ıılıılı cısco

# **Cisco IOS Advantage Webinars**

# Efficient Data Center Design with FabricPath

**Babi Seal and Patrick Warichet** 

We'll get started a few minutes past the top of the hour.

Note: you may not hear any audio until we get started.

### **Speakers & Panelists Introduction**



### Housekeeping

- Submit questions in Q&A panel and send to "All Panelists" Avoid CHAT window for better access to panelists
- For Webex audio, select COMMUNICATE > Join Audio Broadcast
- For Webex call back, click ALLOW Phone button at the bottom of Participants side panel
- Where can I get the presentation? Or send email to: <u>ask\_iosadvantage@cisco.com</u>
- Please complete the post-event Survey
- Join us on September 5 for our next IOS Advantage Webinar: A Closer Look: Comparing Benefits of EIGRP and OSPF www.cisco.com/go/iosadvantage

Communicate

Info

Ouick Start

Participant

+& -

EOAM-C

Share Viev

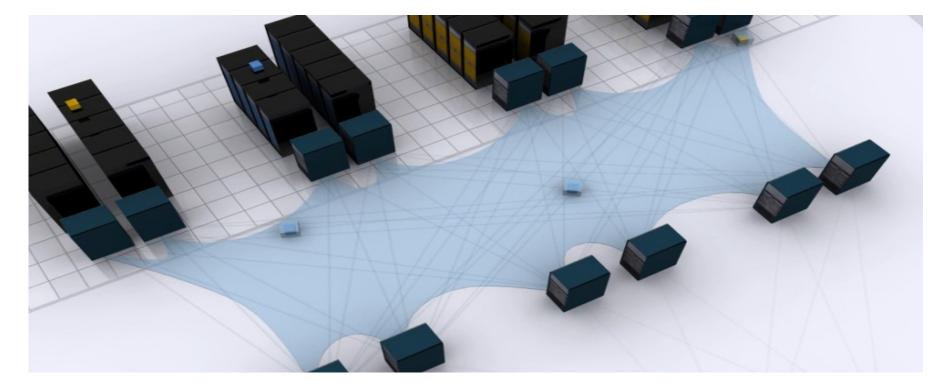
4

Chat

### Agenda

- Why Layer 2 in the Data Center?
- FabricPath solution overview
- Inside the Fabric
- FabricPath Unicast Details
- FabricPath Multicast Details
- FabricPath designs





# Why Layer 2 in the Data Center?

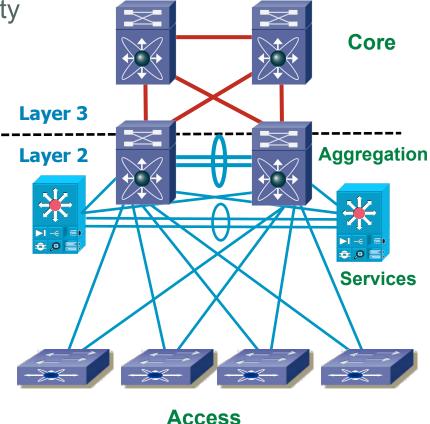


# Why Layer 2 in the Data Center?

#### **Key Decision Factors**

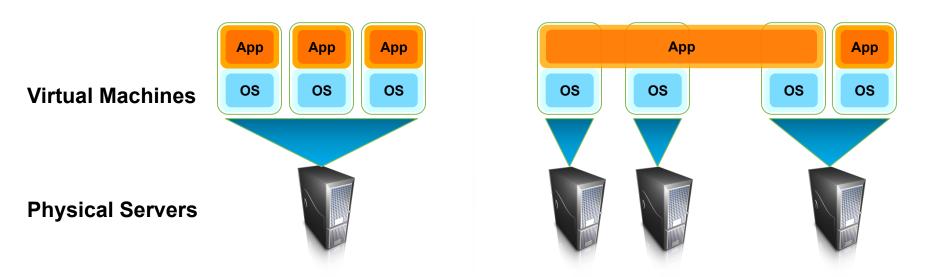
- Some protocols rely on the functionality
- Simple, almost plug and play
- No addressing
- Required for implementing subnets
- Allows easy server provisioning
- Allows virtual machine mobility

### So what changed ?

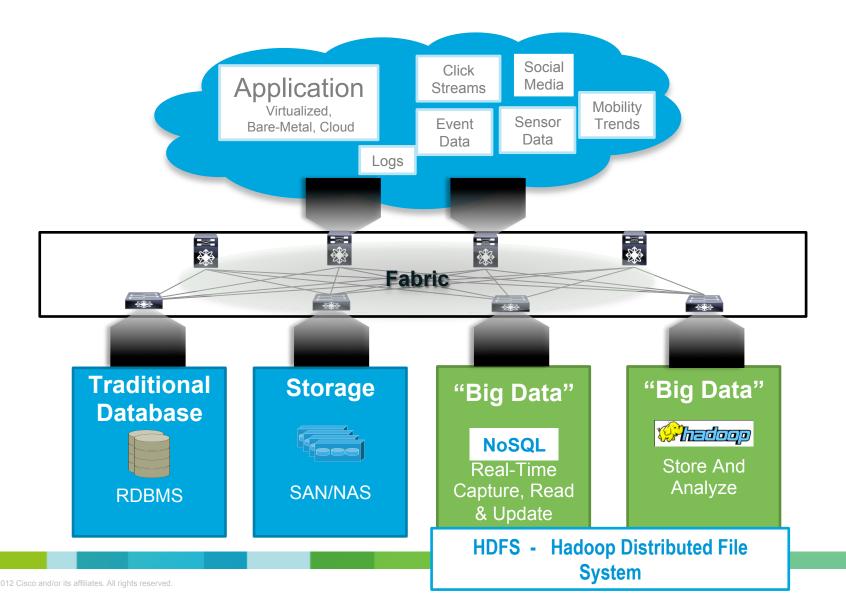


#### Change # 1 – Workload Virtualization Flexibility & Provisioning

- Partitioning
- Physical devices partitioned into virtual devices
- <u>Clustering</u>
- Applications distributed across multiple servers

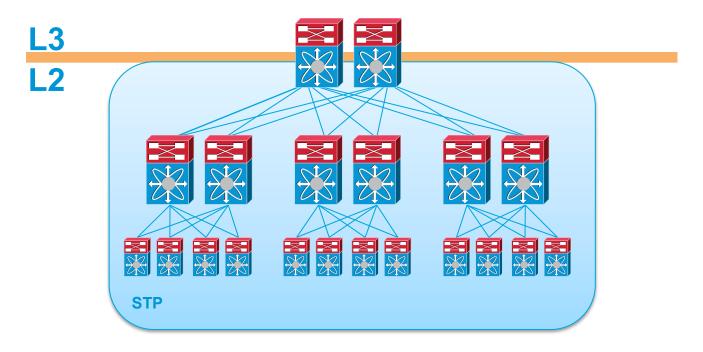


#### Change # 2 – Web 2.0 & Big Data Big Data – Hadoop, NoSQL & HDFS



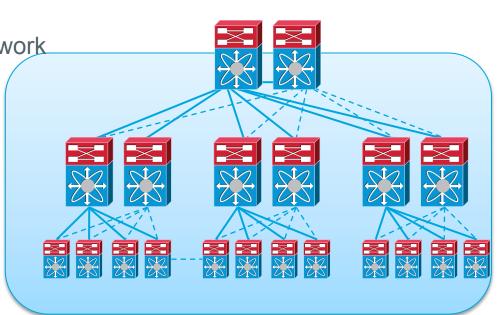
# Possible Solution for End-to-End L2?

#### Just extend STP to the whole network

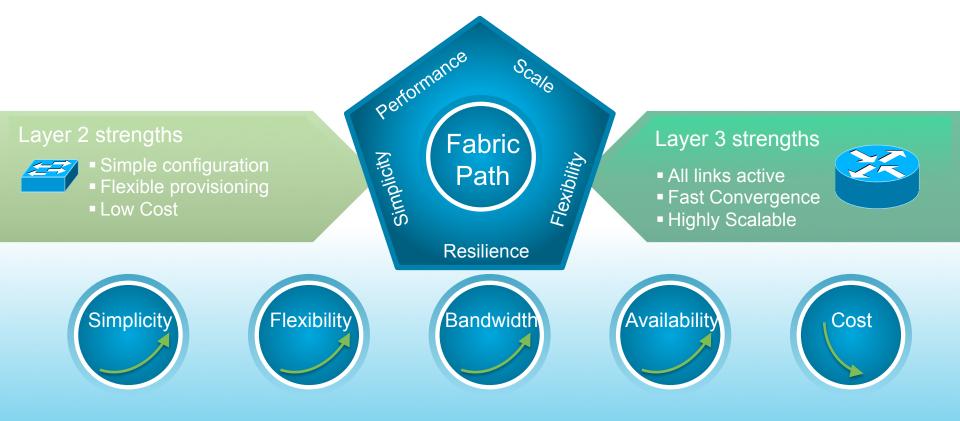


# **Some Layer 2 Limitations**

- 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.



### Introducing Cisco FabricPath An NX-OS Innovation Enhancing L2 with L3



"The FabricPath capability within Cisco's NX-OS offers dramatic increases in network scalability and resiliency for our service delivery data center. FabricPath extends the benefits of the Nexus 7000 in our network, allowing us to leverage a common platform, simplify operations, and reduce operational costs."

