Interface Down
3. Routing Update

~ 8 msec loss

21:38:37.042 UTC: %LINEPROTO-5-UPDOWN: Line protocol on Interface GigabitEthernet3/1, changed state to down
21:38:37.050 UTC: %LINK-3-UPDOWN: Interface GigabitEthernet3/1, changed state to down
21:38:37.050 UTC: IP-EIGRP(Default-IP-Routing-Table:100): Callback: route_adjust GigabitEthernet3/1



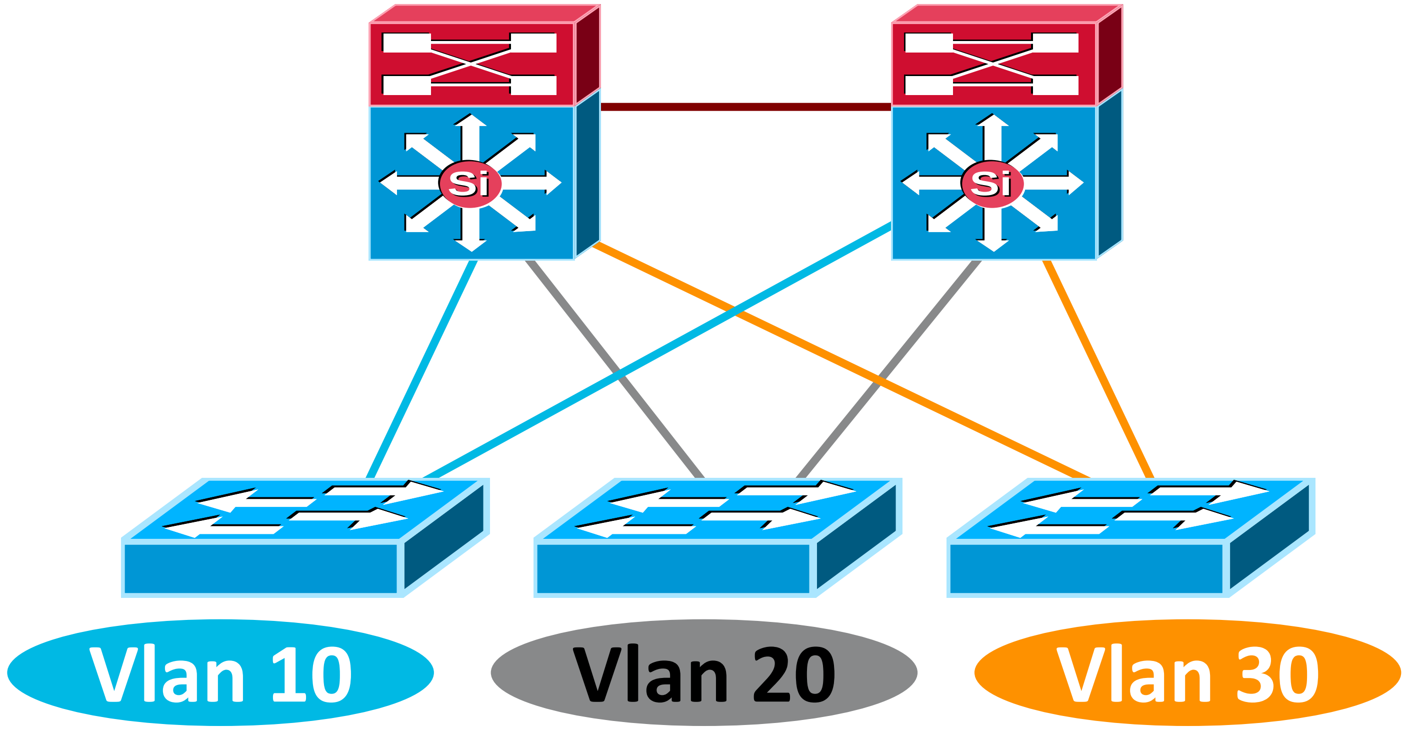
1. Link Down
2. Interface Down
3. Autostate
4. SVI Down
5. Routing Update

~ 150–200 msec loss

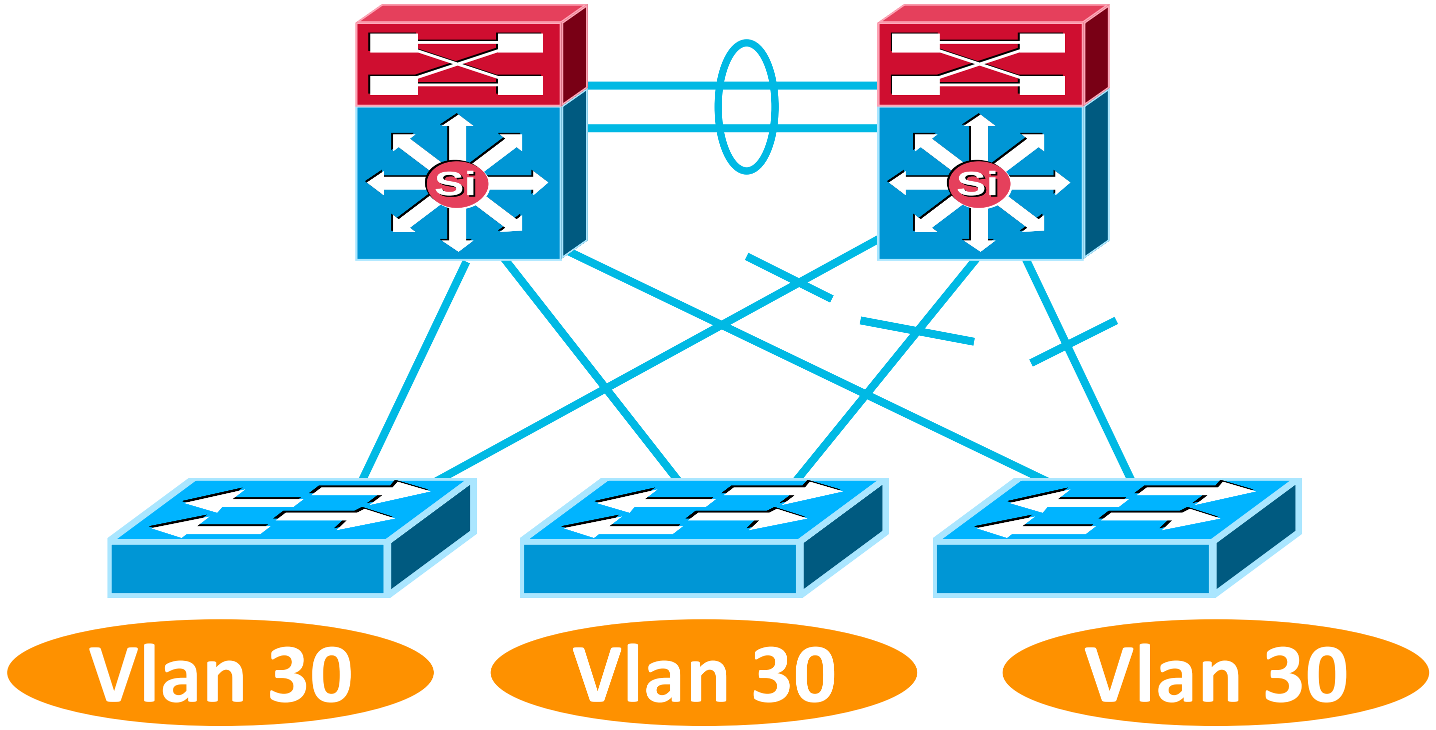
21:32:47.813 UTC: %LINEPROTO-5-UPDOWN: Line protocol on Interface GigabitEthernet2/1, changed state to down
21:32:47.821 UTC: %LINK-3-UPDOWN: Interface GigabitEthernet2/1, changed state to down
21:32:48.069 UTC: %LINK-3-UPDOWN: Interface Vlan301, changed state to down
21:32:48.069 UTC: IP-EIGRP(Default-IP-Routing-Table:100): Callback: route, adjust Vlan301

Multilayer Network Design

Layer 2 Access with Layer 3 Distribution



- Each access switch has unique VLANs
- No Layer 2 loops
- Layer 3 link between distribution
- No blocked links

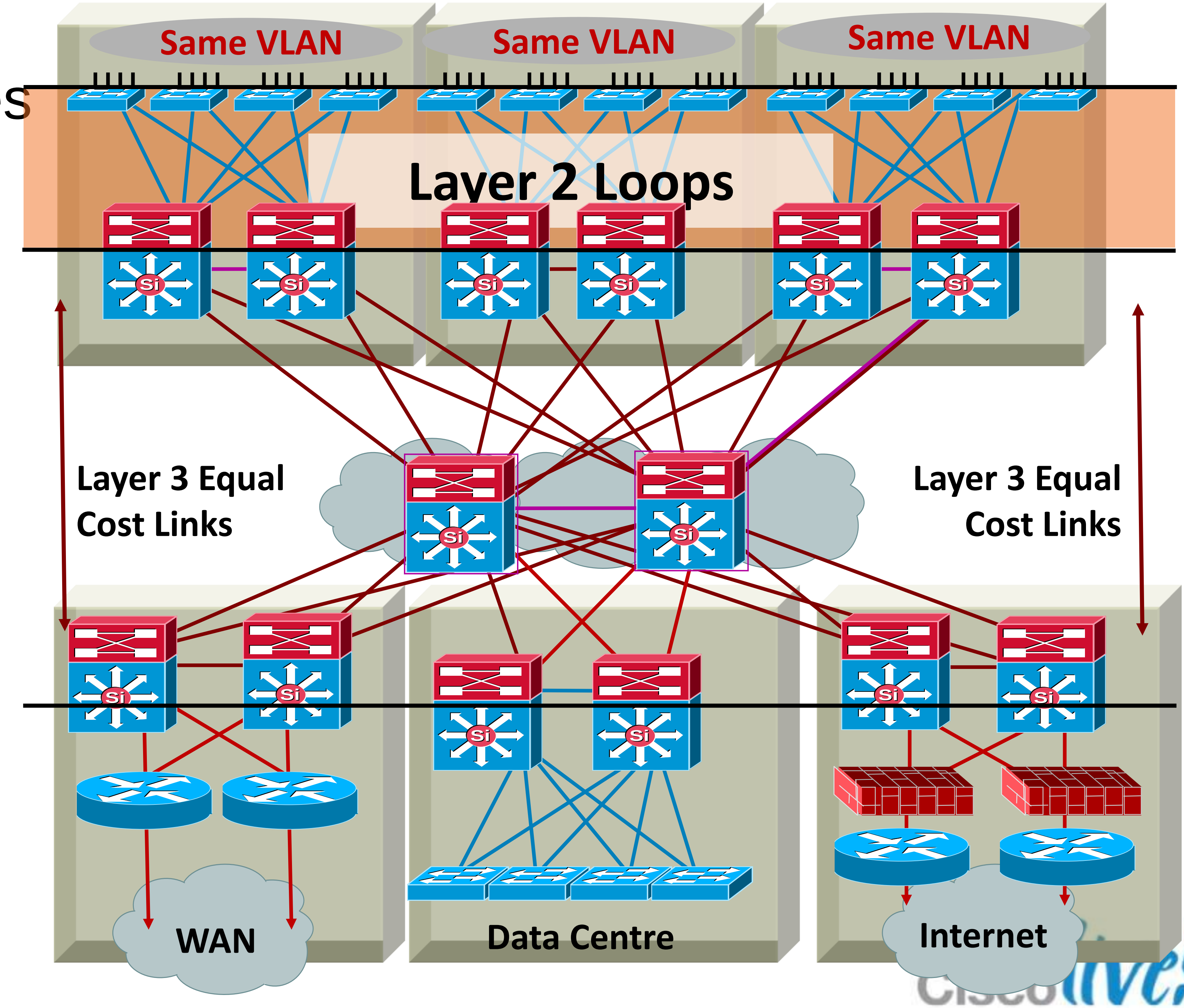


- At least some VLANs span multiple access switches
- Layer 2 loops
- Layer 2 and 3 running over link between distribution
- Blocked links

Best Practices

Spanning Tree Configuration

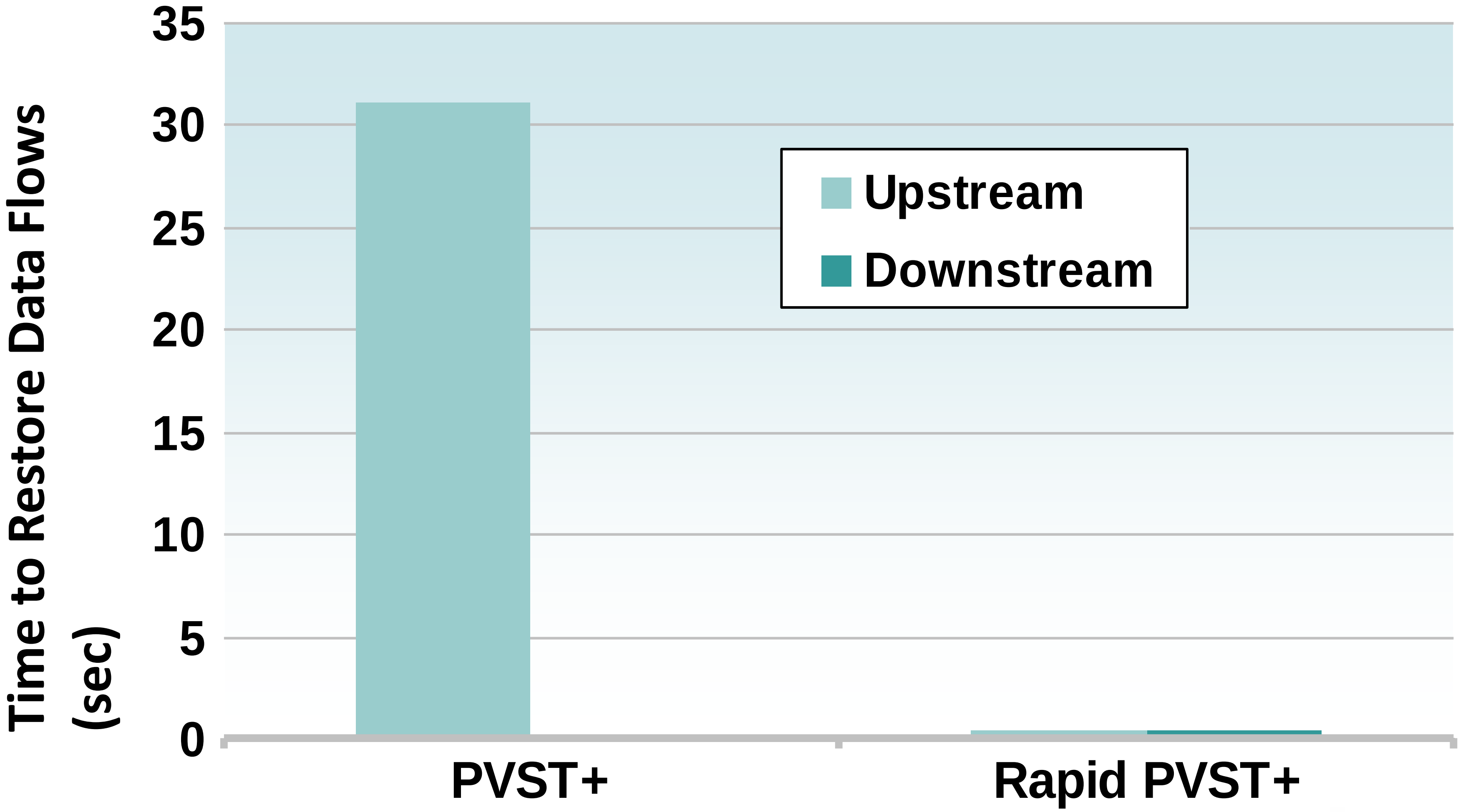
- **Only** span VLAN across multiple access layer switches when you have to!
- Use rapid PVST+ for best convergence
- Required to protect against user side loops
- Required to protect against operational accidents (misconfiguration or hardware failure)
- Take advantage of the spanning tree toolkit



Optimising L2 Convergence

PVST+, Rapid PVST+ or MST

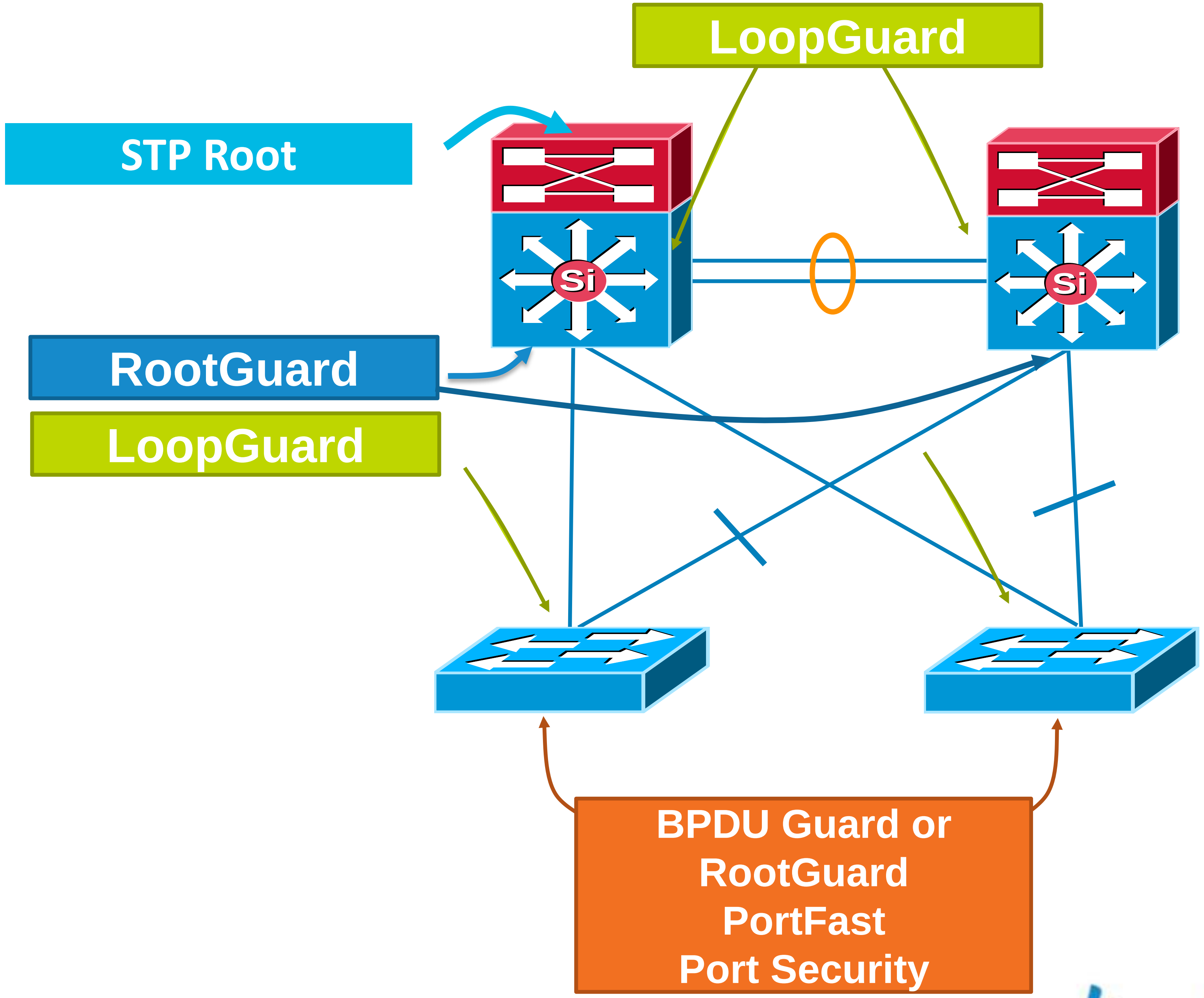
- Rapid-PVST+ greatly improves the restoration times for any VLAN that requires a topology convergence due to link UP
- Rapid-PVST+ also greatly improves convergence time over backbone fast for any indirect link failures
- PVST+ (802.1d)
  - –Traditional spanning tree implementation
- Rapid PVST+ (802.1w)
  - –Scales to large size (~10,000 logical ports)
  - –**Easy to implement, proven, scales**
- MST (802.1s)
  - –Permits very large scale STP implementations (~30,000 logical ports)
  - –**Not as flexible as rapid PVST+**



Layer 2 Hardening

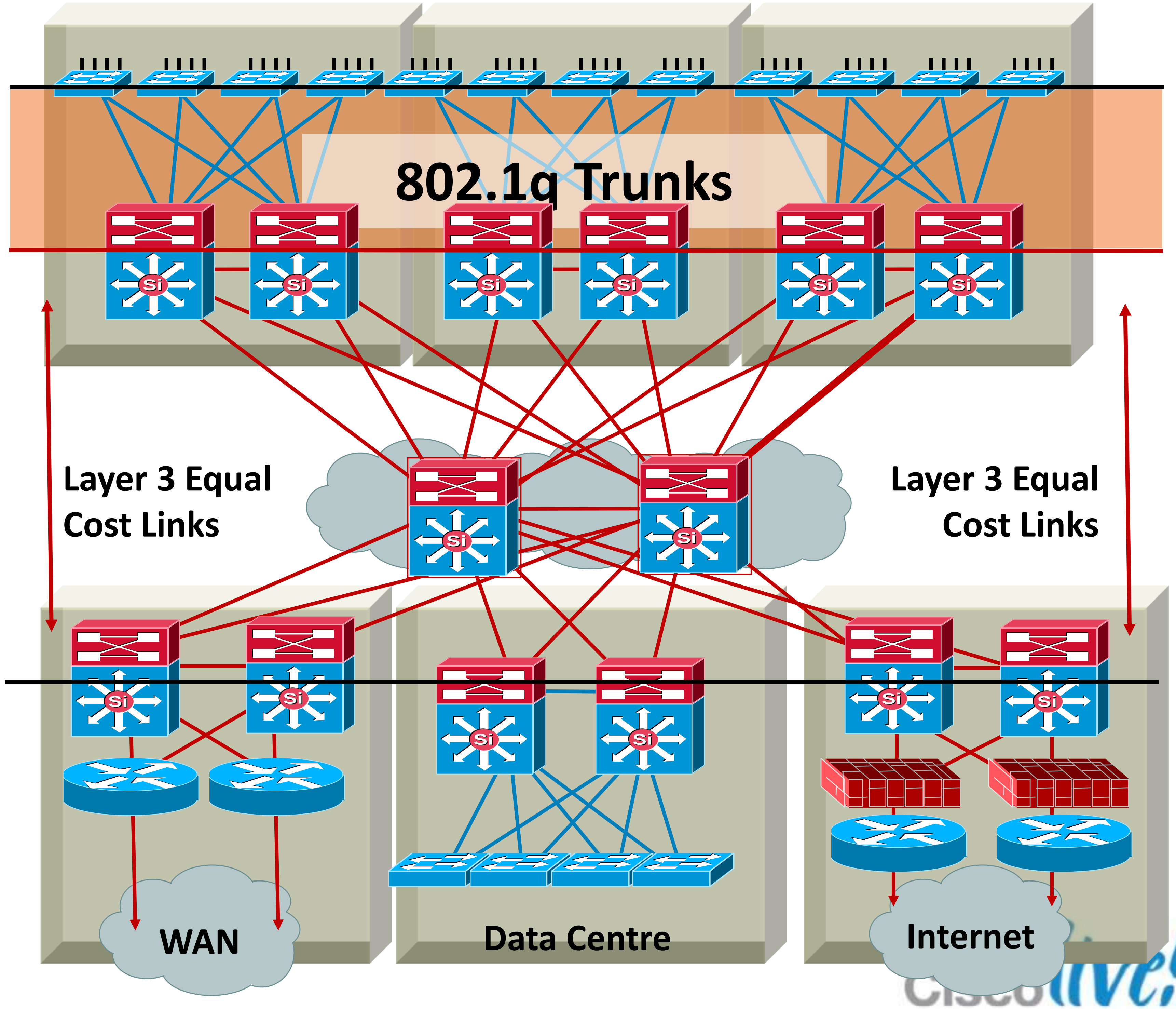
Spanning Tree Should Behave the Way You Expect

- Place the root where you want it
  - Root primary/secondary macro
- The root bridge should stay where you put it
  - RootGuard
  - LoopGuard
  - UplinkFast
  - UDLD
- Only end-station traffic should be seen on an edge port
  - BPDU Guard
  - RootGuard
  - PortFast
  - Port-security



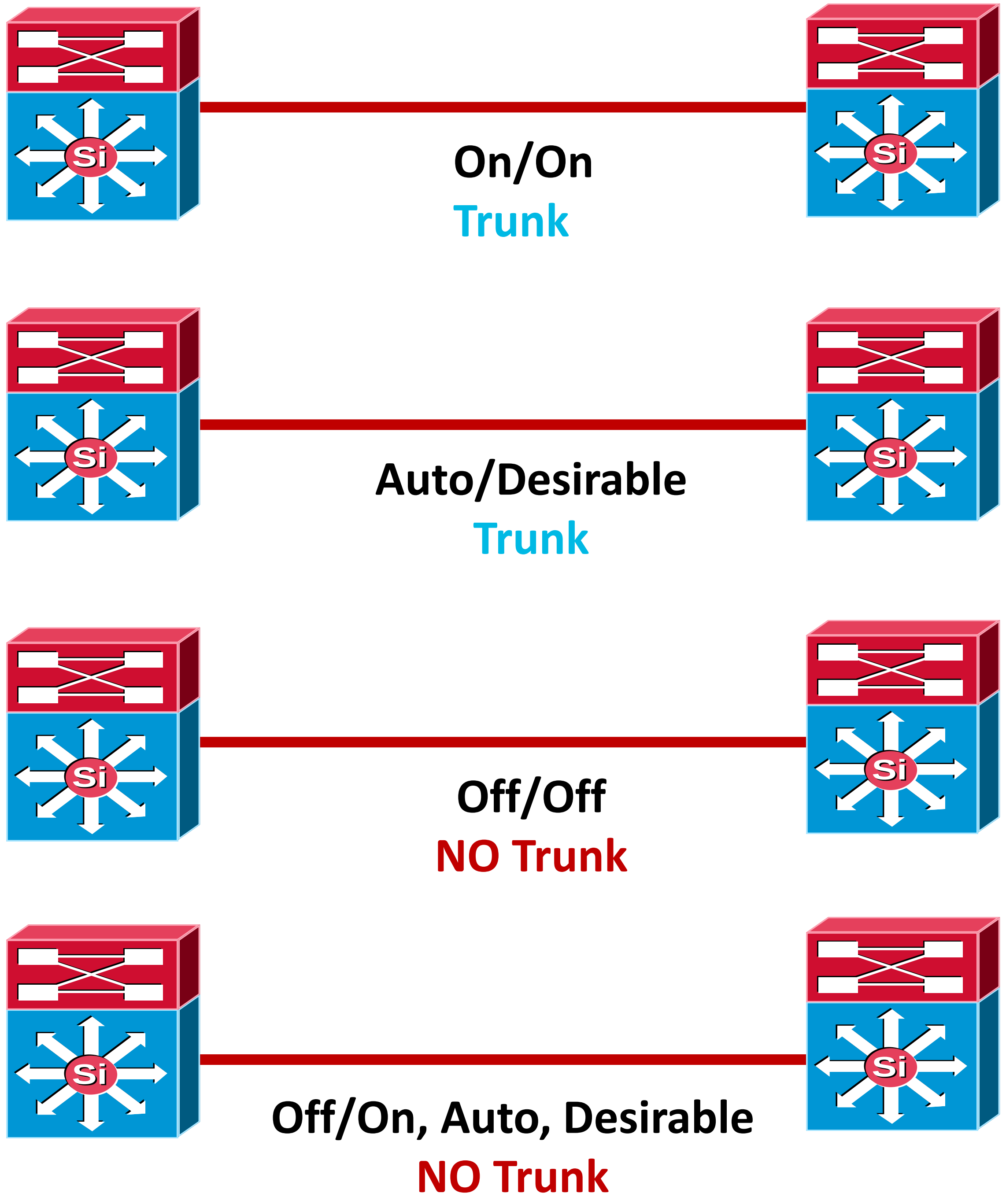
Best Practices—Trunk Configuration

- Typically deployed on interconnection between access and distribution layers
- Use VTP transparent mode to decrease potential for operational error
- Hard set trunk mode to on and encapsulation negotiate off for optimal convergence
- Change the native VLAN to something unused to avoid VLAN hopping
- Manually prune all VLANS except those needed
- Disable on host ports:
  - Cisco IOS: `switchport host`



DTP Dynamic Trunk Protocol

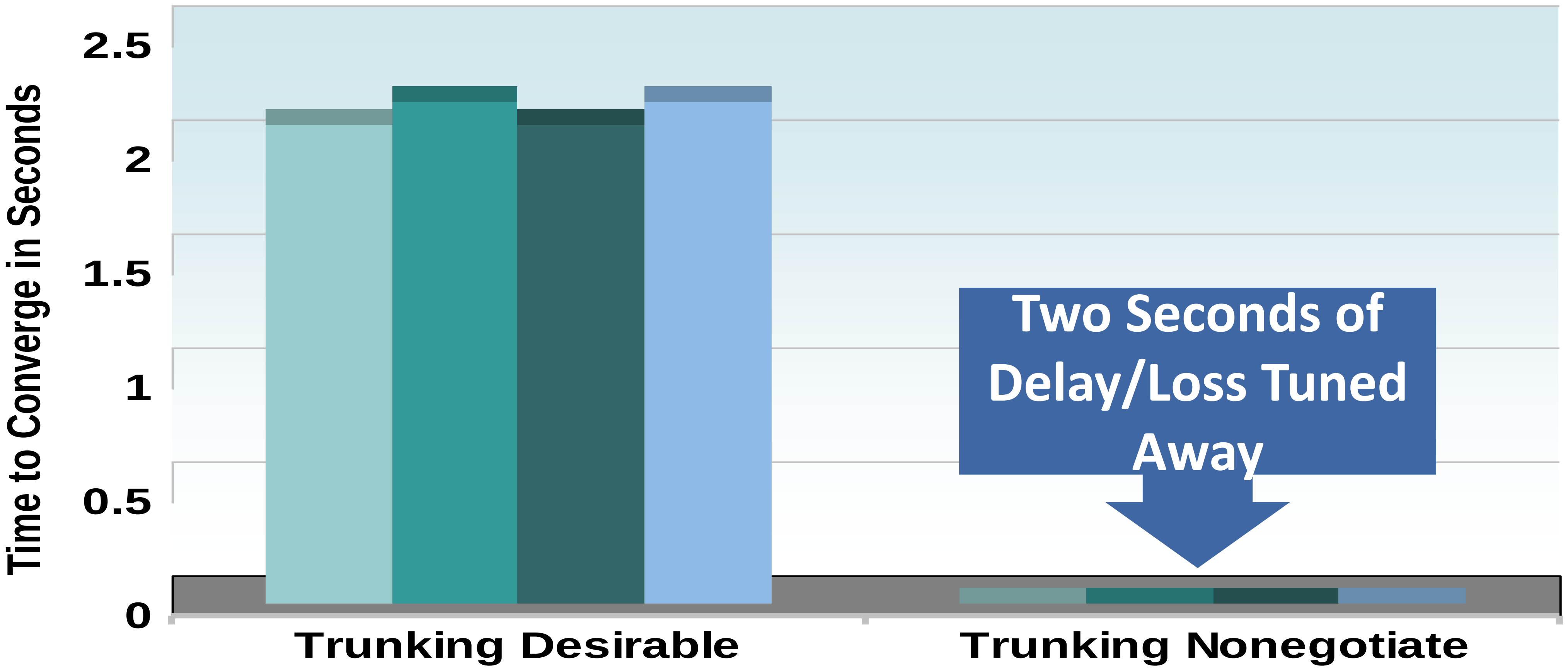
- Automatic formation of trunked switch-to-switch interconnection
  - On: always be a trunk
  - Desirable: ask if the other side can/will
  - Auto: if the other sides asks I will
  - Off: don't become a trunk
- Negotiation of 802.1Q or ISL encapsulation
  - ISL: try to use ISL trunk encapsulation
  - 802.1q: try to use 802.1q encapsulation
  - Negotiate: negotiate ISL or 802.1q encapsulation with peer
  - Non-negotiate: always use encapsulation that is hard set

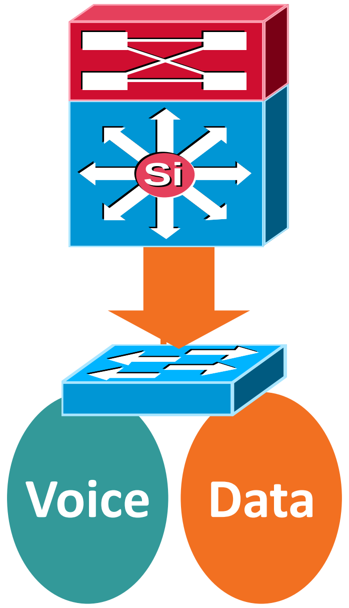


Optimising Convergence: Trunk Tuning

Trunk Auto/Desirable Takes Some Time

- DTP negotiation tuning improves link up convergence time
  - IOS(config-if)# switchport mode trunk
  - IOS(config-if)# switchport nonegotiate





