Global Knowledge®

Expert Reference Series of White Papers

Cisco FabricPath: Is It Switching, Routing or a Bit of Both?

Cisco FabricPath: Is It Switching, Routing or a Bit of Both?

Annette Smallworth, Global Knowledge Cisco Instructor, CCSI, CCNP Data Center, CCNP Routing and Switching

Introduction

Cisco FabricPath has often been described as routing at Layer 2. Anybody who has attended a networking course or read a networking book that teaches the OSI 7 layer model would surely see that statement as a contradiction. They would argue that routing is a Layer 3 function. The purpose of this white paper is to understand Cisco FabricPath and to explain why routing at Layer 2 can be an acceptable description of Cisco FabricPath.

To understand why the phrase "Routing at Layer 2" is used to describe Cisco FabricPath, we first have to revisit the difference between Layer 2 which is switching and Layer 3 which is routing.

Layer 2 Switching

The only Layer 2 LAN protocol to have survived the test of time is Ethernet. It was created in the 1970s by Xerox and later promoted as the standard by DEC, Intel, and Xerox in the early 1980s. It became an IEEE 802.3 Working Group's official standard soon after. Let's remind ourselves what constitutes the Ethernet frame:

Figure 1: Classic Ethernet Frame Format

6 Bytes	6 Bytes	2 Bytes	Variable	4 Bytes
Destination MAC	Source MAC	EtherType	Data (typically IPv4 Packet)	FCS

Only the fields in red above are the Ethernet header and trailer, and because of the limited number of fields, this leads to some issues.

Limitations of Layer 2 Switching

Switches make a forwarding decision by comparing the Layer 2 header to the MAC address table, but how is the MAC address table built? Recall that switches record the source MAC of every frame that comes into a switch, and the port that the frame came in on. This does present a potential problem though because the larger the Layer 2 switched network becomes, the larger the MAC address tables will become.
Remember that to forward a frame, a switch will compare the destination MAC address to the MAC address table and do one of the following with the frame:

- Flood - Broadcasts, multicasts and unknown unicasts
- Forward - Known unicasts
- Filter (drop) - FCS errors, port security violations, etc.

Limitations of Spanning Tree Protocol

Spanning Tree is used to resolve Layer 2 Ethernet loops by blocking one port per loop. Frames will only enter or exit a port that is in the forwarding state, not one that is in the blocking state. How do you determine which ports will be forwarding? Well, that depends on the spanning tree algorithm. In Figure 2 below, there are seven switches with a total of fourteen links between the switches.

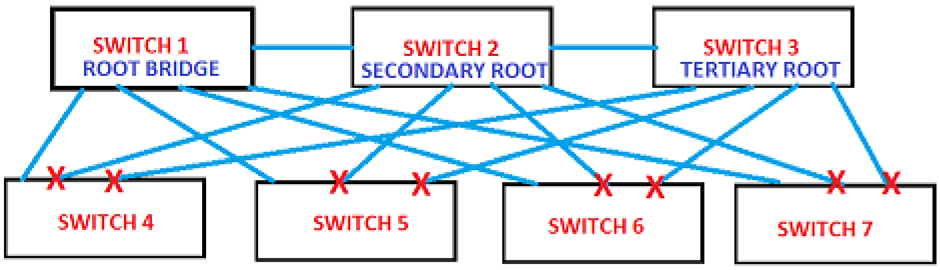

Figure 2: Switched Network

However, because the Spanning Tree algorithm blocks one port per loop, only six of the links are actually being used, therefore the traffic is using less than 50 percent of available bandwidth plus much of the traffic is also traversing suboptimal paths. For example, notice that frames between Switch 3 and Switch 4 would have to traverse Switch 1 and Switch 2, even though there is a link between Switch 3 and Switch 4.

The Dreaded Broadcast Storm

Another issue with Spanning Tree is that because it relies 100 percent on special frames called bridge protocol data units (BPDUs), it "fails open." What that means is that when left at default, if a port does not receive a BPDU it will automatically transition to the forwarding state.

