



# Evolving Data Center Switching

## Part 1. Setting the Stage for Layer 2 Multi-pathing (TRILL)

Brad Hedlund  
Cisco Systems, Inc.

v 1.5

# - About the Author -



## **Brad Hedlund**

Technical Solutions Architect, Data Center  
Cisco Systems, Inc.

CCIE #5530

<http://bradhedlund.com/about/>

Blog: <http://bradhedlund.com>

Twitter: <http://twitter.com/bradhedlund>

E-mail: [bhedlund@cisco.com](mailto:bhedlund@cisco.com)

Comments welcome.

# Why Evolve Data Center Switching?

## Transformational Paradigm Shifts

### From Connectivity to Virtualization

The Server is a fluid object

The virtual machine is the new Server

The physical machine is the new Network

Miniaturization & Scale

Any server, any VLAN, anywhere, anytime

### New Requirements

Large, Flat, Scalable L2 fabrics

Simpler, Smarter Networks

Ultra High Availability

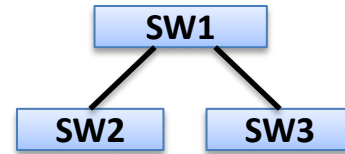
Storage/Ethernet consolidation

More bandwidth

### What's Next?

The "Data Center" becomes a fluid object

What works? What needs to be improved?



# REVISITING CLASSIC ETHERNET

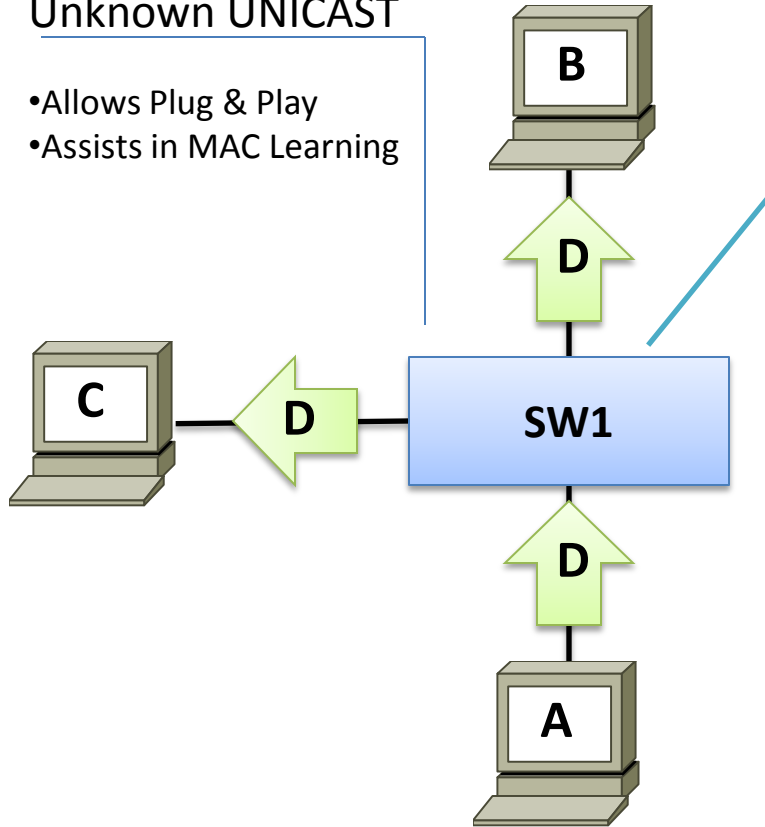
Narrative of this section located at:

<http://bradhedlund.com/2010/05/07/setting-the-stage-for-trill/>

# Flooding Behavior

## Unknown UNICAST

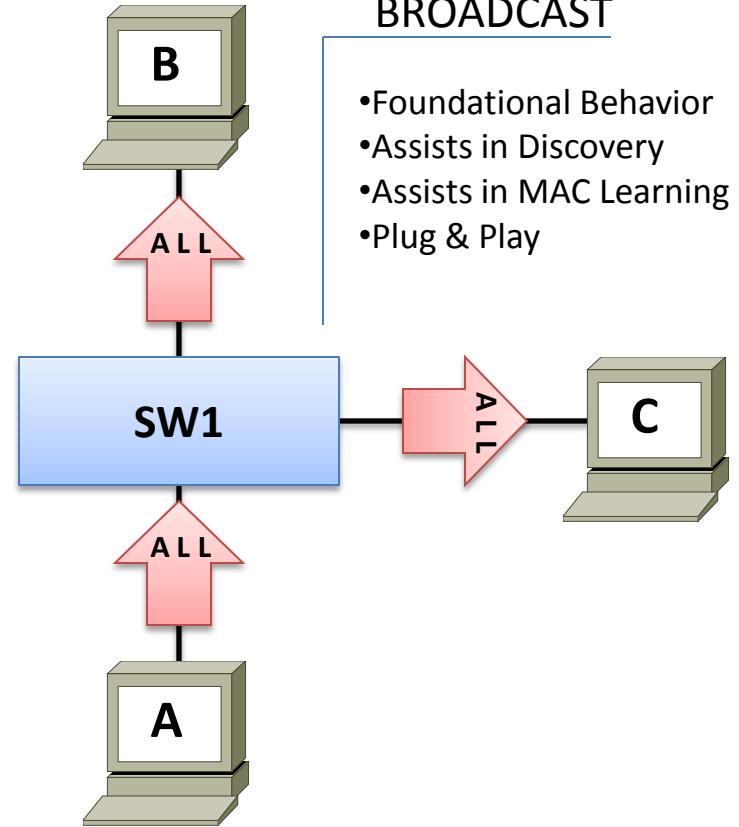
- Allows Plug & Play
- Assists in MAC Learning



