Cisco Expo
2012

Technológia Cisco FabricPath a skúsenosti s jej implementáciou

T-DC1/L2

Marián Klas, Tomáš Ondovčík

Cisco

Prosíme, ptejte se nás

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

Agenda

- FabricPath Solution Overview
- FabricPath Multicast Details
- FabricPath Unicast Details
- FabricPath and TRILL
- FabricPath Design Examples
- FabricPath Case Study

Agenda

- FabricPath Solution Overview
- FabricPath Multicast Details
- FabricPath Unicast Details
- FabricPath and TRILL
- FabricPath Design Examples
- FabricPath Case Study

Some Layer 2 Limitations

- Local problems have network-wide impact, troubleshooting is difficult
- Tree topology provides limited bandwidth
- Tree topology introduces sub-optimal paths
- STP convergence is disruptive
- MAC address tables don’t scale
- Host flooding impacts the whole network

Cisco FabricPath Goal

Switching

- Easy Configuration
- Plug & Play
- Provisioning Flexibility

Routing

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

FabricPath

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

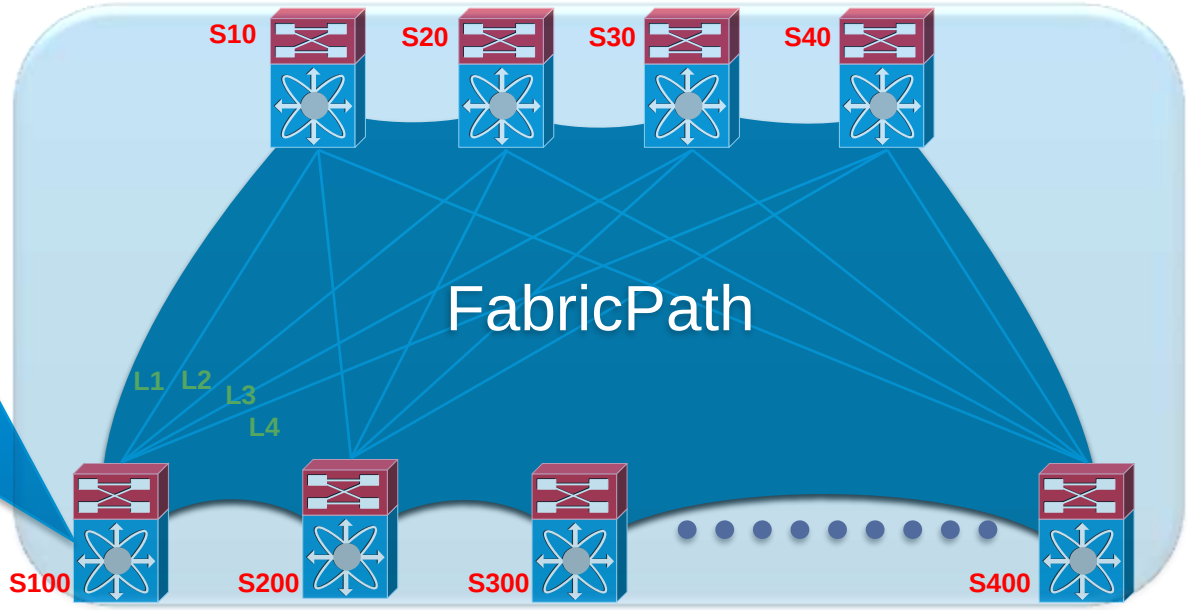
New Control Plane

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

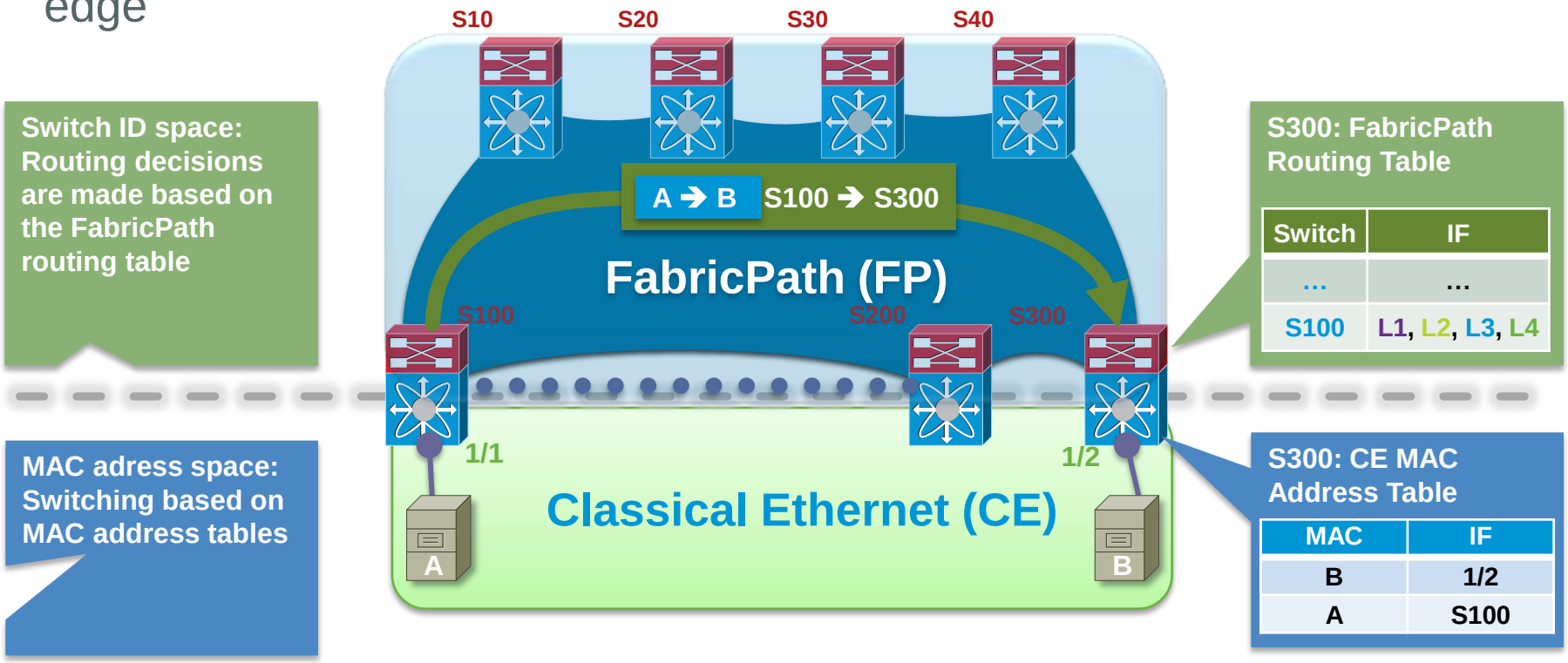
FabricPath Routing Table

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



New Data Plane

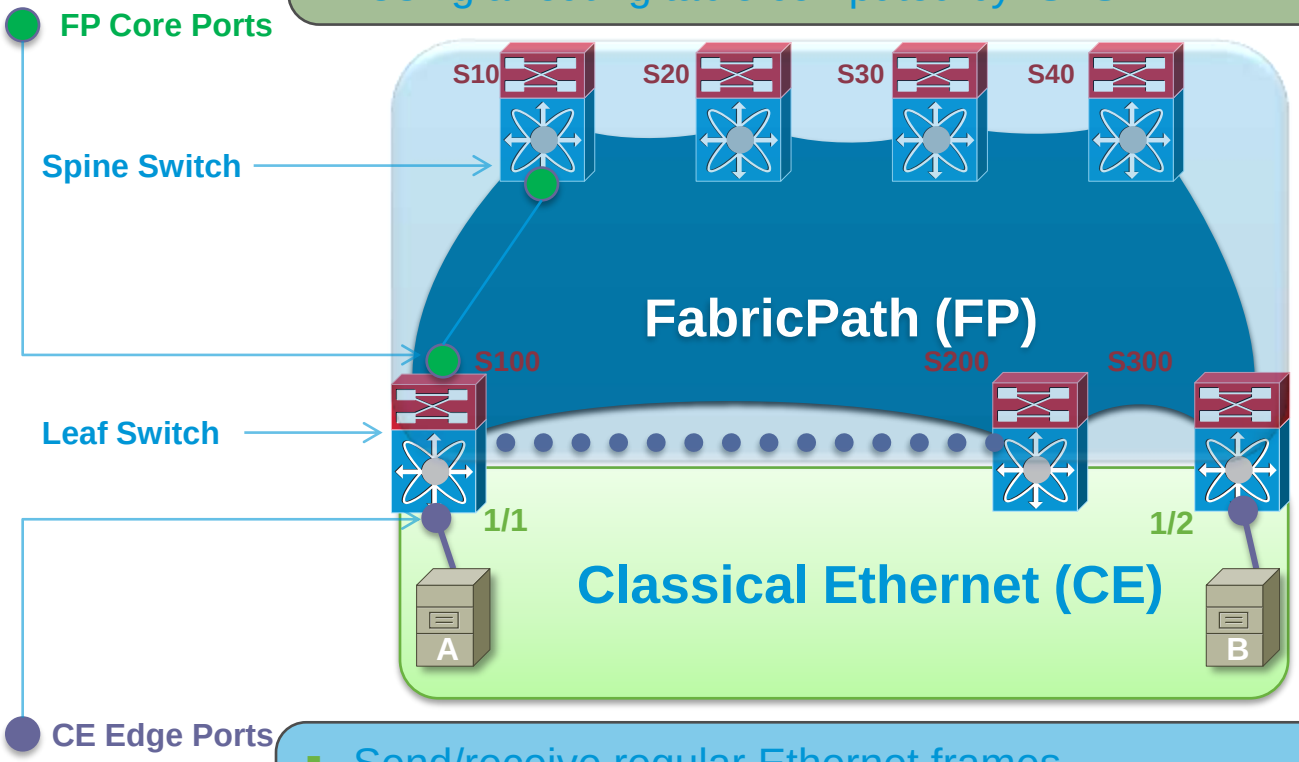
- The association MAC address/Switch ID is maintained at the edge



