Scaling Data Centers with FabricPath and the Cisco FabricPath Switching System

What You Will Learn

Traditional network architectures have been designed to provide high availability for static applications. Industry trends such as server virtualization and massively scalable distributed applications require more flexibility to be able to move freely between physical data center zones and greater bandwidth scalability to support any-to-any communication.

Cisco FabricPath is an innovative Cisco NX-OS technology that can transform the way data center networks are conceived. It brings the benefits of Layer 3 routing to Layer 2 switched networks to build a highly resilient and scalable Layer 2 fabric. Cisco FabricPath is a critical component of the Cisco FabricPath Switching System (FSS), an integrated system solution based on the Cisco Nexus® 7000 Series Switches. You can use Cisco FSS as a foundation for building massively scalable and flexible data centers.

Challenges in Current Network Design

Modern data centers still include some form of Layer 2 switching, partly because of the requirements set by certain solutions, which expect Layer 2 connectivity, but also because of the administrative overhead and the lack of flexibility that IP addressing introduces. Setting up a server in a data center needs planning and implies the coordination of several independent teams: network team, server team, application team, storage team, etc. In a routed network, moving the location of a host requires changing its address, and because some applications identify servers by their IP addresses, changing the location of a server is basically equivalent to starting the server installation process all over again. Layer 2 introduces flexibility by allowing the insertion or movement of a device in a transparent fashion from the perspective of the IP layer. Virtualization technologies increase the density of managed virtual servers in the data center, making better use of the physical resources, but also exacerbating the need for flexible Layer 2 networking.

Although Layer 2 switching may provide the flexibility critical to the operation of a large data center, it also presents some shortcomings compared to a routed solution. The Layer 2 data plane is susceptible to frame proliferation. The forwarding topology, typically but not necessarily computed by the Spanning Tree Protocol (STP), must be loop-free at any cost; otherwise, frames could be replicated at wire speed and affect the entire bridged domain. This restriction prevents Layer 2 from taking full advantage of the available bandwidth in the network, and it often creates suboptimal paths between hosts over the network. Also, because a failure could affect the entire bridged domain, Layer 2 is confined to small islands for risk containment.

Therefore, current data center designs are a compromise between the flexibility provided by Layer 2 and the scaling offered by Layer 3:

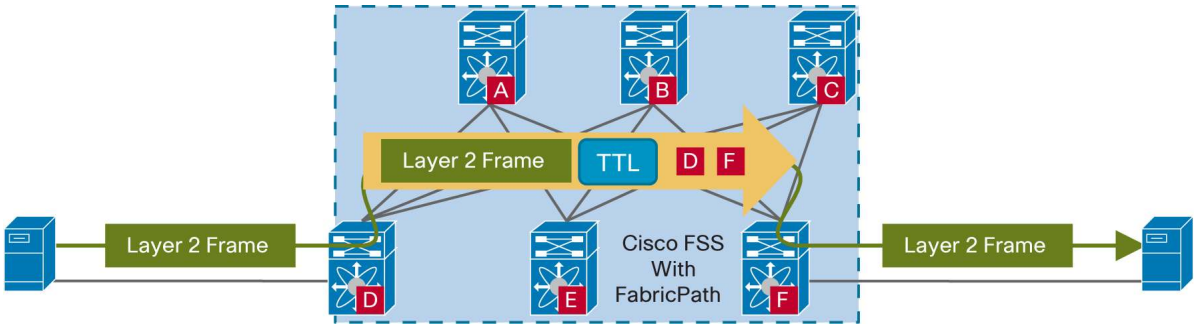
- Limited scale: Layer 2 provides flexibility but cannot scale. Bridging domains are thus restricted to small areas, strictly delimited by Layer 3 boundaries.
- Suboptimal performance: Traffic forwarding within a bridged domain is constrained by spanning-tree rules, limiting bandwidth and enforcing inefficient paths between devices.
- Complex operation: Layer 3 segmentation makes data center designs static and prevents them from matching the business agility required by the latest virtualization technologies. Any change to the original plan is complicated, configuration intensive, and disruptive.