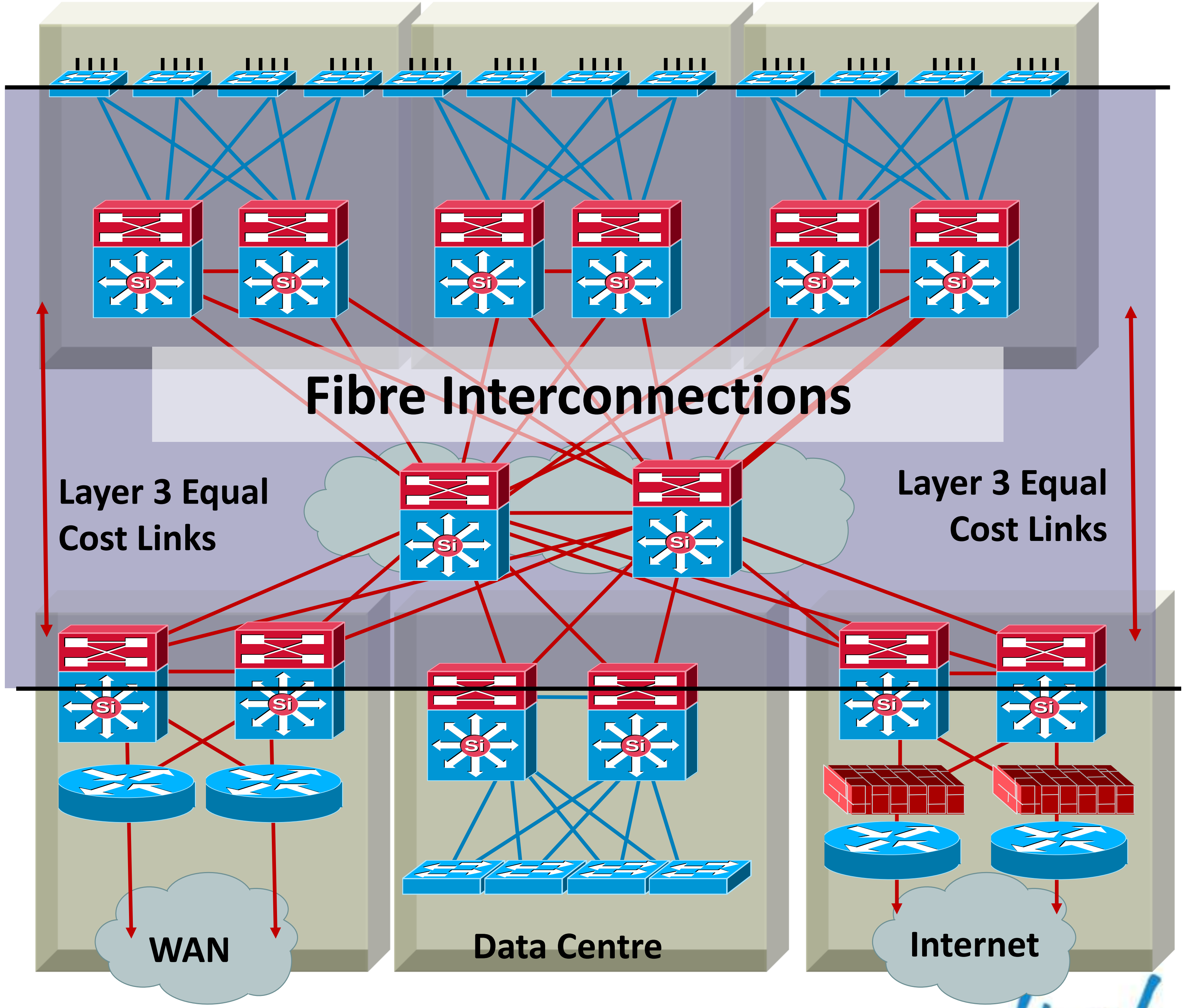
Trunking/VTP/DTP—Quick Summary

- VTP transparent should be used; there is a trade off between administrative overhead and the temptation to span existing VLANS across multiple access layer switches
- One can consider a configuration that uses DTP **ON/ON** and **NO NEGOTIATE**; there is a trade off between performance/HA impact and maintenance and operations implications
- An **ON/ON** and **NO NEGOTIATE** configuration is faster from a link up (restoration) perspective than a desirable/desirable alternative. However, in this configuration DTP is not actively monitoring the state of the trunk and a misconfigured trunk is not easily identified
- It's really a balance between fast convergence and your ability to manage configuration and change control …

Best Practices—UDLD Configuration

- Typically deployed on any fibre optic interconnection
- Use UDLD aggressive mode for most aggressive protection
- Turn on in global configuration to avoid operational error/misses
- Config example
  – Cisco IOS:

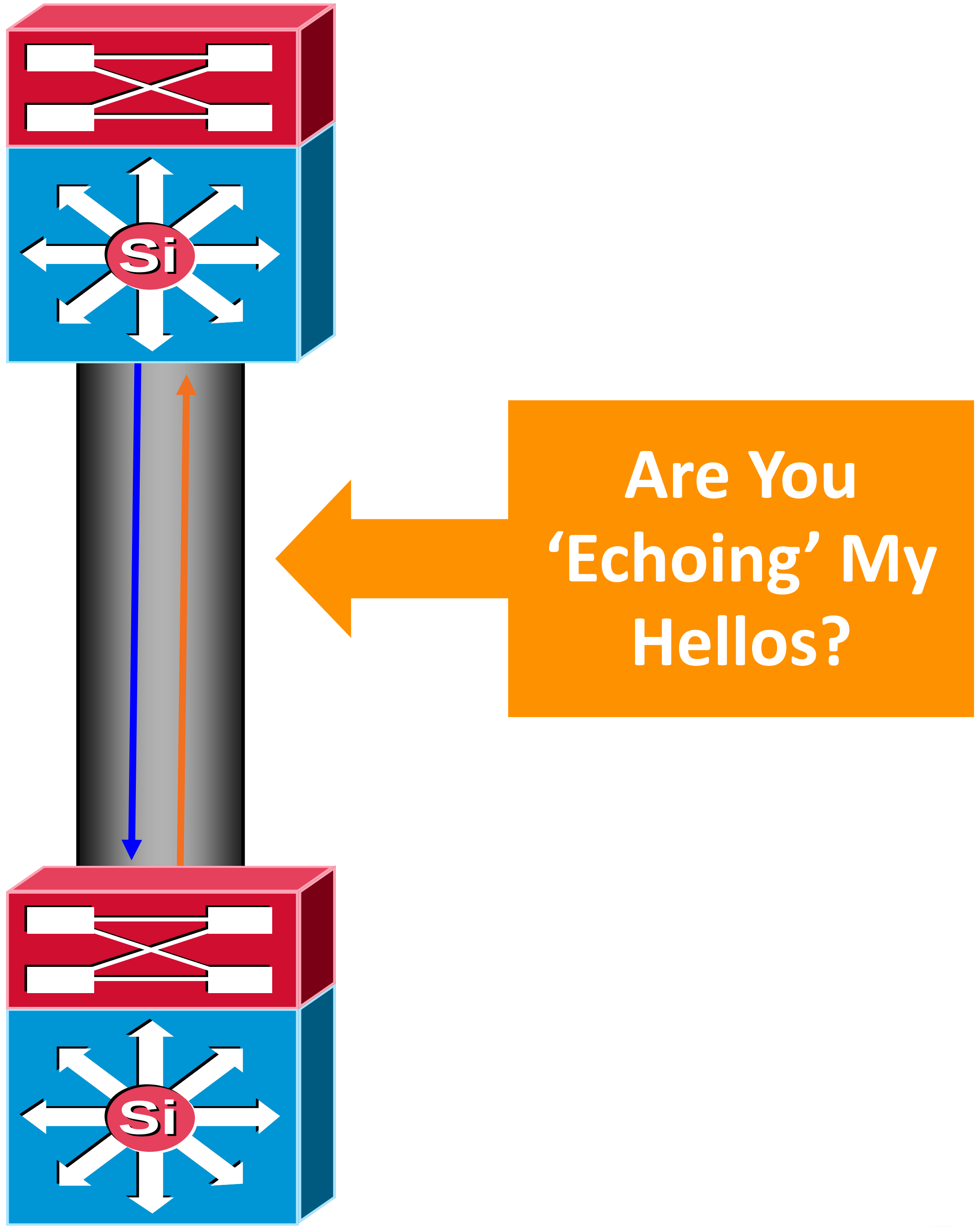
```
udld aggressive
```



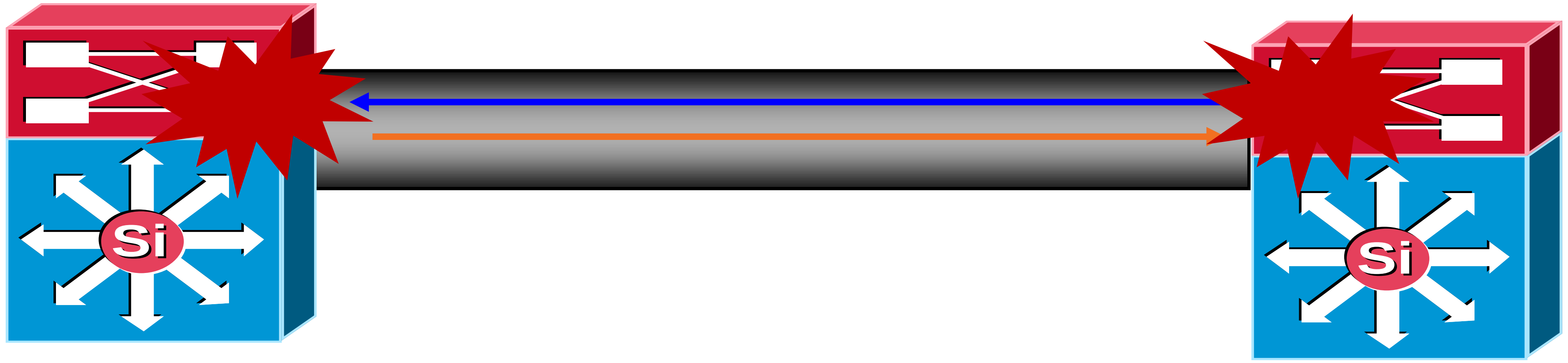
Unidirectional Link Detection

Protecting Against One-Way Communication

- Highly-available networks require UDLD to protect against one-way communication or partially failed links and the effect that they could have on protocols like STP and RSTP
- Primarily used on fibre optic links where patch panel errors could cause link up/up with mismatched transmit/receive pairs
- Each switch port configured for UDLD will send UDLD protocol packets (at L2) containing the port's own device/port ID, and the neighbour's device/port IDs seen by UDLD on that port
- Neighbouring ports should see their own device/port ID (echo) in the packets received from the other side
- If the port does not see its own device/port ID in the incoming UDLD packets for a specific duration of time, the link is considered unidirectional and is shutdown



UDLD Aggressive and UDLD Normal

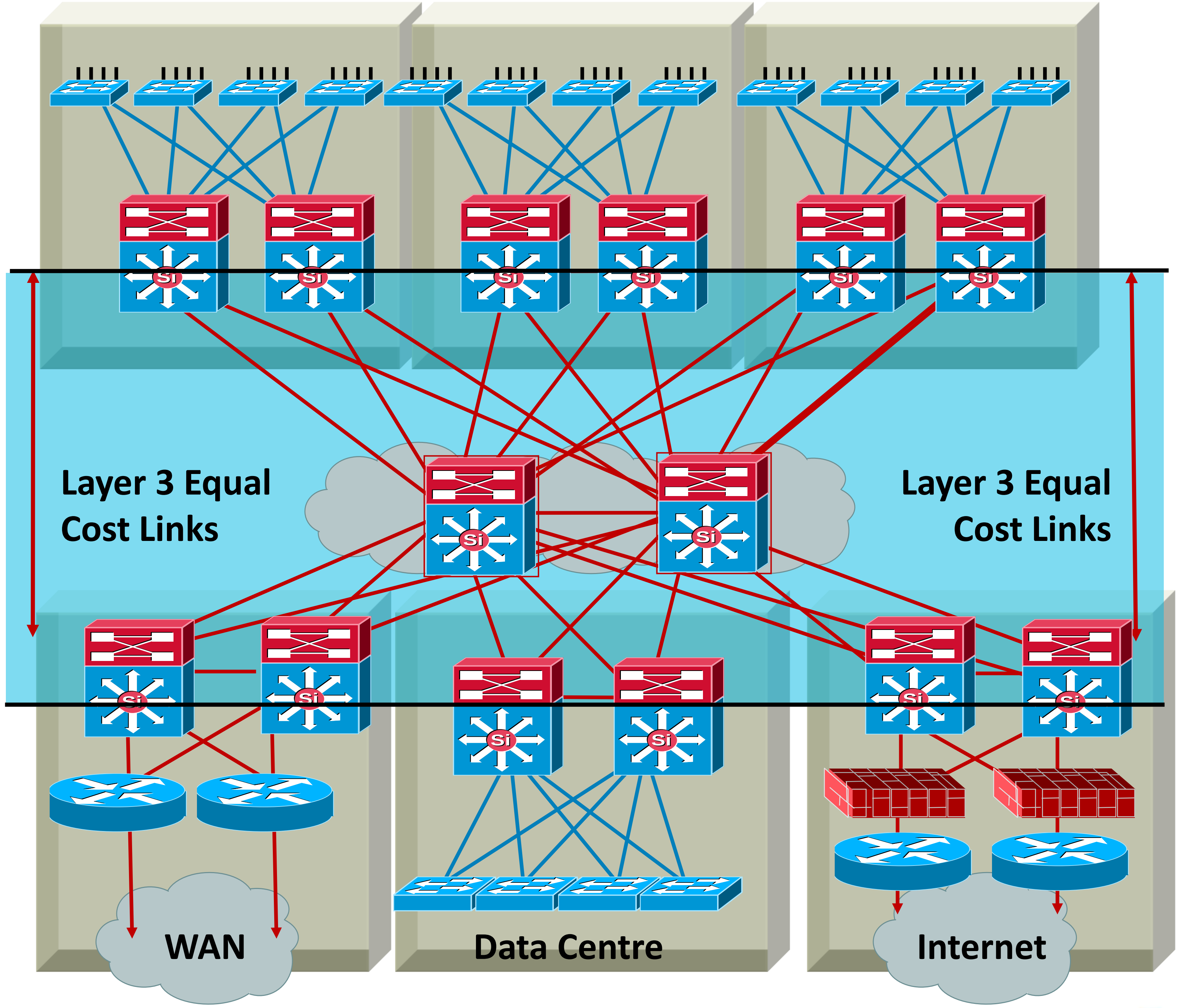


- Timers are the same—15-second hellos by default
- Aggressive Mode—after aging on a previously bi-directional link—tries eight times (once per second) to reestablish connection then err-disables port
- UDLD—Normal Mode—only err-disable the end where UDLD detected other end just sees the link go down
- UDLD—Aggressive—err-disable **both** ends of the connection due to err-disable when aging and re-establishment of UDLD communication fails

Best Practices

EtherChannel Configuration

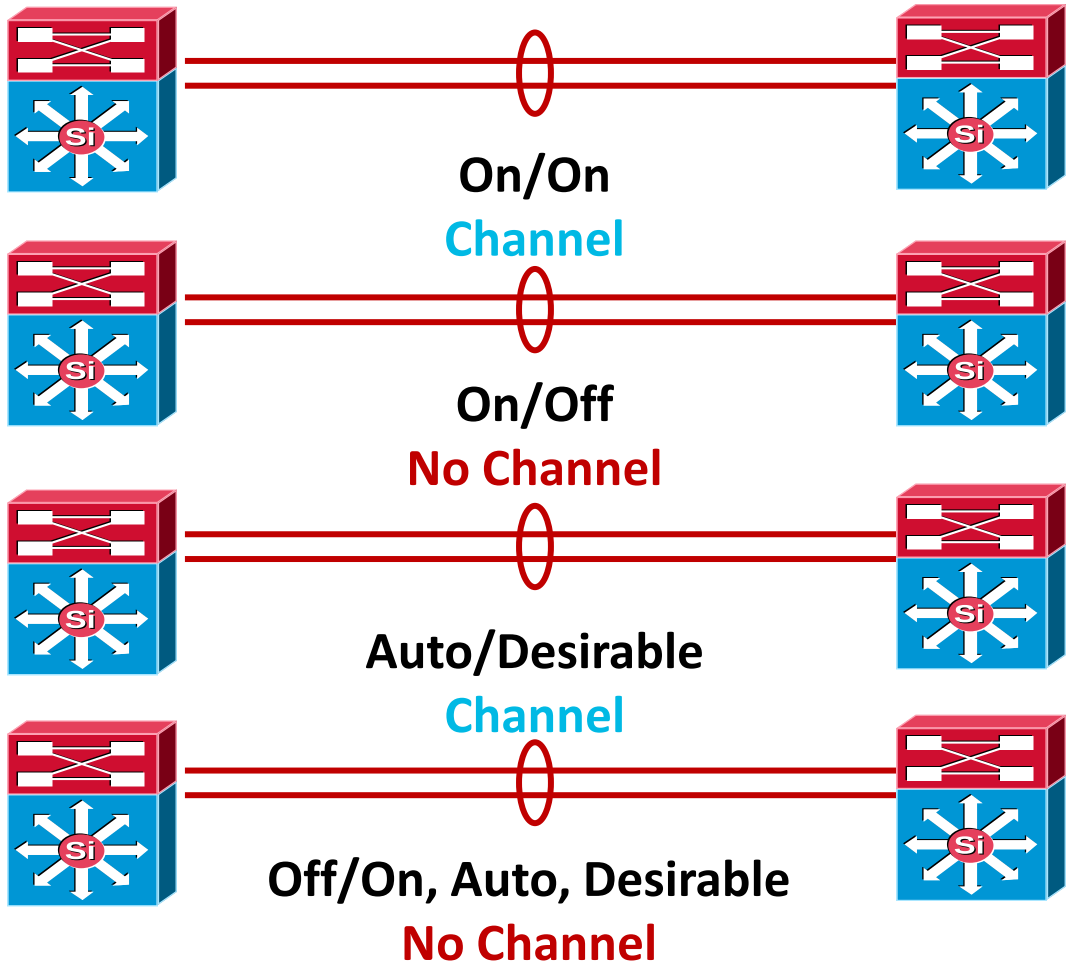
- Typically deployed in distribution to core, and core to core interconnections
- Used to provide link redundancy—while reducing peering complexity
- Tune L3/L4 load balancing hash to achieve maximum utilisation of channel members
- Deploy in powers of two (two, four, or eight)
- Match CatOS and Cisco IOS PAgP settings
- 802.3ad LACP for interop if you need it
- Disable unless needed
  - Cisco IOS: `switchport host`



Understanding EtherChannel

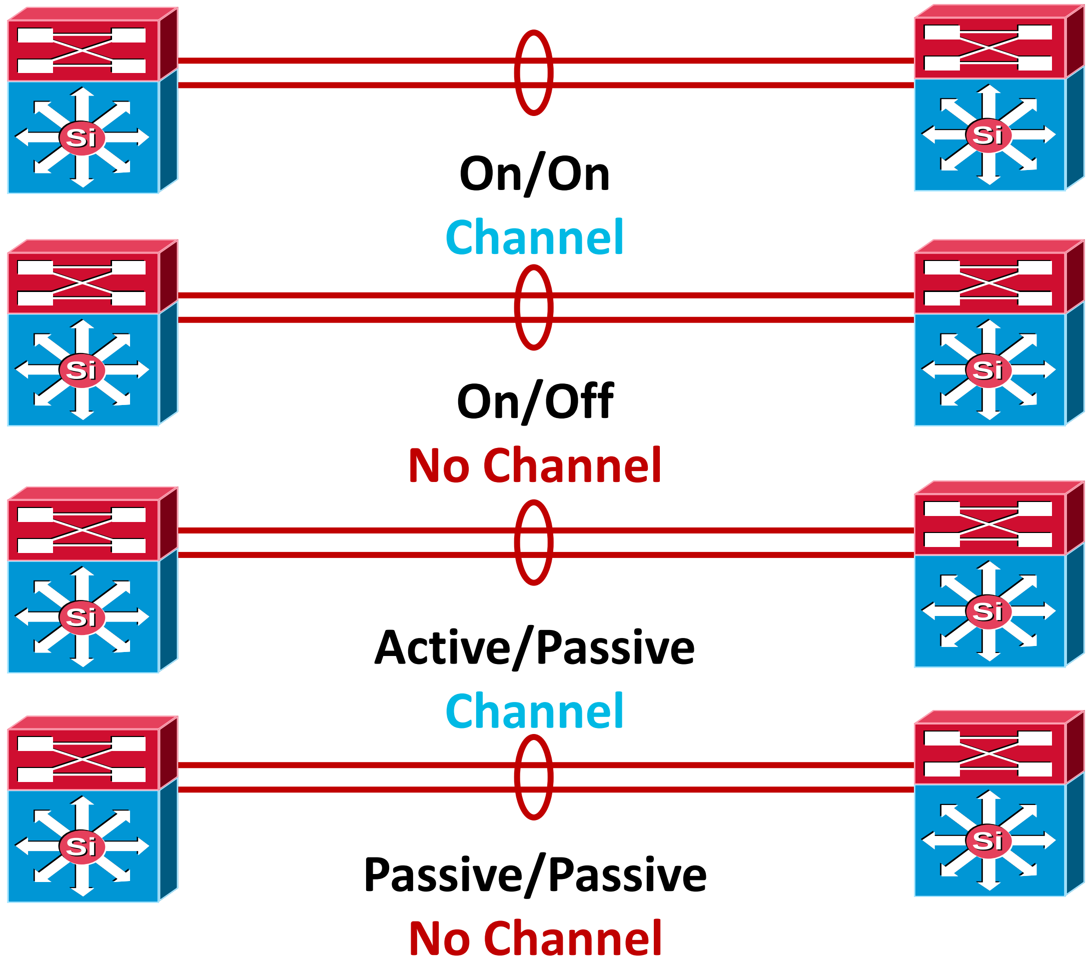
Link Negotiation Options—PAgP and LACP

Port Aggregation Protocol



On: always be a channel/bundle member
Desirable: ask if the other side can/will
Auto: if the other side asks I will
Off: don’t become a member of a channel/bundle

Link Aggregation Protocol



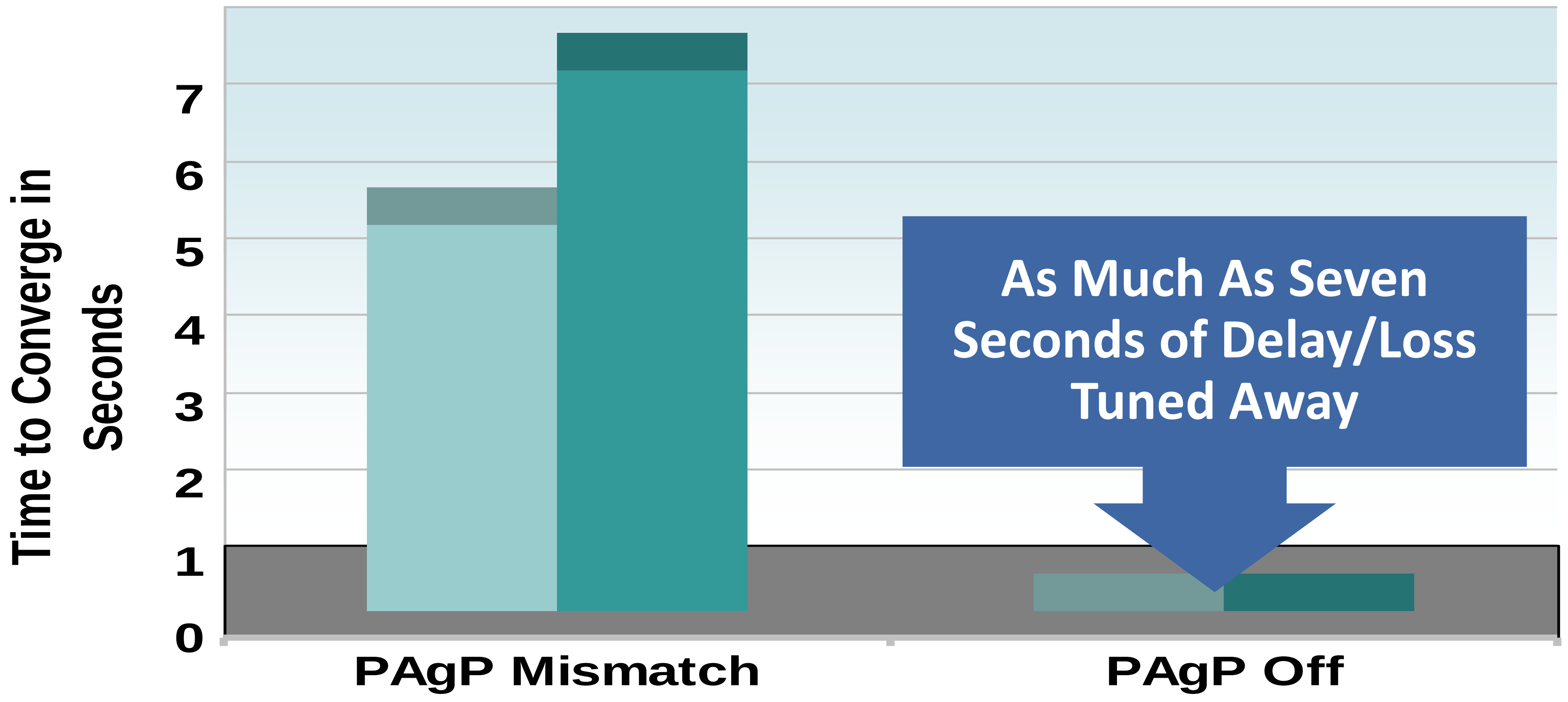
On: always be a channel/bundle member
Active: ask if the other side can/will
Passive: if the other side asks I will
Off: don’t become a member of a channel/bundle

PAgP Tuning

PAgP Default Mismatches

Matching EtherChannel Configuration on Both Sides Improves Link Restoration Convergence Times

```
Channel-group 20 mode desirable
```

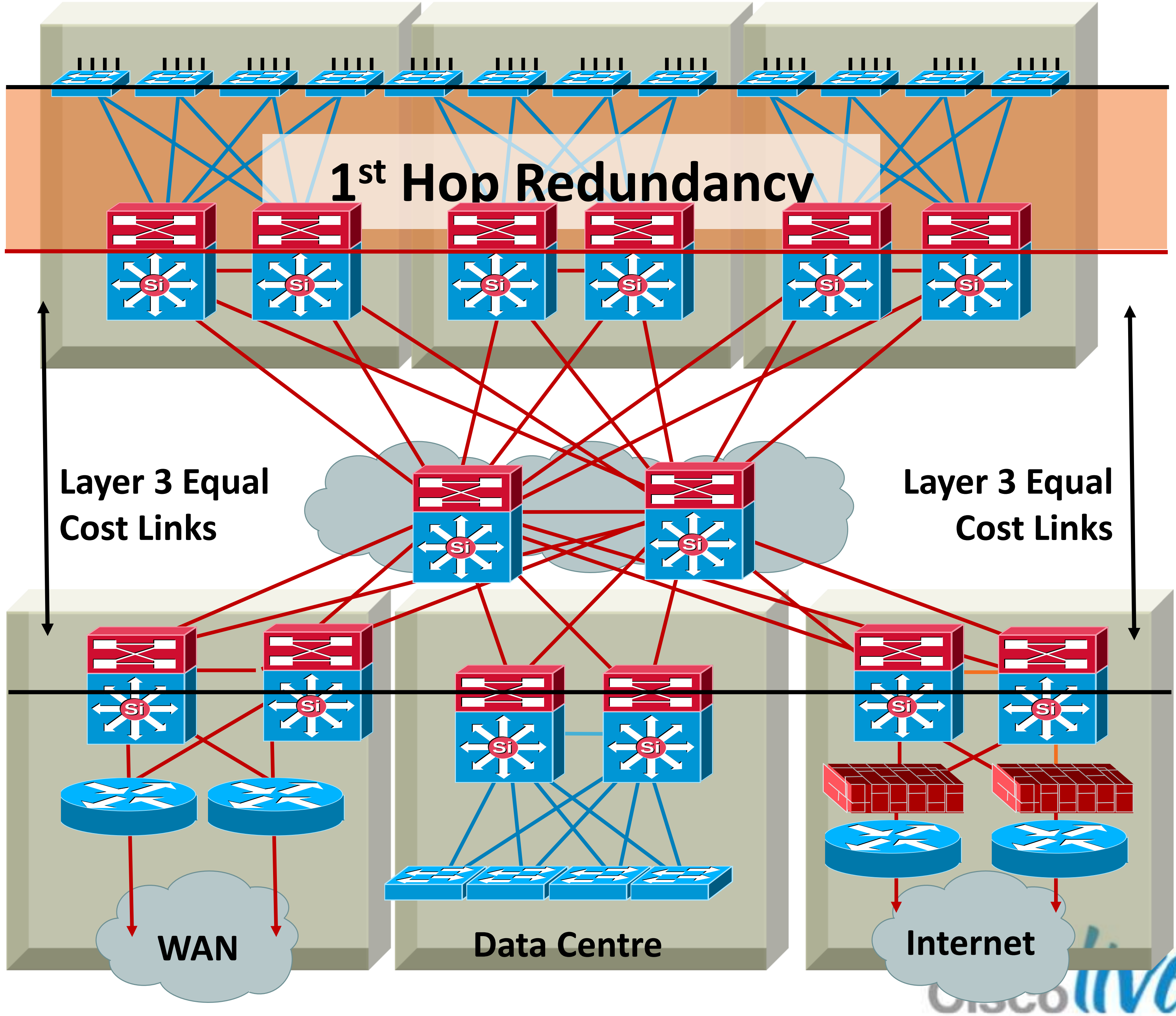


EtherChannels—Quick Summary

- For Layer 2 EtherChannels: Desirable/Desirable is the recommended configuration so that PAgP is running across all members of the bundle **insuring** that an individual link failure will not result in an STP failure
- For Layer 3 EtherChannels: one can consider a configuration that uses ON/ON. There is a trade-off between performance/HA impact and maintenance and operations implications
- An ON/ON configuration is faster from a link-up (restoration) perspective than a Desirable/Desirable alternative. However, in this configuration PAgP is not actively monitoring the state of the bundle members and a misconfigured bundle is not easily identified
- Routing protocols may not have visibility into the state of an individual member of a bundle. LACP and the minimum links option can be used to bring the entire bundle down when the capacity is diminished.
  - –OSPF has visibility to member loss (best practices pending investigation). EIGRP does not…
- When used to increase bandwidth—no individual flow can go faster than the speed of an individual member of the link
- Best used to eliminate single points of failure (i.e., link or port) dependencies from a topology

Best Practices—First Hop Redundancy

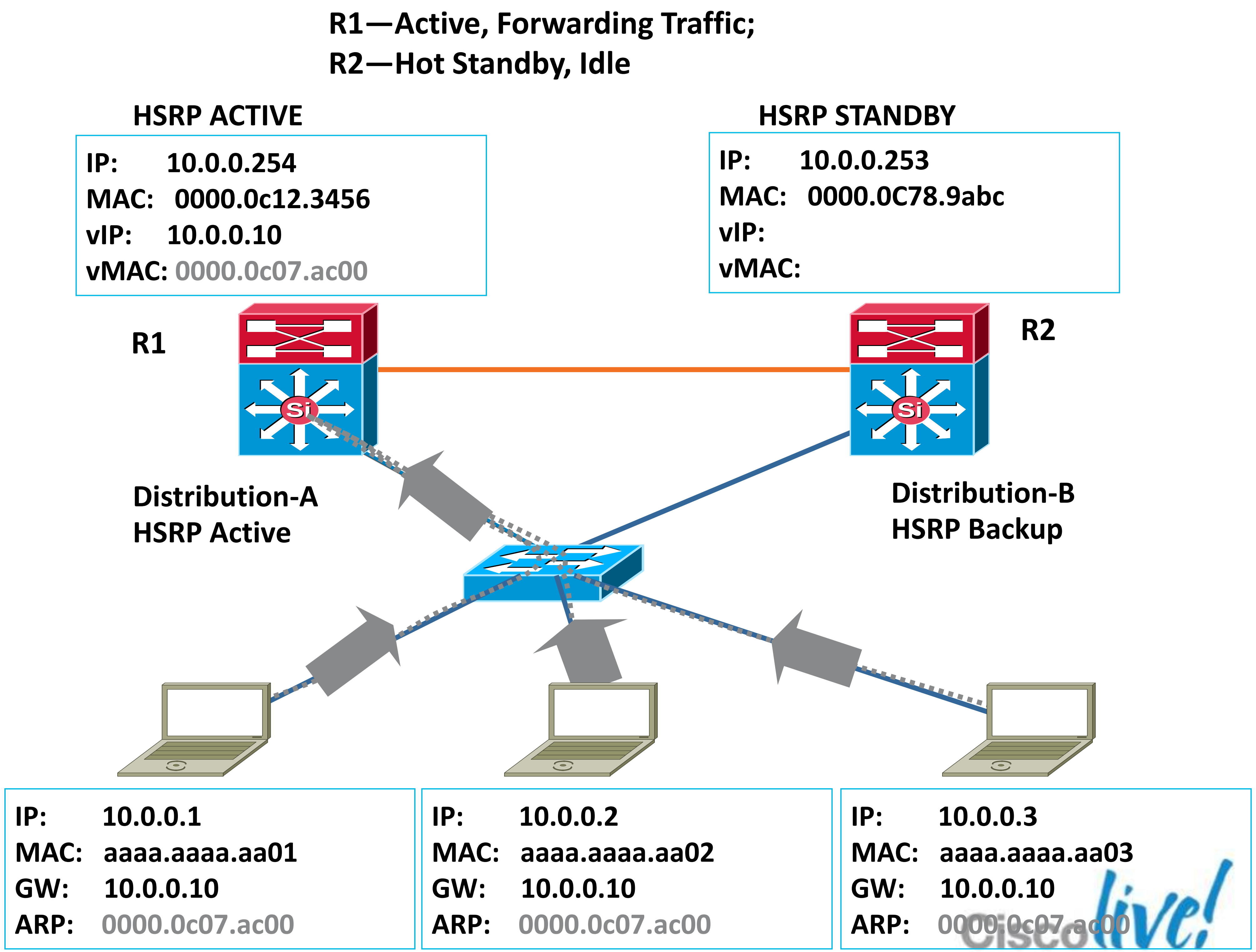
- Used to provide a resilient default gateway/first hop address to end-stations
- HSRP, VRRP, and GLBP alternatives
- VRRP, HSRP, and GLBP provide millisecond timers and excellent convergence performance
- VRRP if you need multivendor interoperability
- GLBP facilitates uplink load balancing
- Preempt timers need to be tuned to avoid black-holed traffic