- Traffic is encapsulated across the Fabric

FabricPath Terminology

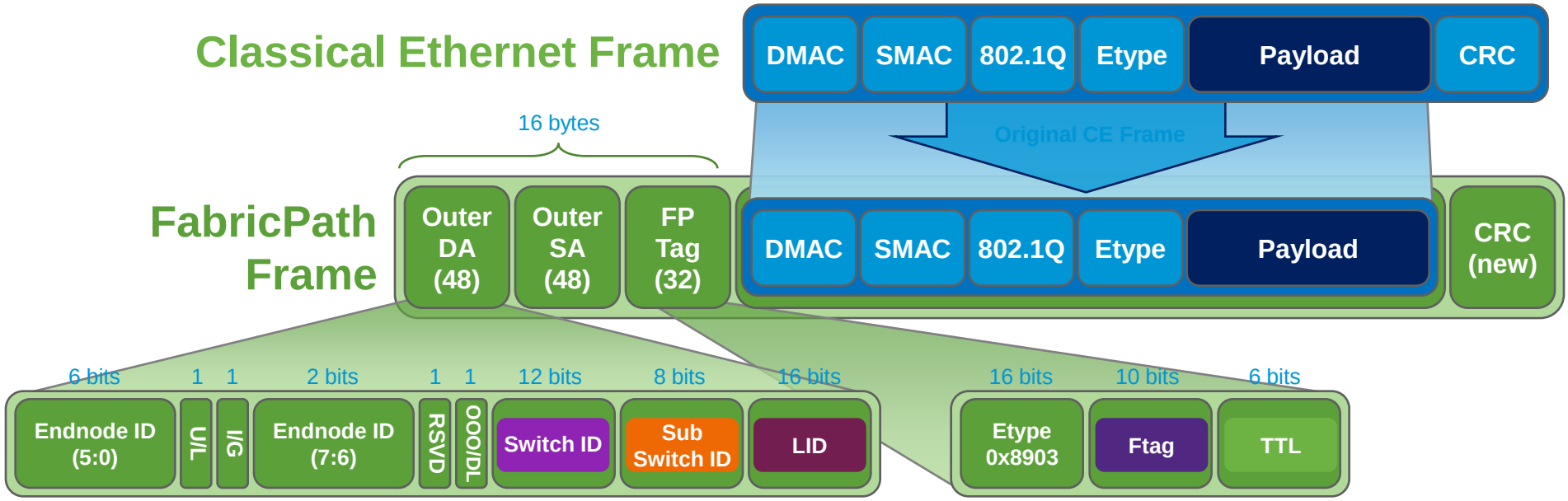
- Send/receive FabricPath frame
- No STP, no MAC learning, no MAC address table
- Using a routing table computed by IS-IS



- Send/receive regular Ethernet frames
- Run STP, do MAC address learning using a MAC address table

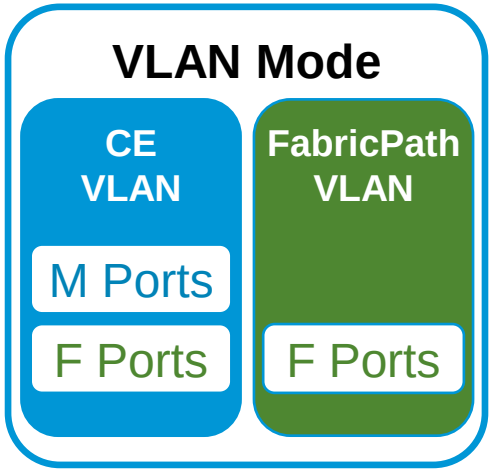
FabricPath Encapsulation

16-Byte MAC-in-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

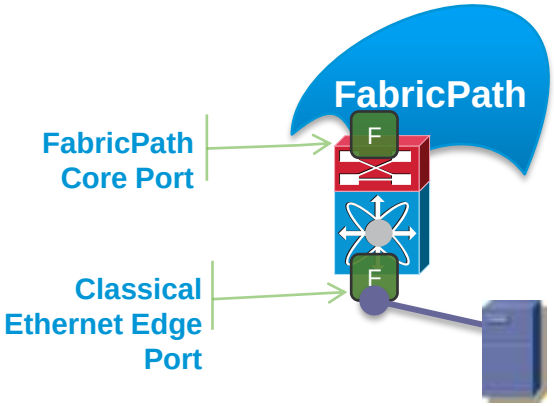
FabricPath VLANs



- The Nexus 7000 features M and F I/O Modules
- **FP Core and CE Edge ports must be on an F module**

- New FabricPath/CE locally significant VLAN mode:

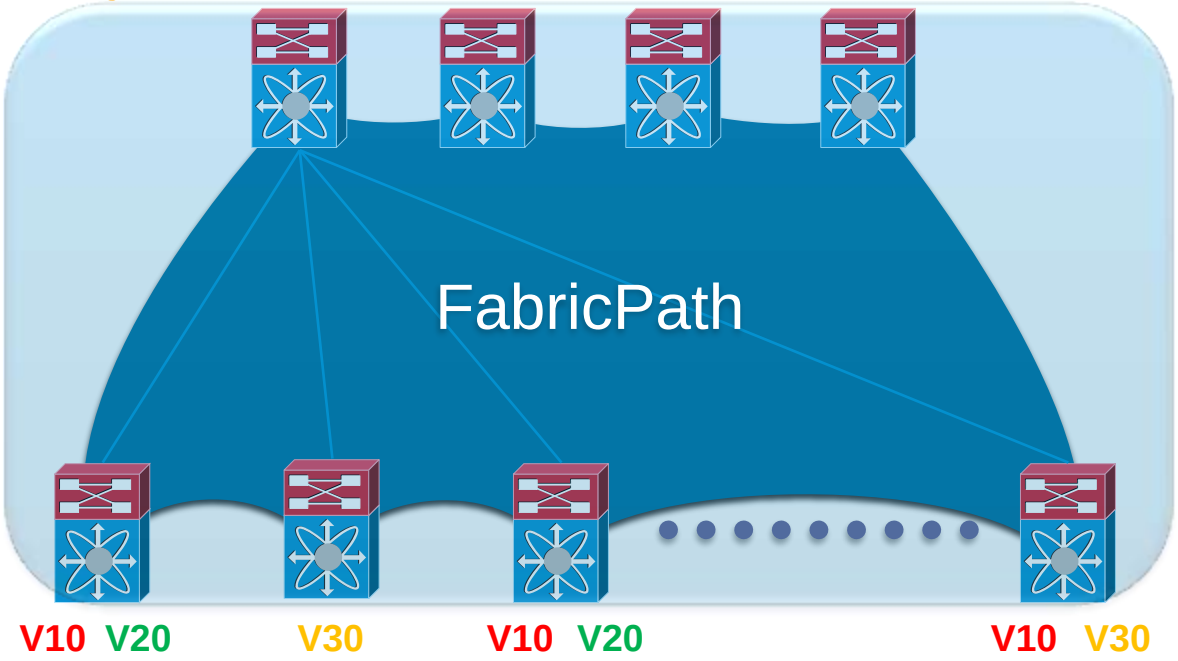
```
S100(config)# vlan 10
S100(config-vlan)# mode ?
  ce           Classical Ethernet VLAN mode
  fabricpath   Fabricpath VLAN mode

S100(config-vlan)# mode fabricpath
S100(config-vlan)#
```

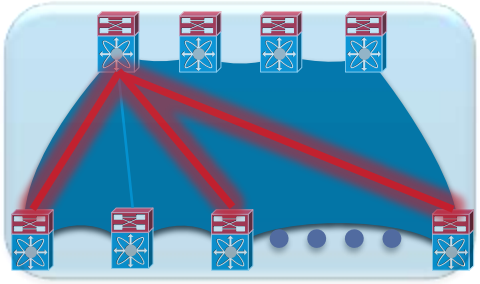


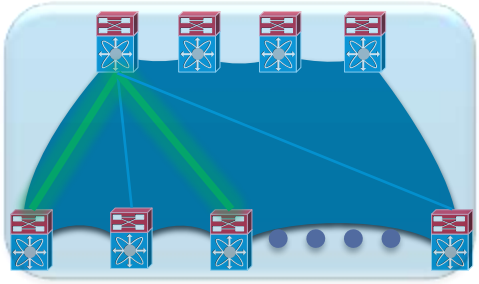
- FabricPath VLANs can only be enabled on F modules
- FabricPath VLANs are also relevant to the Nexus 5500
- Only traffic in FabricPath VLANs can traverse FabricPath domain

