Scaling the DC Architecture: Be Ready for the Cloud Evolution

Steve Day
Technical Marketing Engineer, Europe

Evolving Applications impact on The Cloud Fabric

Increased Dependence on

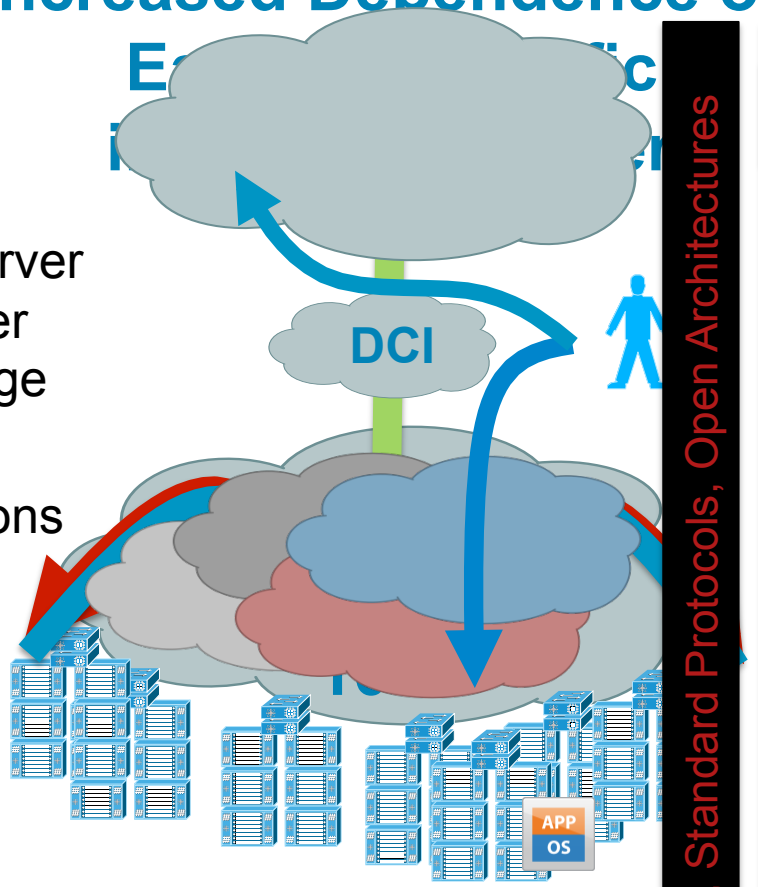
More Cores Per Server
More VMs on Server
Drive for 10G at edge

Clustered Applications

Multi-Tenancy

Disaster Recovery

Workload Mobility



Low Cost, Standard Protocols, Open Architectures

Automated Policy Driven Provisioning and Management

High Density 10G at Edge
40G &100G in Core/Agg
Control Plane Scalability

Non-Blocking Fabric
Predictable Lower Latency

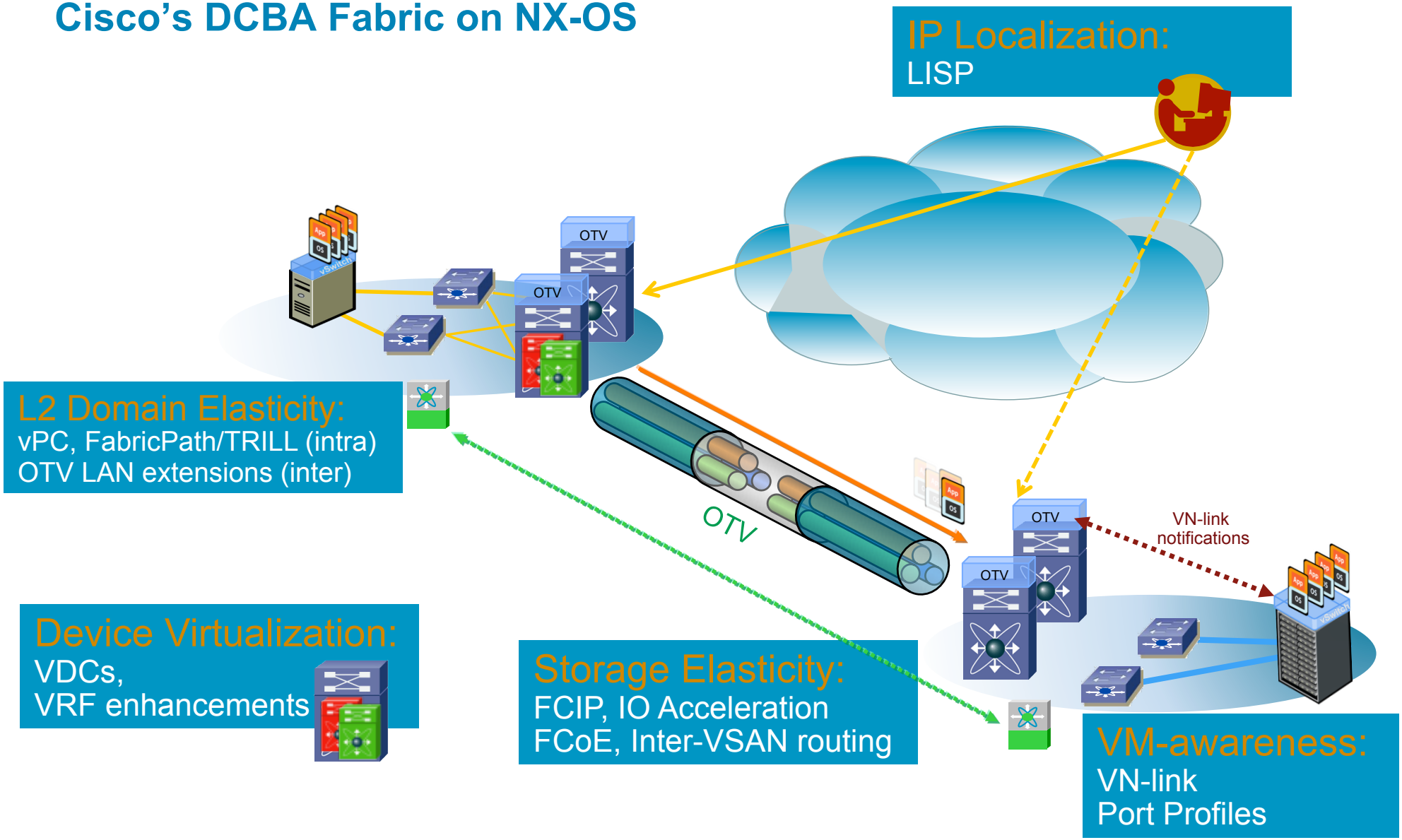
Secure Segmentation

L2 or L3 DCI Connectivity
Storage Extensions

Large L2 Domains
Non-Disruptive Migration

NX-OS Innovations for Cloud & Virtualization

Cisco's DCBA Fabric on NX-OS



Compute resources are part of the cloud, location is transparent to the user

Agenda

- FabricPath
- OTV – Overlay Transport Virtualization
- LISP – Locator ID Separation Protocol
- Summary

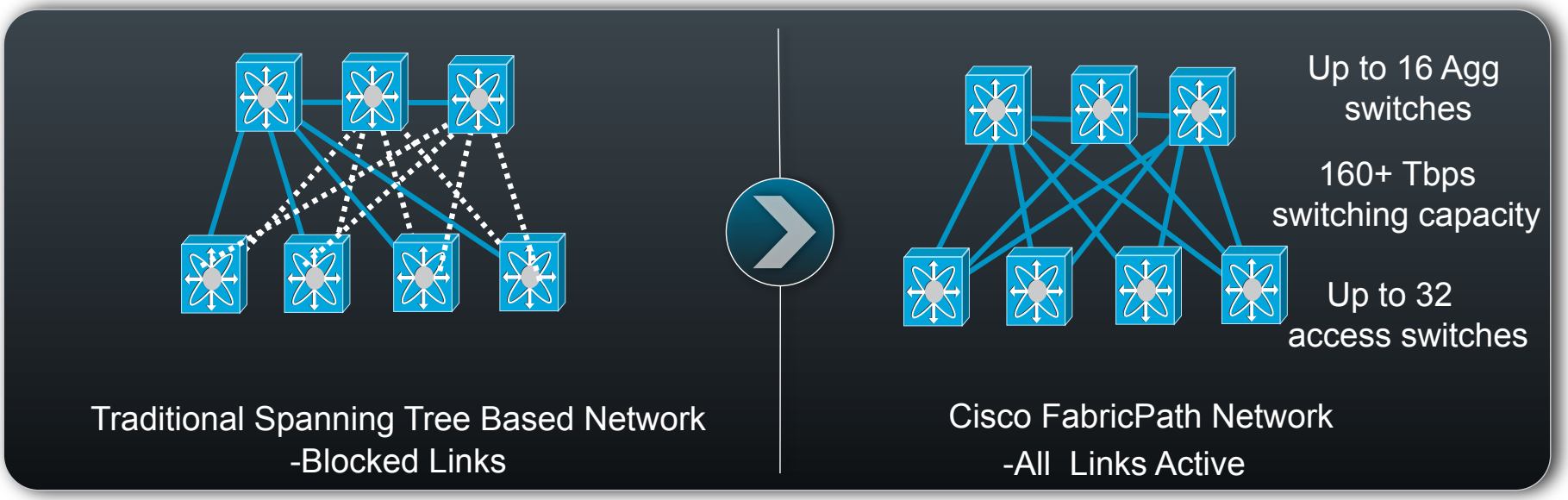
FabricPath

NETWORKWORLD

Oct, 25th 2010

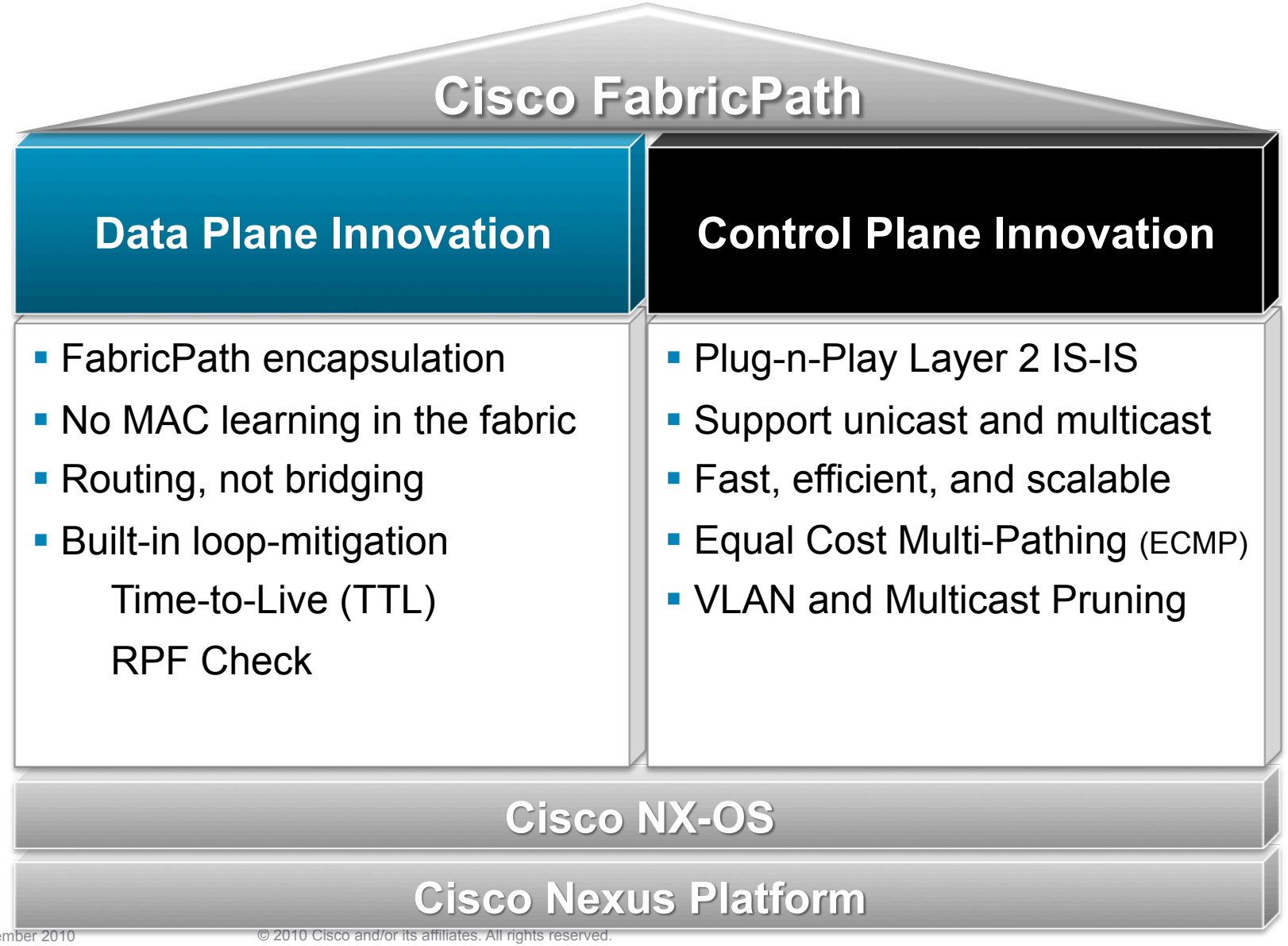
- Cisco FabricPath enables faster, simpler, flatter data center networks : http://www.networkworld.com/reviews/2010/102510-cisco-fabricpath-test.html
- How we tested Cisco FabricPath : http://www.networkworld.com/reviews/2010/102510-cisco-fabricpath-test-how.html?source=NWWNLE_nlt_cisco_2010-10-25

Cisco FabricPath



- Eliminate Spanning tree limitations
- Multi-pathing across all links, high cross-sectional bandwidth
- High resiliency, faster network re-convergence
- Any VLAN, Anywhere in the Fabric

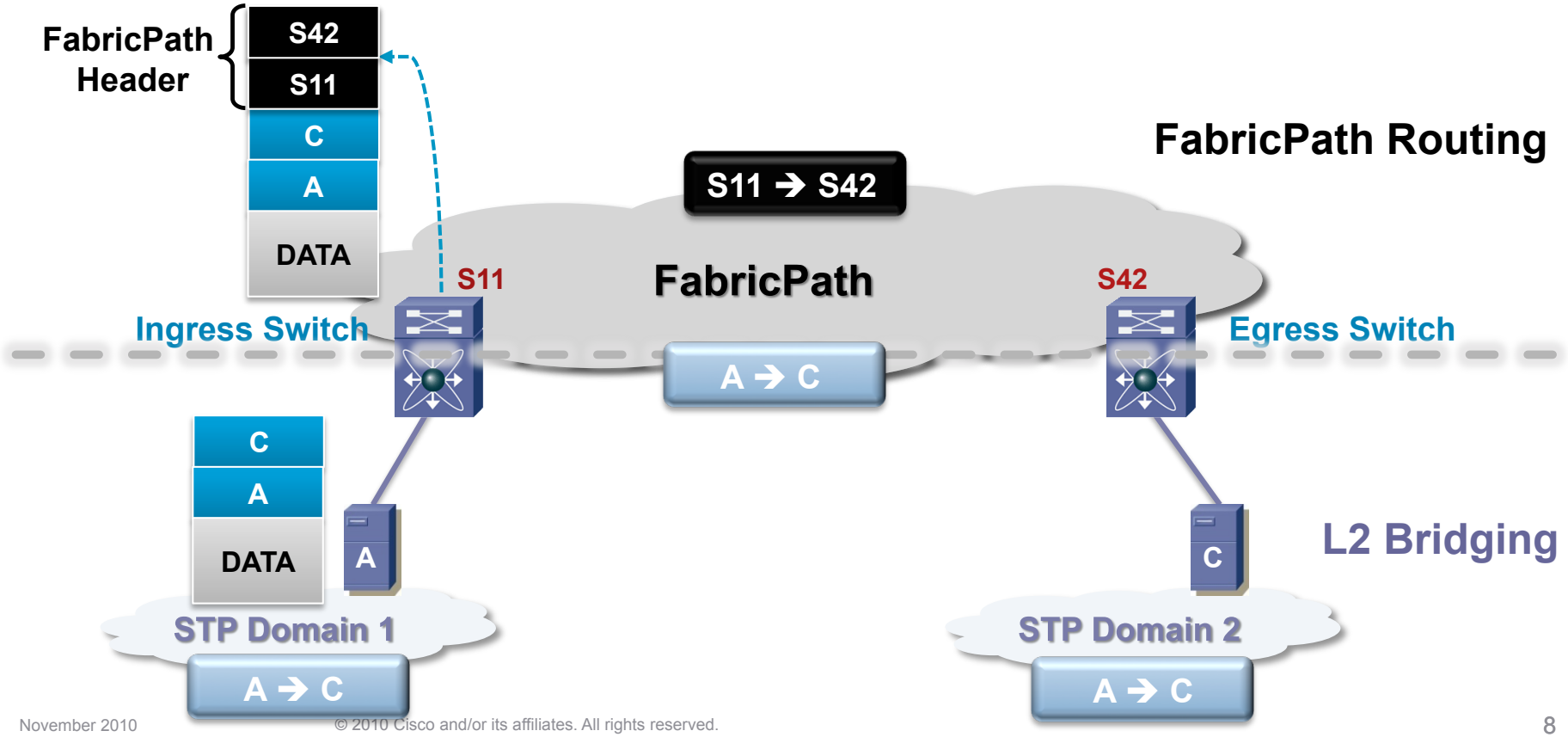
Cisco FabricPath Innovations



Data Plane Operation

Encapsulation to create hierarchical address scheme

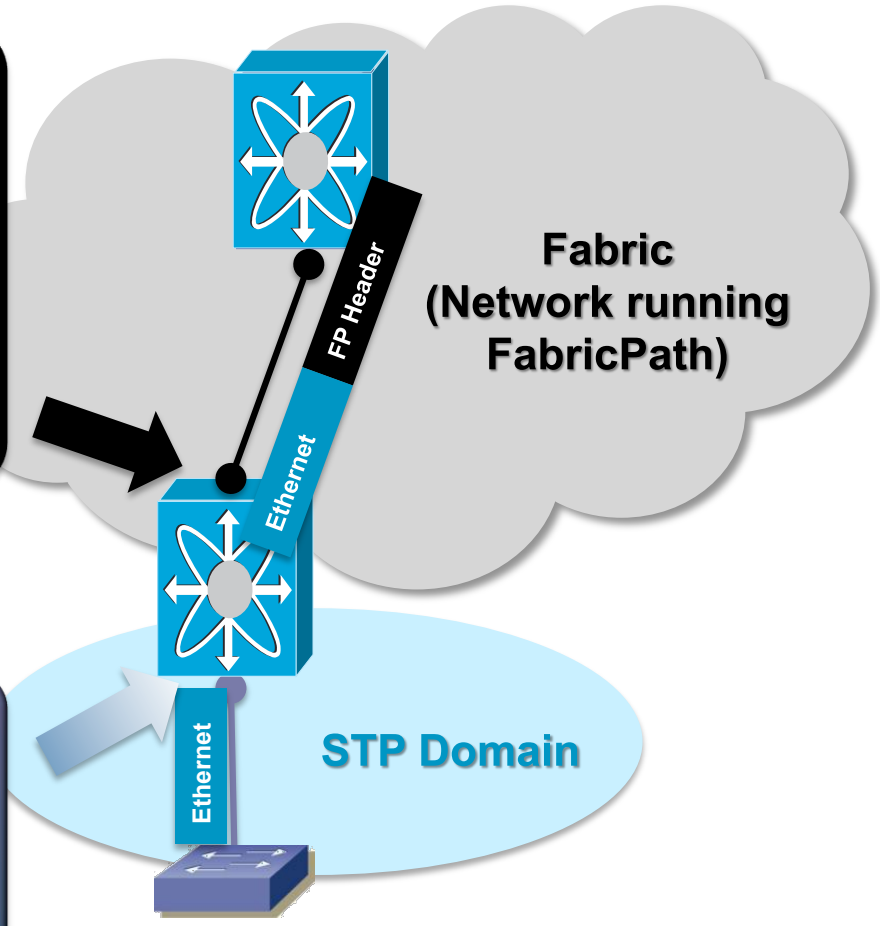
- FabricPath header is imposed by ingress switch
- Ingress and egress switch addresses are used to make the "Routing" decision
- No MAC learning required inside the L2 Fabric