## Plug & Play

### MAC Table

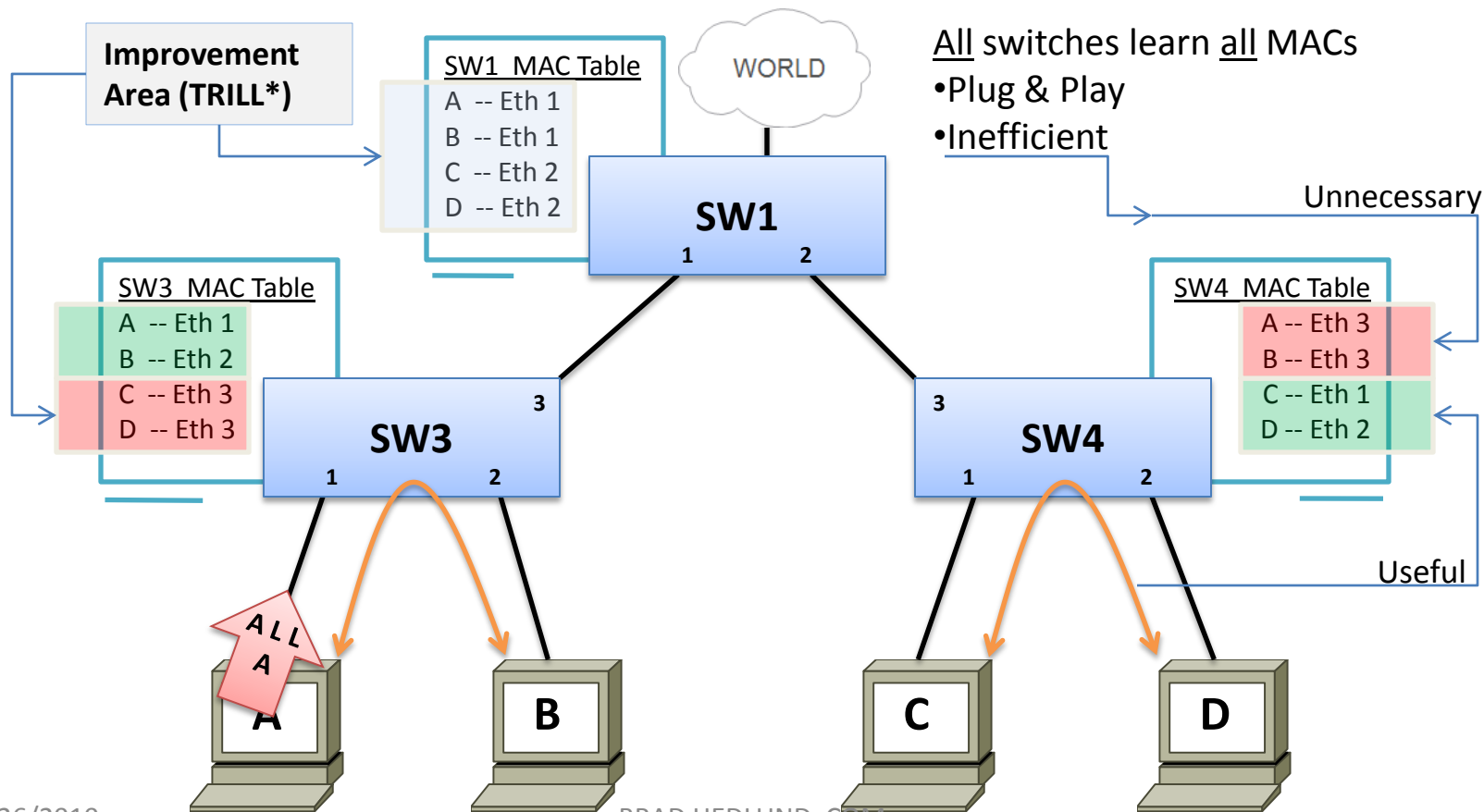
A	Eth 1/1
B	Eth 1/2
C	Eth 1/3



## BROADCAST

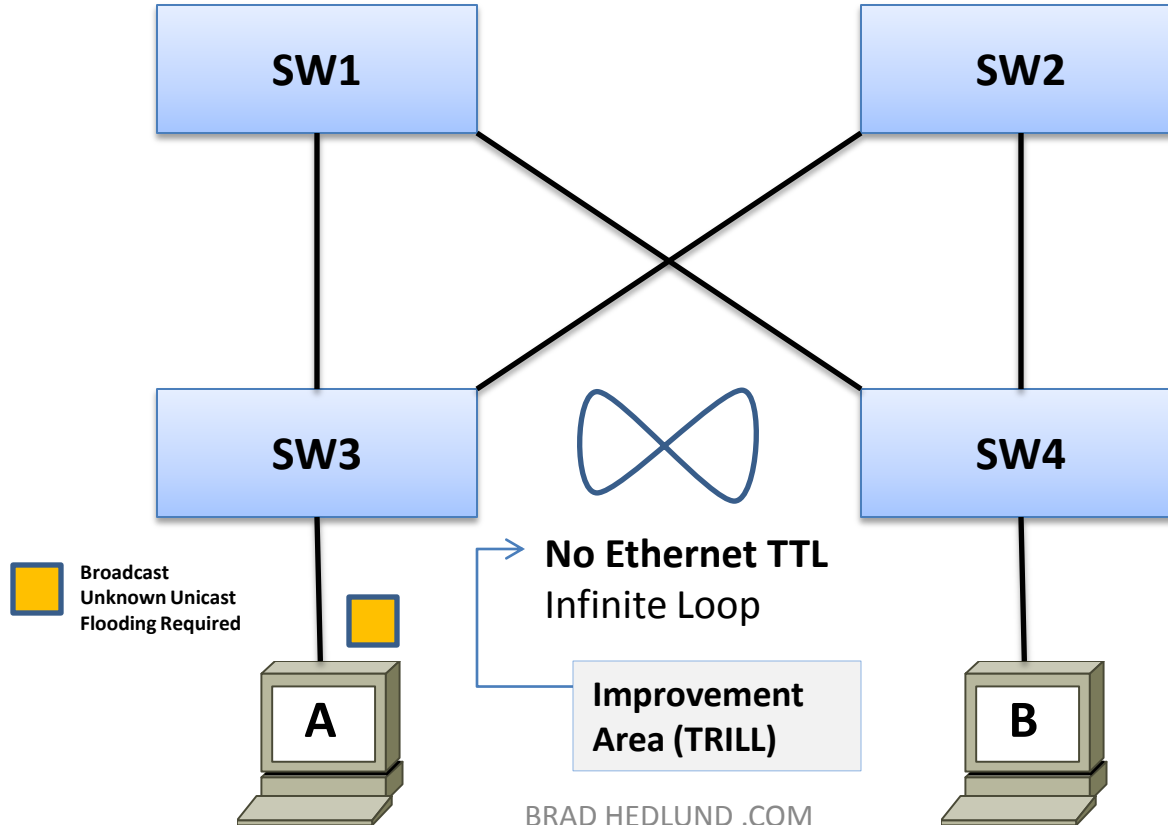
- Foundational Behavior
- Assists in Discovery
- Assists in MAC Learning
- Plug & Play

# Classical Ethernet MAC Learning



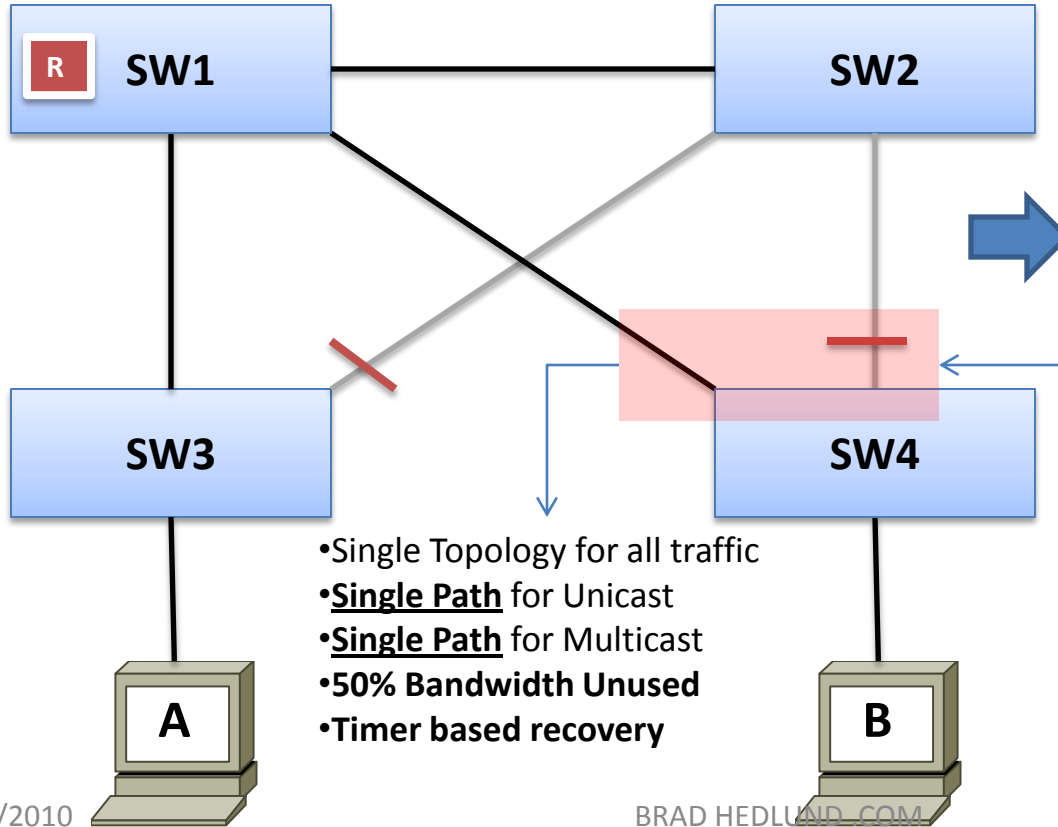
# Flooding with Multiple Paths

Plug & Play loop prevention is needed...

