Cisco FabricPath Technology and Design

BRKDCT-2081

July 10 - July 14, 2011 · Las Vegas, NV

Session Goal

- To provide you with a conceptual and technical understanding of Cisco FabricPath – its operation, its relationship to industry standards, and network design options
- Related sessions:
  - TECDCT-4125: Cisco FabricPath Techtorial (PARTLY REDUNDANT TO THIS SESSION)
  - BRKARC-3470: Cisco Nexus 7000 Hardware Architecture
  - BRKARC-3471: Cisco NX-OS Software Architecture
  - BRKDCT-2951: Deploying Nexus 7000 in Data Center Networks
  - BRKDCT-2048: Deploying Virtual Port Channel in NX-OS
  - BRKDCT-2121: VDC Design and Implementation Considerations with Nexus 7000
  - BRKARC-3472: NX-OS Routing & Layer 3 Switching
  - BRKCRS-3144: Troubleshooting Cisco Nexus 7000 Series Switches
  - LTRDCT-4047: Deploying Nexus 7000/NX-OS Hands-on Lab (Lab)
  - LTRCRT-5205: Configuring Nexus 7000 Virtualization (Lab)

Agenda

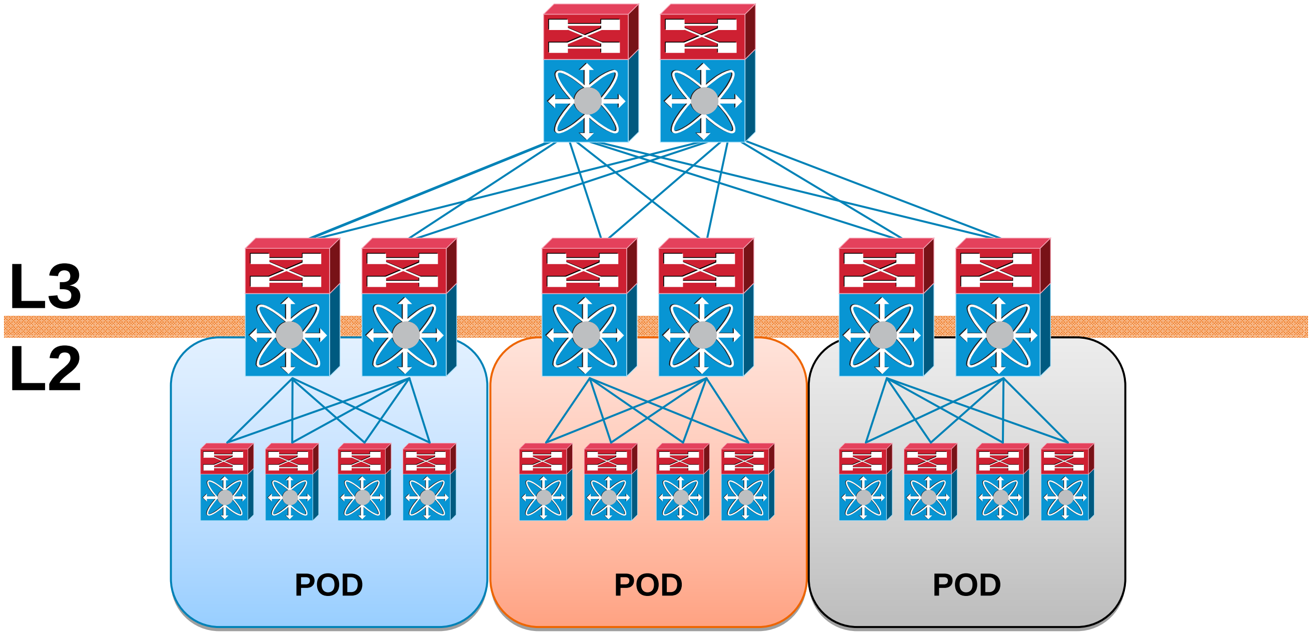
- Introduction to FabricPath
- FabricPath Technical Overview
- Unicast Forwarding
- Multicast Forwarding
- FabricPath Design
- Conclusion

Introduction to FabricPath

Why Layer 2 in the Data Center?

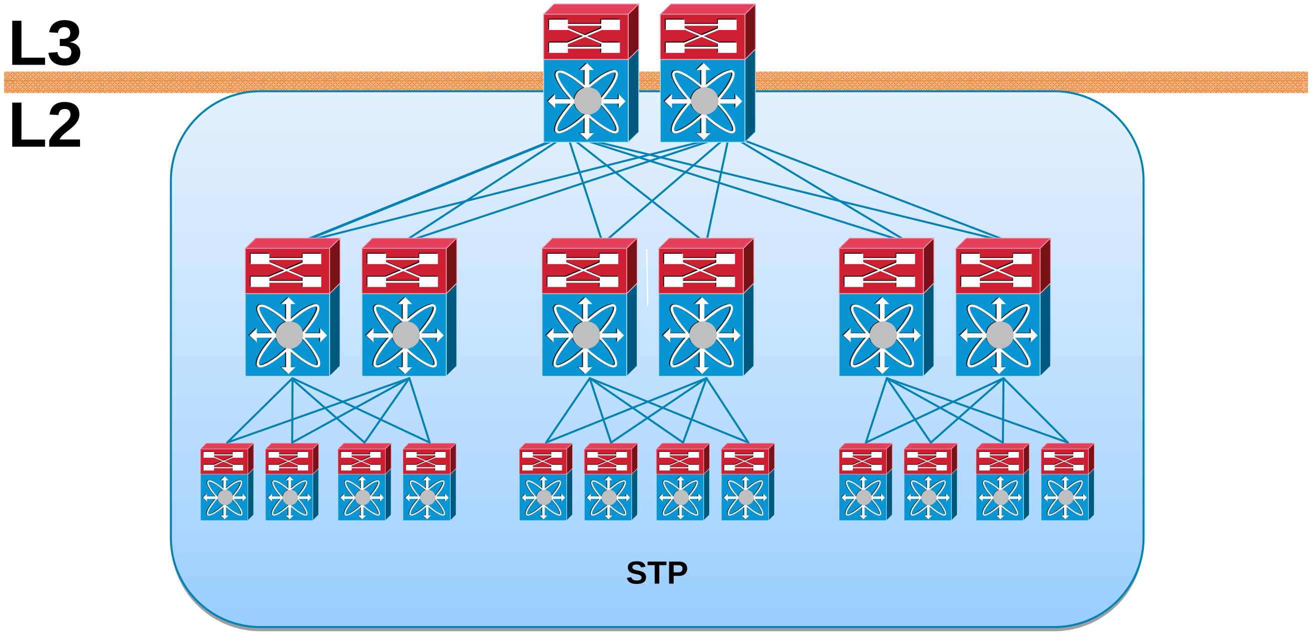
- Because customers request it!
  - Some protocols rely on the functionality
  - Simple, almost plug and play
  - No addressing
  - Required for implementing subnets
  - Allows easy server provisioning
  - Allows virtual machine mobility

Current Data Center Design



- L2 benefits limited to a POD

Possible Solution for End-to-End L2?



- Just extend STP to the whole network

Typical Limitations of L2

- Local STP problems have network-wide impact, troubleshooting is difficult
- STP provides limited bandwidth (no load balancing)
- STP convergence is disruptive
- Tree topologies introduce sub-optimal paths
- MAC address tables don't scale
- Flooding impacts the whole network

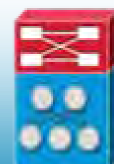
Cisco FabricPath Goal

Switching

- Easy Configuration
- Plug & Play
- Provisioning Flexibility

Routing

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

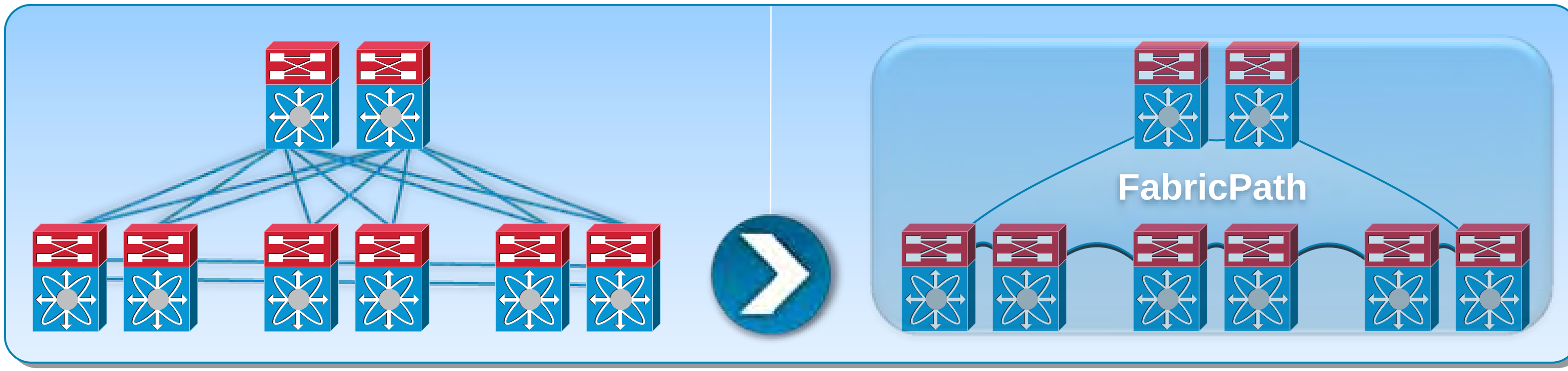


FabricPath

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

FabricPath: An Ethernet Fabric

Turn the network into a Fabric



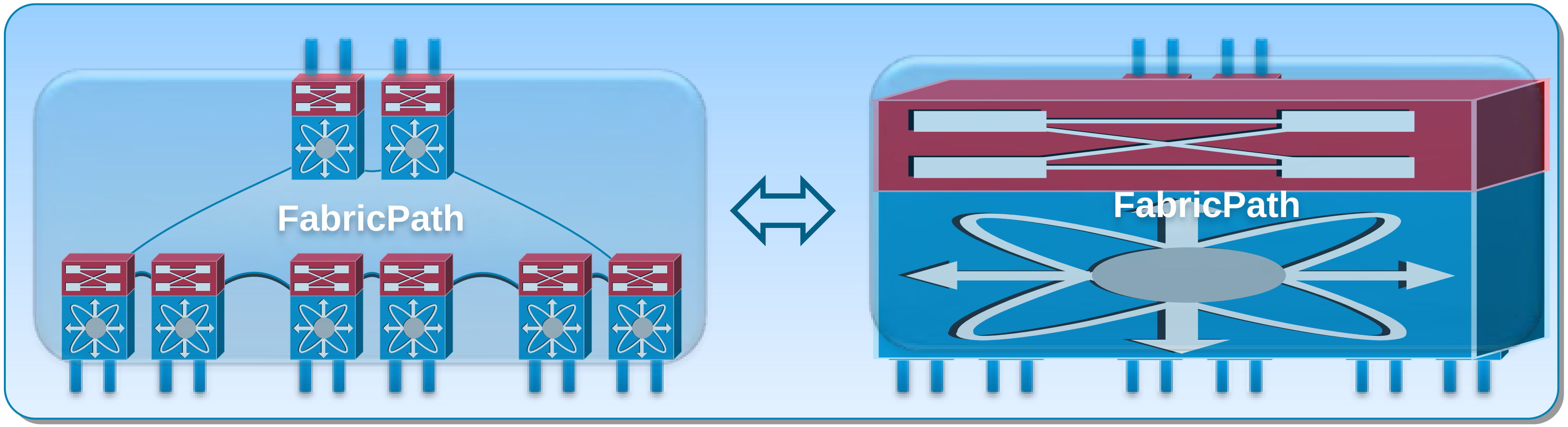
- Connect a group of switches using an **arbitrary** topology
- With a simple CLI, aggregate them into a Fabric:

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

- No STP inside. An open protocol based on L3 technology provides Fabric-wide intelligence and ties the elements together.

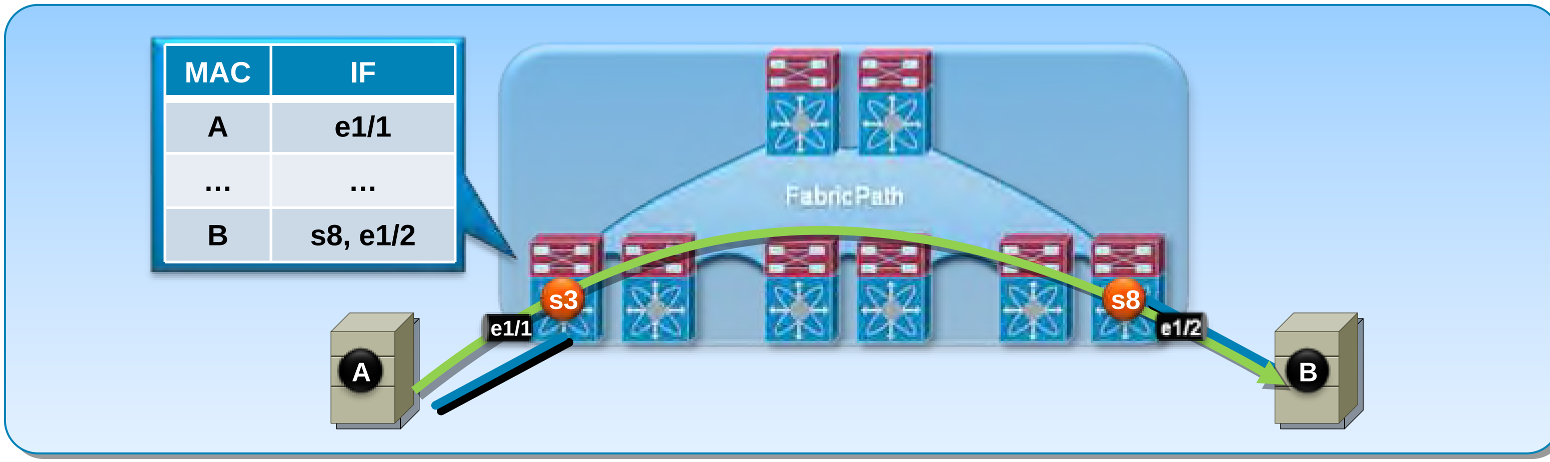
What Is a Fabric?

- Externally, a Fabric looks like a single switch
- Internally, a protocol adds Fabric-wide intelligence and ties the elements together.
  This protocol provides in a plug-and-play fashion:
  - Optimal, low latency connectivity any to any
  - High bandwidth, high resiliency
  - Open management and troubleshooting
- Cisco FabricPath provides additional capabilities in term of scalability and L3 integration



Optimal, Low Latency Switching

Shortest path any-to-any

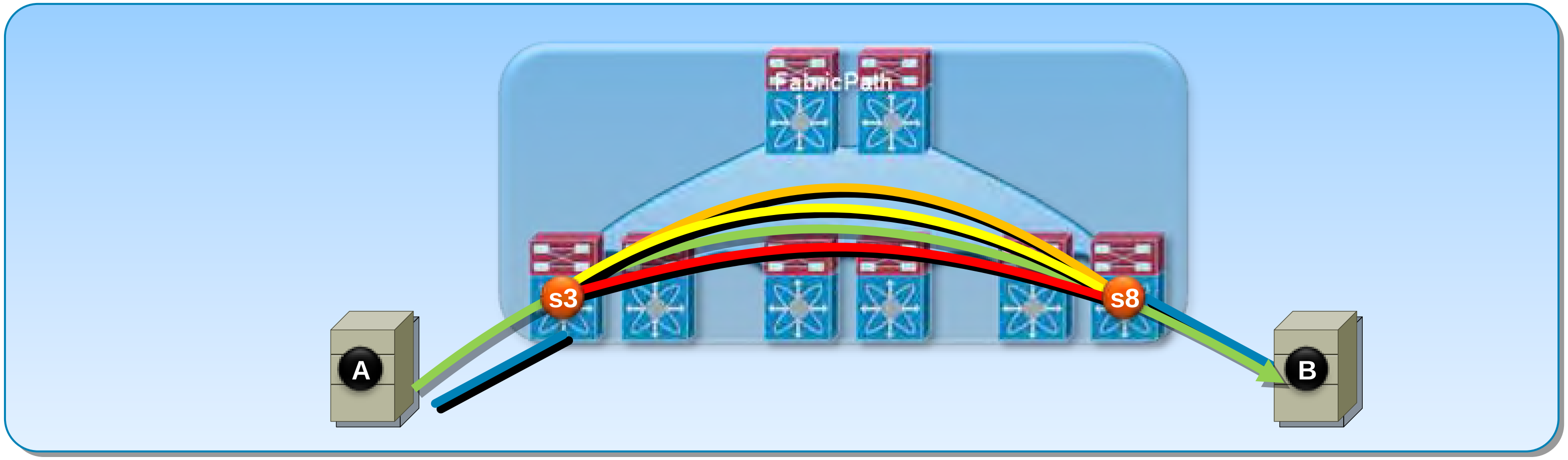


MAC	IF
A	e1/1
…	…
B	s8, e1/2

- Single address lookup at the ingress edge identifies the exit port across the fabric
- Traffic is then switched using the shortest path available
- Reliable L2 and L3 connectivity any to any
  (L2 as if it was within the same switch, **no STP inside**)

High Bandwidth, High Resiliency

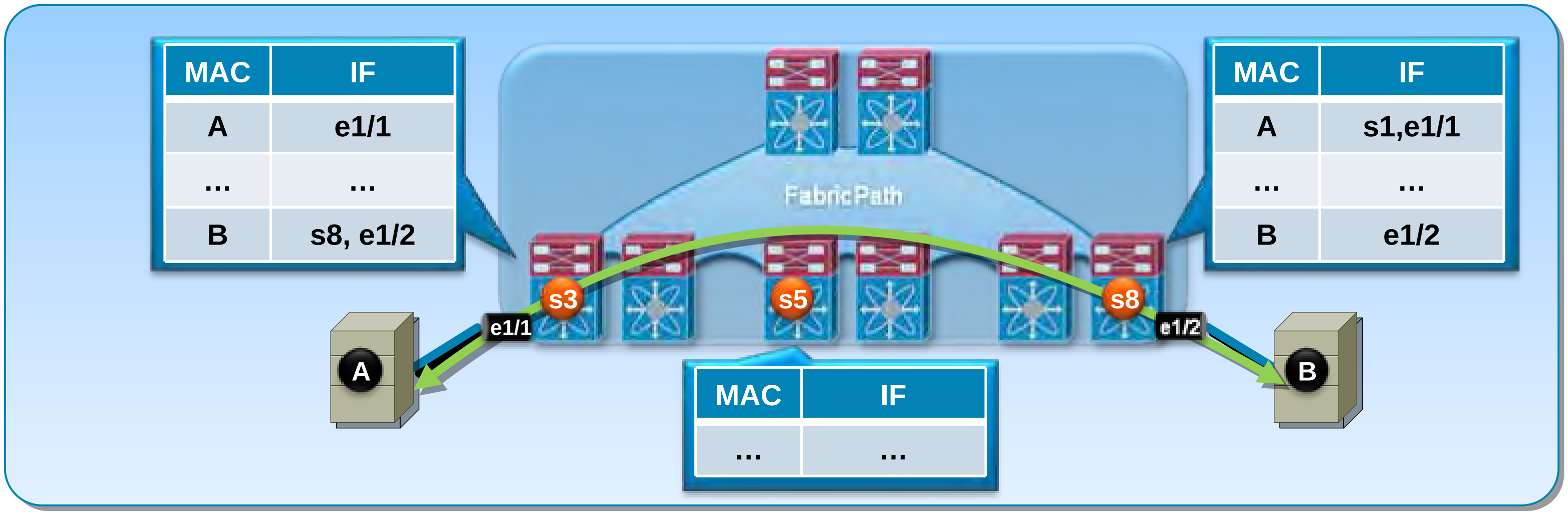
Equal Cost Multi-Pathing



- Multi-pathing (up to 256 links active between any 2 devices)
- Traffic is redistributed across remaining links in case of failure, providing fast convergence