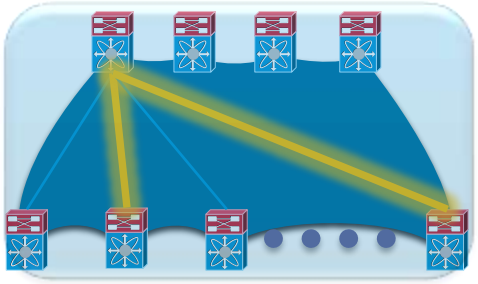
VLAN Pruning

Automatically Handled by IS-IS





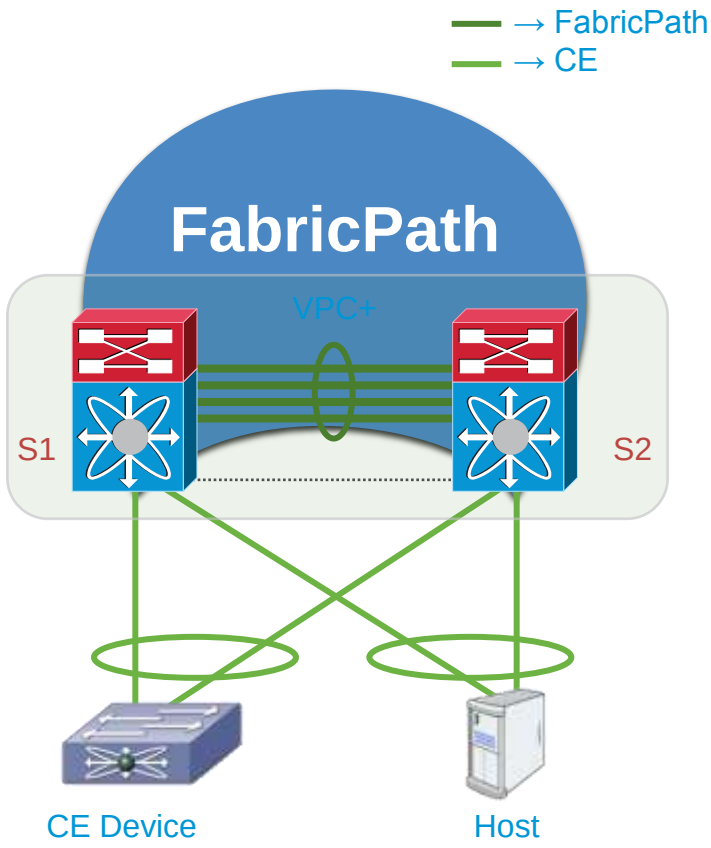




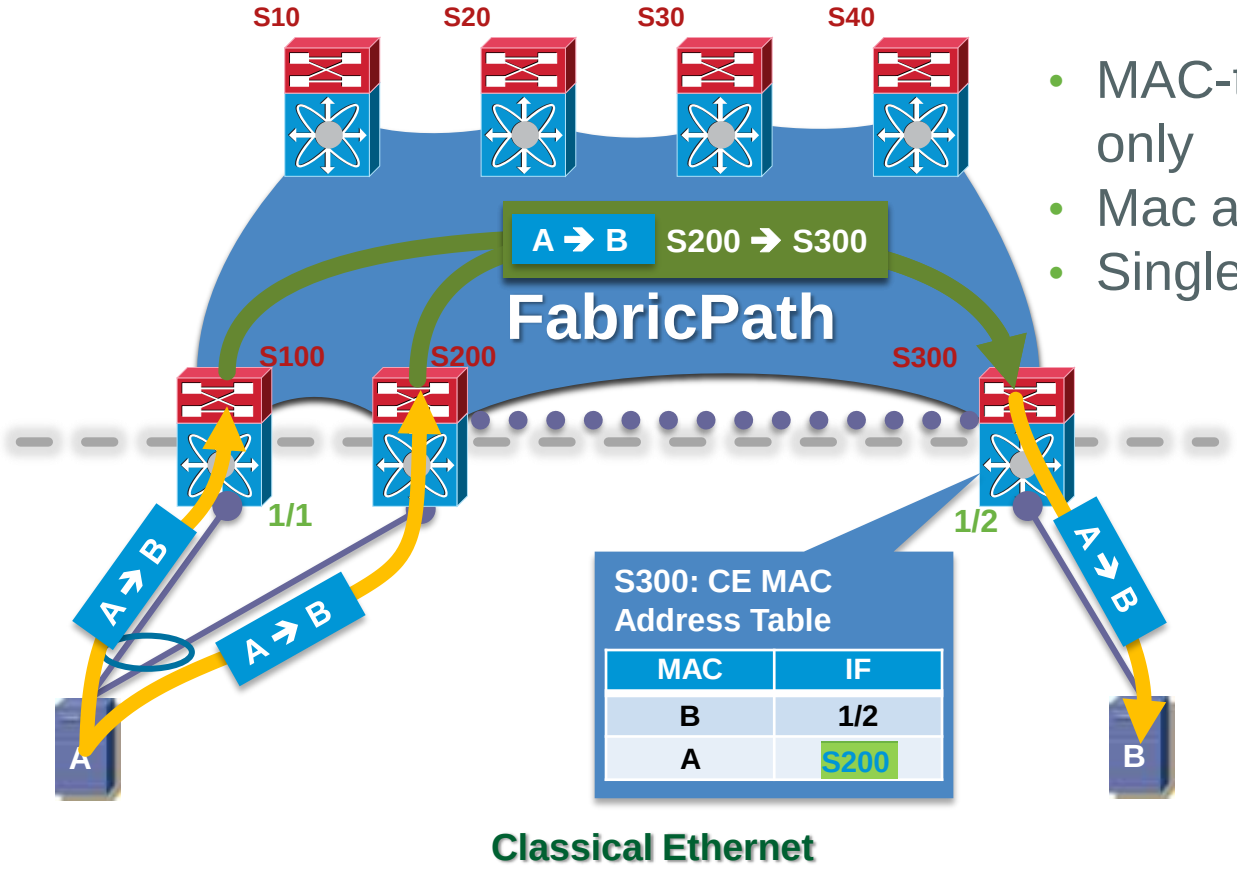
VPC+

Virtual Port Channel in FabricPath Environment

- Allows non FabricPath capable devices to connect redundantly to the fabric using port channels
- Provides active/active FHRP at the FabricPath to L3 boundary
- Configuration virtually identical to standard VPC

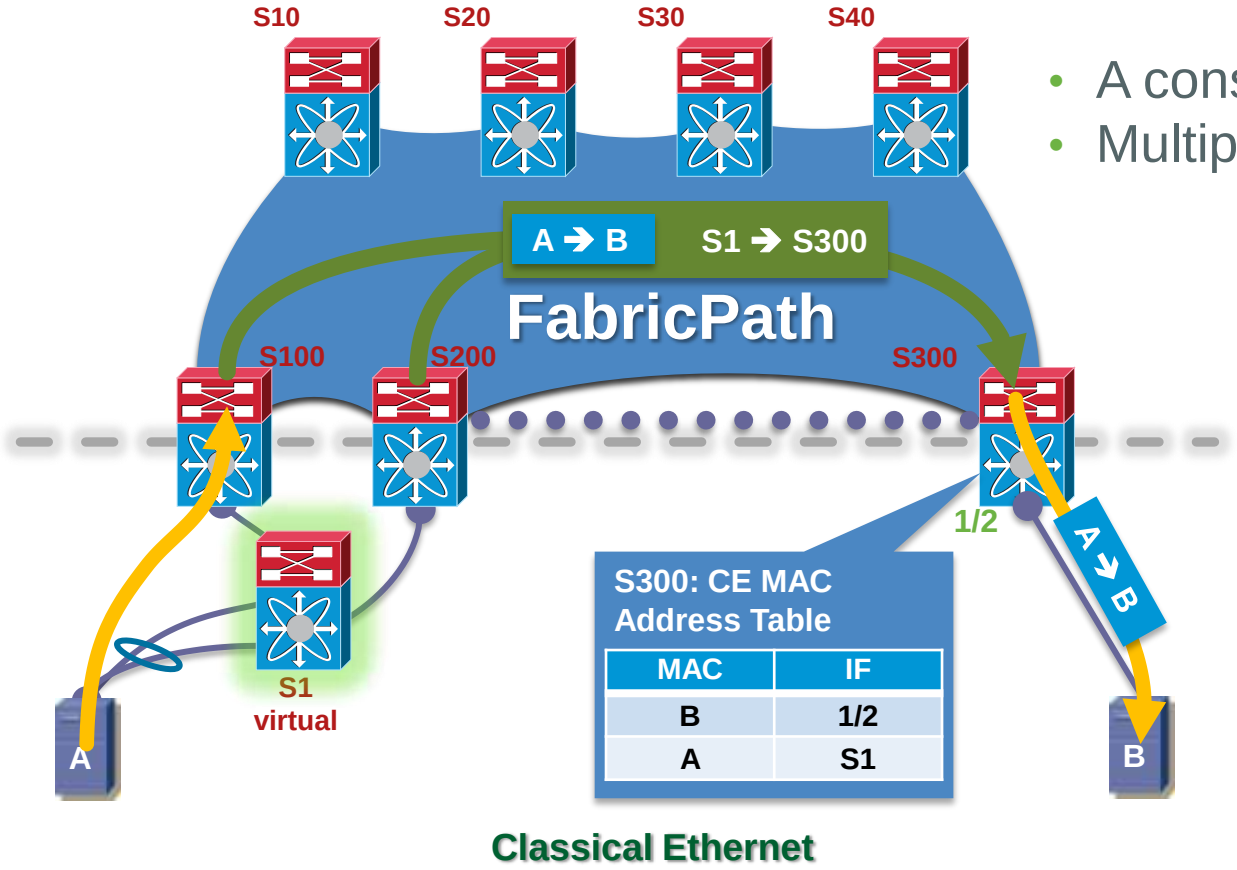


VPC+ Technical Challenges



- MAC-to-SID mapping is 1:1 only
- Mac address flapping on S300
- Single path to A