For example, in Figure 2 above, Switch 4, Switch 5, Switch 6, and Switch 7 have multiple ports in the blocking state because they are receiving BPDUs on multiple ports and hence each switch knows that it is part of a Layer 2 loop. However, if for any reason a switch fails to receive a BPDU on a port because of a unidirectional link failure or issues on the neighboring switch, then the Spanning Tree algorithm would transition the port to the forwarding state, and consequently a Layer 2 frame would go round and round "forever", resulting in a broadcast storm. How long is forever? Not that long, because the Layer 2 network will become unusable.

So to summarize, the issues with Classic Ethernet Switching and Spanning Tree are:

- large MAC address tables
- no load balancing (only one best path between any pair of switches)
- suboptimal paths
- Layer 2 network fails open and broadcast storms

Layer 3 Routing

So why do we not experience these issues with Layer 3 IPv4 networks? Routers make a forwarding decision by comparing the destination IP address to the routing table. But, how is a routing table built?

The only subnets that automatically go into the routing table are the directly connected subnets. Any other subnets are learnt either statically or dynamically, via a routing protocol, and both are in control of the network administrator, so indirectly, the size of the routing table can be controlled by the network administrator.

In considering one of the most common routing protocols, Open Shortest Path First (OSPF), recall it forms a neighbor relationship by using HELLO packets before advertising any subnets between the two neighbors. Therefore, if there are any issues such as an unidirectional link failure, it will not form a neighbor relationship and consequently it will "fail closed" and not advertise any subnets over that link.

Equal Cost Multipath Routing

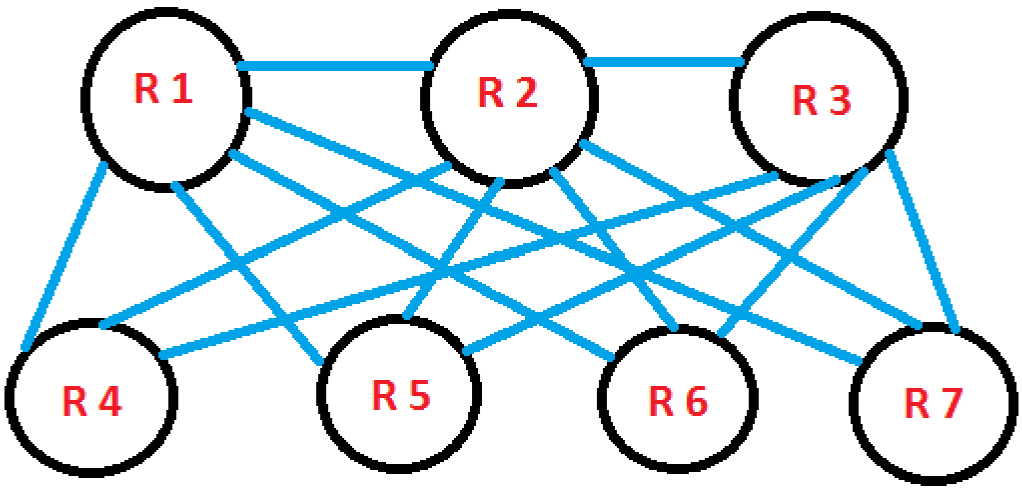

Figure 3: Routed Network

Remember that in the switched network in Figure 2, the Spanning Tree algorithm has blocked many of the links. Compare this to the routed network in Figure 3 above. Notice there are no blocked ports and any traffic will always be using the best path between any pair of routers. Also, recall all interior gateway protocols such as OSPF, Enhanced Interior Gateway Routing Protocol (EIGRP), etc., perform equal cost multipath (ECMP) routing. Meaning, if there are multiple best paths between a pair of routers, all of those best paths will be in the routing table.

Routing Storm?

Have you ever heard of a Routing Storm? No, because routing storms cannot occur. This is due to one tiny, not-so-inconsequential field in the IP Header—the TTL field. Worst case scenario, even if the routing tables were in a mess and a routing loop had occurred, the IP packets could not be forwarded round and round forever, as is the case of Ethernet frames, because the IP TTL field would eventually reach zero, and the packet would be dropped.

At this point you may be thinking, "If routing is so good, why continue with switching?" One primary reason is that Layer 2 switching gives the flexibility of being able to place a device on the same VLAN, same subnet, anywhere in the building or data center, and this flexibility is particularly useful in server environments for example for VM Mobility.