Scalable

Conversational Learning

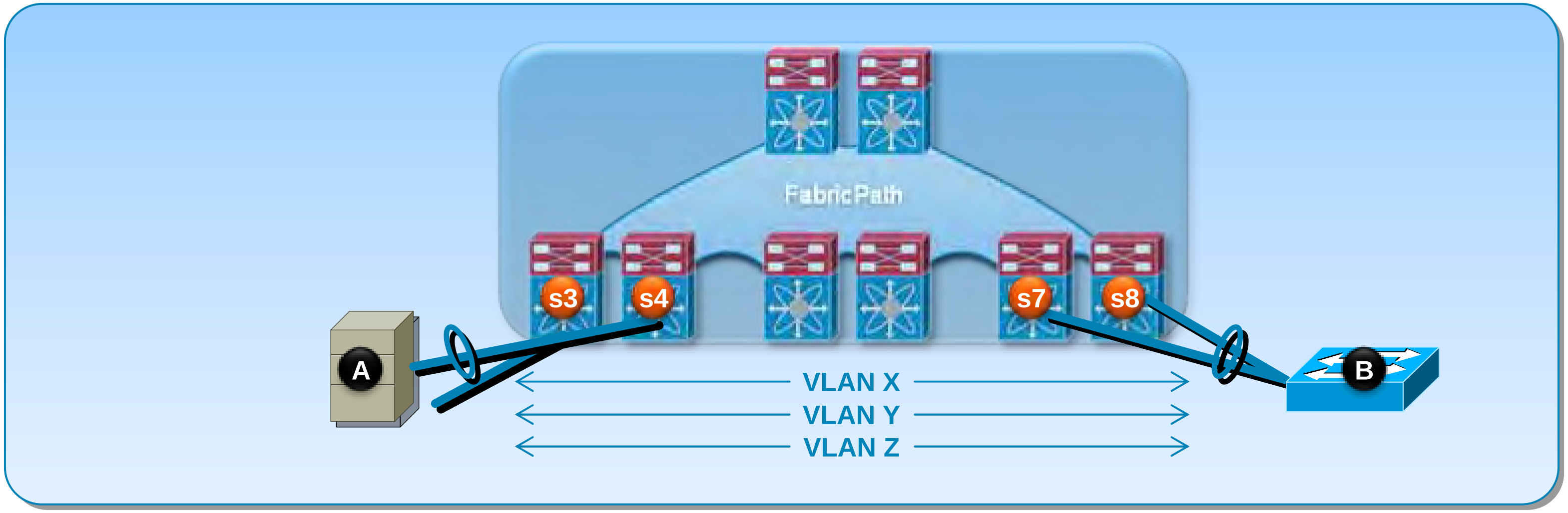


- Per-port MAC address table only needs to learn the peers that are reached across the fabric

A virtually unlimited number of hosts can be attached to the fabric

Layer 2 integration

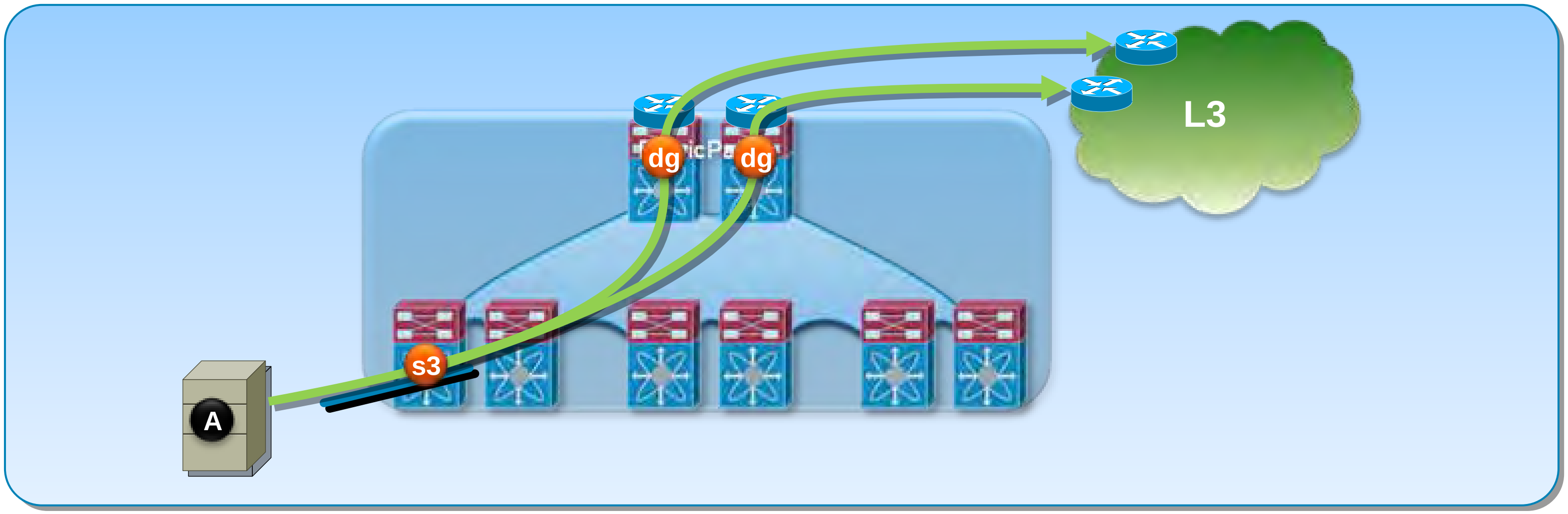
VPC+



- Allows extending VLANs with no limitation (no risks of loop)
- Devices can be attached active/active to the fabric using IEEE standard port channels and without resorting to STP

Edge Device Integration

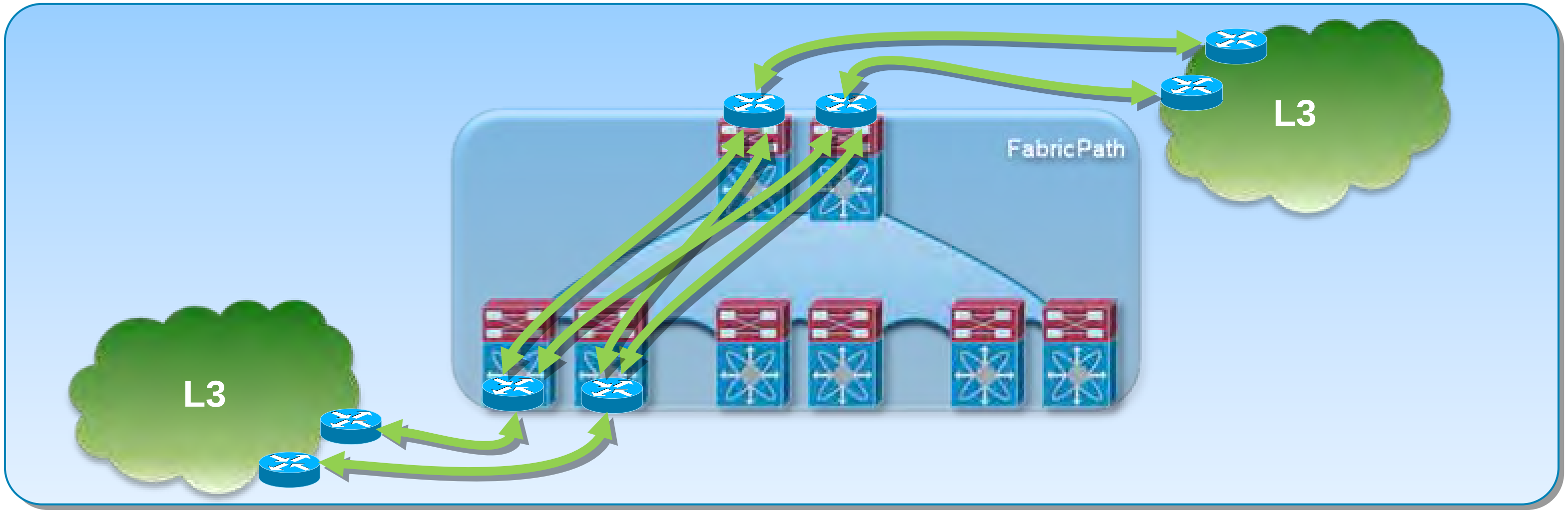
Hosts can leverage multiple L3 default gateways



- Hosts see a single default gateway
- The fabric provide them transparently with multiple simultaneously active default gateways dg
- Allows extending the multipathing from the inside of the fabric to the L3 domain outside the fabric

Layer 3 Integration

XL tables, SVIs anywhere



- The fabric provides seamless L3 integration
- An arbitrary number of routed interfaces can be created at the edge or within the fabric
- Attached L3 devices can peer with those interfaces
- The hardware is capable of handling million of routes

Agenda

- Introduction to FabricPath
- FabricPath Technical Overview
- Unicast Forwarding
- Multicast Forwarding
- FabricPath Design
- Conclusion

FabricPath Technical Overview

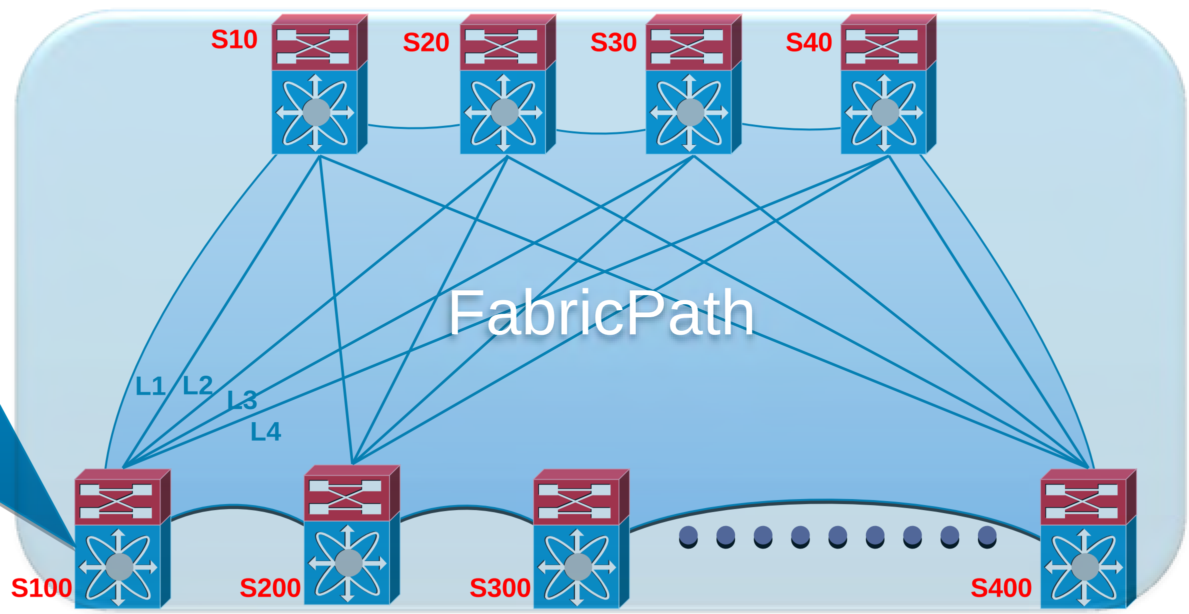
New Control Plane

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

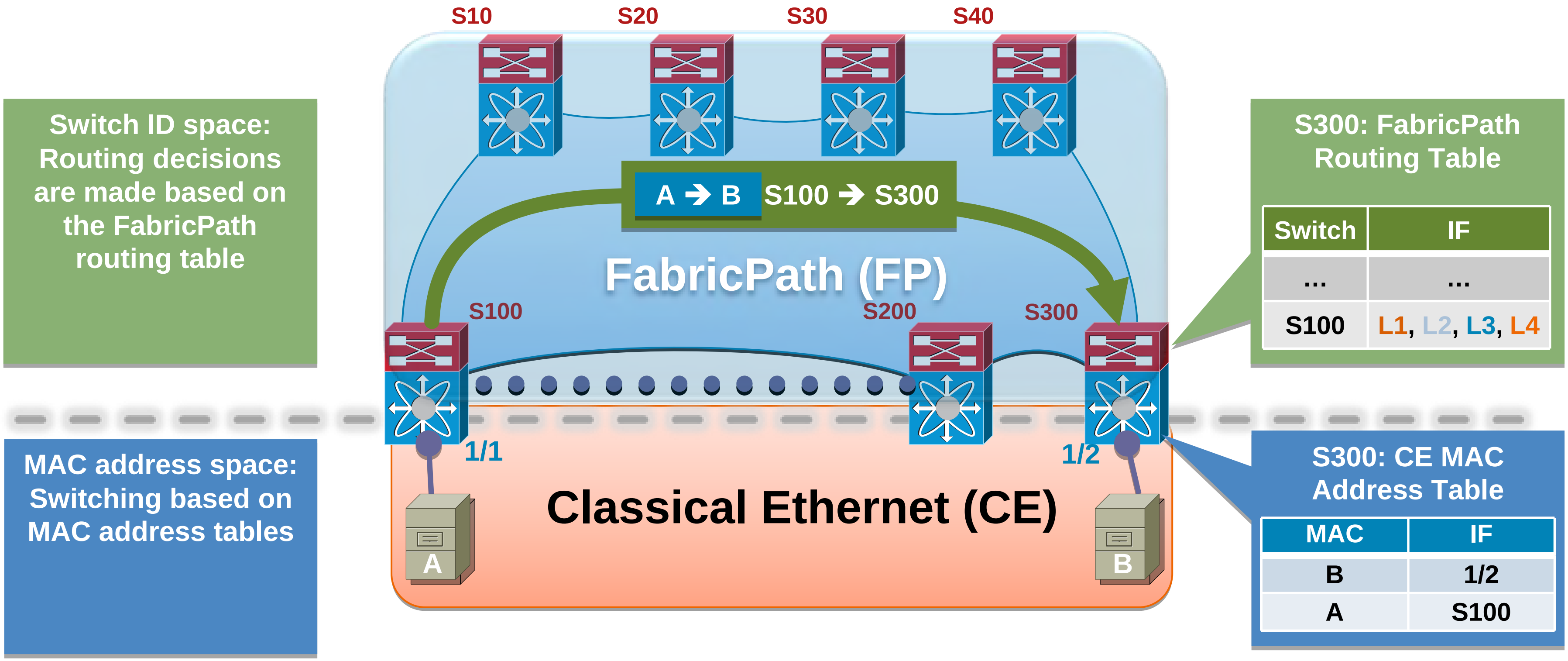
- IS-IS assigns addresses to all FabricPath switches automatically
- Compute shortest, pair-wise paths
- Support equal-cost paths between any FabricPath switch pairs

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

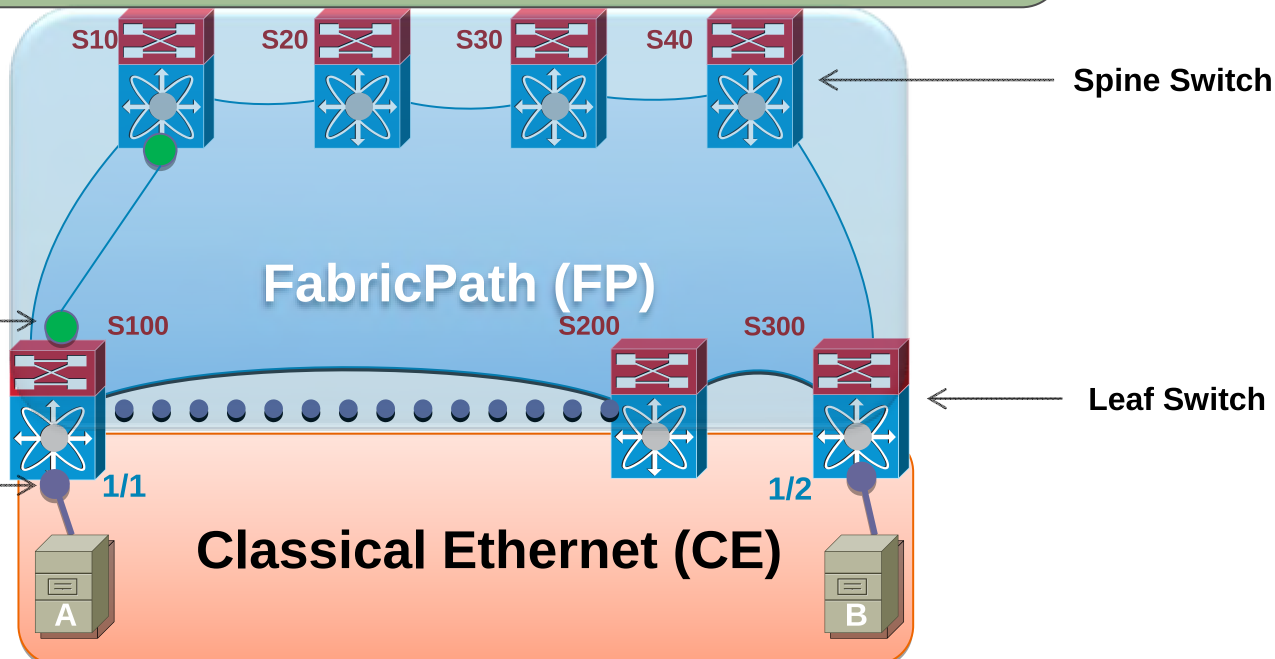


- The association MAC address/Switch ID is maintained at the edge
- Traffic is encapsulated across the Fabric

FabricPath Terminology

- 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’