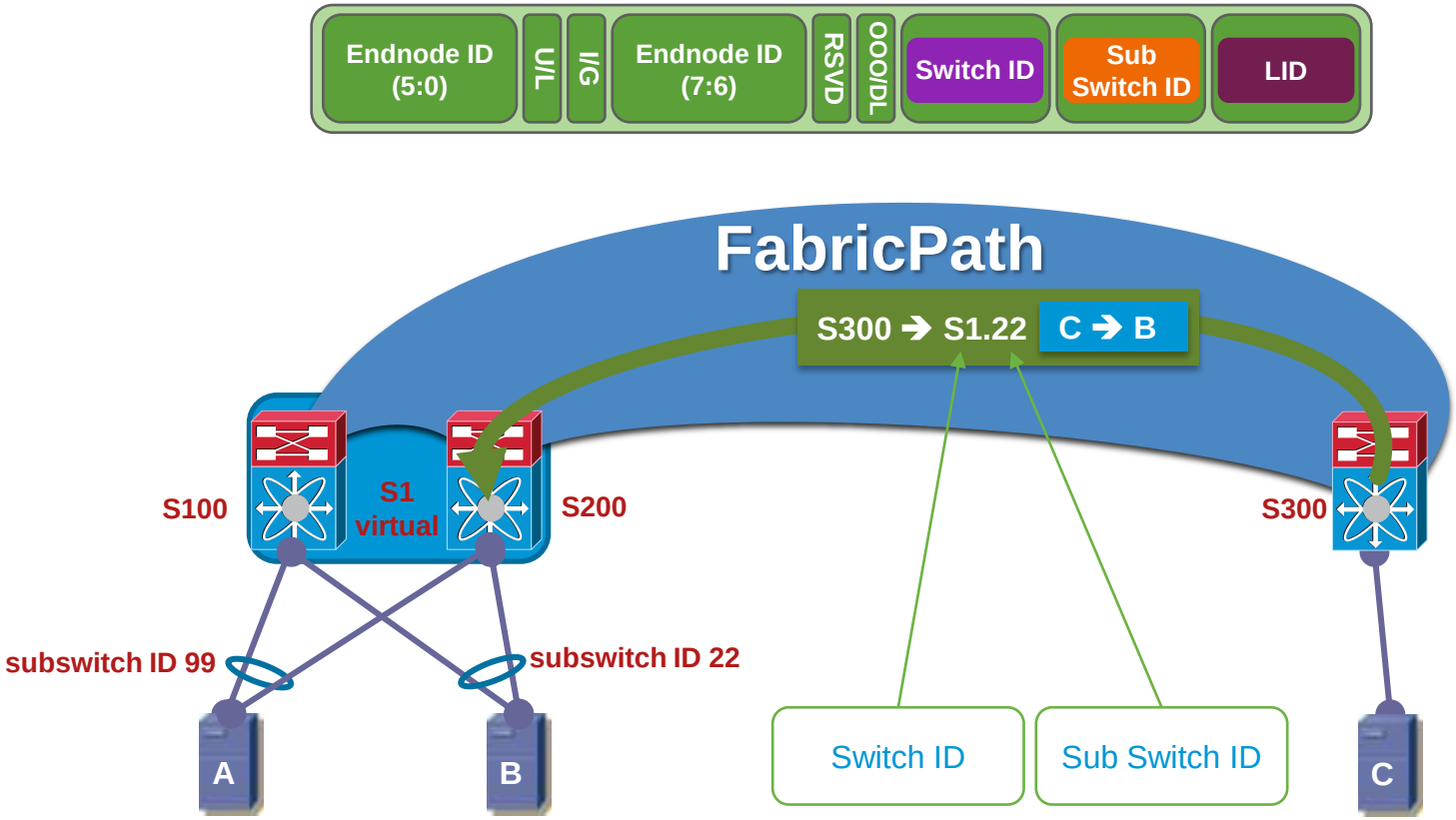
VPC+ Virtual Switch



- A consistently associated to S1
- Multipathing to A

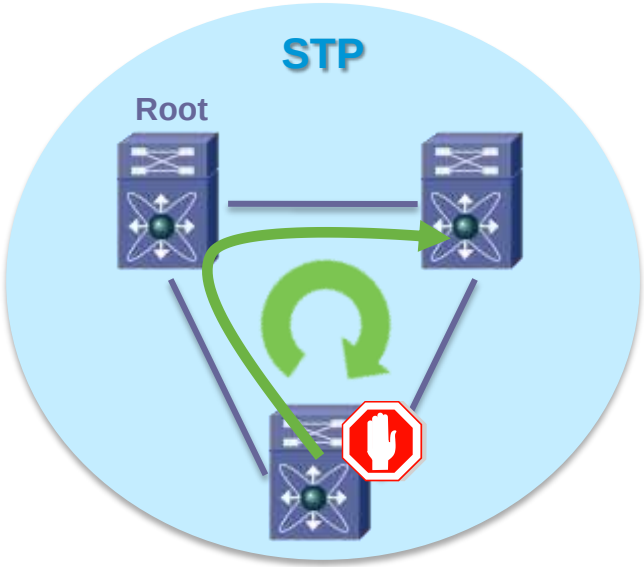
Sub-Switch ID

Identifies a VPC Off a Virtual Switch

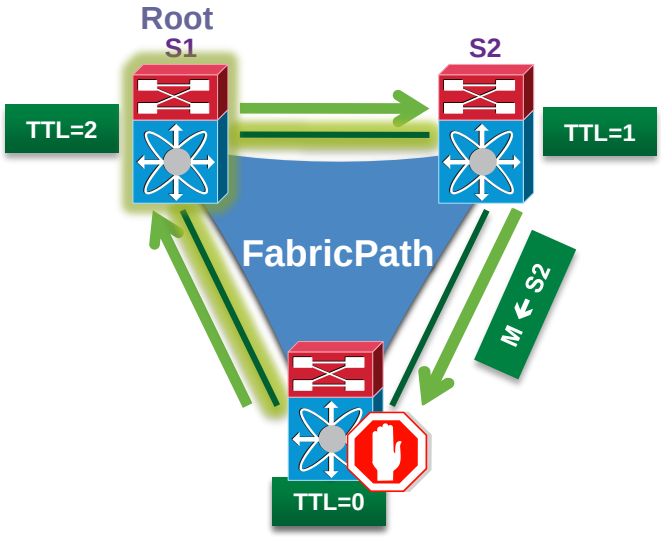


Loop Mitigation with FabricPath

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



- The control protocol is the only mechanism preventing loops
- If STP fails →infinite loop
  - No backup mechanism in the data plane
  - Flooding impacts the whole network

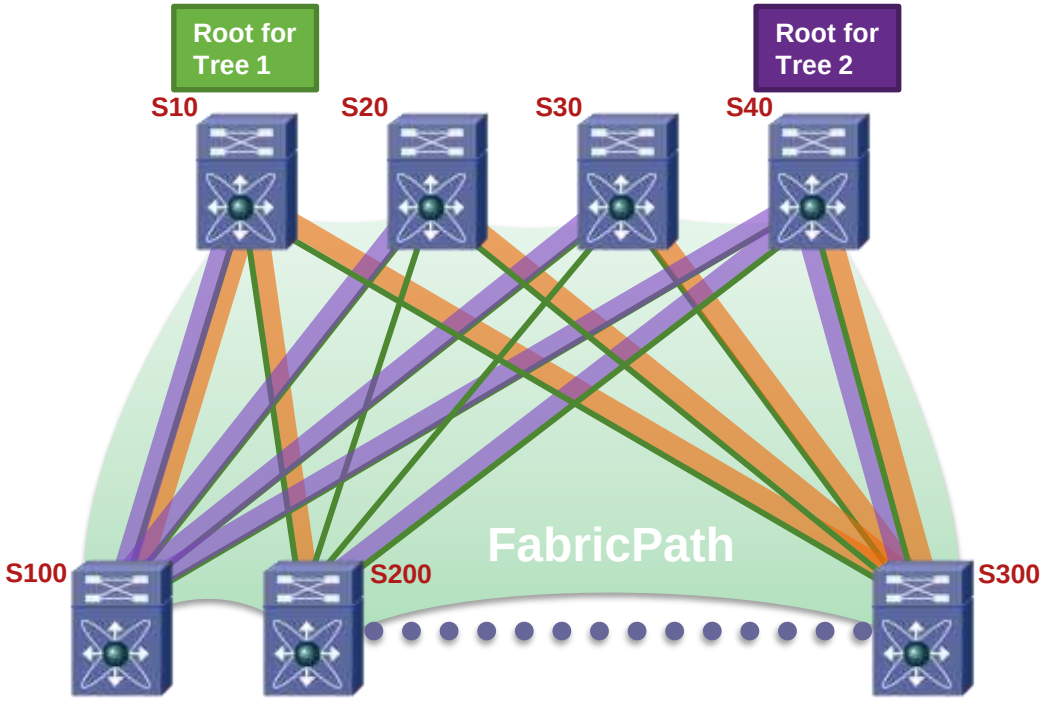


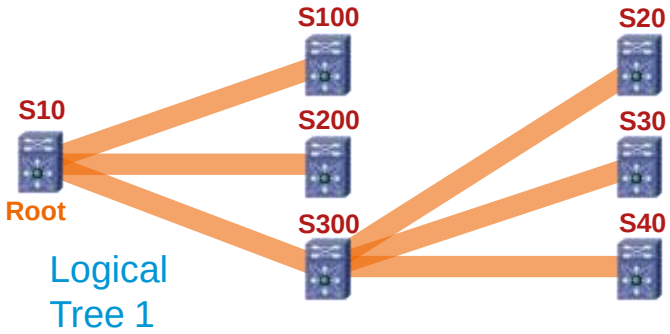
- TTL in FabricPath header
- RPF Check for multi-destination traffic
- The data plane is protecting against loops

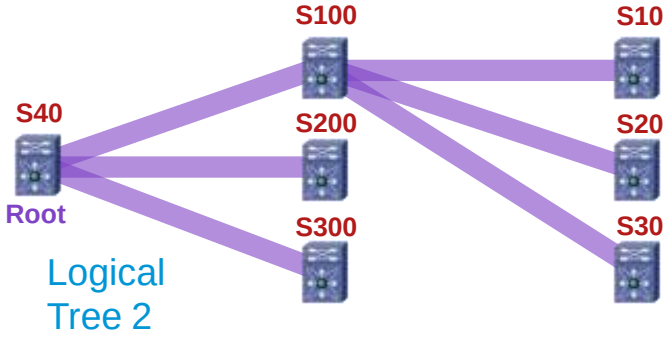
Agenda

- FabricPath Solution Overview
- FabricPath Multicast Details
- FabricPath Unicast Details
- FabricPath and TRILL
- FabricPath Design Examples
- FabricPath Case Study