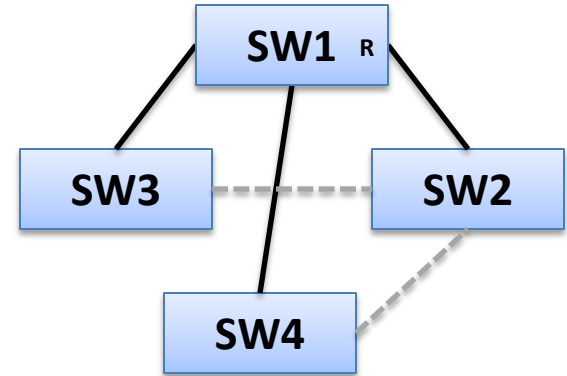


# Classical Ethernet Loop Prevention

## Spanning Tree Protocol (STP)



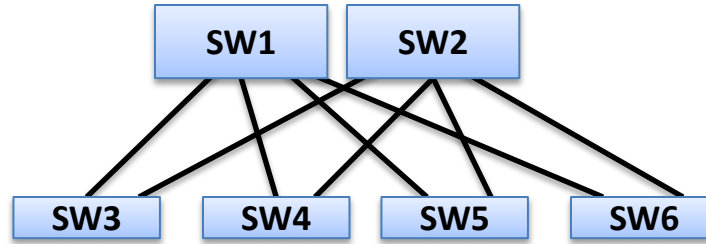
## STP Enforced Tree Topology



- Plug & Play
- Loop Free for flooded traffic

Improvement Areas (TRILL\*)

Challenges, Approaches,



# SCALING THE DATA CENTER NETWORK

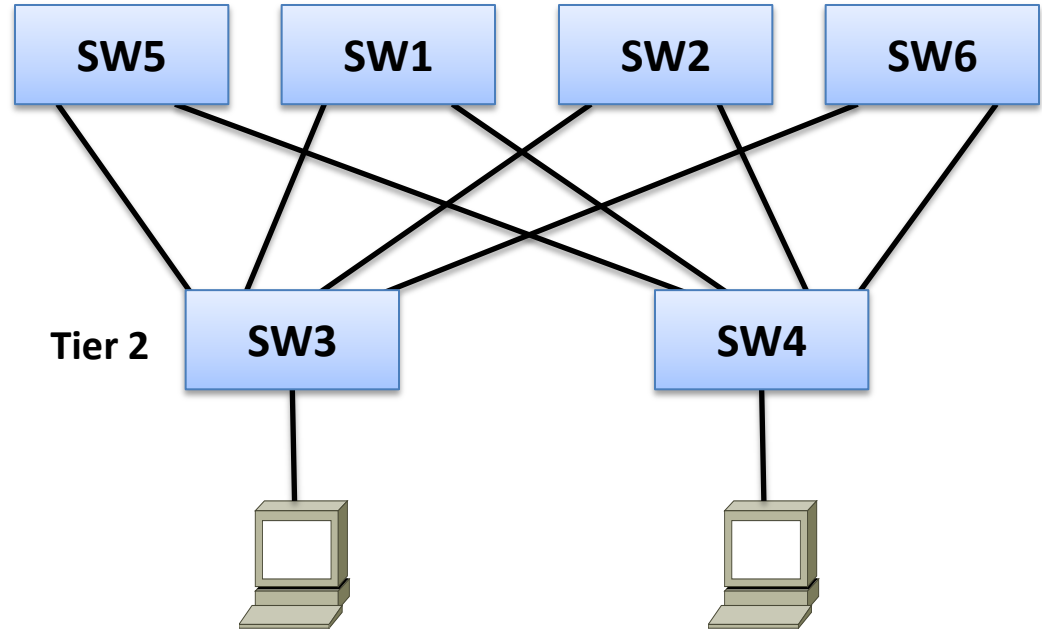
Narrative of this section located at:

<http://bradhedlund.com/2010/05/07/setting-the-stage-for-trill/>

# Scaling out Tier 1

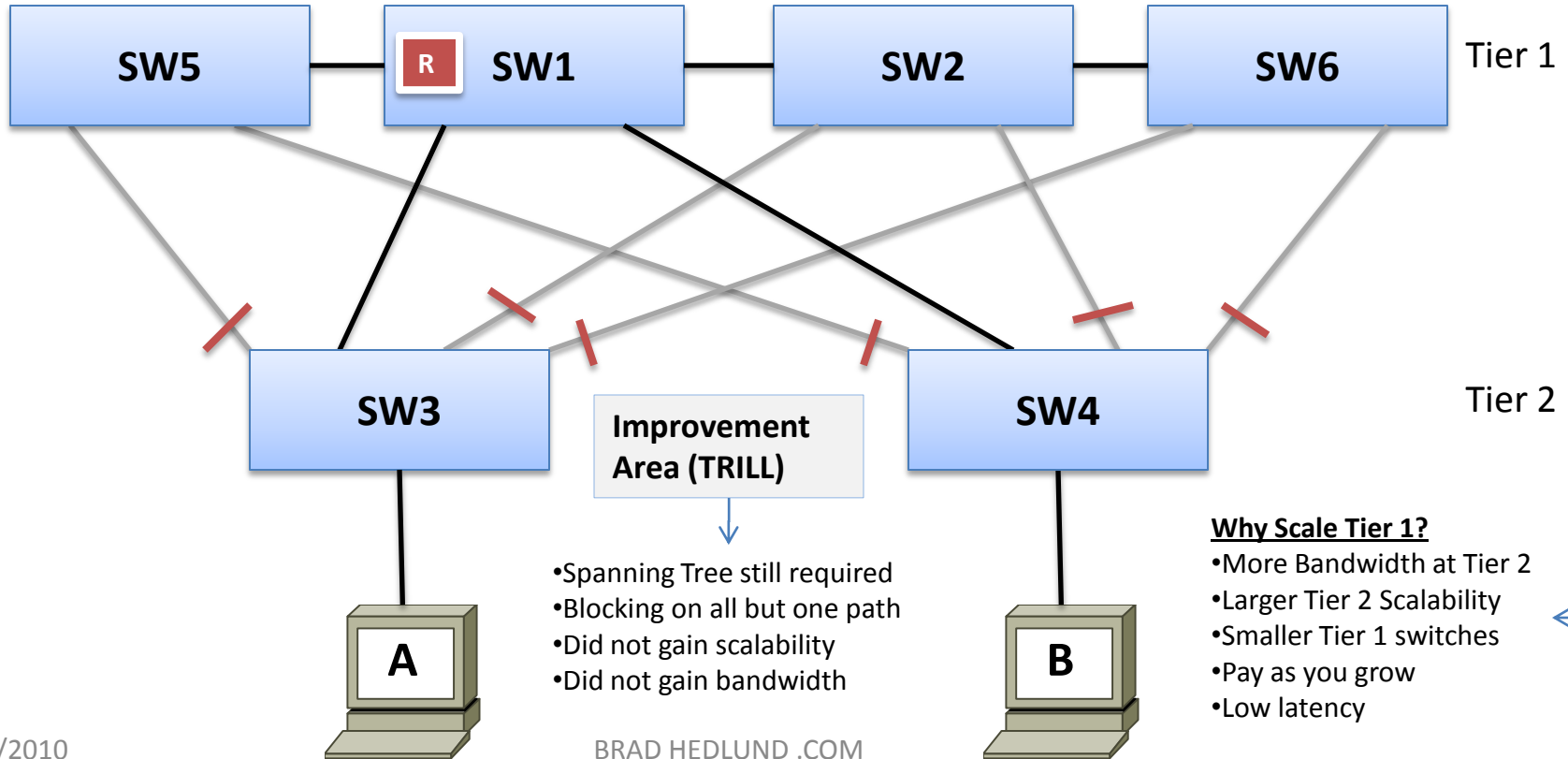
## Why scale out Tier 1?

- More Bandwidth at Tier 2
- Larger Tier 2 Scalability
- Smaller Tier 1 switches
- Spread Risk (RAID)
- Pay as you grow
- Lower latency



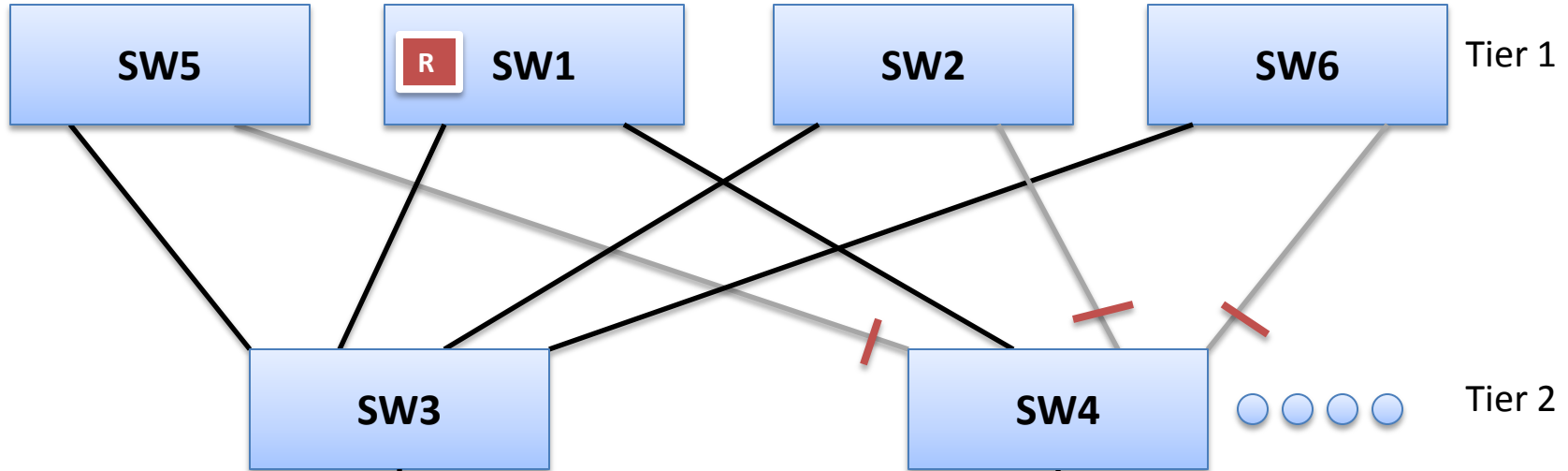
# Scaling out Tier 1 with Classic Ethernet

with Spanning Tree Protocol (STP)