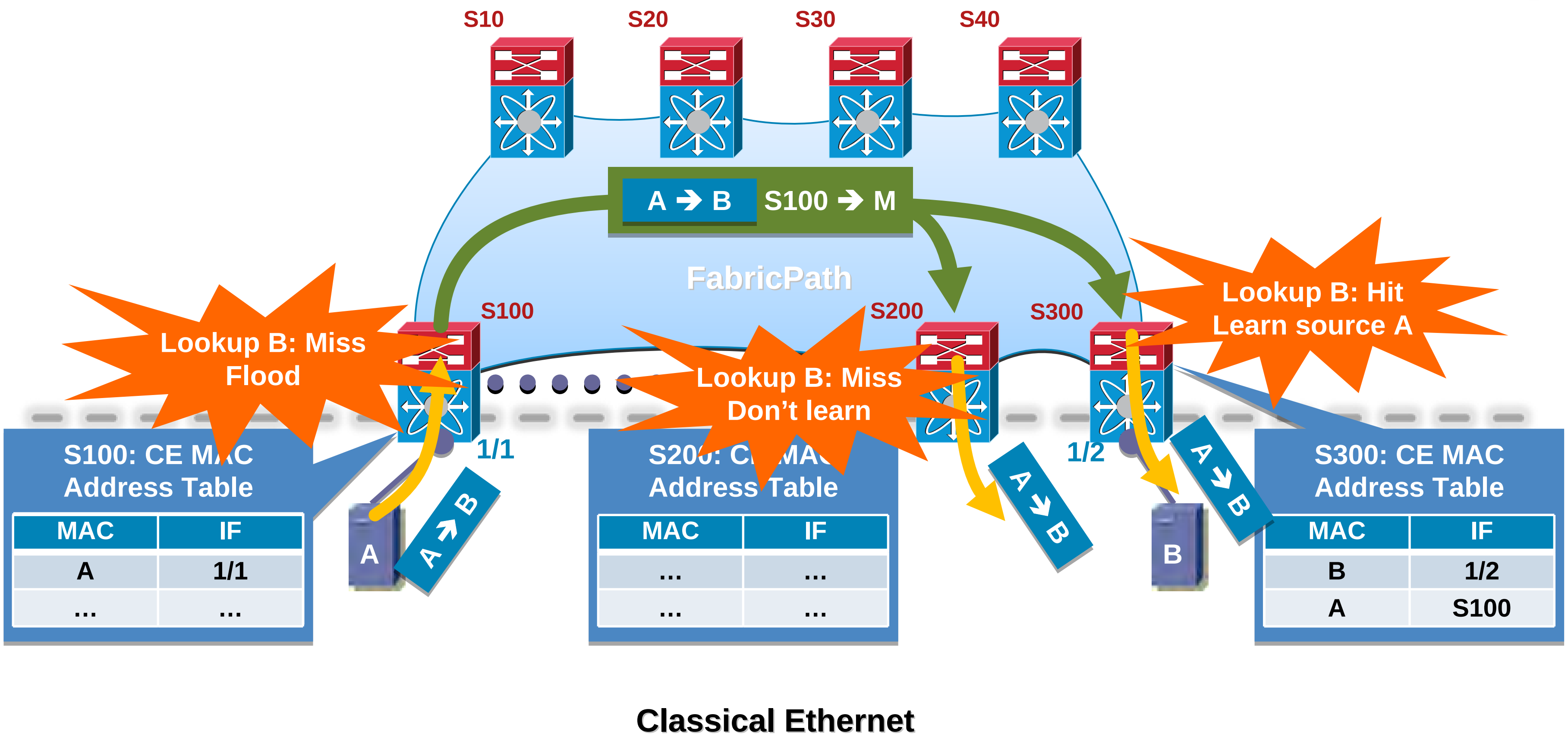
FP Core Ports



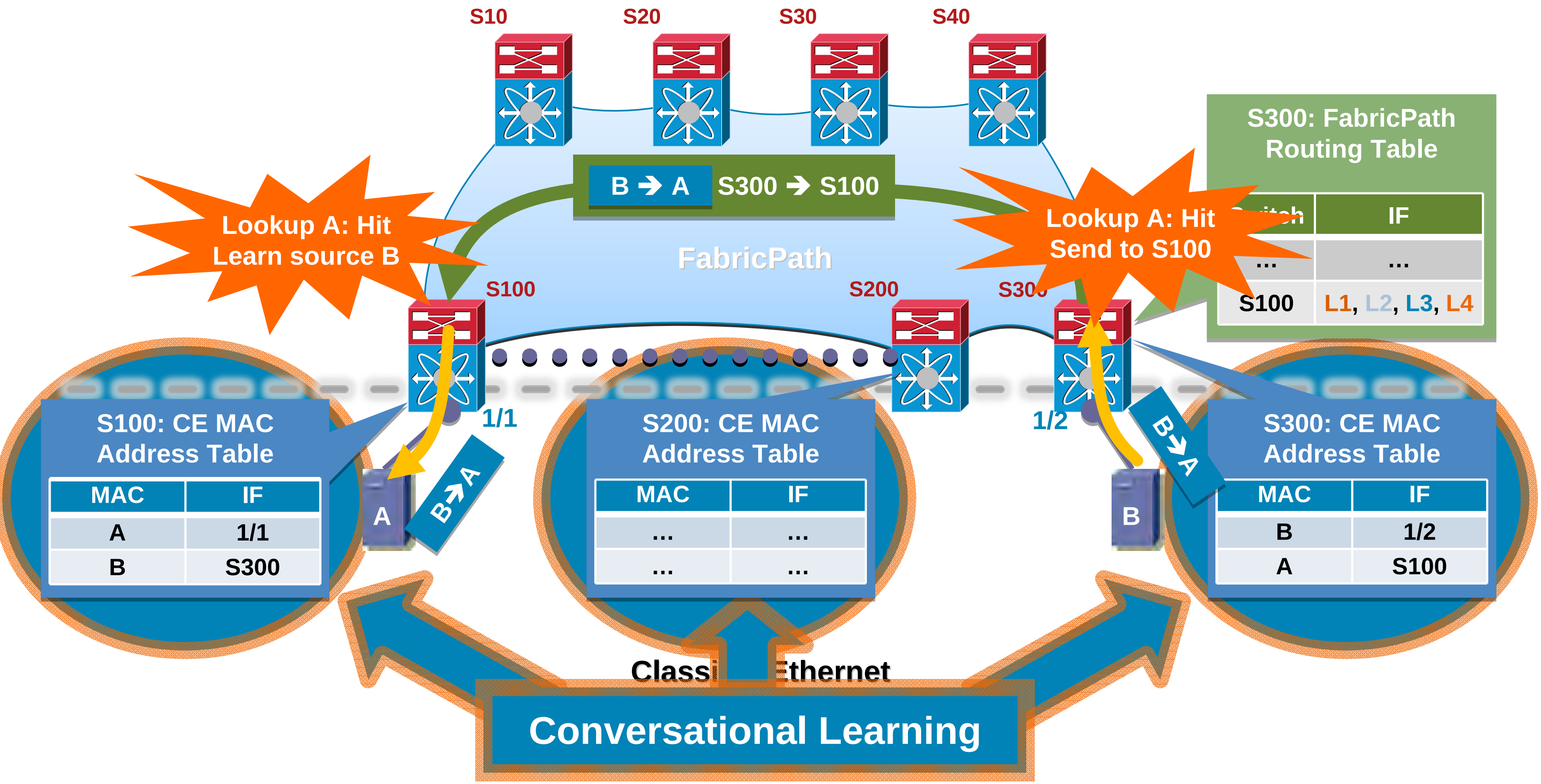
CE Edge Ports

- 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

Unknown Unicast



Known Unicast, Conversational Learning



New Data Plane Means ASIC Support: F1 Series I/O Module

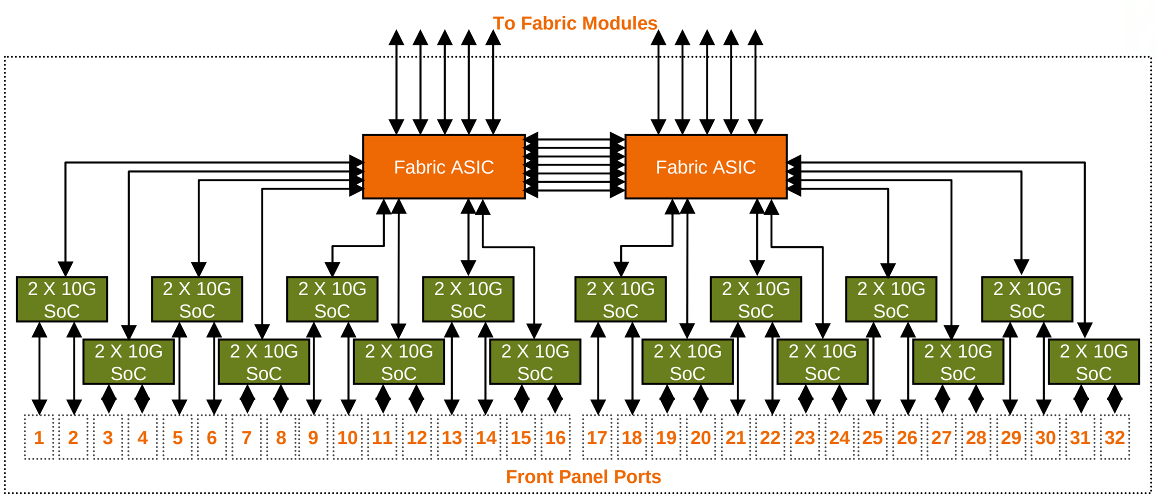
- SFP+ 1/10G I/O module
- Layer 2 forwarding with L3/L4 services (ACL/QoS)
- Multi-protocol – Classic Ethernet/VPC, FabricPath, DCB, FCoE

- High performance

  230Gbps fabric connectivity

  32 line-rate ports per slot with local switching

N7K-F132XP-



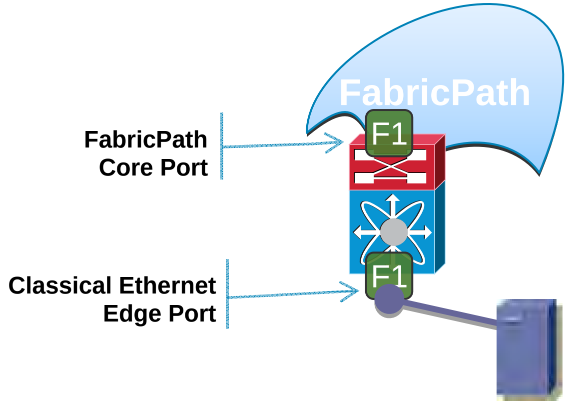
System-on-Chip (SoC)† design

† sometimes called "switch-on-chip"

Only F1 modules switch FabricPath traffic

- The Nexus 7000 features two kinds of IO Modules:
  M series and F series.
- M I/O Modules cannot switch FabricPath traffic
- When running FabricPath,
  FP Core and CE Edge ports must be on an F module
- New FabricPath/CE locally significant VLAN mode:

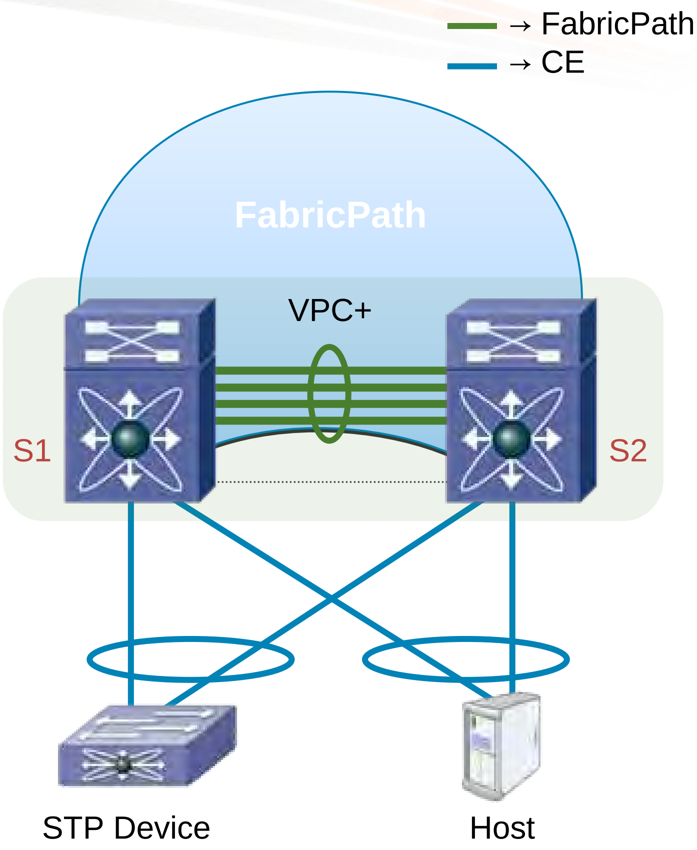
```
S100(config)# vlan 10
S100(config-vlan)# mode ?
  ce          Classical Ethernet VLAN mode
  fabricpath  Fabricpath VLAN mode

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



- FabricPath VLANs can only be enabled on F modules

Introducing VPC+

- Allows dual-homed connections from edge ports into FabricPath domain with active/active forwarding
- Classic Ethernet switches, Layer 3 routers, load-balancers, dual-homed servers, etc.
  - Only requirement is device can form port-channel interface
- Can also provide active/active HSRP
- Configuration virtually identical to standard VPC

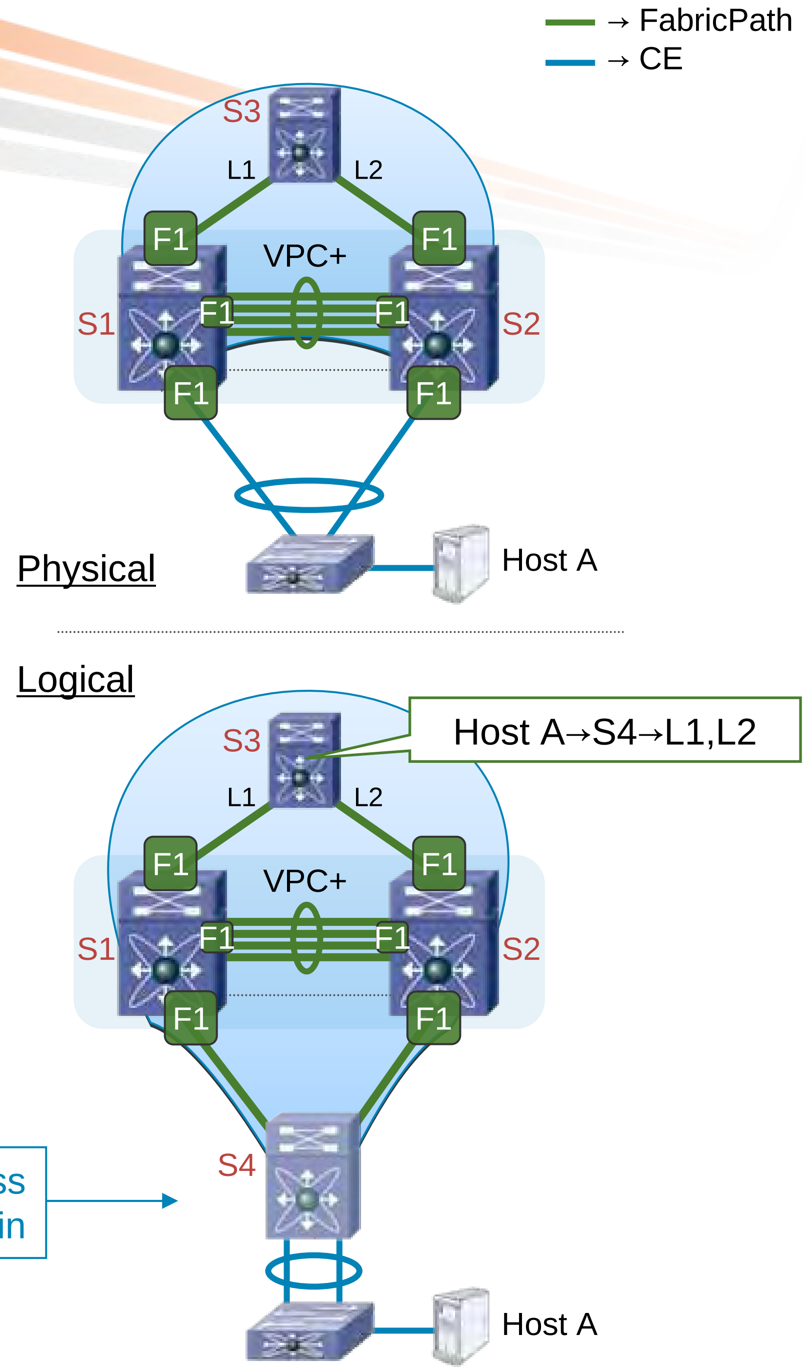


VPC+ Details

- VPC+ peer switches share a “virtual” FabricPath switch ID
- MAC addresses behind VPC+ port-channels appear as “connected” to the virtual switch, not the VPC+ peer switches
- Allows load-balancing within FabricPath domain toward the VPC+ virtual switch
- VPC+ requires F1 modules with FabricPath enabled in the VDC

    Peer-link and all VPC+ connections must be to F1 ports

Virtual “Switch 4” becomes egress switch for Host A in FabricPath domain



VPC vs. VPC+

- A given VDC can be part of VPC domain, or VPC+ domain, but **not** both
- VPC+ only works on F1 modules with FabricPath enabled in the VDC
- Conversion between VPC and VPC+ is disruptive

	VPC	VPC+
Peer-link	M1 ports or F1 ports	F1 ports
Member ports	M1 ports or F1 ports	F1 ports
VLANs	CE or FabricPath VLANs	FabricPath VLANs only
Peer-link switchport mode	CE trunk port	FabricPath core port

Transparent Interconnection of Lots of Links (TRILL)

- IETF standard for Layer 2 multipathing
- Driven by multiple vendors, including Cisco
- TRILL now officially moved from Draft to Proposed Standard in IETF
- Proposed Standard status means vendors can confidently begin developing TRILL-compliant software implementations
- Cisco FabricPath capable hardware is also TRILL capable

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

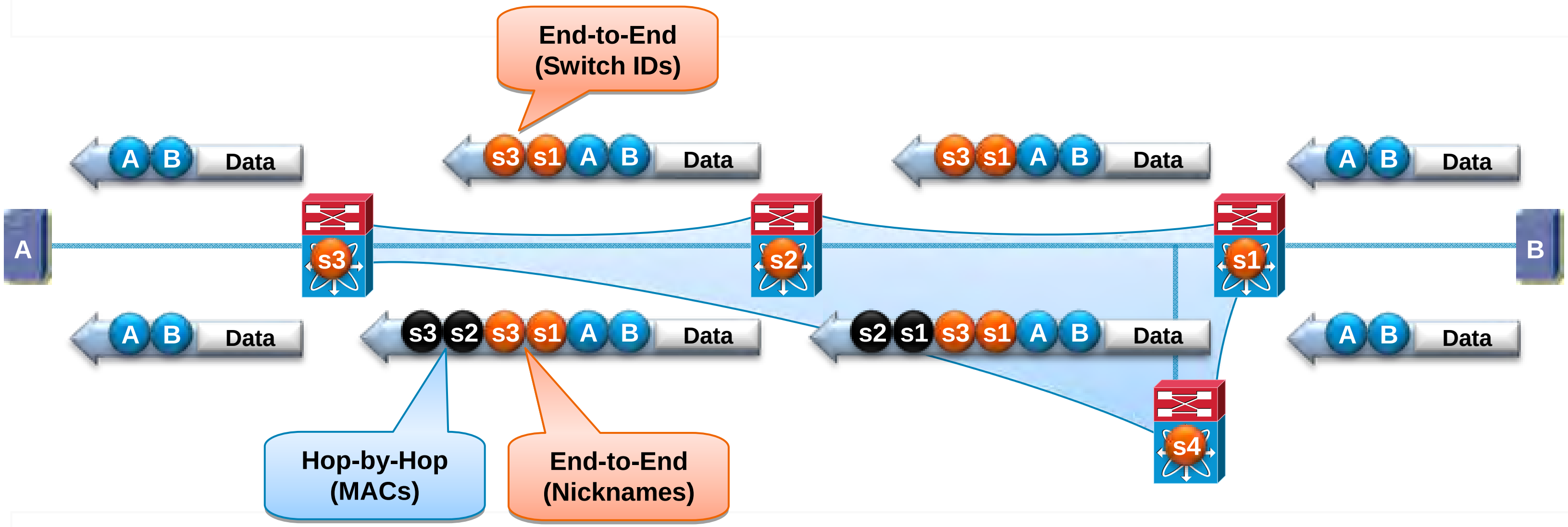
FabricPath vs. TRILL Overview

	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

FabricPath vs. TRILL: Encapsulation

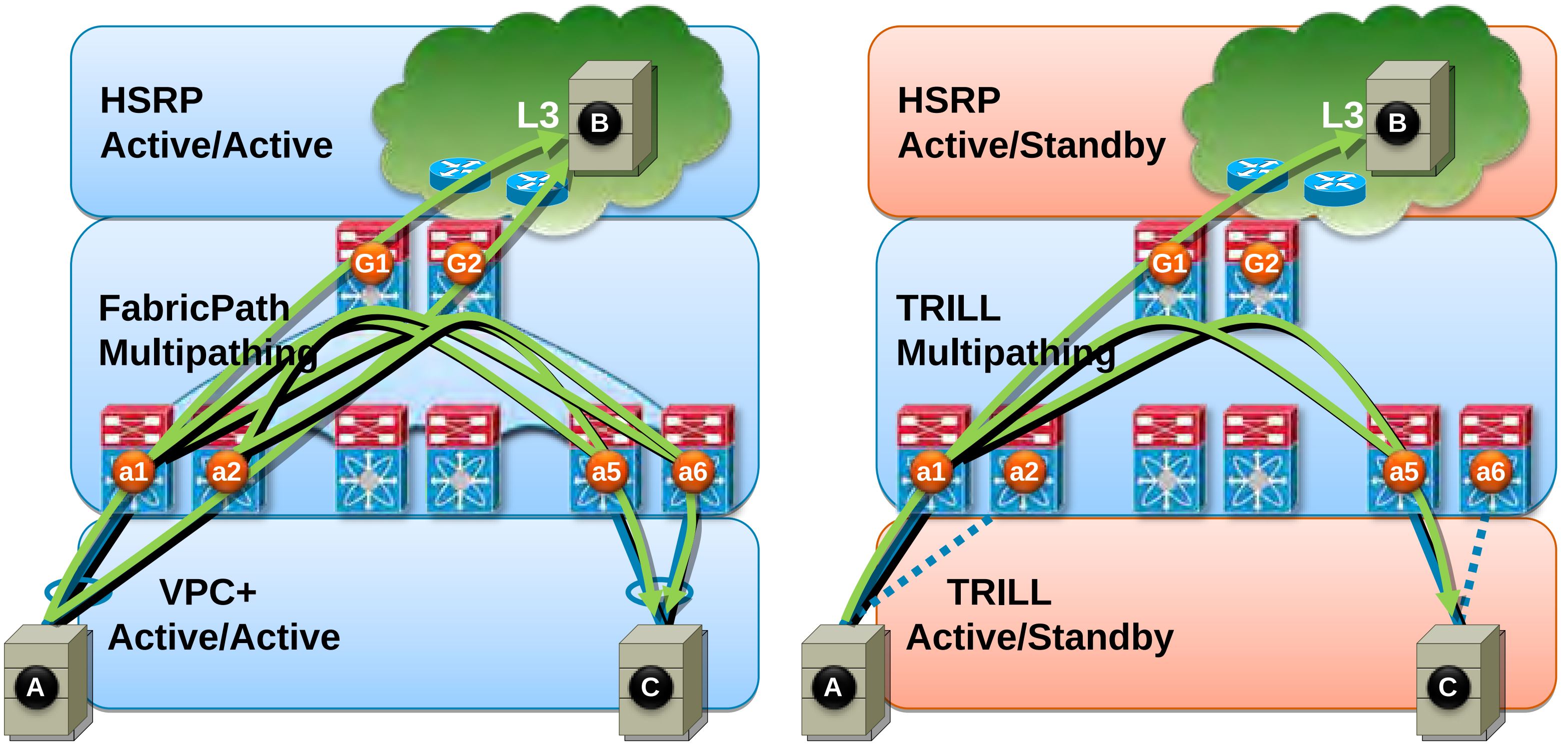
FabricPath Encapsulation



TRILL Encapsulation

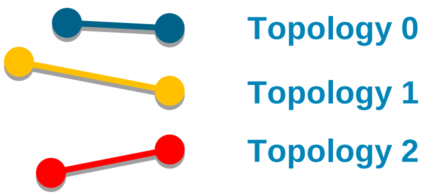
- TRILL devices can communicate over a **shared Ethernet** segment with several peers
- FabricPath has a more **compact frame format** (simpler hardware, lower latency); can only peer on point-to-point links

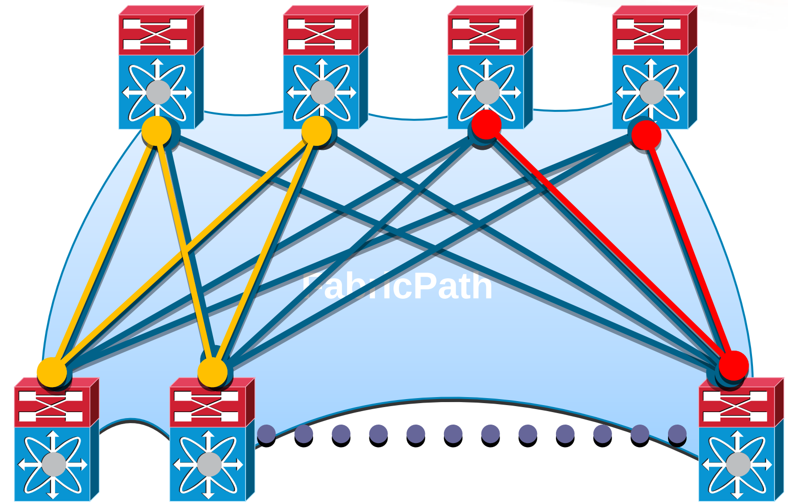
FabricPath vs. TRILL: Multipathing



- End-to-end multipathing (L2 edge, Fabric, L3 edge) provides resiliency and fast convergence

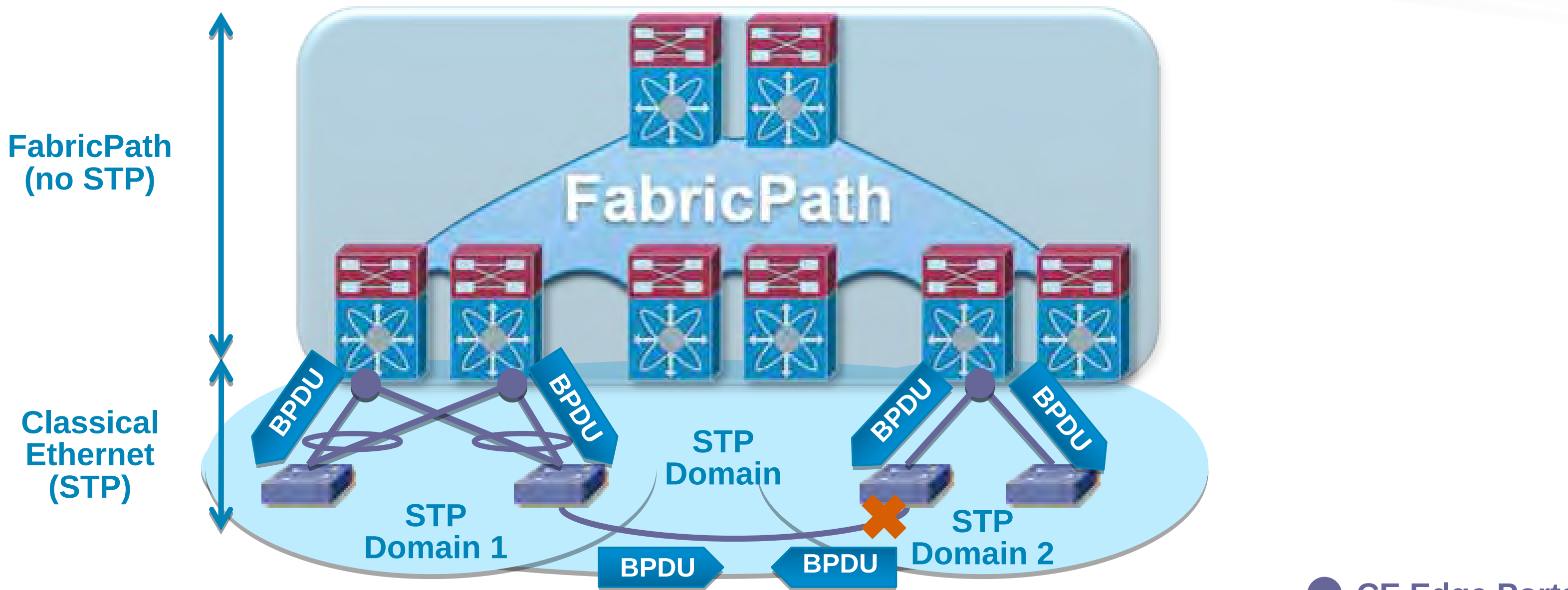
FabricPath vs. TRILL: FabricPath Multiple Topologies





- **Topology**: A group of links in the Fabric
- By default, all the links are part of topology 0
- Other topologies are 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 are used for VLAN pruning, security, traffic engineering etc…

FabricPath vs. TRILL: FabricPath Simple STP Interaction



- The Fabric looks like a single bridge:
  - It sends the same STP information on all edge ports
  - It expects to be the root of the STP for now (edge ports will block if they receive better information)
- **No BPDUs** are forwarded across the fabric
- An optional mechanism allows propagating TCNs if needed

Agenda

- Introduction to FabricPath
- FabricPath Technical Overview
- Unicast Forwarding
- Multicast Forwarding
- FabricPath Design
- Conclusion

Unicast Forwarding

FabricPath Unicast Forwarding

Control plane:

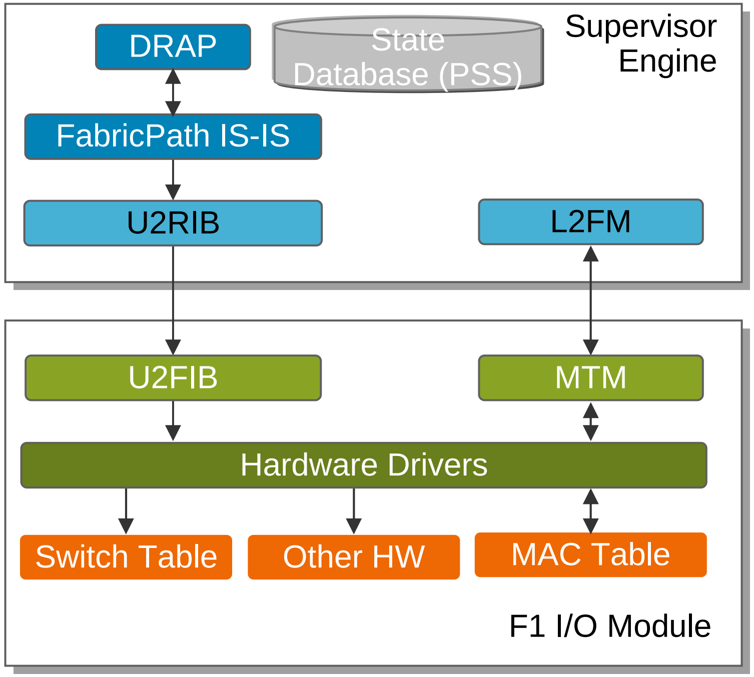
- Routing table – FabricPath IS-IS learns switch IDs (SIDs) and builds routing table
- Multidestination trees – FabricPath IS-IS elects roots and builds multidestination forwarding trees

Data plane:

- MAC table – Hardware performs MAC table lookups to determine destination FabricPath switch (unicast) or to identify multidestination frames
- Switch table – Hardware performs destination SID lookups to forward unicast frames to other switches
- Multidestination table – Hardware selects multidestination tree to forward multidestination frames through network fabric

FabricPath Unicast System Architecture

- Complete separation of control plane and data plane
- Fully modular software implementation of control plane and infrastructure components
- Fully distributed data plane forwarding with hardware-based MAC learning / forwarding and hardware SID / ECMP lookups



Key FabricPath Unicast Processes

Running on the Supervisor Engine:

- **FabricPath IS-IS** – SPF routing protocol process that forms the core of the FabricPath control plane
- **DRAP** – Dynamic Resource Allocation Protocol, extension to FabricPath IS-IS that ensures network-wide unique and consistent Switch IDs and Ftag values
- **U2RIB** – Unicast Layer 2 RIB, containing the "best" unicast Layer 2 routing information
- **L2FM** – Layer 2 forwarding manager, managing the MAC address table

Running on the I/O modules:

- **U2FIB** – Unicast Layer 2 FIB, managing the hardware unicast routing table
- **MTM** – MAC Table Manager, managing the hardware MAC address table

Hardware tables on I/O modules:

- **Switch table** – Contains Switch IDs and next-hop interfaces
- **MAC table** – Contains local and remote MAC addresses
- **Other HW** – Variety of other table memories, hardware registers, etc. required for FabricPath forwarding

FabricPath Routing Table

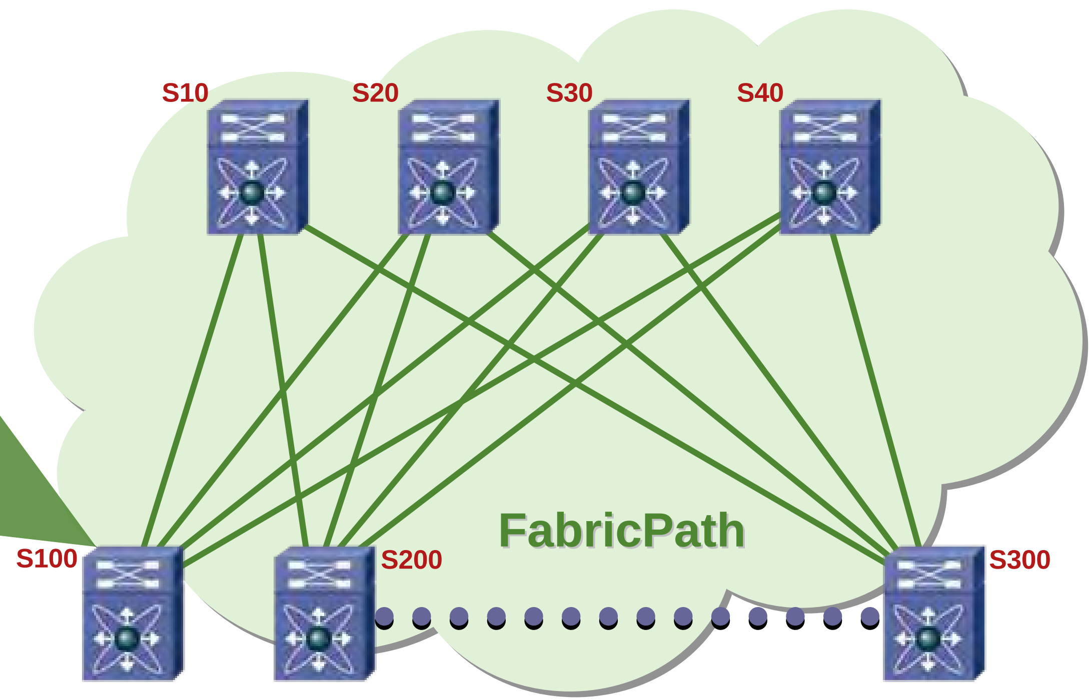
- Describes shortest (best) paths to each Switch ID based on link metrics
- Equal-cost paths supported between FabricPath switches

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



Display IS-IS View of Routing Topology

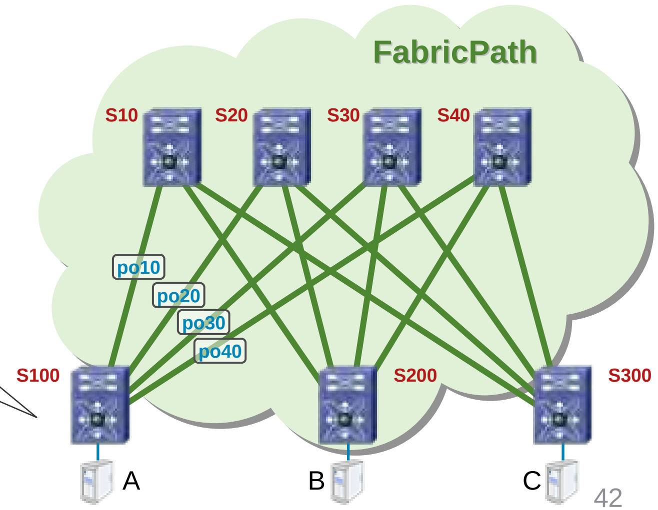
show fabricpath isis route

Destination Switch ID

Next-hop interface(s)

Routing metric

```
S100# sh fabricpath isis route
Fabricpath IS-IS domain: default MT-0
Topology 0, Tree 0, Swid routing table
10, L1
 via port-channel10, metric 20
20, L1
 via port-channel20, metric 20
30, L1
 via port-channel30, metric 20
40, L1
 via port-channel40, metric 20
200, L1
 via port-channel30, metric 40
 via port-channel40, metric 40
 via port-channel20, metric 40
 via port-channel10, metric 40
300, L1
 via port-channel30, metric 40
 via port-channel40, metric 40
 via port-channel20, metric 40
 via port-channel10, metric 40