Cisco FabricPath: Routing at Layer 2

Cisco FabricPath is an innovative Cisco NX-OS feature designed to bring the stability and performance of routing to Layer 2.

Cisco Fabric Path takes over as soon as an Ethernet frame transitions from an Ethernet network (referred to as classical Ethernet) to a FabricPath fabric. Ethernet bridging rules do not dictate the topology and the forwarding principles in a FabricPath fabric. The frame is encapsulated with a FabricPath header, which consists of routable source and destination addresses. These addresses are simply the address of the switch on which the frame was received and the address of the destination switch to which the frame is heading. From there on, the frame is routed until it is reaches the remote switch, where it is deencapsulated and delivered in its original Ethernet format. Figure 1 illustrates this simple process.

Figure 1. Frame Transported Across Cisco FSS Using FabricPath

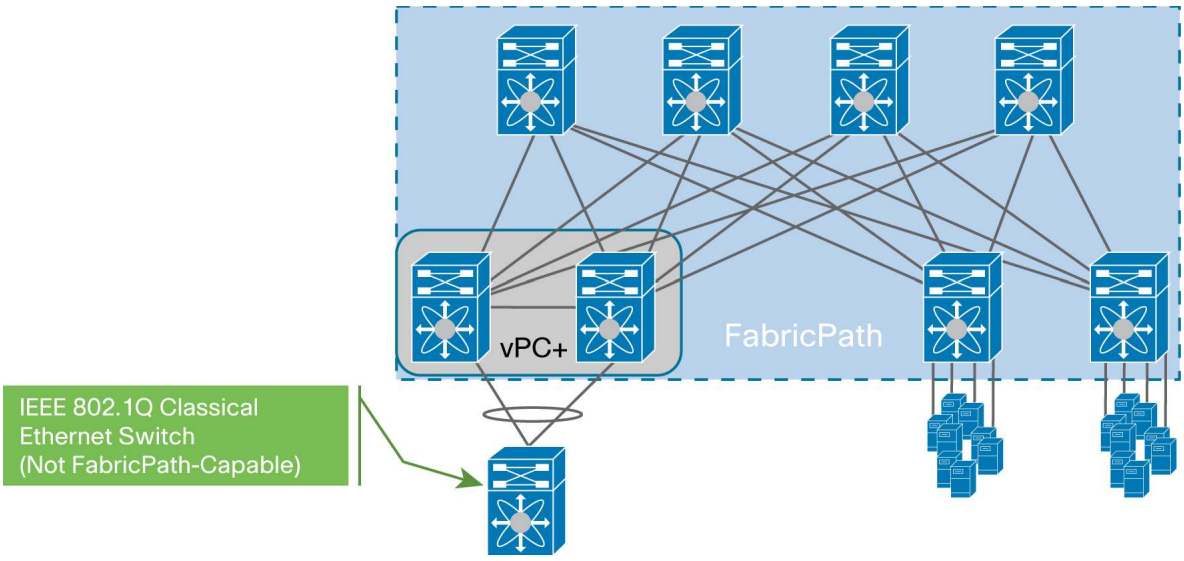


The fundamental difference between FabricPath and classical Ethernet is that with FabricPath, the frame is always forwarded in the core using a known destination address. The addresses of the bridges are automatically assigned, and a routing table is computed for all unicast and multicast destinations. The forwarding process in the fabric never resorts to flooding. The resulting solution still provides the simple and flexible behavior of Layer 2, while using the routing mechanisms that make IP reliable and scalable.

Cisco FabricPath provides the following benefits:

- Simplified network, reducing operating expenses:
  - Cisco FabricPath is extremely simple to configure. In fact, the only necessary configuration consists of distinguishing the core ports, which link the switches, from the edge ports, where end devices are attached. There is no need to tune any parameter to get an optimal configuration, and switch addresses are assigned automatically.
  - A single control protocol is used for unicast forwarding, multicast forwarding, and VLAN pruning. This protocol requires less combined configuration than in an equivalent network that uses the Spanning Tree Protocol, further reducing the overall management needed for the solution.
  - Static network designs make some assumptions about traffic patterns and the locations of servers and services. If those assumptions are incorrect, a situation that often happens after a while, complex redesign may be necessary. A network based on FabricPath can be modified as needed in a nondisruptive manner for the end stations.
  - The capabilities of FabricPath troubleshooting tools surpass those of the tools currently available in the IP world. The ping and traceroute features now offered at Layer 2 can measure latency and test a particular path among the multiple equal-cost paths to a destination within the fabric.
  - A device that is not FabricPath-capable can be attached redundantly to two separate FabricPath bridges with vPC+, providing an easy migration path (see Figure 2 below). Just like vPC, vPC+ relies on EtherChannels to provide multipathing and redundancy without resorting to STP.