# Scaling out Tier 1 with Classic Ethernet

with Spanning Tree Protocol (STP)



## Alternate Topology

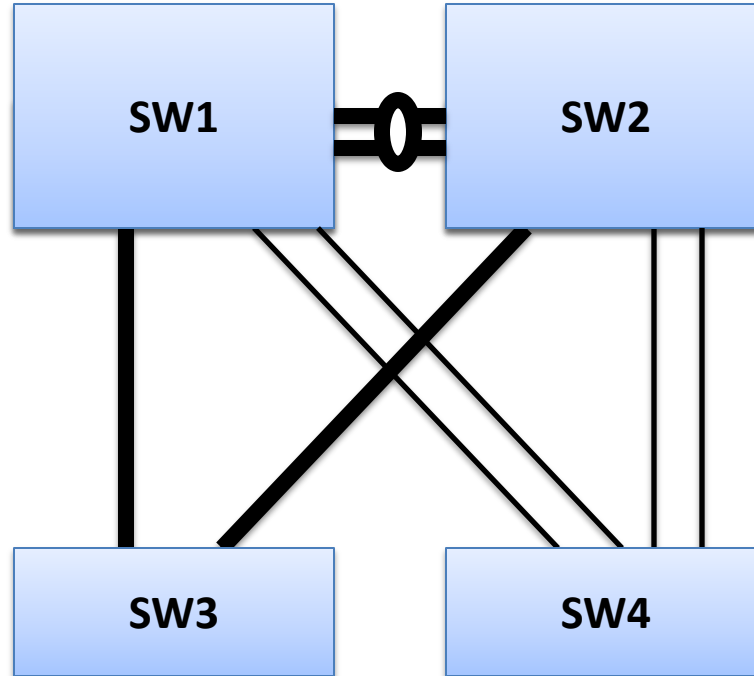
- Unbalanced at Tier 2
- Only 1 switch with all paths
- Did not gain scalability
- Did not gain bandwidth

## Why Scale Tier 1?

- More Bandwidth at Tier 2
- Larger Tier 2 Scalability
- Smaller Tier 1 switches
- Pay as you grow
- Low latency

# Scaling UP Tier 1 with Classic Ethernet

Rigid (2) Switch Tier 1 design constraint forces Scaling UP – not OUT

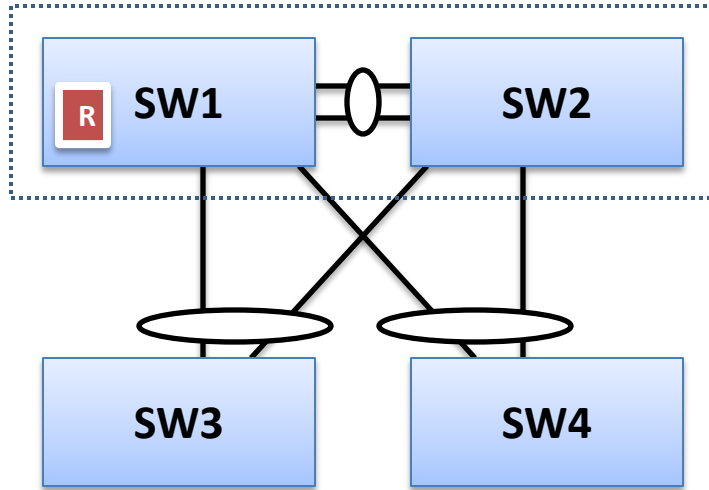


Improvement  
Area (TRILL)

# Multi Path with Classic Ethernet

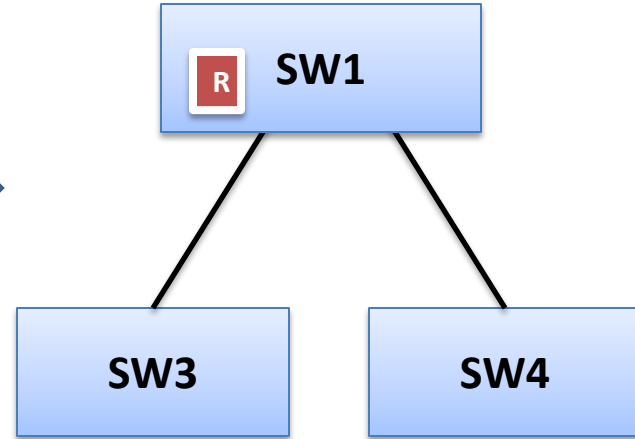
## Multi Chassis Ether Channel (MCEC) aka Virtual Port Channel (vPC)

State Synchronization at Tier 1



Not a trivial accomplishment!  
Several different states must be synced

STP finds a Loop Free Topology  
No Blocking Paths, All links active

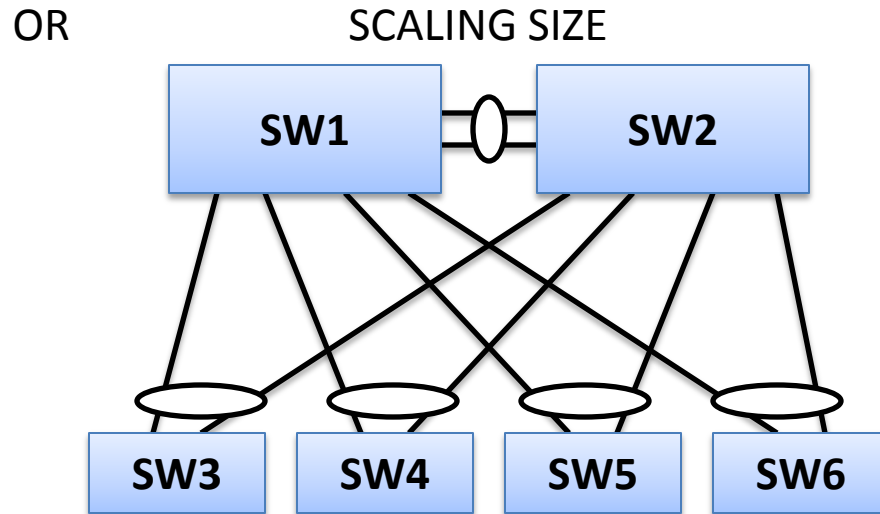
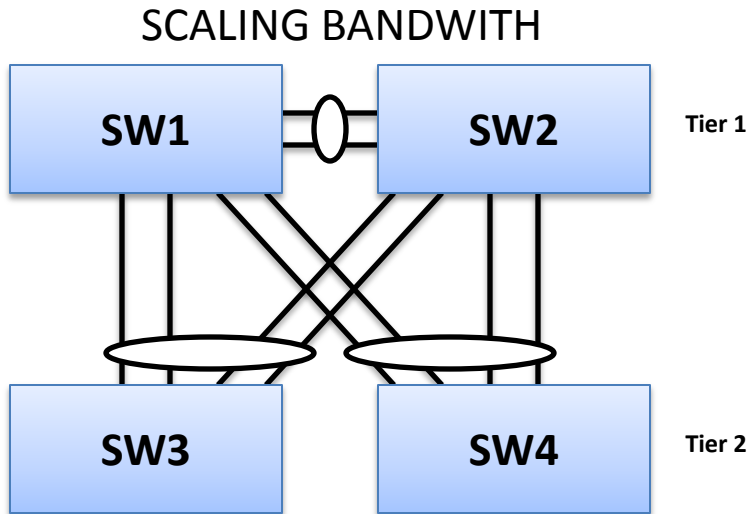


- ✓ STP Roles
- ✓ MAC learning
- ✓ LACP
- ✓ Interface states/config
- ✓ Split Brain detection

Improvement area (TRILL)  
Multi path with minimal  
state/sync complexity

# Scaling Tier 2 with Classical Ethernet

Scale Bandwidth or Size? – Pick one, you can't have both!