Mr. Klaus Schmid, Head of DC Network & Operating, T-Systems International GmbH



### FabricPath NX-OS Solution Roadmap

	CY2011	1H CY2012	1H CY2013
Nexus 7000	<ul> <li>Scale: 128 Switch IDs</li> <li>FabricPath on F2 I/O modules <ul> <li>Introducing native Fabric Extender support</li> <li>No conversational learning</li> </ul> </li> </ul>	<ul> <li>vPC+ on Fabric Extender host interfaces</li> <li>Conversational learning on F2</li> <li>PVLAN on F2 (FabricPath and CE)</li> <li>FabricPath Traceroute (PONG)</li> </ul>	<ul> <li>F1/M1 L2 proxy learning</li> <li>Multiple topologies</li> <li>Overload bit (for least disruptive convergence)</li> <li>route-map and mesh groups for scale</li> <li>Vlan pruning enhancement (based on CE forwarding ports)</li> <li>Support on F2 10GbaseT I/O Modules</li> <li>Affinity to the closest rooted tree for multidestination traffic</li> <li>Anycast HSRP</li> <li>Scale 4K vlans, 512 switch IDs</li> </ul>
Nexus 5500	<ul> <li>Cisco FabricPath</li> <li>16-Way ECMP</li> <li>128 Switch IDs</li> <li>STP Boundary &amp; Termination</li> <li>vPC+ (FabricPath Boundary)</li> <li>L2 proxy learning</li> <li>PIM SSM over vPC+</li> </ul>	SHIPPING NX-OS 5.2 • Multi-Topology Support (2)	<ul> <li>Anycast HSRP</li> <li>Overload bit (for least disruptive convergence)</li> <li>FabricPath Traceroute (PONG)</li> </ul>

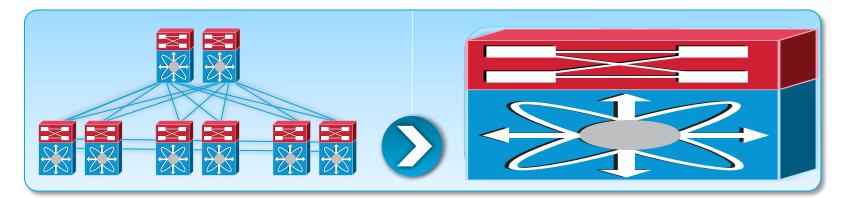


# FabricPath Solution Overview

			Agenda	Back	Е	
			Why Layer 2 in the Data Center?			
			FabricPath Solution Overview			
			Inside the Fabric			
			FabricPath Unicast Details           FabricPath Multicast Details		tails	
					etails	
			[	abricPath Design	S	

# FabricPath, an Ethernet Fabric

#### Shipping Since 2010: turn your network into a Fabric

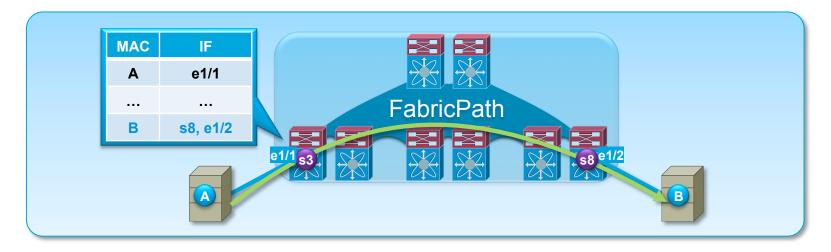


- Connect a group of switches using an arbitrary topology
- With a simple CLI, aggregate them into a Fabric:

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

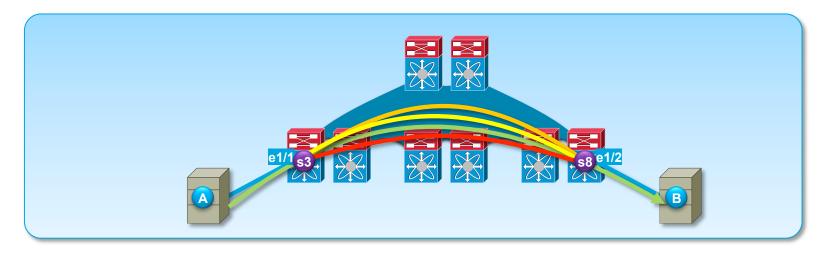
 An open protocol based on Layer 3 technology provides Fabric-wide intelligence and ties the elements together

# **Optimal, Low Latency Switching**



- 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 connectivity any to any (as if it was the same switch, no STP inside)

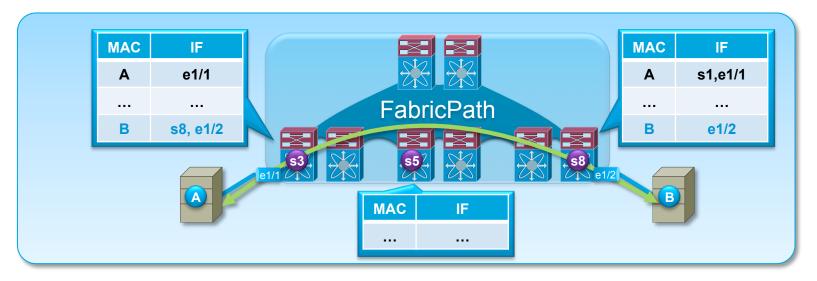
### High Bandwidth, High Resiliency Equal Cost MultiPathing (ECMP)



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

# Mac Address Table Scale

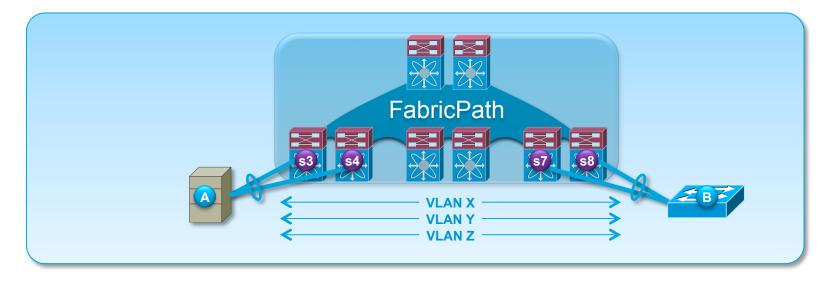
#### **Conversational Learning**



• 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

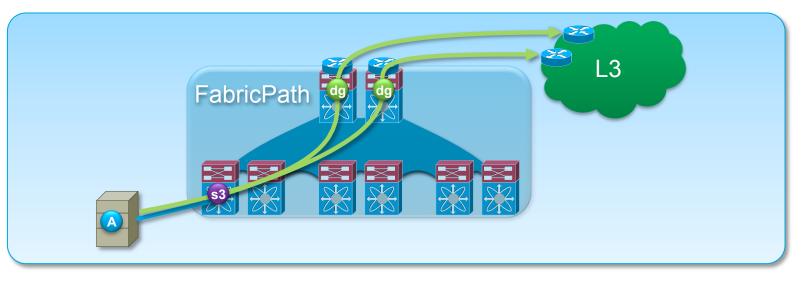
# Layer 2 integration vPC+



- 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 Devices Integration**

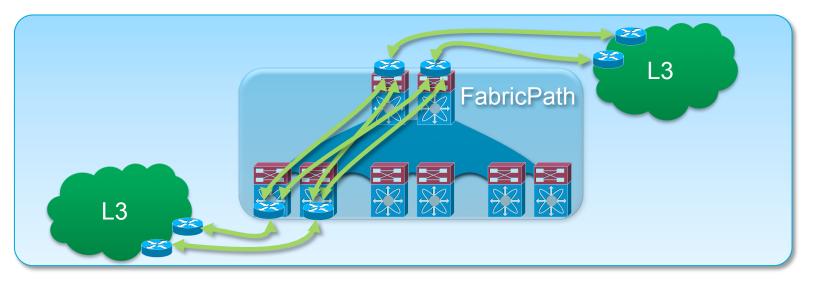
Hosts Can Leverage Multiple Active L3 Default Gateways



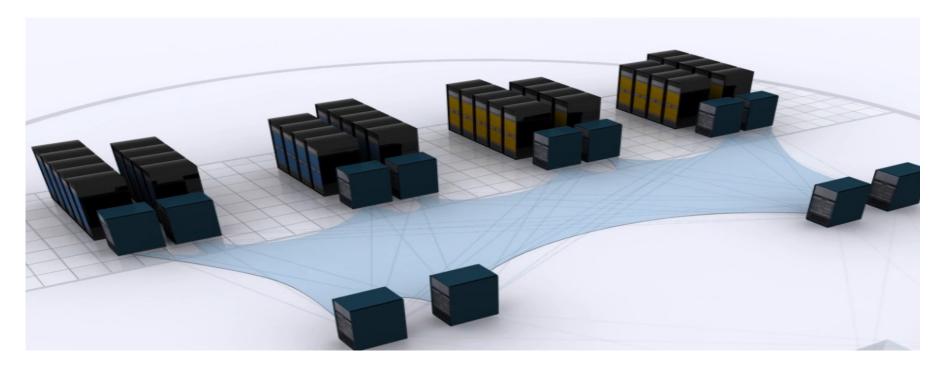
- Hosts see a single default gateway
- The fabric provide them transparently with multiple simultaneously active default gateways
- Allows extending the multipathing from the inside to the fabric to the L3 domain outside the fabric