S100#
```



Display U2RIB View of Routing Topology

show fabricpath route

```
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#
```

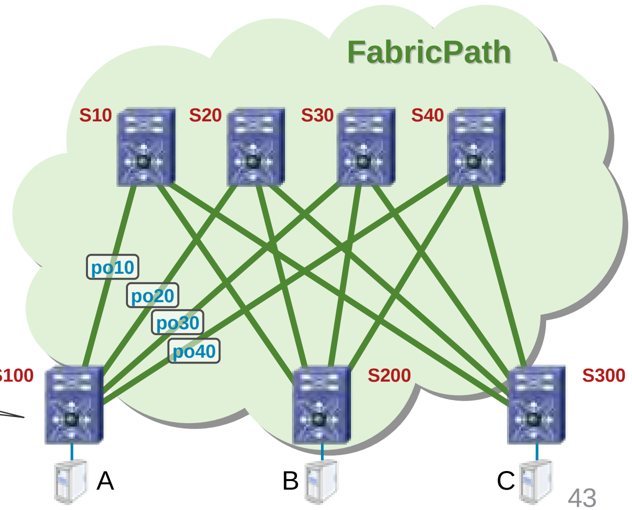
Topology (ftag), Switch ID, Sub-Switch ID

Administrative distance, routing metric

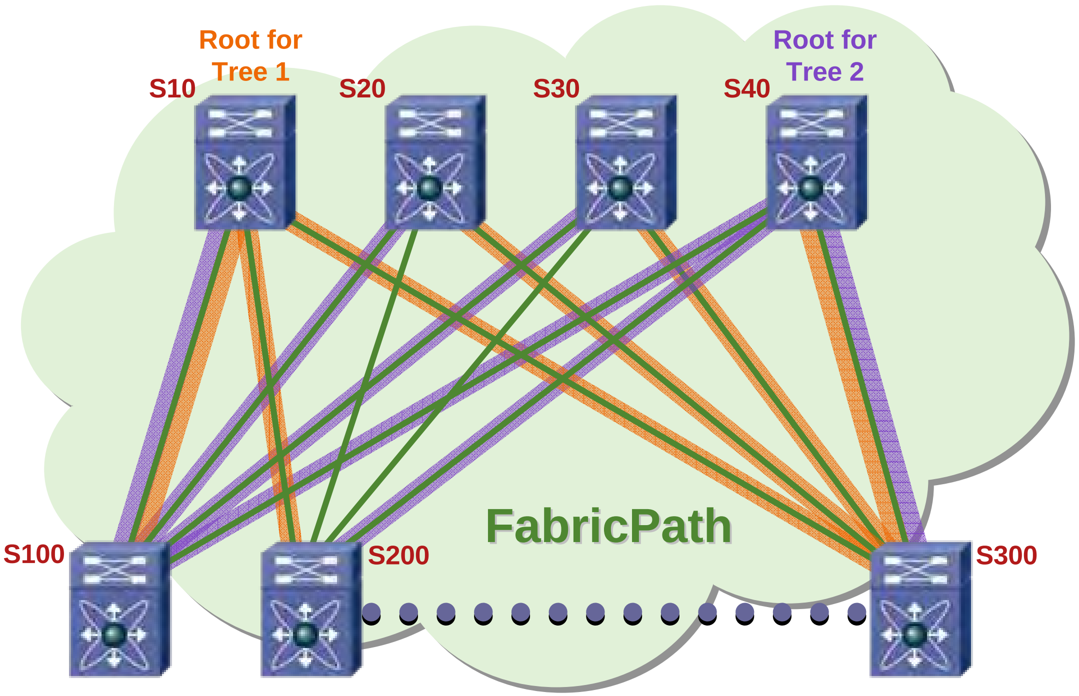
Route age

Client protocol

Next-hop interface(s)

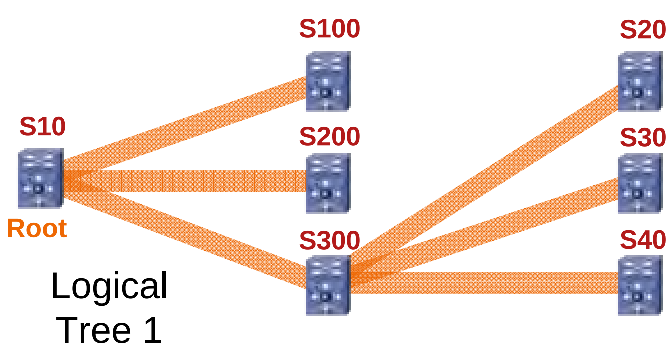


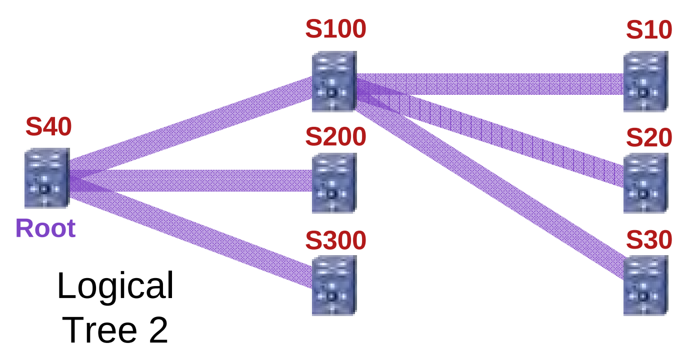
FabricPath Multidestination Trees



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

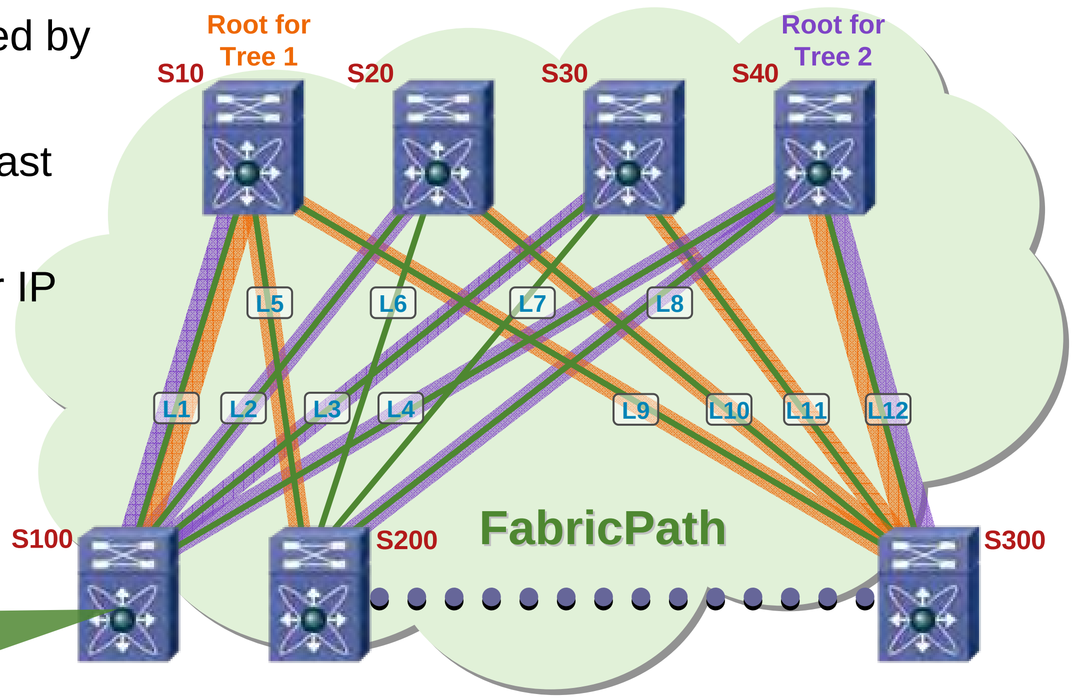
  Two multidestination trees supported in NX-OS release 5.1





Multidestination Trees and Role of the Ingress FabricPath Switch

- Ingress FabricPath switch determines which tree to use for each flow
- Other FabricPath switches forward based on tree selected by ingress switch
- Broadcast and unknown unicast typically use first tree
- Hash-based tree selection for IP multicast, with several configurable hash options



Multidestination Trees on Switch 100

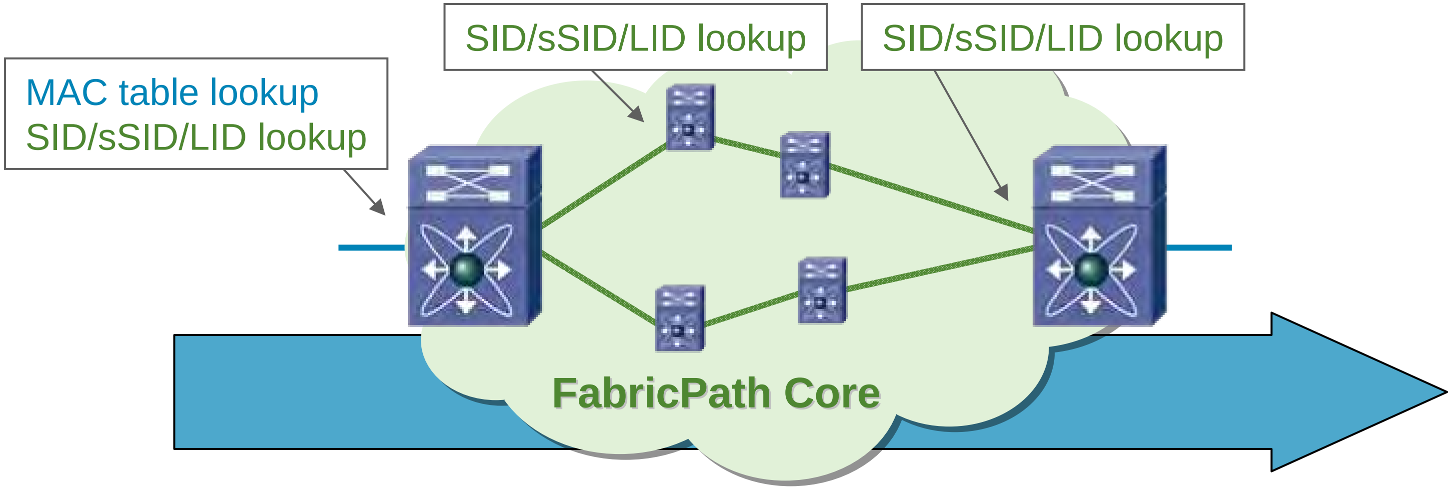
Tree	IF
1	L1
2	L1,L2,L3,L4

How Are Multidestination Roots Selected?

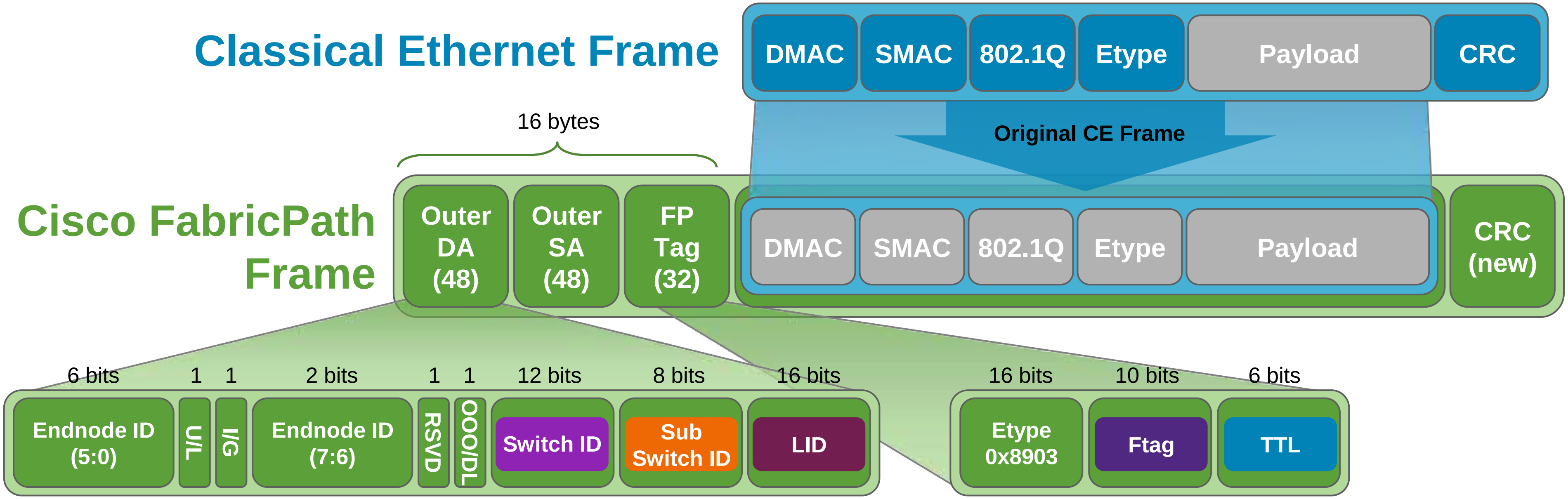
- FabricPath network elects a single root switch for the first (broadcast) multidestination tree in the topology
- All FabricPath switches announce their root priority in Router Capability TLV
- Switch with highest priority value becomes root for the tree
  - Highest system ID, then highest Switch ID value, used in event of a tie
- Broadcast root determines roots of additional multicast trees and announces them in Router Capability TLV
  - Multicast roots spread among available switches to balance load
  - Selection based on same criteria as above

Unicast Data-Plane Forwarding Decisions

- Each FabricPath switch builds Switch ID (L2 routing) table
- Ingress FabricPath switch:
  - MAC table lookup identifies destination Switch-ID (SID), sub-Switch ID (sSID), and local ID (LID)
  - SID/sSID/LID lookup determines next-hop interface(s) for destination SID
- Core FabricPath switch:
  - SID/sSID/LID lookup determines next-hop interface(s) for destination SID
- Egress FabricPath switch:
  - SID/sSID/LID lookup determines output port



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

What Are the Switch ID and Sub-Switch ID?

- Every FabricPath switch assigned a **Switch ID** (12 bits)
  - Network provisions Switch IDs automatically
  - User has option to manually configure Switch ID, but must ensure all switches in FabricPath domain have unique value
- Encoded in “Outer MAC addresses” of FabricPath MAC-in-MAC frames
- FabricPath network automatically detects conflicting Switch IDs and prevents data path initialization on FabricPath interfaces
- **Sub-Switch ID** (8 bits) used in VPC+ to identify specific VPC+ bundle associated with VPC+ Virtual Switch ID
  - Must be unique within each VPC+ Virtual Switch domain

What is the LID?

- **Local ID** – Identifies the exact port which sourced the frame, or to which the frame is destined
- Encoded in “Outer MAC addresses” of FabricPath MAC-in-MAC frames
- Egress FabricPath switch uses LID to determine output interface
  - Removes requirement for MAC learning on FabricPath core ports
- LID is locally significant – has no meaning to switches other than the one originating it

What Is the Ftag?

- **Forwarding tag** – Unique 10-bit number identifying topology and/or distribution tree
- Encoded in FabricPath Tag field in FabricPath MAC-in-MAC frames
- For unicast packets, identifies which FabricPath IS-IS topology to use
- For multidestination packets (broadcast, multicast, unknown unicast), identifies which distribution tree to use

Putting It All Together – Host A to Host B

(1) Broadcast ARP Request

Multidestination Trees on Switch 10

4 Ftag →

Tree	IF
1	po100,po200,po300
2	po100

Multidestination Trees on Switch 100

3 Broadcast →

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

FabricPath MAC Table on S100

MAC	IF/SID
A	e1/13 (local) 2

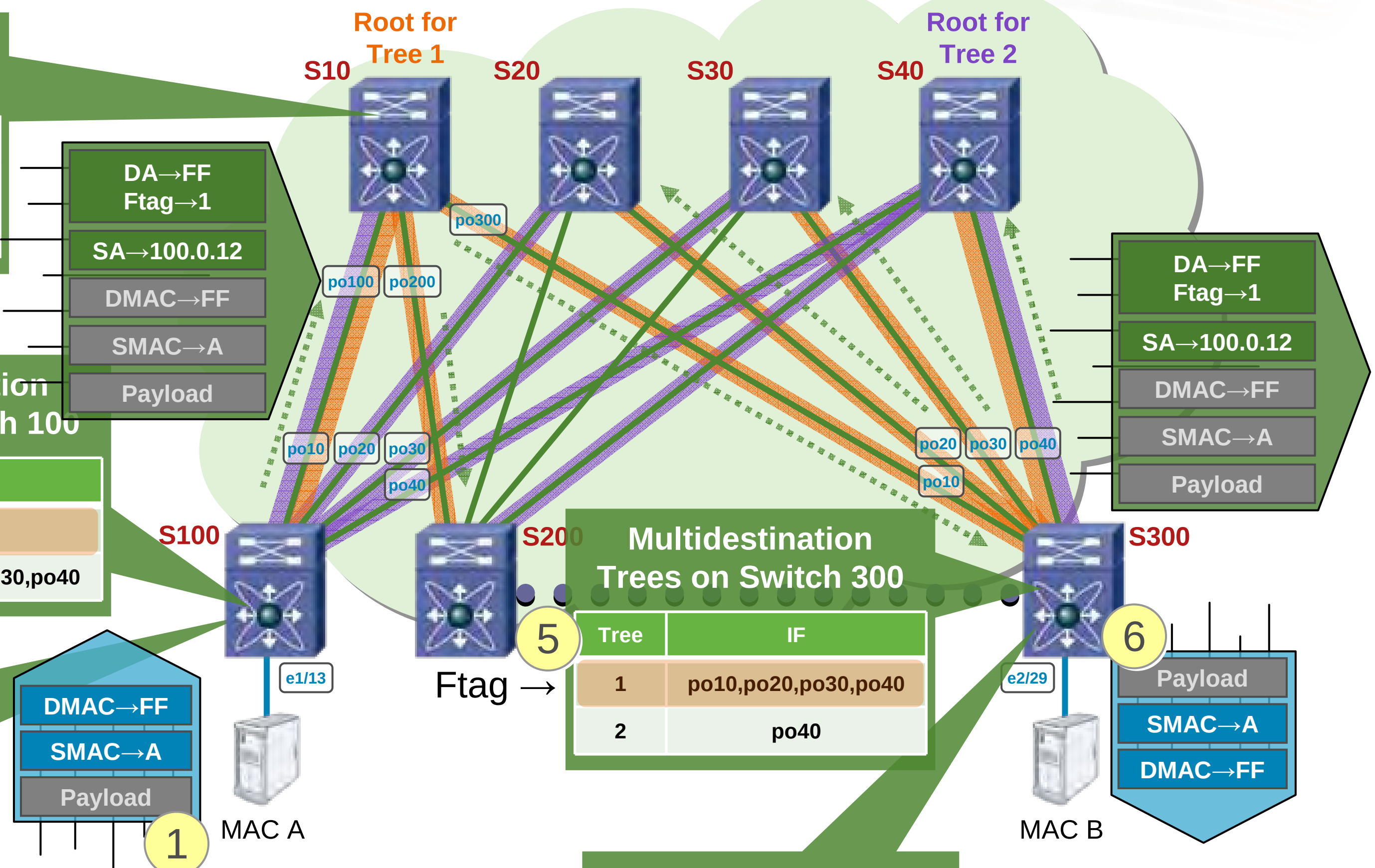
Multidestination Trees on Switch 300

5 Ftag →

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

FabricPath MAC Table on S200

MAC	IF/SID



Learn MACs of directly-connected devices unconditionally

Don't learn MACs from flood frames

Broadcast Forwarding

- Ingress FabricPath switch determines which tree to use – broadcast typically uses first Tree ID (Ftag 1)
- Outer Destination MAC remains all-ones
- Other FabricPath switches honor Tree ID selected by ingress switch (Tree 1 in this case)
- Egress FabricPath switch removes FabricPath header and floods in VLAN based on broadcast

MAC Address Tables After Broadcast ARP

- 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

S100#
```