Figure 2. Connecting Devices That Are Not FabricPath-Capable with vPC+



- Reliability based on proven technology
  - FabricPath uses a control protocol built on top of the powerful Intermediate System-to-Intermediate System (IS-IS) routing protocol, an industry standard that provides fast convergence and that has been proven to scale up to the largest service provider environments. Nevertheless, no specific knowledge of IS-IS is required in order to operate a FabricPath network.
  - Loop prevention and mitigation is available in the data plane, helping ensure safe forwarding that cannot be matched by any transparent bridging technology. The FabricPath frames include a time-to-live (TTL) field similar to the one used in IP, and a Reverse Path Forwarding (RPF) check is also applied.
- Efficiency and high performance
  - Because equal-cost multipath (ECMP) can be used the data plane, the network can use all the links available between any two devices. The first-generation hardware supporting FabricPath can perform 16-way ECMP, which, when combined with 16-port 10-Gbps PortChannels, represents a potential bandwidth of 2.56 terabits per second (Tbps) between switches.
  - Frames are forwarded along the shortest path to their destination, reducing the latency of the exchanges between end stations compared to a spanning tree-based solution.
  - MAC addresses are learned selectively at the edge, allowing to scale the network beyond the limits of the MAC address table of individual switches.

Introducing Cisco FabricPath Switching System

The virtualization trend started at the edge of the data center. Server virtualization allows consolidation of several servers as virtual machines on a single physical host to increase its usage. Cisco FabricPath Switching System (FSS) is a foundation for building scalable fabrics. It combines the benefits of FabricPath, a critical Cisco NX-OS capability, with the performance advantages of the Cisco Nexus 7000 F1 Series Module to build a high-performance system. Cisco FSS is offered as a set of validated solution bundles that simplify how customers can build out their data center infrastructure. Cisco FSS can provide optimal bandwidth between any two ports regardless of their physical locations. Furthermore, the fabric is built on a new technology that does not suffer from the scaling restrictions of traditional transparent bridging. As a result, instead of being restricted to a rack or a pod, a particular VLAN can be extended to any port in any switch without any configuration overhead. From the perspective of the devices directly attached to it, Cisco FSS behaves as a transparent network - as if all the devices attached to the Cisco FSS were connected to the same switch.

Cisco FSS supports IEEE Data Center Bridging (currently CEE-DCB and soon IEEE-DCB when the standard is finalized), allowing the convergence of all the data center I/O technologies.

Cisco FSS offers significant benefits:

- Simplicity: Cisco FSS offers a validated foundation for building scalable data centers. Solution design guides and packaged product bundles make it easy to architect large network fabrics.
- High performance: Cisco FSS provides optimal use of the bandwidth available; all the shortest paths between any two devices can be used concurrently. Eliminating network tiers can reduce the number of switch hops, thereby reducing the latency in a FSS. The breakthrough hardware innovations of the Cisco Nexus 7000 F-Series Modules enable customers to build a large low-latency fabric.
- Agility: Cisco FSS provides isolation between the data center design and its users. The fabric can be modified in real time with no traffic interruption or server reconfiguration.
  - To gain more bandwidth, just add links between the switches.
  - To scale out the data center, just add switches.