# Layer 3 Integration

### Large tables (XL Hardware), SVIs Anywhere



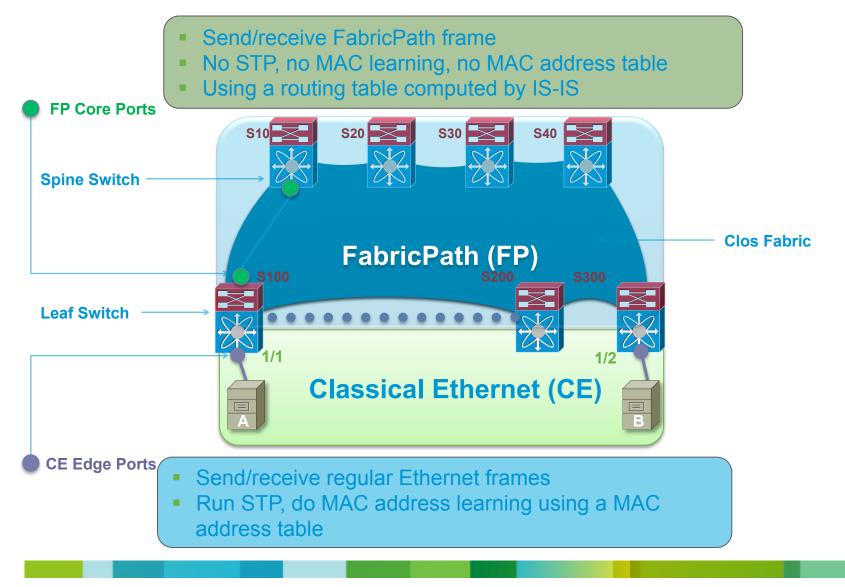
- The fabric provides seamless L3 integration
- An arbitrary number of routed interfaces can be created at the edge or within the fabric
- Attached L3 devices can peer with those interfaces
- The hardware is capable of handling million of routes



# Inside the Fabric

				Agenda	Back	End	
				Why Layer 2 in the Data Center?			
				Fabri	cPath Solution Ove	erview	
				Fab	ricPath Unicast De	etails	
				Fabr	icPath Multicast D	etails	
					FabricPath Design	S	

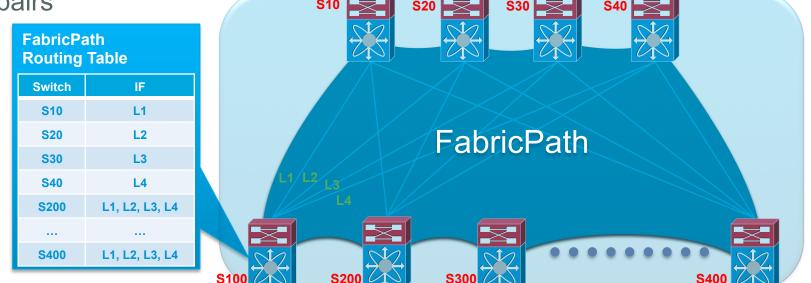
# FabricPath & Clos Fabric Terminology



# **New Control Plane**

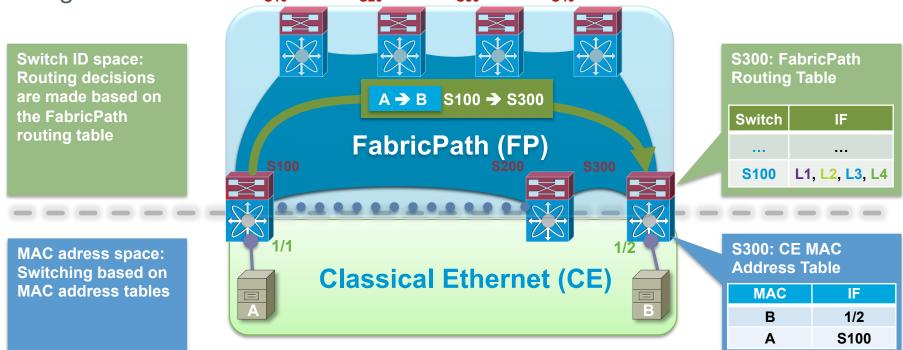
### Plug-n-Play L2 IS-IS Manages Forwarding Topology

- IS-IS assigns addresses to all FabricPath switches automatically
- Compute shortest, pair-wise paths



# **New Data Plane**

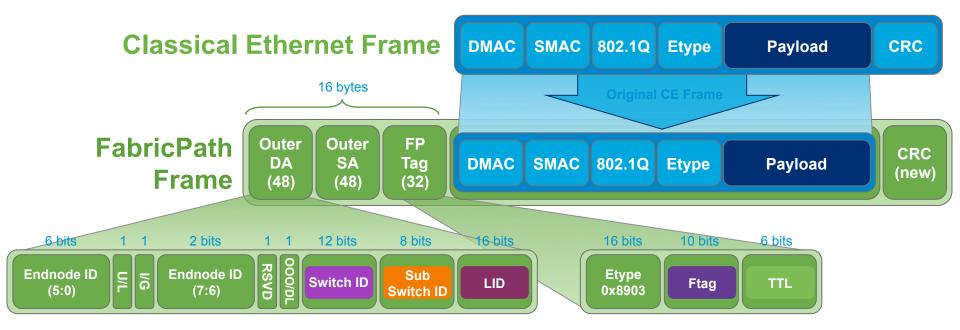
The association MAC address/Switch ID is maintained at the edge
 \$10
 \$20
 \$30
 \$40



• Traffic is encapsulated across the Fabric

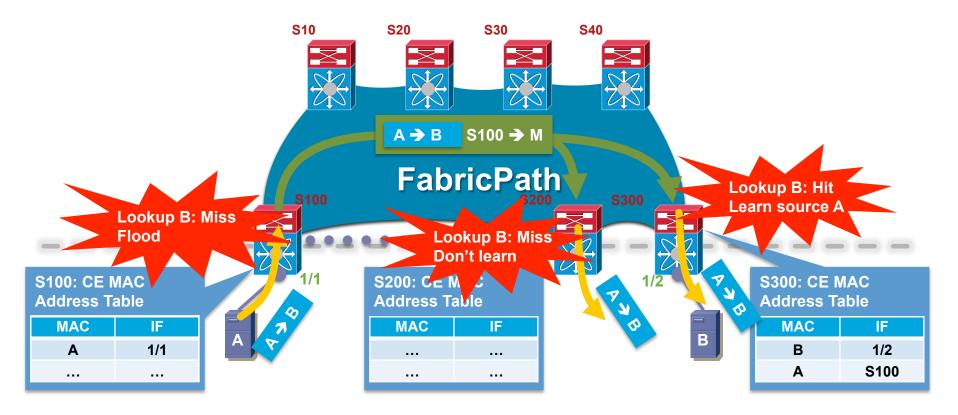
# FabricPath Encapsulation

### 16-Byte MAC-in-MAC Header



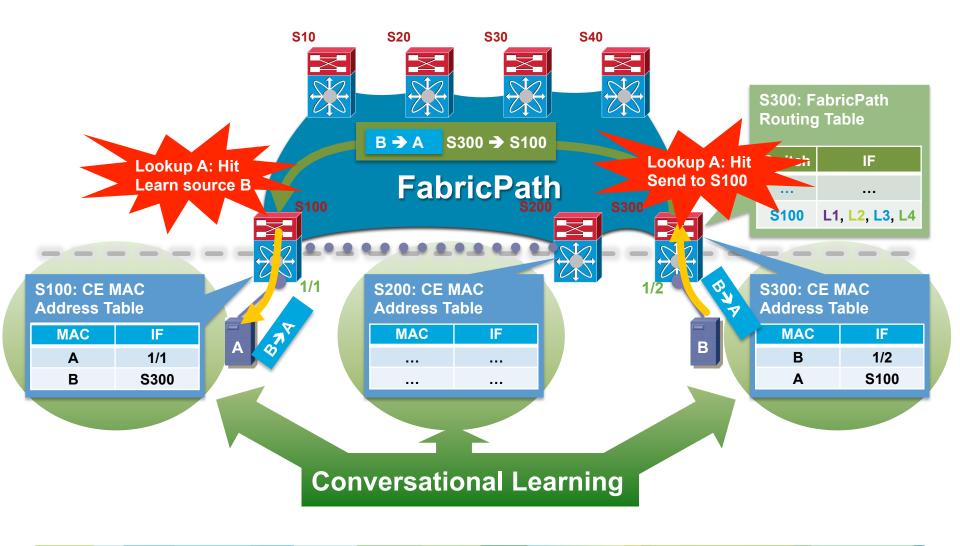
- Switch ID Unique number identifying each FabricPath switch
- Sub-Switch ID Identifies devices/hosts connected via VPC+
- LID Local ID, identifies the destination or source interface
- Ftag (Forwarding tag) Unique number identifying topology and/or distribution tree
- **TTL** Decremented at each switch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