Multicast Trees Determination

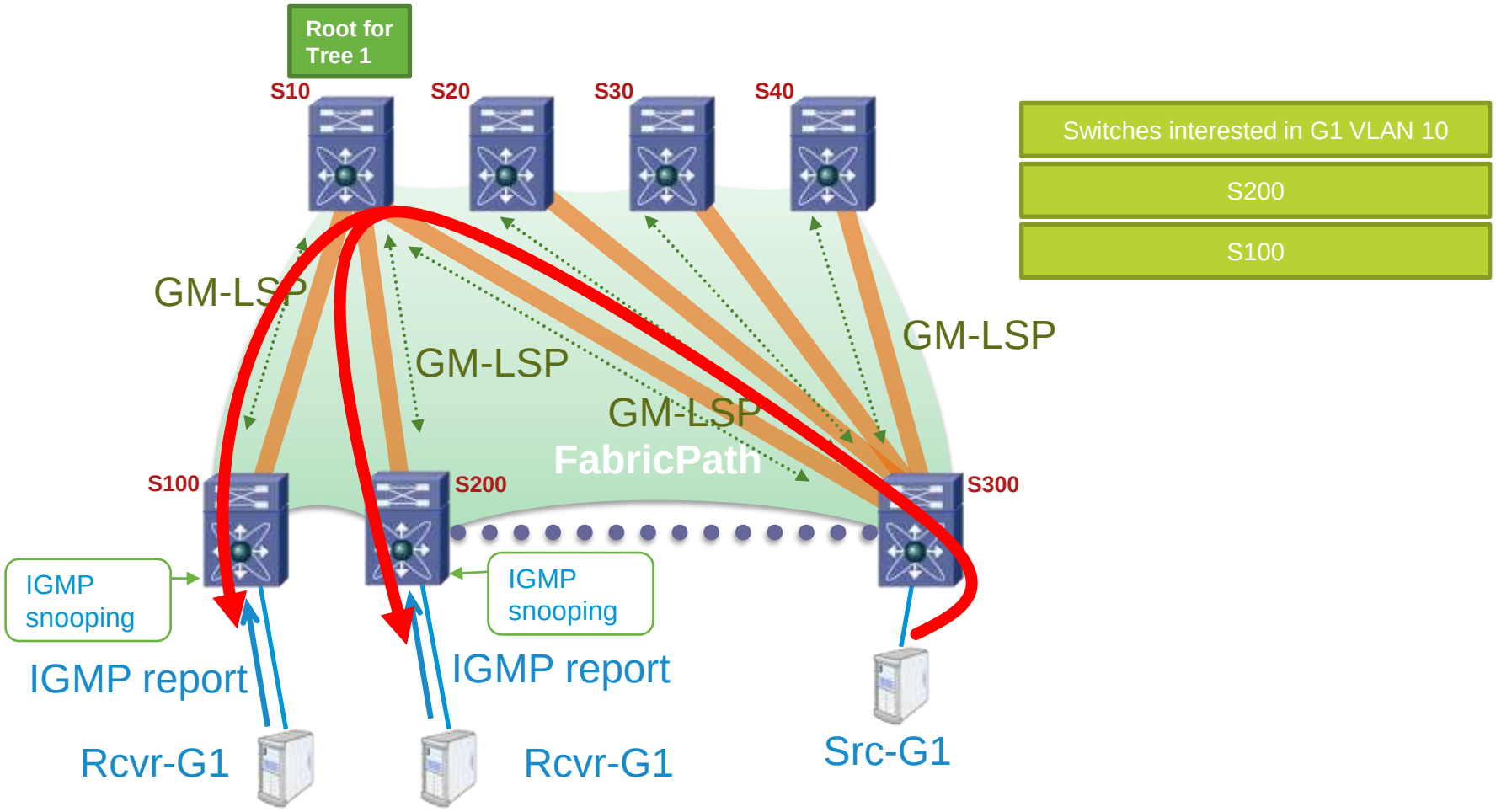






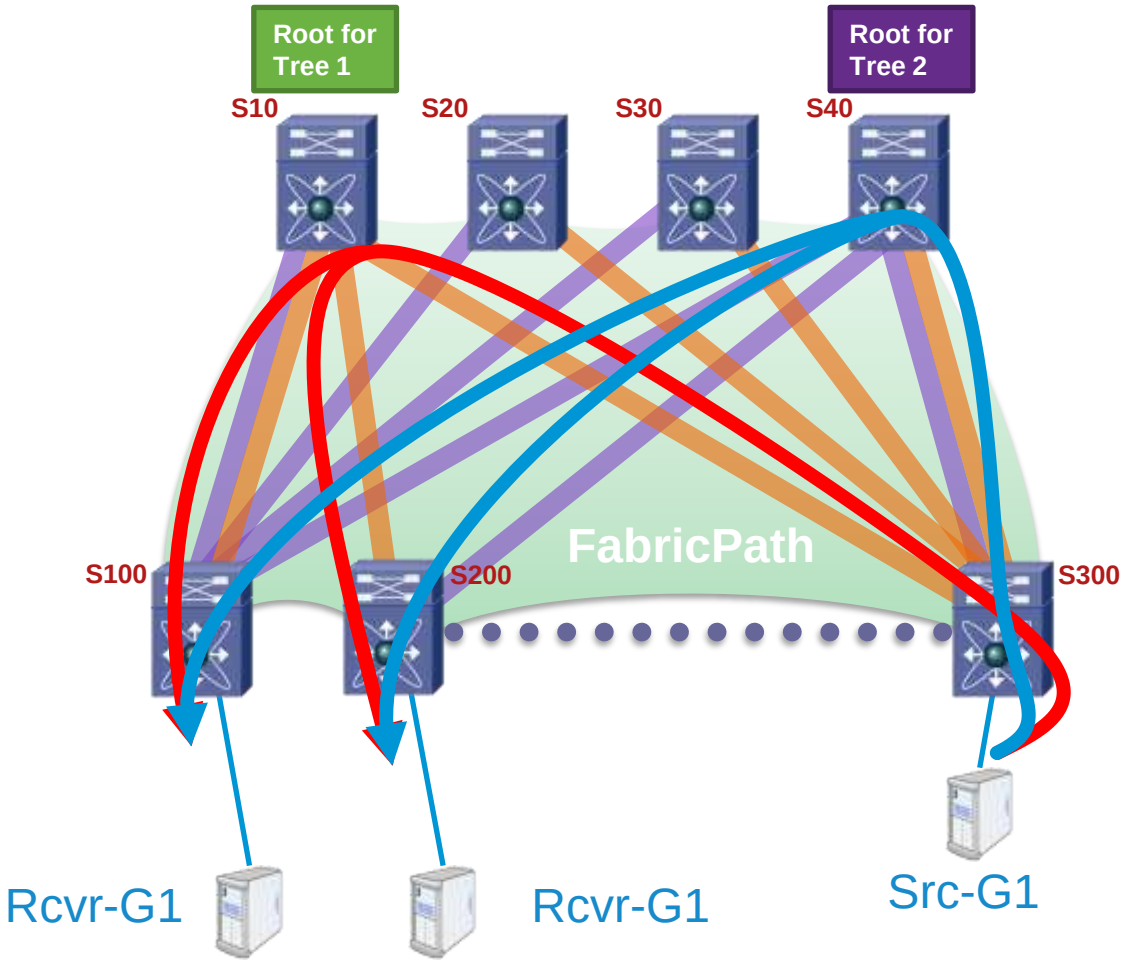
- Switch with highest priority value becomes root for primary tree
  Highest system ID, then highest Switch ID value, in case of a tie
- Primary root designates different secondary root(s) ensuring path variety.

Multicast Tree Pruning



- IS-IS Group Membership LSPs contain multicast forwarding information

Multicast Load Balancing

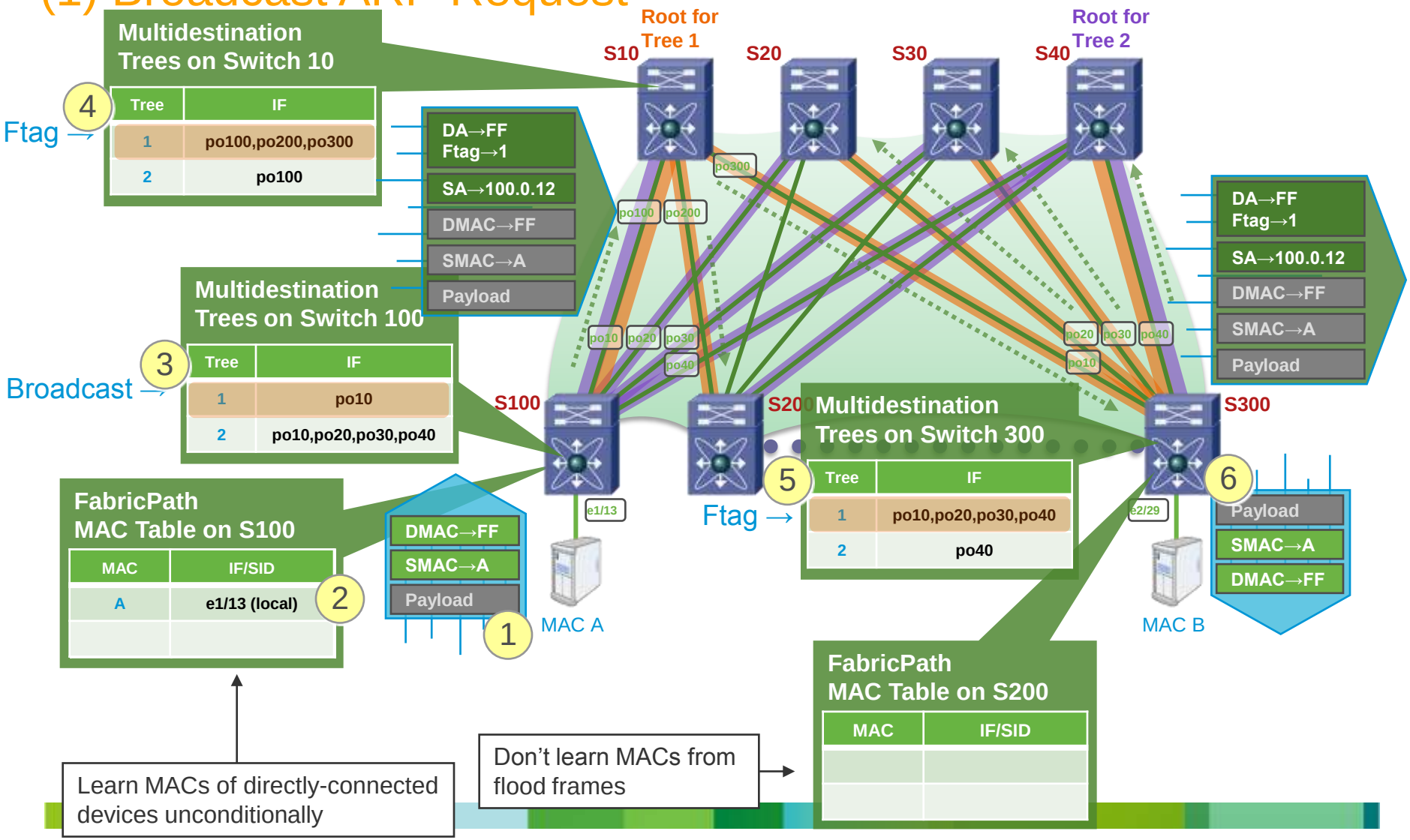


Agenda

- FabricPath Solution Overview
- FabricPath Multicast Details
- FabricPath Unicast Details
- FabricPath and TRILL
- FabricPath Design Examples
- FabricPath Case Study

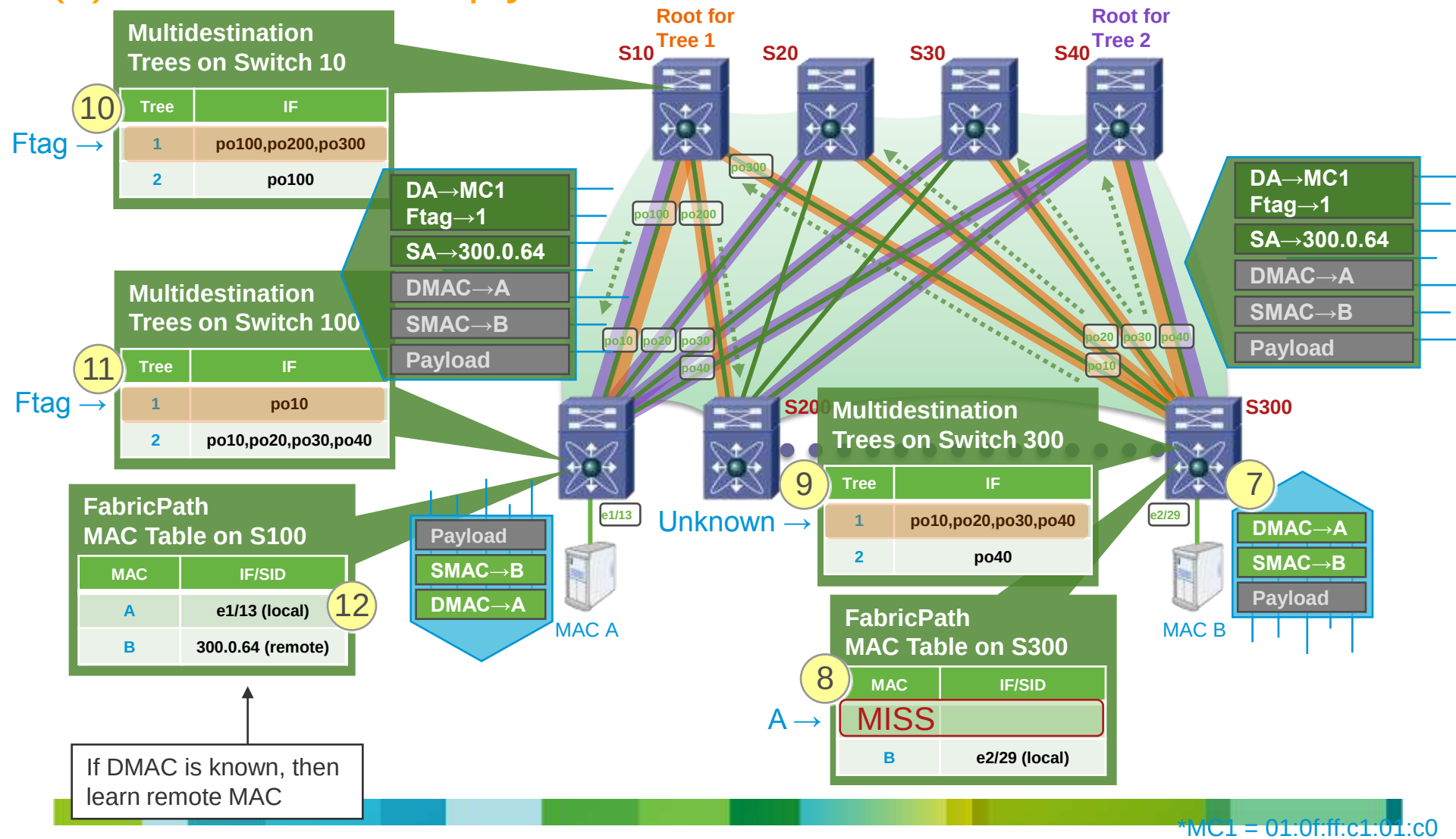
Putting It All Together – Host A to Host B