FabricPath Terminology

FabricPath Port

- *Interfaces connected to another FabricPath Port*
- *Send/receive traffic with FabricPath header*
- *Does not run the spanning tree protocol*
- *No 'MAC Learning'*
- *Exchange topology info via L2 ISIS*
- *Forwarding based on 'Switch Routing Table'*

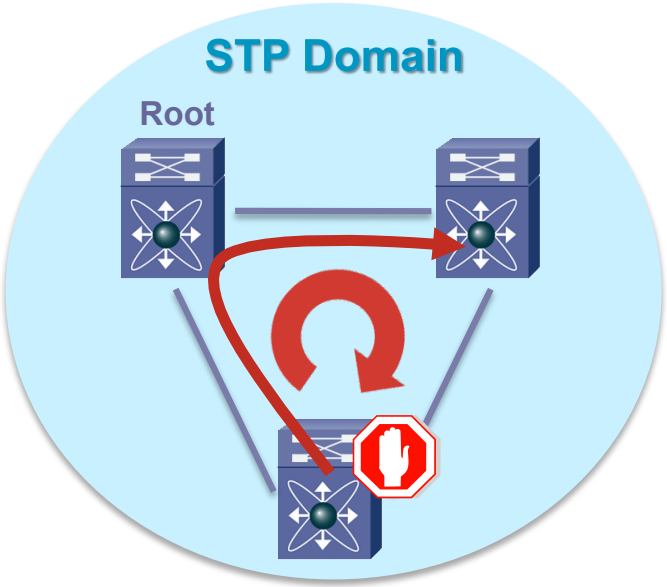


Classic Ethernet (CE) Port

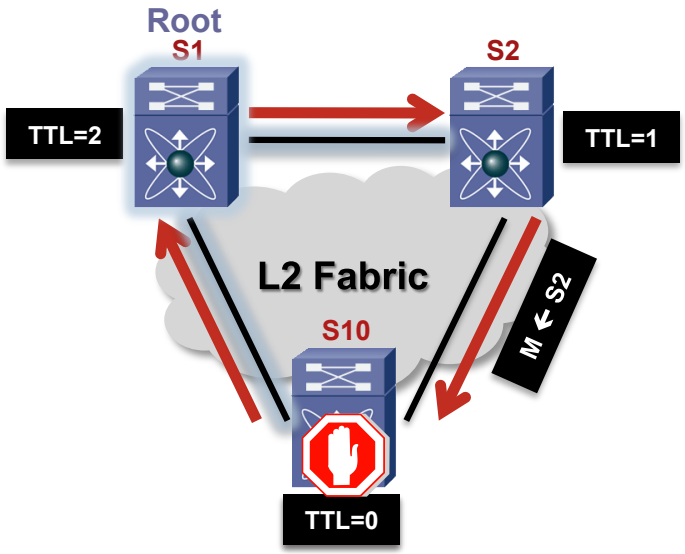
- Interfaces connected to non-FabricPath devices
- Send/receive Ethernet frames
- Participated in STP
- Uses a mac address table

Loop Mitigation with FabricPath

Time To Live (TTL) and Reverse Path Forwarding (RPF) Check

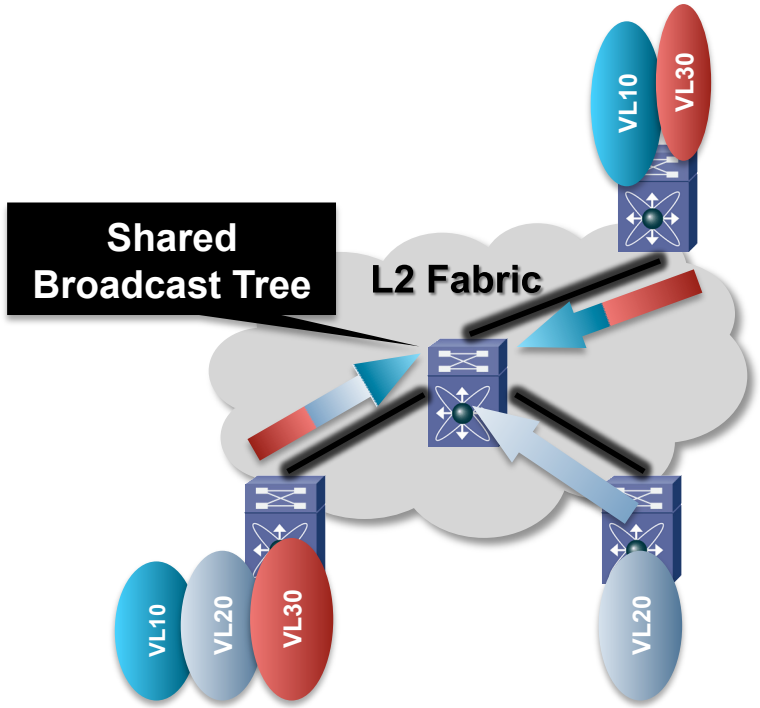


- Control protocol is the only mechanism preventing loops
- If STP fails -> infinite loop
  - No backup mechanism in the data plane
  - Complete network melt-down as the result of flooding

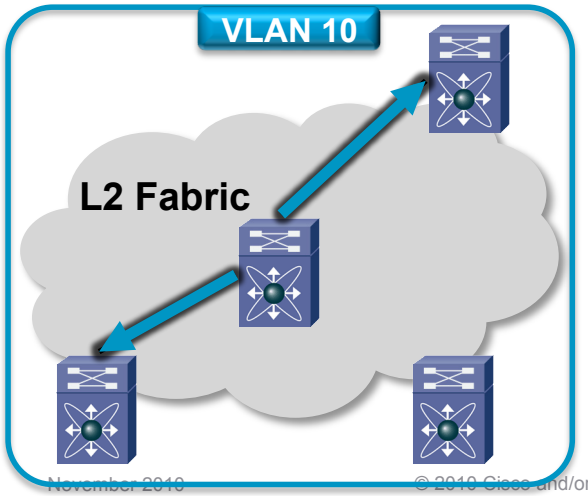


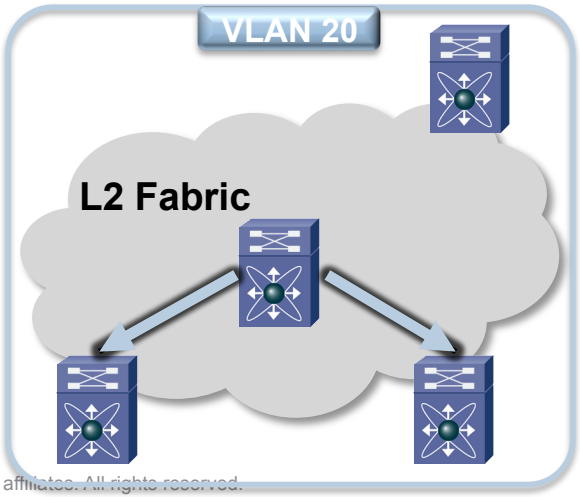
- TTL in FabricPath header
- Decrement by 1 at each hop
- Frames with TTL =0 are discarded
- RPF check for multicast based on “tree” info

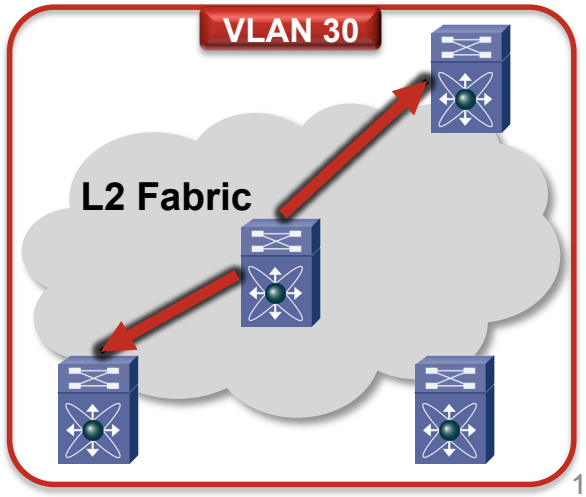
VLAN Pruning in L2 Fabric



- Switches indicate *'locally interested VLANs'* to the rest of the L2 Fabric
- Broadcast traffic for any VLAN only sent to switches that have requested for it

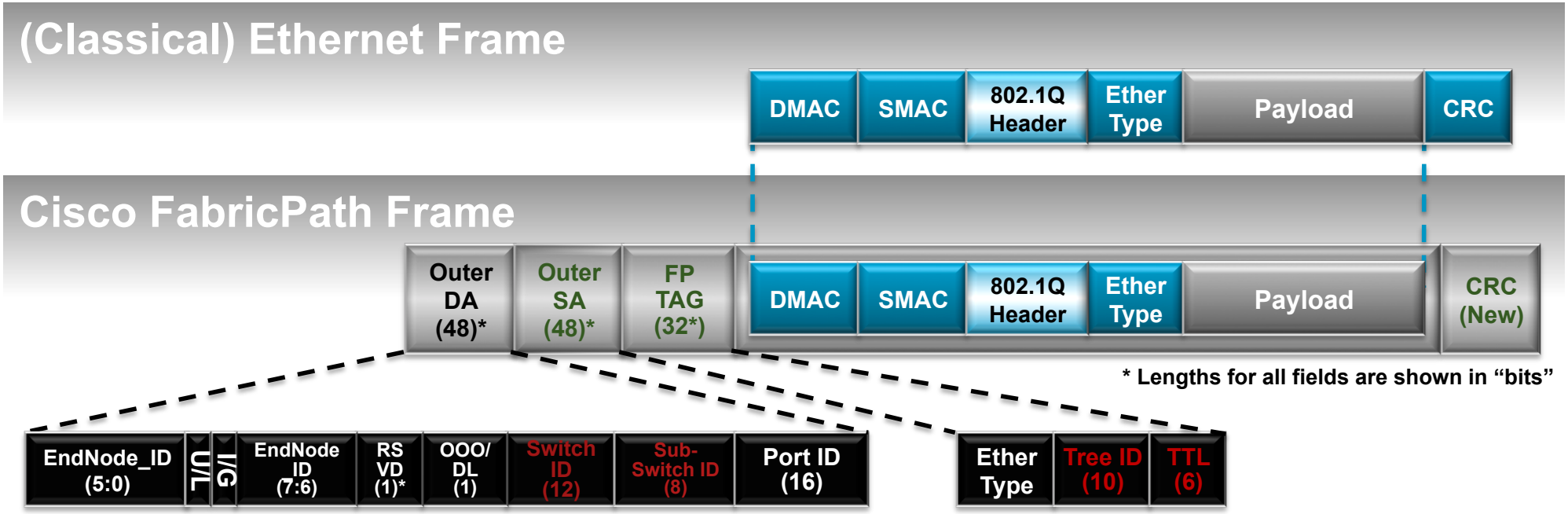






Native FabricPath Encapsulation

16-bytes header provide fields to help create hierarchical L2 address space and facilitate feature enhancements



- ***Switch ID:*** 12-bit number identifying a particular device in the L2 fabric.
- ***Sub-Switch ID:*** Combined with Switch ID to identify vPC+ behind a pair of peer-switches
- ***Tree ID:*** Unique number assigned to help identify each distribution "Tree"

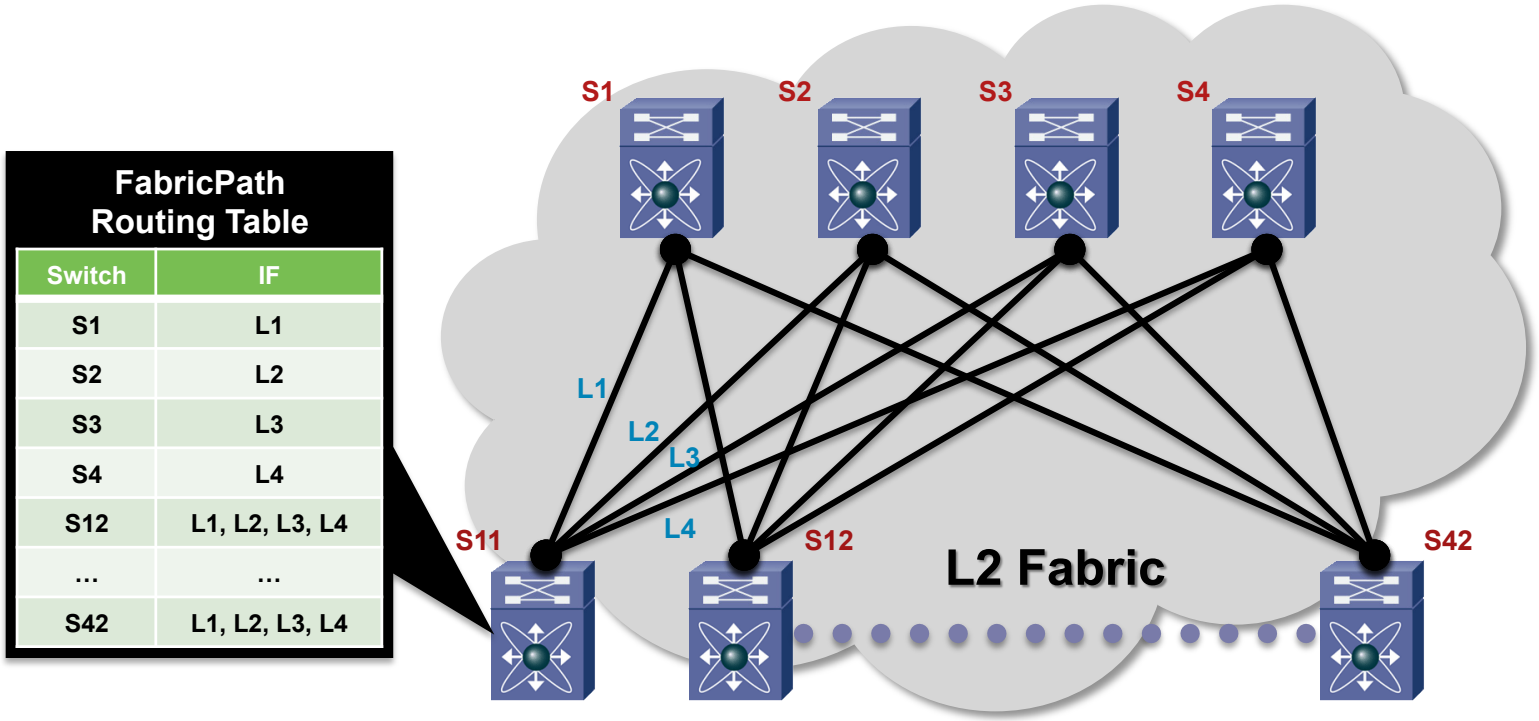
Forwarding Tag (Ftag): mainly used to identify multicast trees

- ***TTL:*** Decremented at each hop, protection against temporary loops in the data plane

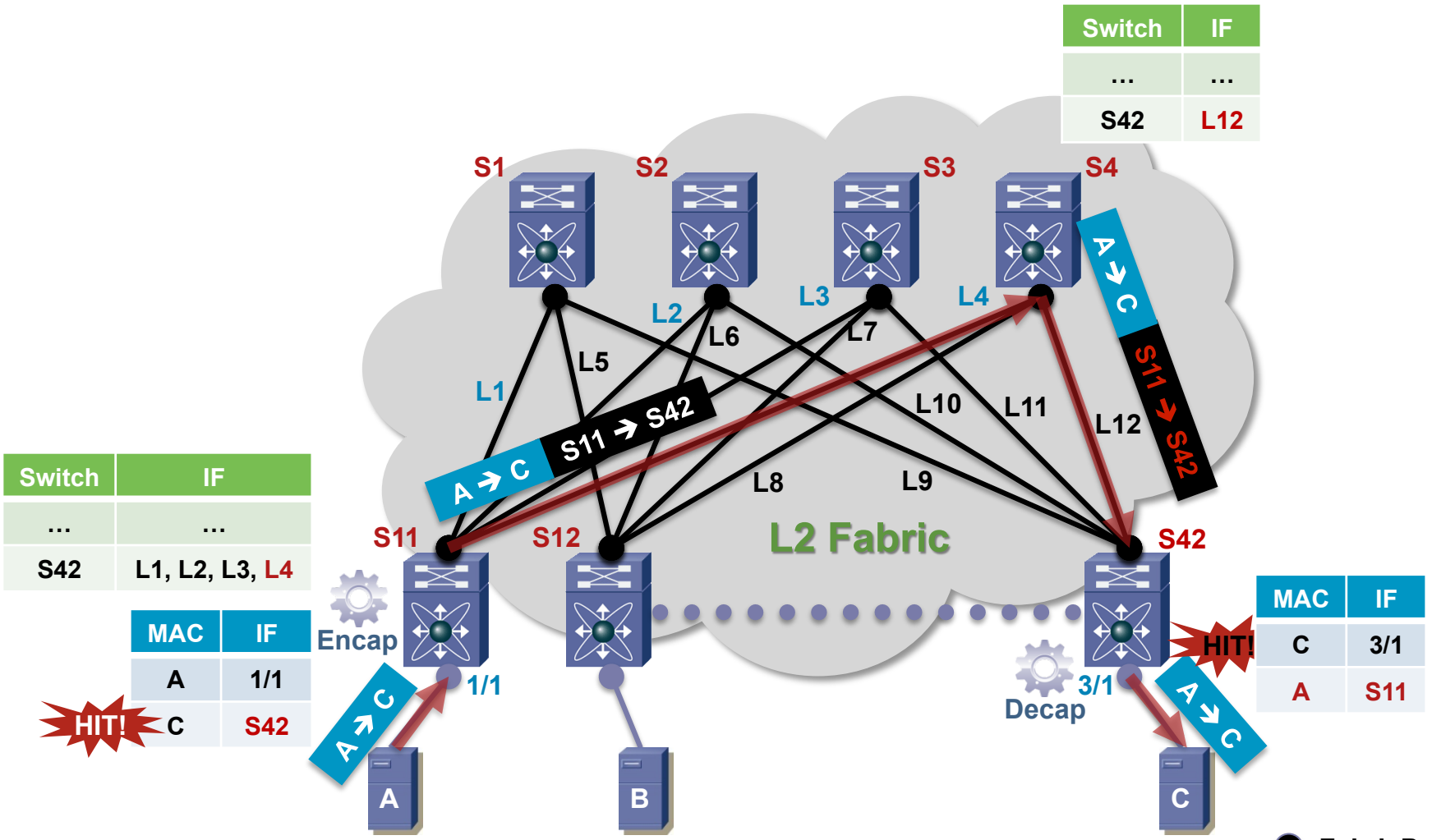
Control Plane Operation

Plug-N-Play L2 IS-IS is used to manage forwarding topology

- Assigned switch addresses to all FabricPath enabled switches automatically (no user configuration required)
- Compute shortest, pair-wise paths
- Support equal-cost paths between any FabricPath switch pairs



