White Paper

Scale Data Centers with Cisco FabricPath

What You Will Learn

Traditional network architectures are designed to provide high availability for static applications, while 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, a crucial element in Cisco Unified Fabric, is an innovative Cisco NX-OS Software technology that addresses these current and emerging requirements that are transforming the conception of data center networks. It brings the stability and performance of Layer 3 routing to Layer 2 switched networks to build a highly resilient and scalable Layer 2 fabric. Cisco FabricPath is 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, 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.

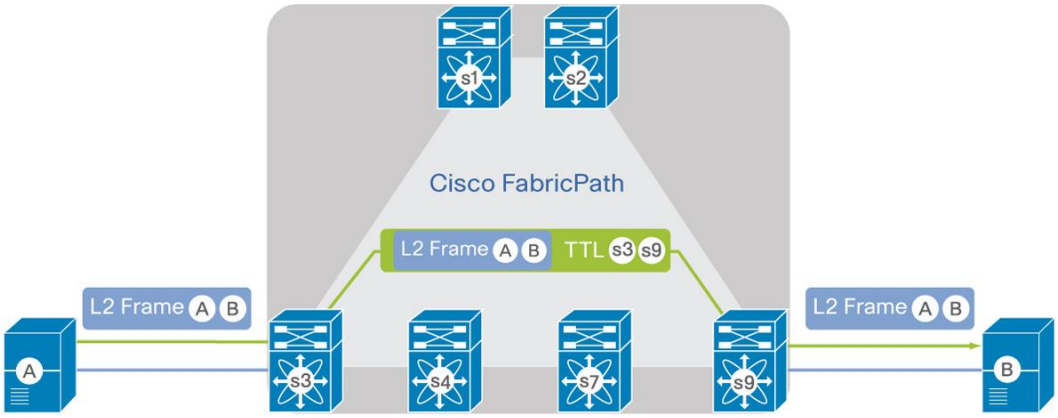
Introducing Layer 2 Fabrics with Cisco FabricPath

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 utilization. Cisco FabricPath delivers the foundation for building a scalable fabric - a network that itself looks like a single virtual switch from the perspective of its users. This property is achieved by providing optimal bandwidth between any two ports regardless of their physical locations. Also, because Cisco FabricPath does not suffer from the scaling restrictions of traditional transparent bridging, a particular VLAN can be extended across the whole fabric, reinforcing this notion of a single virtual switch. Note that if Cisco FabricPath is a Layer 2 technology, the fabric still maintains the Layer 3 capabilities of the Cisco Nexus® Family of switches and provides tight routing integration.

Cisco FabricPath Routes Traffic within the Fabric

Cisco FabricPath brings the stability and performance of routing to Layer 2. Cisco FabricPath takes over as soon as an Ethernet frame transitions from an Ethernet network (referred to as Classical Ethernet) to a Cisco FabricPath fabric. Ethernet bridging rules do not dictate the topology and the forwarding principles in a Cisco FabricPath fabric. The frame is encapsulated with a Cisco 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 Using Cisco FabricPath



The fundamental difference between Cisco FabricPath and Classical Ethernet is that with Cisco 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 resulting solution still provides the simple and flexible behavior of Layer 2, while using the routing mechanisms that make IP reliable and scalable.

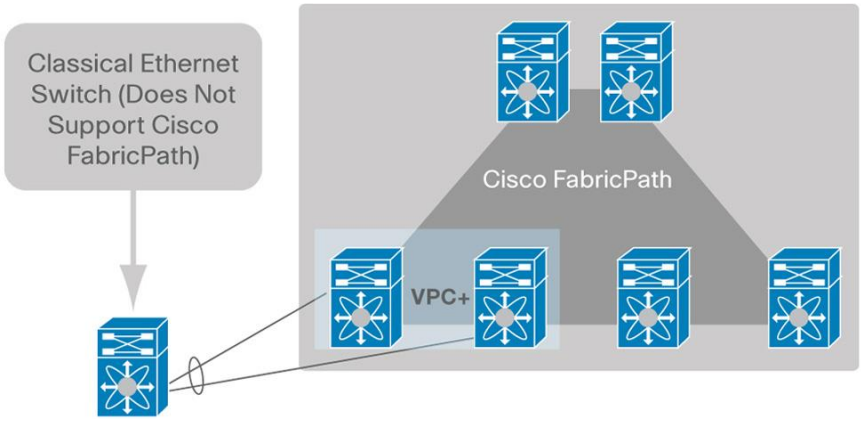
Cisco FabricPath introduces a dramatic change in the data plane, and dedicated hardware is required to implement the functions with low latency. The Cisco Nexus 7000 F-Series I/O modules and the Cisco Nexus 5500 Platform are capable of running Cisco FabricPath as well IEEE Data Center Bridging (DCB) and Fibre Channel over Ethernet (FCoE). Because Cisco Nexus switches also integrate transparently with Layer 3 routing, the resulting fabric can run all the different data center I/O technologies concurrently and efficiently.