Cisco FabricPath Switching System Use Cases

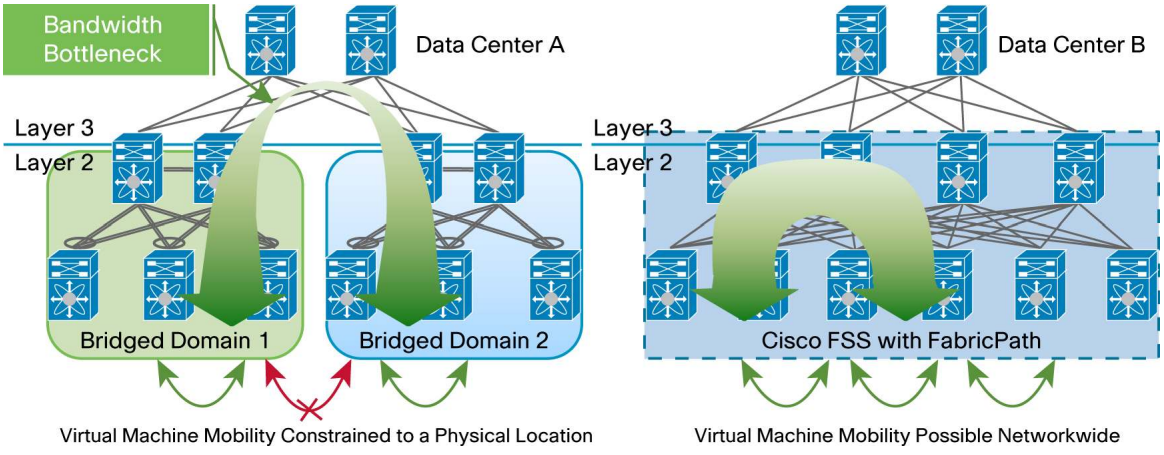
The value proposition of Cisco FSS - to create simple, scalable, and efficient Layer 2 domains - is applicable to many network scenarios. Some of these are presented here.

Typical Data Center Design

Today's data centers are generally divided in pods, like data center A in Figure 2. Within a pod, a bridged domain provides some operational flexibility. The best solution available is currently based on the Cisco virtual PortChannel (vPC), a Cisco NX-OS technology which allows each access switch to be connected redundantly to a pair of aggregation switches without relying on the Spanning Tree Protocol for redundancy and load balancing. (For more information about Cisco vPC technology, see http://www.cisco.com/en/US/prod/collateral/switches/ps9441/ps9402/white_paper_c11-516396.html.)

Data center B in Figure 3 represents a design based on Cisco FSS using exactly the same number of links and switches as data center A. The cabling layout is different, though. In data center A, each access switch is connected through a 4-port EtherChannel to two aggregation switches in a vPC domain. In data center B, each access switch is connected through a single uplink to four aggregation switches.

Figure 3. General-Purpose Data Center



This particular Cisco FSS design is just one of the many possible, selected in order to provide a simple illustration of the benefits that this new architecture brings in the following areas:

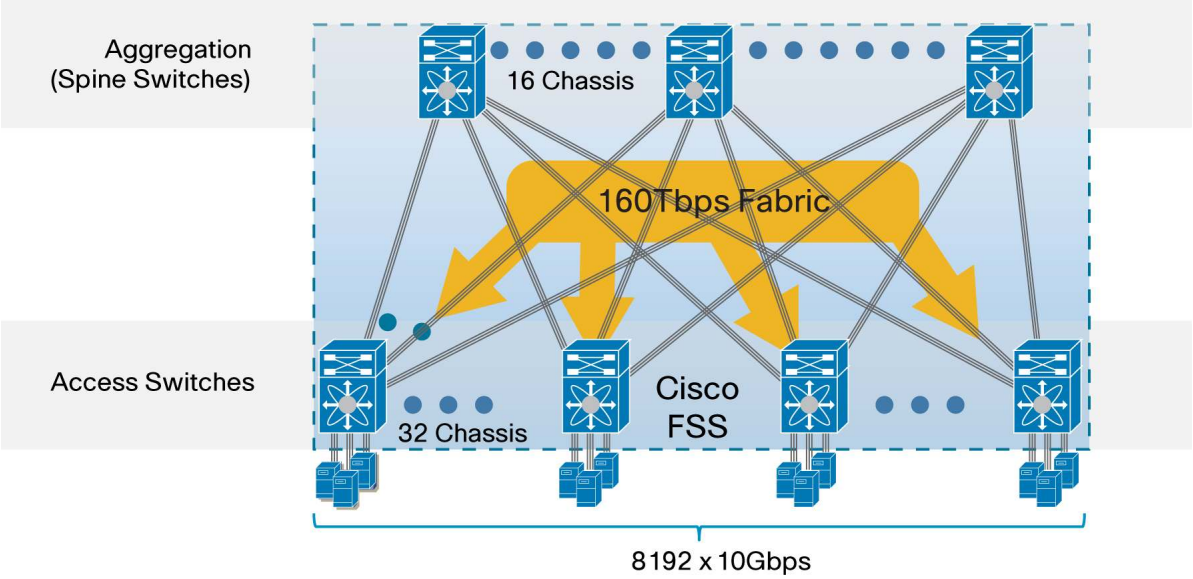
- Configuration
  - In data center A, each switch plays a different role and requires a different PortChannel, spanning-tree, or IP configuration. The configuration is generally not only switch-dependent but also VLAN-dependent.
  - Data center B, which relies on Cisco FSS, does not require VLAN or switch-specific configuration. Links are not even bundled into PortChannels.
- Layer 2 reachability
  - Access switches in data center A are separated into two pods that can communicate only with each other at Layer 3. Servers are statically assigned to a particular pod, and moving them is complex administratively - and may be impossible if the predetermined sizes of the pods are no longer adequate.
  - Access switches in data center B can reach each other at Layer 2, enabling simple administration and movement of virtual machines in seconds. A server does not have to be physically located in a particular pod, making provisioning easy and dynamic.
- Bandwidth
  - Each access switch in data center A has 40 Gbps of bandwidth available to its peers in the same pod, but shares a limited bandwidth through the upper layer when trying to access peers in the other pod.
  - In data center B, each access switch has 40 Gbps of bandwidth available to any peer in the network, using the shortest path possible.
- Resiliency
  - In data center A, the loss of an aggregation switch reduces by 50 percent the bandwidth available for the affected access switches.
  - The failure of an aggregation switch decreases the available bandwidth at the access by 25 percent in data center B.

High-Performance Computing

Data centers for high-performance computing are designed so that servers can communicate with each other with little oversubscription. In a spanning tree-based network, the Layer 3 boundary is generally located close to the root switch, allowing for provision of significant aggregated throughput for north/south traffic. However, lateral east-west traffic is typically highly oversubscribed because transparent bridging ultimately forwards traffic along a spanning tree, meaning that there can be only a single forwarding link between any two bridges. This restriction puts a hard limit on the bisectional bandwidth of the network.

Cisco FabricPath lifts this restriction using ECMP. The network represented in Figure 4 can be implemented with the first generation of Cisco Nexus 7000 F-Series I/O Modules, which will be available at the launch of Cisco FSS.

Figure 4. High-Performance Fabric with First-Generation Cisco Nexus 7000 F-Series I/O Modules



On each access switch, uplinks are all 16-port PortChannels leading to 16 different aggregation (or spine) switches; they provide 2.56 Tbps of connectivity to the fabric and bisectional bandwidth. The access switches also include 256 x 10-Gbps access ports that have nonblocking reachability to their peers anywhere in the network. This design, including 8192 10-Gbps ports, can be extended even further by introducing some form of oversubscription or by using upcoming 40- and 100-Gbps uplinks.

Conclusion

Some form of Layer 2 is required for the operation of modern, highly virtualized data centers. However, the scale of bridging domains is limited by some transparent bridging data-plane constraints. Cisco FSS, based on FabricPath technology, combines the flexibility of Layer 2 with the scaling and performance characteristics of routing and provides a solution that is simple, scalable, and efficient.

Americas Headquarters
Cisco Systems, Inc.
San Jose, CA

Asia Pacific Headquarters
Cisco Systems (USA) Pte. Ltd.
Singapore

Europe Headquarters
Cisco Systems International BV
Amsterdam, The Netherlands

Cisco has more than 200 offices worldwide. Addresses, phone numbers, and fax numbers are listed on the Cisco Website at **www.cisco.com/go/offices.**

Printed in USA C11-605488-00 06/10