FabricPath Forwarding: Unicast



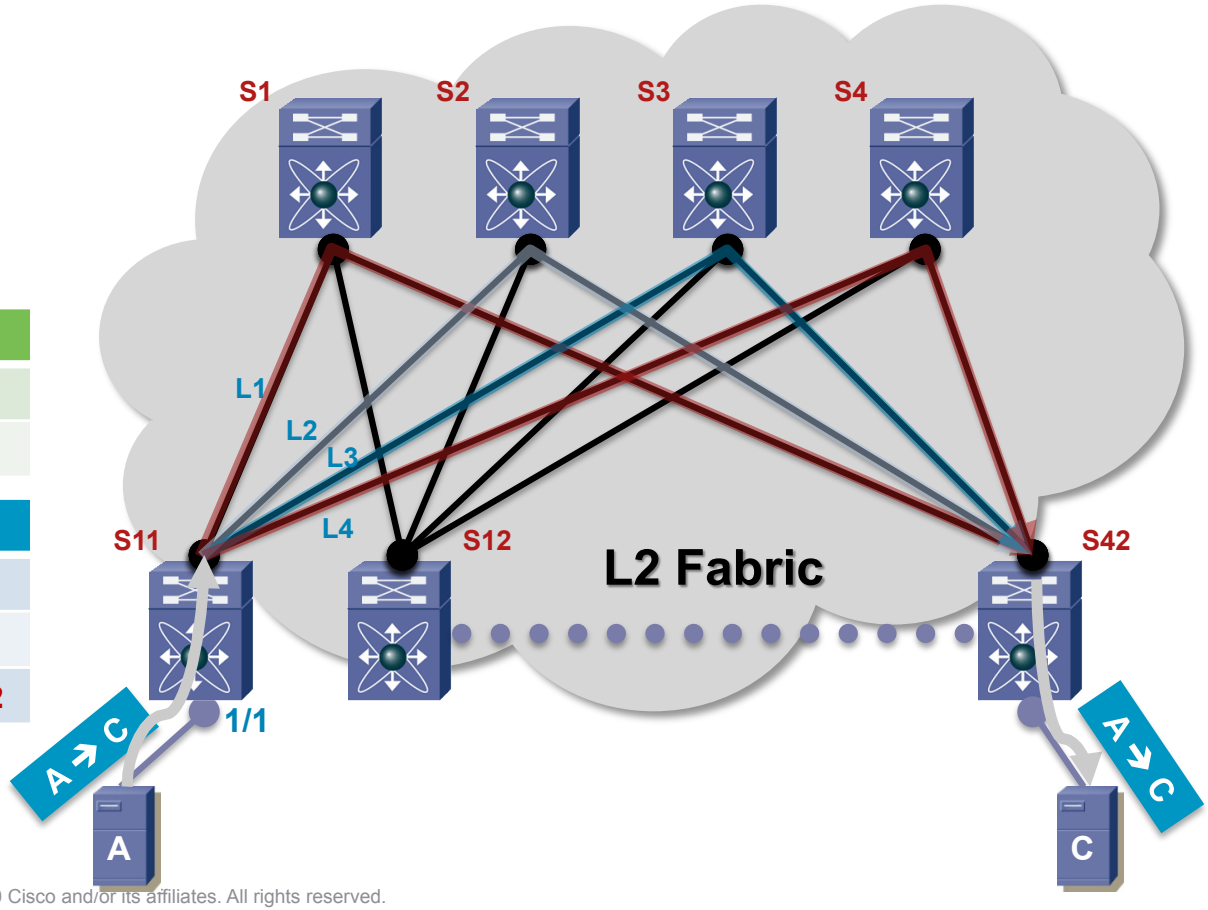
Unicast Equal Cost Multipathing

Forwarding decision based on 'FabricPath Routing Table'

- Support more than 2 paths (16 way ECMP) across the Fabric
- Increase bi-sectional bandwidth beyond port-channel
- High availability with N+1 path redundancy

Switch	IF
…	…
S42	L1, L2, L3, L4

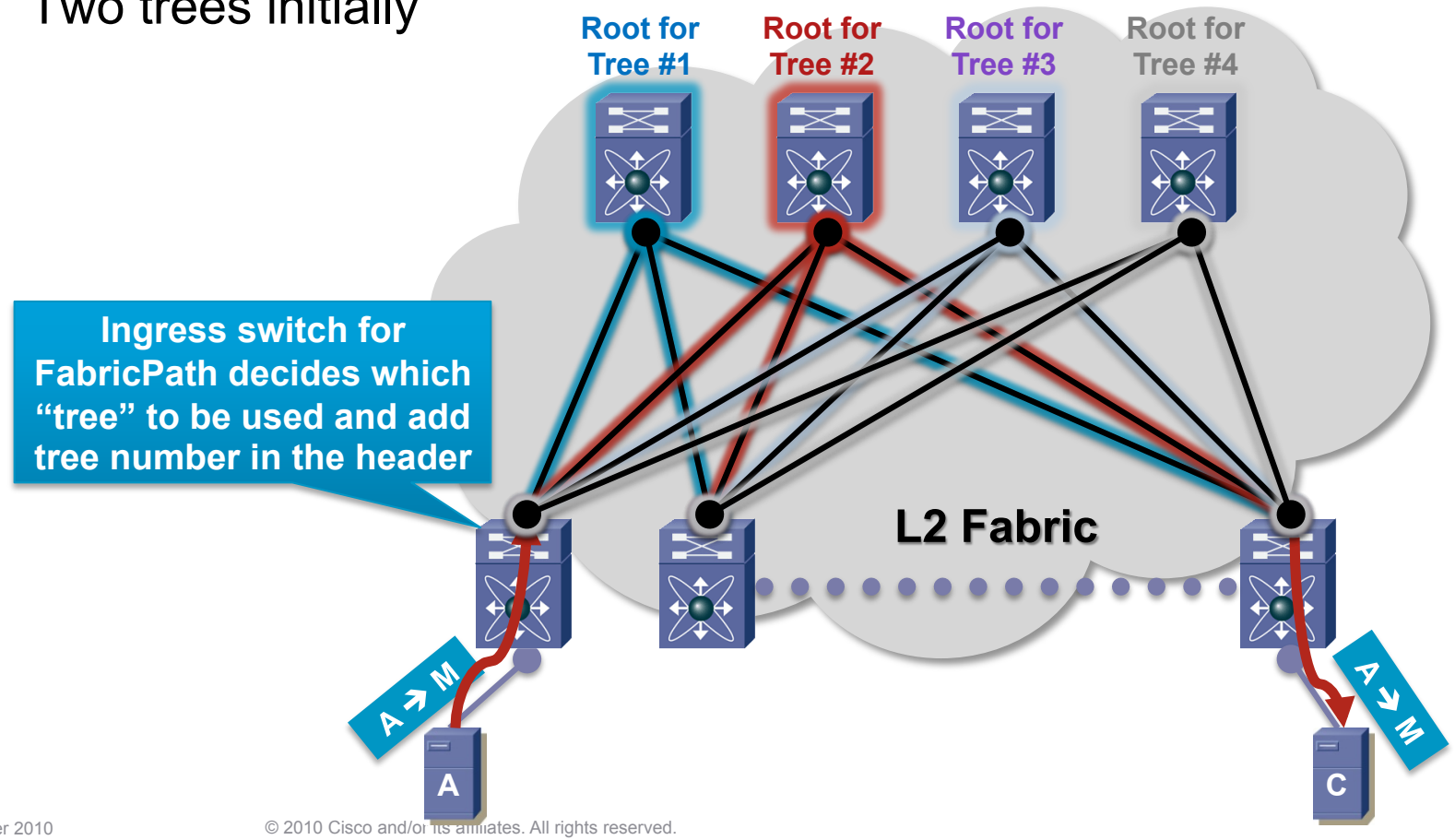
MAC	IF
A	1/1
…	…
C	S42



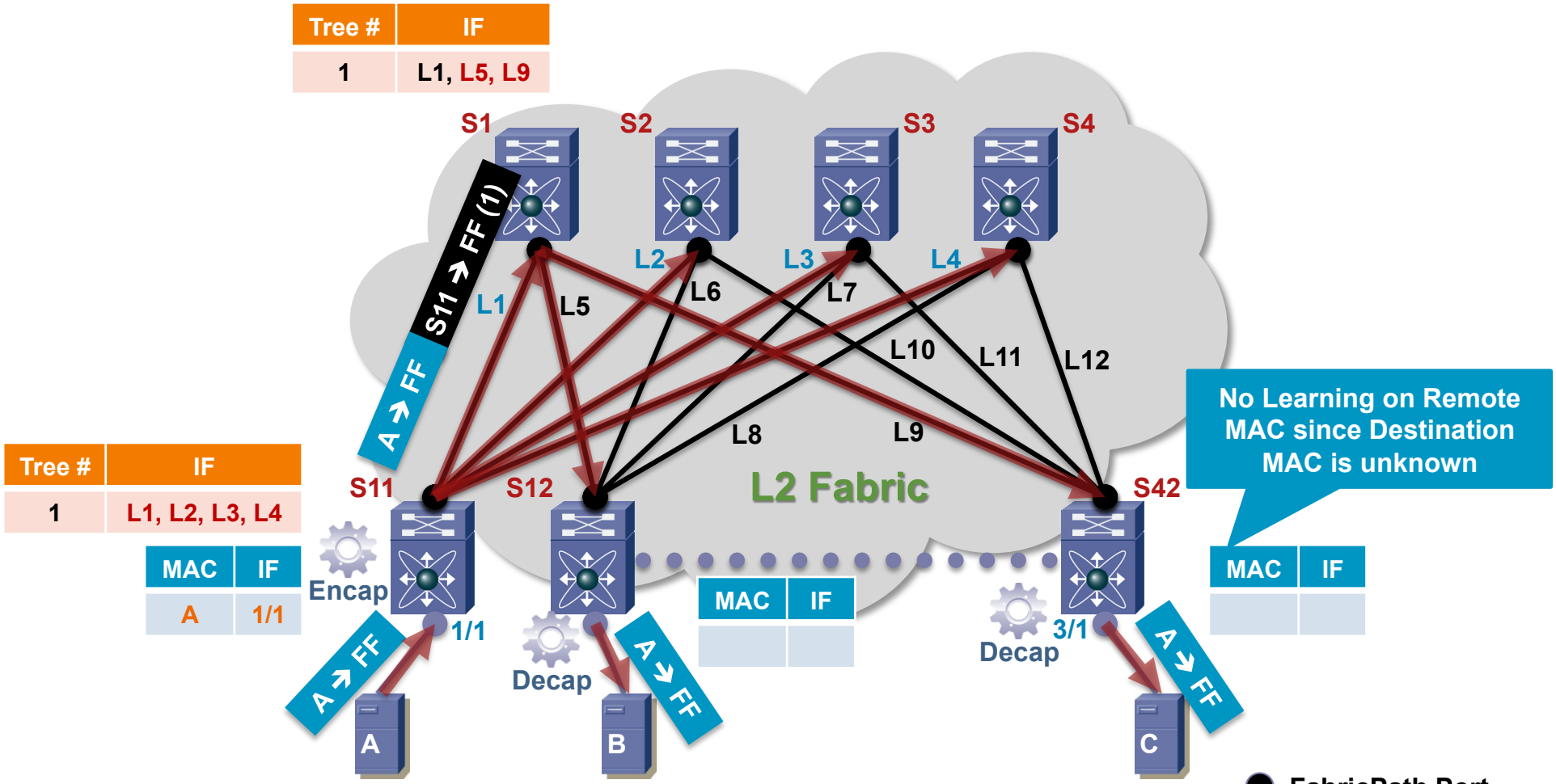
FabricPath Forwarding: Multicast

Forwarding through distinct ‘Trees’

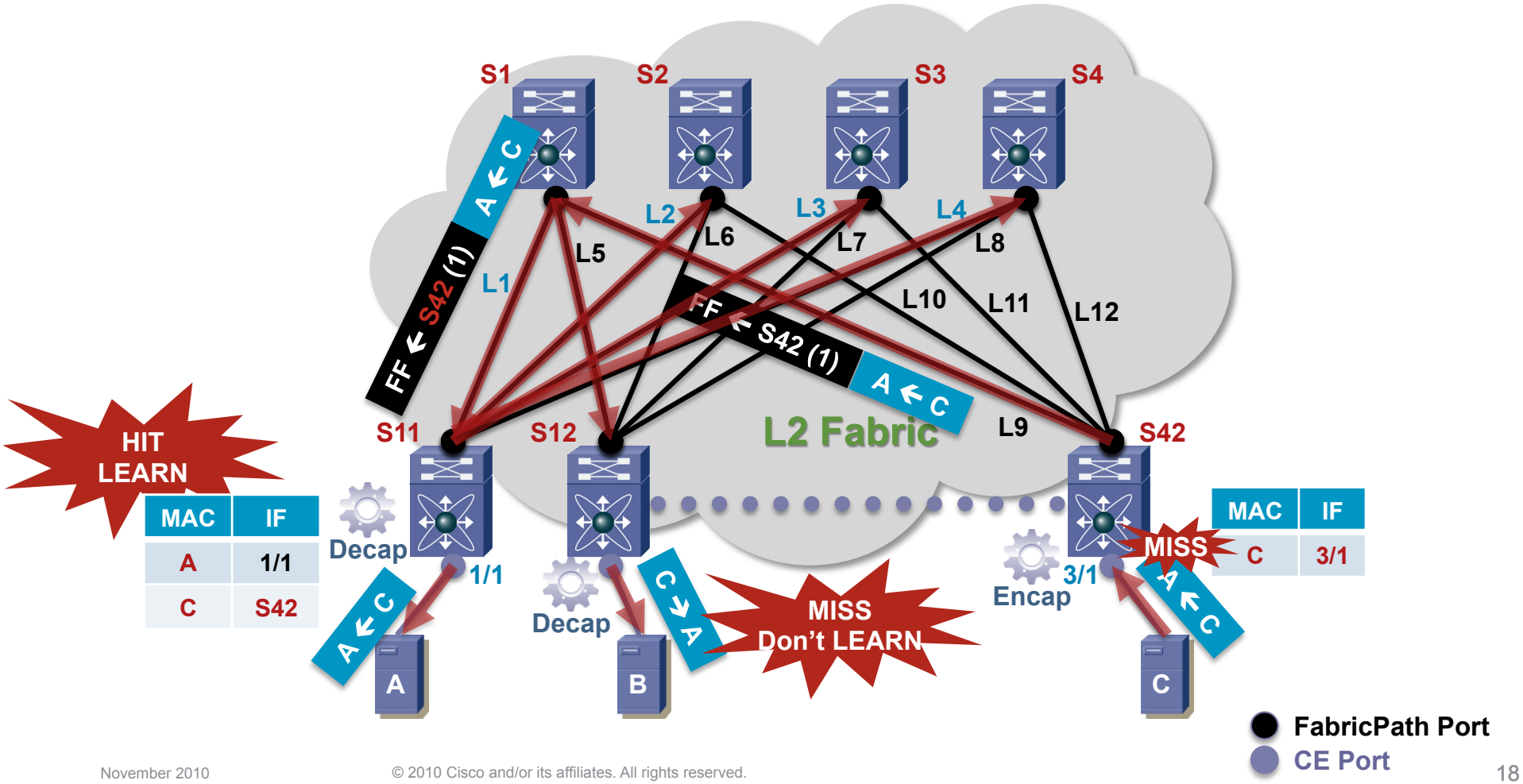
- Several ‘Trees’ are rooted in key location inside the fabric
- All Switches in L2 Fabric share the same view for each ‘Tree’
- Multicast traffic load-balanced across these ‘Trees’
- Two trees initially