Cisco FabricPath Benefits

Cisco FabricPath provides the following benefits:

- **Simplicity, 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. The Cisco FabricPath solution requires less combined configuration than an equivalent Spanning Tree Protocol-based network, further reducing the overall management cost.
  - 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 Cisco FabricPath can be modified as needed in a nondisruptive manner for the end stations.
  - The capabilities of Cisco 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 does not support Cisco FabricPath can be attached redundantly to two separate Cisco FabricPath bridges with enhanced virtual PortChannel (vPC+) technology, providing an easy migration path (Figure 2). Just like vPC[1], vPC+ relies on PortChannel technology to provide multipathing and redundancy without resorting to Spanning Tree Protocol.

Figure 2. Connecting Devices That Do Not Support Cisco FabricPath with vPC+



[1] For more information about Cisco NX-OS vPC, see http://www.cisco.com/en/US/prod/collateral/switches/ps9441/ps9402/white_paper_c11-516396.html.

- **Scalability based on proven technology**
  - Cisco 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 Cisco 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 Cisco 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 Cisco FabricPath can perform 16-way ECMP, which, when combined with 16-port 10-Gbps port channels, 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.

Cisco FabricPath Use Cases

The value proposition of Cisco FabricPath - to create simple, scalable, and efficient Layer 2 domains - is applicable to many network scenarios. Since Cisco FabricPath began shipping in October 2010, Cisco customers have been implementing a wide variety of network designs, from full-mesh to ring topologies. Some of these use cases are presented in this section.

Cisco FabricPath in a Typical Data Center Design

Cisco FabricPath is often associated with scalability and performance. However, today's data centers are generally built around small Layer 2 blocks, called pods. An example of such a network is data center A, represented in Figure 3. Within a pod, switching is handled by Cisco NX-OS vPC technology. vPC provides an active-active environment that does not depend on Spanning Tree Protocol and that converges quickly in the event of failure. Because vPC seems sufficient at this scale, it is important to note some other aspects of Cisco FabricPath that makes it attractive in this scenario:

- Cisco FabricPath is simple to configure and to manage. There is no need to identify a pair of peers or configure PortChannels. All the devices in the fabric have the same role and same minimal configuration.
- Cisco FabricPath is flexible and does not require a particular topology. Even if the network is currently cabled for the classic triangle vPC topology, Cisco FabricPath can accommodate any design that might be needed in the future.
- Cisco FabricPath does not use or even extend Spanning Tree Protocol. Even a partial introduction of Cisco FabricPath has a beneficial effect on the network because it segments the span of Spanning Tree Protocol. Because it is an optimization of Classical Ethernet, vPC still requires Spanning Tree Protocol on top of it to address certain scenarios.
- Cisco FabricPath can be extended easily without degrading operations. Adding a switch or a link in a Cisco FabricPath fabric does not result in a single frame loss. It is thus possible to start with a small network and extend it gradually, as needed.

Scaling the Typical Data Center Design with Cisco FabricPath

The previous section showed the benefits of introducing Cisco FabricPath as a direct replacement for vPC. This section demonstrates how Cisco FabricPath can bring additional significant scalability, availability, and flexibility improvement by reorganizing the cabling of an existing data center. Figure 3 shows two data centers using the exact same number of links and switches. In data center A, each access switch is connected through a 4-port PortChannel to two aggregation switches in a vPC domain. In data center B, which supports Cisco FabricPath, each access switch is instead connected through a single uplink to four aggregation switches.