### **Unknown Unicast**



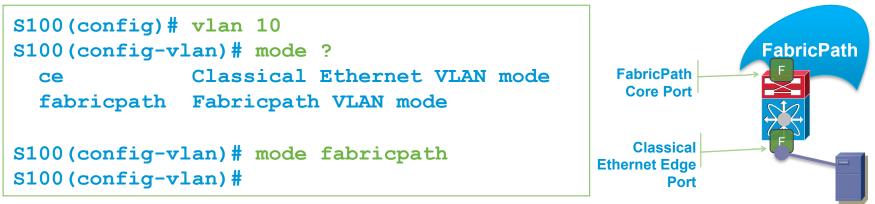
**Classical Ethernet** 

# Known Unicast, Conversational Learning



# 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:

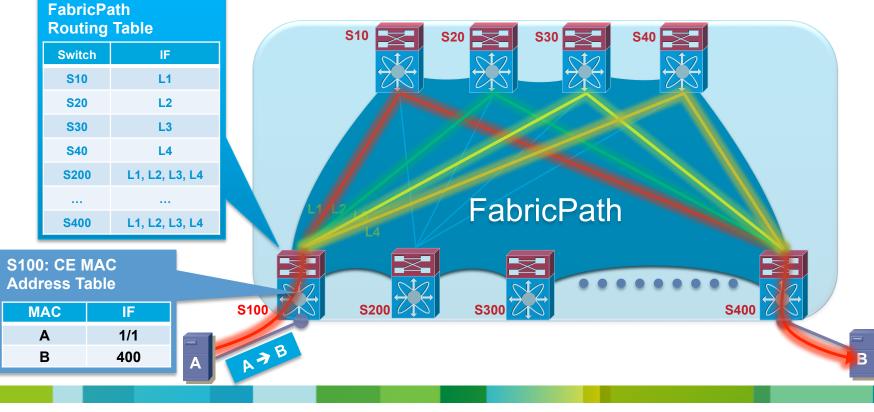


- FabricPath VLANs can only be enabled on F modules
- FabricPath VLANs are also relevant to the Nexus 5500

# **Equal Cost Multipathing**

### Traffic Forwarding Based on a Routing Table

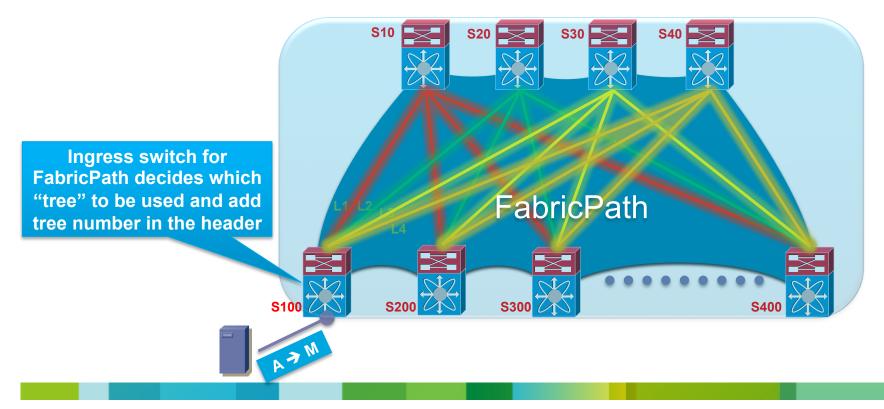
- IS-IS assigns addresses to all FabricPath switches automatically
- Compute shortest, pair-wise paths
- Support equal-cost paths between any FabricPath switch pairs



# **Multicast Traffic**

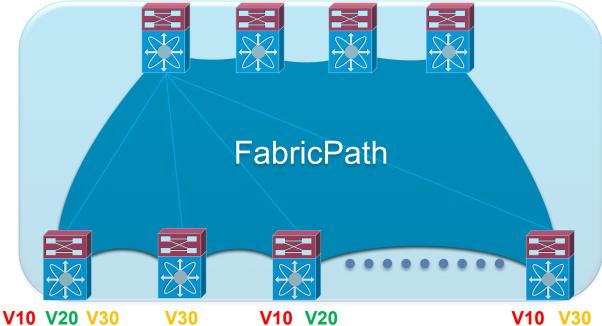
#### Load Balancing on a Per-Tree Basis

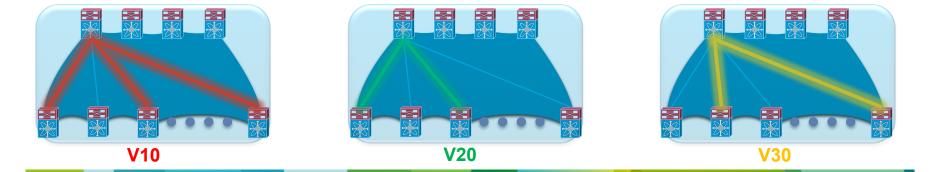
- IS-IS computes several trees automatically
- Location of the root switches can be configured
- Multicast traffic is pinned to a tree at the edge



# **VLAN Pruning**

#### Automatically Handled by IS-IS

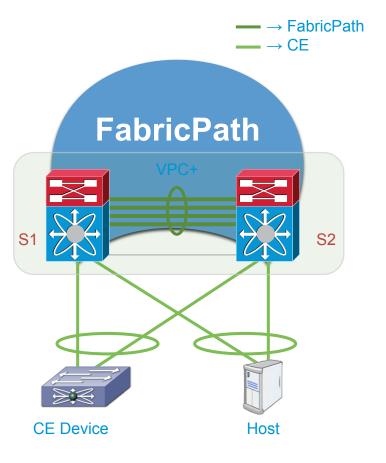




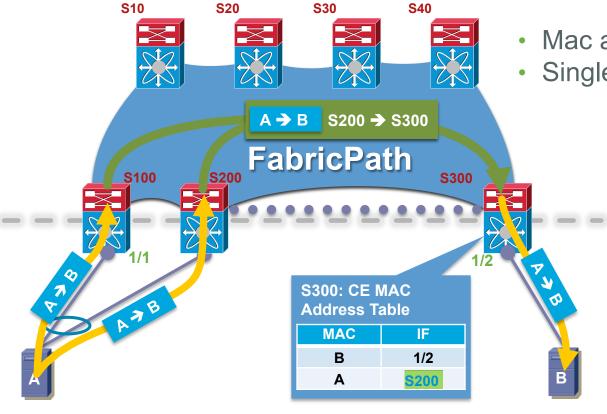
### VPC+

#### Virtual Port Channel in FabricPath Environment

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



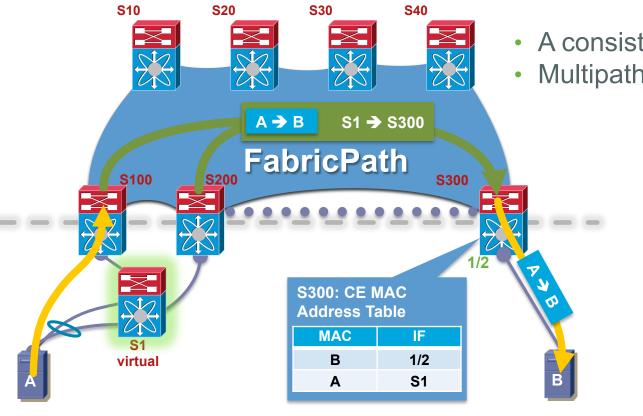
### **VPC+ Technical Challenges**



**Classical Ethernet** 

- Mac address flapping on S300
- Single path to A

### **VPC+ Virtual Switch**



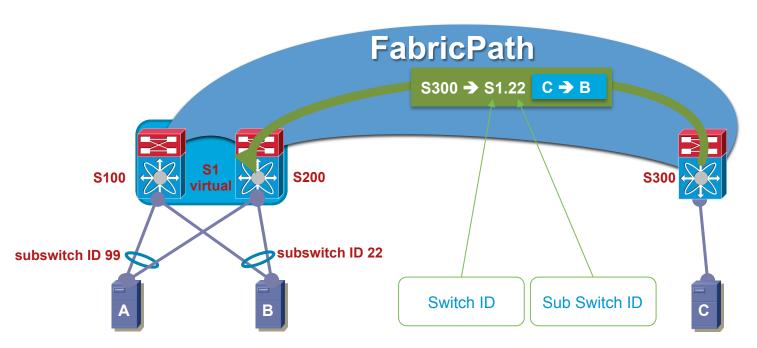
- A consistently associated to S1
- Multipathing to A

**Classical Ethernet** 

# Sub-Switch ID

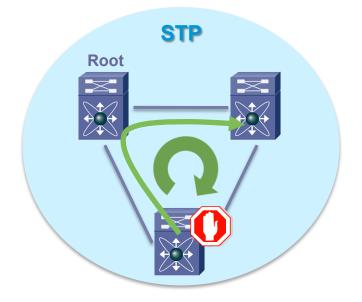
### Identifies a VPC Off a Virtual Switch



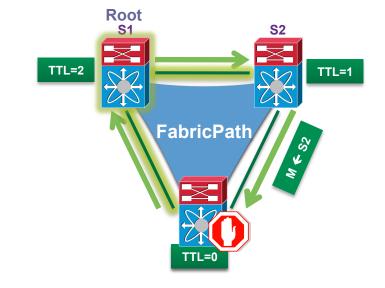


# Loop Mitigation with FabricPath

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

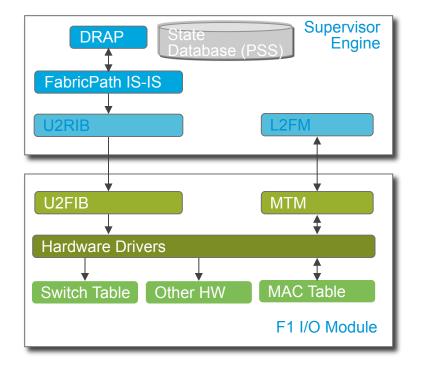


## **Unicast Forwarding Details**

		Agenda	Back	End
		Why La	yer 2 in the Data (	Center?
		Fabric	Path Solution Ove	erview
			Inside the Fabric	
		Fabr	icPath Unicast De	tails
		Fabri	cPath Multicast De	etails
		F	abricPath Designs	3