FabricPath Combines the Benefit of Switching and Routing

The chart below is a reminder of the characteristics of switching, routing and FabricPath. Notice that FabricPath combines the benefits of switching and routing.

Switching	Routing	FabricPath (FP)
Flexibility to place equipment anywhere	Limited Flexibility due to Subnet constraints	Flexibility to move equipment
No ECMP	ECMP	ECMP
Suboptimal Paths	Optimal Paths	Optimal Paths
Broadcast Storms	No Broadcast Storms (because of IP TTL)	No Broadcast Storms (because of FP TTL)

Figure 4

No, that we have reminded ourselves of the differences between switching and routing, let's move on to discover how FabricPath combines the two. First, it must be emphasized that FabricPath is 100 percent Layer 2 Ethernet. There is no IP within FabricPath. However, where the confusion may arise is that FabricPath uses features and terms that we are familiar with at Layer 3.

FabricPath Control Plane

The primary FabricPath control plane protocol is Intermediate System to Intermediate System (IS-IS), which some only know as a traditional routing protocol to advertise IP subnets, but IS-IS is so much more than that. IS-IS was developed by DEC as part of the Open System Interconnection (OSI) model. It is a link state algorithm that is similar to OSPF. However, there are two features that are particularly good about IS-IS.

- IS-IS is encapsulated directly within Layer 2. Meaning it does not use or need IP to work.
- IS-IS makes use of TLV (type, length, and value), so that developers can modify it relatively easily to advertise different variables. In FabricPath the TLV are used to advertise switch IDs and the cost of directly connected links.

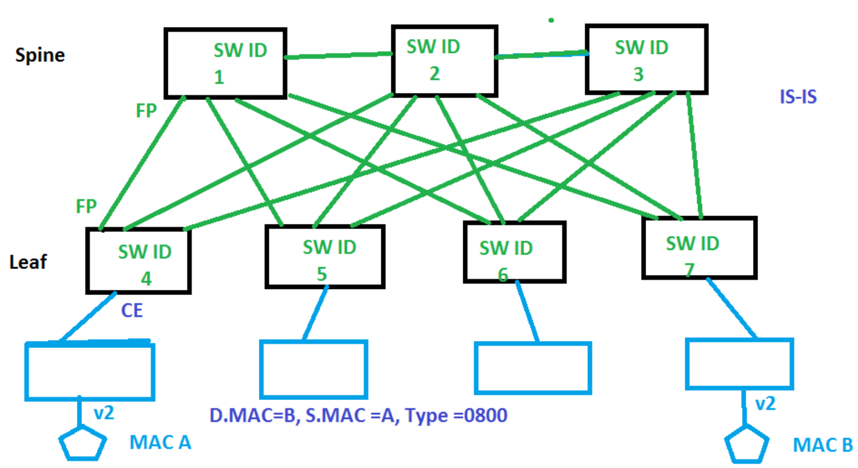

Figure 5: FabricPath Network

In Figure 5, all ports are Ethernet, but the green links represent those that are running FabricPath (FP), the blue links represent those that are not running FabricPath, which from this point on are known as Classic Ethernet (CE). As I describe the FabricPath Control Plane in the bullet points below, note the similarities to OSPF:

- Each FabricPath Switch has a unique number called a Switch ID (SW ID)
    - The SW ID can be left as a random number, but best practice to manually configure (similar to OSPF Router ID)

- IS-IS HELLO frames form neighbor relationships between adjacent Switches
    - similar to OSPF HELLO packets

- Each FabricPath Switch learns the topology via IS-IS Link State Packets (LSPs)
    - similar to OSPF LSAs

- Each FabricPath Switch runs the Dijkstra SPF algorithm to compute the FabricPath Routing Table
    - The best path between two switches is the one with the lowest cost, which by default relates to bandwidth
    - If multiple equal cost best paths exist then they will all be added to the FabricPath Routing table—hence ECMP
    - This process is similar to OSPF Dijkstra algorithm and best path selection