(1) Broadcast ARP Request



Putting It All Together – Host A to Host B

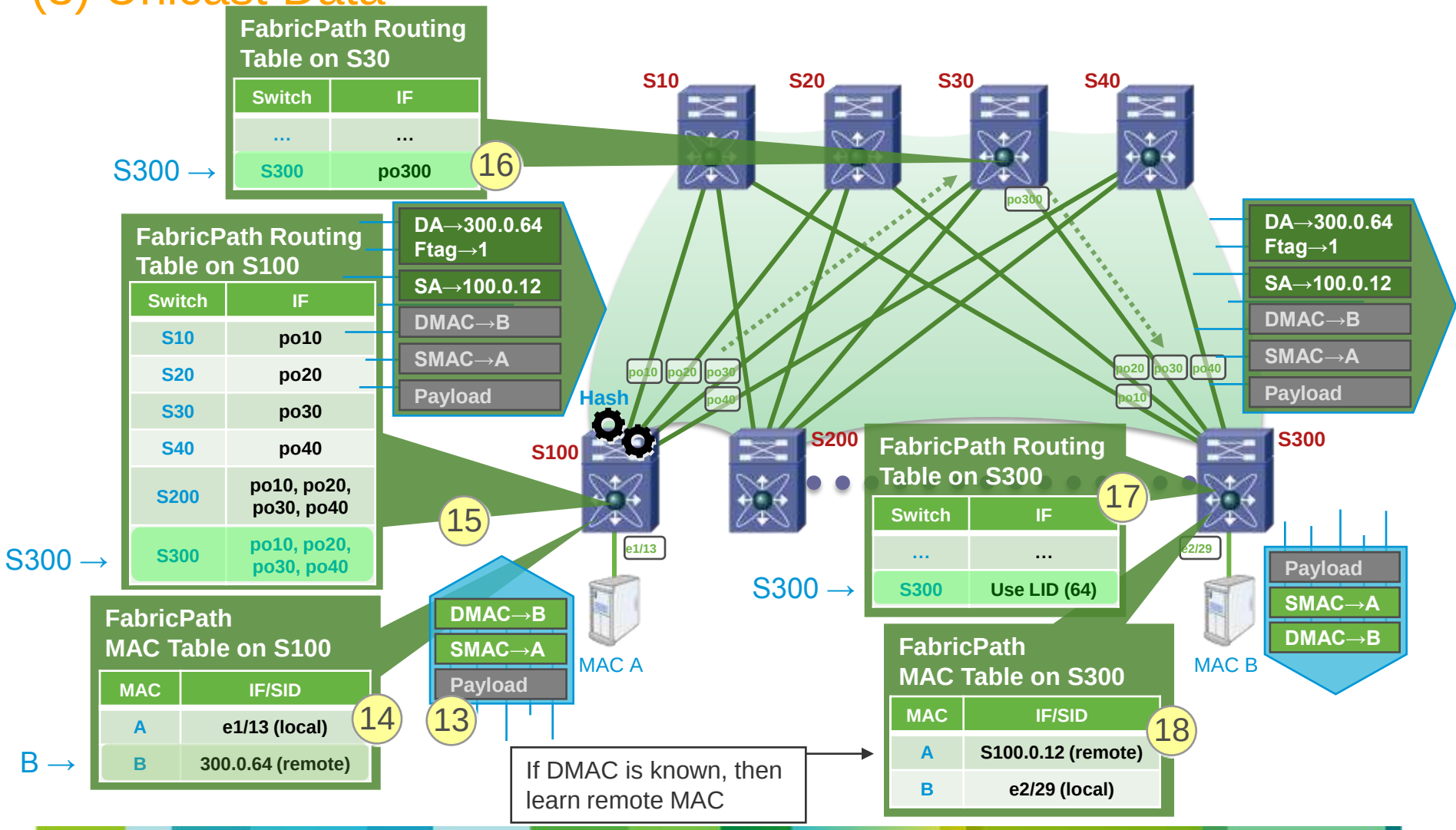
(2) Unicast ARP Reply



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

Putting It All Together – Host A to Host B

(3) Unicast Data



Agenda

- FabricPath Solution Overview
- FabricPath Multicast Details
- FabricPath Unicast Details
- FabricPath and TRILL
- FabricPath Design Examples
- FabricPath Case Study

TRILL
Transparent Interconnection of Lots of Links

- IETF standard for Layer 2 multipathing
- Driven by multiple vendors, including Cisco
- TRILL is now an IETF proposed standard
- FabricPath capable hardware is also TRILL capable

http://datatracker.ietf.org/wg/trill/

FabricPath vs. TRILL Overview

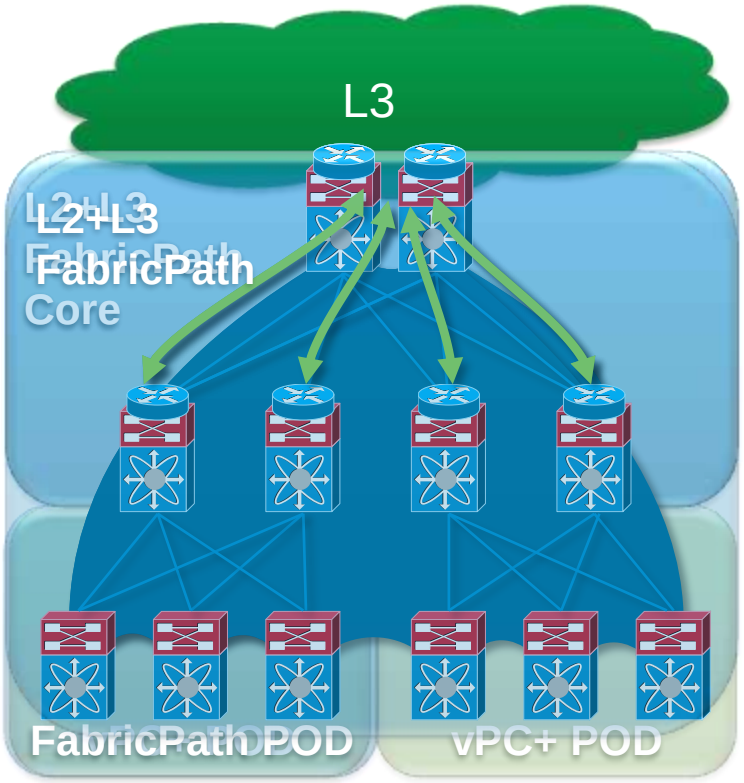
	FabricPath	TRILL
Frame routing (ECMP, TTL, RPFC etc…)	Yes	Yes
vPC+	Yes	No
FHRP active/active	Yes	No
Multiple topologies	Yes	No
Conversational learning	Yes	No
Inter-switch links	Point-to-point only	Point-to-point OR shared

- FabricPath will provide a TRILL mode with a software upgrade (hardware is already TRILL capable)
- Cisco will push FabricPath specific enhancements to TRILL

Agenda

- FabricPath Solution Overview
- FabricPath Multicast Details
- FabricPath Unicast Details
- FabricPath and TRILL
- FabricPath Design Examples
- FabricPath Case Study

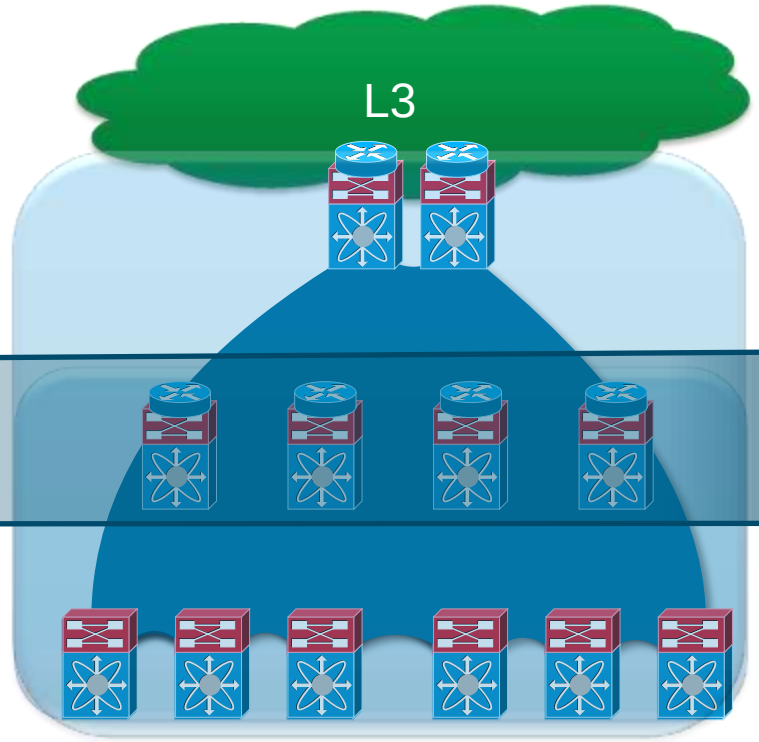
FabricPath Evolution



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

Tier Consolidation with FabricPath

"Flattening"