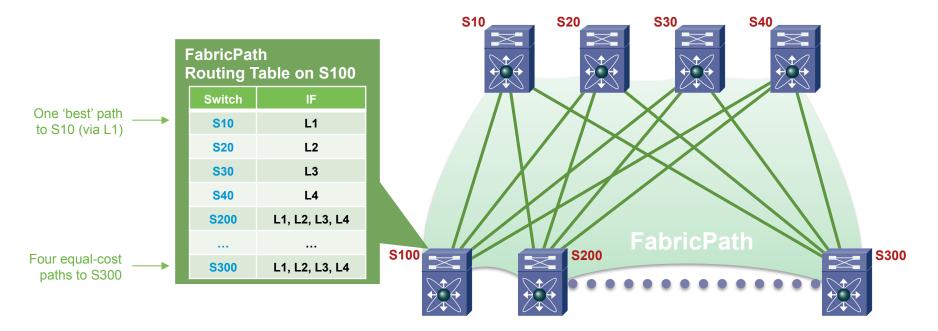
## FabricPath Unicast System Architecture

- 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

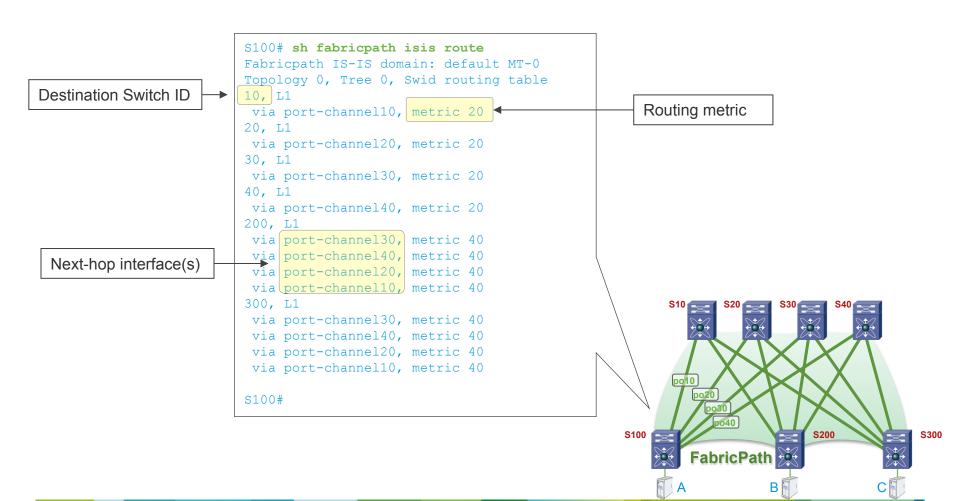


## FabricPath Routing Table

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

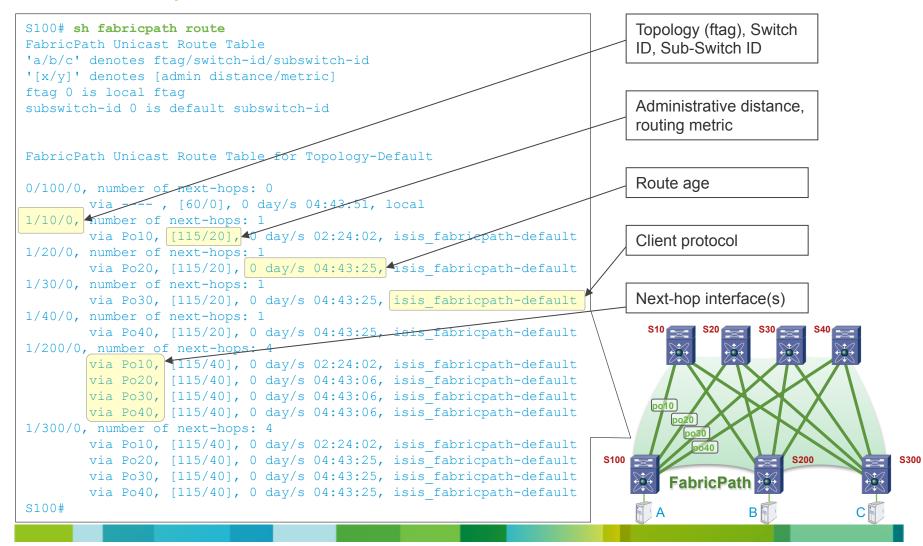


# Display IS-IS View of Routing Topology show fabricpath isis route

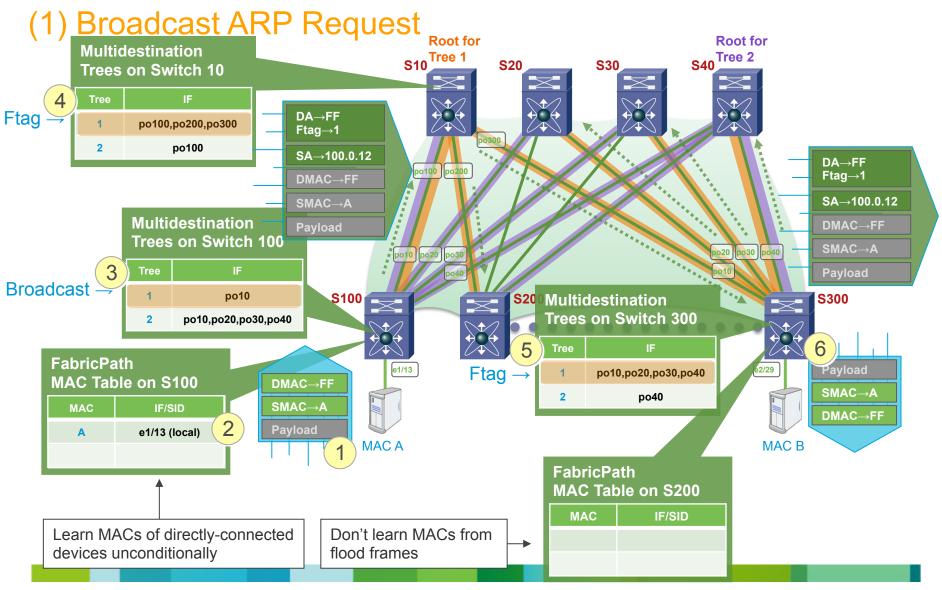


#### **Display U2RIB View of Routing Topology**

#### show fabricpath route



#### Putting It All Together – Host A to Host B

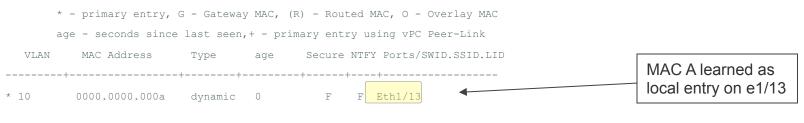


#### MAC Address Tables After Broadcast ARP

• S100:

#### S100# sh mac address-table dynamic

#### Legend:



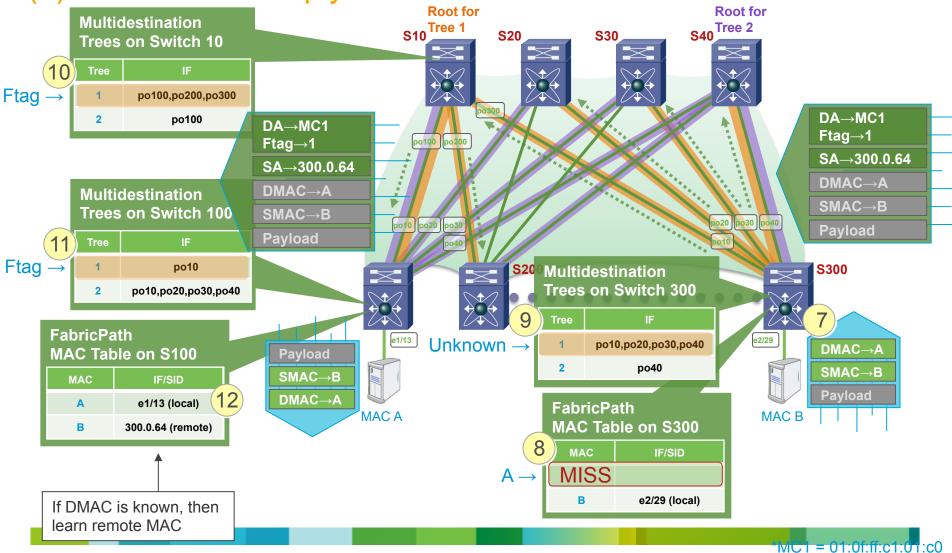
S100#

#### • S10 (and S20, S30, S40, S200):

#### S10# sh mac address-table dynamic Legend: \* - primary entry, G - Gateway MAC, (R) - Routed MAC, O - Overlay MAC MAC A not learned age - seconds since last seen, + - primary entry using vPC Peer-Link on other switches VLAN MAC Address Type age Secure NTFY Ports/SWID.SSID.LID S10# 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