First Hop Redundancy with HSRP

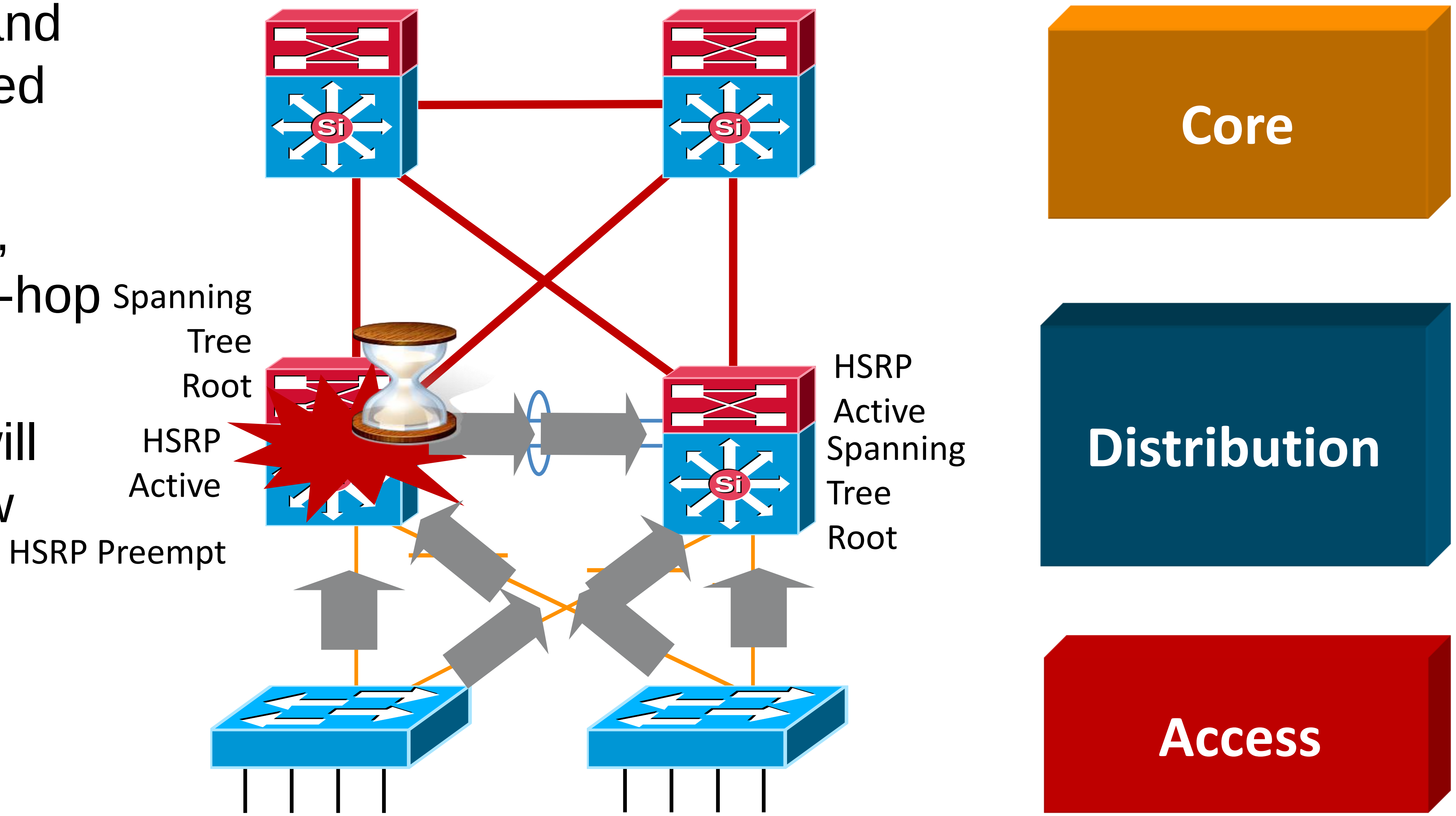
RFC 2281 (March 1998)

- A group of routers function as one virtual router by sharing **one** virtual IP address and one virtual MAC address
- One (active) router performs packet forwarding for local hosts
- The rest of the routers provide hot standby in case the active router fails
- Standby routers stay idle as far as packet forwarding from the client side is concerned



Why You Want HSRP Preemption

- Spanning tree root and HSRP primary aligned
- When spanning tree root is re-introduced, traffic will take a two-hop path to HSRP active
- HSRP preemption will allow HSRP to follow spanning tree topology



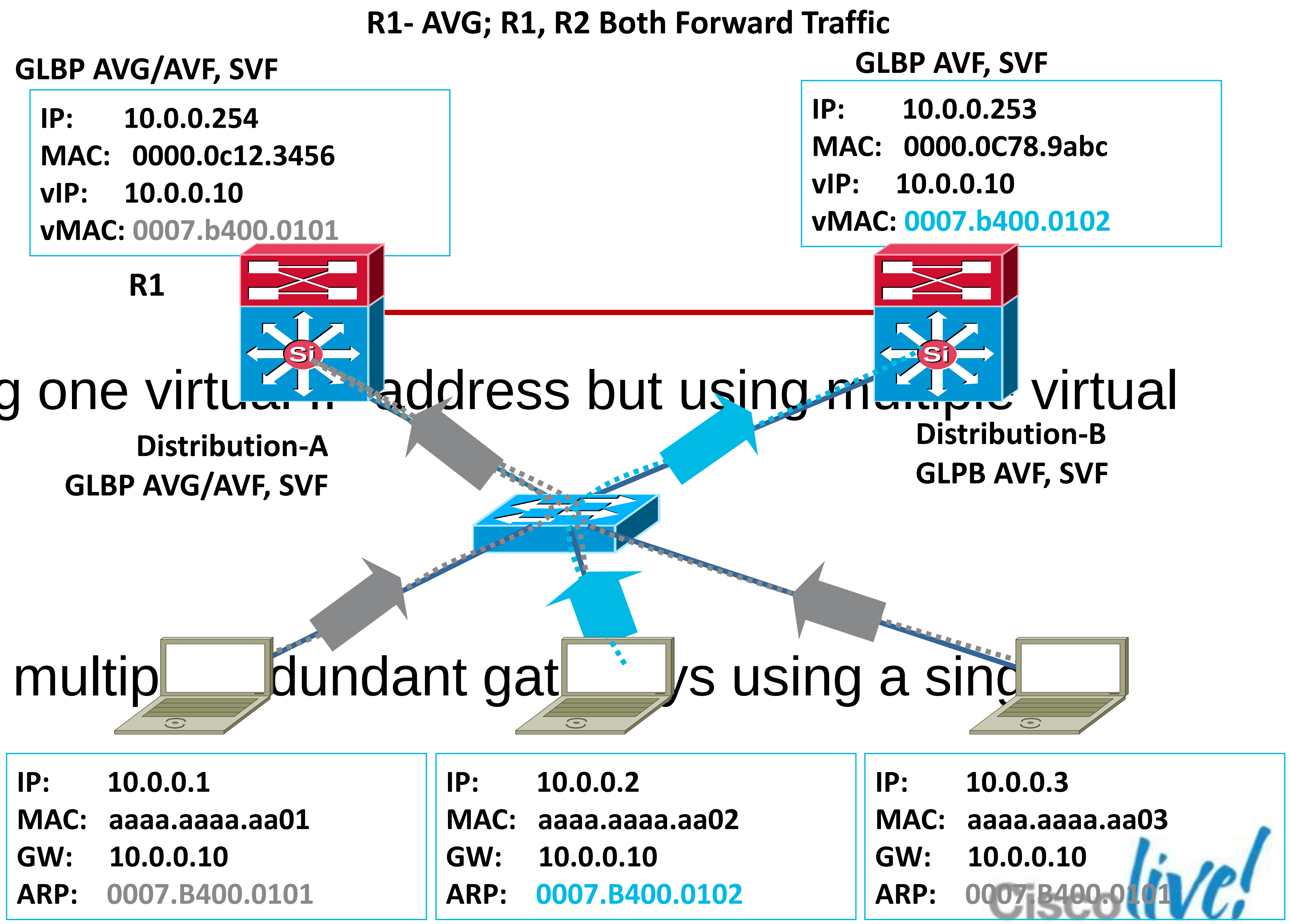
Without Preempt Delay HSRP Can Go Active Before Box Completely Ready to Forward Traffic: L1 (Boards), L2 (STP), L3 (IGP Convergence)

```
standby 1 preempt delay minimum 180
```


First Hop Redundancy with GLBP

Cisco Designed, Load Sharing, Patent Pending

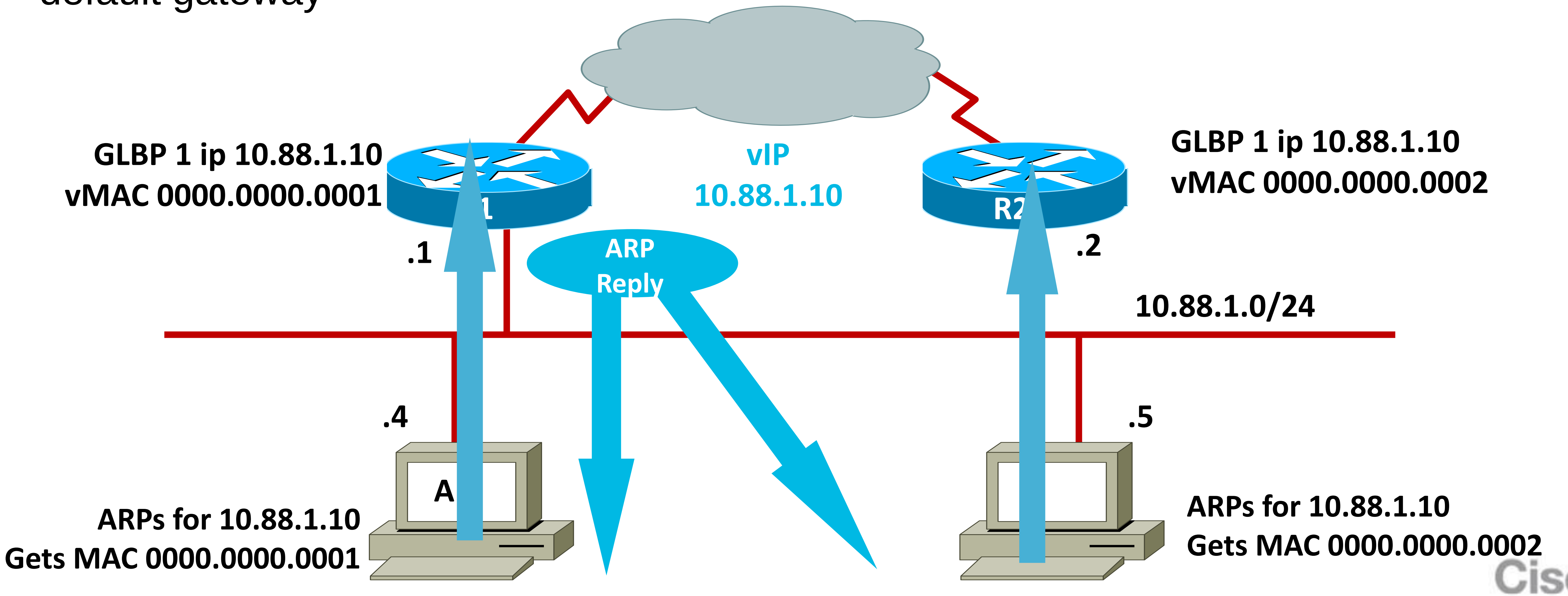
- All the benefits of HSRP plus load balancing of default gateway → utilises all available bandwidth
- A group of routers function as one virtual router by sharing one virtual IP address but using multiple virtual MAC addresses for traffic forwarding
- Allows traffic from a single common subnet to go through multiple redundant gateways using a single virtual IP address



First Hop Redundancy with Load Balancing

Cisco Gateway Load Balancing Protocol (GLBP)

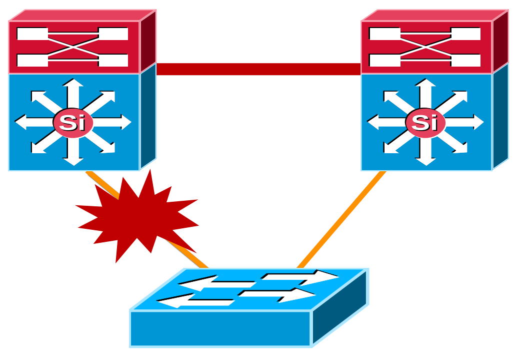
- Each member of a GLBP redundancy group owns a unique virtual MAC address for a common IP address/default gateway
- When end-stations ARP for the common IP address/default gateway they are given a load-balanced virtual MAC address
- Host A and host B send traffic to different GLBP peers but have the same default gateway



Optimising Convergence: VRRP, HSRP, GLBP

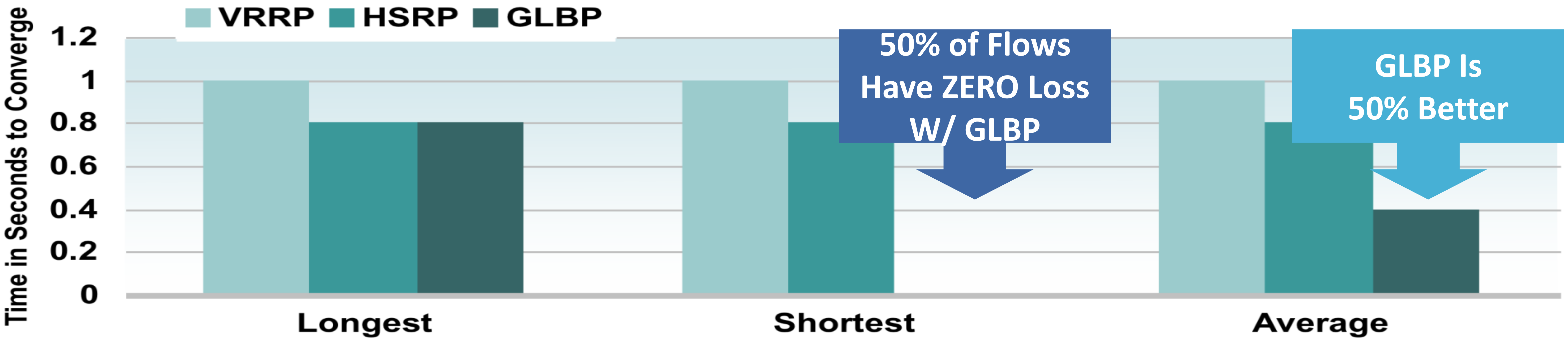
Mean, Max, and Min—Are There Differences?

- VRRP not tested with sub-second timers and all flows go through a common VRRP peer; mean, max, and min are equal
- HSRP has sub-second timers; however all flows go through same HSRP peer so there is no difference between mean, max, and min
- GLBP has sub-second timers and distributes the load amongst the GLBP peers; so 50% of the clients are not affected by an uplink failure



Distribution to Access Link Failure

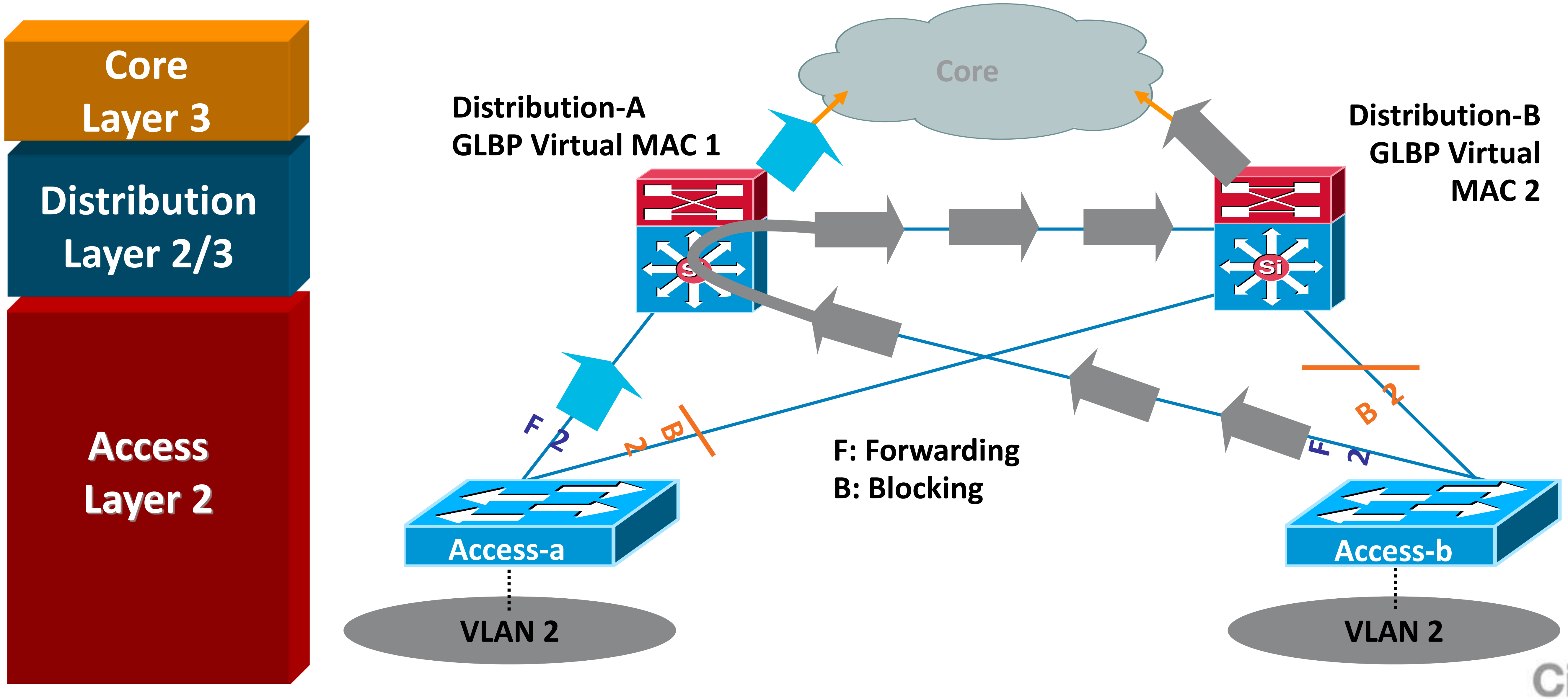
Access to Server Farm



If You Span VLANS, Tuning Required

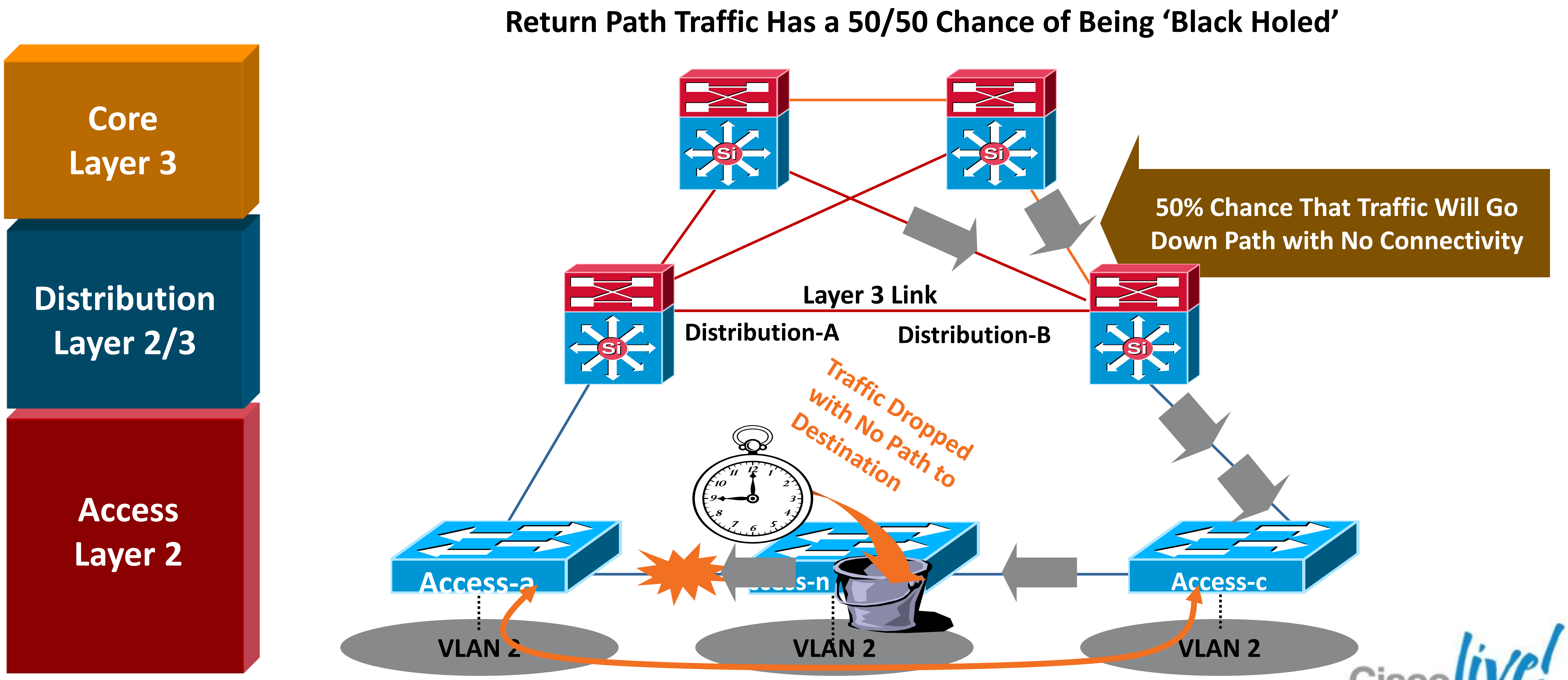
By Default, Half the Traffic Will Take a Two-Hop L2 Path

- Both distribution switches act as default gateway
- Blocked uplink caused traffic to take less than optimal path



Daisy Chaining Access Layer Switches

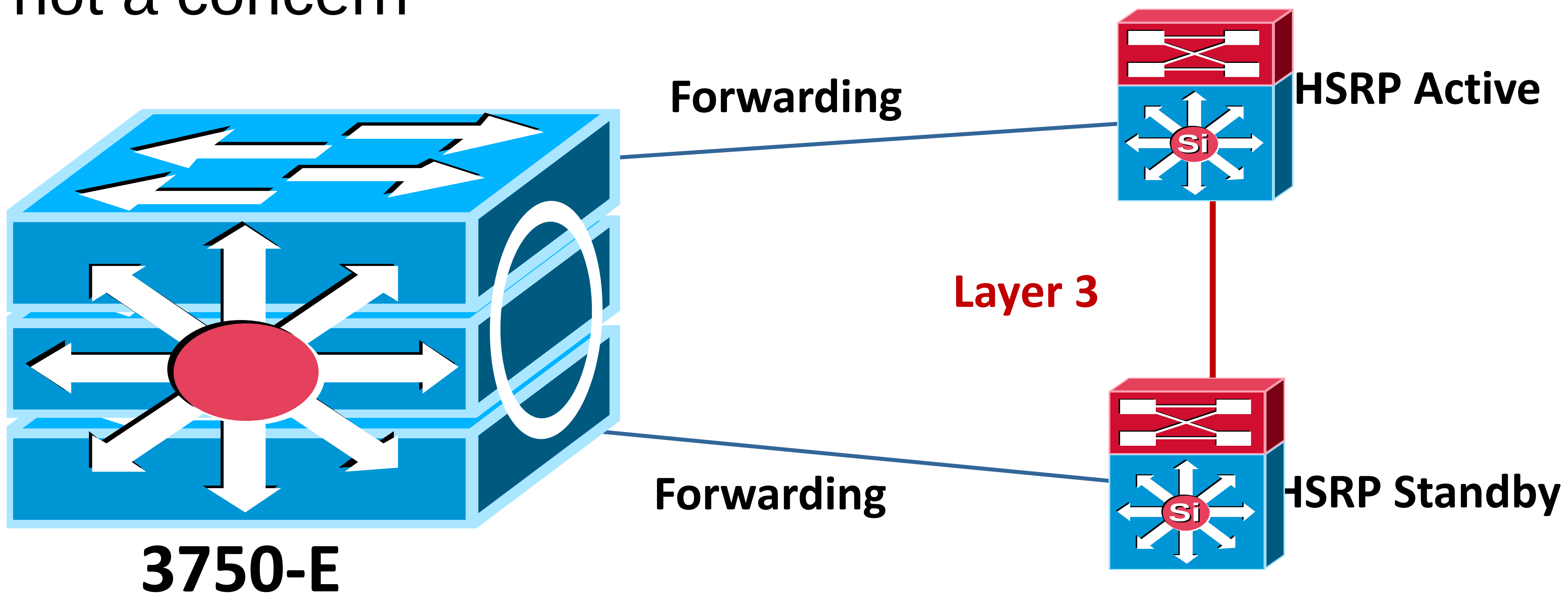
Avoid Potential Black Holes



Daisy Chaining Access Layer Switches

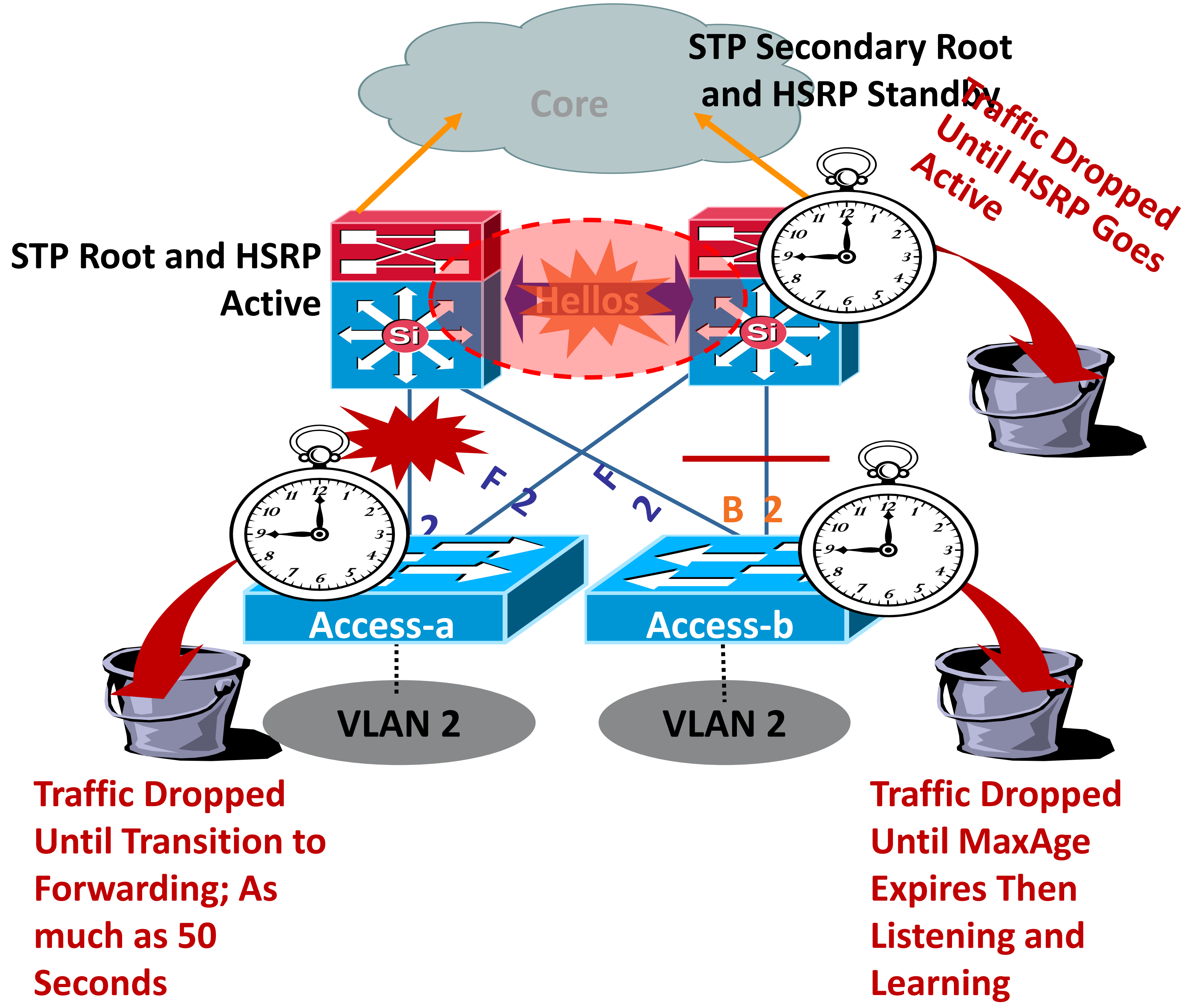
New Technology Addresses Old Problems

- **Stackwise/Stackwise-Plus** technology eliminates the concern
  - – Loopback links not required
  - – No longer forced to have L2 link in distribution
- If you use modular (chassis-based) switches, these problems are not a concern



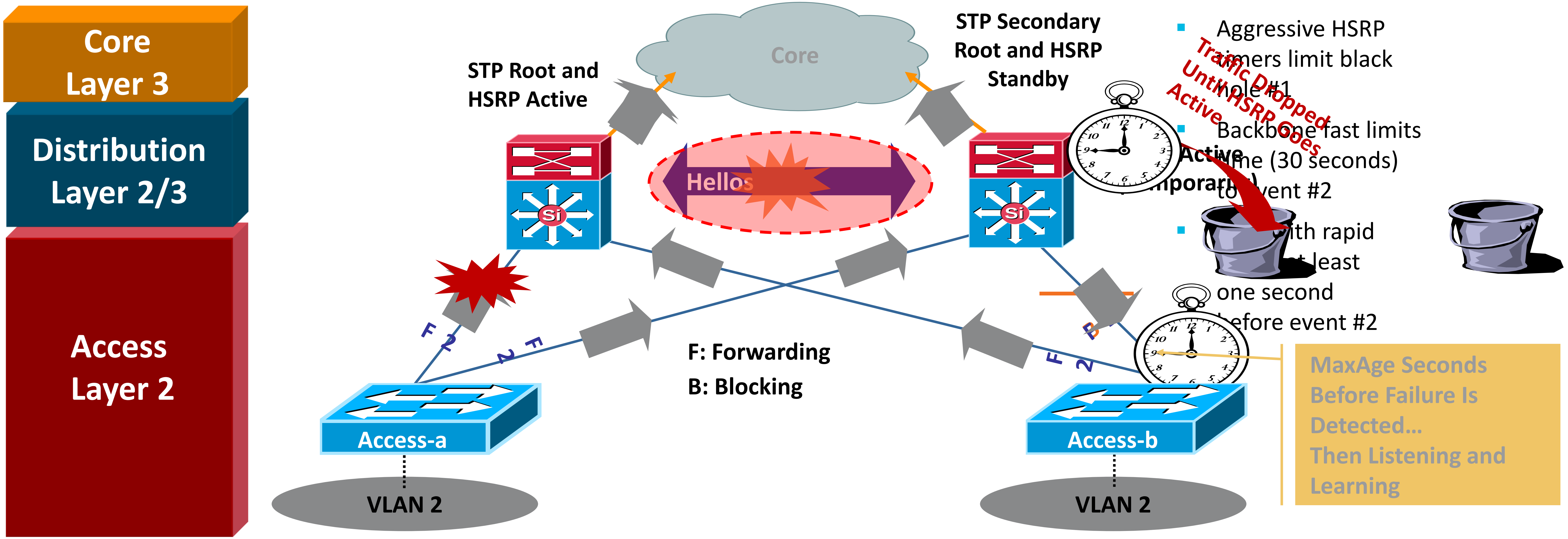
What Happens if You Don’t Link the Distributions?

- STPs slow convergence can cause considerable periods of traffic loss
- STP could cause non-deterministic traffic flows/link load engineering
- STP convergence will cause Layer 3 convergence
- STP and Layer 3 timers are independent
- Unexpected Layer 3 convergence and reconvergence could occur
- Even if you do link the distribution switches dependence on STP and link state/connectivity can cause HSRP irregularities and unexpected state transitions



What if You Don’t?

Black Holes and Multiple Transitions …



- Aggressive HSRP timers limit black hole #1
- Backbone fast limits time (30 seconds) to event #2
- With rapid PVST+ at least one second before event #2

Blocking link on access-b will take 50 seconds to move to forwarding → traffic black hole until HSRP goes active on standby HSRP peer

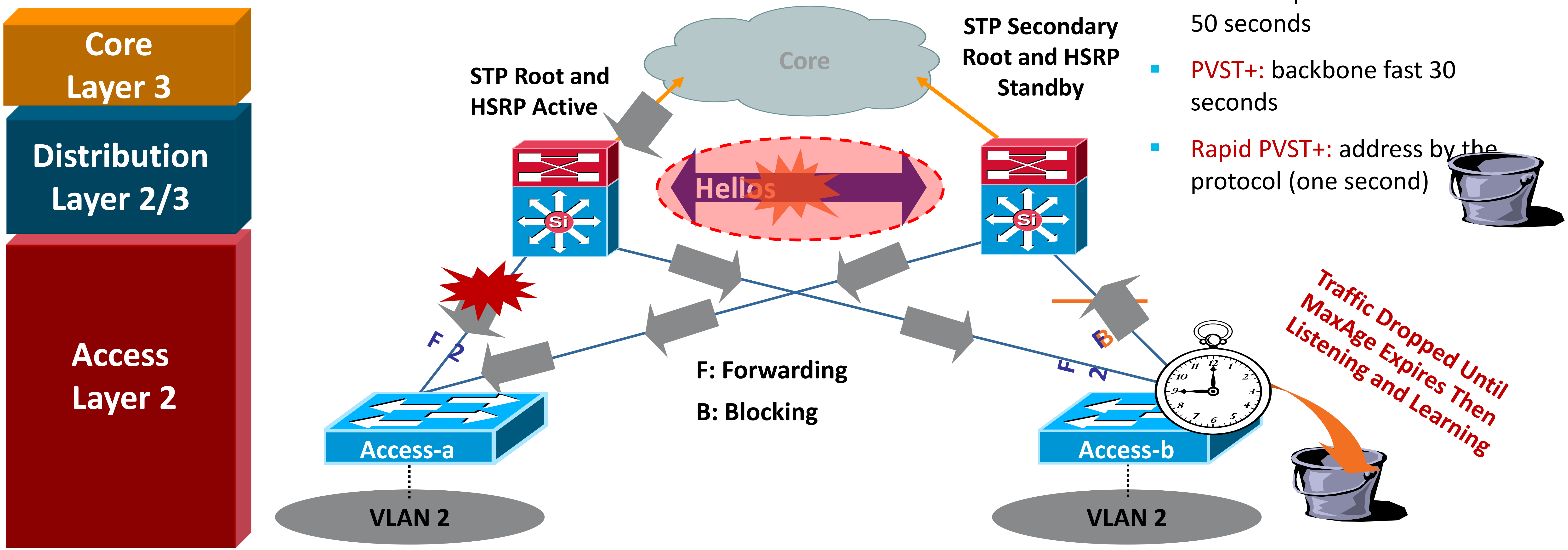
After MaxAge expires (or backbone fast or Rapid PVST+) converges HSRP preempt causes another transition

Access-b used as transit for access-a’s traffic

What If You Don't?