FabricPath Forwarding: Broadcast

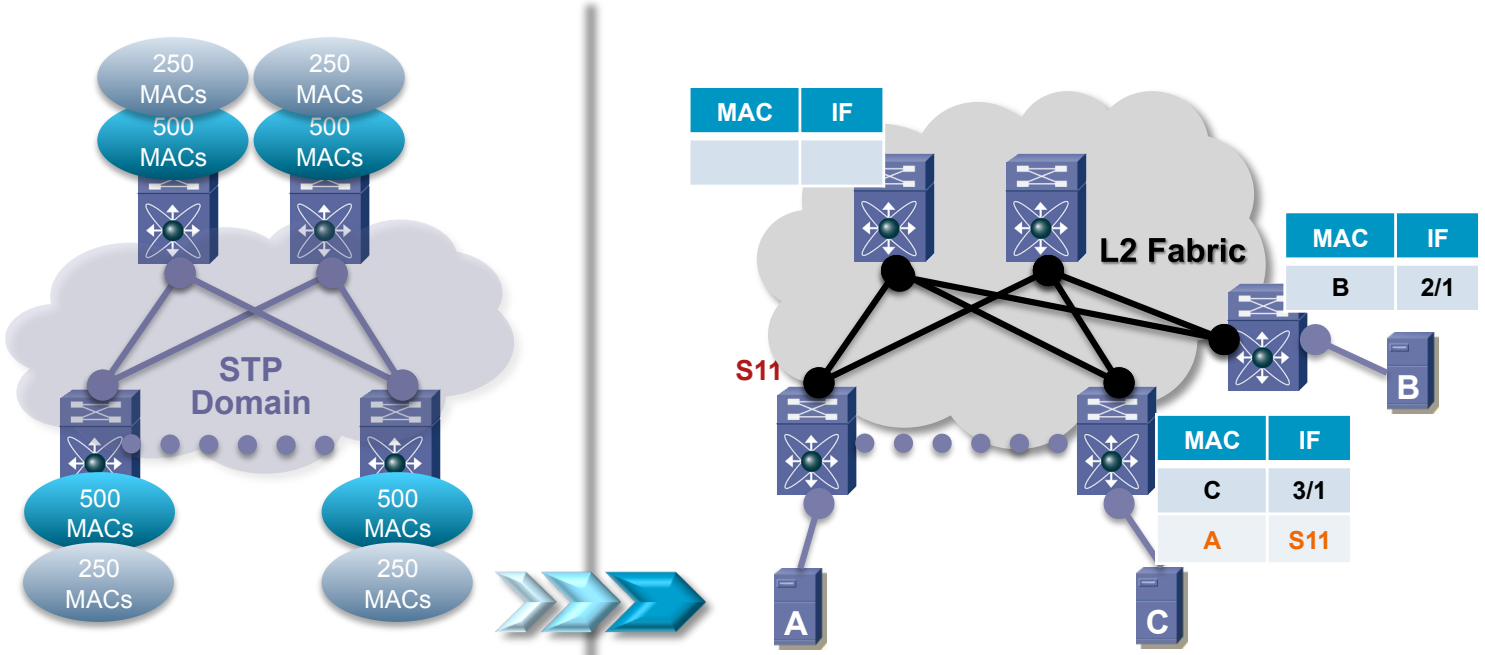


FabricPath Forwarding: Unknown Unicast



Conversational MAC Learning

Optimized Resource Utilization

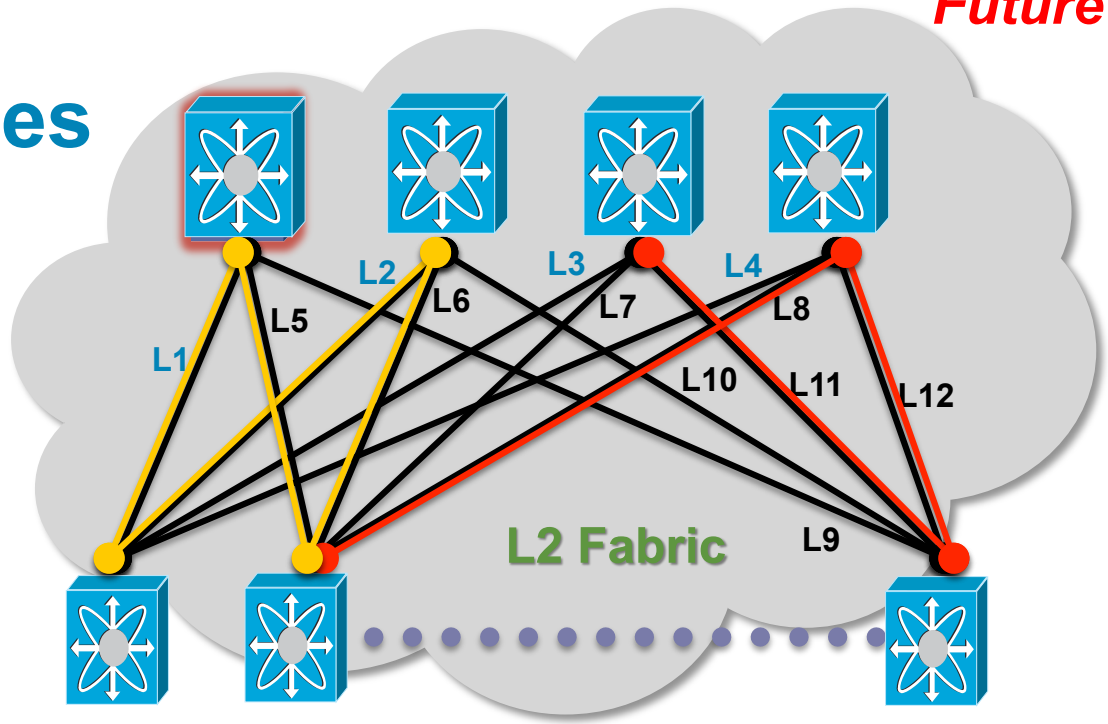


- *All* MACs need to be learned on *every* Switch
- Large L2 domains and virtualization present challenges to MAC Table scalability

- **Local MAC:** Source-MAC Learning only happens to traffic received on *CE Ports*
- **Remote MAC:** Source-MAC for traffic received on *FabricPath Ports* are only learned if Destination-MAC is already known as Local

Future

Multiple Topologies

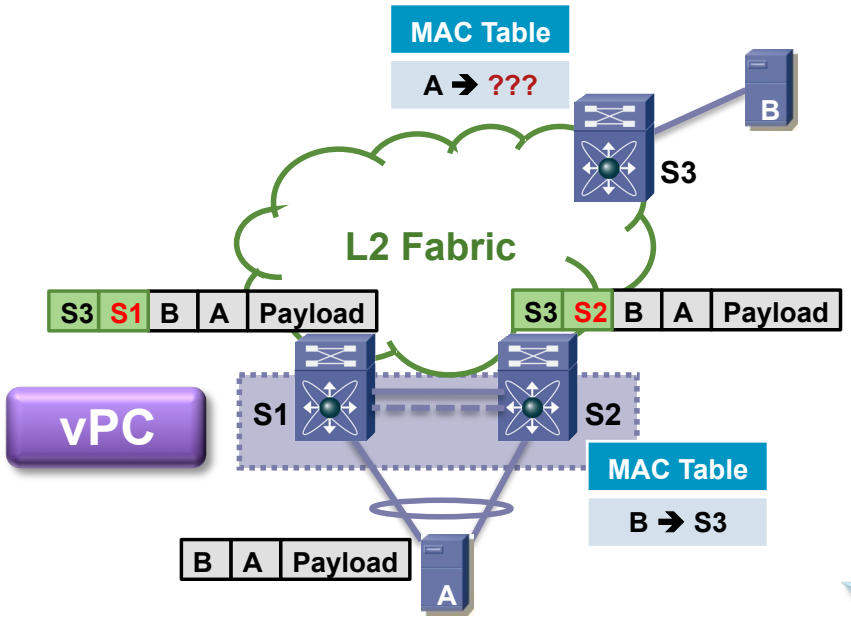


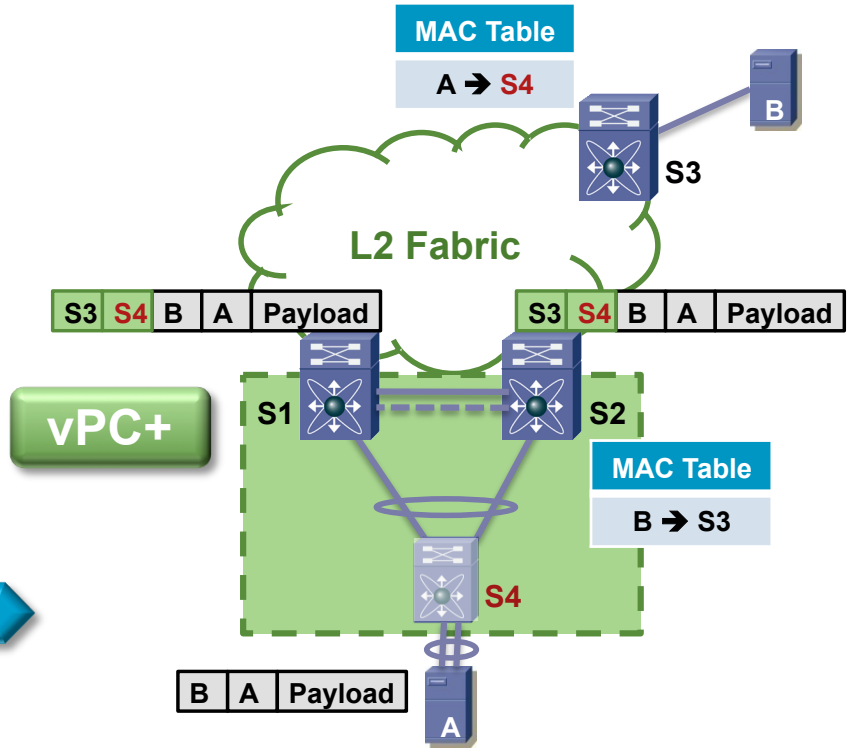
Topology: A group of links in the Fabric.
By default, all the links are part of topology 0.

- Other topologies can be created by assigning a subset of the links to them.
- A link can belong to several topologies
- A VLAN is mapped to a unique topology

Topologies can be used for traffic engineering, security etc…

vPC Enhancement for FabricPath



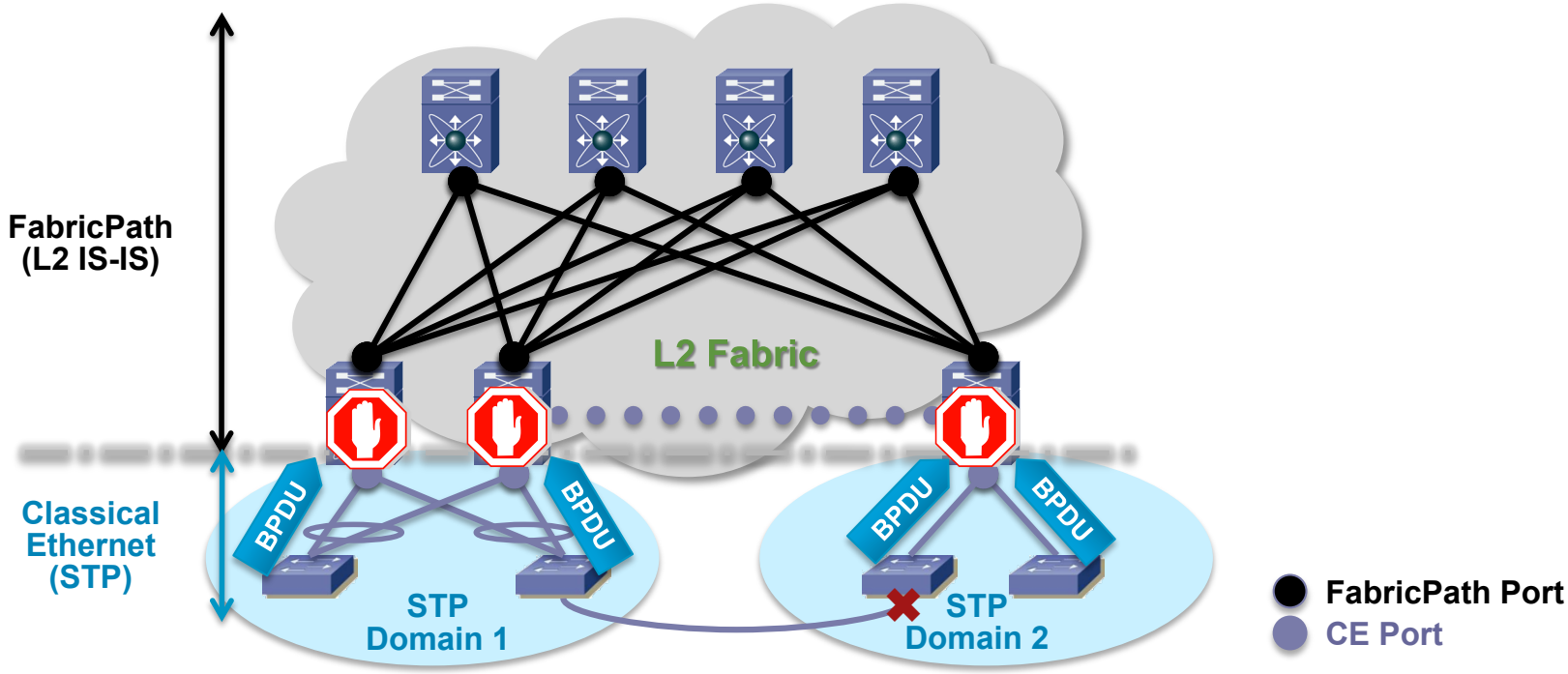


For Switches at L2 Fabric Edge

- vPC is still required to provide active/active L2 paths for dual-homed CE devices or clouds
- However, MAC Table only allows 1-to-1 mapping between MAC and Switch ID

- Each vPC domain is represented by an unique ‘Virtual Switch’ to the rest of L2 Fabric
- Switch ID for such ‘Virtual Switch’ is then used as Source in FabricPath encapsulation

STP Boundary Termination



- L2 Fabric is presented as a single bridge to all connected CE devices
- *L2 Fabric should be the root for all connected STP domains.* CE ports will be put into blocking state when 'better BPDU' is received ("rootguard")
- No BPDUs are forwarded across the fabric (terminated on CE ports)

Platform Availability

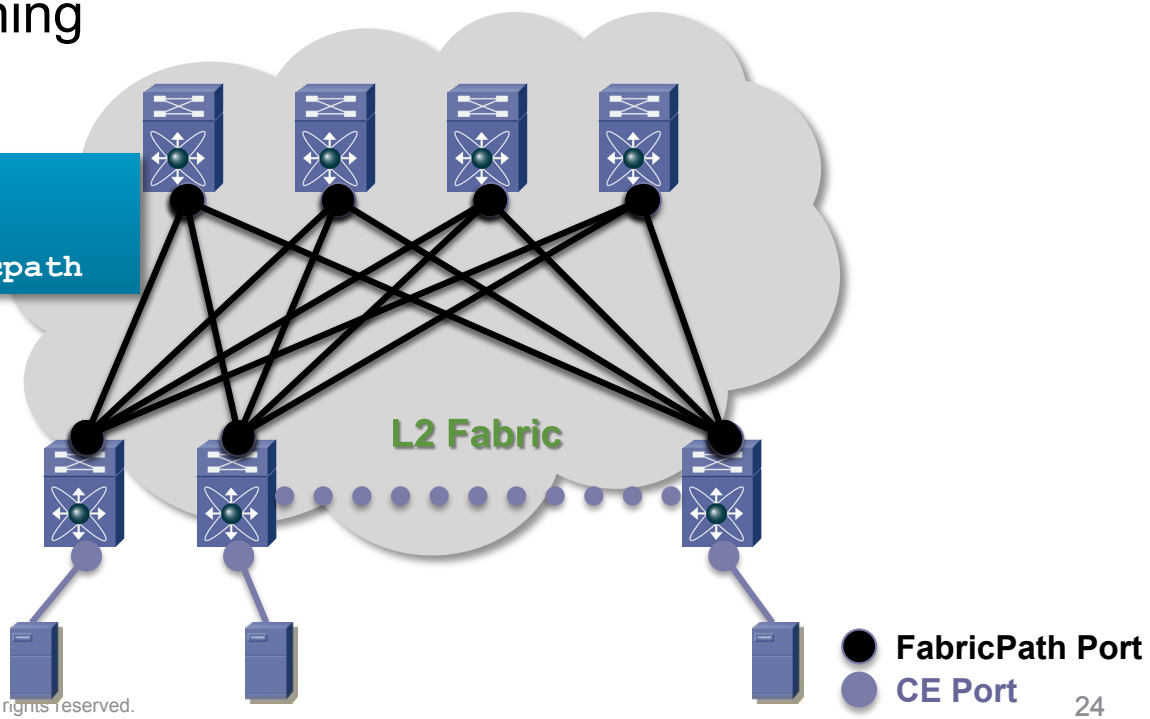
- Nexus 7000
  - Supported on F1-series module in all chassis
  - Supported from NX-OS 5.1(1), released 25th October 2010
  - Manuals and Release Notes on cisco.com
- Nexus 5000 Family
  - Nexus 55xx hardware is ready, Nexus 50xx is not
  - Software support planned* for Q3 CY11

* Please verify

FabricPath Configuration

- No L2 IS-IS configuration required
- New ‘feature-set’ keyword allows loading the code and starting the multiple services required by FabricPath (e.g. L2 IS-IS, LLDP, etc.)
- Simplified operational model – only 3 commands to get FabricPath up and running

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

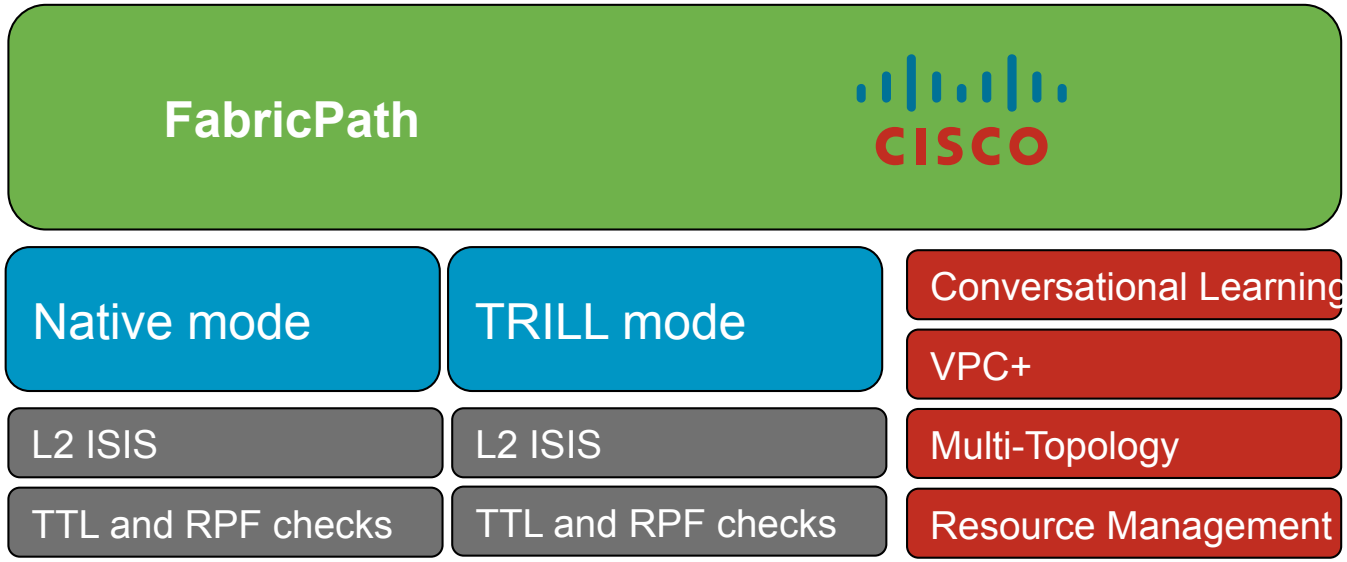


What Is the Relationship between FabricPath and TRILL?

- **FabricPath** is an **umbrella term** for a set of Layer 2 multipathing technologies
- FabricPath's initial release runs in a Native mode that is Cisco-specific, using proprietary encapsulation and control-plane elements
- Once the TRILL standard is complete, FabricPath will offer a TRILL-compliant mode for third-party interoperability. This will be achieved by a simple software upgrade.
- Nexus 7000 F1 I/O modules and Nexus 5500 HW are capable of running both FabricPath and TRILL modes

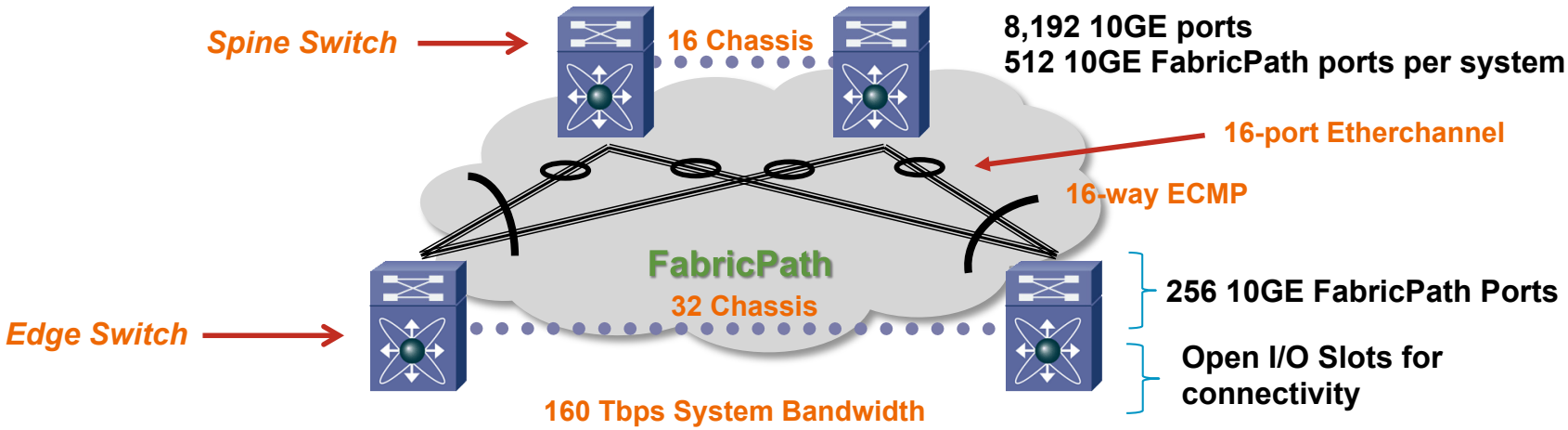
Any Benefit to Run Cisco FabricPath vs. TRILL?