#### Putting It All Together – Host A to Host B

#### (2) Unicast ARP Reply

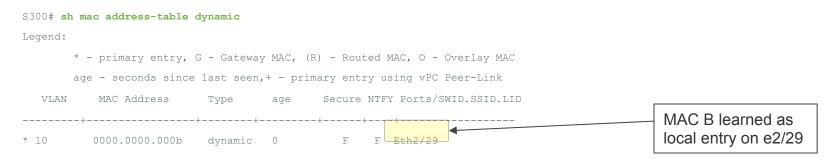


#### MAC Address Tables After Unicast ARP Reply

#### • S100:

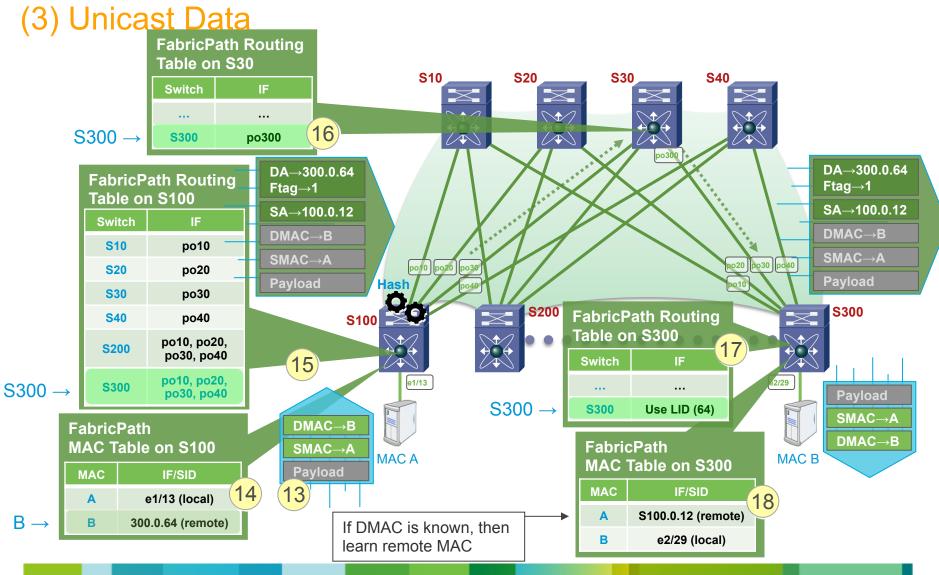
S100# <b>sh</b>	mac address-table (	dynamic				
Legend:						
*	- primary entry,	G - Gateway	/ MAC, (	R) – Rout	ted MAC, O - Overlay MAC	
ē	ge - seconds since	last seen,	+ - pri	mary ent:	ry using vPC Peer-Link	
VLAN	MAC Address	Туре	age	Secure	NTFY Ports/SWID.SSID.LID	
	+	++-		-+	++	
* 10	0000.0000.000a	dynamic	90	F	F Eth1/13	S100 learns MAC B as
10	0000.0000.000b	dynamic	60	F	F 300.0.64	remote entry reached
						through S300
S100#					L	

#### • S300:



S300#

#### Putting It All Together – Host A to Host B



### **MAC Address Tables After Unicast Data**

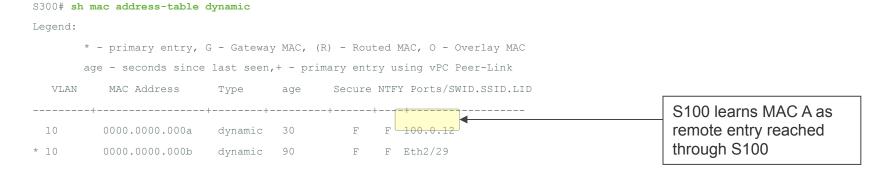
#### • S100:

S100# sh mac address-table dynamic

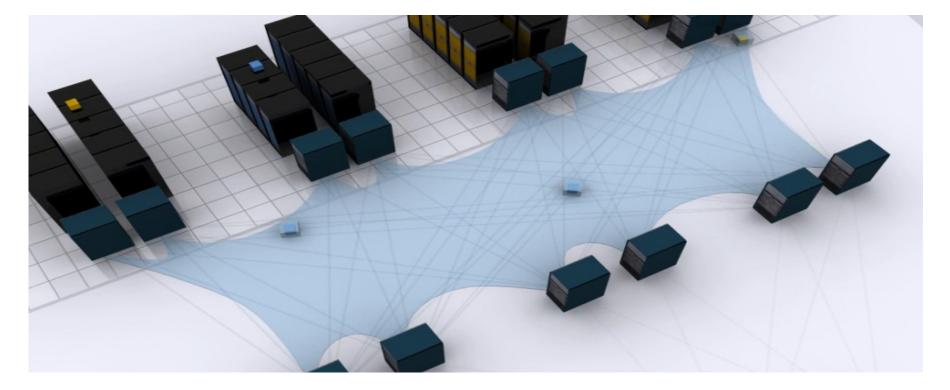
Legend:						
*	- primary entry,	G - Gateway	y MAC, (1	R) – Rout	ed	MAC, O - Overlay MAC
a	ge – seconds since	last seen,	,+ - prin	mary entr	y u	sing vPC Peer-Link
VLAN	MAC Address	Туре	age	Secure	NTF	Y Ports/SWID.SSID.LID
	+	++-		-++		-+
* 10	0000.0000.000a	dynamic	90	F	F	Eth1/13
10	0000.0000.000b	dynamic	60	F	F	300.0.64

S100#

#### • S300:



S300#



## **Multicast Forwarding Details**

FabricPath Unicas       FabricPath Multicas					Agenda	Back	End
Inside the Fat FabricPath Unicas FabricPath Multicas					Why L	ayer 2 in the Data	Center?
FabricPath Unicas       FabricPath Multicas					Fabr		
FabricPath Multicas						Inside the Fabric	
Fabric Path Design of the State						FabricPath Design	

## FabricPath IP Multicast

#### Control plane:

Build several multidestination trees Run IGMP snooping on FabricPath edge switches Advertise receivers location with dedicated LSPs

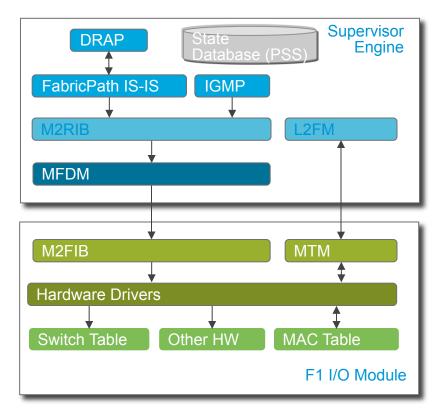
#### • Data plane (hardware):

Selects which multidestination tree or each flow based on hash function

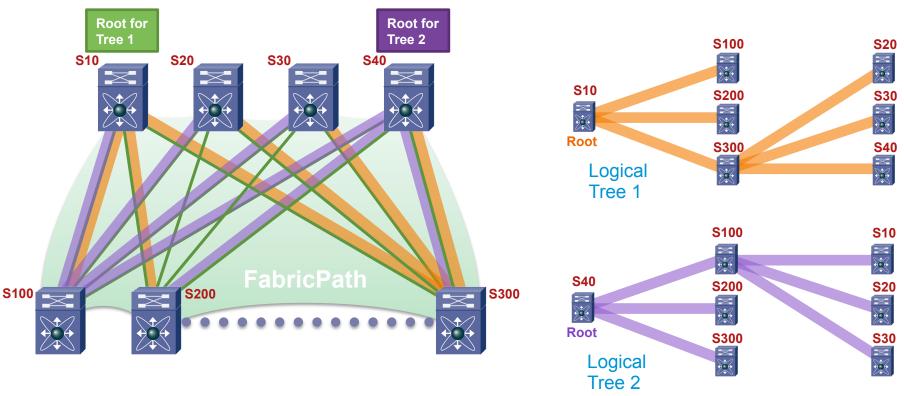
Forward traffic along selected tree

#### 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

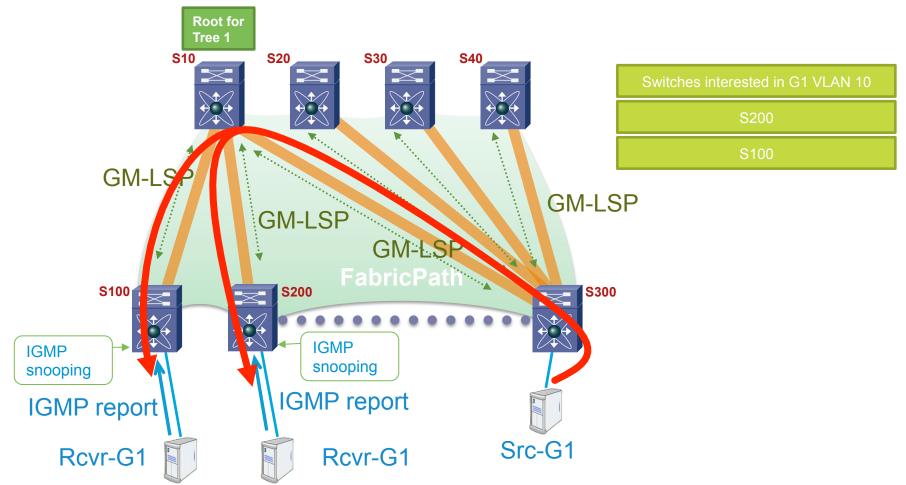


### **Multicast Trees Determination**



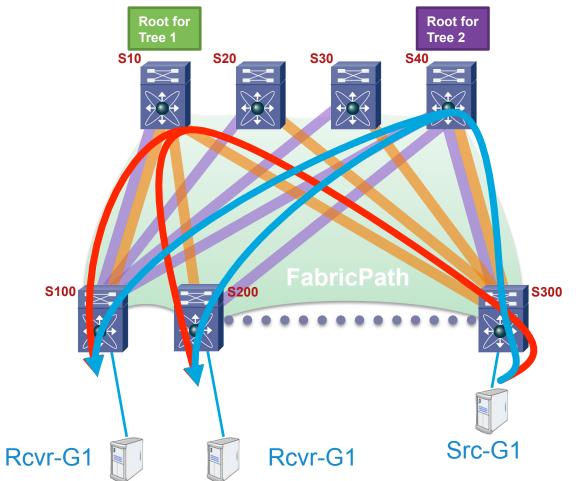
- Switch with highest priority value becomes root for primary tree Highest system ID, then highest Switch ID value, in case of a tie
- Primary root designates different secondary root(s) ensuring path variety.