Return Path Traffic Black Holed …

- 802.1d: up to 50 seconds
- PVST+: backbone fast 30 seconds
- Rapid PVST+: address by the protocol (one second)



Blocking link on access-b will take 50 seconds to move to forwarding → return traffic black hole until then

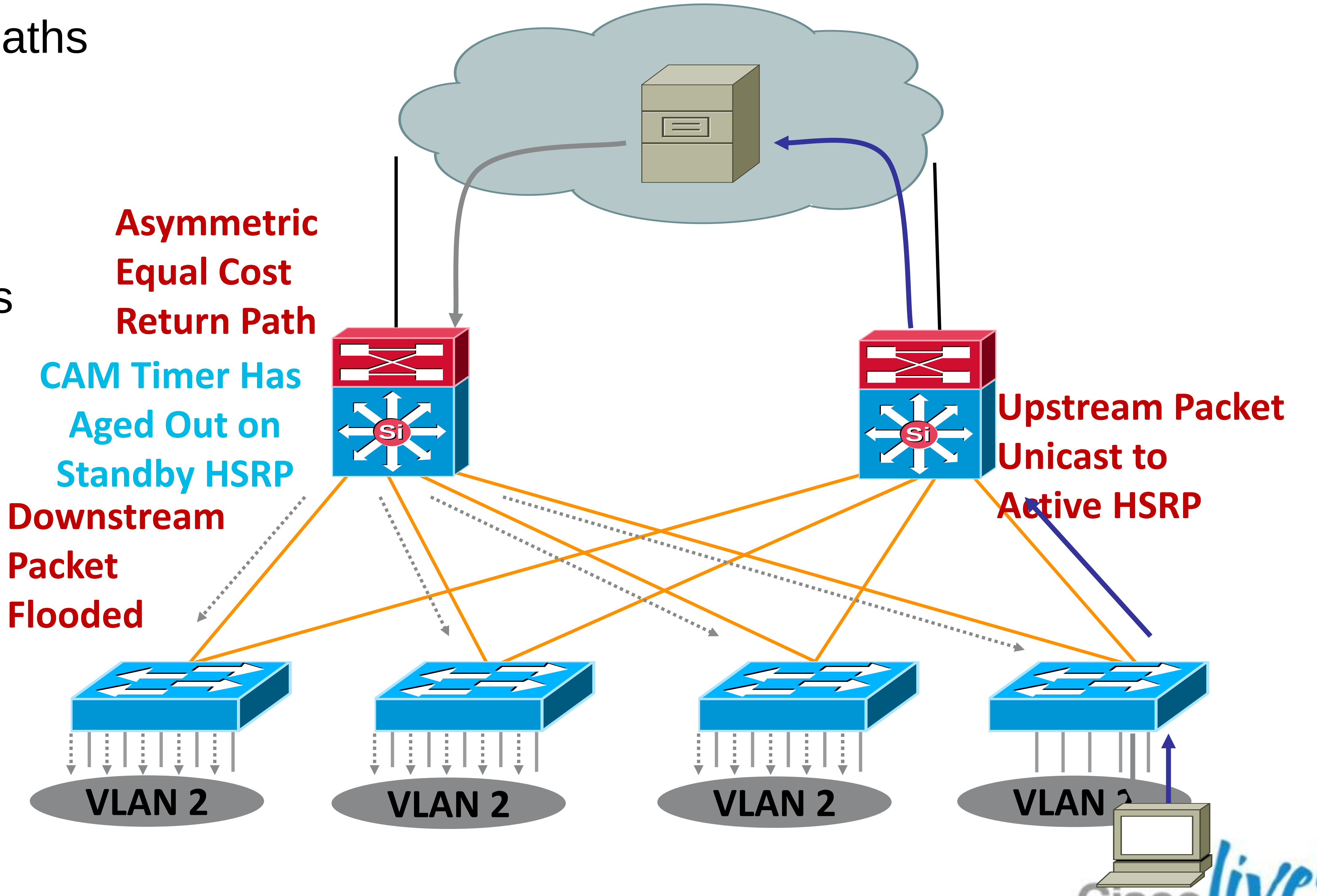
Asymmetric Routing (Unicast Flooding)

Affects redundant topologies with shared L2 access

One path upstream and two paths downstream

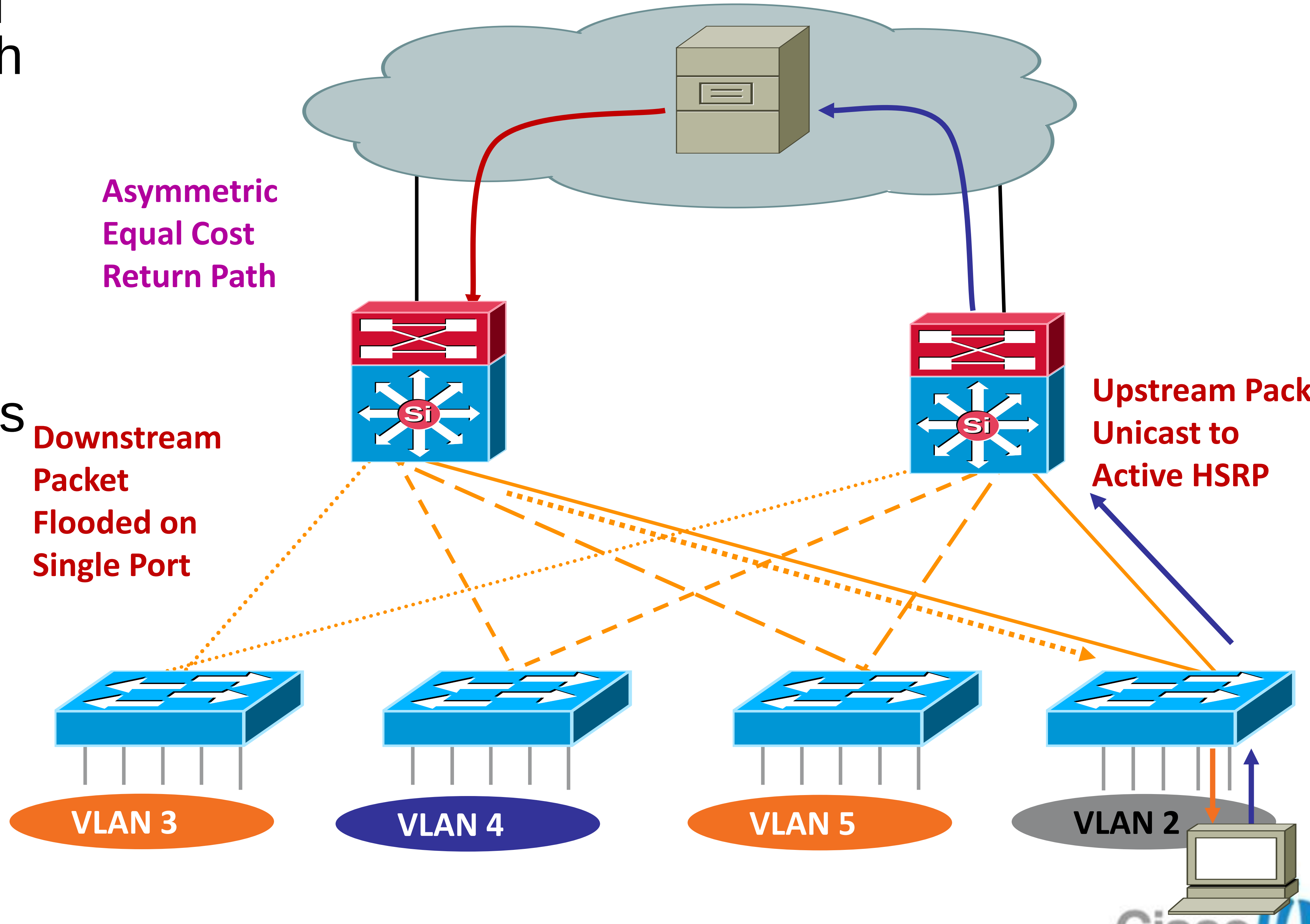
CAM table entry ages out on standby HSRP

Without a CAM entry packet is flooded to all ports in the VLAN



Best Practices Prevent Unicast Flooding

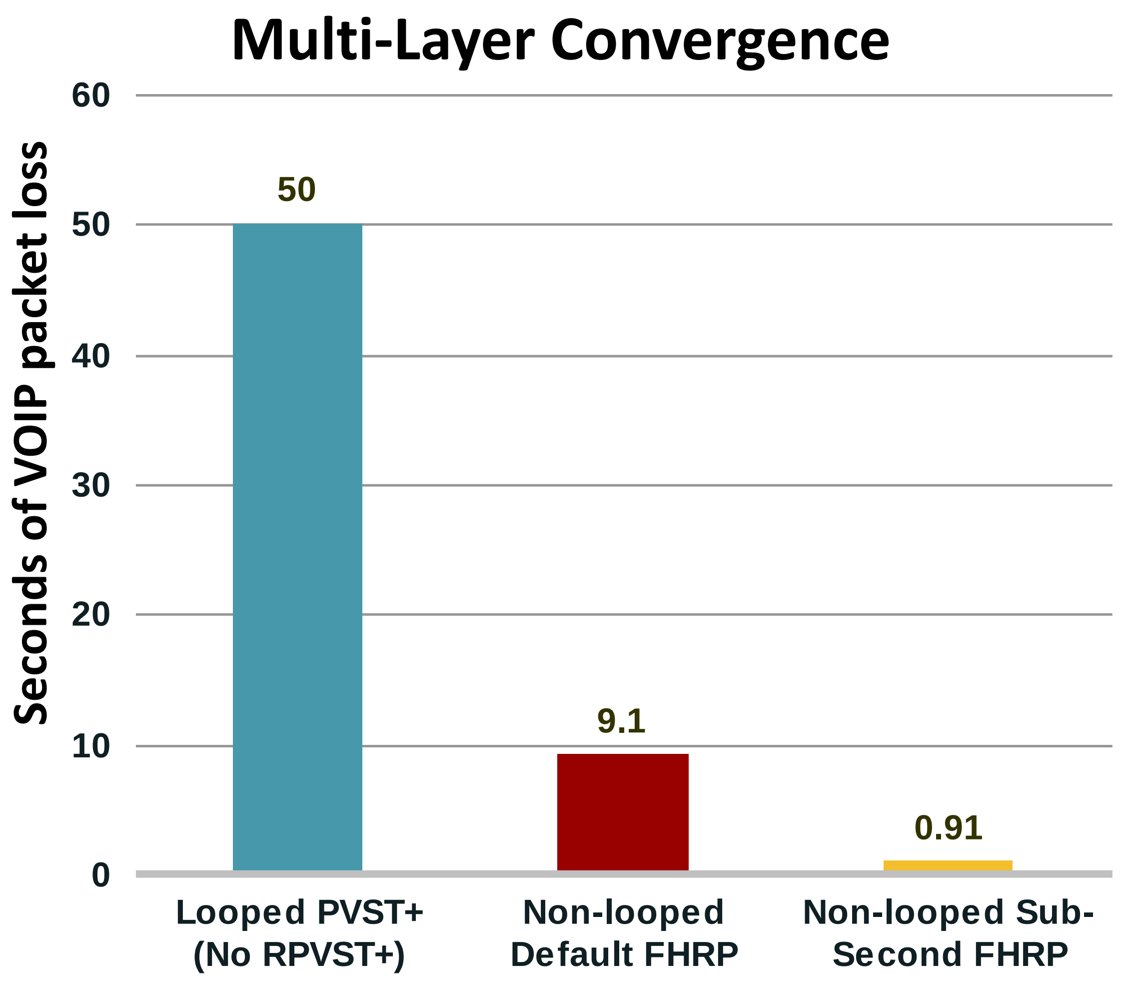
- Assign one unique data and voice VLAN to each access switch
- Traffic is now only flooded down one trunk
- Access switch unicasts correctly; no flooding to all ports
- If you have to:
  - Tune ARP and CAM aging timers; CAM timer exceeds ARP timer
  - Bias routing metrics to remove equal cost routes

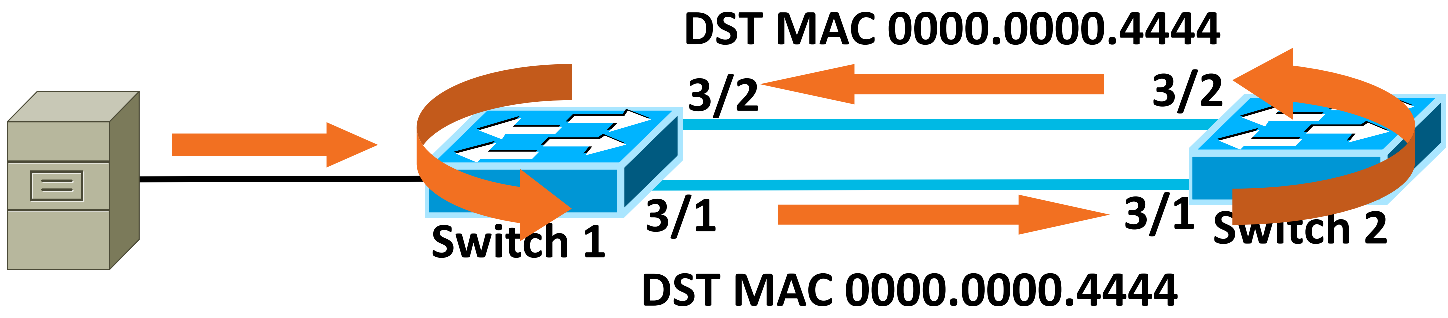


Multi-Layer Network Design

Good Solid Design, But –

- Utilises multiple Control Protocols
  - Spanning Tree (802.1w), HSRP / GLBP, EIGRP, OSPF
- Convergence is dependent on multiple factors –
  - FHRP – 900msec to 9 seconds
  - Spanning Tree – Up to 50 seconds
- Load balancing –
  - Asymmetric forwarding
  - HSRP / VRRP – per subnet
  - GLBP – per host
- Unicast flooding in looped design
- STP, if it breaks badly, has no inherent mechanism to stop the loop



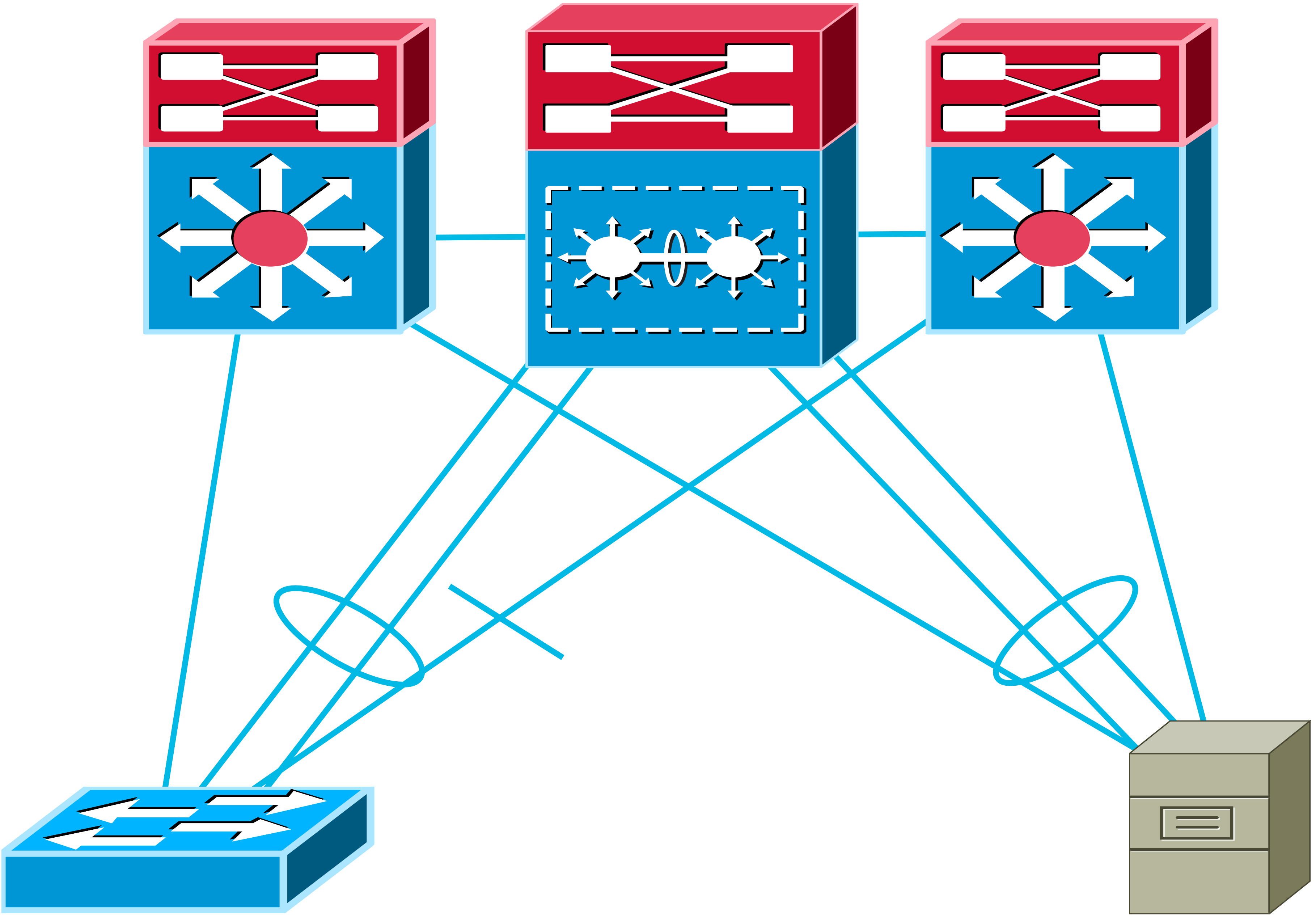


Virtual Switching System (VSS) Designs

Virtual Switching System

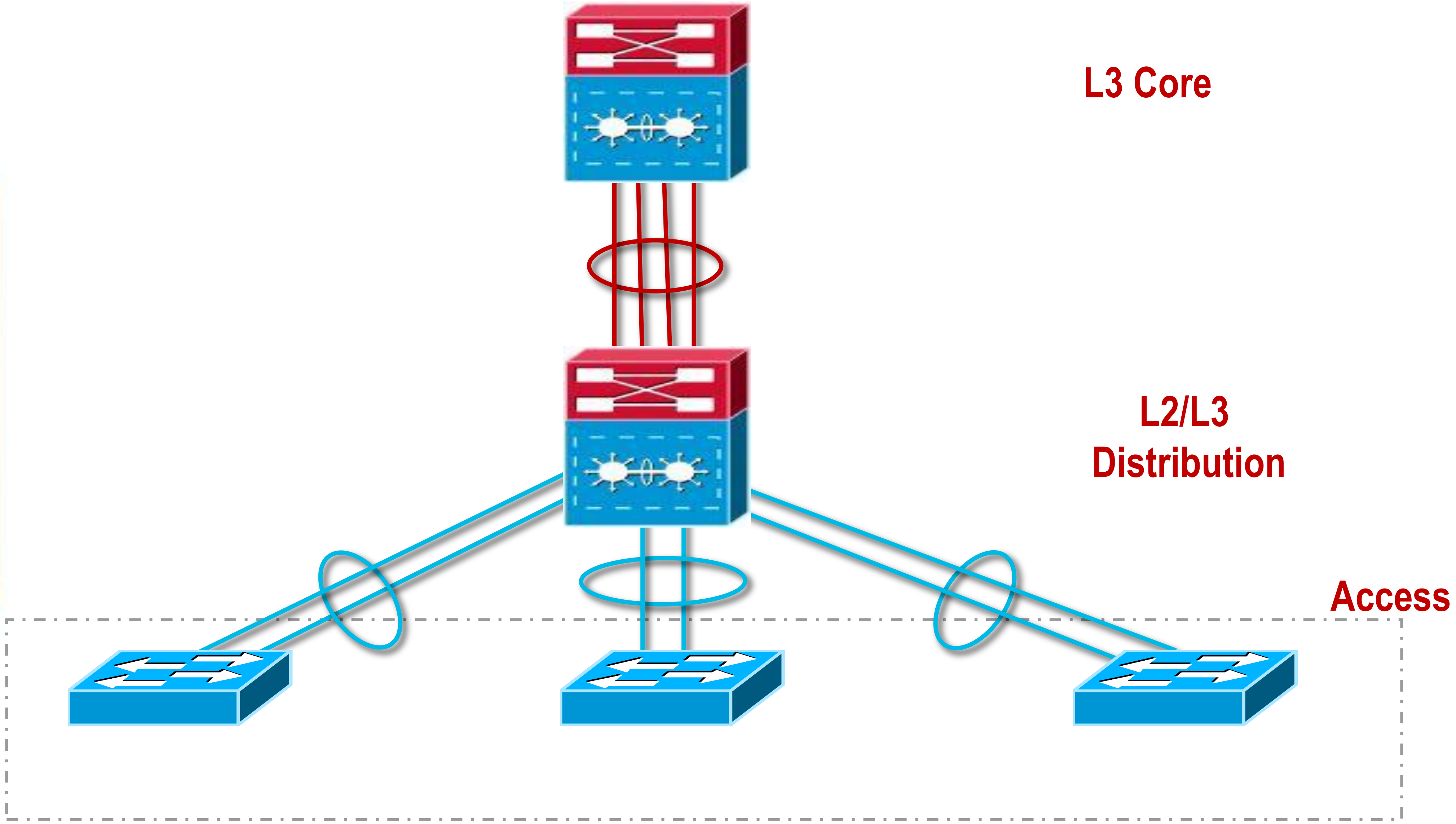
Traditional Design

VSS Design



Virtual Switching System

VSS Enterprise Campus



Reduced routing neighbours, Minimal L3 reconvergence

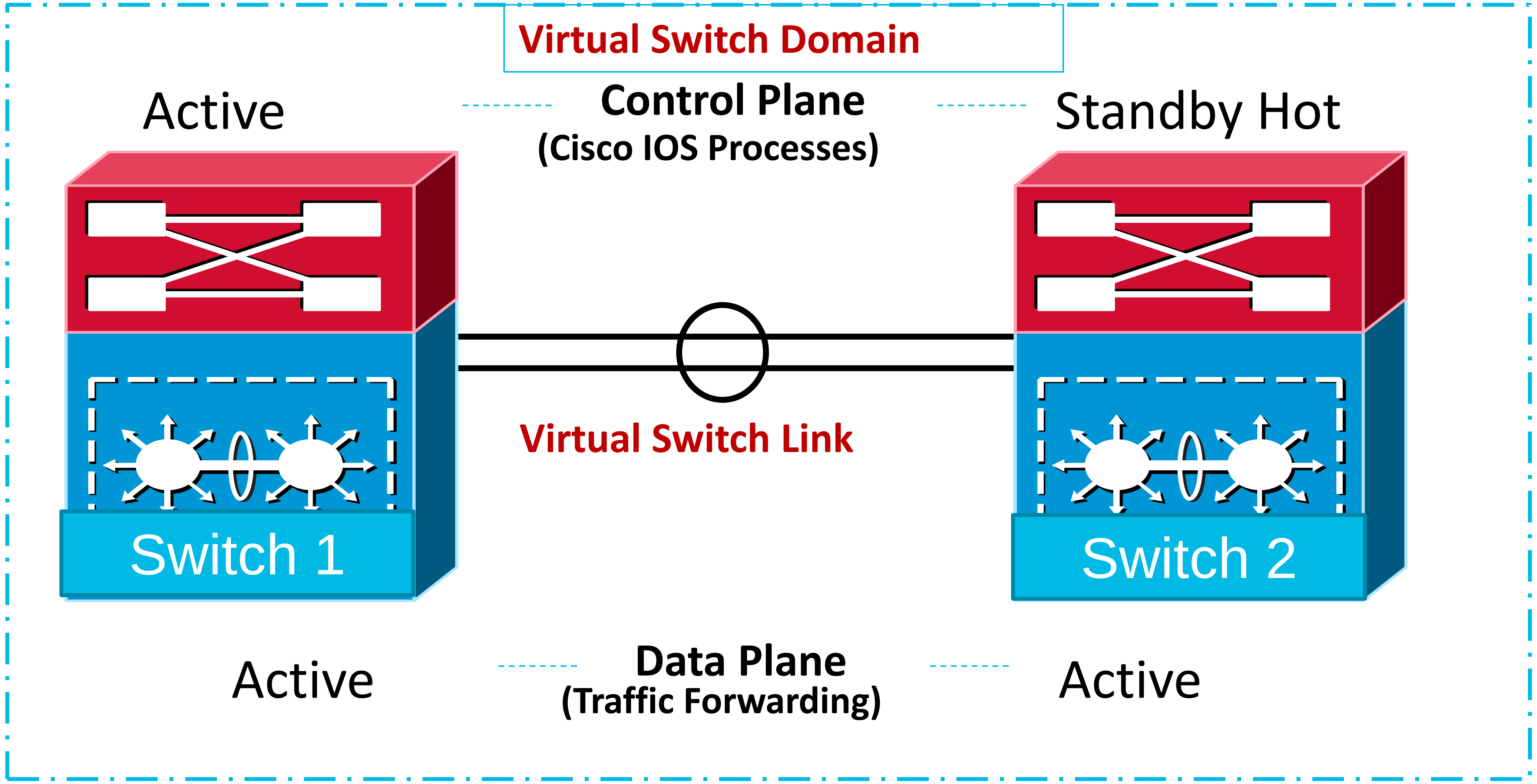
No FHRPs
No Looped topology
Policy Management

Multiple active uplinks per VLAN, No STP convergence

VSS Simplifies the Configuration

Standalone Switch 1 (Coordinated Configuration)	Standalone Switch 2 (Coordinated Configuration)	VSS (One simplified configuration)
Spanning Tree Configuration		
! Enable 802.1d per VLAN spanning tree enhancements. spanning-tree mode pvst spanning-tree loopguard default no spanning-tree optimize bpdu transmission spanning-tree extend system-id spanning-tree uplinkfast spanning-tree backbonefast spanning-tree vlan 2,4,6,8,10 priority 24576!	! Enable 802.1d per VLAN spanning tree enhancements. spanning-tree mode pvst spanning-tree loopguard default no spanning-tree optimize bpdu transmission spanning-tree extend system-id spanning-tree uplinkfast spanning-tree backbonefast spanning-tree vlan 3,5,7,9,11 priority 24576!	! Enable 802.1d per VLAN spanning tree enhancements spanning-tree mode rapid-pvst no spanning-tree optimize bpdu transmission spanning-tree extend system-id spanning-tree vlan 2-11 priority 24576
L3 SVI Configuration (sample for 1 VLAN)		
! Define the Layer 3 SVI for each voice and data VLAN interface Vlan4 description Data VLAN ip address 10.120.4.3 255.255.255.0 no ip redirects no ip unreachables ! Reduce PIM query interval to 250 msec ip pim query-interval 250 msec ip pim sparse-mode load-interval 30 ! Define HSRP default gateway with 250/800 msec hello/hold standby 1 ip 10.120.4.1 standby 1 timers msec 250 msec 800 ! Set preempt delay large enough to allow network to stabilize before HSRP ! switches back on power on or link recovery standby 1 preempt delay minimum 180 ! Enable HSRP authentication standby 1 authentication cisco123	! Define the Layer 3 SVI for each voice and data VLAN interface Vlan4 description Data VLAN ip address 10.120.4.3 255.255.255.0 no ip redirects no ip unreachables ! Reduce PIM query interval to 250 msec ip pim query-interval 250 msec ip pim sparse-mode load-interval 30 ! Define HSRP default gateway with 250/800 msec hello/hold standby 1 ip 10.120.4.1 standby 1 timers msec 250 msec 800 ! Set preempt delay large enough to allow network to stabilize before HSRP ! switches back on power on or link recovery standby 1 preempt delay minimum 180 ! Enable HSRP authentication standby 1 authentication cisco123	! Define the Layer 3 SVI for each voice and data VLAN interface Vlan4 description Data VLAN ip address 10.120.2.1 255.255.255.0 no ip redirects no ip unreachables ip pim sparse-mode load-interval 30

VSS Architecture Concepts



VSS Control Plane

Active / Standby Model

Switch 1 Console (Active)

```
vss#
vss#
vss#
vss#
vss#show switch virtual
Switch mode                   : Virtual Switch
Virtual switch domain number  : 10
Local switch number           : 1
Local switch operational role: Virtual Switch Active
Peer switch number            : 2
Peer switch operational role  : Virtual Switch Standby
vss#
```

Switch 2 Console (Standby Hot)

```
vss-sdby> enable
Standby console disabled

vss-sdby>
```

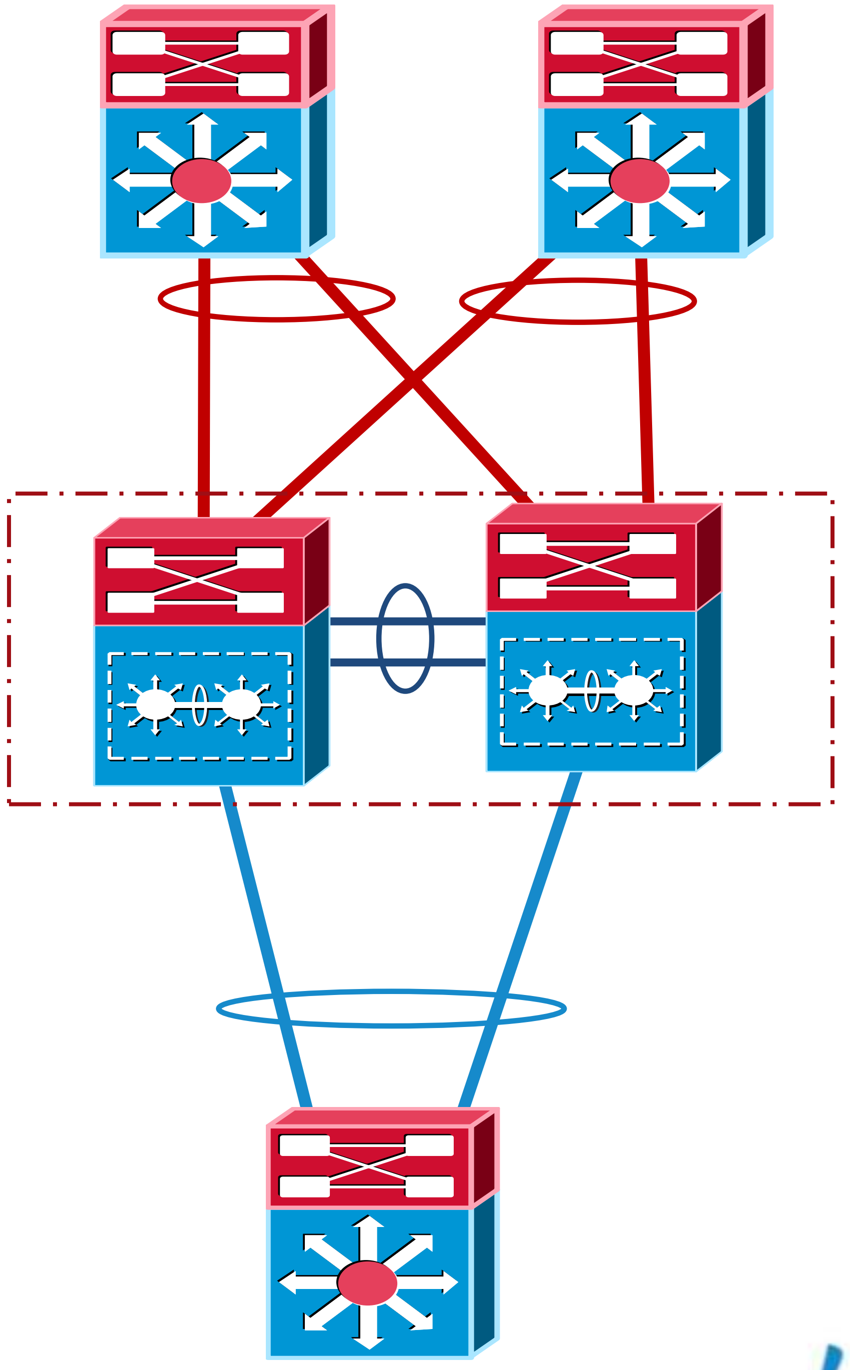
- The switch in Active redundancy mode will maintain the single configuration file for the VSS and sync it to the Standby switch
- Only the console interface on the Active switch is accessible, the Standby console is prohibited from user access

VSS Data Plane

Active – Active

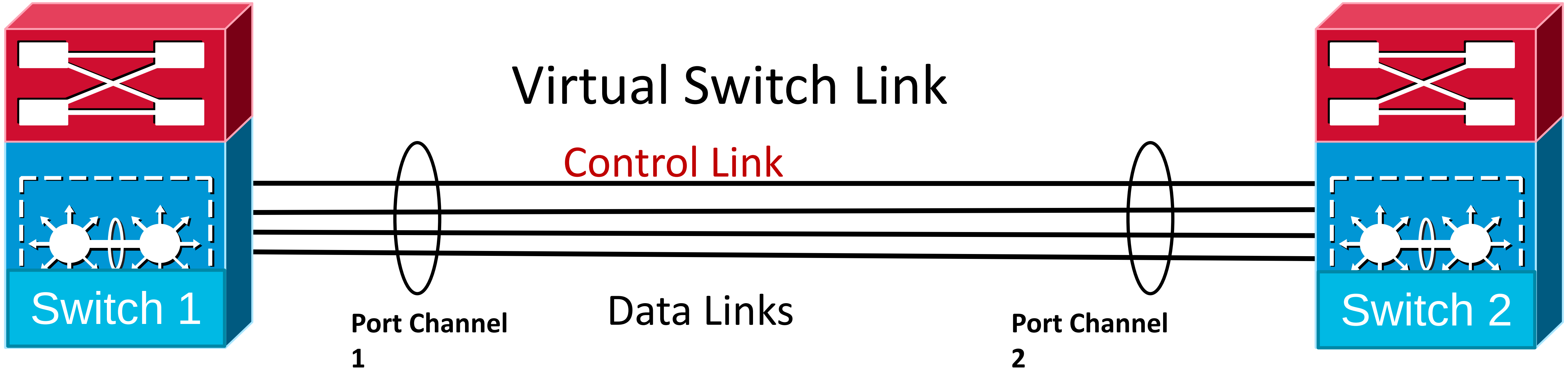
- **Both data and forwarding planes are active**
- Standby supervisor and all line cards are actively forwarding
- No STP blocking ports due to Etherchannel uplinks

```
VSS# show switch virtual redundancy
My Switch Id = 1
Peer Switch Id = 2
<snip>
Switch 1 Slot 5 Processor Information :
-----------------------------------------------
  Current Software state = ACTIVE
<snip>
              Fabric State = ACTIVE
       Control Plane State = ACTIVE
Switch 2 Slot 5 Processor Information :
-----------------------------------------------
  Current Software state = STANDBY HOT (switchover
   target)
 <snip>
              Fabric State = ACTIVE
       Control Plane State = STANDBY
```