- Yes!



Use Case: High Performance Compute

Building Large Scalable Compute Clusters



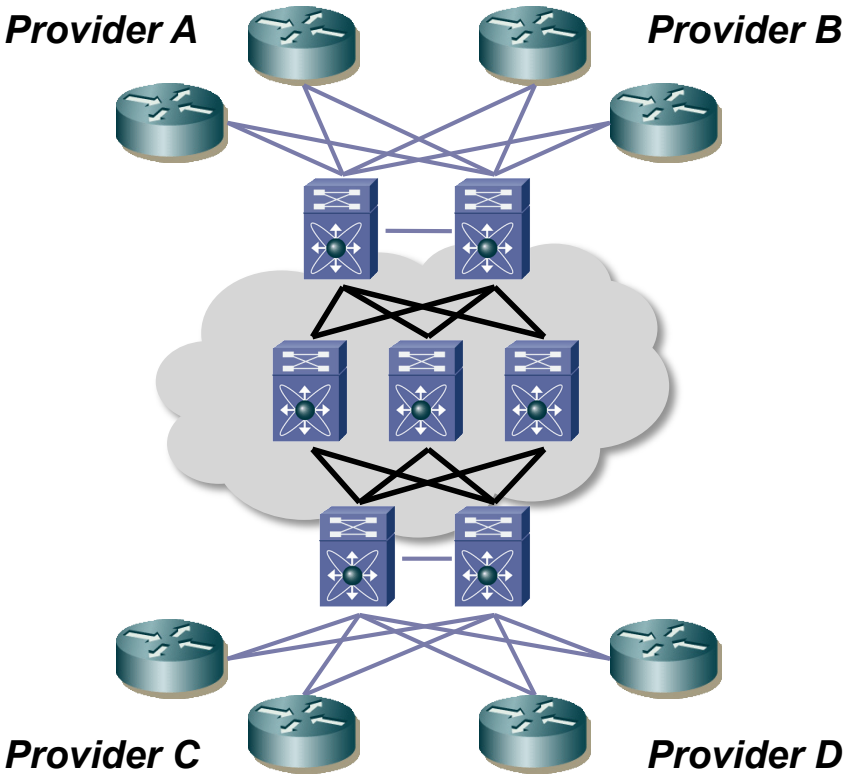
HPC Requirements

- HPC Clusters require high-density of compute nodes
- Minimal over-subscription
- Low server to server latency

FabricPath Benefits for HPC

- FabricPath enables building a high-density fat-tree network
- Fully non-blocking with FabricPath ECMP & port-channels
- Minimize switch hops to reduce server to server latencies

Use Case: L2 Internet Exchange Point



IXP Requirements

- Layer 2 Peering enables multiple providers to peer their internet routers with one another
- 10GE non-blocking fabric
- Scale to thousands of ports

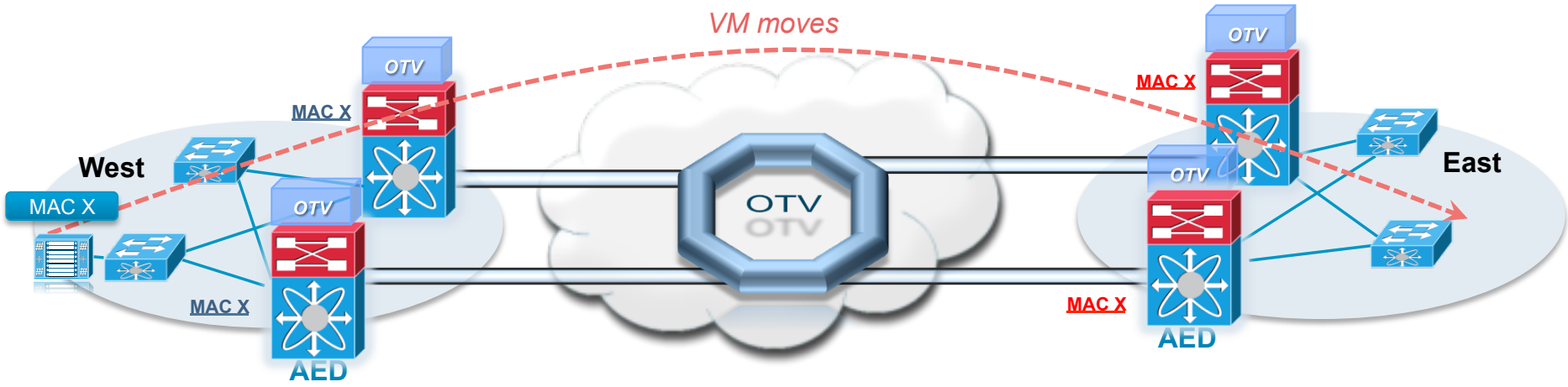
FabricPath Benefits for IXP

- Transparent Layer 2 fabric
- Scalable to thousands of ports
- Bandwidth not limited by chassis / port-channel limitations
- Simple to manage, economical to build

OTV

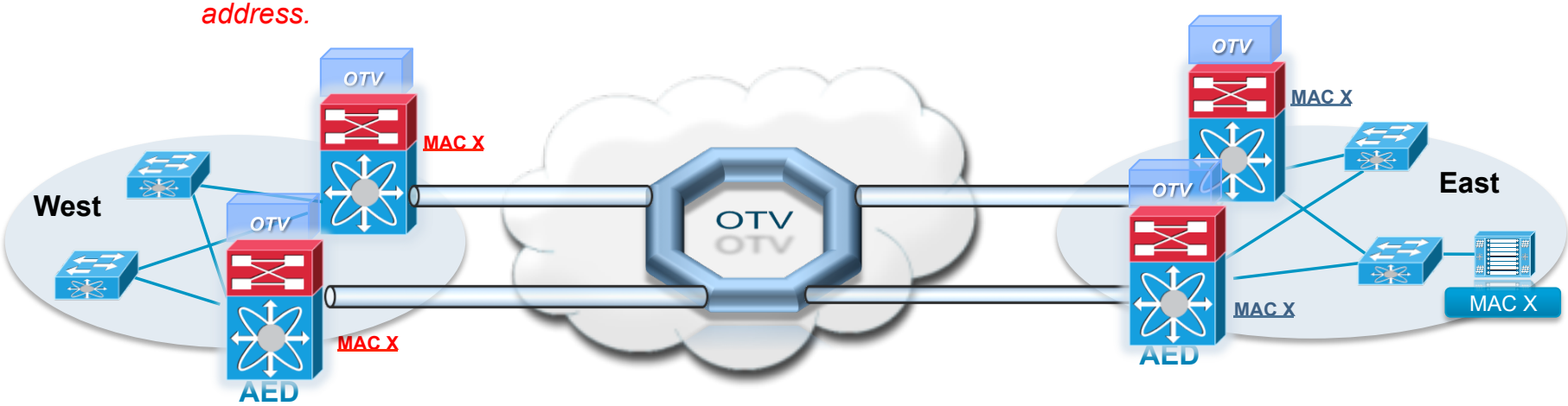
SearchServerVirtualization.com
BEST OF vmworld 2010
GOLD AWARD
HARDWARE
FOR
VIRTUALIZATION

Virtual Machine Mobility



Local MAC = **Blue**
Remote MAC = **Red**

Site West see MAC X advertisement with a better metric from site East and change them to remote MAC address.



SearchServerVirtualization.com
BEST OF vmworld 2010
GOLD AWARD
HARDWARE FOR VIRTUALIZATION

OTV wins Best of VMworld 2010 Gold Award

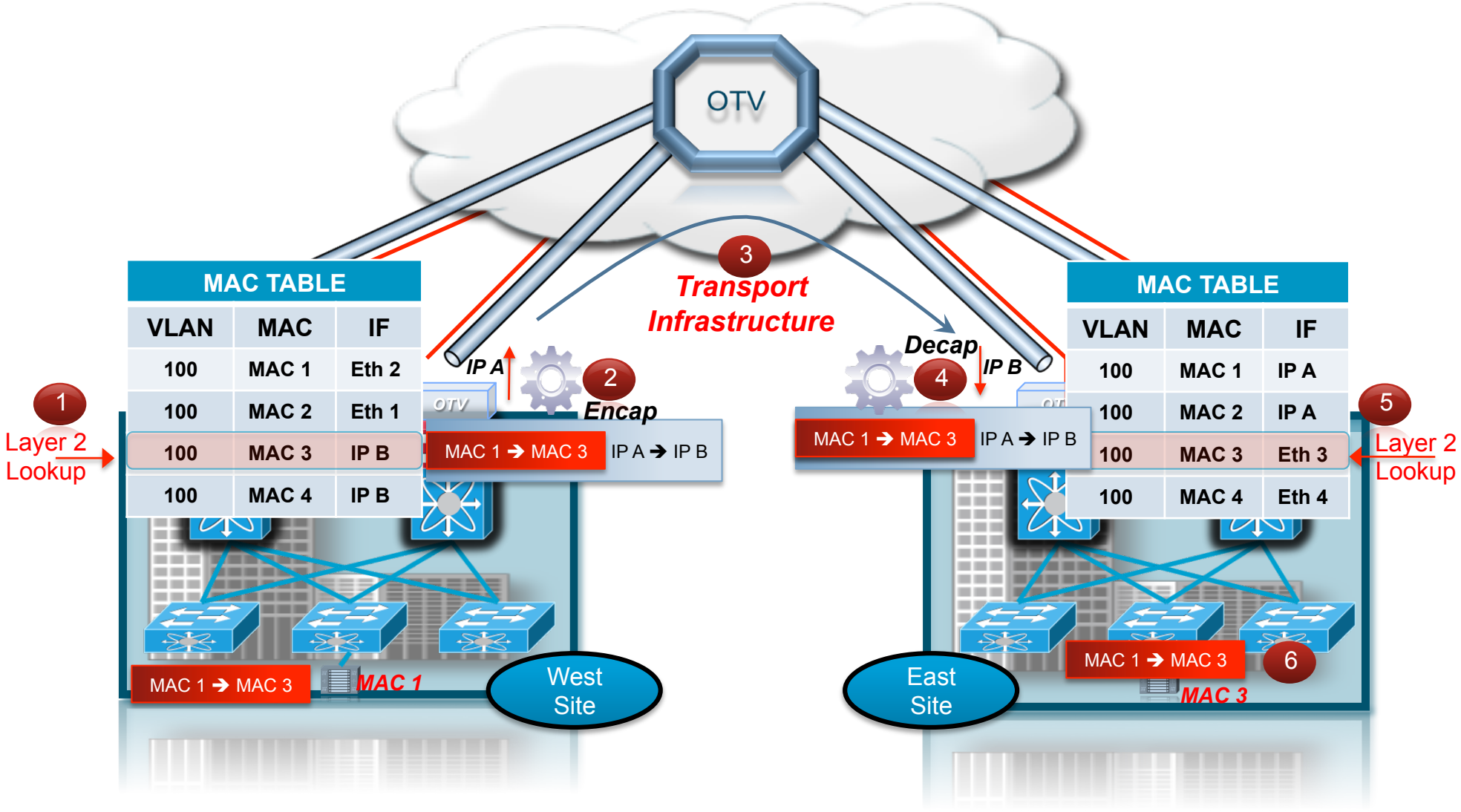
We are thrilled to announce that **Cisco Nexus 7000** won the prestigious Best Of VMworld 2010 award in the **Hardware for Virtualization** category for **Overlay Transport Protocol (OTV)**.

The panel of judges for the Best of VMworld 2010 awards was a mix of industry experts, IT consultants and TechTarget editors. 200 entrants were scored on innovation, the value provided by the product, performance, reliability and ease of use.

What the judges said: "*Cisco Nexus 7000 Overlay Transport Virtualization lets you extend data networks across data centers, which has tremendous benefits for multi-site disaster recovery.*"

Go here for more information about the award
Best of VMworld 2010 Awards

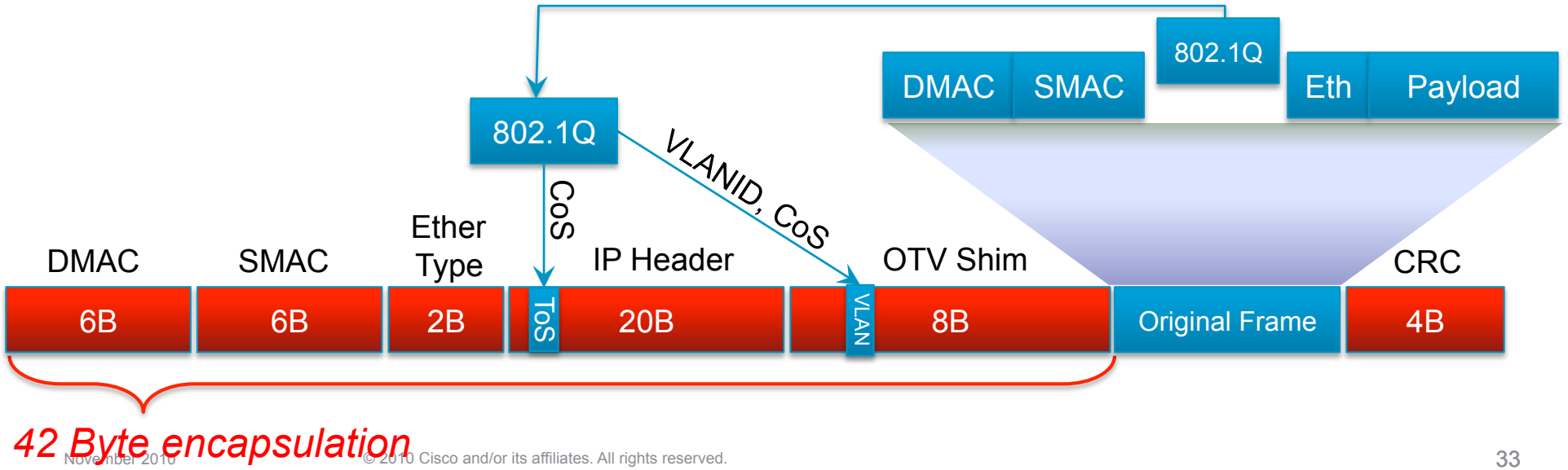
OTV Data Plane: Inter-Site Packet Flow



OTV Encapsulation

Consideration

- OTV adds a 42 Byte IP encapsulation
- The OTV shim header contains VLAN ID, Overlay number and CoS
- The OTV Edge Devices do NOT perform packet fragmenting and reassembling. A packet failing the MTU is dropped by the Forwarding Engine
- Make sure that *[xB + 42B] < DCI MTU*... where *x* = Size of original packet



OTV QoS for IP Traffic

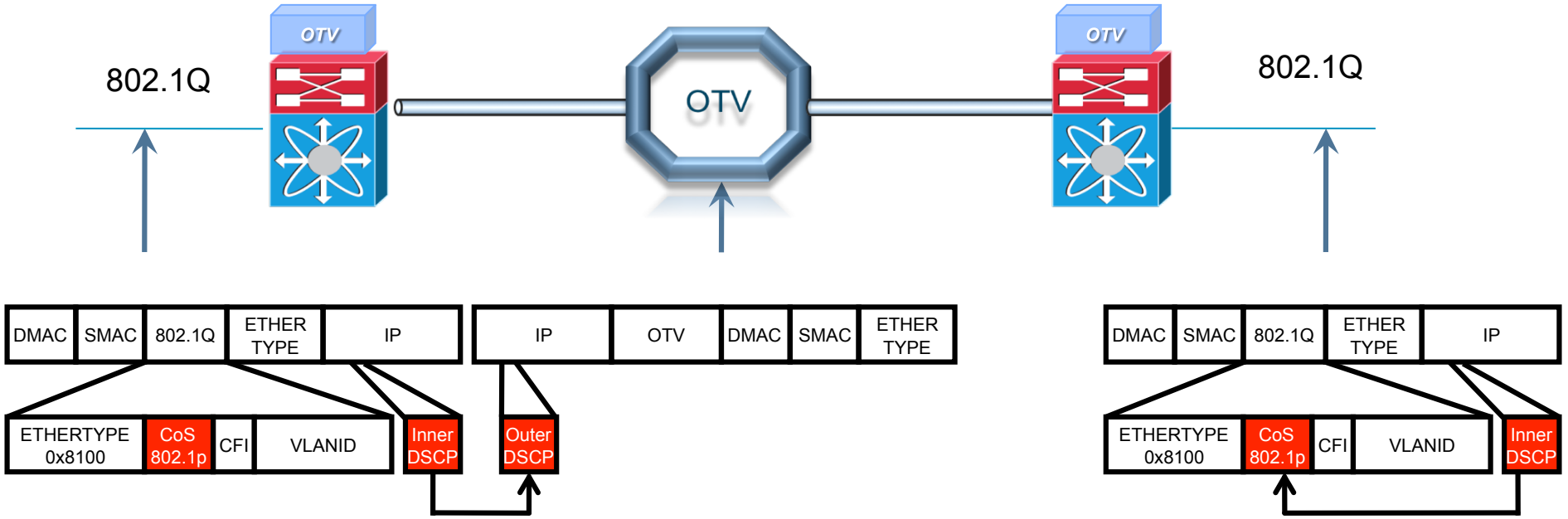
Default Behavior in NX-OS 5.0(3)

- Site sending the traffic (Encap side):

  The original (inner) DSCP value is copied to “outer” DSCP

- Site receiving the traffic (Decap side):