## **Multicast Tree Pruning**

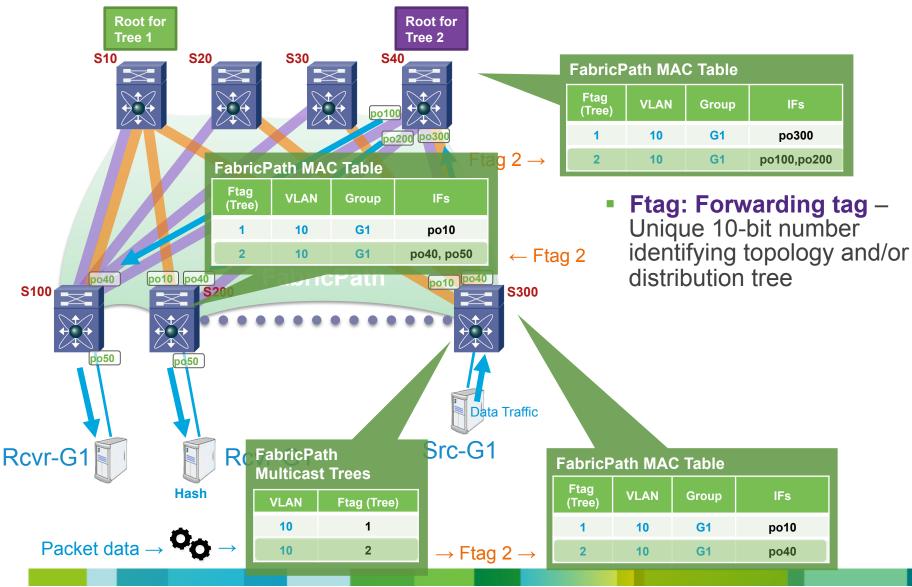


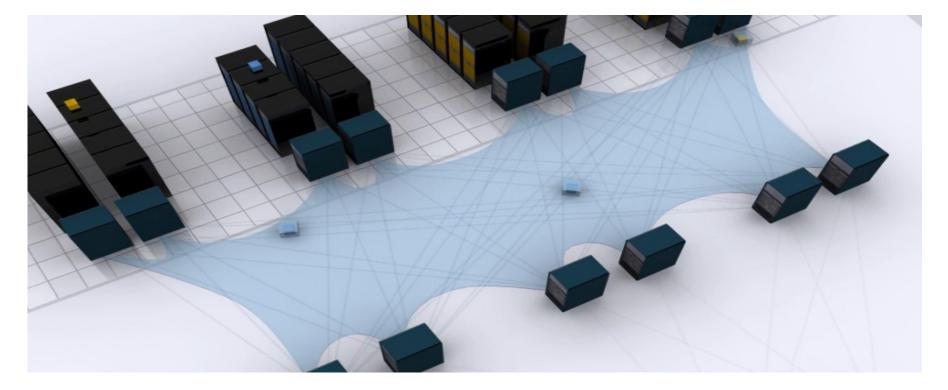
 IS-IS Group Membership LSPs contain multicast forwarding information

## **Multicast Load Balancing**



## Multicast Data Plane Step by Step

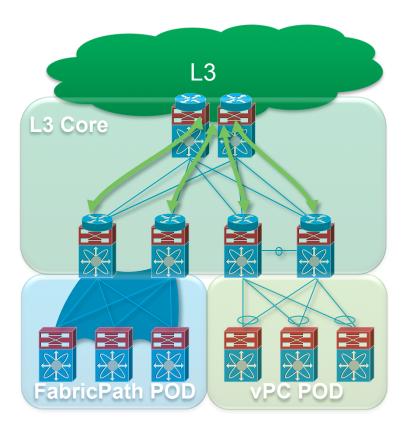




## FabricPath Designs

		Agenda	Back	Enc
		Why La	ayer 2 in the Data C	Center?
		Fabri	cPath Solution Ove	rview
			Inside the Fabric	
		Fab	ricPath Unicast De	tails
		Fabr	icPath Multicast De	etails
			FabricPath Designs	3

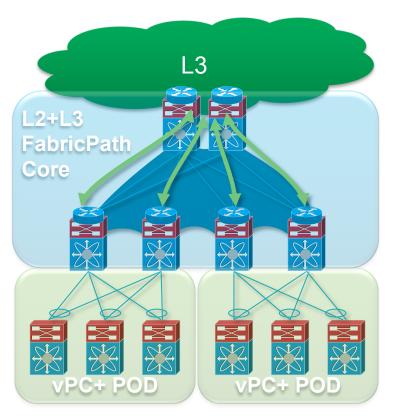
#### Classical POD with FabricPath FabricPath vs. vPC/STP



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

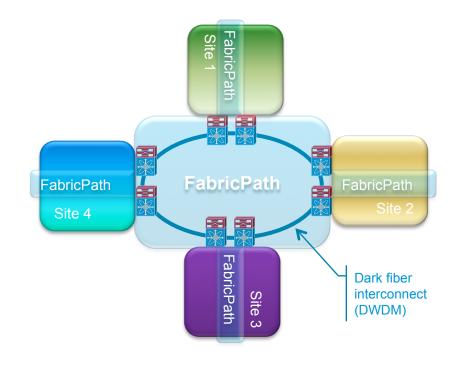
## FabricPath Core

#### **Efficient POD Interconnect**



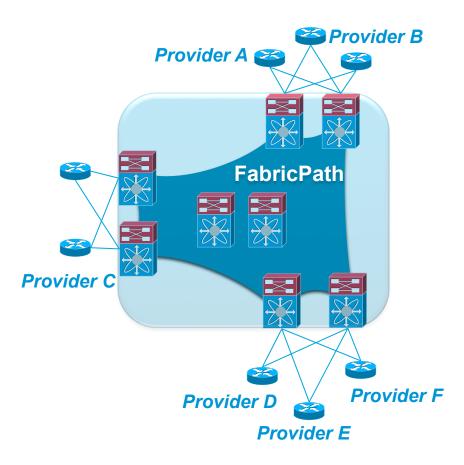
- FabricPath in the Core
- VLANs can terminate at the distribution or extend between PODs.
- STP is not extended between PODs, remote PODs or even remote data centers can be aggregated.
- Bandwidth or scale can be introduced in a non-disruptive way

## 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 table scaling
- VLANs can be selectively extended/terminated

## Internet Exchange Point (IXP)



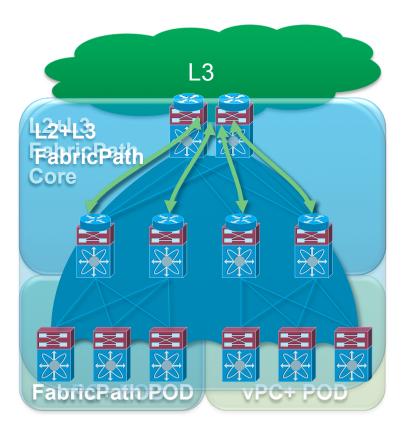
#### **IXP** Requirements

- Layer 2 Peering
- 10GE non-blocking Fabric
- Scale to thousands of ports

#### FabricPath Benefits for IXP

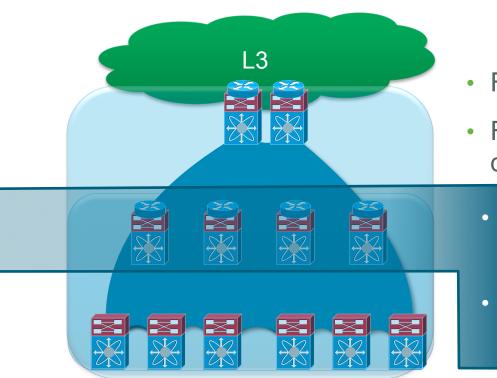
- Layer 2 Fabric
- Non-blocking up to thousands 10GE ports
- Simple to manage
- No design constraint, easy to grow

### **FabricPath Evolution**



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

### Tier Consolidation with FabricPath "Flattening"

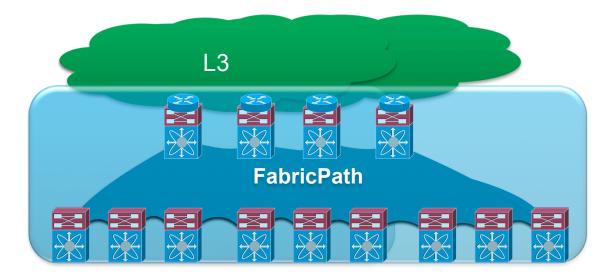


- FabricPath in the Core
- FabricPath extended down to the leaves
  - There is enough bandwidth and port density on the core Nexus 7000s for aggregating the whole network.
  - There is no need for a distribution layer for POD isolation