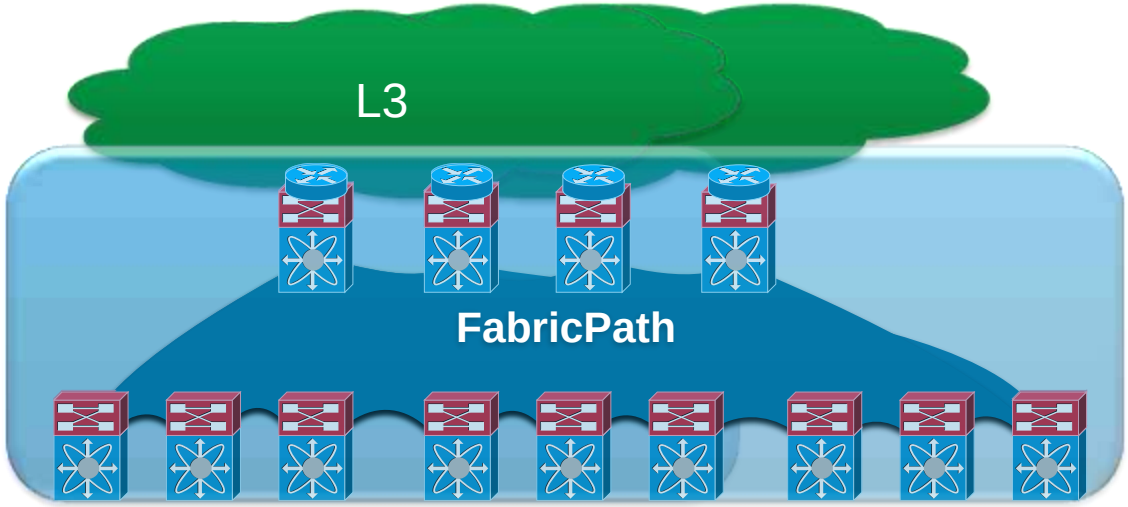
- FabricPath in the Core
- FabricPath extended down to the leaves

- There is enough bandwidth and port density on the core Nexus 7000s for aggregating the whole network.
- There is no need for a distribution layer for POD isolation

FabricPath Flexibility

The Network Can Evolve With No Disruption

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



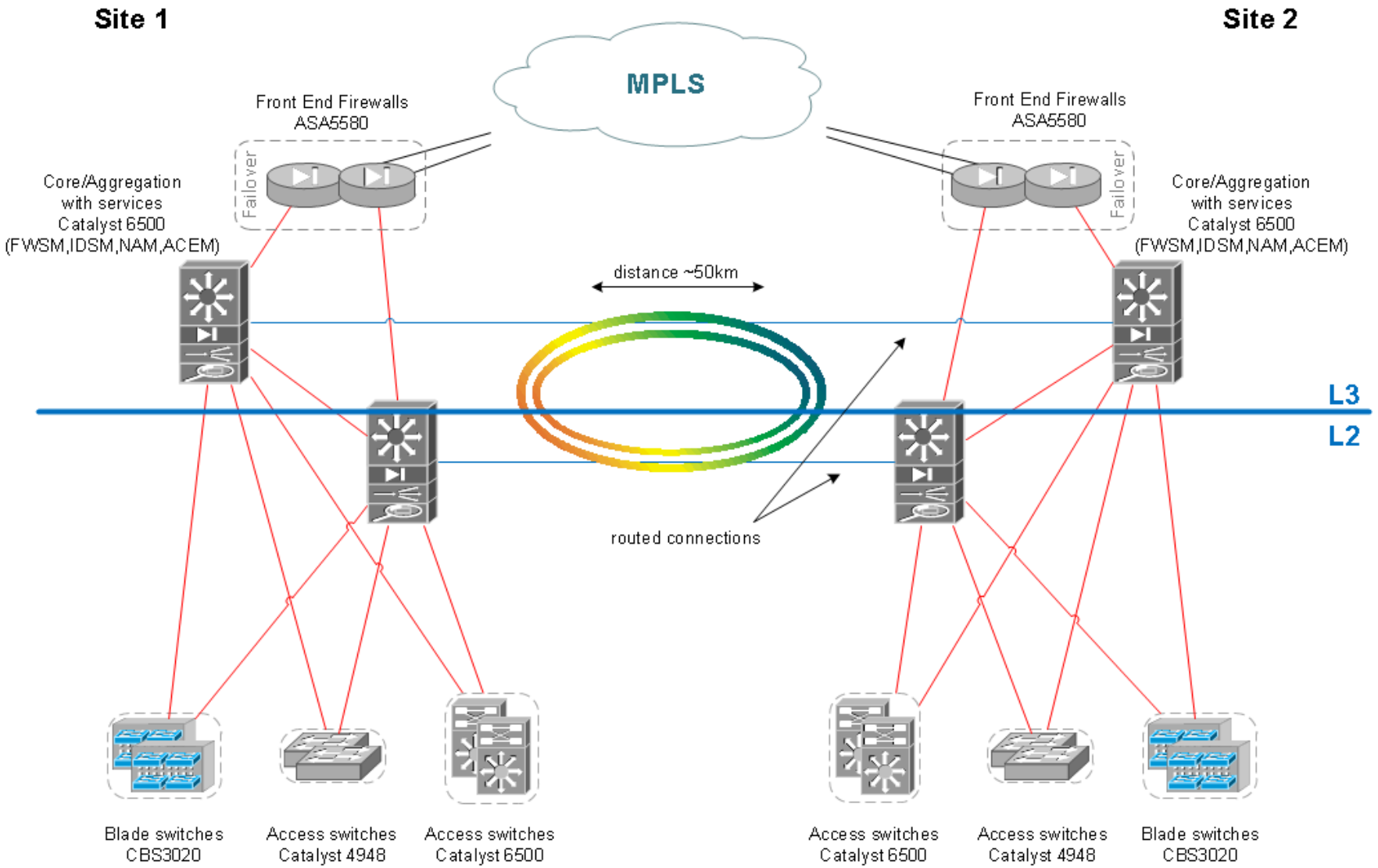
Agenda

- FabricPath Solution Overview
- FabricPath Multicast Details
- FabricPath Unicast Details
- FabricPath and TRILL
- FabricPath Design Examples
- FabricPath Case Study

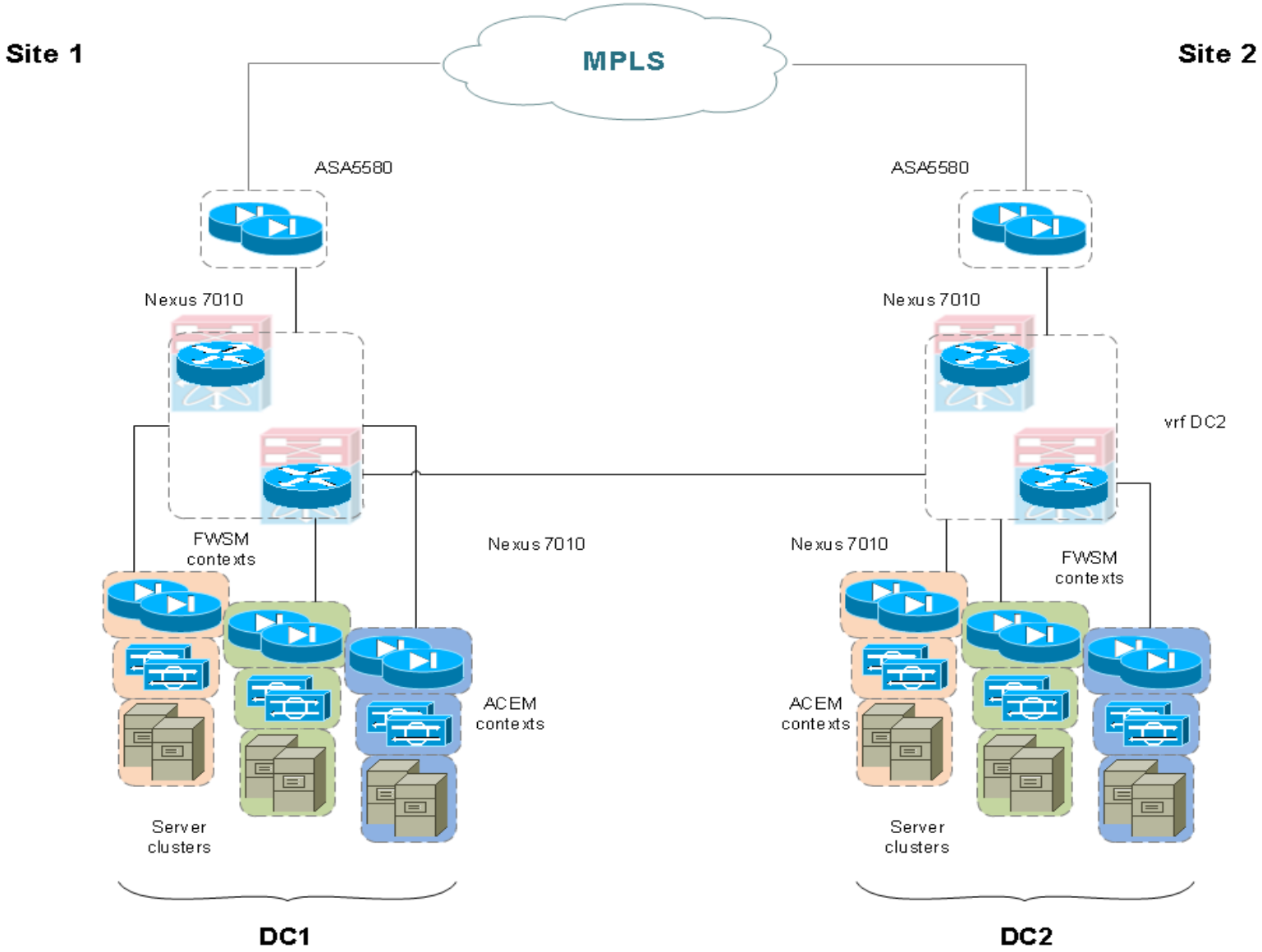
… before

- 2 sites – DC1, DC2
  - distance ~50km
  - active / backup
  - only L3 connection (STP issues in the past)
- each site:
  - collapsed core architecture
    - core/aggregation layer - Cisco Catalyst 6500 with service modules
      - L3 routing between FWSM contexts, sites, to/from ext. networks
    - access layer – Cisco Catalyst 3020 (Blade) / 4948 / 6500
  - service modules
    - virtualized FWSM (front, mid, backend) – def GW for servers
    - virtualized ACE module (front, mid, backend) – SLB
    - NAM module – monitoring, troubleshooting
    - IDSM – security monitoring
  - frontend FW – Cisco ASA 5580

... before – physical topology



... before – logical topology



New requirements

- Support for new L2 clusters spread out over 2 sites
  - Company decision to migrate from Active/Backup to Active/Active DCs architecture
  - Reliable L2 connection between sites required
- New network architecture must fully support:
  - Existing applications running at Active/Backup DCs architecture
  - New applications implementation into new Active/Active DCs architecture
  - Seamless migration from old to new DCs architecture
- Scalability, reliability and performance improvement
  - Due to DCs consolidation new network architecture must support future growth
- Multi-tenancy support – end-to-end virtualization
- Limited budget