Now perhaps you are beginning to see why FabricPath is often referred to as "Routing at Layer 2," especially when looking at a section of a FabricPath routing example below:

```
S4# show 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
1/7/0, number of next-hops: 3
        via Po10, [115/40], 0 day/s 19:07:39, isis_fabricpath-default
        via Po20, [115/40], 0 day/s 19:07:40, isis_fabricpath-default
        via Po30, [115/40], 0 day/s 19:07:40, isis_fabricpath-default
```

Now that IS-IS is running at the control plane for FabricPath, two of the Layer 2 issues have been solved. This means no more suboptimal paths and there is now Equal Cost Multipath routing. But that still leaves two issues to be resolved: broadcast storms and MAC address table size. These two issues are resolved by the FabricPath data plane.

FabricPath Data Plane

Another phrase used to describe FabricPath is "MAC in MAC," however, this phrase describes the FabricPath data plane. Notice that through FabricPath the Classic Ethernet frame is encapsulated in another Ethernet frame.

FabricPath Header				Classic Ethernet Frame				
6 Bytes	6 Bytes	2 Bytes	2 Bytes	6 Bytes	6 Bytes	2 Bytes	Variable	4 Bytes
Outer Destination Address (ODA)	Outer Source Address (OSA)	EType 0x8903	FTAG and **TTL**	Destination MAC	Source MAC	EtherType (Typically 0x0800)	Data (typically IPv4 Packet)	New FCS

Figure 6: FabricPath Frame

In the FabricPath Header

The ODA contains the Destination SW ID and Destination Port ID. The OSA contains the source switch ID and source port ID. Note that in the FabricPath header there is a TTL field. By default this value starts at 32, and is decremented every time the frame goes through a FabricPath switch. So now a third issue of Layer 2 switching has been resolved. In the worst-case scenario, even if a FabricPath loop ever occurred, a FabricPath frame would not loop forever, therefore no broadcast storms in FabricPath. Hooray!

Classic MAC Address Learning versus Conversational Learning

We are left with one Layer 2 Switching issue to resolve: the size of the MAC address table. I imagine many of you are thinking, "The size of my MAC address table has never been a problem." However, recall that is because most companies limit Layer 2 to the access layer. The purpose of FabricPath is to allow massive Layer 2 domains; therefore MAC address table scalability is required. The solution is conversational learning, which is automatically implemented on FabricPath ports.

To understand conversational learning, compare it to Classic Ethernet learning:

Classic Ethernet MAC Address Learning (which occurs on CE Ports)

- Records the source MAC of every frame and the port it came in on

Conversational MAC Address Learning (which occurs on FP Ports)

- Only records the source MAC of the frame if the destination MAC is already in the MAC address table

Therefore on an FP port the source MAC will never be recorded from broadcast, multicast, or unknown unicast frames, hence keeping the MAC address table smaller. In FabricPath a switch learns a MAC address on a "need-to-know" basis.

Conclusion

Cisco FabricPath combines the benefits of Layer 2 Switching and Layer 3 Routing. It allows for the scalability and flexibility of Layer 2, plus supporting optimal paths, equal cost multipath routing and a Time to Live (TTL) field, which are traditional Layer 3 components. So to answer the question in the title of this white paper, "Is FabricPath Switching, Routing or a bit of both?" I leave it to you to debate and decide!

Learn More

Learn more about how you can improve productivity, enhance efficiency, and sharpen your competitive edge through training.

DCUFI - Implementing Cisco Data Center Unified Fabric v5.0
DCUFT - Troubleshooting Cisco Data Center Unified Fabric v5.0
DCNX5K - Implementing the Cisco Nexus 5000 and 2000 v2.0
DCNX7K - Configuring Cisco Nexus 7000 Switches v3.0

Visit **www.globalknowledge.com** or call **1-800-COURSES (1-800-268-7737)** to speak with a Global Knowledge training advisor.

About the Author

Annette Smallworth has worked in the IT and networking field for 35 years. For ten years she was employed by DEC as a Computer and Network Engineer. She has been a Cisco Network Trainer for the past fifteen years, specializing in routing and switching and Data Center. She has multiple Cisco certifications including CCSI, CCNP R&S, and CCNP DC.