## FabricPath Flexibility

#### The Network Can Evolve With No Disruption

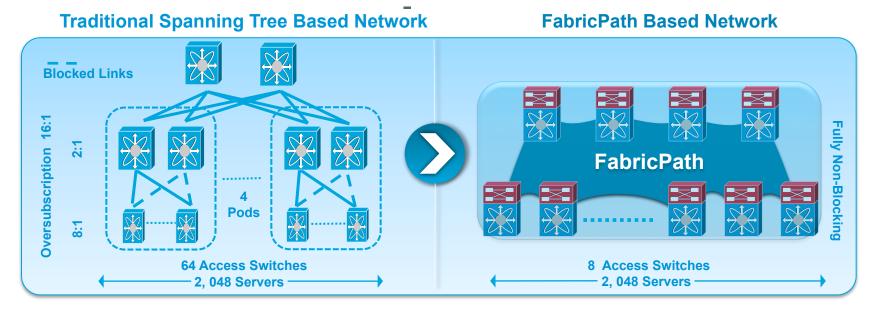
- Need more edge ports?
- Need more bandwidth?
- $\rightarrow$  Add more leaf switches
- $\rightarrow$  Add more links and spines



## Scaling with FabricPath

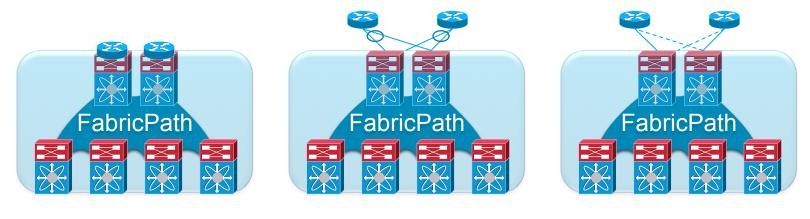
#### Example: 2,048 x 10GE Server Design

- 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)



#### Layer 3/Service Integration, 2 Spine Switches Business as Usual...

• Existing designs and best practices apply easily



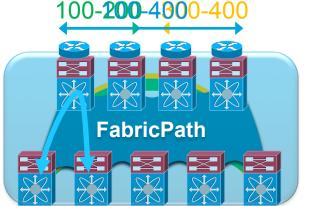
SVIs, active/active

vPC+ external

Active/Standby

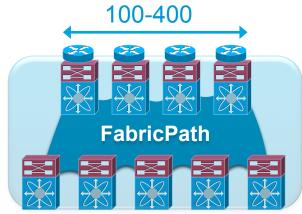
## Layer 3/Service Integration, >2 Spine Switches

#### Trying to adapt the existing model



#### Splo MBPANs

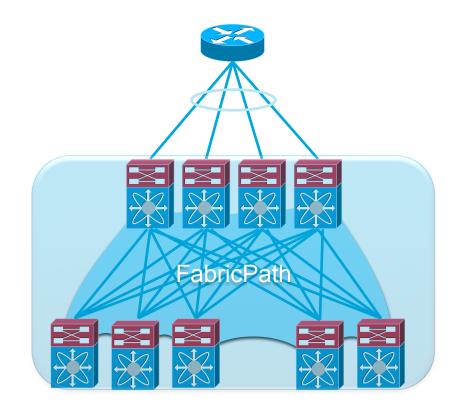
- Bostespoilanizattora
- sinteriel/gate/wayfic can
- bessugceptinhard load balancing



#### Anycast FHRP

- All active
- Available in the future for routing
- Might not be available soon for service modules/appliances

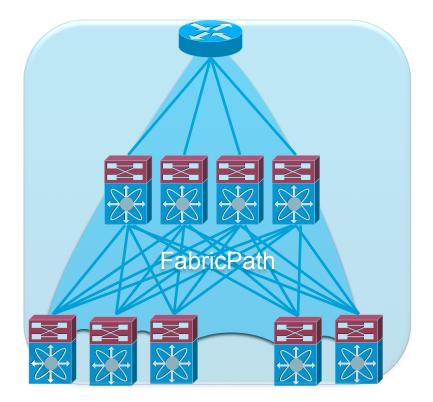
### Other L3/Service Integration Options N-Way VPC+



- Would allow to connect any device supporting port channeling
- Not in the roadmap

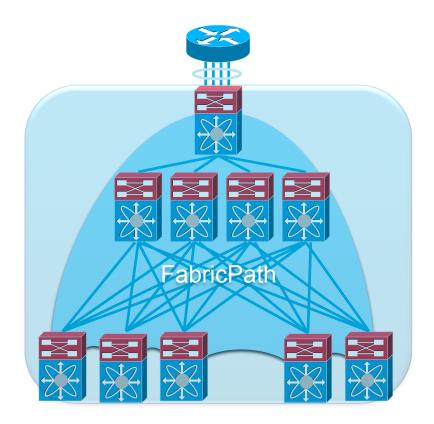
## **Other L3/Service Integration Options**

#### FabricPath Capable Router/Appliance



- Possible with Nexus platforms as routers
- Other appliances not FabricPath capable

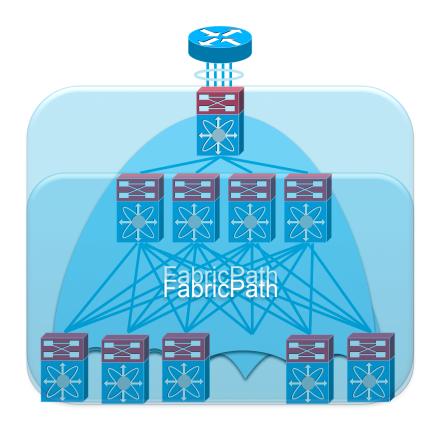
#### Other L3/Service Integration Options Insert a FabricPath Switch



- Allows non-FabricPath capable devices to be attached to the fabric in an optimal way
- 2 hops to reach the router/ appliance

   → Nexus 7000 F series I/O modules and Nexus 5500 have lower latency anyway
- Thanks to FabricPath, enough bandwidth can be provisioned to the router/ appliance

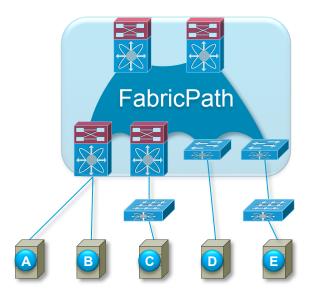
#### Other L3/Service Integration Options Attach to a Leaf Switch



- Allows non-FabricPath capable devices to be attached to the fabric in an optimal way
- 2 hops to reach the router/ appliance

   → Nexus 7000 F series I/O modules and Nexus 5500 have lower latency anyway
- Thanks to FabricPath, enough bandwidth can be provisioned to the router/ appliance

### FabricPath Edge Connectivity Options



1/10G connectivity to Nexus 7000 F1 I/O Module CX-1 cabling provide cost effective solution

- 1/10G connectivity to Nexus 7000 F2 I/O Module
- C 1/10G connectivity to Fabric Extender attached to Nexus 7000 F2 I/O Module
- 1/10G connectivity to Nexus 5500

1/10G connectivity to Fabric Extender attached to Nexus 5500

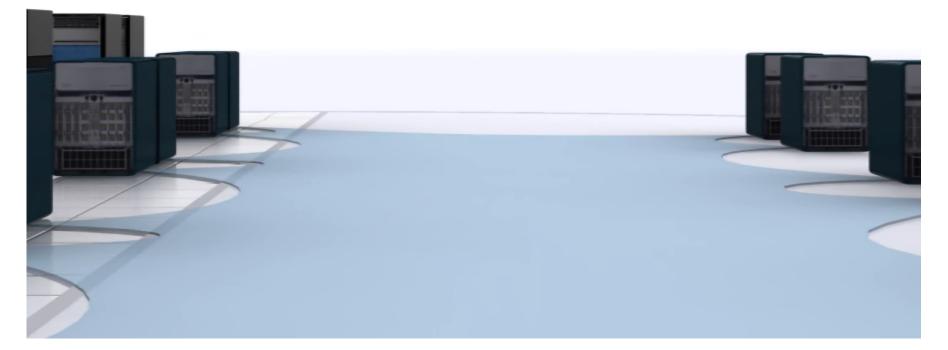


## Conclusion

		Agenda Why Lay	Back yer 2 in the Data Co	End enter?
			Path Solution Over	view
			Inside the Fabric cPath Unicast Deta	ails
			Path Multicast Def	
		F	abricPath Designs	

## **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)



## TRILL



#### TRILL



What is TRILL?

- TRILL, Transparent Interconnection of Lots of Links
- IETF standard for Layer 2 multipathing (since summer 2011)
- Pushed by Cisco but by other competitors too.

Why aren't we doing TRILL already?

- We delivered FabricPath before the standard, and some critical features are still missing to TRILL
- Our FabricPath capable hardware is also TRILL capable
- We will provide TRILL as a FabricPath mode

## 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

Cisco will push FabricPath specific enhancements to TRILL

- Thank you!
- Please complete the post-event survey.
- Join us September 5 for our next webinar:
   A Closer Look: Comparing Benefits of EIGRP and OSPF

To register, go to www.cisco.com/go/iosadvantage

### Thank you.

#