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
---------+-----------------+--------+---------+------+----+------------------

S10#
```

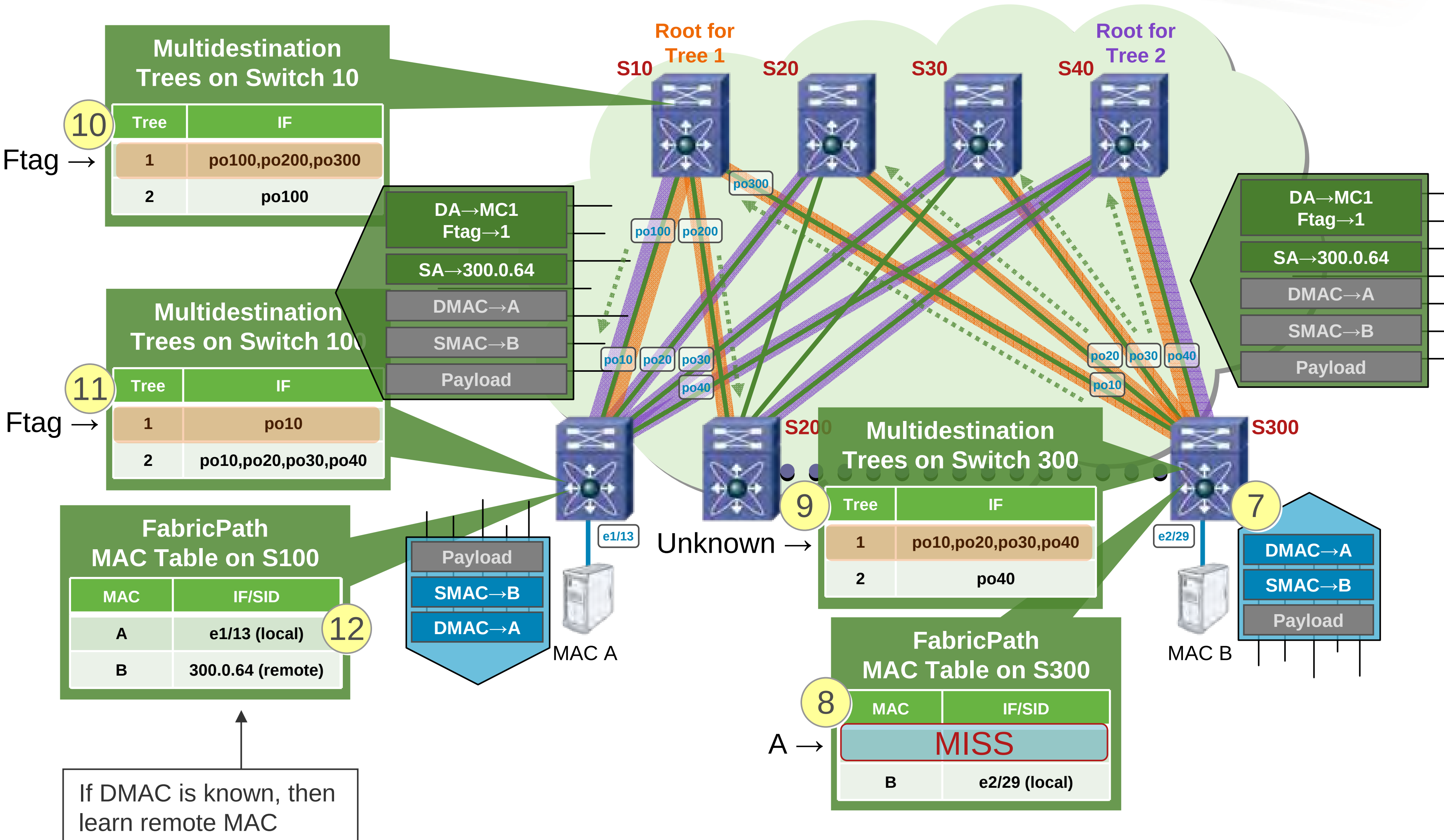
MAC A not learned on other switches

- 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) Unicast ARP Reply



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

Unknown Unicast Forwarding

- Ingress FabricPath switch determines which tree to use – unknown unicast typically uses first Tree ID (Ftag 1)
- Outer Destination MAC set to well-known “flood to fabric” multicast address (MC1)*
- Other FabricPath switches honor Tree ID selected by ingress switch (Tree 1 in this case)
- Egress FabricPath switch removes FabricPath header and floods in VLAN based on unknown unicast

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

MAC Address Tables After Unicast ARP Reply

- 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

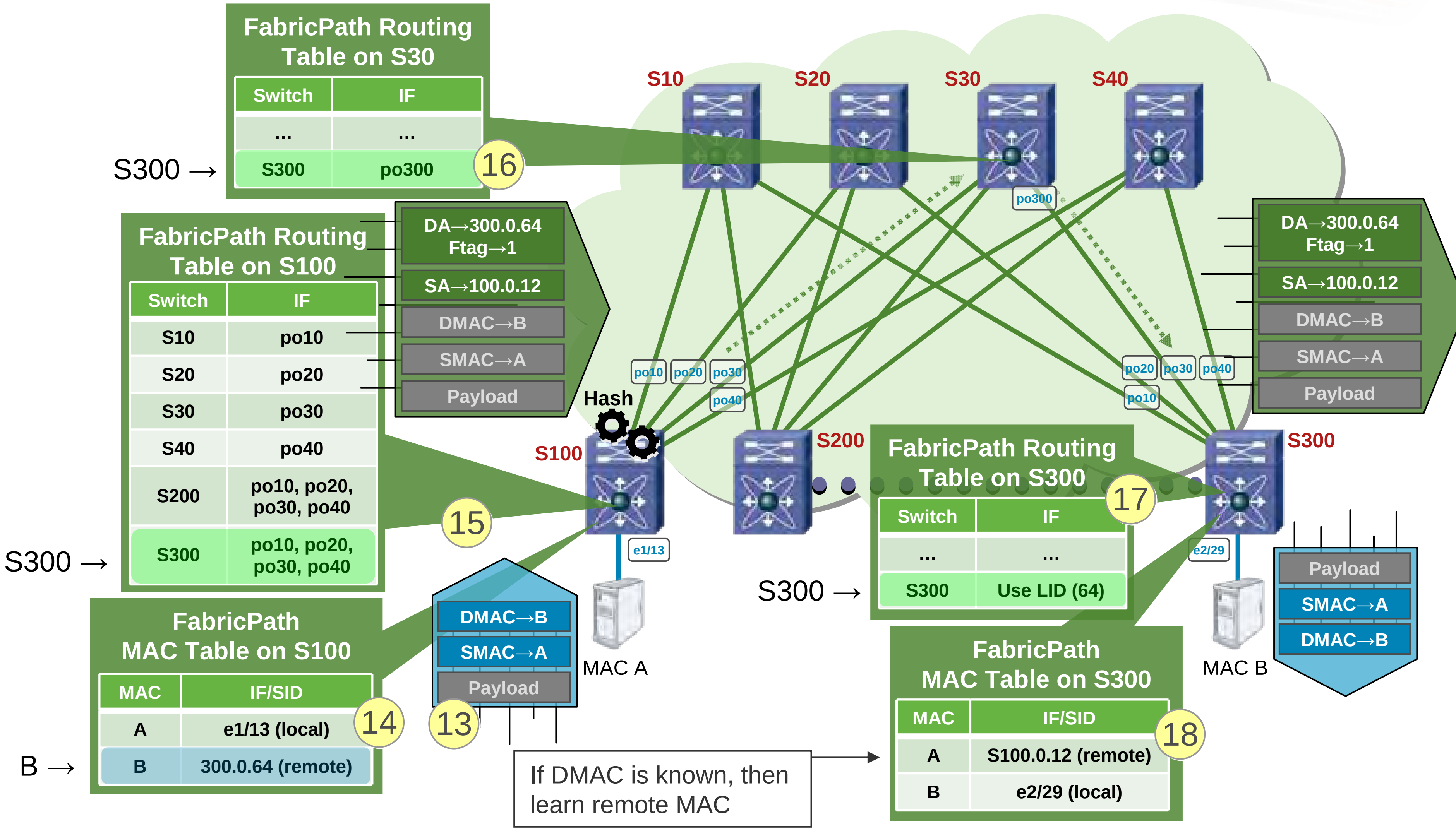
- S300:

```
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

(3) Unicast Data



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)

MAC Address Tables After Unicast Data

- 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

Agenda

- Introduction to FabricPath
- FabricPath Technical Overview
- Unicast Forwarding
- Multicast Forwarding
- FabricPath Design
- Conclusion

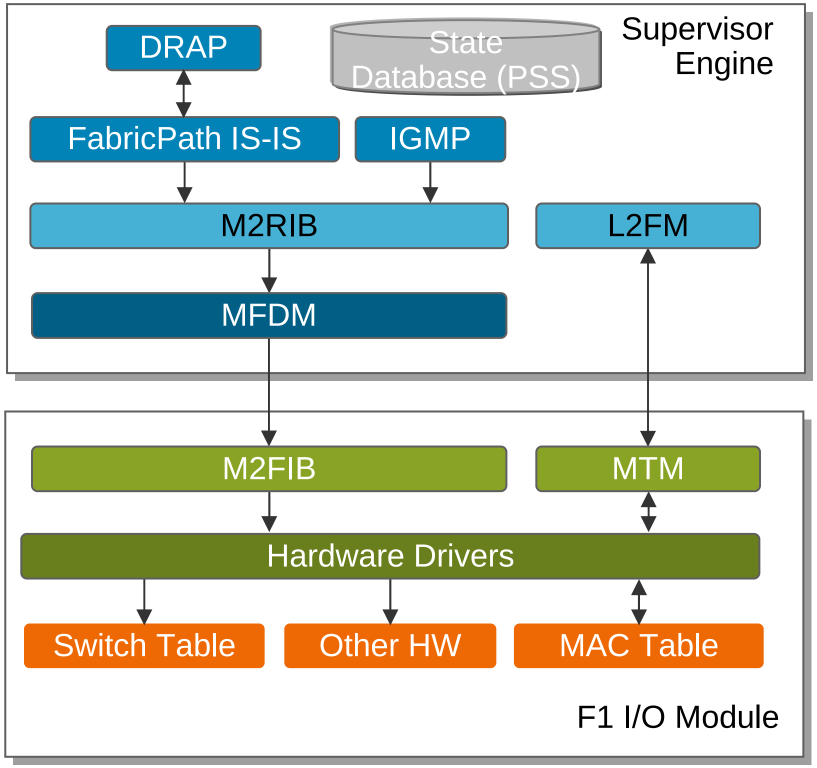
Multicast Forwarding

FabricPath IP Multicast

- **Control plane:**
  - IGMP snooping operates as usual in FabricPath edge switches
  - FabricPath IS-IS learns multicast group membership from IGMP snooping on edge switch
  - FabricPath edge switch announces group interest by using GM-LSPs, creating “pruned trees” for each group on each multidestination tree
- **Data plane:**
  - Hardware selects which multidestination tree to use for each flow based on hash function
  - Once tree selected, traffic constrained to pruned tree for that IP multicast group, based on MAC table lookup

FabricPath IP Multicast System Architecture

- Complete separation of control plane and data plane
- Fully modular software implementation for control plane and infrastructure components
- Fully distributed data plane forwarding with hardware-based MAC learning



Key FabricPath Multicast Processes

Running on the Supervisor Engine:

- **FabricPath IS-IS** – SPF routing protocol process that forms the core of the FabricPath control plane
- **DRAP** – Dynamic Resource Allocation Protocol, extension to FabricPath IS-IS that ensures network-wide unique and consistent Switch IDs and Ftag values
- **IGMP** – Provides IGMP snooping support for building multicast forwarding database
- **M2RIB** – Multicast Layer 2 RIB, containing the "best" multicast Layer 2 routing information
- **L2FM** – Layer 2 forwarding manager, managing the MAC address table
- **MFDM** – Multicast forwarding distribution manager, providing shim between platform-independent control-plane processes and platform-specific processes on I/O modules

Running on the I/O modules:

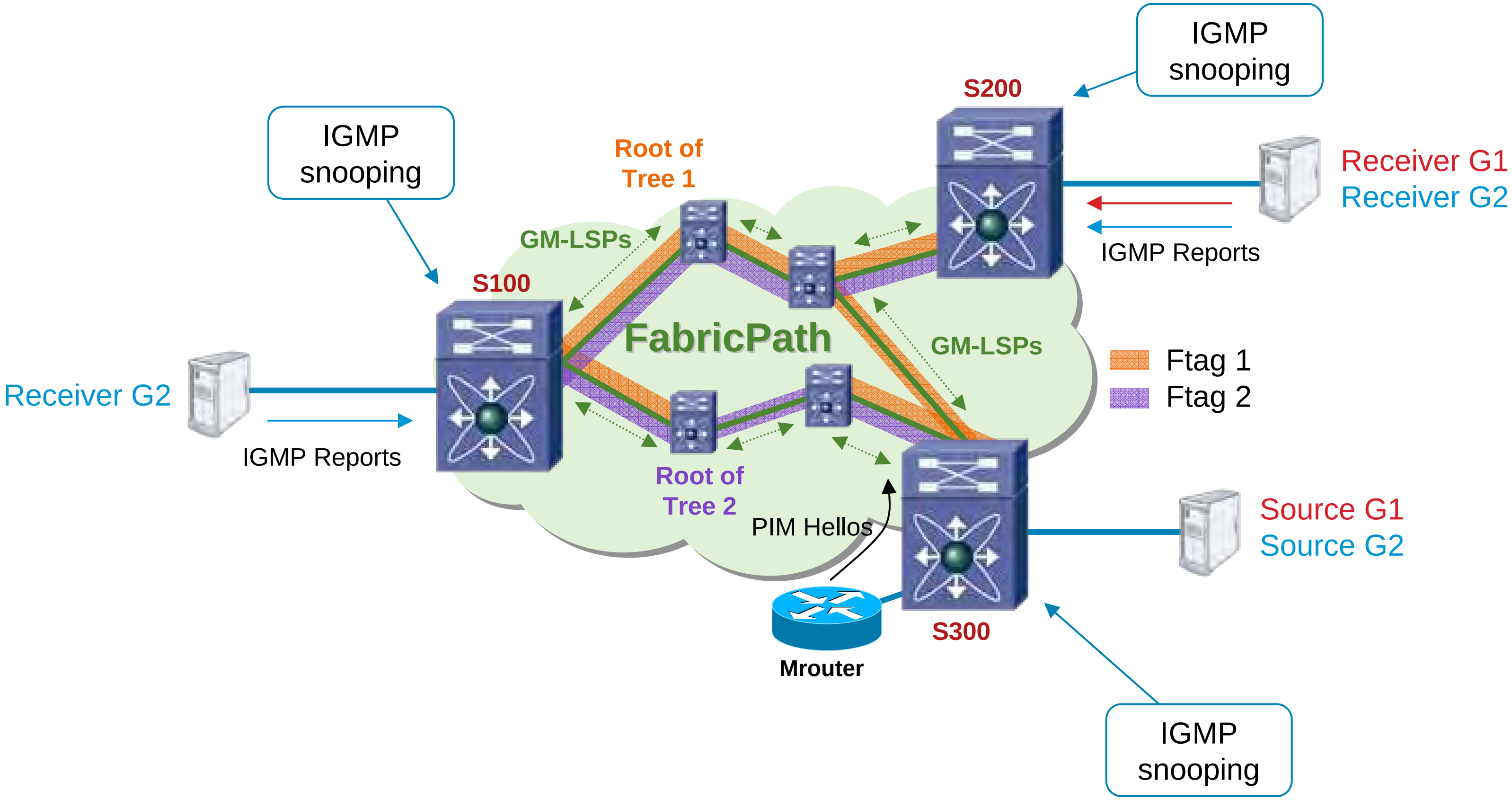
- **M2FIB** – Multicast Layer 2 FIB, managing the hardware multicast routing table
- **MTM** – MAC table manager, managing the hardware of the MAC address table

Hardware tables on I/O modules:

- **Switch table** – Contains Switch IDs and next-hop interfaces
- **MAC table** – Contains local and remote MAC addresses
- **Other HW** – Variety of other table memories, hardware registers, etc. required for FabricPath forwarding

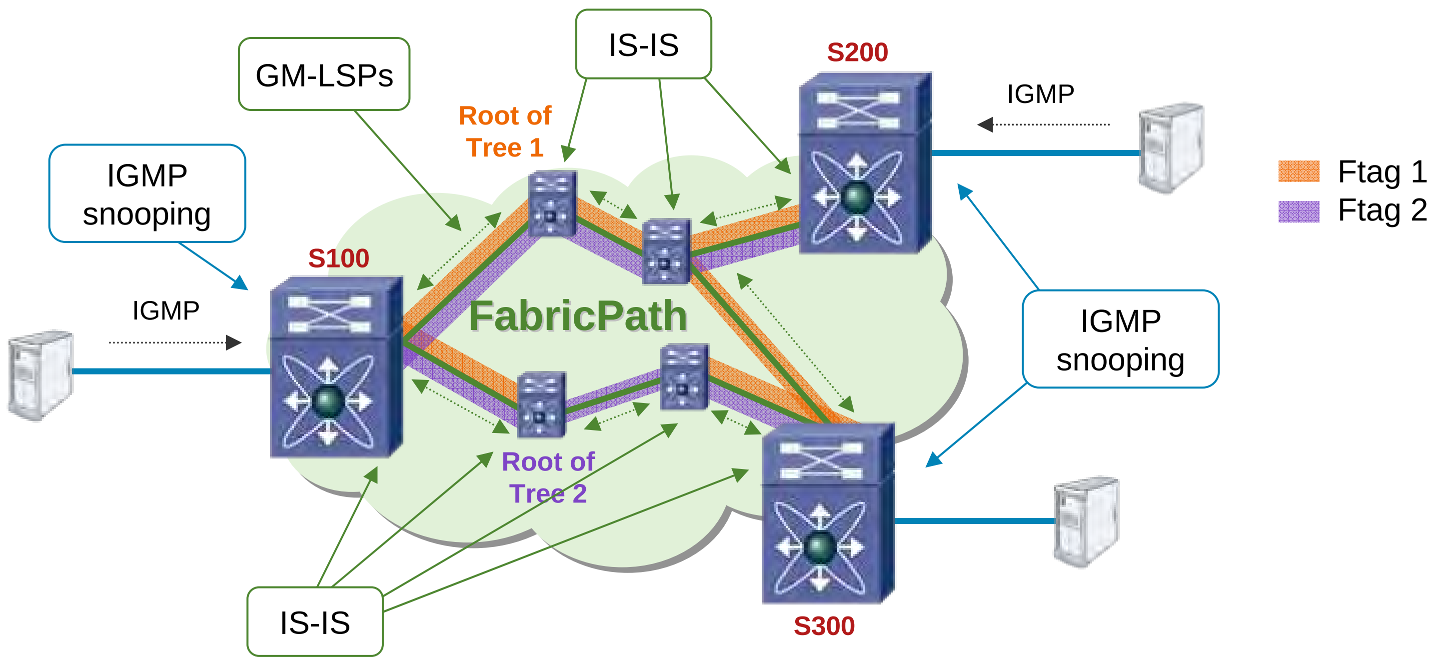
IGMP Snooping in FabricPath

- IGMP snooping learns of interested receivers on FabricPath edge switches
- Membership tracked on CE ports based on receiving IGMP reports/leaves
  - Only locally connected receivers tracked on a given edge switch
- Group membership advertised in FabricPath IS-IS using GM-LSPs



What Are GM-LSPs?