  The original (inner) DSCP is used to populate CoS



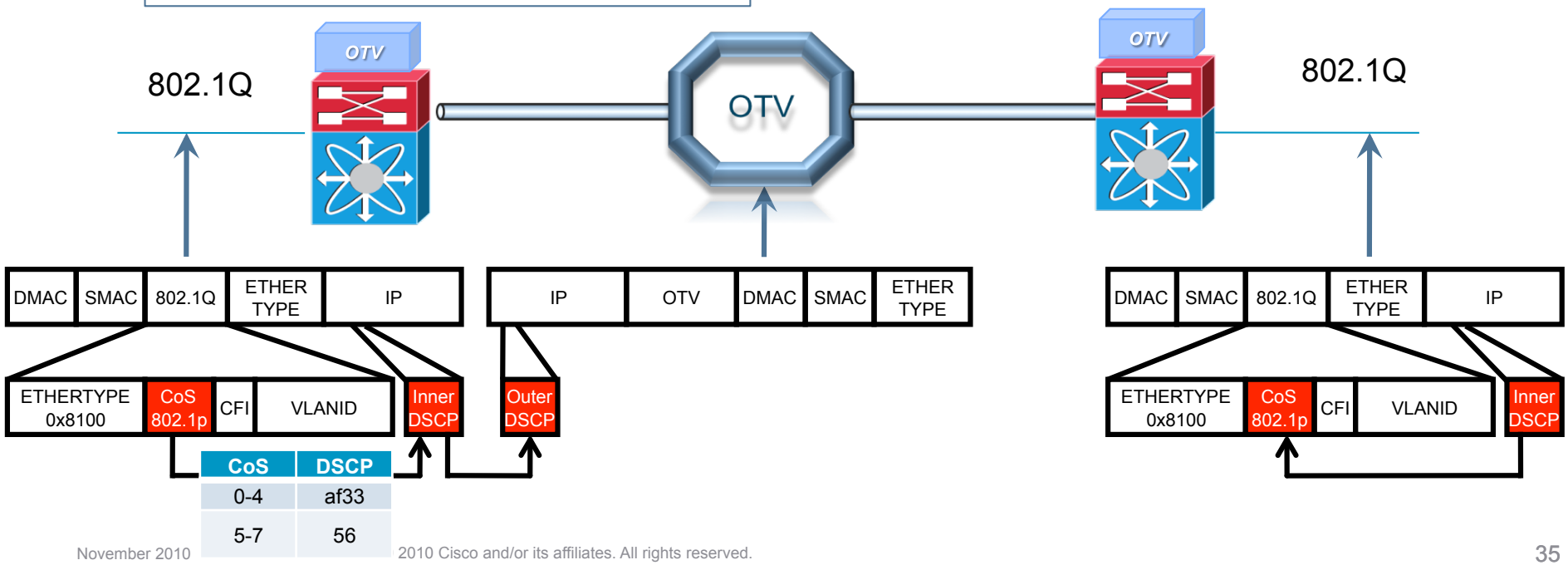
OTV QoS for IP Traffic

CoS Mapping in NX-OS 5.0(3)

- On the internal interfaces you can configure Cos-to-DSCP maps

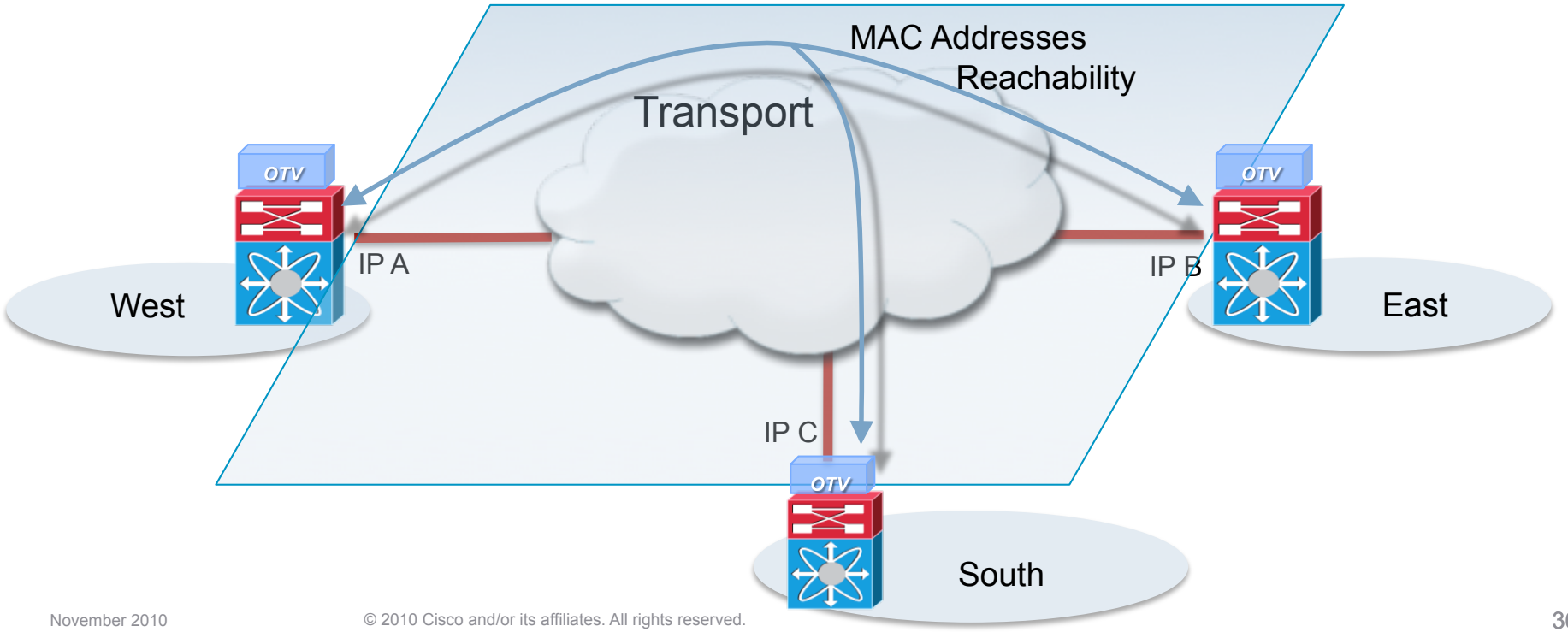
```
class a
   match cos 0-4
class b
   match cos 5-7

policy-map otv-cos-2-dscp
  class a
    set dscp af33
  class b
    set dscp 56
```

```
!applying to extended vlans
vlanx-y
  service-policy input otv-cos-2-dscp
```



Building the MAC tables
The OTV Control Plane

- The OTV control plane **proactively advertises** MAC reachability (control-plane learning)
- The MAC addresses are advertised in the **background** once OTV has been configured
- *No protocol specific configuration is required*



OTV Control Plane

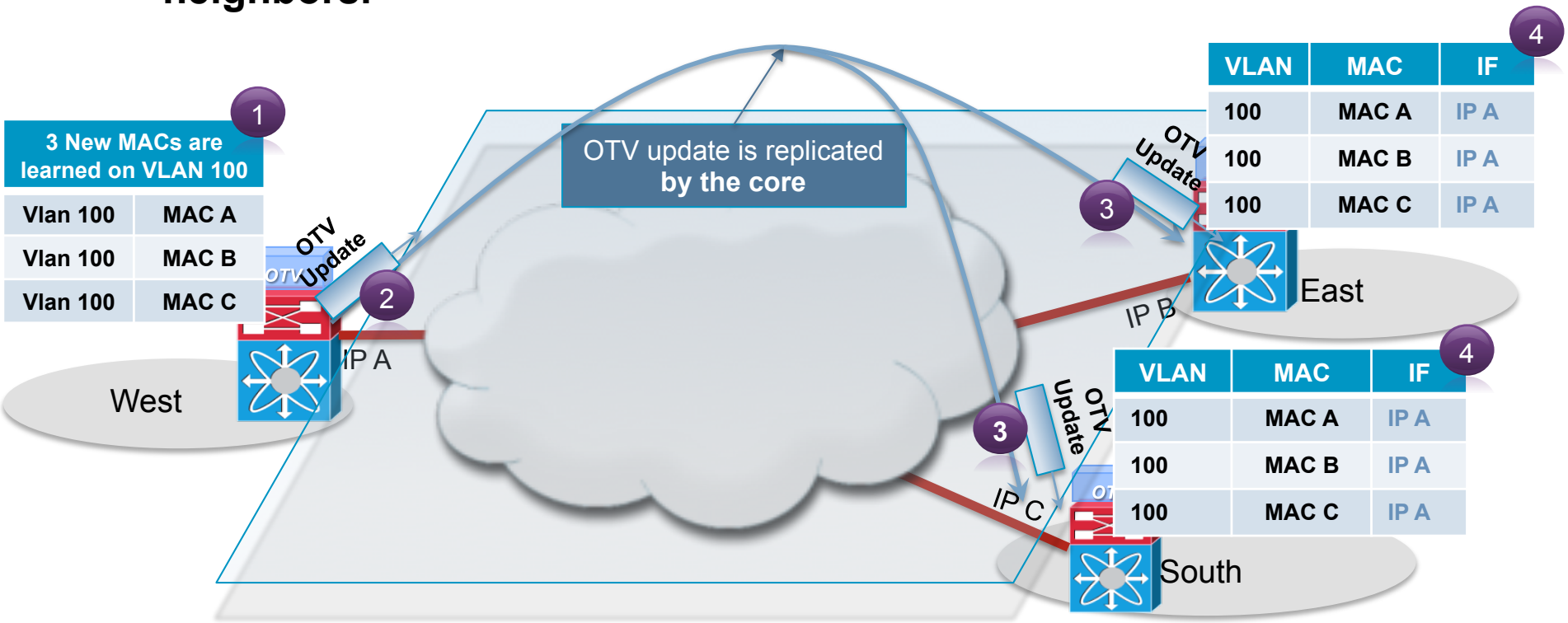
Neighbor Discovery and Adjacency Formation

- Before any MAC address can be advertised the *OTV Edge Devices must build a neighbor relationship with each other*
- The neighbor relationship can be built over:
  - a **multicast-enabled** transport infrastructure
  - an **unicast-only** transport infrastructure
- Multicast preferred:
  - More efficient use of resources
- In 5.0(3) and 5.1.1, OTV supports only multicast-enabled transports
- For a non-multicast enabled core, use the Adjacency Server mechanism (roadmap)

OTV Control Plane

MAC Address Advertisements (Multicast-Enabled Transport)

- When an Edge Device learns a new MAC address it advertises it together with its associated VLAN IDs and the IP address of the Join-interface
- A single OTV update can contain multiple MACs from different VLANs
- With a multicast-enabled transport a **single update reaches all neighbors.**



Multicast Groups in the Core

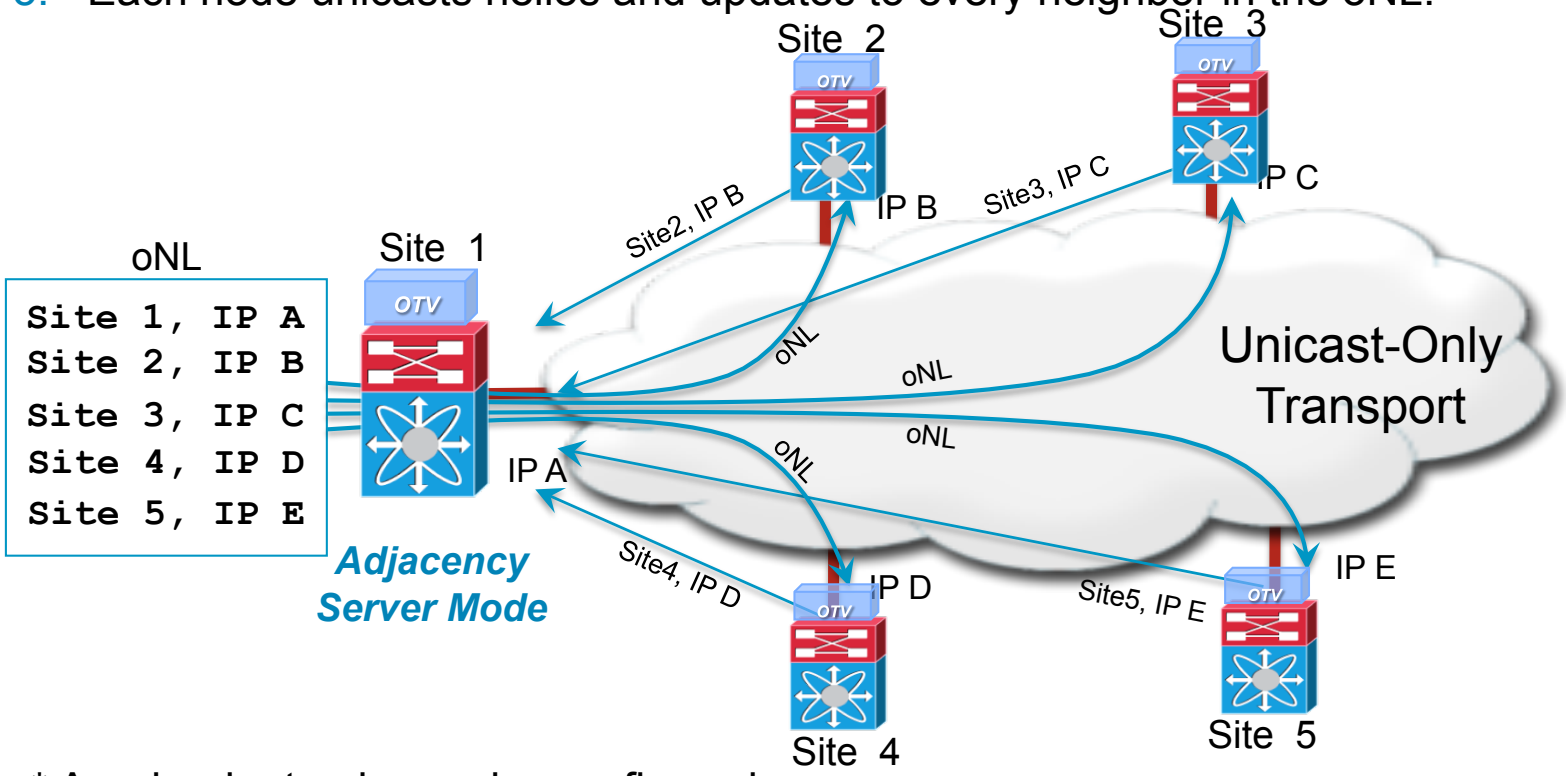
- Control group – Single PIM-SM or PIM-bidir group used to form adjacencies and exchange MAC reachability information
- Data groups – Range of SSM groups used to carry multicast data traffic generated by the sites

Future

OTV Control Plane

Neighbor Discovery (Unicast-Only Transport)

1. One of the OTV Edge Devices (ED) is configured as an Adjacency Server (AS)*.
2. All EDs are configured to register to the AS: send their site-id and IP address.
3. The AS builds a list of neighbor IP addresses: **overlay Neighbor List (oNL).**
4. The AS unicasts the oNLto every neighbor.
5. Each node unicasts hellos and updates to every neighbor in the oNL.



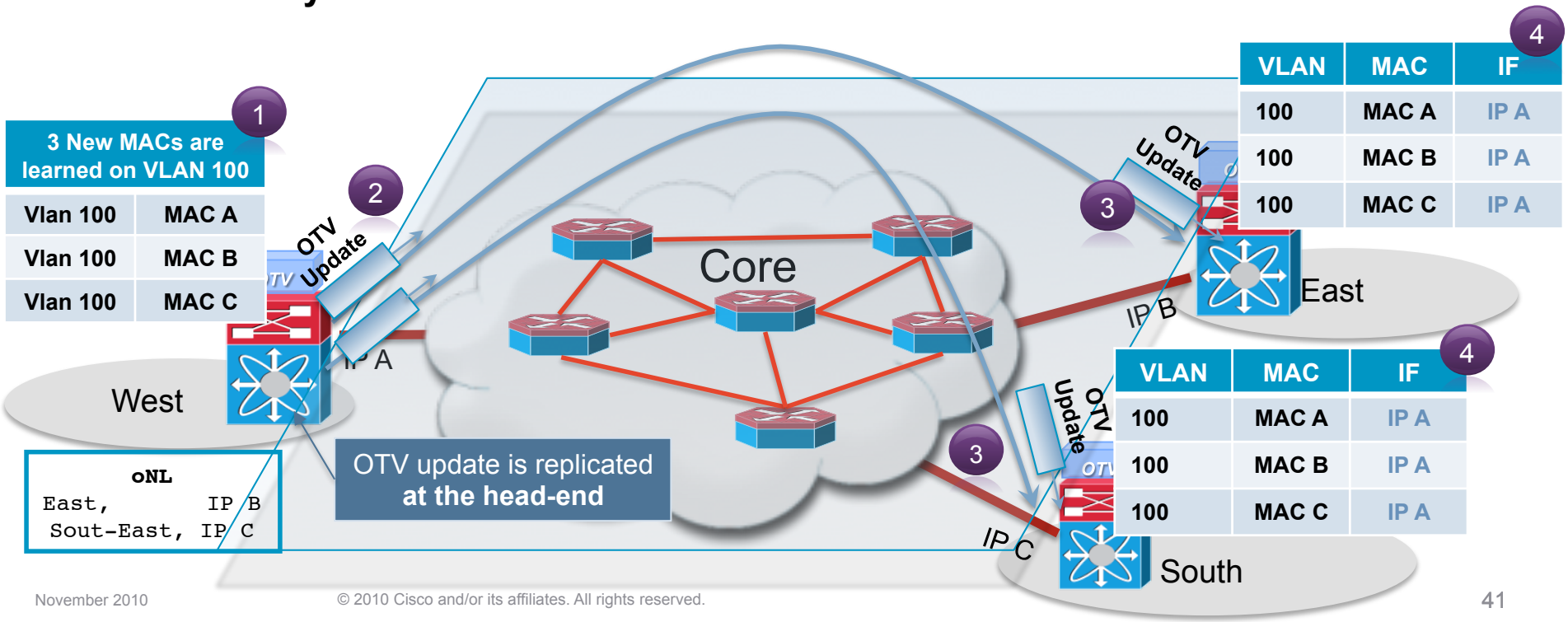
* A redundant pair may be configured

Future

OTV Control Plane

MAC Advertisements (Unicast-Only Transport)

- Every time an Edge Device learns a new MAC address, the OTV control plane will advertise it together with its associated VLAN IDs and IP next hop.
- The IP next hops are the addresses of the Edge Devices through which these MACs are reachable in the core.
- A single OTV update can contain multiple MAC addresses for different VLANs.
- **A single update needs to be created for each destination EDs present on the Overlay.**



OTV Provides Layer 2 Fault Isolation

- Spanning Tree
  - Each site has its own independent STP domain
  - Any STP issues are localised
  - Local site's STP domain is configured and operates as usual
- Unknown Unicasts
  - Are no longer needed
  - MAC moves are advertised, not flooded
  - Mechanism for handling silent hosts
- ARP Traffic
  - ARP snooping and caching at L3 boundary as usual
- All the above without configuration