Proposed solution

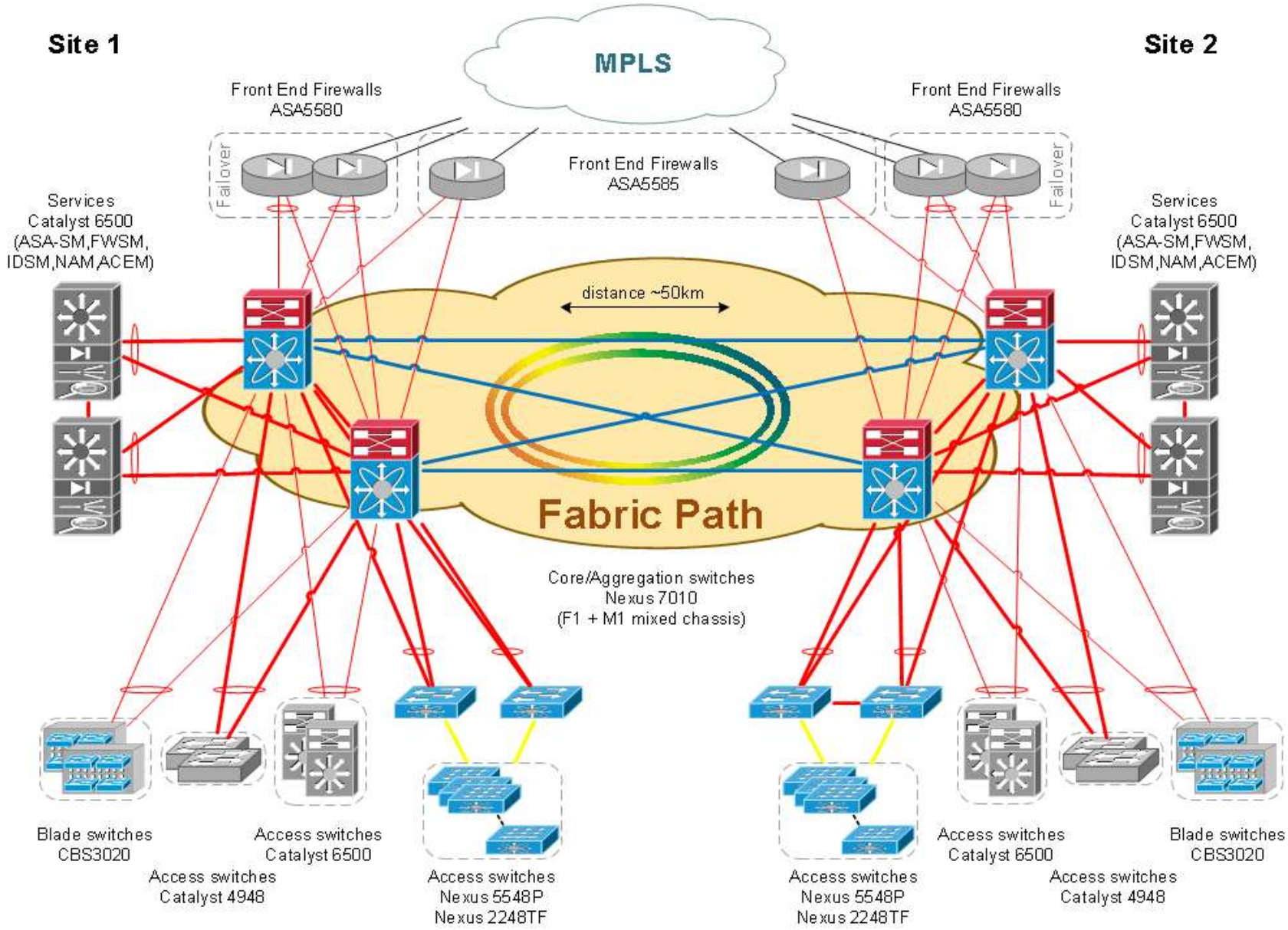
- Maximum reuse of existing HW
- Maintain existing collapsed core architecture within each sites
  - Core/Aggregation layer – Nexus 7010 (F1 + M1 mixed chassis)
  - Service layer - existing Catalyst 6500 reuse
    - new 10G line cards, only fabric enabled modules
  - Access layer
    - Nexus 5548P + Nexus 2248TF extenders
    - Existing Catalyst 4948, 6500 switches, Blade switches
- Leveraging existing VLANs, IP addressing scheme
- **FabricPath** in the core/aggregation and for DCI
  - vPC+ for access, service layer, FWs, ... connectivity
  - No FabricPath in Access switches Nexus 5500 (possible with today NX-OS releases)

Proposed solution

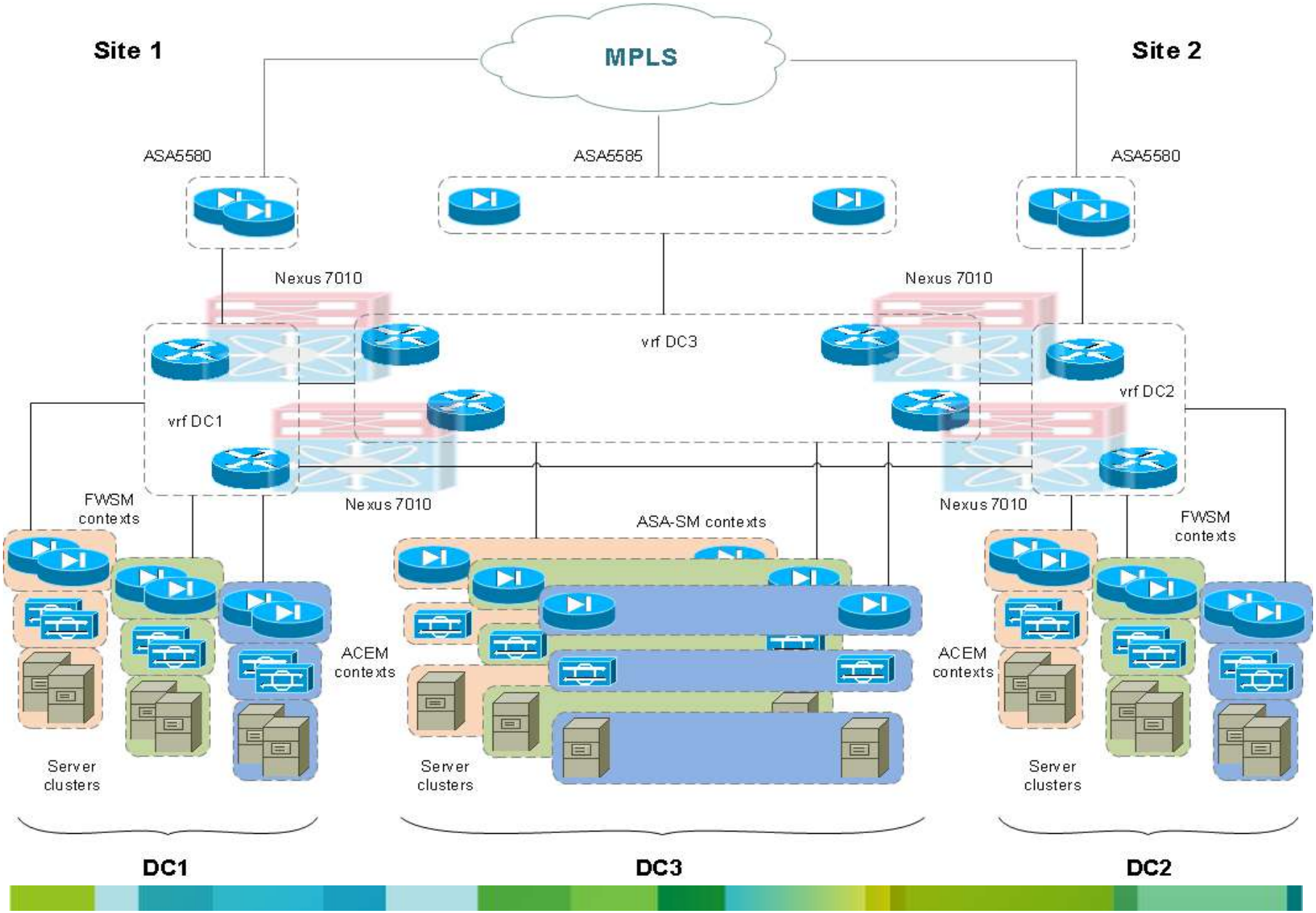
- Introduction of new logical DC3 for supporting new Active/Active architecture
  - Same physical Core/Aggregation, Services and Access switches
  - Same logical topology for DC3 as in DC1 and DC2
    - Failover pairs of services devices (FWs, SLBs) spread out over 2 sites
  - VRF based virtualization (no VDCs) at L3, Fabric Path at L2
    - PBR (Policy-based routing) for Inter-VRF routing

- Nexus 1000V + VSG within VMware environment for end-to-end virtualization support (separate project)

... after – physical topology



... after – logical topology



Lessons learned

- FabricPath
  - Multicast/Broadcast/Unknown unicast
    - Root for Multidestination Tree requires all VLANs
    - Initially we don't have DC1 VLANs on DC2 and vice versa
    - Multiple topology feature in 6.2.x will provide solution
  - STP interaction
    - FabricPath switches must be STP root
      - same lowest priority on all FabricPath L2 GW
      - STP Root-Guard like functionality on CE ports within FabricPath
    - optional TCN propagation through FabricPath (not used in our case)
  - Multicast – enable fabricpath multicast load-balance (best practice)
    - Root for Multidestination Tree #1 on vPC+ Primary switch,
      Root for Multidestination Tree #2 on secondary (best practice)

Lessons learned

vPC+ default behavior and enhancements

- New virtual switch within FabricPath topology (effective load/balancing)
- Different loop/avoidance mechanism than vPC (no more VSL check)
  - vPC+ peer-links are FabricPath core ports -> TTL check
- No vPC and vPC+ in the same system/VDC
- All VLANs in FabricPath mode
- Active/Active HSRP
  - HSRP Active on vPC+ Primary switch (best practice)
  - Peer/link failure (dual-active exclude vlan <VLAN list >)
- vPC+ orphan-port suspend
  - Active/Standby failover in case of peer-link failure
- Raise the FabricPath IS-IS metric for vPC+ peer-link

- Pre-migration in-depth lab testing !
- NRFU (Network Ready for Use) testing

Otázky a odpovědi

- Twitter www.twitter.com/CiscoCZ
- Talk2Cisco www.talk2cisco.cz/dotazy
- SMS 721 994 600

- Zveme Vás na **Ptali jste se…** v sále **LEO**
  1.den 17:45 – 18:30
  2.den 16:30 – 17:00

Otázky a odpovědi

- Twitter www.twitter.com/CiscoCZ
- Talk2Cisco www.talk2cisco.cz/dotazy
- SMS 721 994 600

- Zveme Vás na **Ptali jste se…** v sále **LEO**
  1.den 17:45 – 18:30
  2.den 16:30 – 17:00

Prosíme, ohodnoťte tuto přednášku.