- Group Membership LSPs contain multicast forwarding information
  - Called Multicast Group PDU in TRILL
- Build Layer 2 multicast forwarding state for FabricPath core ports
  - Per-group IGMP snooping state created only at FabricPath edge switches
- Flooded to other FabricPath switches to advertise which edge switches need which multicast groups

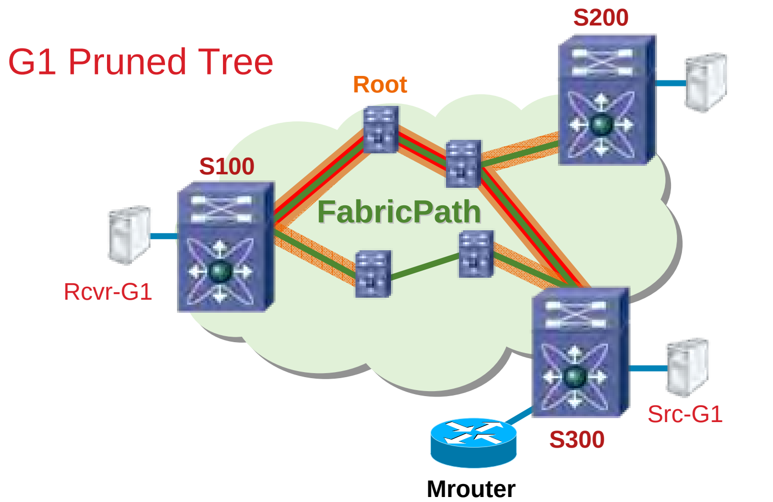


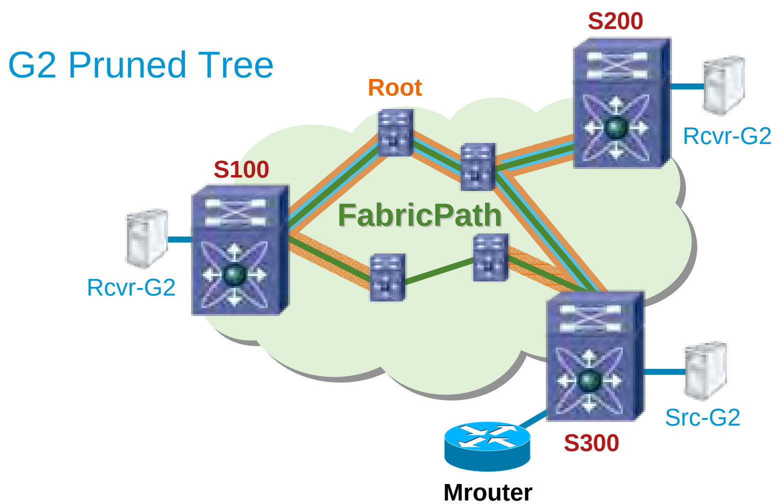
FabricPath IP Multicast Control Plane

IS-IS Creates Pruned Forwarding Trees Using GM-LSPs

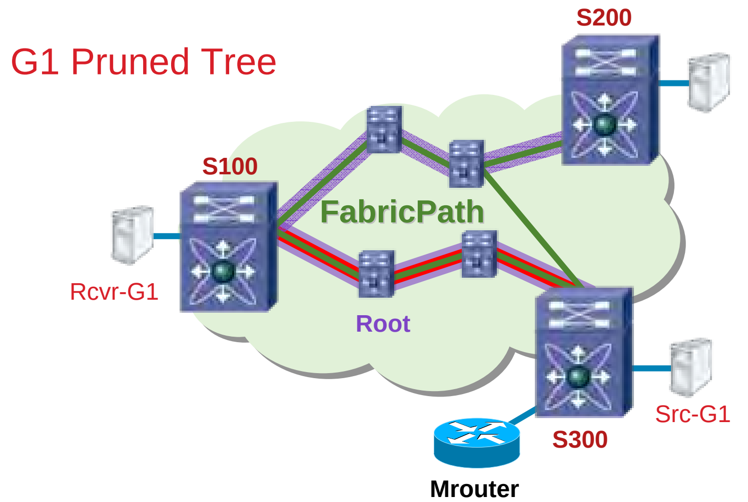
Ftag 1
Ftag 2

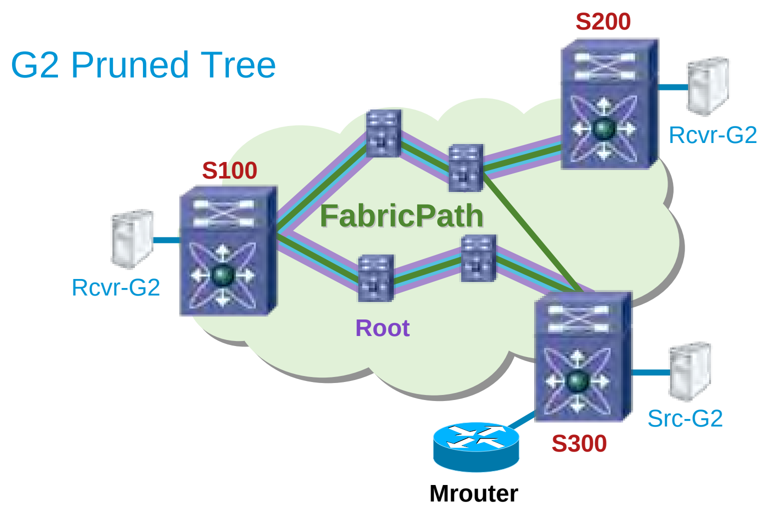
Multidestination Tree 1





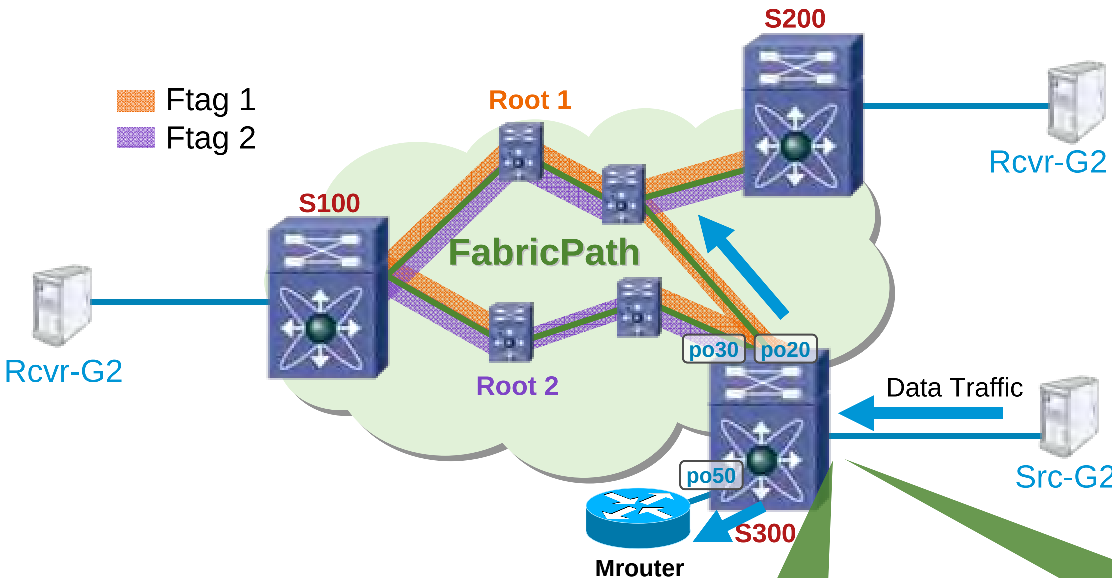
Multidestination Tree 2





FabricPath IP Multicast Data Plane

Tree Selection and MAC Table Lookup on Ingress Switch – Ftag 1



Hash

Packet data → → → Tree 1 →

FabricPath Multicast Trees

VLAN	Tree (Ftag)
10	1
10	2

FabricPath MAC Table

Tree (Ftag)	VLAN	Group	SID	IFs
1	10	G2	S100,S200	po20, po50
2	10	G2	S100,S200	po30, po50

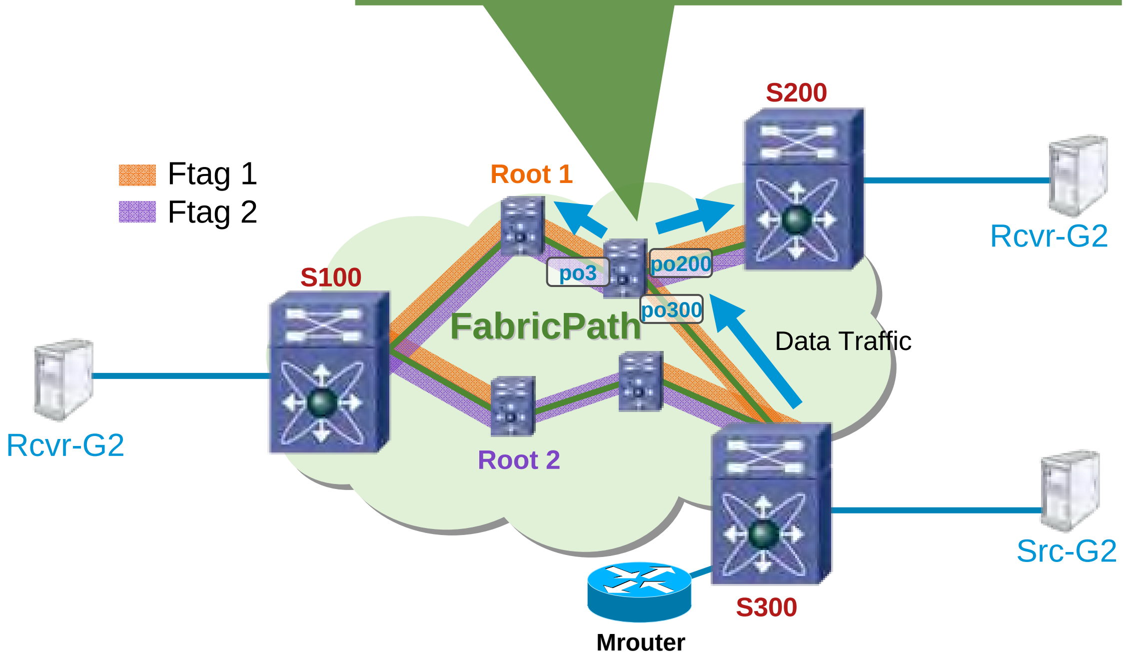
FabricPath IP Multicast Data Plane

MAC Table Lookup on Core Switch – Ftag 1

FabricPath MAC Table

Tree (Ftag)	VLAN	Group	SID	IFs
1	10	G2	S100,S200	po3,po200
2	10	G2	S100,S200	po3,po200

Tree 1 →



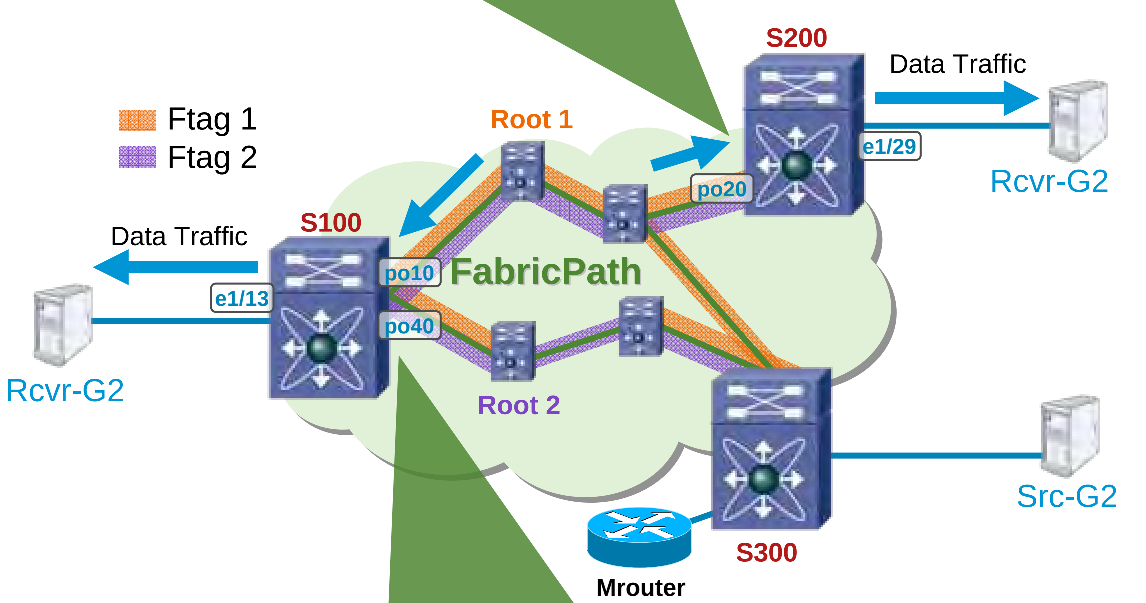
FabricPath IP Multicast Data Plane

MAC Table Lookup on Egress Switches – Ftag 1

FabricPath MAC Table

Tree 1 →

Tree (Ftag)	VLAN	Group	SID	IFs
1	10	G2	S100,S200	po20,e1/29
2	10	G2	S100,S200	po20,e1/29



FabricPath MAC Table

Tree 1 →

Tree (Ftag)	VLAN	Group	SID	IFs
1	10	G2	S100,S200	po10,e1/13
2	10	G2	S100,S200	po10,e1/13

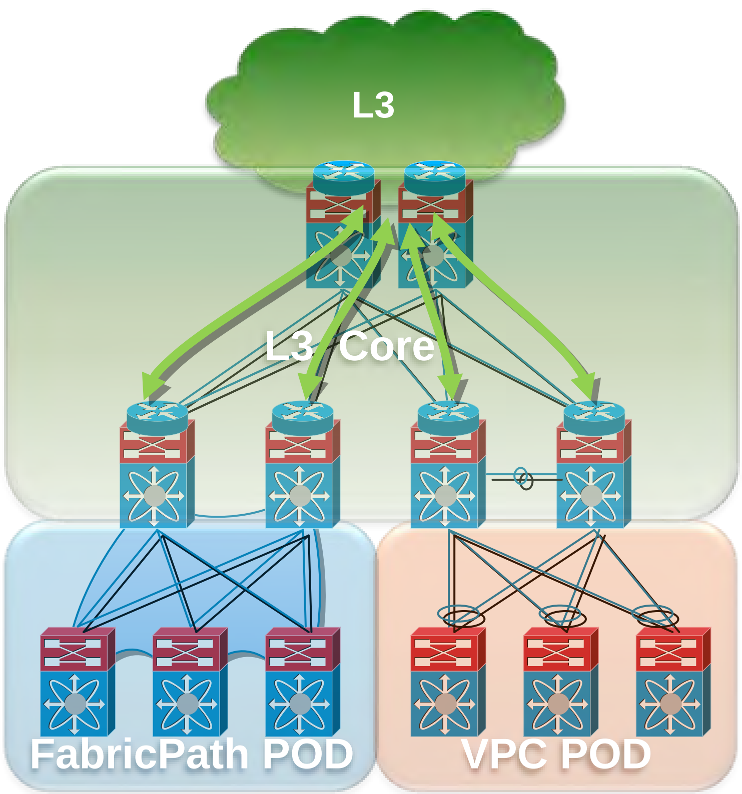
Agenda

- Introduction to FabricPath
- FabricPath Technical Overview
- Unicast Forwarding
- Multicast Forwarding
- FabricPath Design
- Conclusion

FabricPath Design

Classical POD with FabricPath

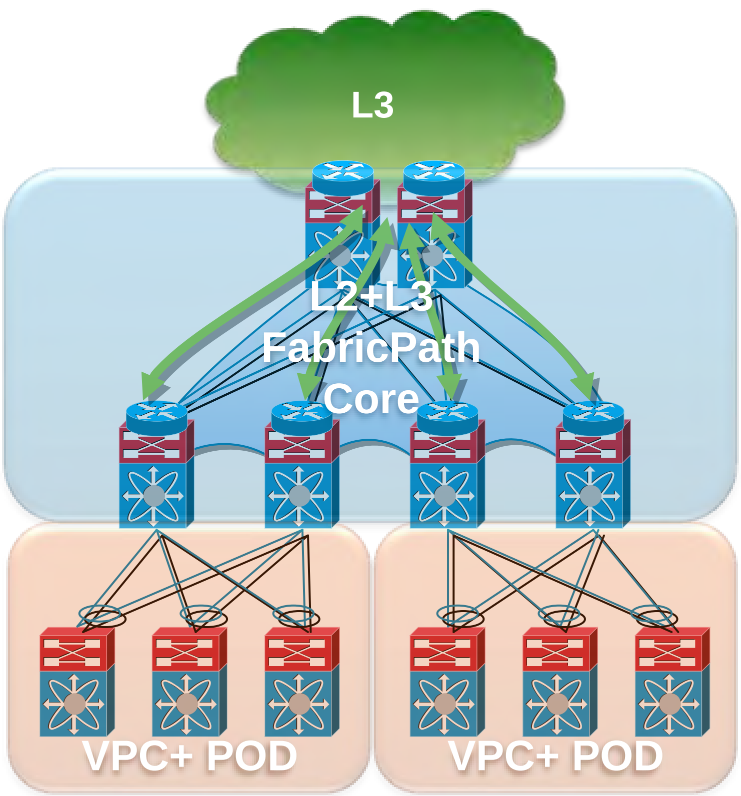
FabricPath vs. VPCs/STP



- Simple configuration (no peer link, no pair of switches, no port channels)
- Total flexibility in design and cabling
- Seamless L3 integration
- No STP, no traditional bridging (no topology changes, no sync to worry about, no risk of loops)
- Scale mac address tables with conversational learning
- Unlimited bandwidth, even if hosts are single attached
- Can extend easily and without operational impact