OTV Automated Multi-homing

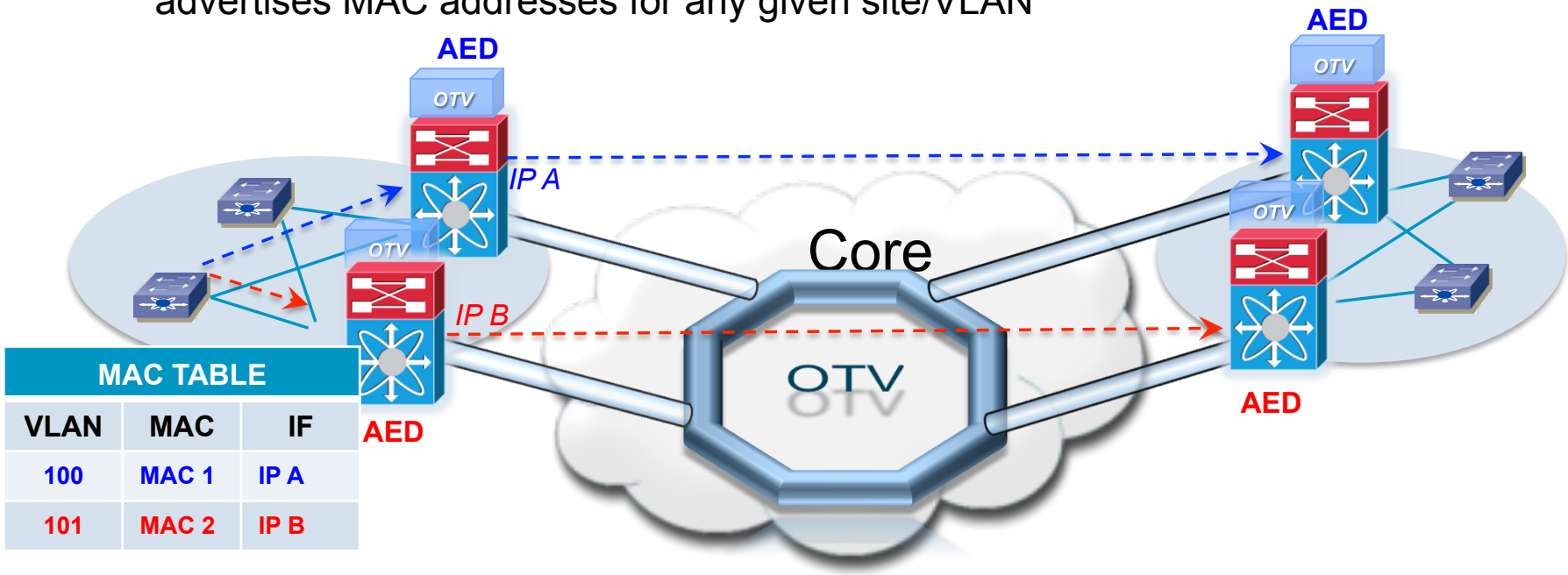
Per-VLAN Load Balancing

- The detection of the multi-homing is fully automated and it does not require additional protocols and configuration
- In each site OTV elects one of the Edge Devices to be the ***Authoritative Edge Device*** (AED) for a subset of the extended VLANs

  e.g. in a dual-homed site the VLANs will be split into odd and even VLANs
- The AED:

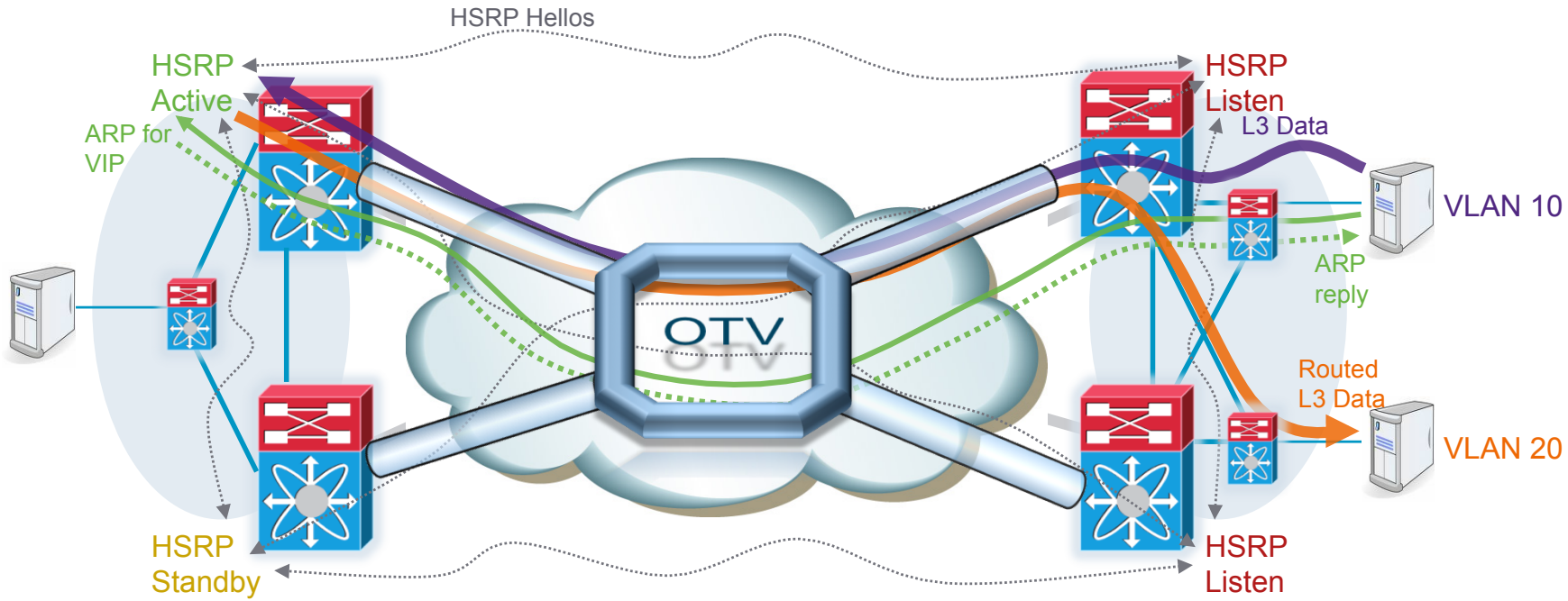
  forwards traffic to and from the overlay

  advertises MAC addresses for any given site/VLAN



Egress Routing with VLAN Extension

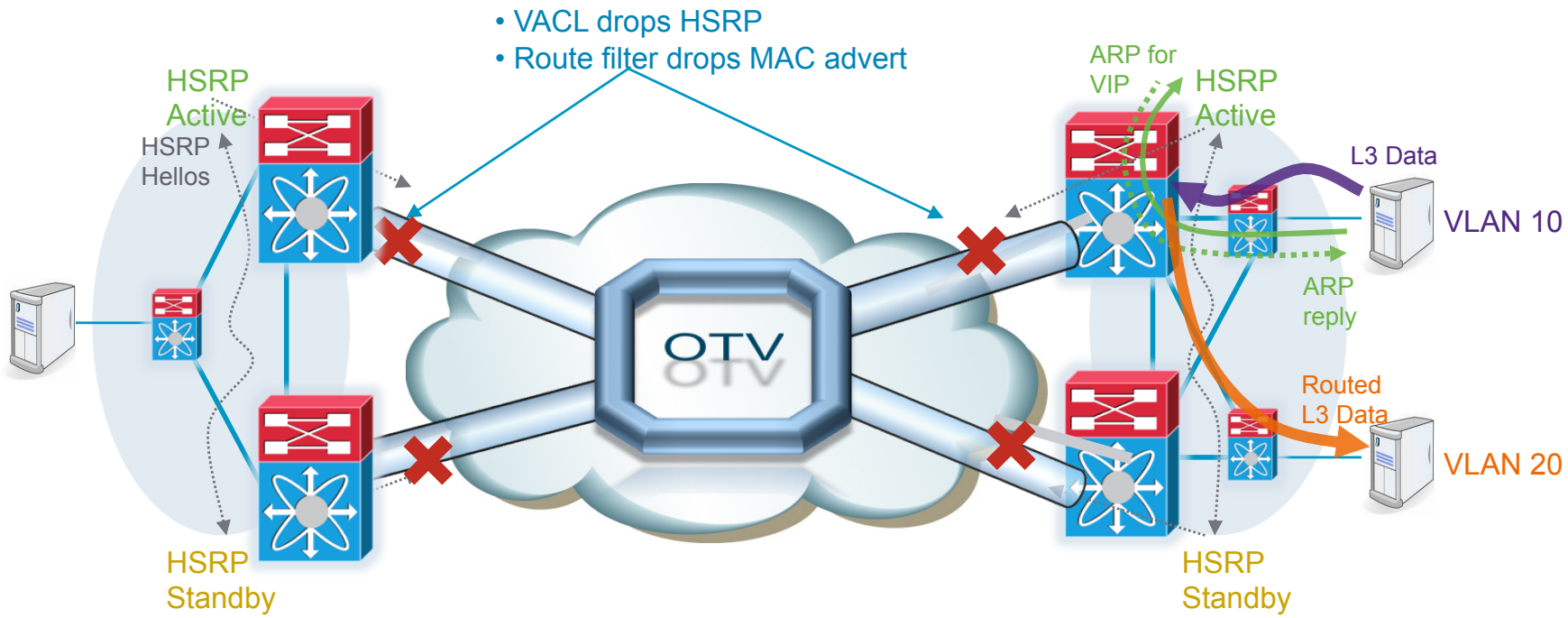
- Extended VLAN typically has associated HSRP group
- Only one HSRP router active, with all servers pointing to HSRP VIP as default gateway
- Result: sub-optimal (trombone) routing



Egress Routing with VLAN Extension

OTV FHRP Filtering Solution

- Filter FHRP with combination of VACL and OTV MAC route filter in OTV VDC
- Still have one HSRP group with one VIP, but now have active router at each site for optimal first-hop routing
- Native FHRP filtering in OTV planned for future release



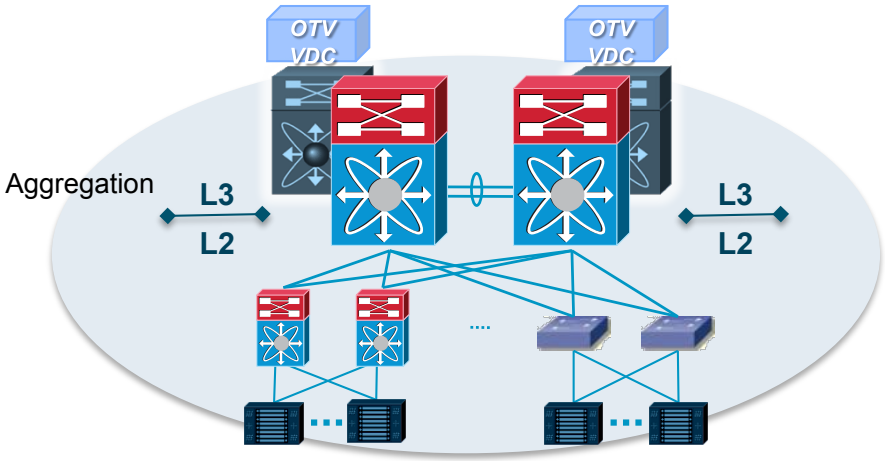
Ingress Path Optimization Techniques

- Available Now:
  - Active/Standby subnet advertisement
  - Reverse Health Injection (RHI): Host based /32 announcement
  - ACE/GSS: DNS based Global Site Selection
  - Whitepapers on cisco.com
- Coming in 2011
  - Locator/ID Separation Protocol – LISP

OTV Edge Device at the Aggregation

OTV–SVI Separation

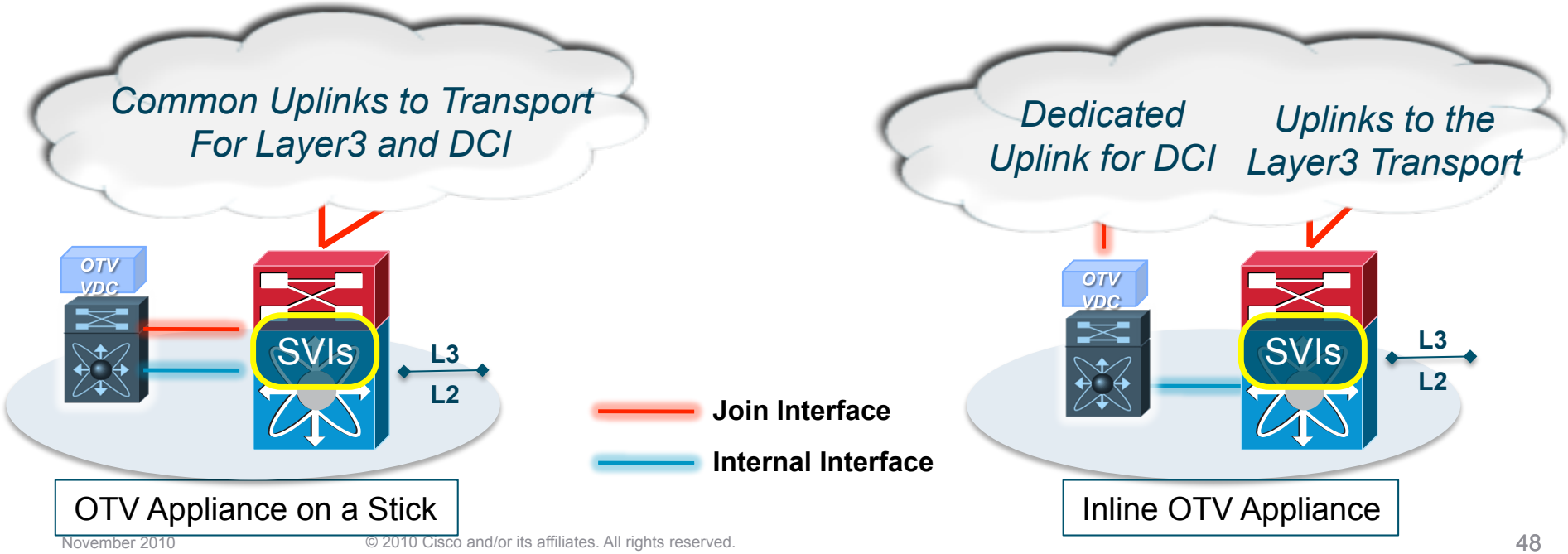
- At present, a VLAN cannot both be extended with OTV and have an SVI in the same device
- The separation between OTV and SVIs can be provided:
  - through the topology (e.g. SVI on firewall)
  - or by using a separate VDC for the OTV DCI



OTV VDC

OTV VDC Models

- Two different deployment models are considered for the OTV VDC based on the availability of uplinks to the DCI Transport:
  - OTV Appliance on a Stick
  - Inline OTV Appliance
- From an OTV functionality prospective there is NOT difference between the two models. The Inline OTV Appliance can provide better convergence results

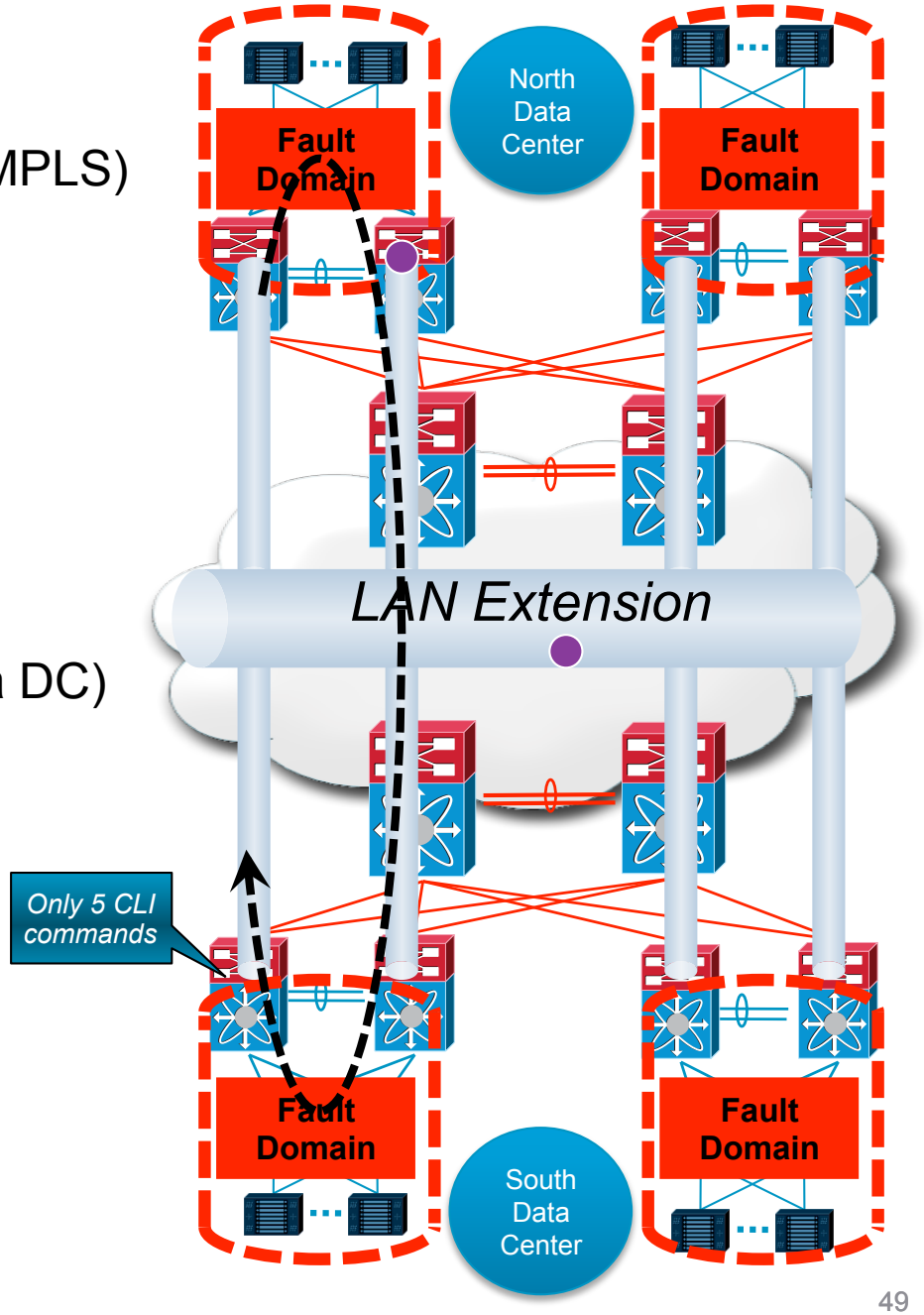


OTV Appliance on a Stick

Inline OTV Appliance

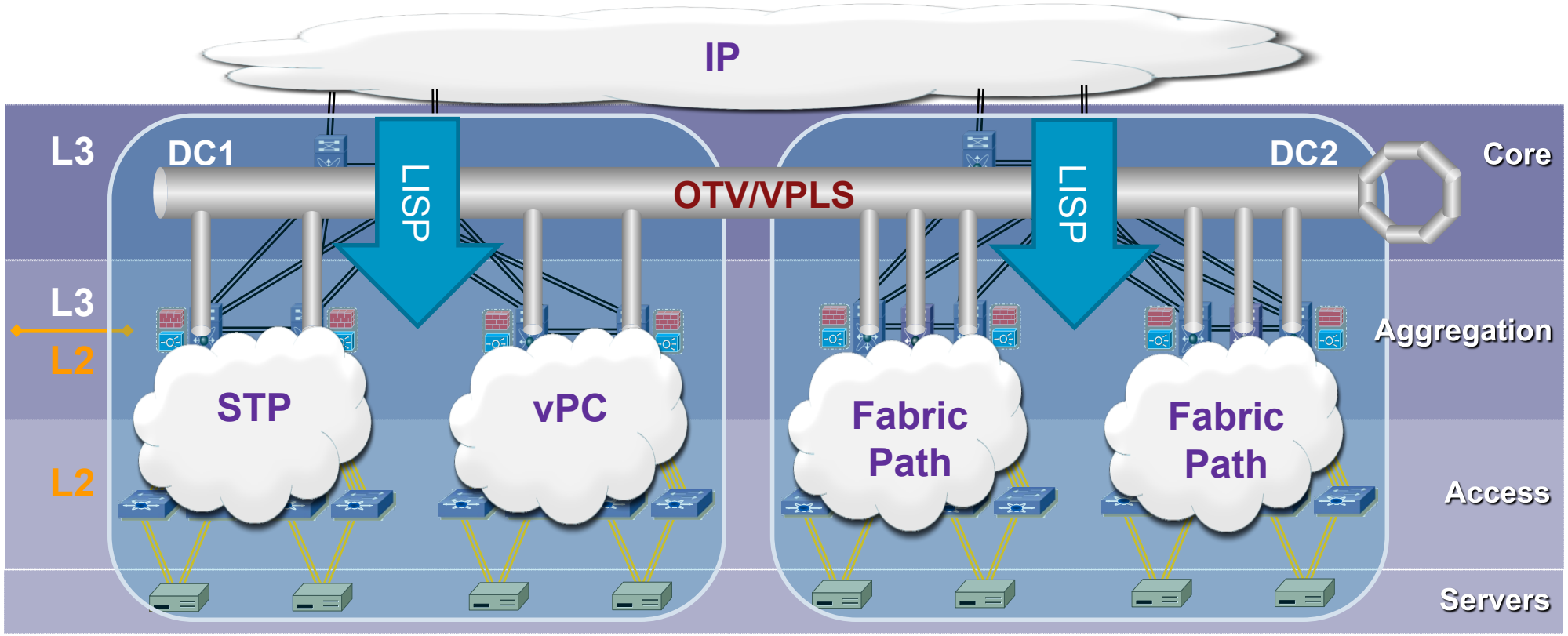
Summary
Real Problems Solved by OTV

- Extensions over any transport (IP, MPLS)
- Preserves the Failure boundary
- Site independence / isolation
- Optimal BW utilization
  (no head-end replication)
- Resiliency/multihoming
- Built-in end-to-end loop prevention
- Multisite connectivity (inter and intra DC)
- Scalability
  - VLANs, sites, MACs
  - ARP, broadcasts/floods
- Operations simplicity