Virtual Switching System Architecture

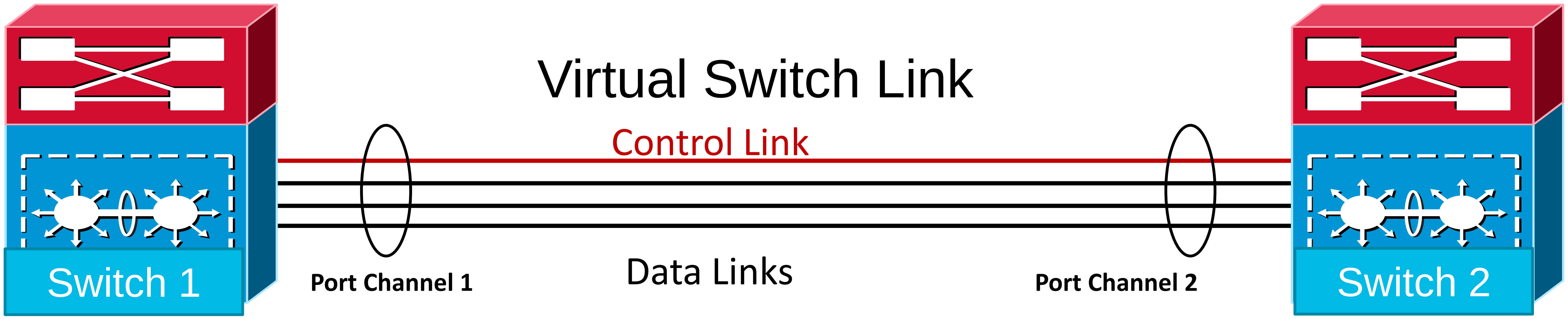
Virtual Switch Link (VSL)




```
interface Port-channel1
 no switchport
 no ip address
 switch virtual link 1
 mls qos trust cos
 no mls qos channel-consistency
!
interface Port-channel2
 no switchport
 no ip address
 switch virtual link 2
 mls qos trust cos
 no mls qos channel-consistency
```

Virtual Switch Link

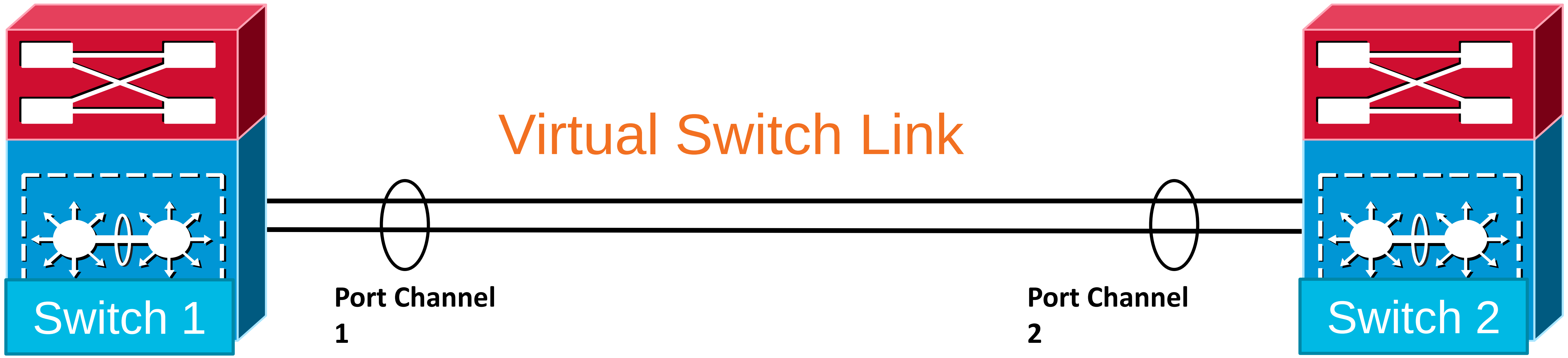
VSL Header



All traffic traversing the VSL link is encapsulated with a 32 byte "Virtual Switch Header" containing ingress and egress switchport indexes, class of service (COS), VLAN number, other important information from the layer 2 and layer 3 header

Virtual Switch Link

Initialisation



1 **Pre-parse config file and bring up VSL interfaces**

2 **Link Management Protocol (LMP) used to track and reject Unidirectional Links, Exchange Chassis ID and other information between the 2 switches**

3 **Role Resolution Protocol (RRP) used to determine compatible Hardware and Software versions to form the VSL as well as determine which switch becomes Active and Hot Standby from a control plane perspective**

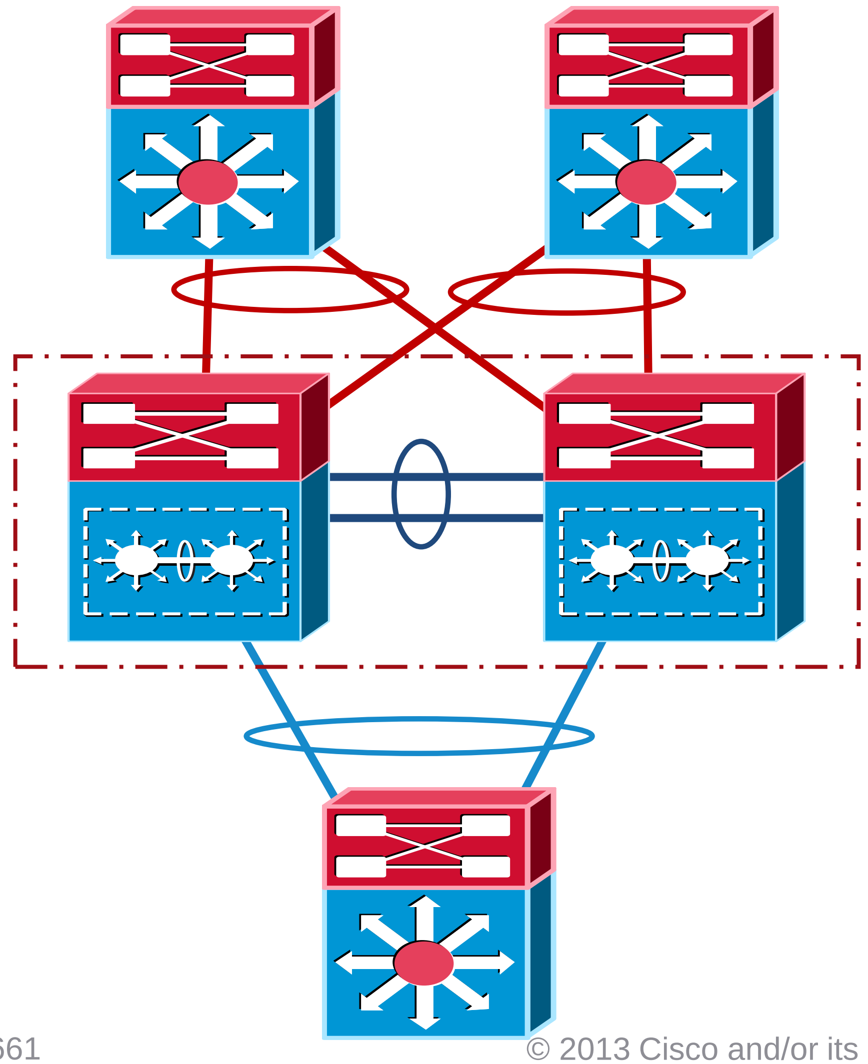
Virtual Switching System Architecture

Traffic Forwarding Enhancements

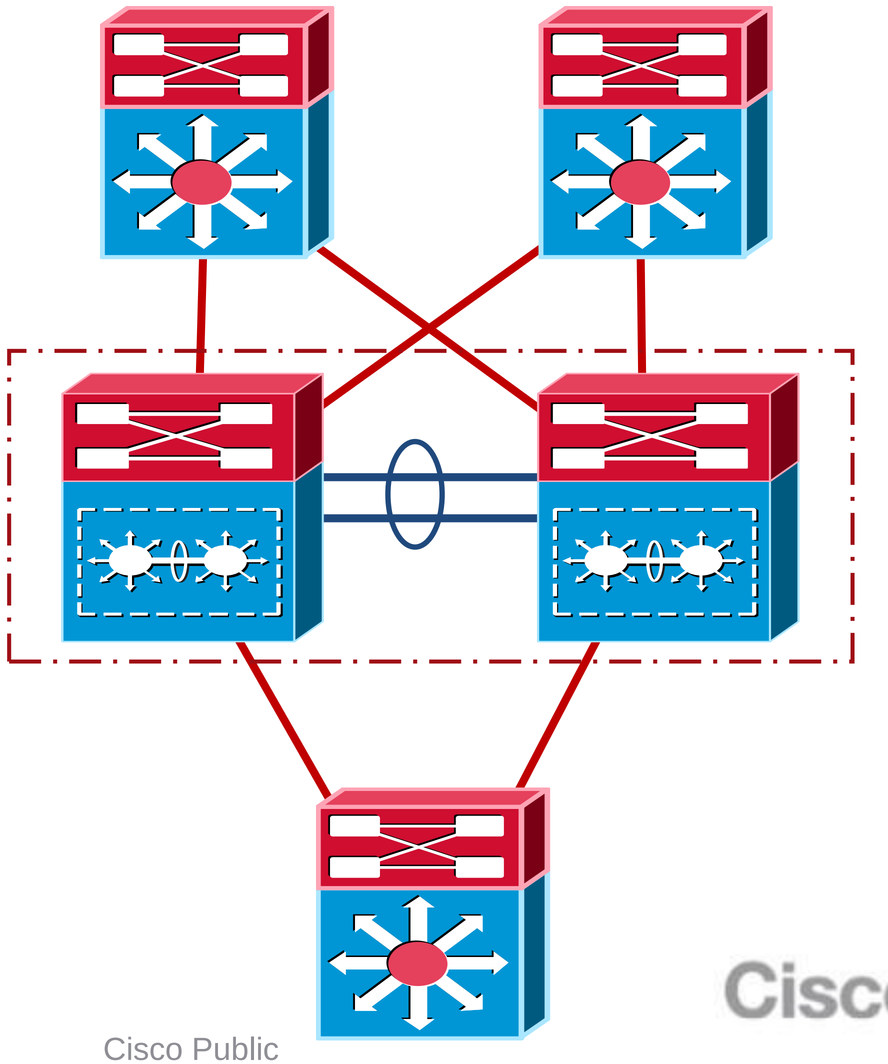
For a VSS, Etherchannel and L3 ECMP forwarding will always favor locally attached interfaces

- **Deterministic Traffic patterns**
- **Removes the need to send traffic over the VSL**

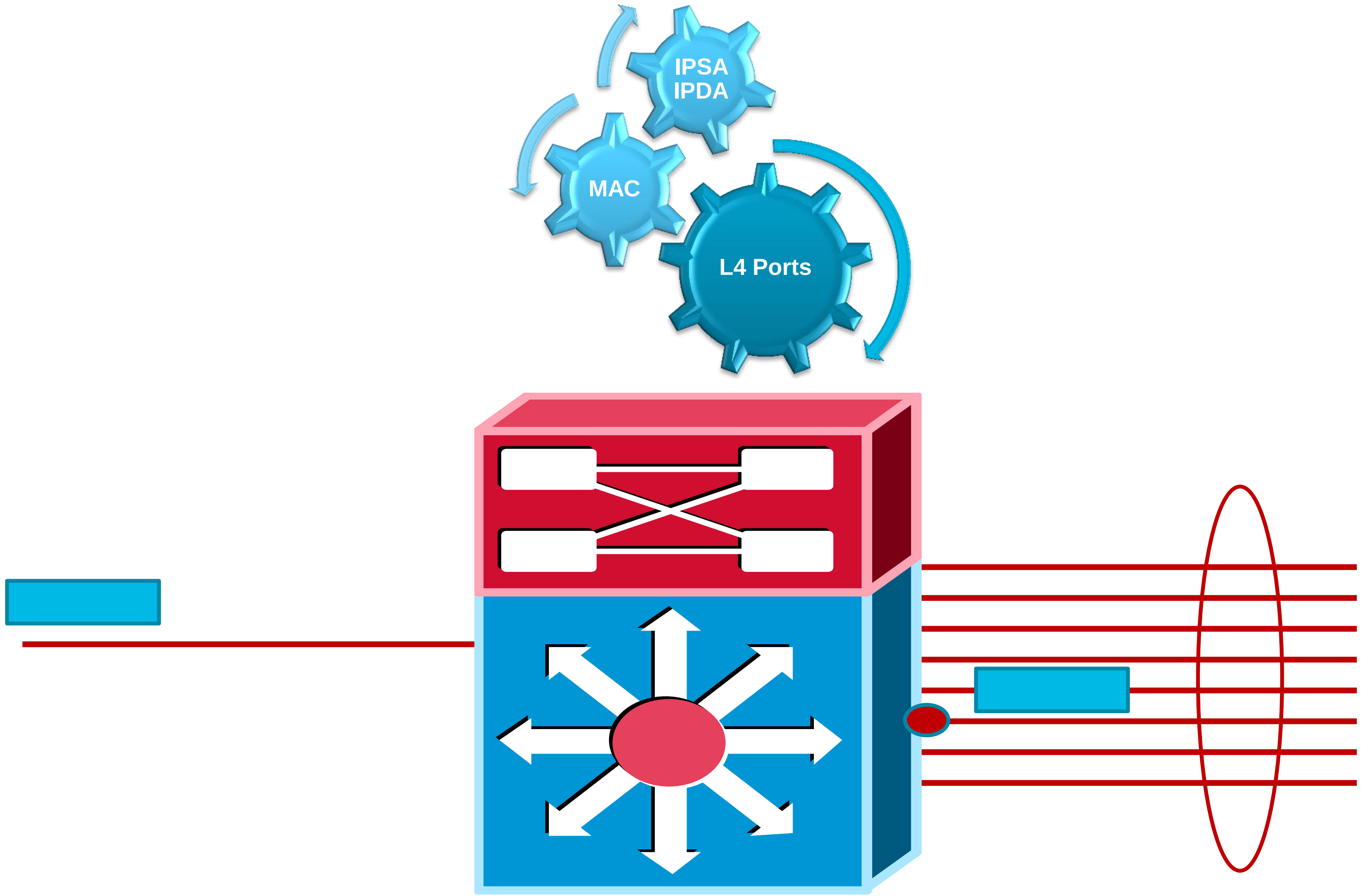
Multichassis Etherchannel (MEC)



L3 Equal Cost Multi-Path Routing (ECMP)



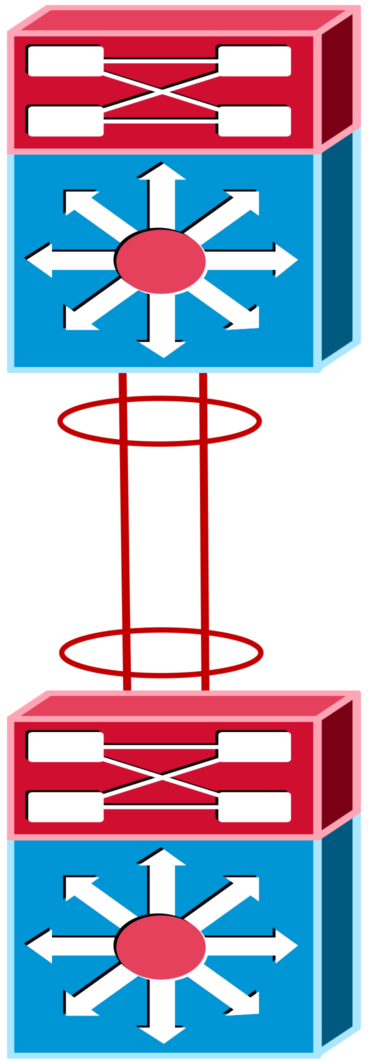
Etherchannel Traffic Load Balancing



Virtual Switching System Architecture

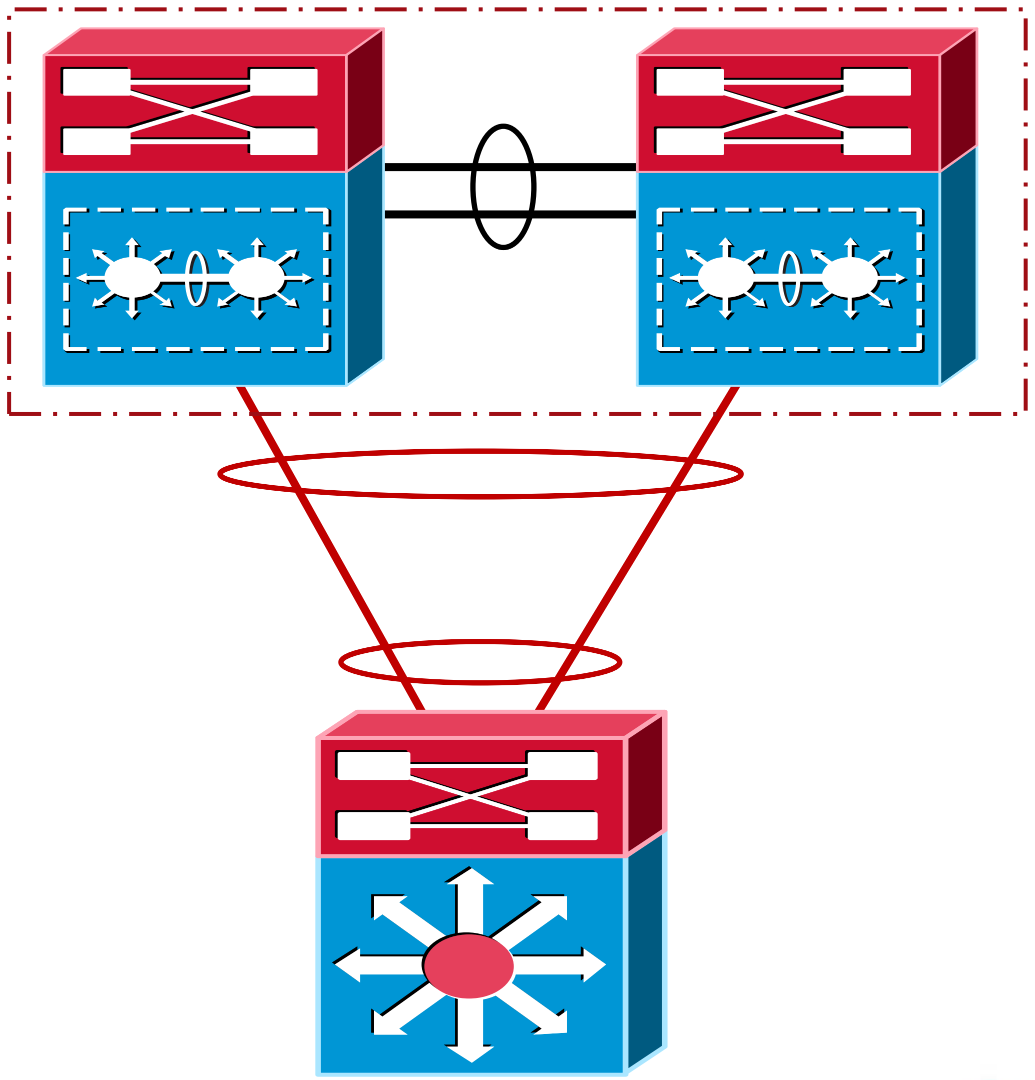
Multichassis EtherChannel (MEC)

Traditional Etherchannel



Multichassis Etherchannel (MEC)

One logical link partner, but two physical chassis



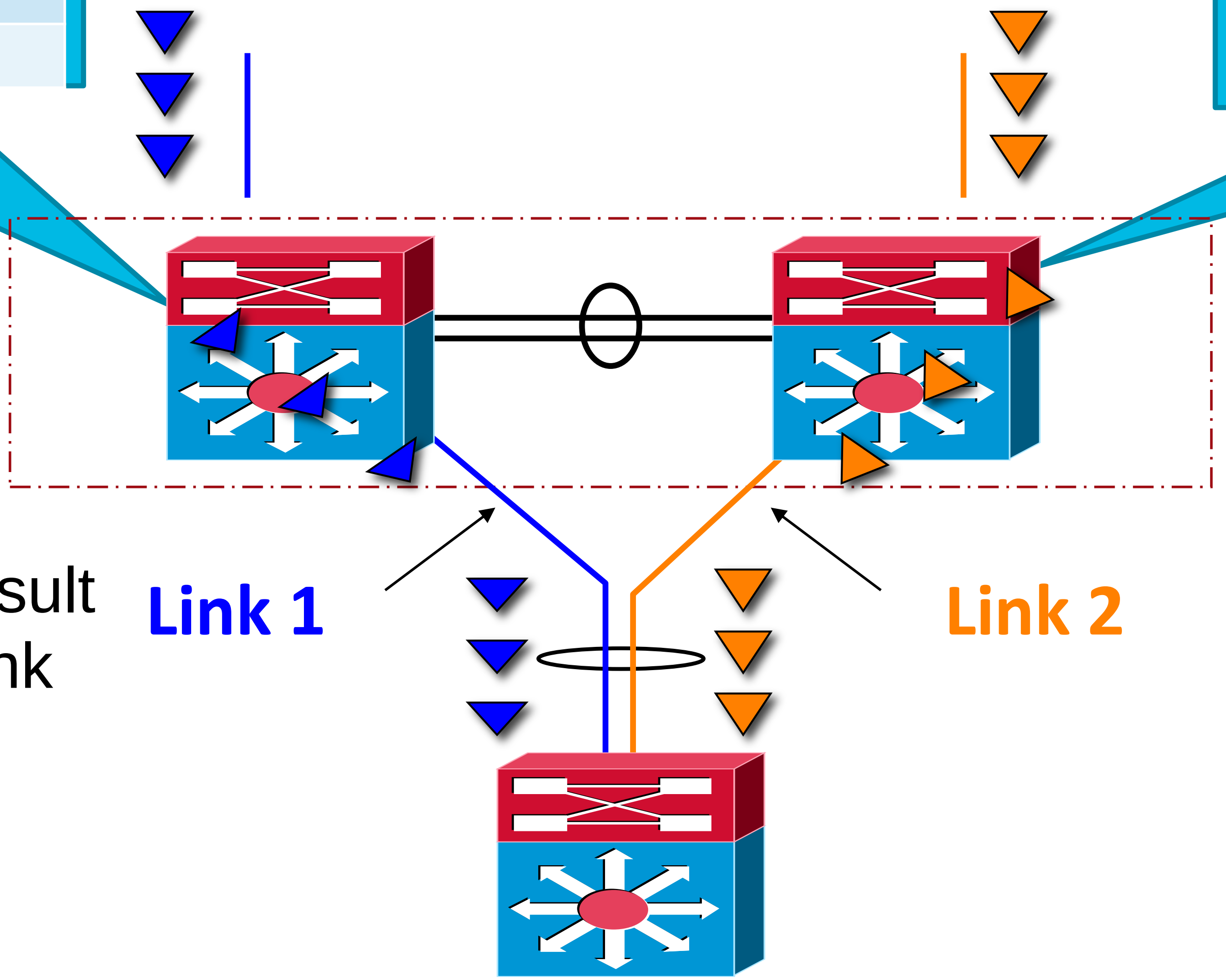
Virtual Switching System Architecture

EtherChannel Hash for MEC

Etherchannel hashing algorithms are modified in VSS to always favor locally attached interfaces

Logical Interface	Physical Interface	Result Bundle Hash (RBH) Value
PO-1	T 1/1/1	0,1,2,3,4,5,6,7
PO-1	T2/1/1	

Logical Interface	Physical Interface	Result Bundle Hash (RBH) Value
PO-1	T 1/1/1	
PO-1	T2/1/1	0,1,2,3,4,5,6,7

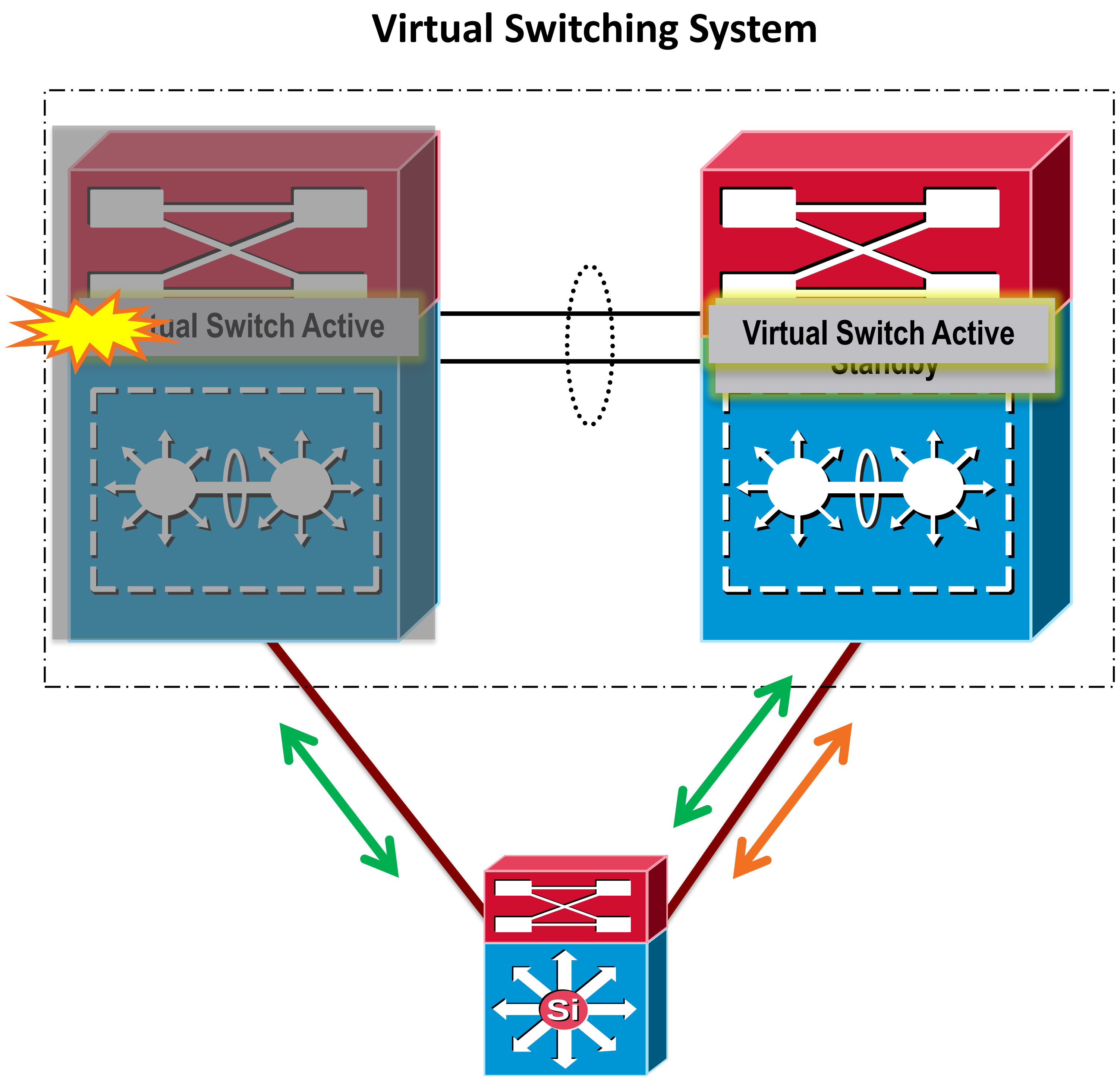


Blue Traffic flow will result in **Link 1** in the MEC link bundle

Orange Traffic flowwill result in **Link 2** in the MEC link bundle

Virtual Switching System

Inter Chassis NSF/SSO



1 Virtual Switch Active incurs a supervisor outage

2 Standby Supervisor takes over as Virtual switch Active

Virtual Switch Standby initiates graceful restart

Non Stop forwarding of packets will continue using hardware entries as Switch-2 assumes active role

NSF aware neighbours exchange updates with Virtual Switch Active

High Availability

NSF/SSO or Graceful Restart Configuration

- Non Stop Forwarding or Graceful Restart configuration is required to maintain forwarding along last known good paths
- Configuration is L3 routing protocol dependant

Example : OSPF Configuration

```
VSS#config t
VSS(config)#router ospf 1
VSS(config-router)#nsf

VSS#show ip ospf
 Routing Process "ospf 10" with ID 192.168.2.1
 Start time: 00:15:29.344, Time elapsed: 23:12:03.484
 Supports only single TOS(TOS0) routes
 External flood list length 0
 Non-Stop Forwarding enabled
 IETF NSF helper support enabled
 Cisco NSF helper support enabled
 Reference bandwidth unit is 100 mbps
```

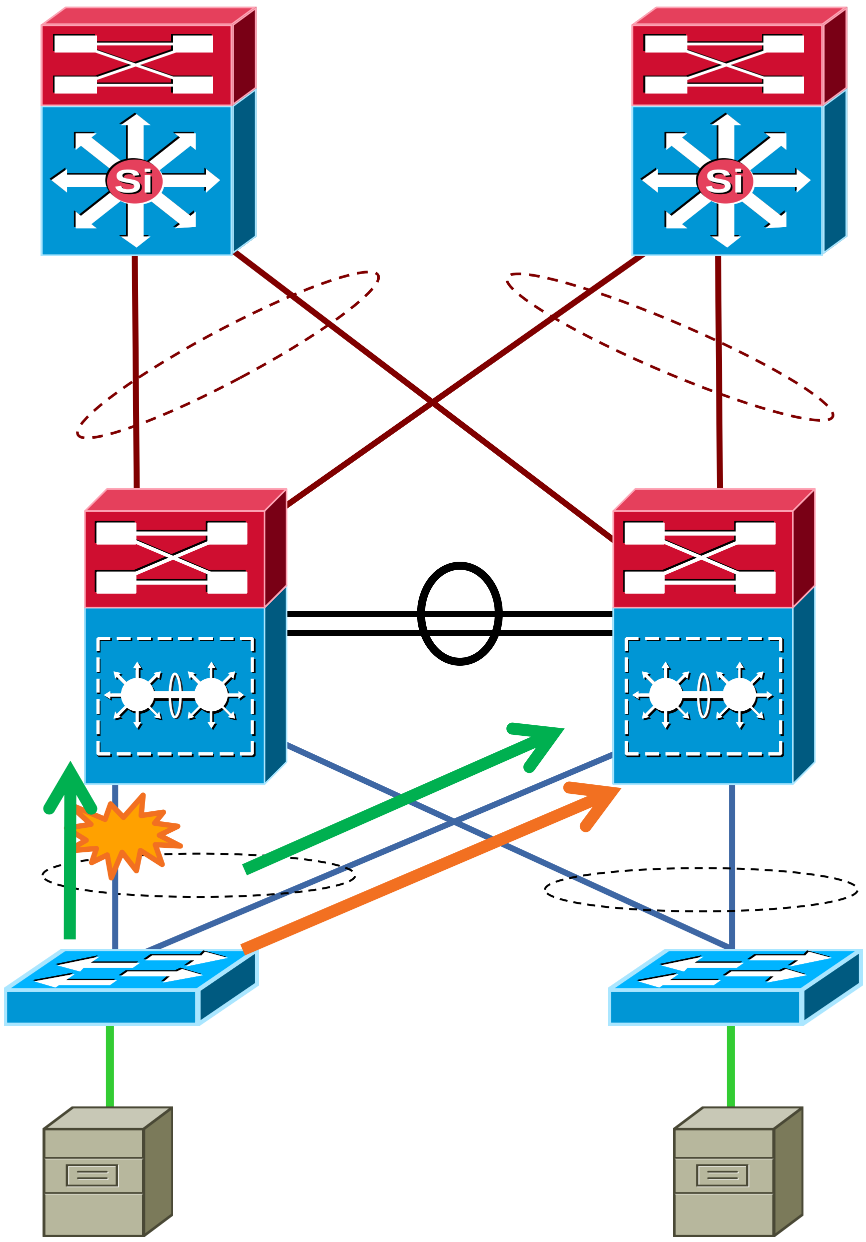
High Availability

Failure of MEC member – Upstream Traffic

- Convergence is determined by Access device link fail detection and Etherchannel convergence

- Etherchannel convergence - typically 200ms

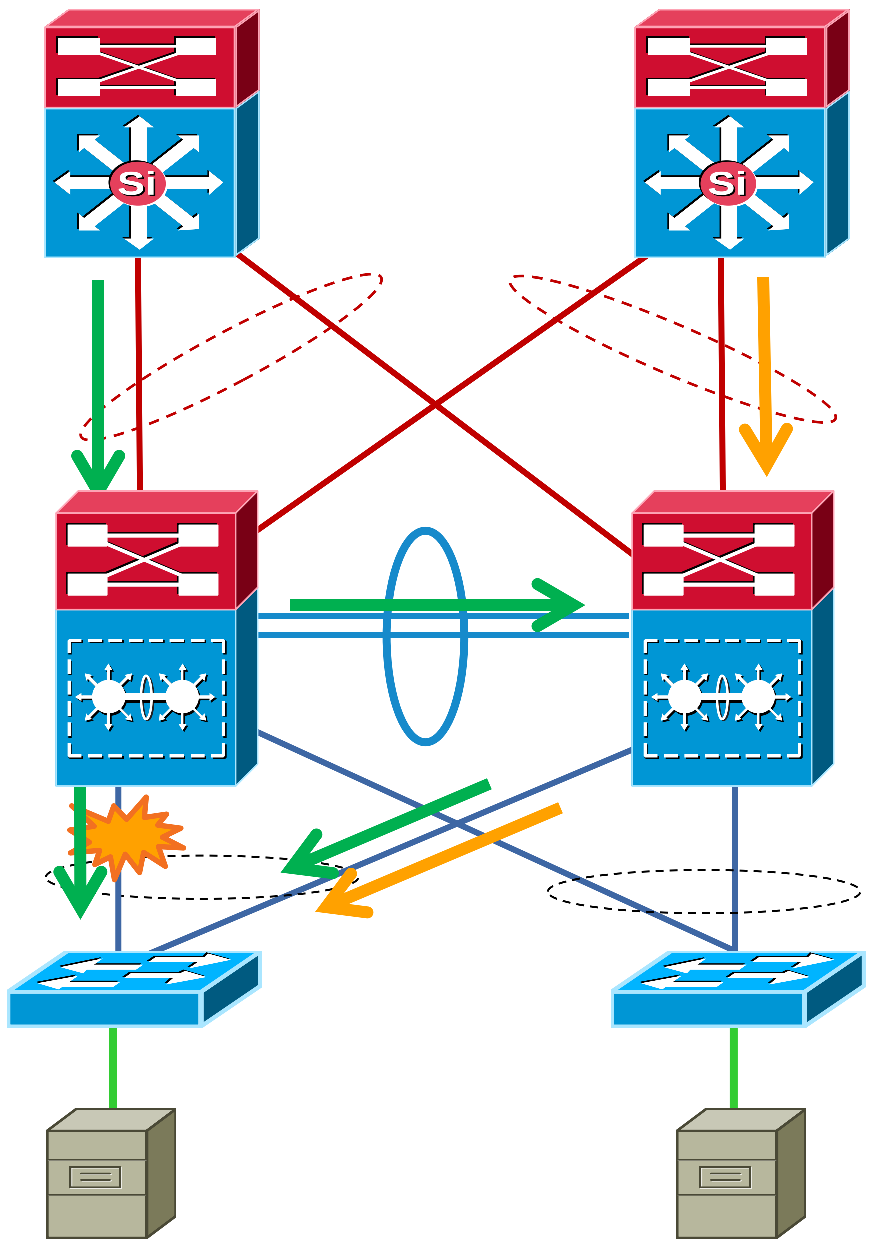
- Typically only the flows on the failed link are effected