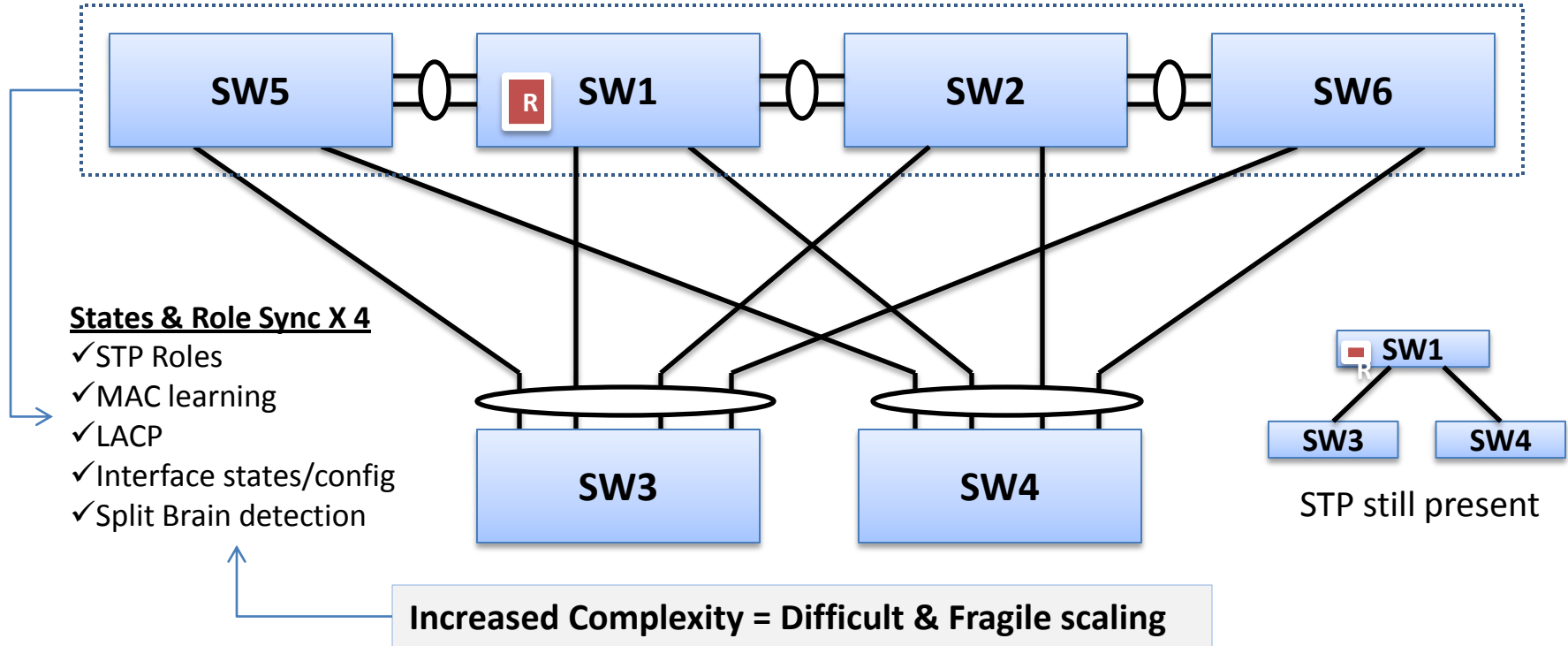


Trade-off between Bandwidth *or* Size  
Tier 1 switch density key scaling factor

Improvement  
Area (TRILL)

# Scaling out Tier 1 with MCEC

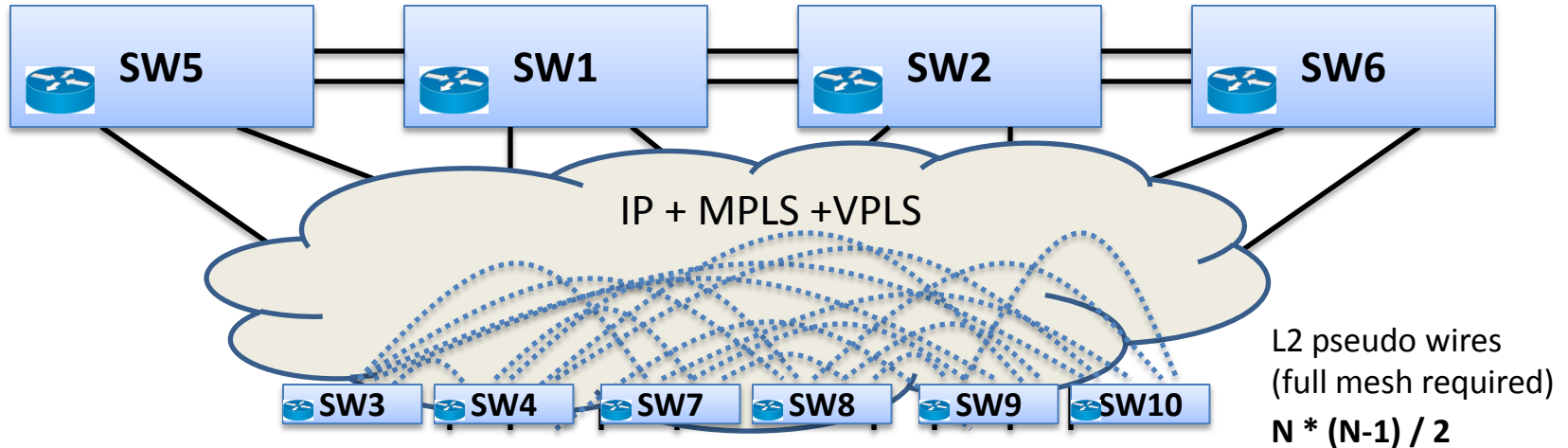
Increasingly Complex State Synchronization across (4) or more Tier 1 switches



# Scaling out Tier 1 with MPLS

Replacing L2 switching with L3 + MPLS and L2 pseudo wire full mesh Overlay via VPLS

Sound complicated? That's because it is.



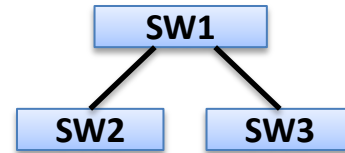
Complex overlay of L2 services over L3

MPLS skill sets?  
Configuration Intense

NOT Plug & Play!

**Increased Complexity = Difficult & Fragile scaling**

An Introduction



# TRILL - LAYER 2 MULTI PATHING

Narrative of this section to be posted at:

<http://www.internetworkexpert.org/topic/trill/>

# Design Goals for TRILL

## Switching



- Minimal Configuration
- Plug & Play
- Auto Discovery
- Auto Learning
- Flat Addressing
- Spanning Tree Protocol (STP)
- Slow Convergence
- Single Path
- Edge-to-Root Rigid Design
- Single Multicast Tree
- Constrained Scaleability