LISP

IP Path Optimization for Workload Distribution

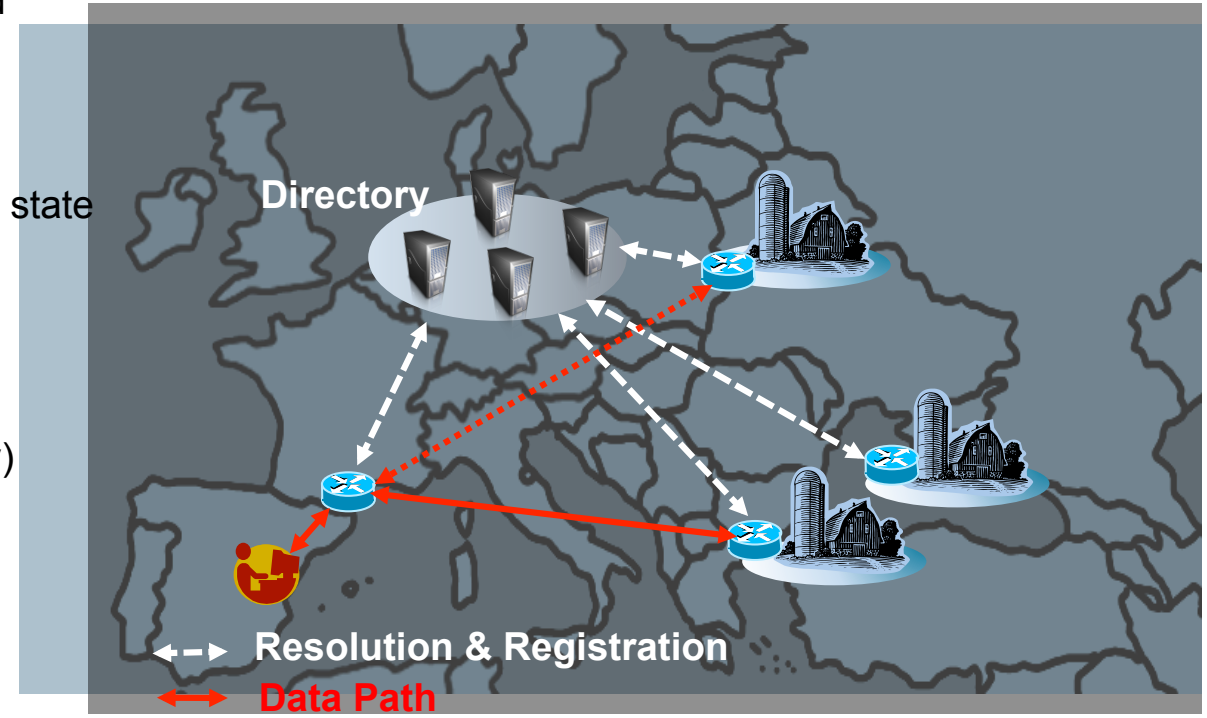


- LANs extend across PODs and/or DCs
- Extended subnets create suboptimal routing
- LISP allows the optimization of the routing and aids IP mobility

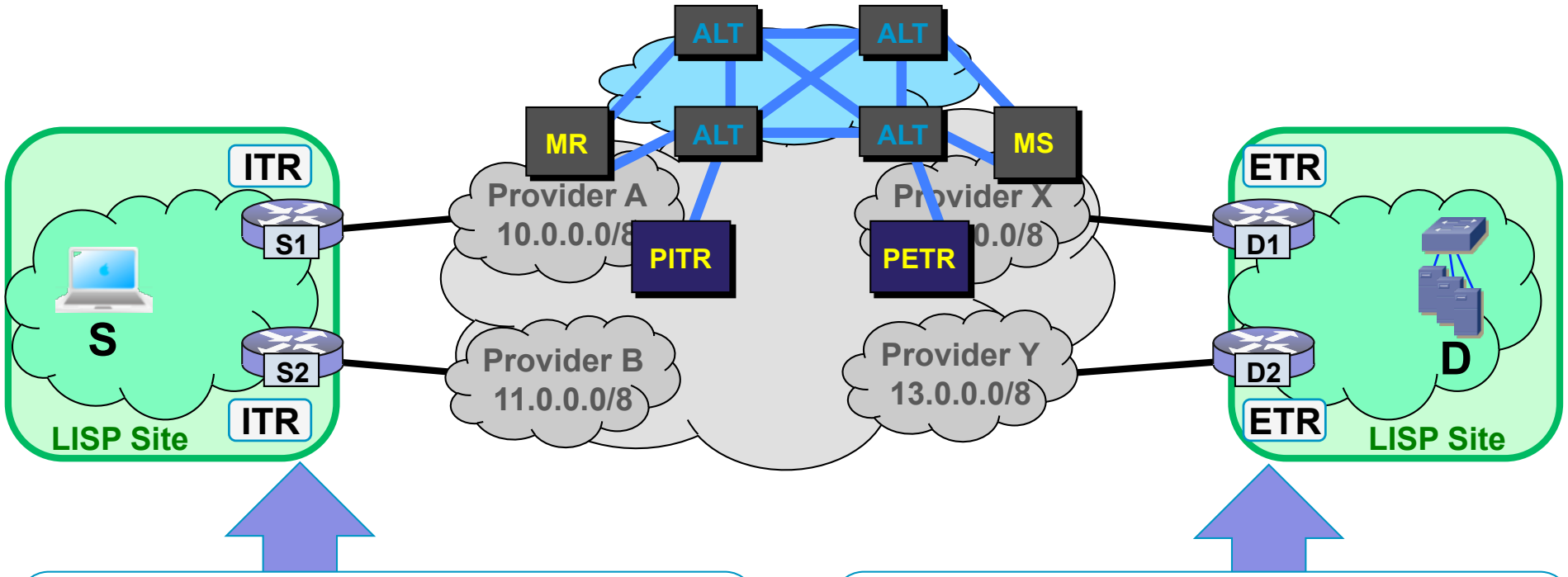
LISP Next Gen Routing Architecture

Locator-ID Separation Protocol

- Today: an IP Address = Identity + Location bundled together
- LISP decouples Identity (Host IP) from Location (Gateway IP)
- ID to Location mappings are kept in an 'out-of-band' Directory
- Traffic is routed in the core based solely on location
  - Traffic is IP in IP encapsulated
- LISP Benefits
  - Internet & Intranet Scalability
  - Reduction of Routing Table IP state
  - Flexible Routing Policy
  - Prefix Portability
  - Seamless Mobility
  - VPN semantics (multi-tenancy)
  - IPv4/IPv6 co-existence



LISP Components
Ingress / Egress Tunnel Router (xTR)



ITR – Ingress Tunnel Router

- Receives packets from site-facing interfaces
- Encapsulates to remote LISP site (or natively forwards to non-LISP site)

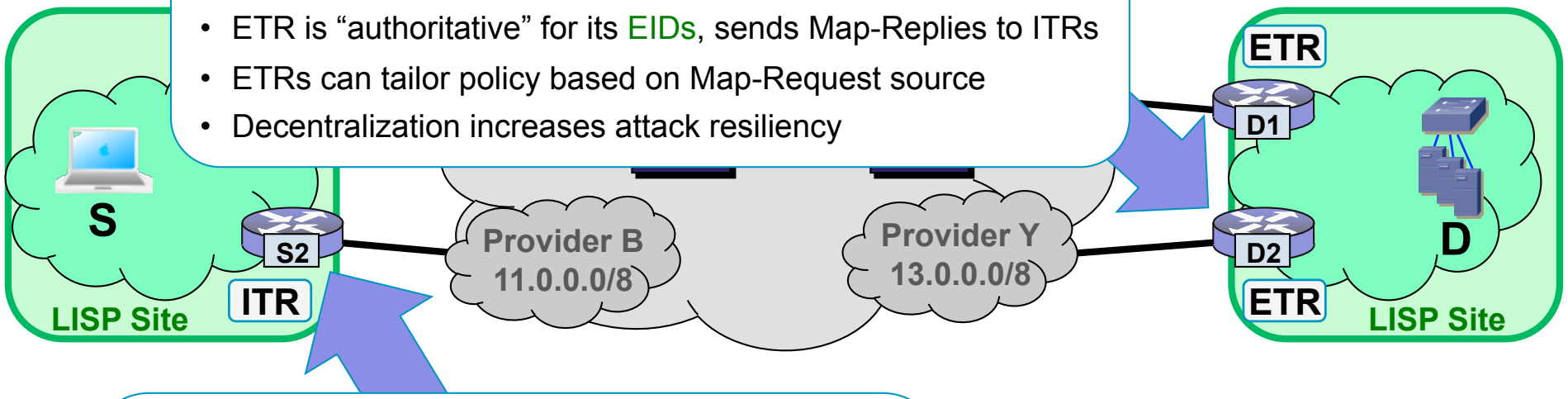
ETR – Egress Tunnel Router

- Receives packets from core-facing interfaces
- De-caps and delivers packets to local **EIDs** at the site

Control Plane
Local Mapping Database & Map Cache

LISP Site Mapping-Database

- EID-to-RLOC mappings in all ETRs for local LISP site
- ETR is “authoritative” for its EIDs, sends Map-Replies to ITRs
- ETRs can tailor policy based on Map-Request source
- Decentralization increases attack resiliency



LISP Map Cache

- “Lives” on ITRs and only stores mappings for sites to which ITR is currently sending packets.
- Map-Cache populated by sending Map-Requests through ALT and receiving Map-Replies from ETRs
- ITRs must respect Map-Reply policy, including TTLs, RLOC up/down status, RLOC priorities/weights

Workload Distribution / VM-Mobility

Future

Use case

Needs:

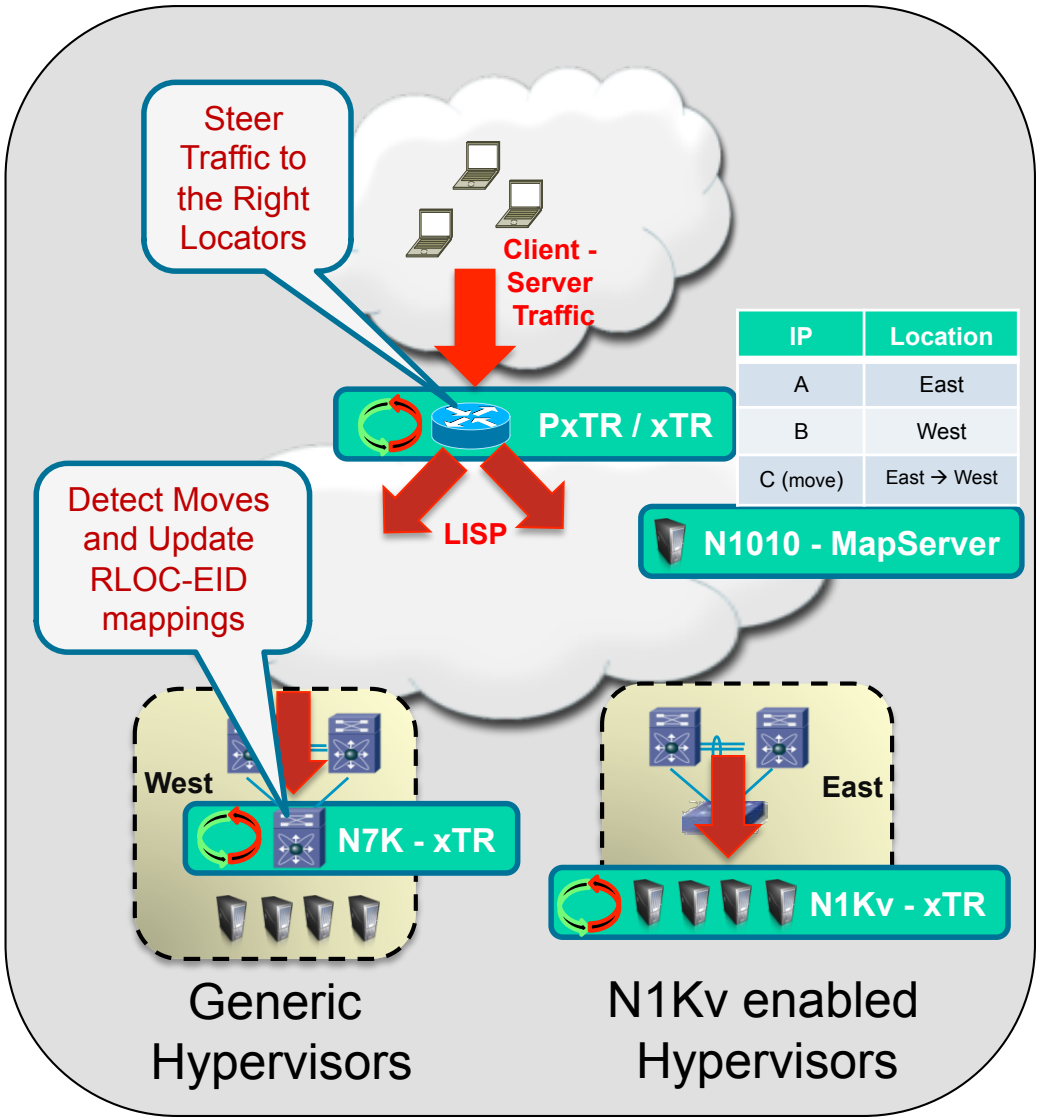
- VM-Mobility
  - Move Detection
  - Dynamically update EID-to-RLOC mappings
  - Traffic Redirection

LISP Solution:

- xTR Functions on Hypervisor, Access or Aggregation switches

Benefits:

- Integrated Mobility
- Direct Path (no triangulation)
- Connections maintained across move
- No routing re-convergence
- No DNS updates required
- Global Scalability (cloud bursting)
- IPv4/IPv6 Support
- ARP elimination

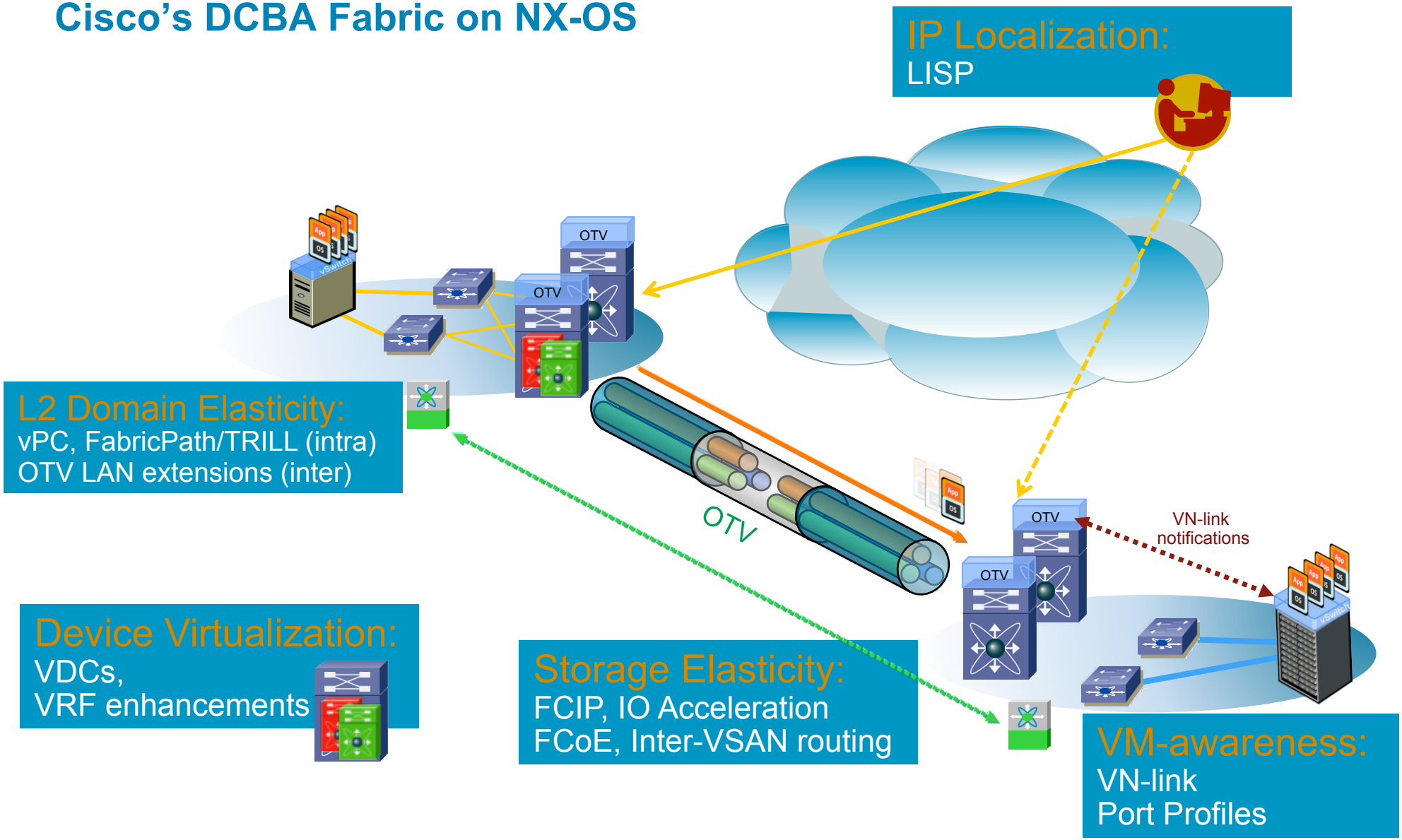


Agenda

- FabricPath
- OTV – Overlay Transport Virtualization
- LISP – Locator ID Separation Protocol
- Summary

NX-OS Innovations for Cloud & Virtualization

Cisco's DCBA Fabric on NX-OS



Compute resources are part of the cloud, location is transparent to the user