Figure 3. General-Purpose Data Center

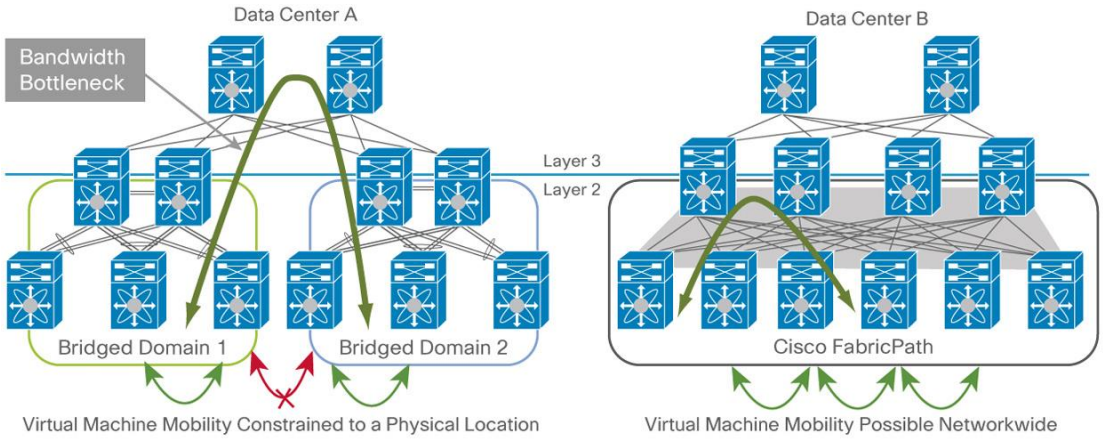


Table 1. Benefits of Cisco FabricPath in the Data Center

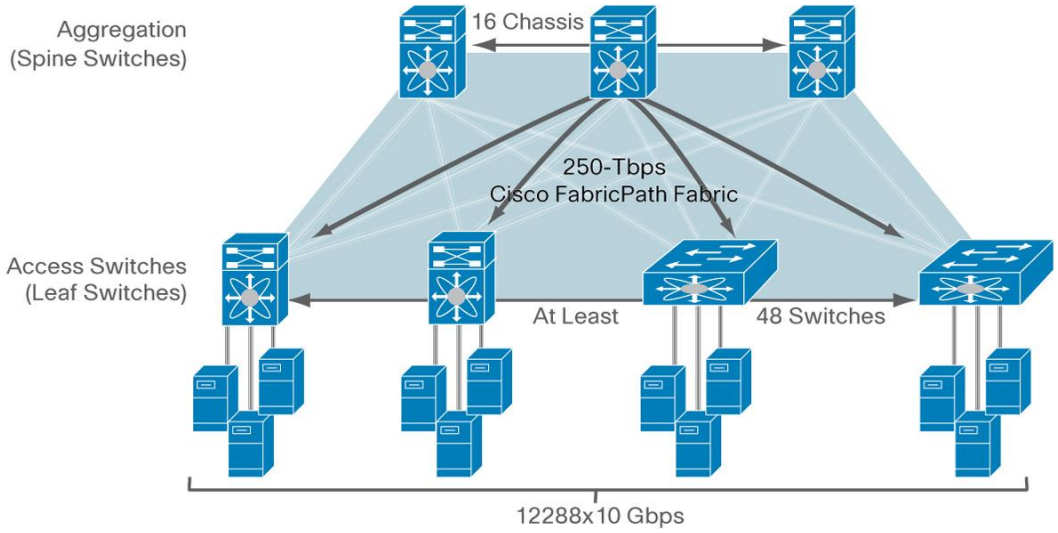
	Spanning Tree Protocol or vPC-Based Data Center A	Data Center B, Which Supports Cisco FabricPath
Configuration	Each switch plays a different role and requires a different PortChannel, spanning tree, or IP configuration. The configuration is generally not only switch dependent but also VLAN dependent.	Simpler: VLAN or switch-specific configuration is not needed. Links are not even bundled into PortChannels.
Layer 2 reachability	Access switches 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.	Wider: Access switches 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 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.	Higher: Each access switch has 40 Gbps of bandwidth available to any peer in the network, using the shortest path possible.
Availability	The loss of an aggregation switch reduces by 50% the bandwidth available for the affected access switches.	Higher: The failure of an aggregation switch decreases the available bandwidth at the access switch by only 25%.

Cisco FabricPath and 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. With current hardware, Cisco FabricPath supports 16-way ECMP. Therefore, up to 16 paths can be active between any two devices in the network. Because each of those 16 paths can itself be a 16-port PortChannel, the solution can in fact provide 2.56 Tbps of bisectional bandwidth. Figure 4 shows a possible network design for providing a nonblocking fabric taking advantage of Cisco FabricPath capabilities. This topology is called a Clos fabric and is simply an extreme case of data center B, shown in Figure 3. Here, a group of 16 aggregation devices (spine switches), provide 16 different paths between the access switches (leaf switches).

Figure 4. High-Performance Fabric with Cisco Nexus 7000 Series and Cisco Nexus 5500 Series



Using the Cisco Nexus 7000 18-Slot Switch and F2 -Series I/O modules on the spine, the current hardware can deliver a nonblocking Layer 2 fabric of more than 12,000 x 10-Gps ports, an industry record. With the Cisco Nexus 7000 18-Slot Switch at the leaf, the whole fabric could be implemented with as few as 64 devices. An arbitrary combination of Cisco Nexus 7000 Series and Cisco Nexus 5500 Series switches is, however, possible, providing a wide variety of implementation options.

The port scalability can be further increased by combining Cisco FabricPath with Cisco Nexus 2000 Series Fabric Extenders at the edge, or by introducing some oversubscription at the access layer.

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 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 for traditional, virtualized, or cloud environments.

For More Information

http://www.cisco.com/en/US/netsol/ns1151/index.html.

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 C11-605488-02 01/12