## TRILL



### The best of Switching and Routing

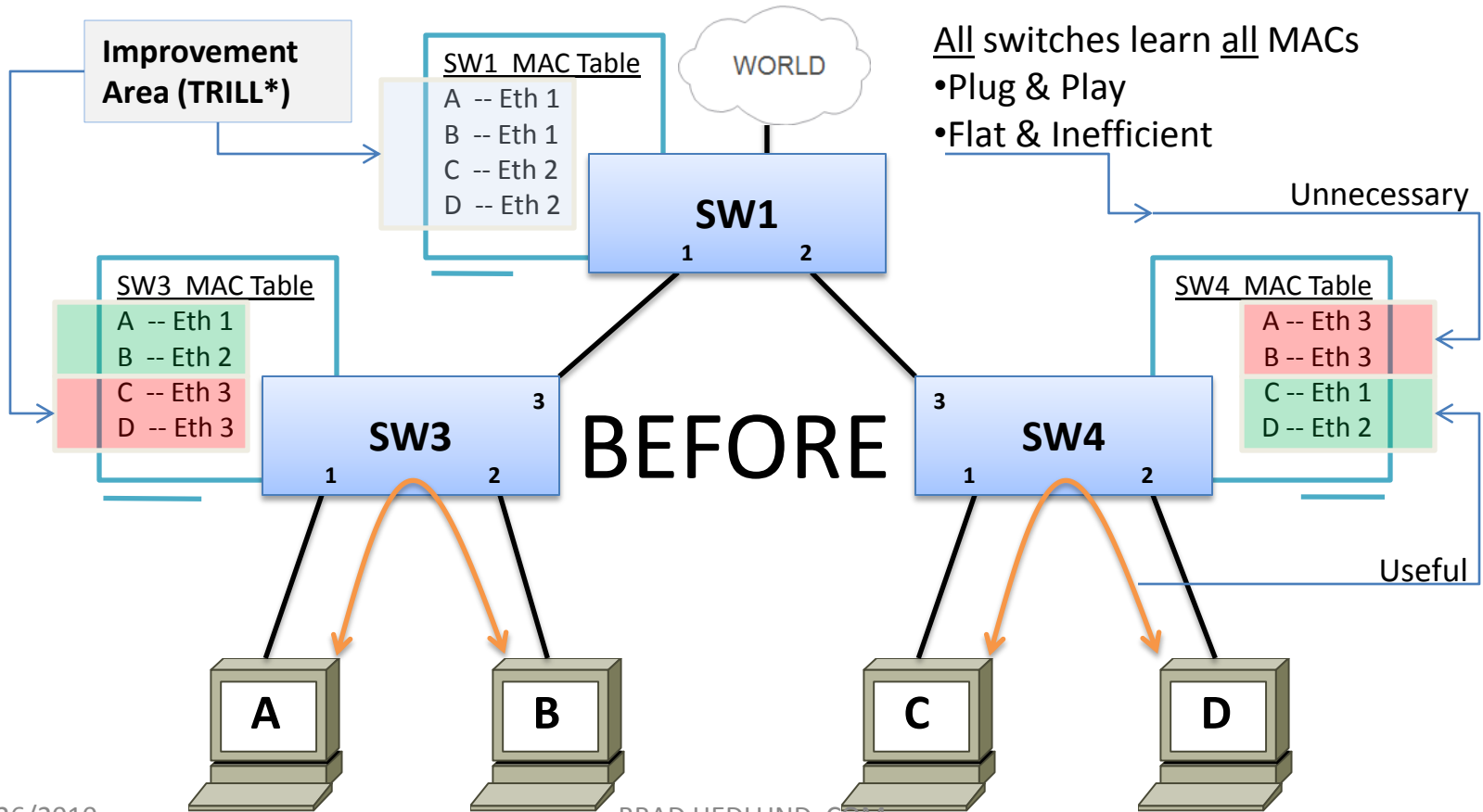
- Minimal Configuration
- Plug & Play
- Auto Discovery
- Efficient MAC Learning
- Multiple Paths
- Load Balancing
- Any-to-any Flexible Design
- Highly Scalable
- Fast Convergence

## Routing

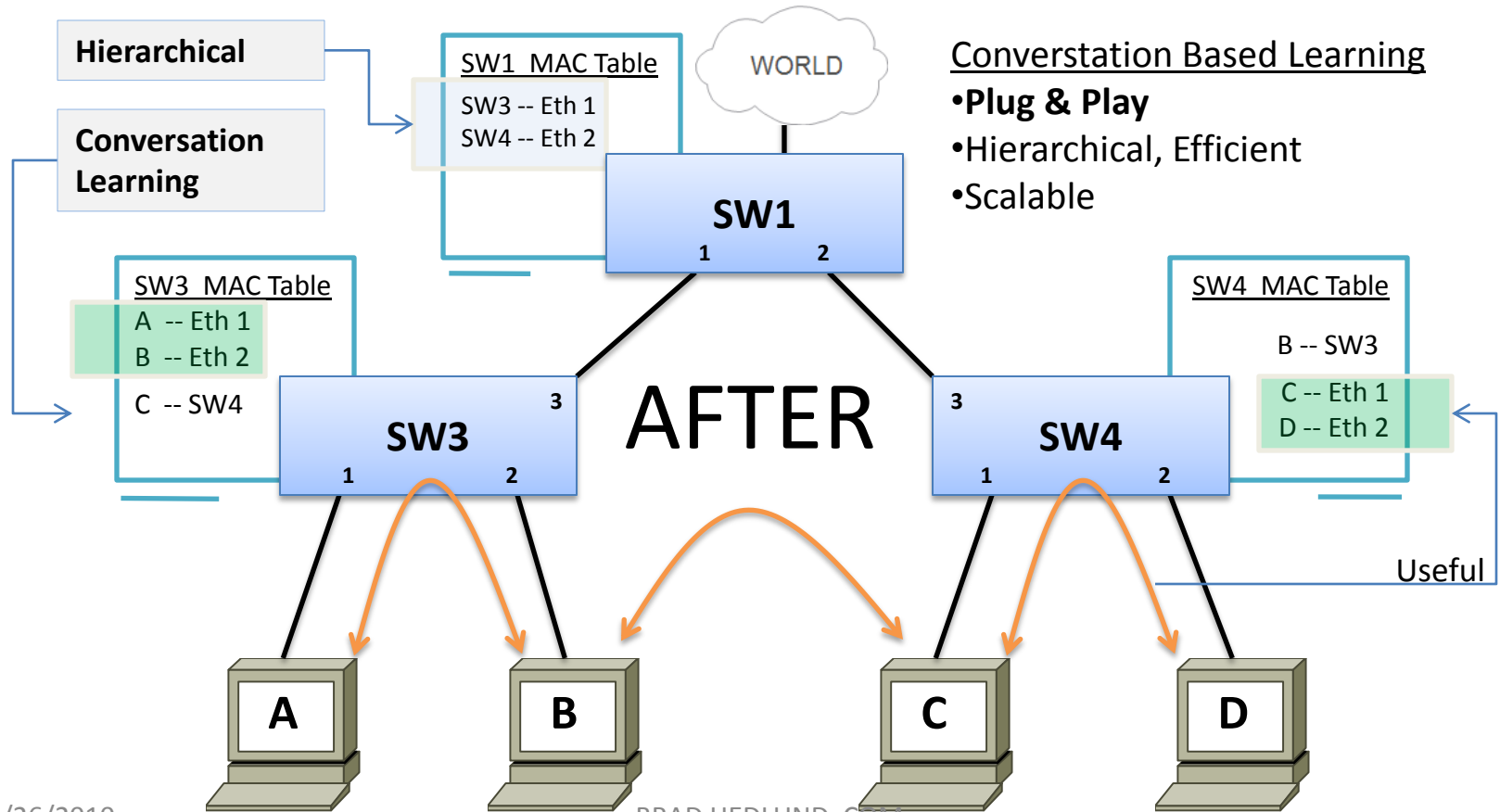


- Configuration Intense
- Configured Learning
- Configured Discovery
- Plan & Play
- Fast Convergence
- Multiple Paths
- Load Balancing
- Multiple Multicast Trees
- Hierarchical Forwarding
- Any-to-any Flexible Design
- Highly Scalable