High Availability

Failure of MEC member – Downstream Traffic

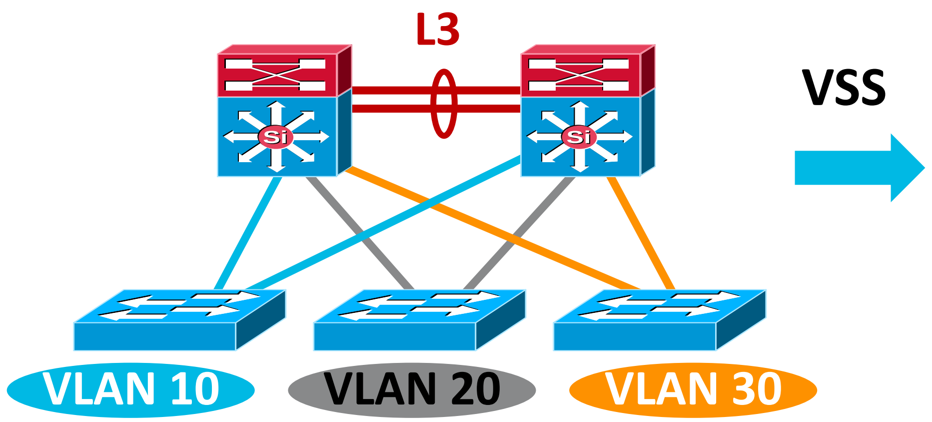
- Convergence is determined by VSS
- VSS Etherchannel convergence
  - Typically Sub - 200ms
  - Only the flows on the failed link are effected

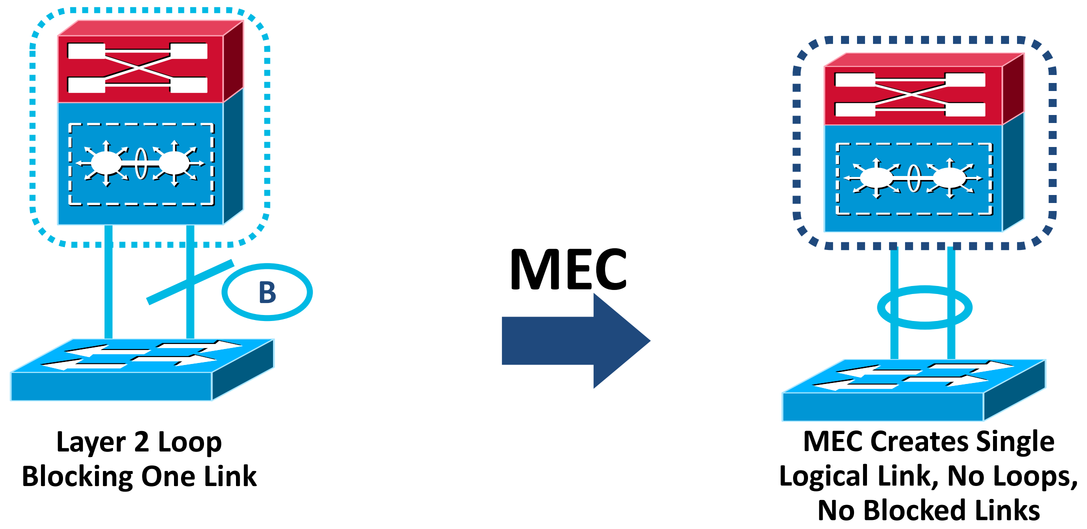


VSS-Enabled Campus Design

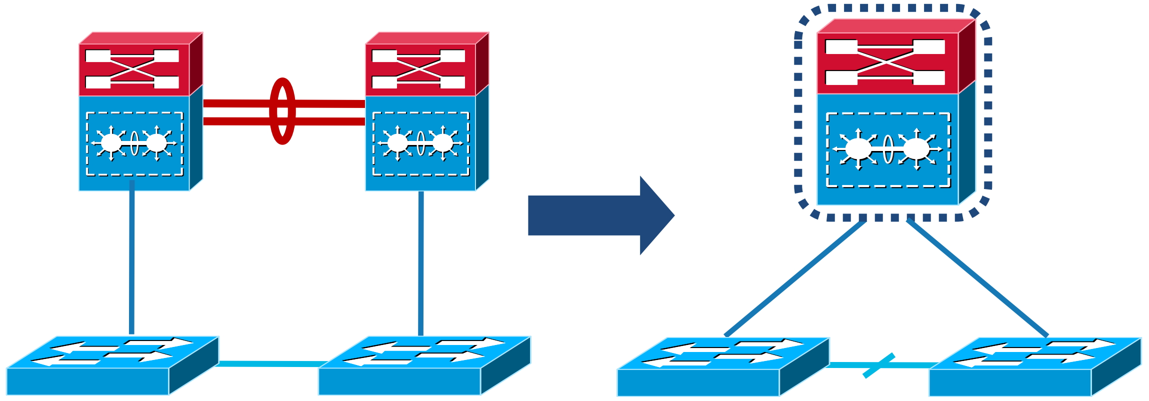
Multi-Layer Topology Considerations

- Optimised multi-layer topology (U / V shape) where VLANs do not span closets
- Deploying VSS in such topology without MEC reintroduces STP loops in the networks
- Use of MEC is recommended any time two L2 links from the same devices connected to VSS





- **U shape design with VSS**
  - Loop – blocked link
  - Downstream traffic goes over VSL link
- **Solution is to use**
  - MEC or
  - Cross-stack EtherChannel

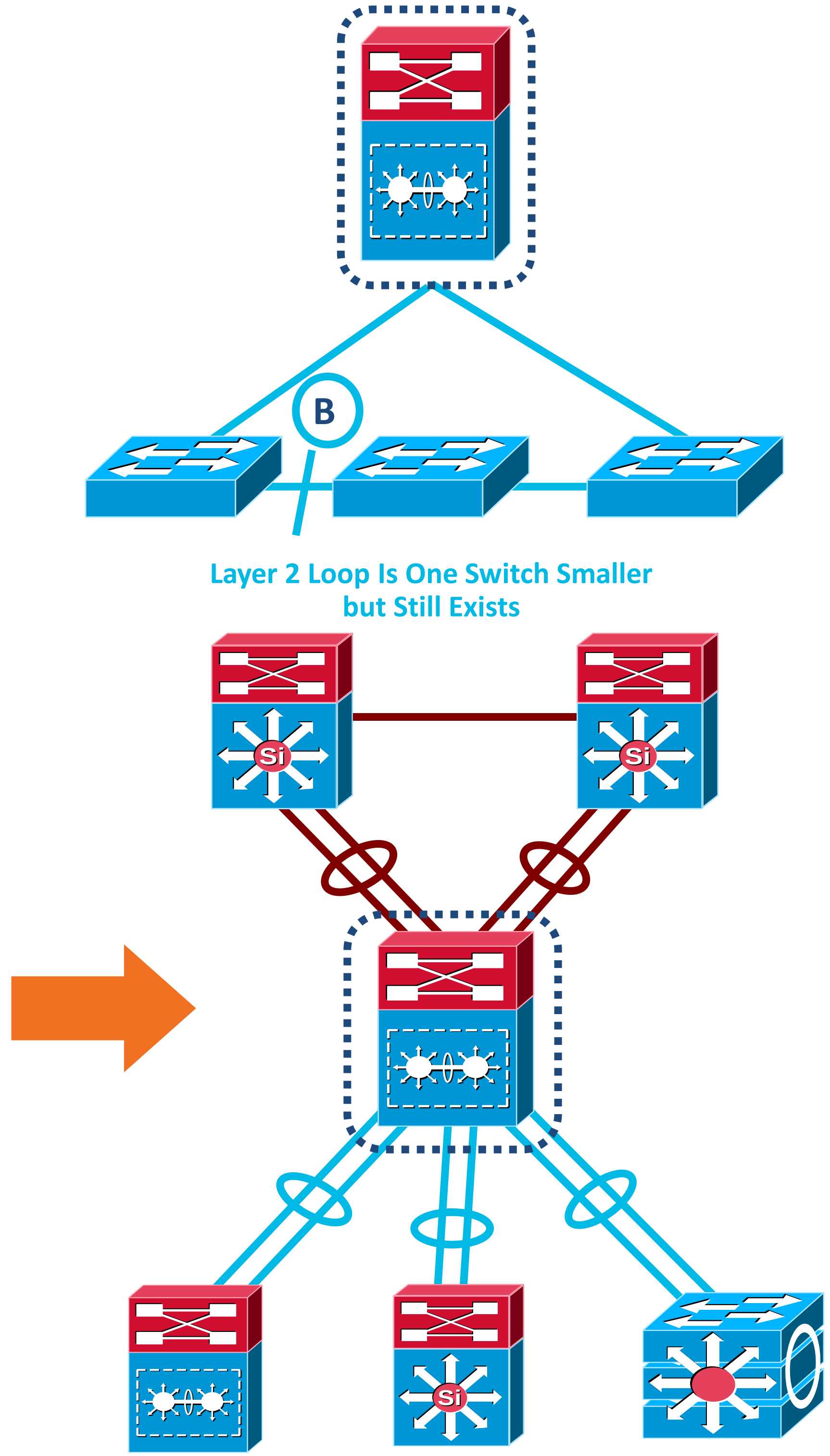


U Shape Topologies Introduces Loop with VSS

VSS-Enabled Campus Design

Multi-Layer Topology Considerations (cont.)

- Daisy chained access introduced L2 loop with an STP blocked link
- Traffic recovery times are determined by Spanning Tree recovery in the event of link or node failures
- Similarly connecting two VSS pair to a single access layer switch will also introduce the loop
- Always use star shaped topology with MEC from each device connected to VSS to
  - Avoid loops
  - Best convergence



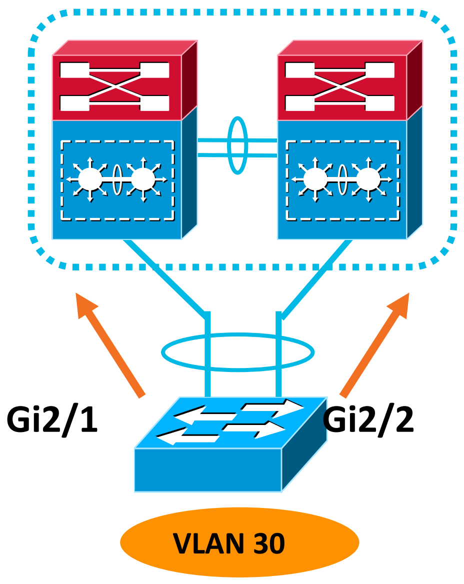
VSS-Enabled Campus Design

PAgP and LACP Best Practices

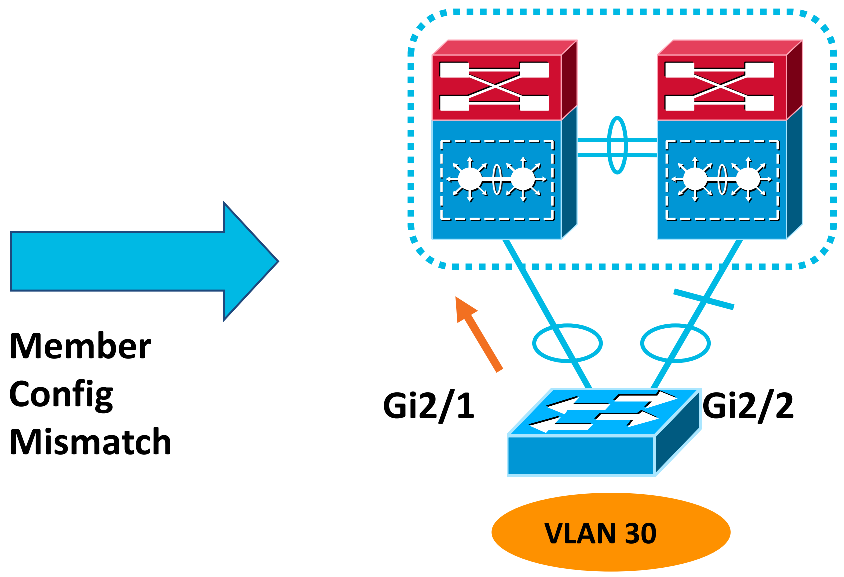
- MEC links on both member switches are managed by ACTIVE control-plane running PAgP / LACP
  - All the rules and properties of EtherChannel applies to MEC such as negotiation, link characteristics (port-type, trunk), QoS, etc.
- Do not use "on" and "off" options with PAgP or LACP protocol negotiation
  - PAgP – Run Desirable-Desirable with MEC links LACP – Run Active-Active with MEC links
- Use Default PAgP and LACP hello timer
- Do not use min-link features of LACP with VSS
- When connecting to NX-OS device – DISABLE graceful convergence in NX-OS – "no lacp graceful-convergence"

```
6500-VSS# show etherchannel 20 summary | inc Gi
Po20(SU)      LACP     Gi2/1(P)      Gi2/2(P)
6500-VSS# show spanning-tree | inc Po20
Po20              Root FWD 3       128.1667 P2p
6500-VSS# config t
VSS(config)# int gi2/2
VSS(config-if)# switchport nonegotiate
VSS(config-if) # shut
VSS(config-if)# no shut
%EC-SPSTBY-5-CANNOT_BUNDLE_LACP: Gi2/2 is not compatible with aggregators in channel
20 and cannot attach to them (trunk mode of Gi2/2 is trunk, Gi2/1 is dynamic)
%EC-SP-5-BUNDLE: Interface Gi2/2 joined port-channel Po20B ! A system generated port-
channel
6500-VSS# show etherchannel 20 summary | inc Gi
Po20(SU)      LACP     Gi2/1(P)
Po20B(SU)     LACP     Gi2/2(P) ! Bundled in separate system-generated port-channel  interface

6500-VSS# show spanning-tree | inc Po20
Po20       Root FWD   4       128.1667     P2p
Po20B      Altn BLK   4       128.1668     P2p ! Individual port running STP is blocked
```



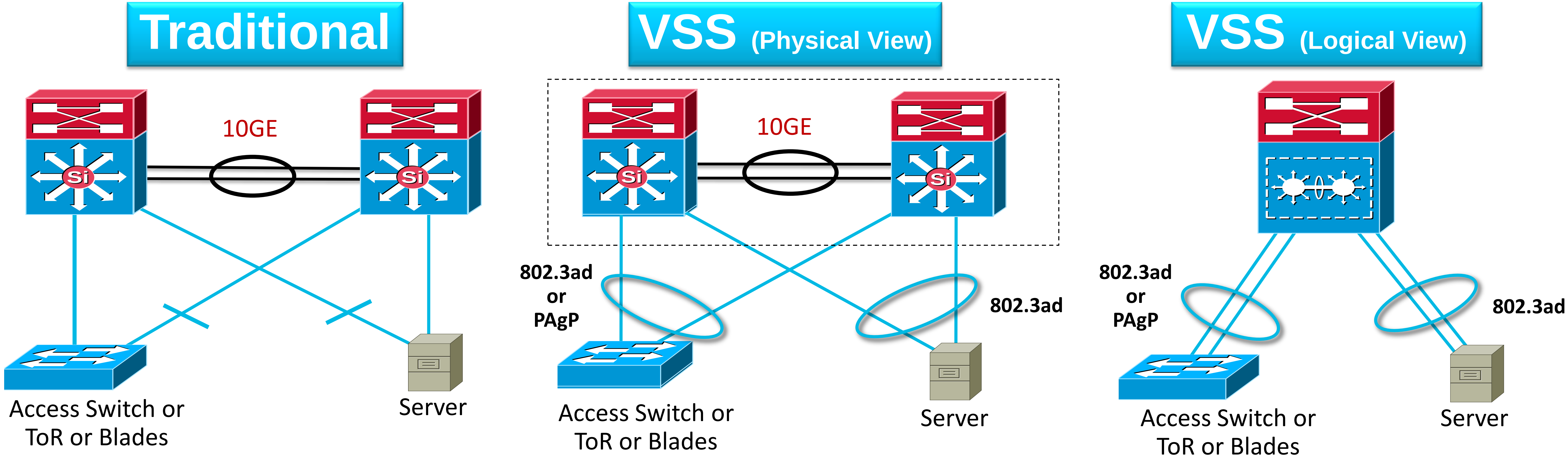
Normal LACP



Individual L2 Path

Virtual Switching System

Benefits Summary



Simplifies operational Manageability via Single point of Management, Non-loop design, minimise reliance on STP, eliminate FHRP etc

Scales system capacity with Active-Active Multi-Chassis Etherchannel (802.3ad/PAgP), no blocking links due to Spanning Tree

Minimises traffic disruption from switch or uplink failure with Deterministic subsecond Stateful and Graceful Recovery (SSO/NSF)

Fabric Path Designs

Cisco FabricPath

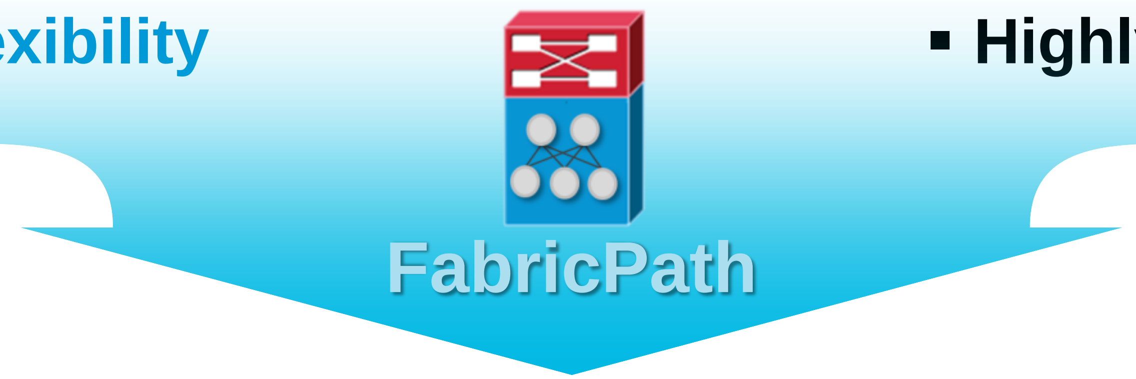
NX-OS Innovation Enhancing L2 and L3

Switching

- Easy Configuration
- Plug & Play
- Provisioning Flexibility

Routing

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



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

Cisco FabricPath

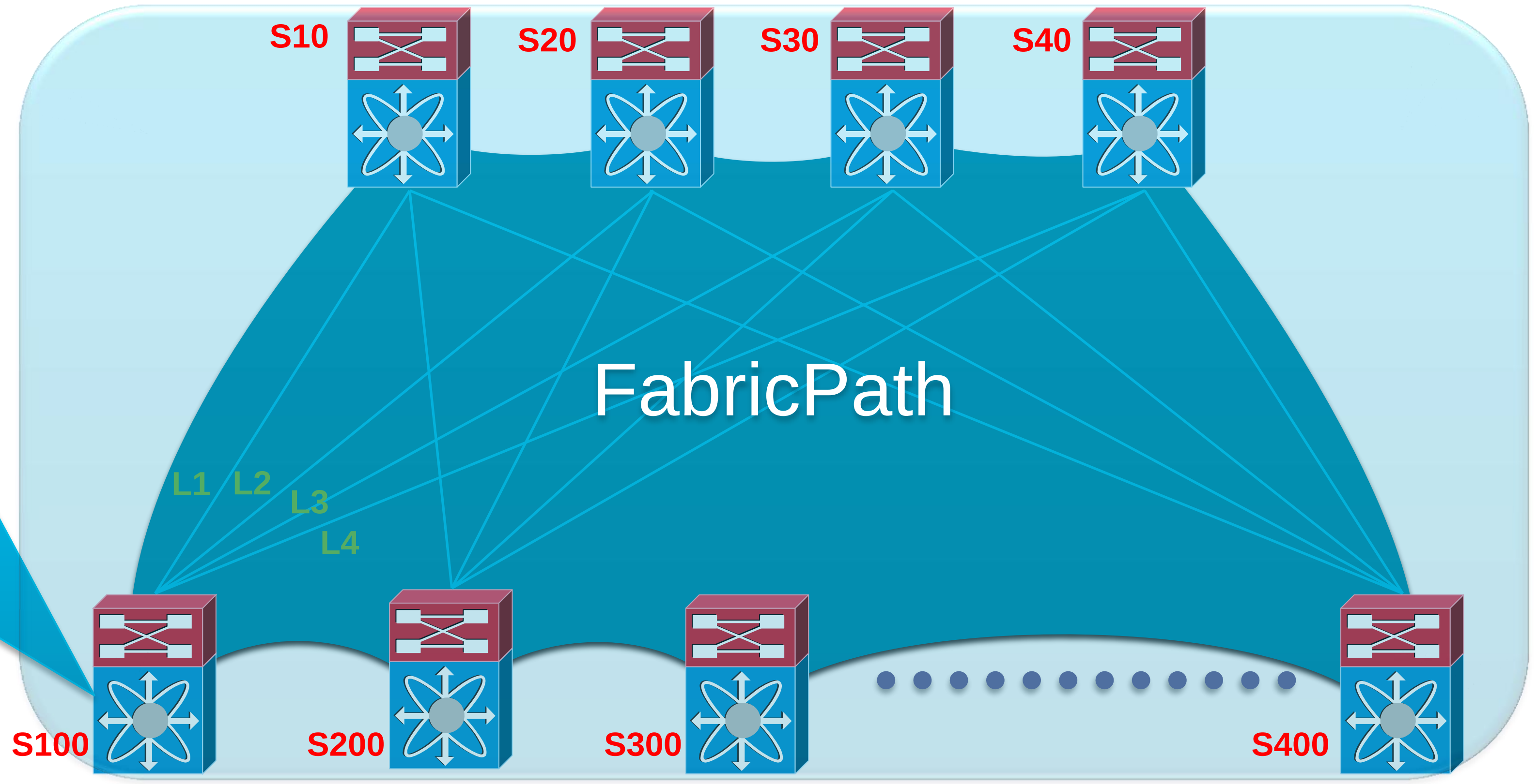
A New Control Plane – IS-IS

Plug-n-Play L2 IS-IS manages forwarding topology

- IS-IS assigns addresses to all FabricPath switches automatically
- Compute shortest, pair-wise paths
- Support equal-cost paths between any FabricPath switch pairs

FabricPath Routing Table

Switch	IF
S10	L1
S20	L2
S30	L3
S40	L4
S200	L1, L2, L3, L4
…	…
S400	L1, L2, L3, L4



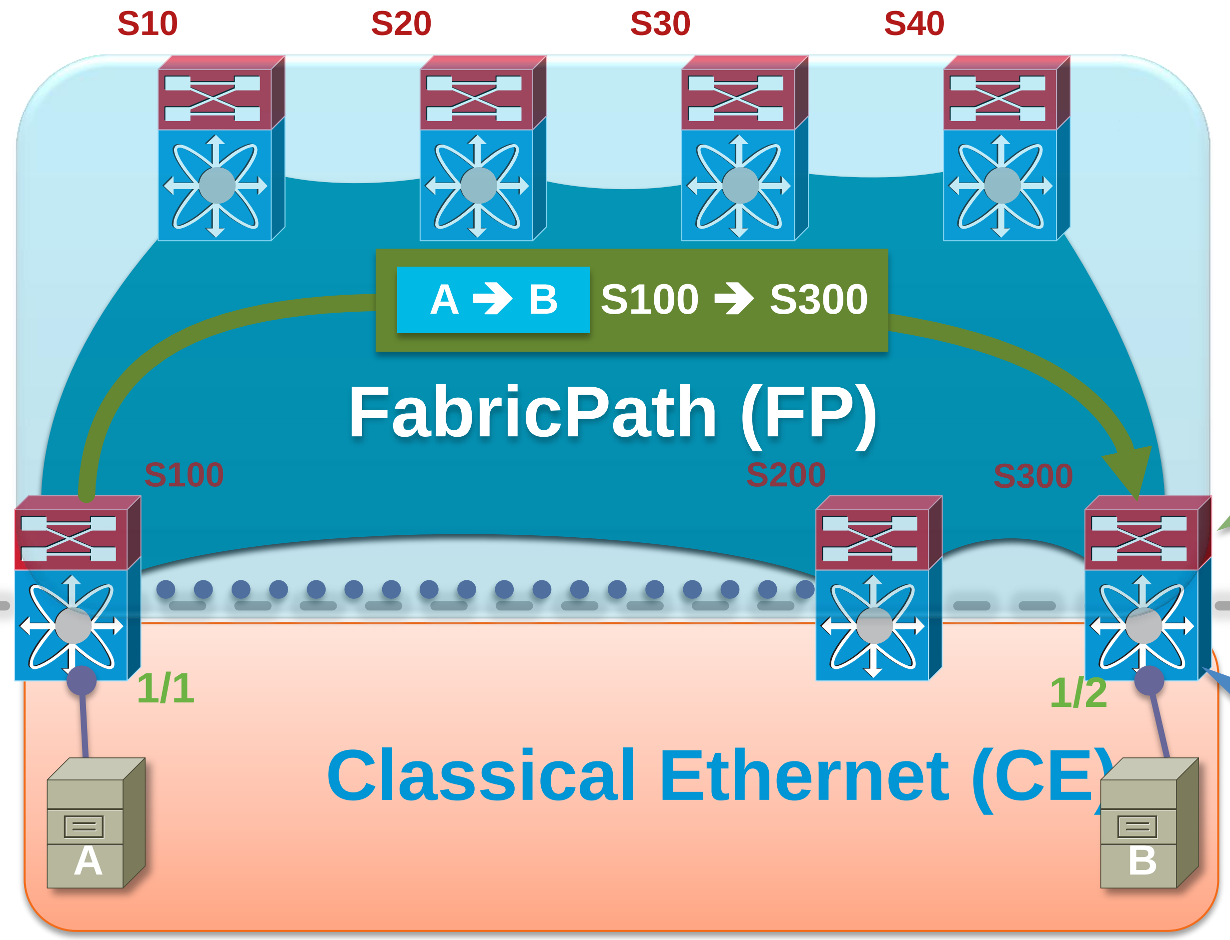
Cisco FabricPath

A New Data Plane

- The association MAC address/Switch ID is maintained at the edge

Switch ID space: Routing decisions are made based on the FabricPath routing table

MAC address space: Switching based on MAC address tables



S300: FabricPath Routing Table

Switch	IF
…	…
S100	L1, L2, L3, L4

S300: CE MAC Address Table

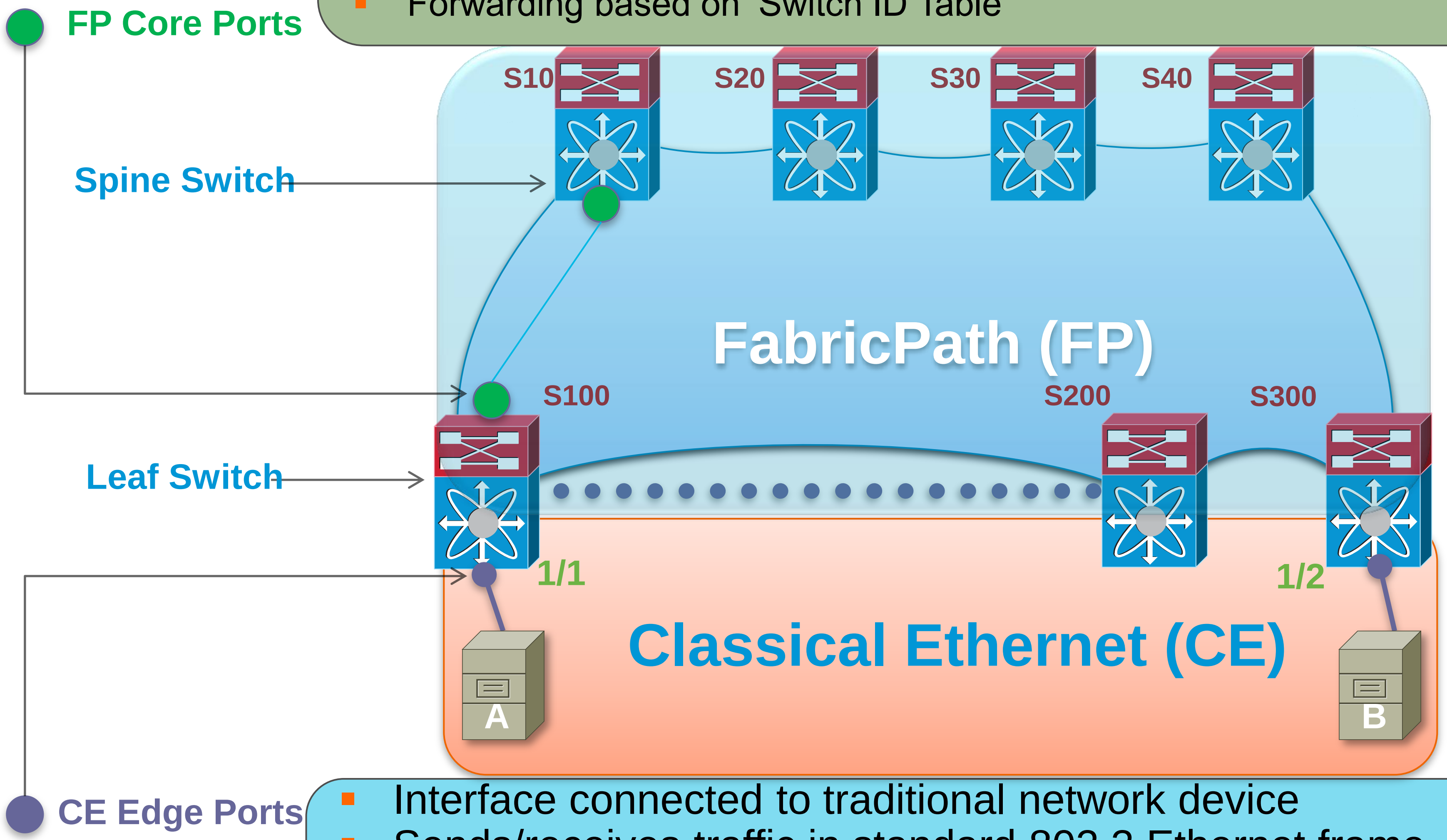
MAC	IF
B	1/2
A	S100

- Traffic is encapsulated across the Fabric

Cisco FabricPath

Terminology

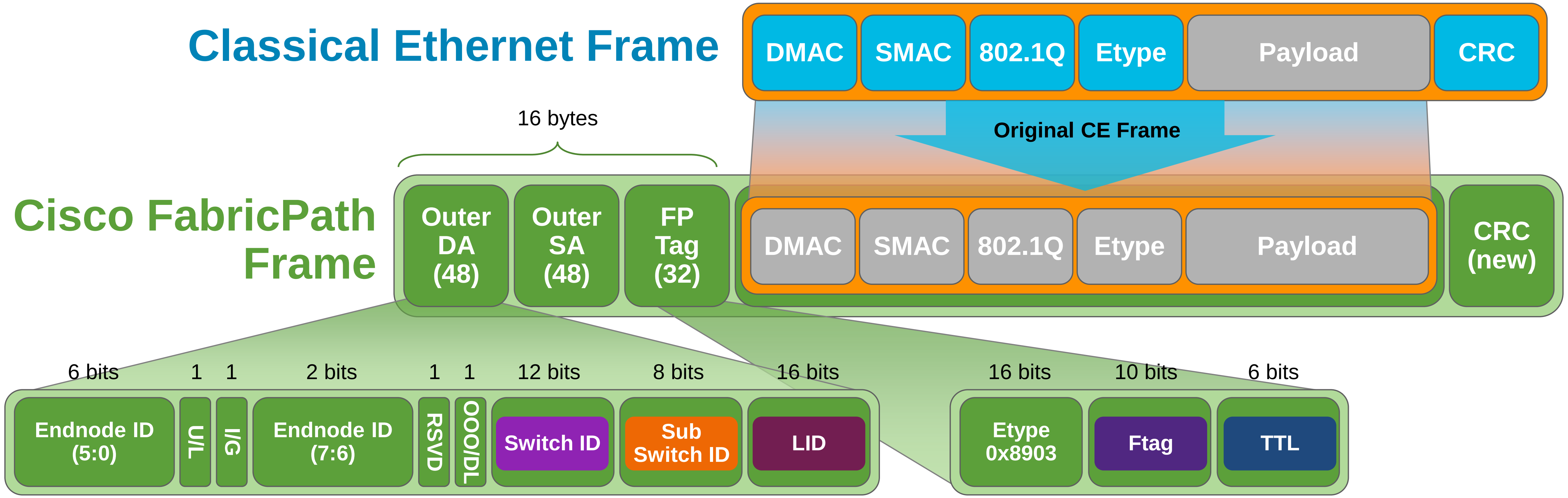
- Interface connected to another FabricPath device
- Sends/receives traffic with FabricPath header
- Does not run spanning tree
- Does not perform MAC learning!
- Exchanges topology info through L2 ISIS adjacency
- Forwarding based on ‘Switch ID Table’