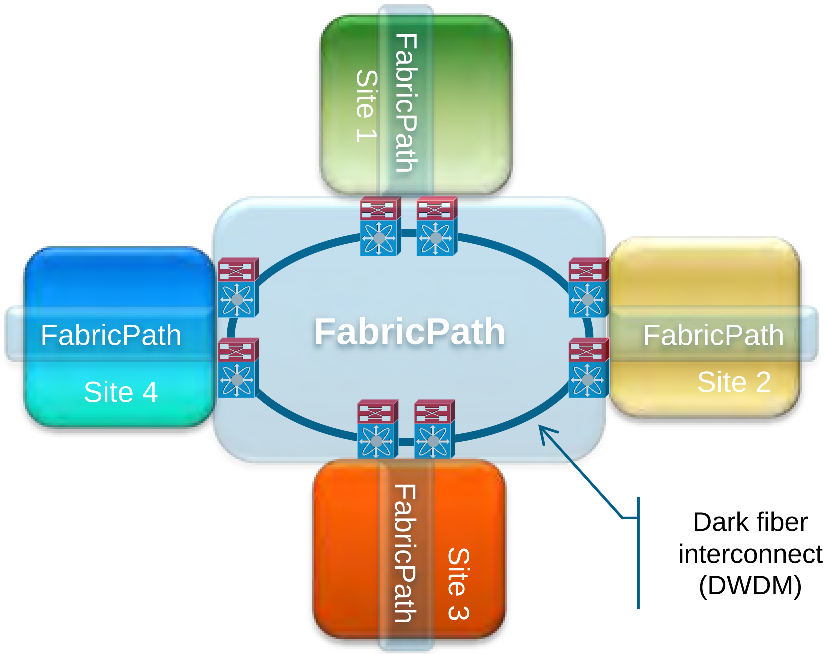
FabricPath Core

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 centers can be aggregated.
- Bandwidth or scale can be introduced in a non-disruptive way

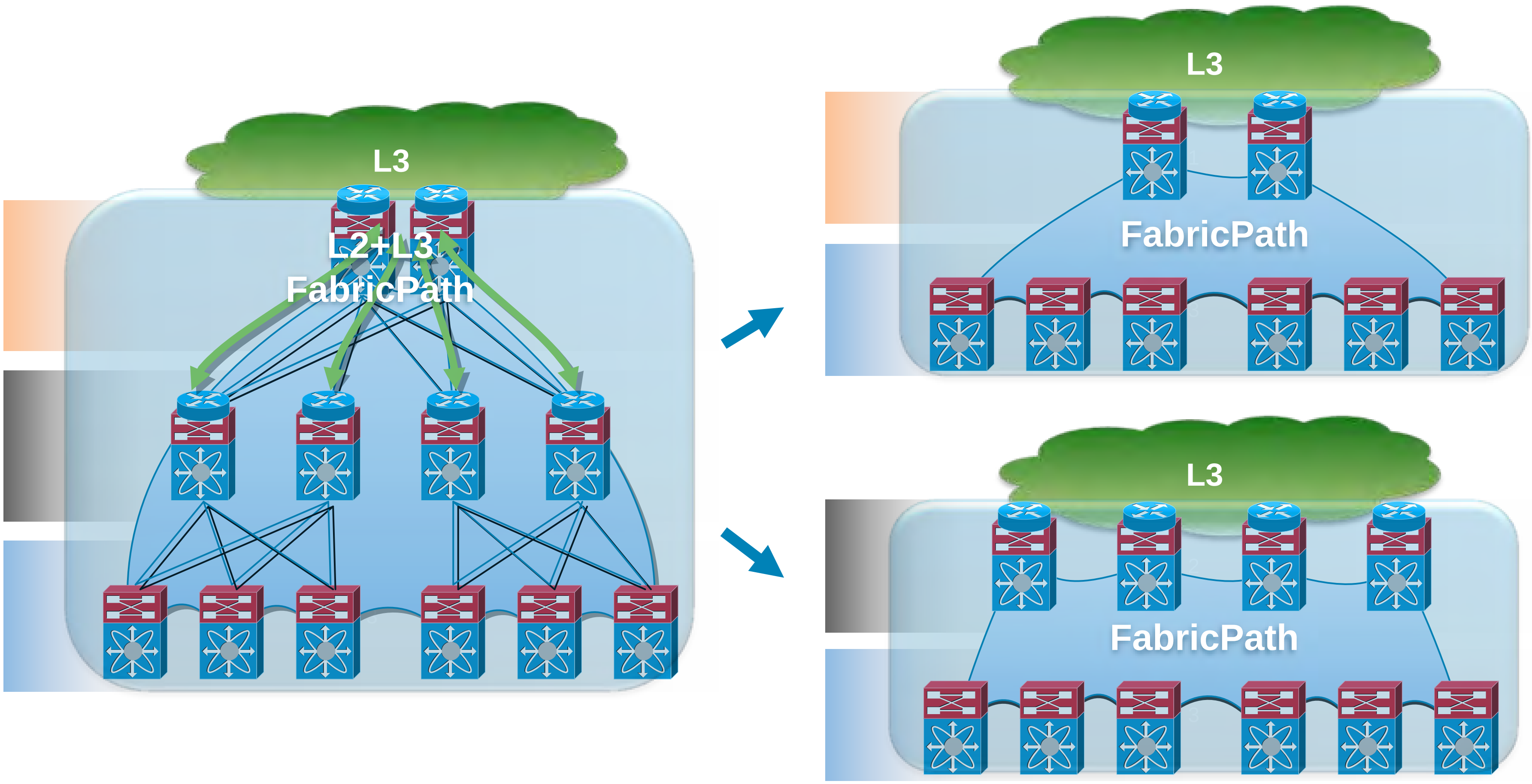
FabricPath as Site Interconnect



- Requires dark fiber
- Arbitrary interconnect topology (not dependent of port channels)
- Any number of sites
- High bandwidth, fast convergence
- Spanning tree isolation
- Mac address scaling
- VLANs can be selectively extended while others can be terminated and routed over the interconnect

Combining FabricPath PODs and Core

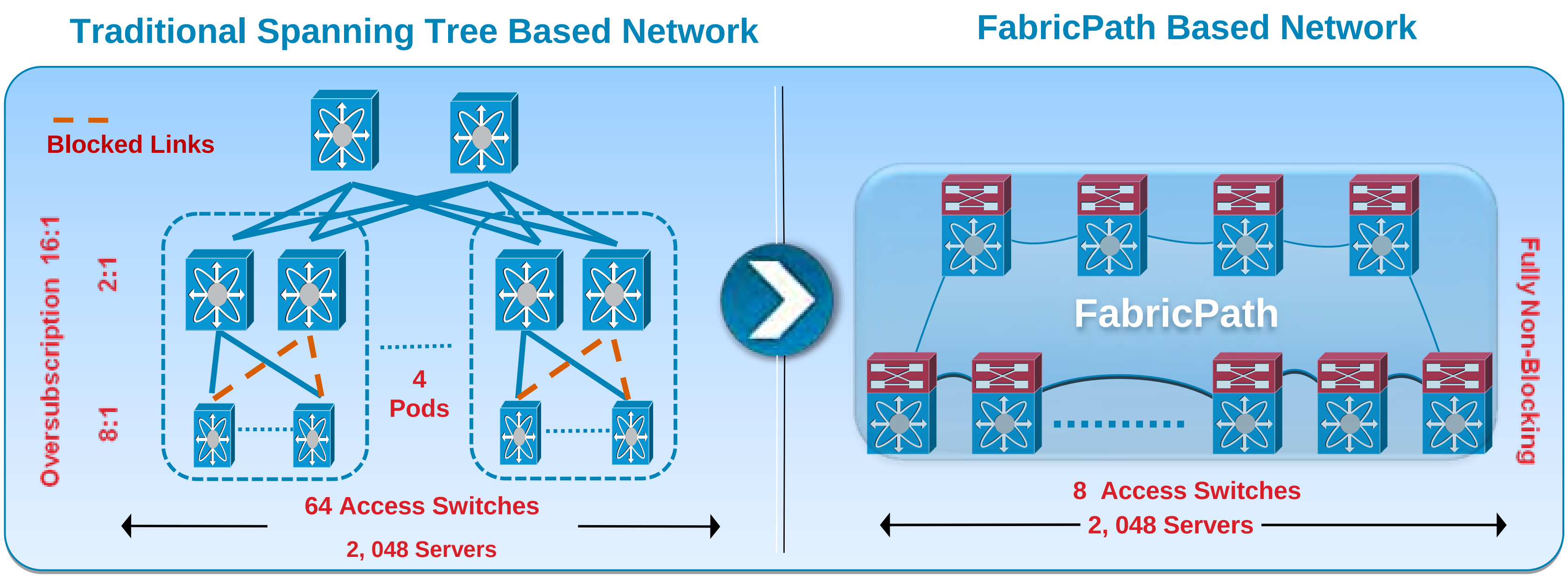
Allows Tier Consolidation



Scaling with FabricPath

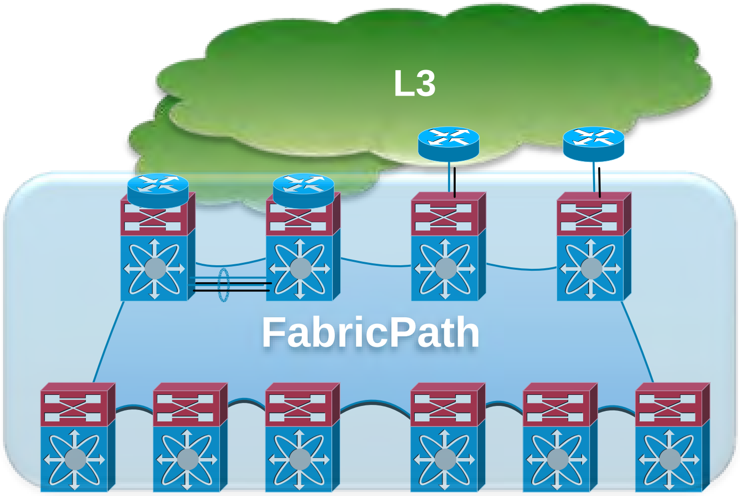
Example: 2,048 X 10GE Server Design

- 16X improvement in bandwidth performance
- 6 to 1 consolidation (from 74 managed devices to 12 devices)
- 2X+ increase in network availability
- Simplified IT operations
  - fewer devices, VLANs anywhere
  - management simplicity



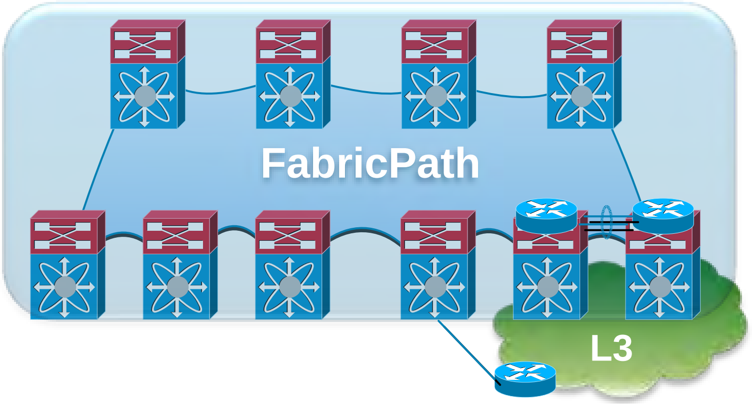
Layer 3 Integration – Integrate at the Spine

- Straightforward with two spine switches
- Considerations with more than two spines:
  - HSRP: Traffic polarized to spines on a per VLAN basis (South-North)
  - GLBP to distribute servers to different default gateways
  - Anycast FHRP future solution



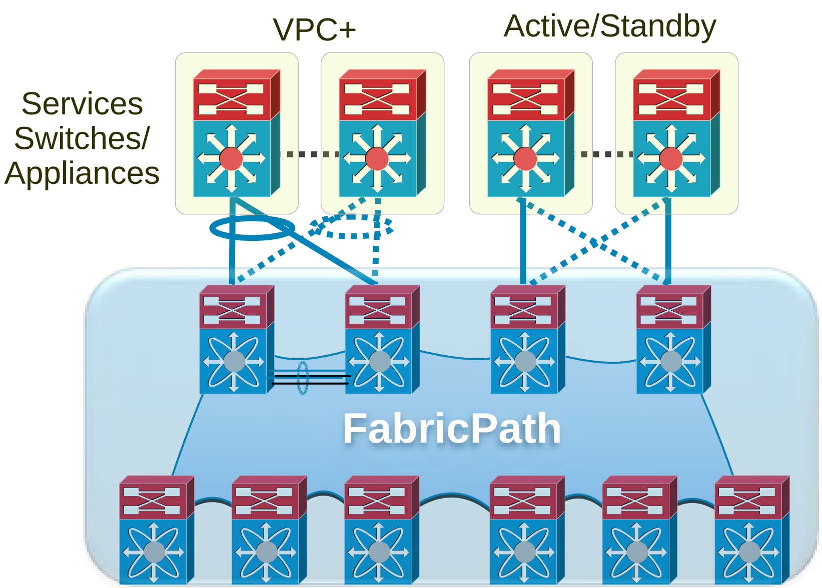
Layer 3 Integration – Integrate at the Leaf

- Provides a "cleaner" spine design
- Traffic distributed equally across spines (no hot spots)
- Increased number of hops to reach gateway (latency)



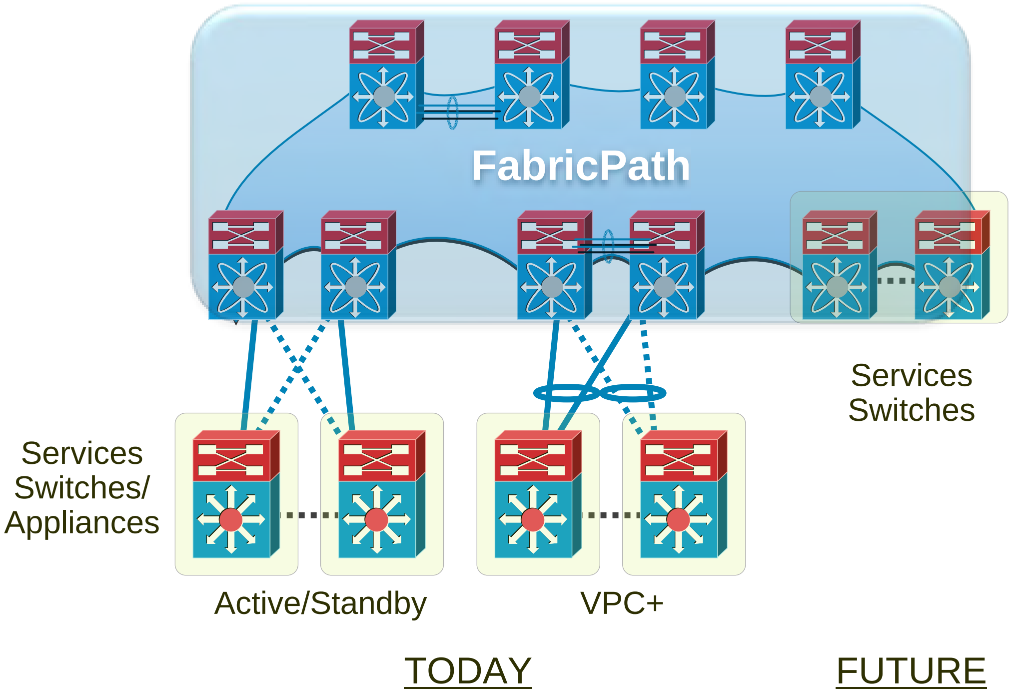
Services Integration – Integrate at the Spine

- Meshes well with current design practices
- Considerations with more than two spines:
  - Forwarding "just works" – MAC and SID forwarding gets traffic to correct spine(s)
  - Consider possible bandwidth constraints – traffic requiring services vectors through only those spines where services connected



Services Integration – Integrate at the Leaf

- Provides a "cleaner" spine design
- Ensures no bottleneck at spine (bottleneck at services nodes still possible)
  - Traffic distributed equally across all spines



Agenda

- Introduction to FabricPath
- FabricPath Technical Overview
- Unicast Forwarding
- Multicast Forwarding
- FabricPath Design
- Conclusion

Conclusion

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)
- FabricPath provides design flexibility
  - New routing and services deployment options

Conclusion

- Thank you for your time today!
- You should now have a thorough understanding of FabricPath concepts, technology, and design considerations!
- **Any questions?**

Reference: Acronym Decoder

- ACL–Access Control List
- ADJ–Adjacency
- ASIC–Application Specific Integrated Circuit
- CE–Classic Ethernet
- CF–Compact Flash
- CLI–Command Line Interface
- CMP–Connectivity Management Processor (lights-out)
- CoPP–Control Plane Policing
- COS–Class of Service
- CP–Control Processor (main CPU)
- DCB–Data Center Bridging
- DCI–Data Center Interconnect
- DSCP–Differentiated Services Code Point
- ECMP–Equal Cost Multi Path
- EOBC–Ethernet Out-of-Band Channel
- ETS–Enhanced Transmission Selection
- FCoE–Fiber Channel over Ethernet
- FE–Forwarding Engine
- FEX–Fabric Extender (Nexus 2000 family)
- FIB–Forwarding Information Base
- FP–FabricPath
- FRU–Field Replaceable Unit
- FTAG–Forwarding Tag
- GM-LSP–Group Membership LSP
- GRE–Generic Route Encapsulation
- HSRP–Hot Standby Router Protocol
- IGMP–Internet Group Management Protocol
- IPC–Inter Process Communication
- IS-IS–Integrated System-to-Integrated System
- LED–Light Emitting Diode
- LID–Local ID
- LOU–Logical Operation Unit
- LSP–Link-State PDU
- MET–Multicast Expansion Table
- NDE–NetFlow Data Export
- OIF–Output Interface
- OIL–Output Interface List
- PACL–Port ACL
- PBR–Policy-Based Routing
- PFC–Priority Flow Control (per-priority pause)
- PIM–Protocol Independent Multicast
- POD–Pool of Devices
- QoS–Quality of Service
- RACL–Router ACL
- RE–Replication Engine
- RPF–Reverse Path Forwarding
- RU–Rack Unit
- SFP+–10G-capable Small-Formfactor Pluggable
- SID–Switch ID
- SoC–System-on-chip/switch-on-chip
- sSID–Sub-Switch ID
- STP–Spanning-Tree Protocol
- SVI–Switched Virtual Interface (VLAN interface)
- TCAM–Ternary CAM
- TLV–Type, Length, Value
- TRILL–Transparent Interconnection of Lots of Links
- uRPF–Unicast RPF
- VACL–VLAN ACL
- VDC–Virtual Device Context
- VOQ–Virtual Output Queuing
- VPC–Virtual Port Channel with Classic Ethernet
- VPC+–Virtual Port Channel with FabricPath
- VRF–Virtual Routing and Forwarding
- VRRP–Virtual Router Redundancy Protocol
- WRED–Weighted Random Early Detection
- WRR–Weighted Round Robin
- XL–Refers to forwarding engine with larger FIB and ACL TCAMs

Complete Your Online Session Evaluation

- Receive 25 Cisco Preferred Access points for each session evaluation you complete.
- Give us your feedback and you could win fabulous prizes. Points are calculated on a daily basis. Winners will be notified by email after July 22nd.
- Complete your session evaluation online now (open a browser through our wireless network to access our portal) or visit one of the Internet stations throughout the Convention Center.
- Don’t forget to activate your Cisco Live and Networkers Virtual account for access to all session materials, communities, and on-demand and live activities throughout the year. Activate your account at any internet station or visit [www.ciscolivevirtual.com.](http://www.ciscolivevirtual.com)

Learn. Connect. Collaborate. together.

Visit the Cisco Store for Related Titles

http://theciscostores.com

Thank you.