# MAC Learning -- Evolved

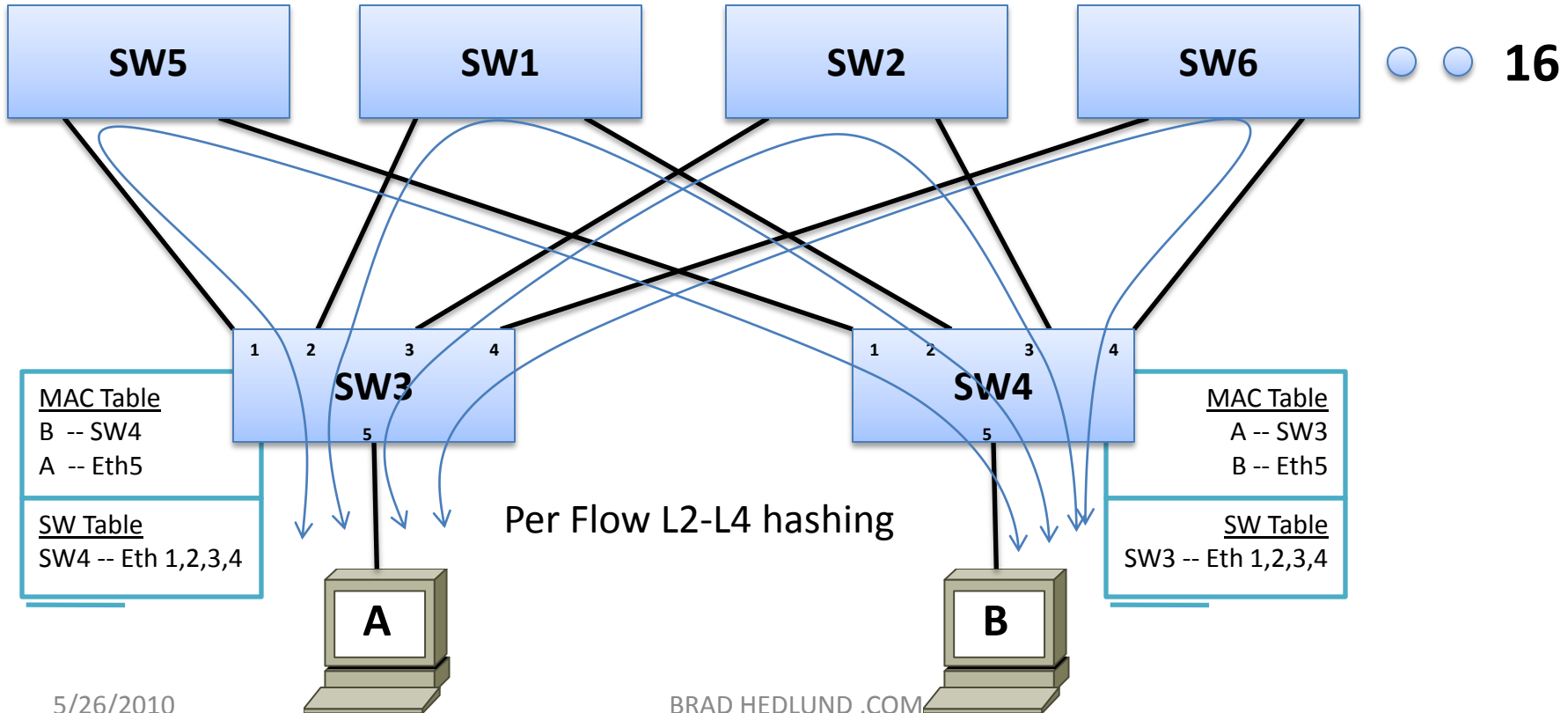


# MAC Learning -- Evolved



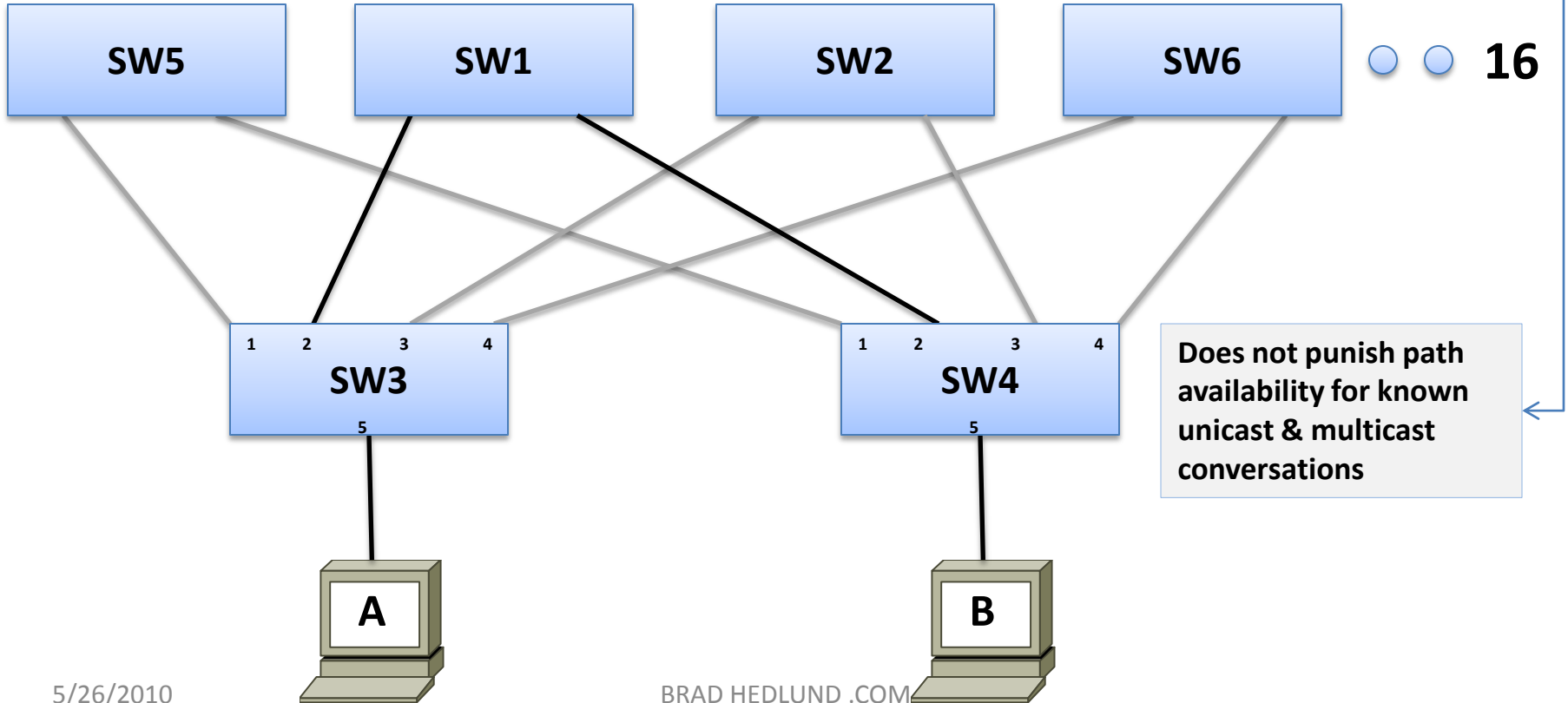
# Multi Pathing -- Evolved

16-way Equal Cost Multi Path (ECMP) Layer 2 Forwarding



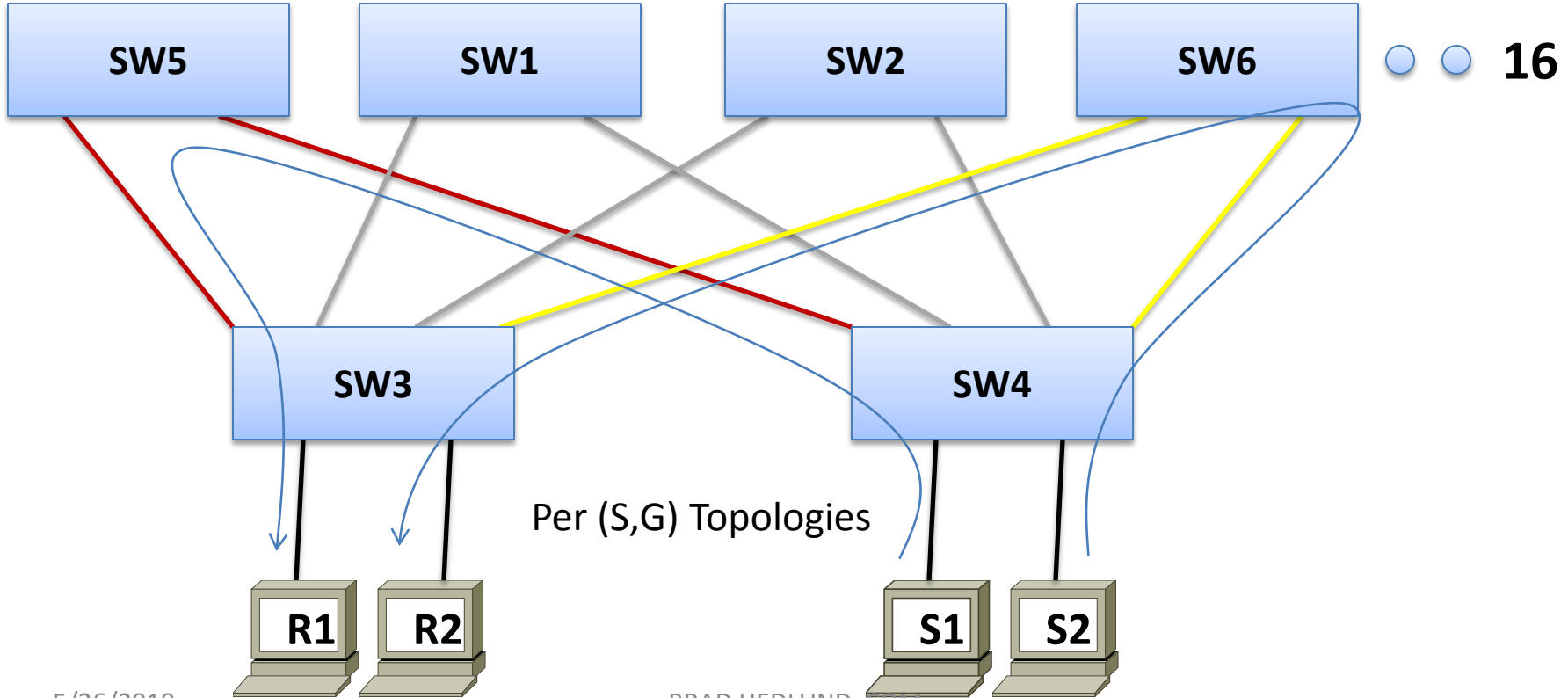
# Loop Free Flooding -- Evolved

A unique loop free forwarding topology for Broadcast & Unknown Unicast ONLY



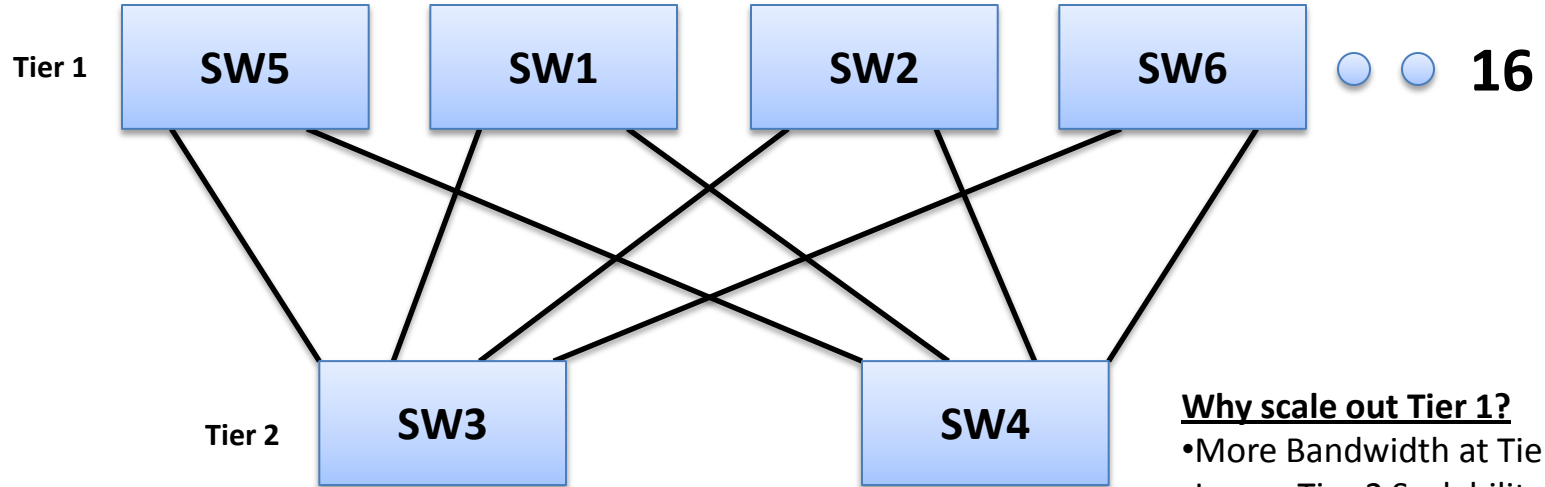
# Multicast -- Evolved

More Bandwidth for Multicast – All possible Topologies Used



# Scaling out Tier 1 with TRILL

Plug & Play L2 fabric with L3 Scalability and Robustness



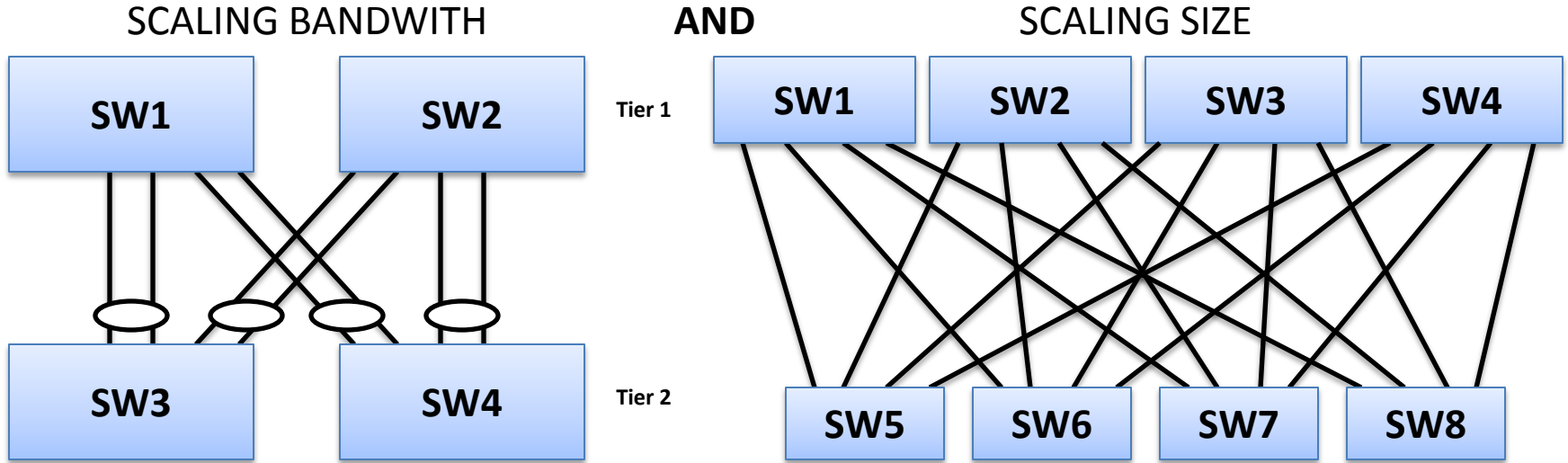
- ✓ No complicated MCEC state sync
- ✓ No Spanning Tree
- ✓ No complicated MPLS overlay
- ✓ Simple Configuration
- ✓ Flexible any-to-any design

## Why scale out Tier 1?

- More Bandwidth at Tier 2
- Larger Tier 2 Scalability
- Smaller Tier 1 switches
- Spread Risk (RAID)
- Pay as you grow
- Low latency