- Interface connected to traditional network device
- Sends/receives traffic in standard 802.3 Ethernet frame format
- Participates in STP domain
- Forwarding based on MAC table

FabricPath Encapsulation

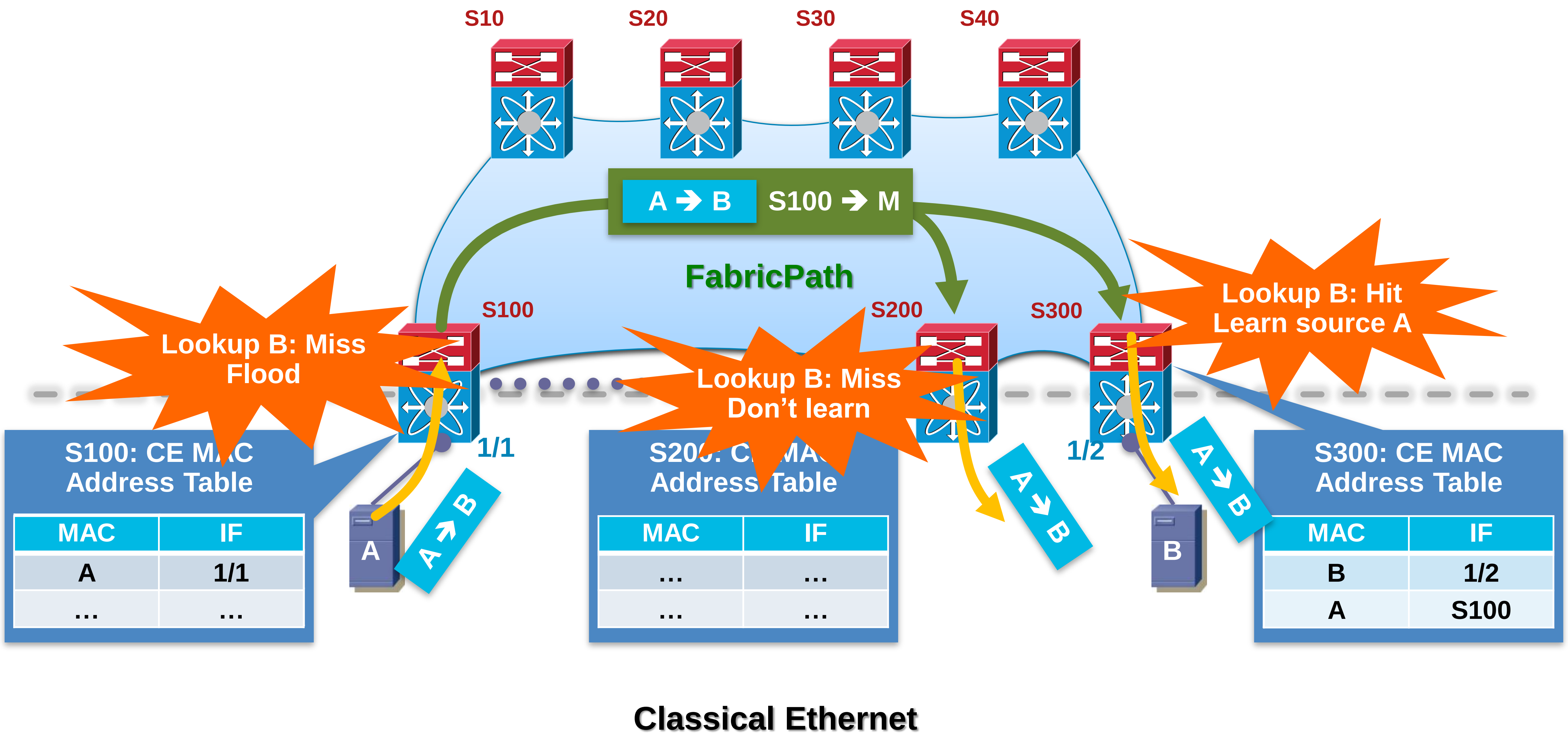
16-Byte MAC-N-MAC Header



- **Switch ID** – Unique number identifying each FabricPath switch
- **Sub-Switch ID** – Identifies devices/hosts connected via VPC+
- **LID** – Local ID, identifies the destination or source interface
- **Ftag** (Forwarding tag) – Unique number identifying topology and/or distribution tree
- **TTL** – Decremented at each switch hop to prevent frames looping infinitely

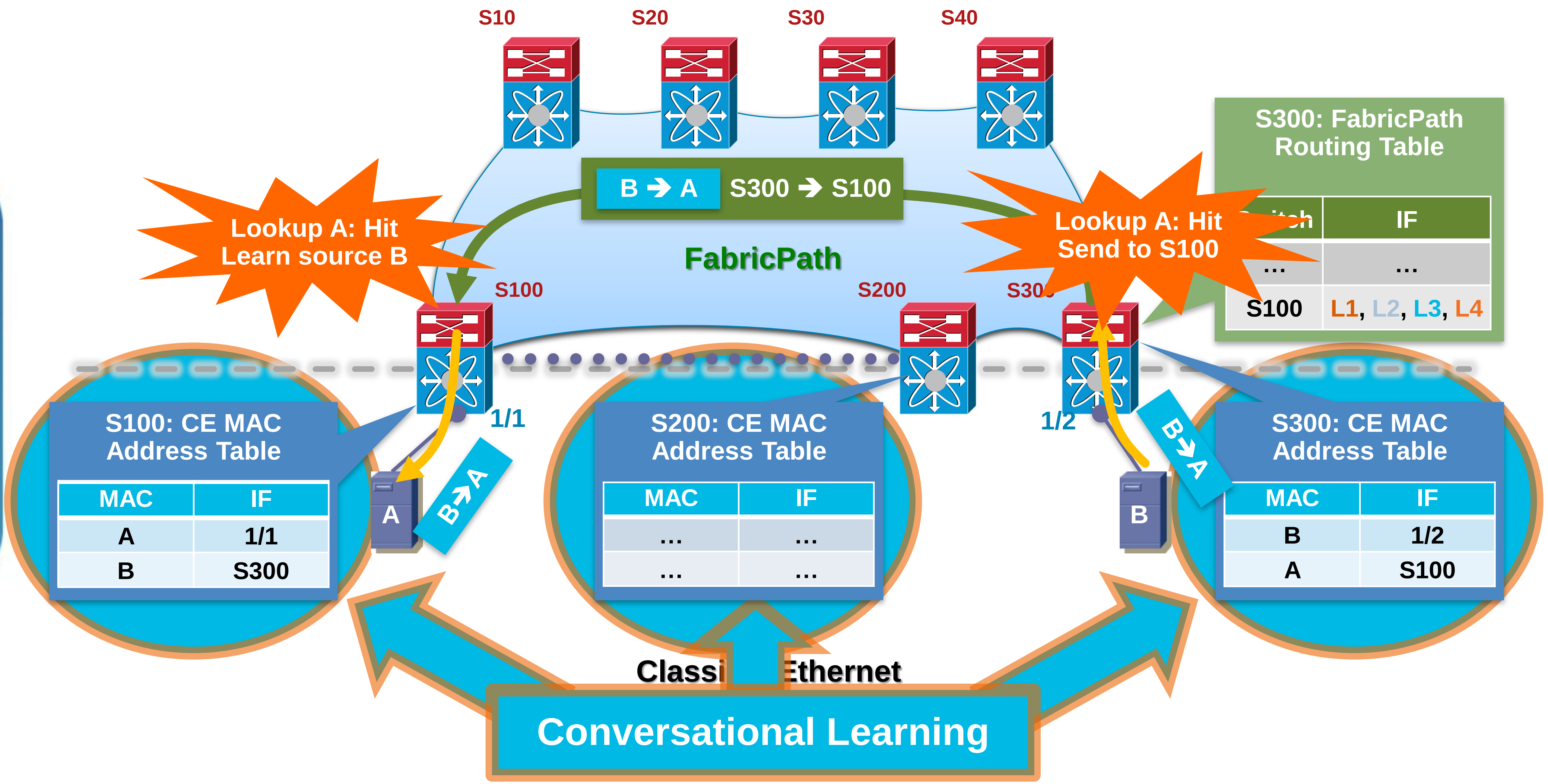
FabricPath – Key Concept #1

Conversational MAC Learning



FabricPath – Key Concept #1

Conversational MAC Learning



FabricPath – Key Concept #2

It's a Routed Network

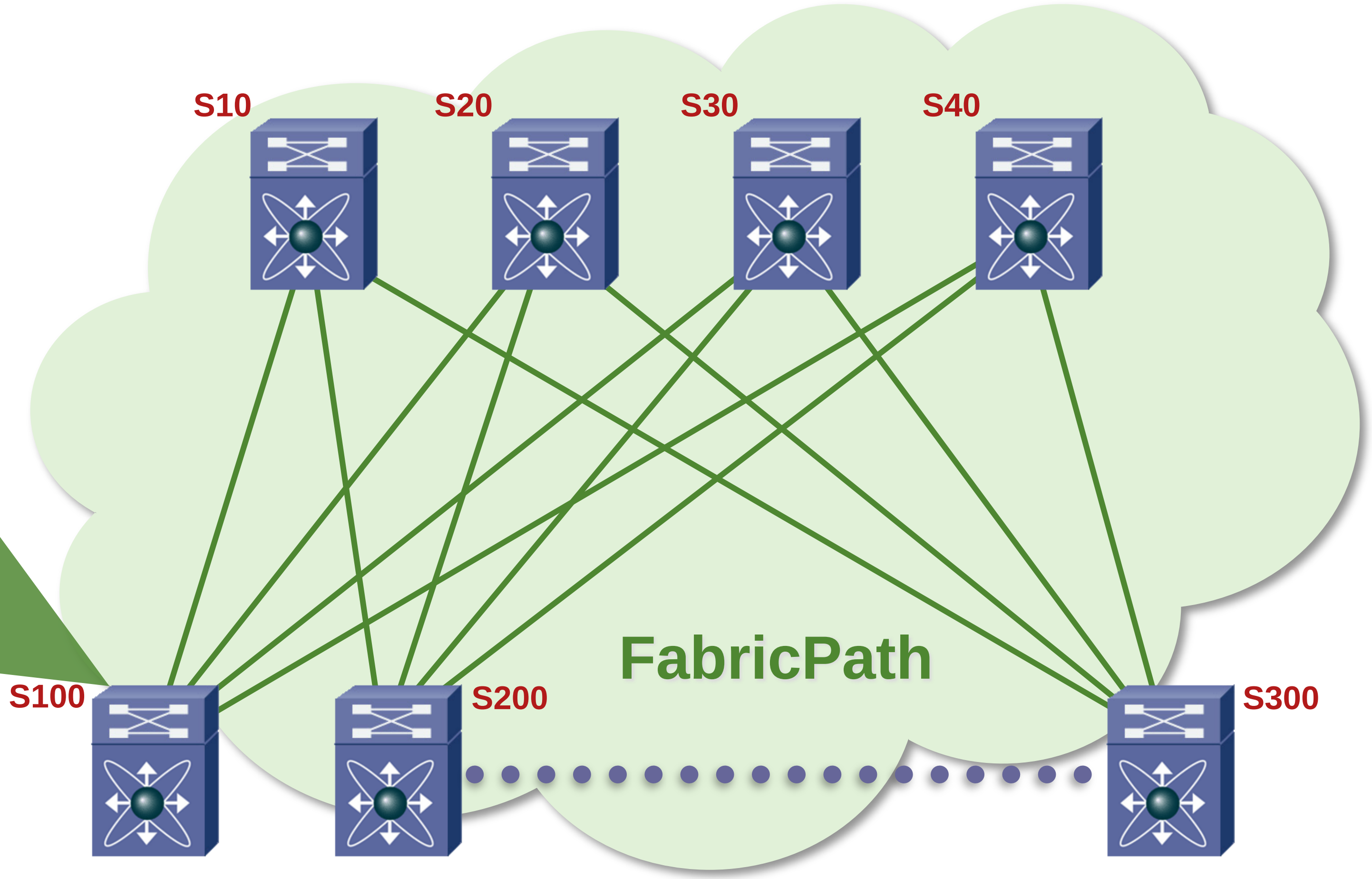
- Describes shortest (best) paths to each Switch ID based on link metrics
- Equal-cost paths supported between FabricPath switches

One 'best' path to S10 (via L1) →

Four equal-cost paths to S300 →

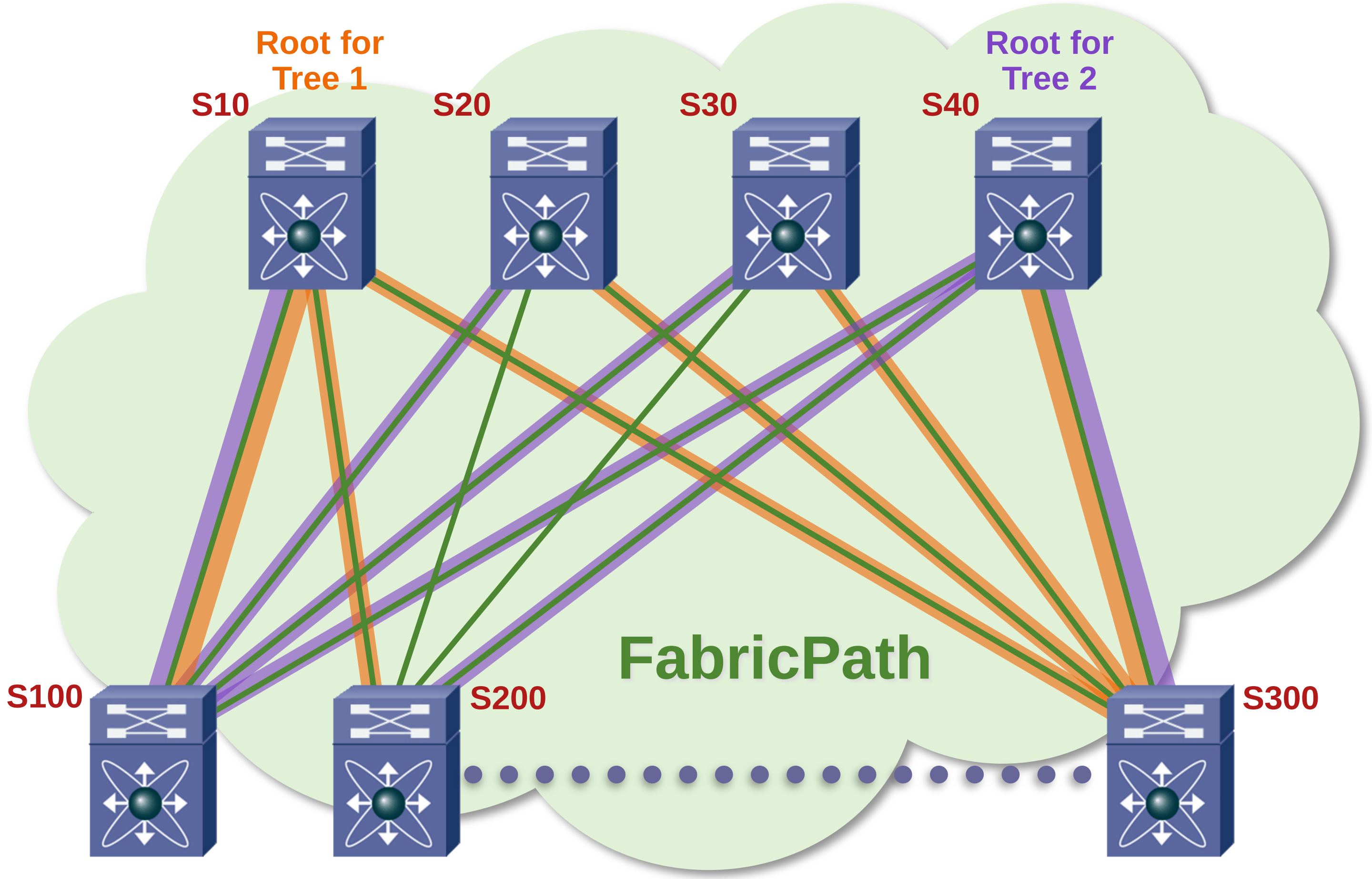
FabricPath Routing Table on S100

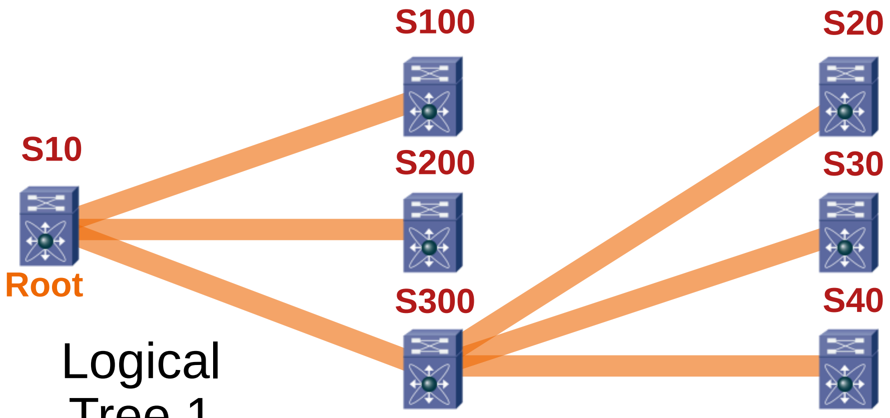
Switch	IF
S10	L1
S20	L2
S30	L3
S40	L4
S200	L1, L2, L3, L4
…	…
S300	L1, L2, L3, L4



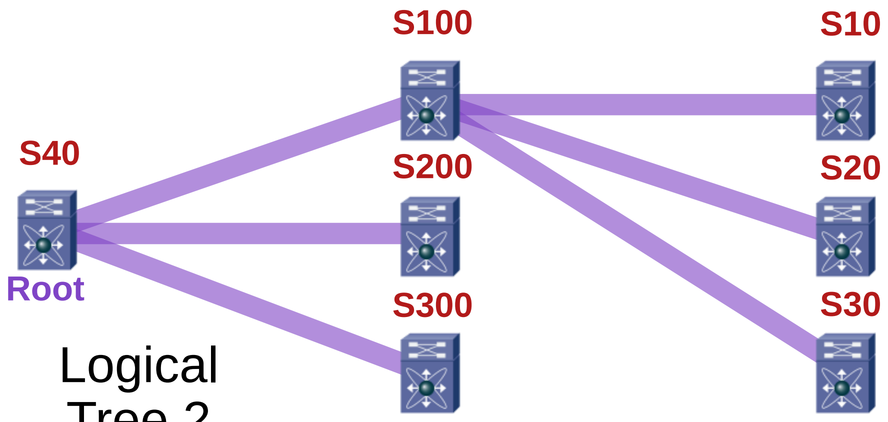
FabricPath – Key Concept #3

Multicasting





Logical Tree 1

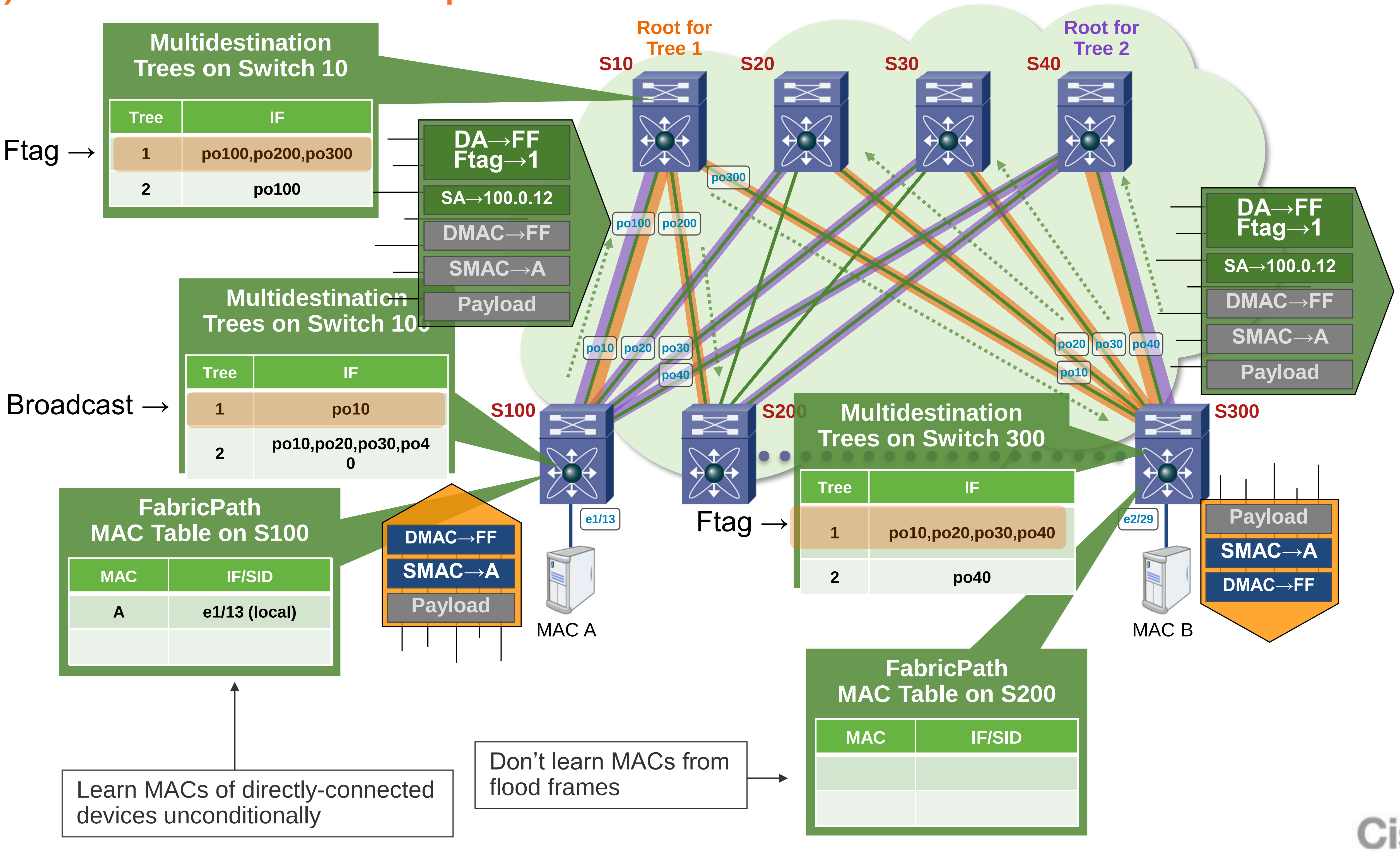


Logical Tree 2

- Multi-destination traffic constrained to loop-free trees touching all FabricPath switches
- Root switch elected for each multi-destination tree in the FabricPath domain
- Loop-free tree built from each Root assigned a network-wide identifier (Ftag)
- Support for multiple multi-destination trees provides multipathing for multi-destination traffic
  - Two multi-destination trees supported in NX-OS release 5.1

Putting It All Together – Host A to Host B

(1) Broadcast ARP Request



Multidestination Trees on Switch 10

Tree	IF
1	po100,po200,po300
2	po100

Multidestination Trees on Switch 100

Tree	IF
1	po10
2	po10,po20,po30,po40

Multidestination Trees on Switch 300

Tree	IF
1	po10,po20,po30,po40
2	po40

FabricPath MAC Table on S100

MAC	IF/SID
A	e1/13 (local)

FabricPath MAC Table on S200

MAC	IF/SID

Putting It All Together – Host A to Host B

MAC Address Table After the First ARP Frame

For Your Reference

- **S100:**

```
– S100# sh mac address-table dynamic
– Legend:
–         * - primary entry, G - Gateway MAC, (R) - Routed MAC, O - Overlay MAC
–         age - seconds since last seen,+ - primary entry using vPC Peer-Link
–    VLAN     MAC Address      Type      age     Secure NTFY Ports/SWID.SSID.LID
– ---------+-----------------+--------+---------+------+----+------------------
– * 10       0000.0000.000a    dynamic   0           F    F  Eth1/13
```

MAC A learned as local entry on e1/13

- **S10 (and S20, S30, S40, S200):**

```
– S10# sh mac address-table dynamic
– Legend:
–         * - primary entry, G - Gateway MAC, (R) - Routed MAC, O - Overlay MAC
–         age - seconds since last seen,+ - primary entry using vPC Peer-Link
–    VLAN     MAC Address      Type      age     Secure NTFY Ports/SWID.SSID.LID
– ---------+-----------------+--------+---------+------+----+------------------
```

MAC A not learned on other switches

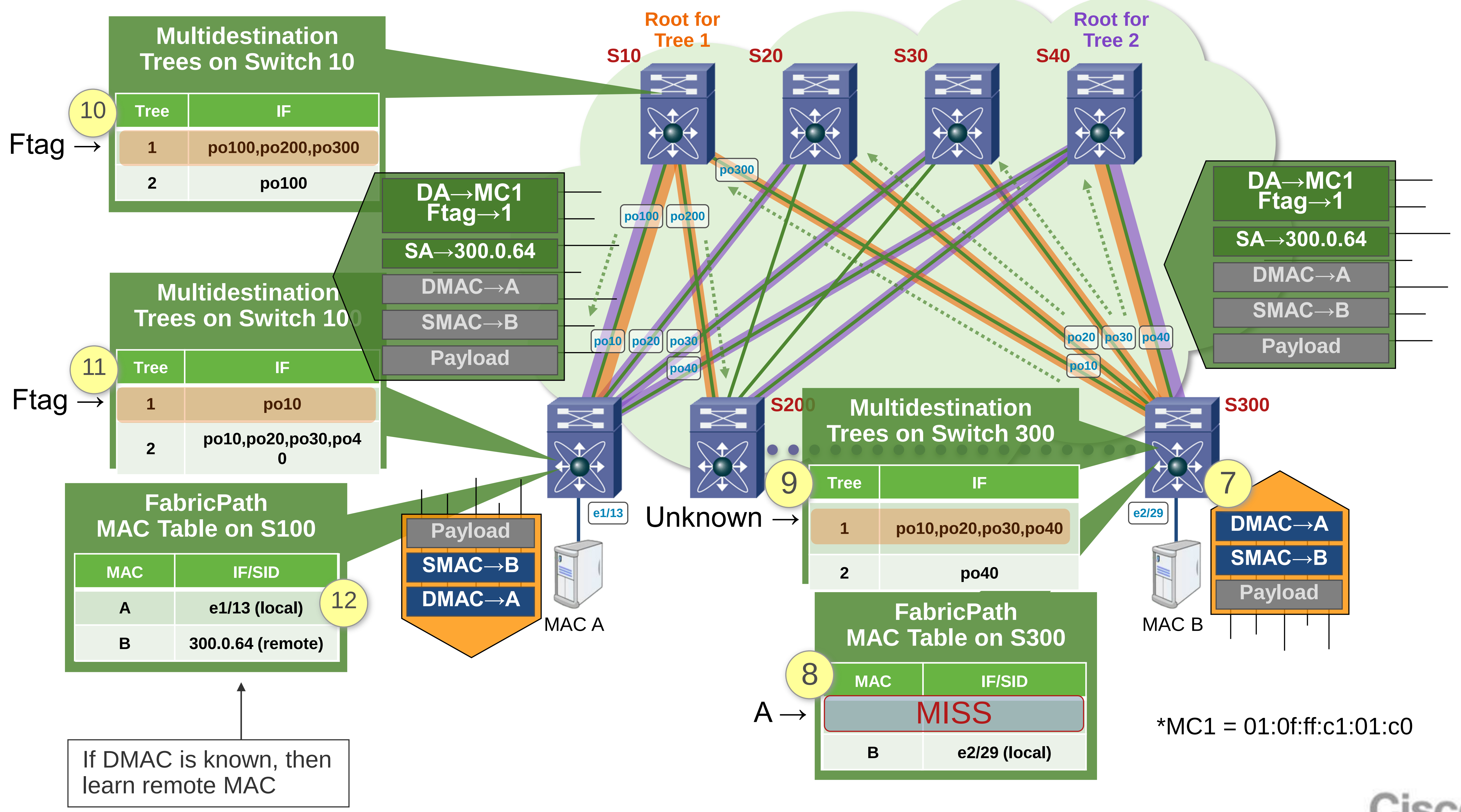
- **S300:**

```
– S300# sh mac address-table dynamic
– Legend:
–         * - primary entry, G - Gateway MAC, (R) - Routed MAC, O - Overlay MAC
–         age - seconds since last seen,+ - primary entry using vPC Peer-Link
–    VLAN     MAC Address      Type      age     Secure NTFY Ports/SWID.SSID.LID
– ---------+-----------------+--------+---------+------+----+------------------

– S300#
```

Putting It All Together – Host A to Host B

(2) Broadcast ARP Reply



Multidestination Trees on Switch 10

Tree	IF
1	po100,po200,po300
2	po100

Multidestination Trees on Switch 100

Tree	IF
1	po10
2	po10,po20,po30,po40

FabricPath MAC Table on S100

MAC	IF/SID
A	e1/13 (local)
B	300.0.64 (remote)

If DMAC is known, then learn remote MAC

Multidestination Trees on Switch 300

Tree	IF
1	po10,po20,po30,po40
2	po40

FabricPath MAC Table on S300

MAC	IF/SID
MISS	
B	e2/29 (local)

*MC1 = 01:0f:ff:c1:01:c0

Putting It All Together – Host A to Host B

MAC Address Table After the First ARP Frame

For Your Reference

- **S100:**

```
S100# sh mac address-table dynamic
Legend:
        * - primary entry, G - Gateway MAC, (R) - Routed MAC, O - Overlay MAC
        age - seconds since last seen,+ - primary entry using vPC Peer-Link
   VLAN     MAC Address      Type      age     Secure NTFY Ports/SWID.SSID.LID
---------+-----------------+--------+---------+------+----+------------------
* 10       0000.0000.000a    dynamic   90         F    F  Eth1/13
  10       0000.0000.000b    dynamic   60         F    F  300.0.64

S100#
```

S100 learns MAC B as remote entry reached through S300

- **S300:**

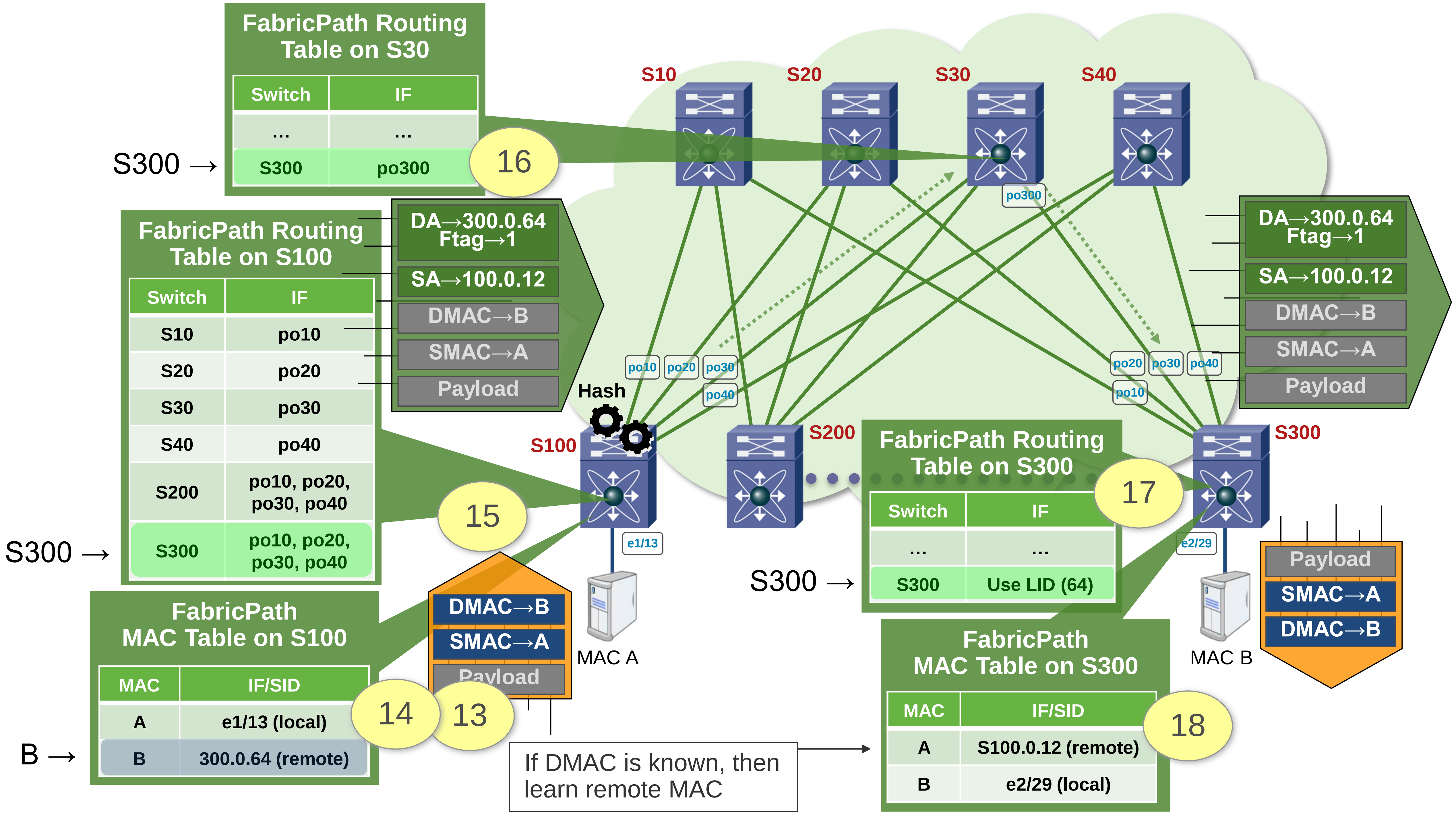
```
S300# sh mac address-table dynamic
Legend:
        * - primary entry, G - Gateway MAC, (R) - Routed MAC, O - Overlay MAC
        age - seconds since last seen,+ - primary entry using vPC Peer-Link
   VLAN     MAC Address      Type      age     Secure NTFY Ports/SWID.SSID.LID
---------+-----------------+--------+---------+------+----+------------------
* 10       0000.0000.000b    dynamic   0          F    F  Eth2/29

S300#
```

MAC B learned as local entry on e2/29

Putting It All Together – Host A to Host B

Unicast Data – Routed



FabricPath Routing Table on S30

Switch	IF
…	…
S300	po300

FabricPath Routing Table on S100

Switch	IF
S10	po10
S20	po20
S30	po30
S40	po40
S200	po10, po20, po30, po40
S300	po10, po20, po30, po40

FabricPath MAC Table on S100

MAC	IF/SID
A	e1/13 (local)
B	300.0.64 (remote)

FabricPath Routing Table on S300

Switch	IF
…	…
S300	Use LID (64)

FabricPath MAC Table on S300

MAC	IF/SID
A	S100.0.12 (remote)
B	e2/29 (local)

Putting It All Together – Host A to Host B

Unicast Forwarding

For Your Reference

- **S100:**

```
S100# sh mac address-table dynamic
Legend:
        * - primary entry, G - Gateway MAC, (R) - Routed MAC, O - Overlay MAC
        age - seconds since last seen,+ - primary entry using vPC Peer-Link
   VLAN     MAC Address      Type      age     Secure NTFY Ports/SWID.SSID.LID
---------+-----------------+--------+---------+------+----+------------------
* 10       0000.0000.000a    dynamic   90         F     F  Eth1/13
  10       0000.0000.000b    dynamic   60         F     F  300.0.64

S100#
```

- **S300:**

```
S300# sh mac address-table dynamic
Legend:
        * - primary entry, G - Gateway MAC, (R) - Routed MAC, O - Overlay MAC
        age - seconds since last seen,+ - primary entry using vPC Peer-Link
   VLAN     MAC Address      Type      age     Secure NTFY Ports/SWID.SSID.LID
---------+-----------------+--------+---------+------+----+------------------
  10       0000.0000.000a    dynamic   30         F     F  100.0.12
* 10       0000.0000.000b    dynamic   90         F     F  Eth2/29

S300#
```

S100 learns MAC A as remote entry reached through S100

Putting It All Together – Host A to Host B

Unicast Forwarding

```
S100# sh fabricpath route
FabricPath Unicast Route Table
'a/b/c' denotes ftag/switch-id/subswitch-id
'[x/y]' denotes [admin distance/metric]
ftag 0 is local ftag
subswitch-id 0 is default subswitch-id


FabricPath Unicast Route Table for Topology-Default

0/100/0, number of next-hops: 0
        via ---- , [60/0], 0 day/s 04:43:51, local
1/10/0, number of next-hops: 1
        via Po10, [115/20], 0 day/s 02:24:02, isis_fabricpath-default
1/20/0, number of next-hops: 1
        via Po20, [115/20], 0 day/s 04:43:25, isis_fabricpath-default
1/30/0, number of next-hops: 1
        via Po30, [115/20], 0 day/s 04:43:25, isis_fabricpath-default
1/40/0, number of next-hops: 1
        via Po40, [115/20], 0 day/s 04:43:25, isis_fabricpath-default
1/200/0, number of next-hops: 4
        via Po10, [115/40], 0 day/s 02:24:02, isis_fabricpath-default
        via Po20, [115/40], 0 day/s 04:43:06, isis_fabricpath-default
        via Po30, [115/40], 0 day/s 04:43:06, isis_fabricpath-default
        via Po40, [115/40], 0 day/s 04:43:06, isis_fabricpath-default
1/300/0, number of next-hops: 4
        via Po10, [115/40], 0 day/s 02:24:02, isis_fabricpath-default
        via Po20, [115/40], 0 day/s 04:43:25, isis_fabricpath-default
        via Po30, [115/40], 0 day/s 04:43:25, isis_fabricpath-default
        via Po40, [115/40], 0 day/s 04:43:25, isis_fabricpath-default
S100#
```

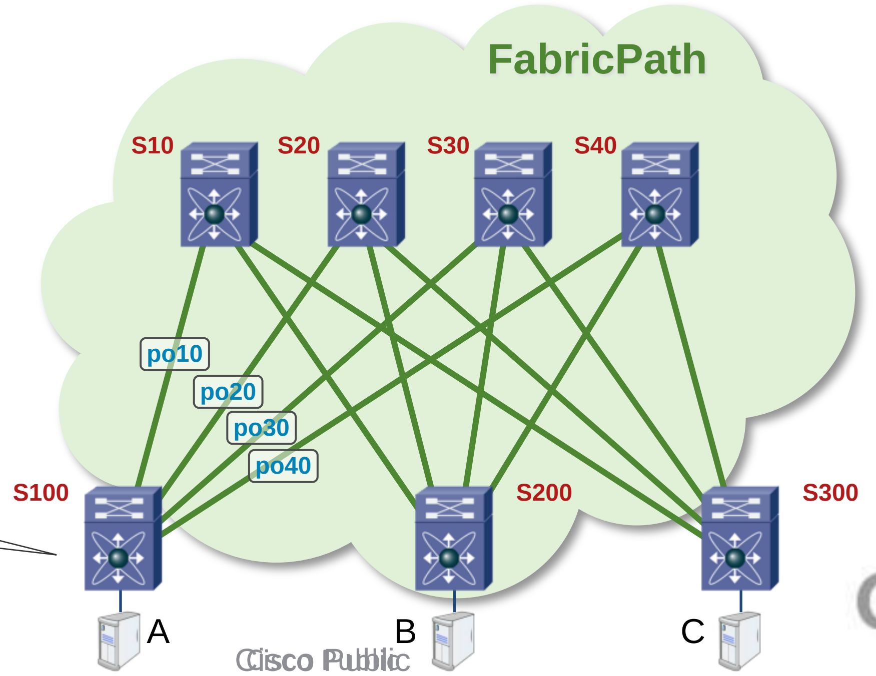
Topology (ftag), Switch ID, Sub-Switch ID

Administrative distance, routing metric

Route age

Client protocol

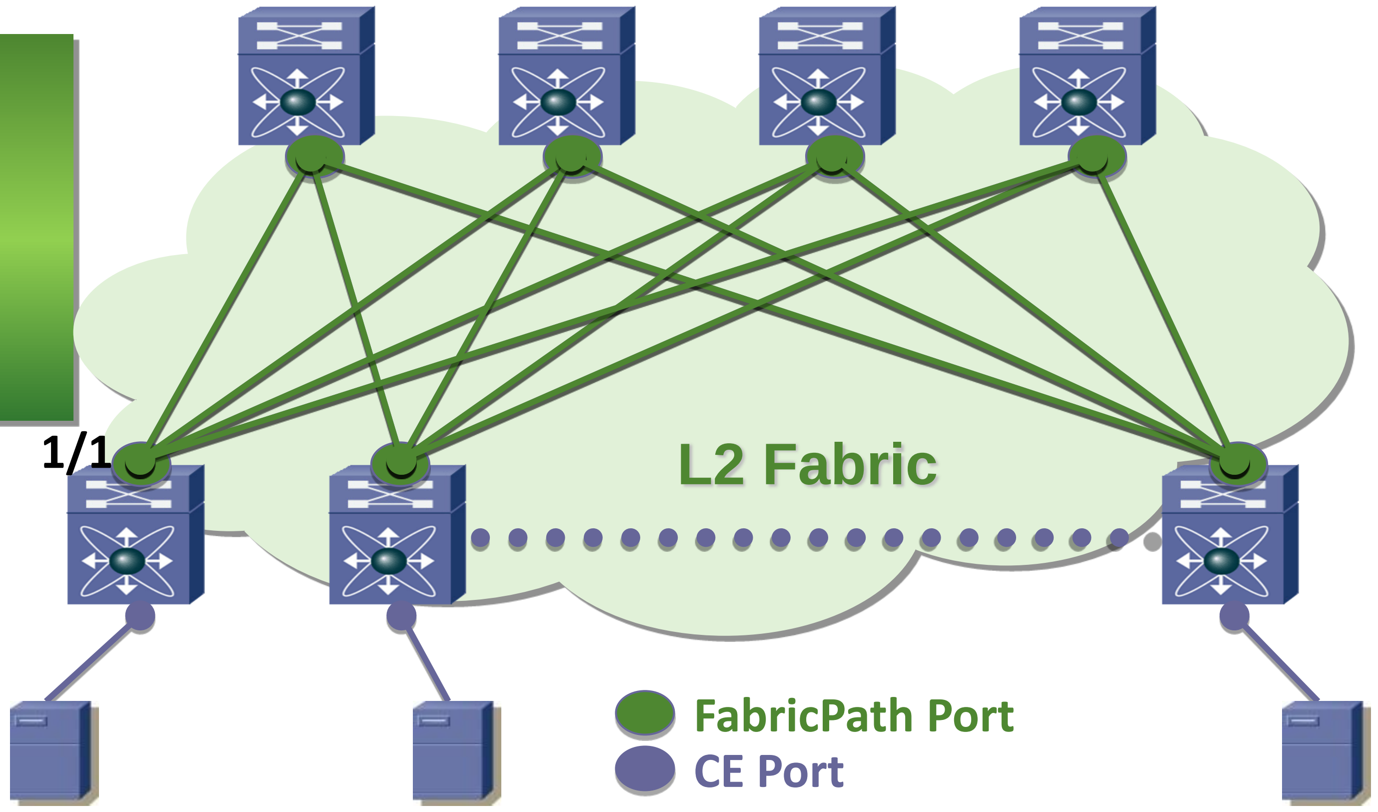
Next-hop interface(s)



FabricPath is Simple

- No L2 IS-IS configuration required
- Single control protocol for unicast, multicast, vlan pruning

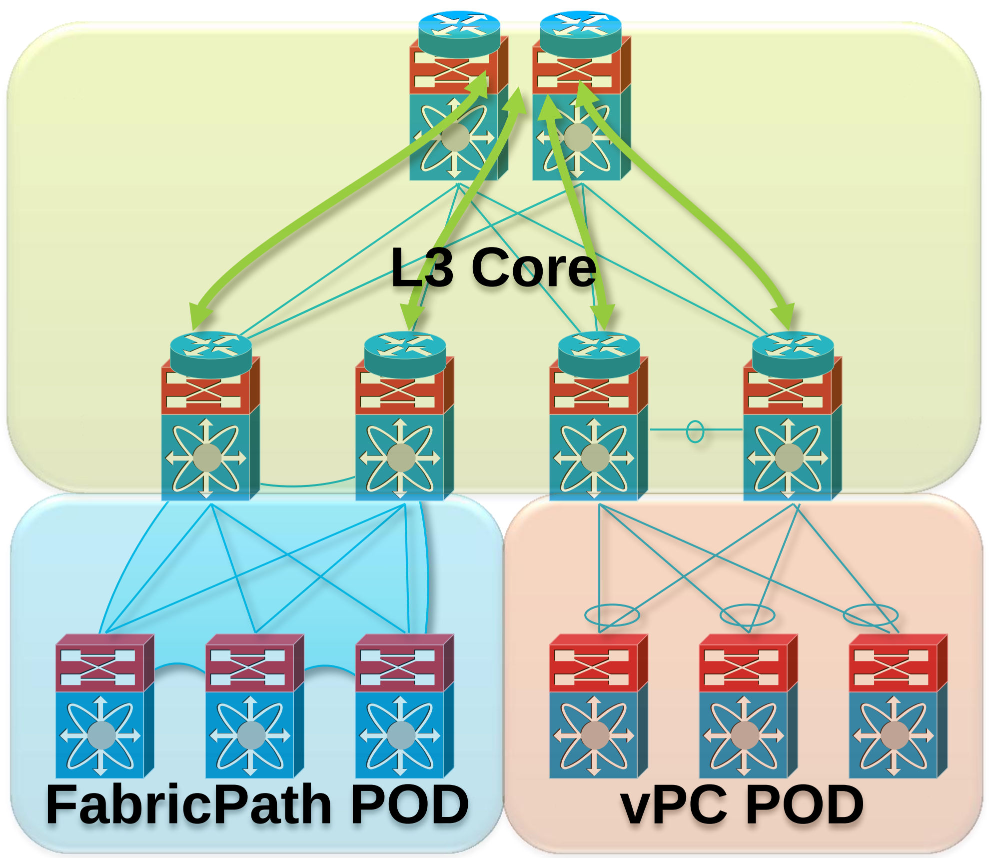
```
N7K(config)# feature-set fabricpath
N7K(config)# fabricpath switch-id <#>
N7K(config)# interface ethernet 1/1
N7K(config-if)# switchport mode
fabricpath
```

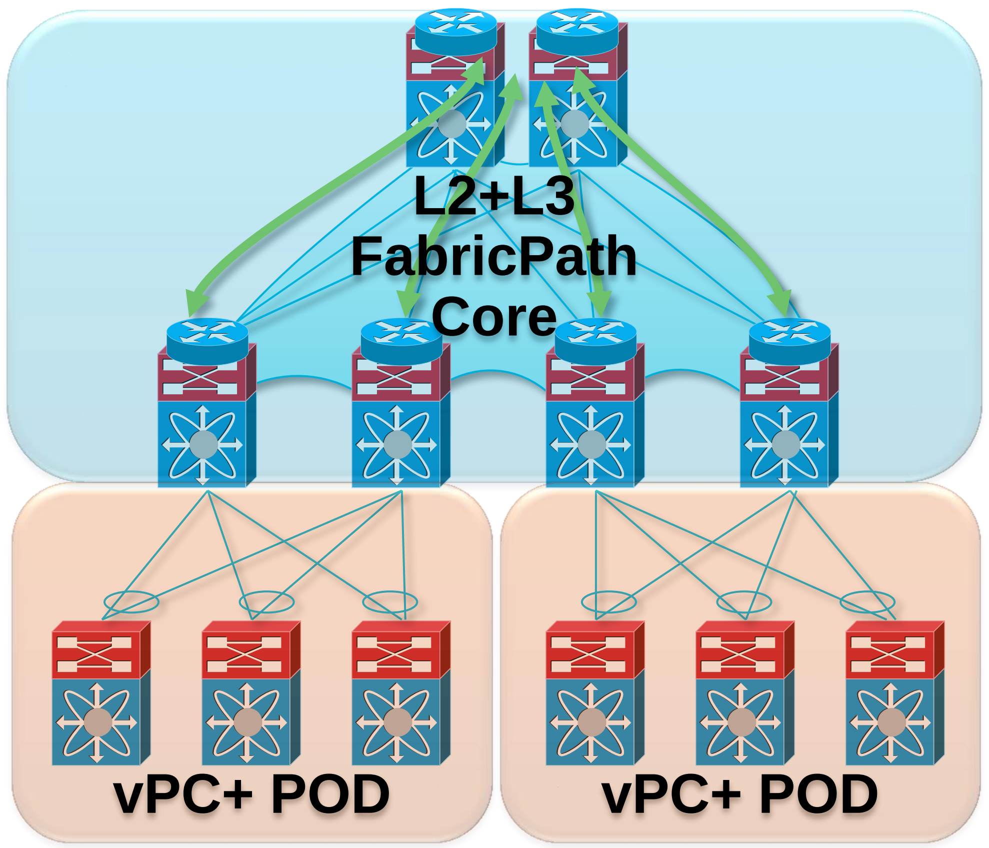


FabricPath Design

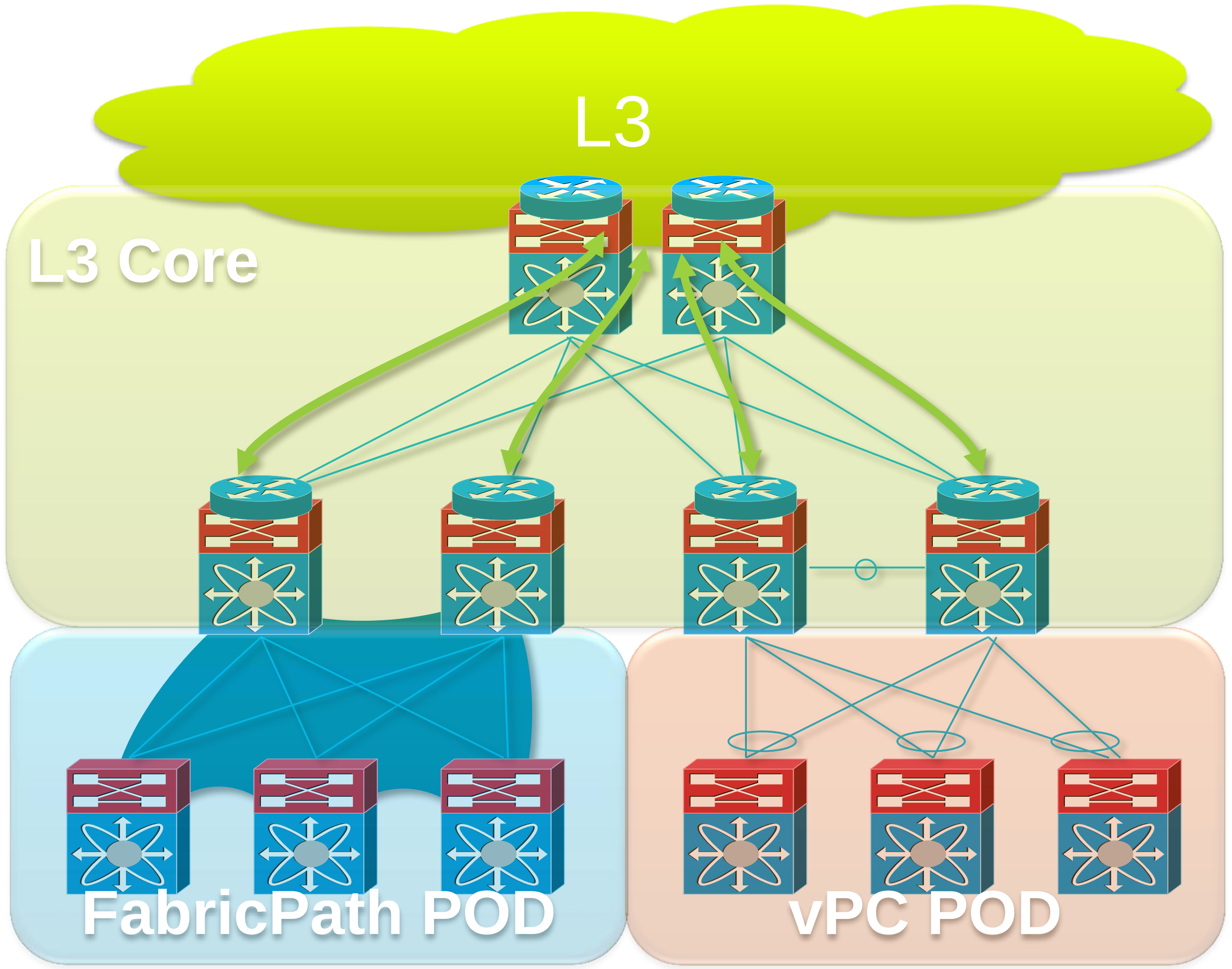
Layer 2 Routing

- FabricPath is not just intended for large scale topologies
- Useful for access to aggregation layer 2 configuration - ‘L2 Routed Access’
- Data Centre Interconnect
- Routed Topology allows variations on the design to meet
  the specific Data Centre topology requirement - CLOS, Ring, Tiers, …





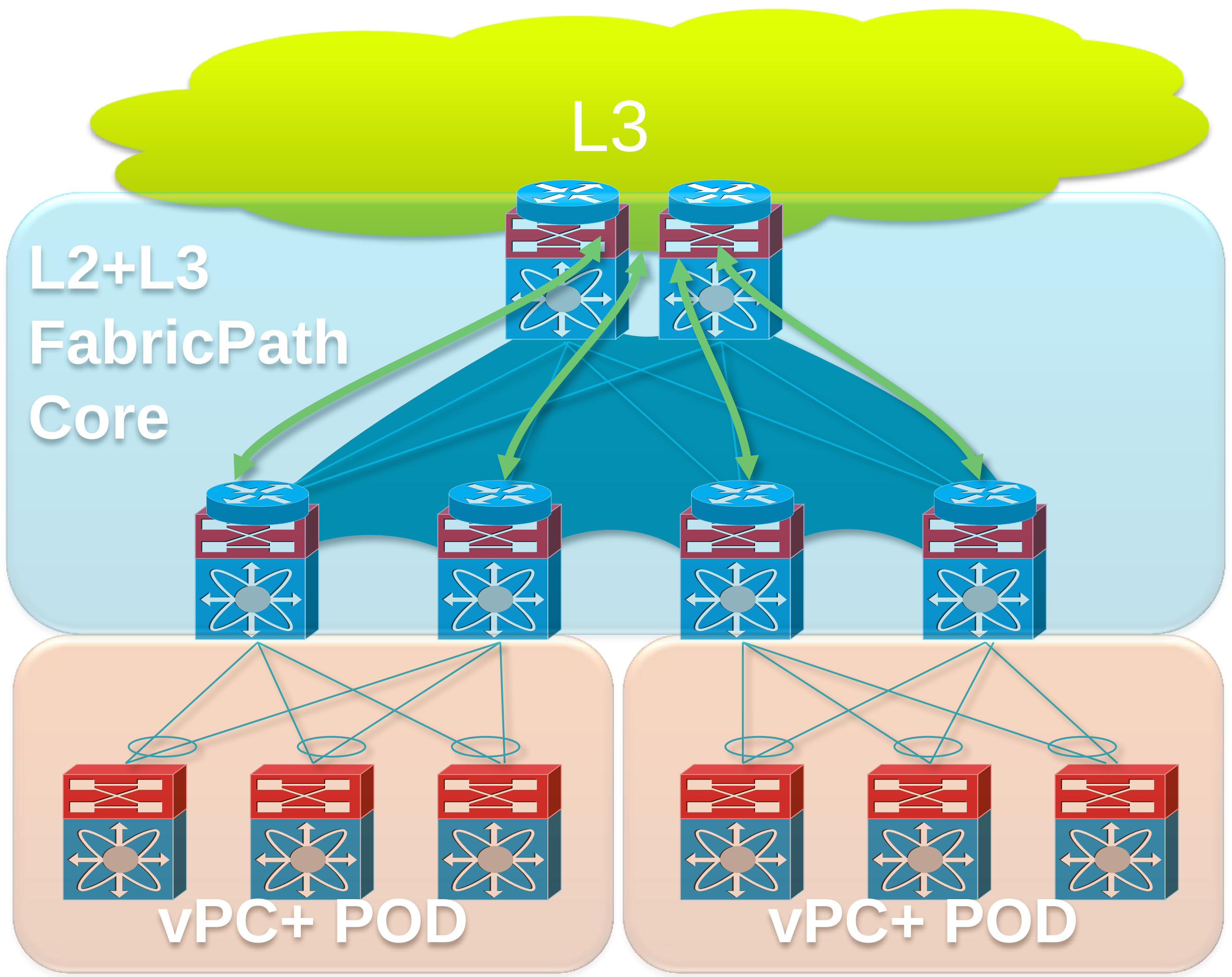
Fabric Path Design – Classical Fabric Path and vPC



- Simple configuration
- No constraint in the design
- Seamless L3 integration
- No STP, no traditional bridging
- Mac address table scaling
- Virtually unlimited bandwidth
- Can extend easily and without operational impact

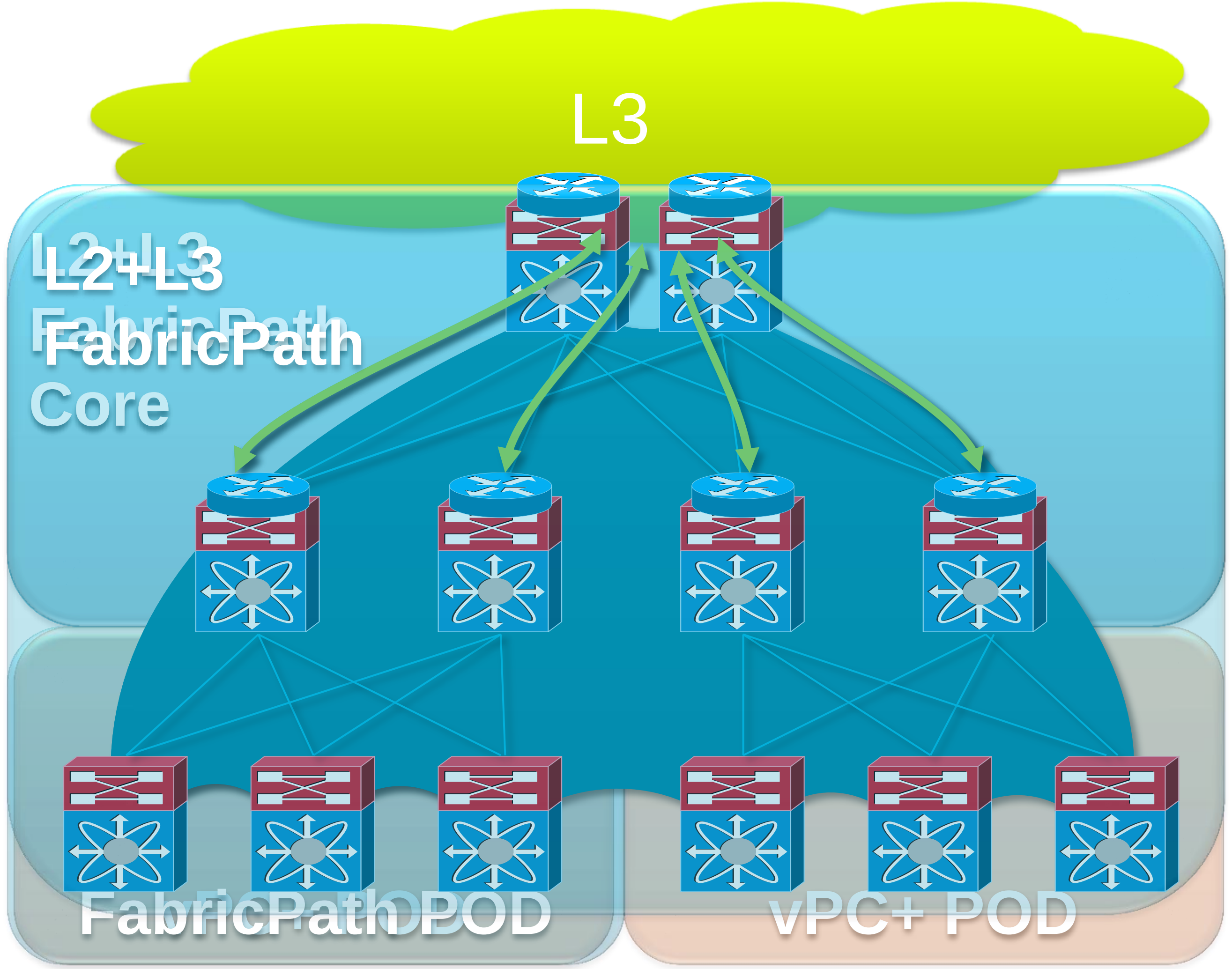
Fabric Path Design - Core

Efficient POD Interconnect



- FabricPath in the Core
- VLANs can terminate at the distribution **or** extend between PODs.
- STP is not extended between PODs, remote PODs or even remote data centres can be aggregated.
- Bandwidth or scale can be introduced in a non-disruptive way

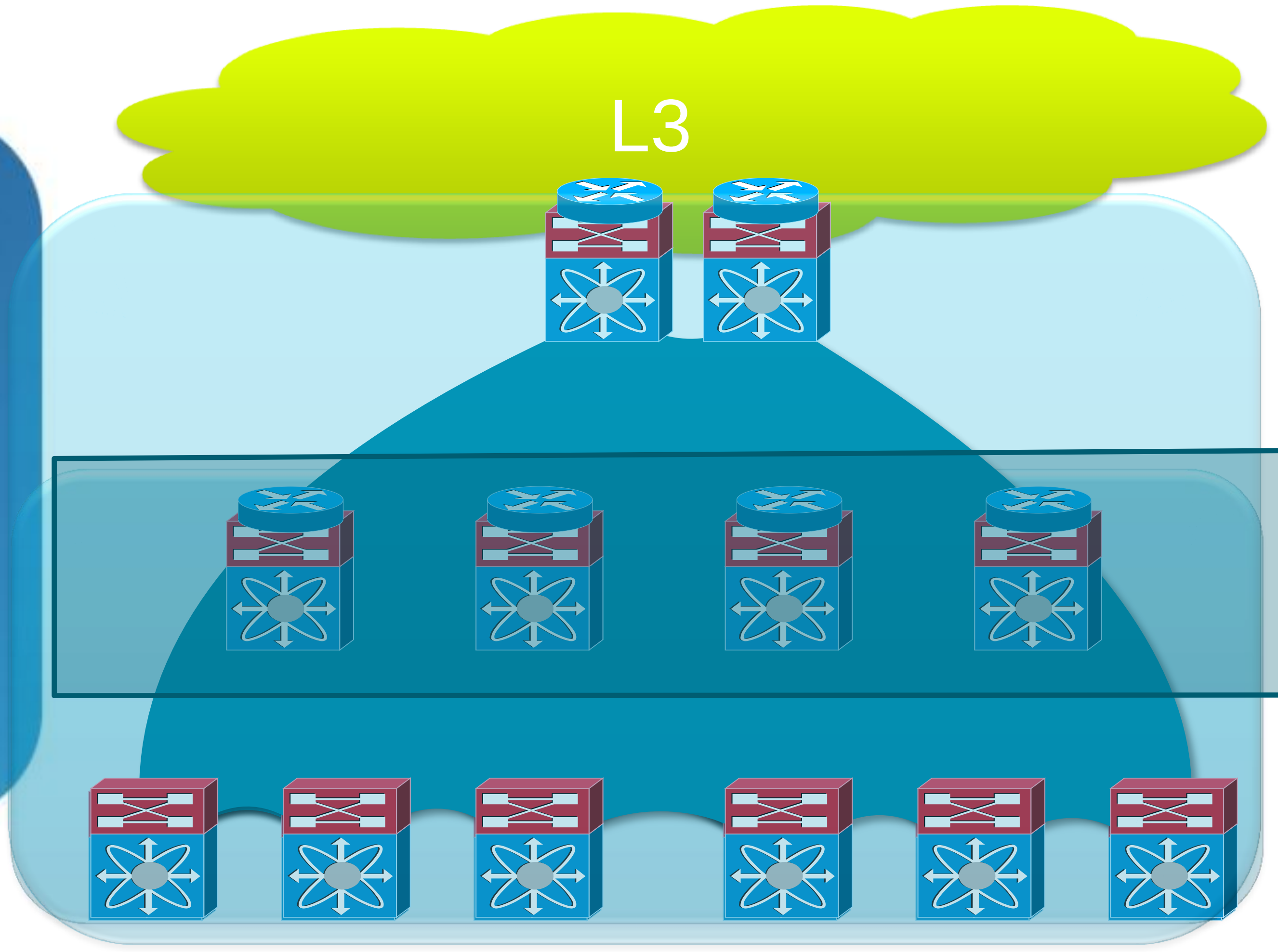
Fabric Path Design - Evolution



- FabricPath in the Core
- FabricPath extended down to the leaves

Fabric Path Design

Lets “Flat” the Network

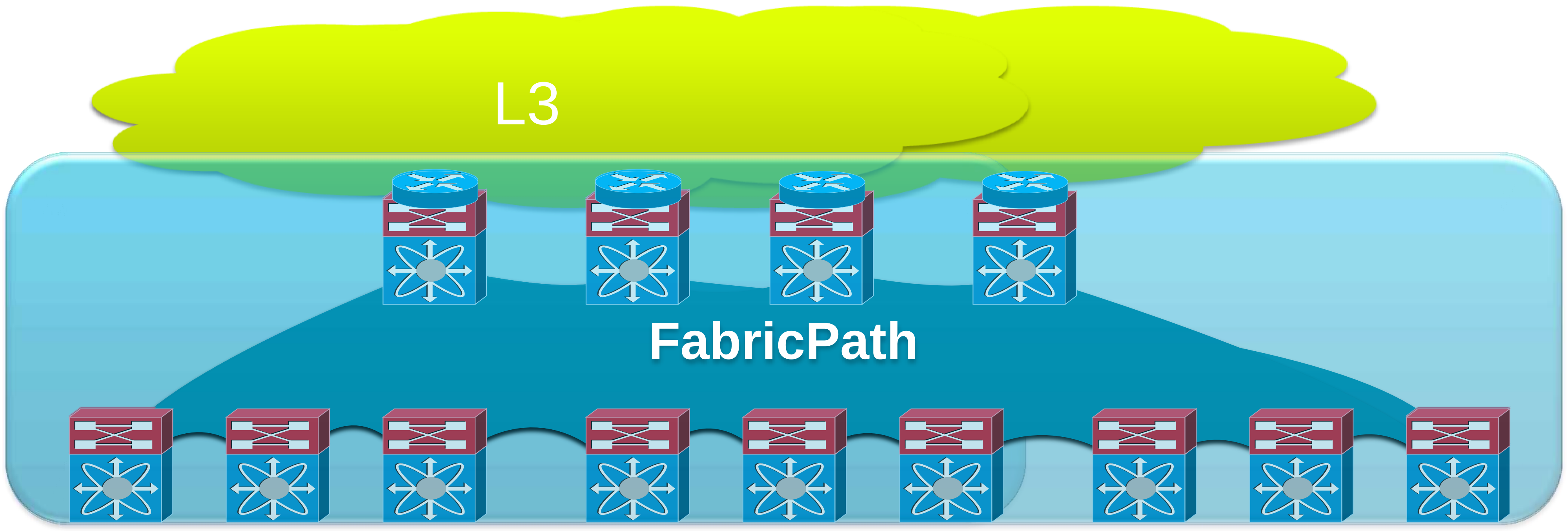


- FabricPath in the Core
- FabricPath extended down to the leaves
  - There is enough bandwidth and port density on the core Nexus 7000s or Nexus 6004s for aggregating the whole network.
    There is no need for a distribution layer for POD isolation

Fabric Path Design - Flexibility

The Network Can Evolve With No Disruption

- Need more edge ports? → Add more leaf switches
- Need more bandwidth? → Add more links and spines



Key Takeaways

- FabricPath is simple, keeps the attractive aspects of Layer 2
  - Transparent to L3 protocols
  - No addressing, simple configuration and deployment
- FabricPath is efficient
  - High bi-sectional bandwidth (ECMP)
  - Optimal path between any two nodes
- FabricPath is scalable
  - Can extend a bridged domain without extending the risks generally associated to Layer 2 (frame routing, TTL, RPFC)

Q & A

Complete Your Online Session Evaluation

Give us your feedback and receive a Cisco Live 2013 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 8 March 12:00pm-2:00pm

Don't forget to activate your Cisco Live 365 account for access to all session material, communities, and on-demand and live activities throughout the year. Log into your Cisco Live portal and click the "Enter Cisco Live 365" button. www.ciscoliveaustralia.com/portal/login.ww