# Scaling Tier 2 with TRILL

Flexible design can scale both Bandwidth and Size



Minimal Trade-off between Bandwidth *or* Size  
**Tier 1 switch density less of a scaling factor**

# Part 1. Summary

## Plug and Play Layer 2 with Layer 3 Scalability and Robustness

### Addressing scalability

- ✓ Conversational MAC Learning\*
- ✓ Hierarchical MAC Forwarding\*

### Bandwidth scalability

- ✓ 16 active paths
- ✓ Multiple Multicast Topologies\*

### Domain Size scalability

- ✓ Flexible 16-switch Tier 1 design options

### Robust L3 characteristics

- ✓ Link State Topology Awareness
- ✓ Fast Convergence

### Plug & Play

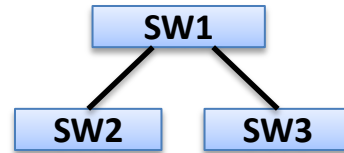
- ✓ Auto Learning
- ✓ Simple Configuration\*\*

\* Not present in current RFC 5556 (TRILL)

\* Present in Cisco specific enhancements, Cisco driving PARs to improve current RFC

\*\* (3) NX-OS commands per Nexus switch, subject to change

<http://bradhedlund.com/feed/>



**STAY TUNED FOR MORE DETAILS...**