Getting Started with Cisco Nexus 9000 Series Switches in the Small-to-Midsize Commercial Data Center

Carly Stoughton
## Contents

**Introduction**  ........................................................................................................................................ 4  
- What You Will Learn ................................................................. 4  
- Disclaimer .................................................................................. 4  
- Why Cisco Nexus 9000 Series Switches ................................. 4  
- About Cisco Nexus 9000 Series Switches ............................... 5  

**Cisco ACI Readiness** ............................................................................................................................... 6  
- What Is Cisco ACI? ................................................................. 6  
- Converting Cisco Nexus 9000 NX-OS Mode to ACI Mode .......... 7  

**Data Center Design Evolution** ............................................................................................................... 8  
- Traditional Data Center Design ................................................ 8  
- Spanning Tree Support ............................................................ 9  
- Commercial Collapsed Designs ............................................... 9  
- Layer 2 versus Layer 3 Implications ....................................... 10  
- Cisco Layer 2 Design Evolution: Virtual PortChannels ............ 12  
- Virtual Overlays ...................................................................... 13  
- Spine-Leaf Data Center Design ............................................. 14  
- Overlay Design ....................................................................... 15  
- Additional Cisco Nexus 9000 Feature Support ....................... 17  
- Quality of Service ................................................................... 17  
- Multicast .................................................................................. 19  
- SNMP and Supported MIBs .................................................... 19  

**Sample Commercial Topologies** ......................................................................................................... 20  
- Cisco Nexus 9500 Platform Product Line ............................... 20  
- Cisco Nexus 9300 Platform Product Line ............................... 21  
- Design A: Two Switches ......................................................... 22  
- Design B: Two Aggregation and Two Access Switches .......... 23  
- Design C: Two Aggregation and Four Access Switches ........ 25  
- Design D: Four Aggregation and Four Access Switches with Spine-Leaf ........................................ 25  

**Integration with Existing Networks** ...................................................................................................... 26  
- Access Layer Device Connectivity ......................................... 26  
- Design A Integration ............................................................... 26  
- Design B Integration ............................................................... 27  
- Design C and D Integration .................................................... 28  
- Fabric Extender Support ......................................................... 30  
- Storage Design ....................................................................... 32  
- End-State Topology ............................................................... 35  

**Example: Cisco Nexus 9000 Series Design and Configuration** .............................................................. 36  
- Hardware and Software Specifications .................................. 36  
- Basic Network Setup ............................................................ 37  
- ECMP Routing Design .......................................................... 39  
- Server NIC Teaming Design and Configuration .................... 40  
- Overlay Design and Configuration ......................................... 44  
- Why VXLAN? ........................................................................... 44  
- How Does VXLAN Work? ....................................................... 45  
- Multicast Considerations ...................................................... 45  
- Spine (Aggregation) Multicast Configuration ......................... 46  
- Leaf (Access) Multicast Configuration ................................... 47  
- Using VXLAN in Commercial Topology ............................... 48  
- Service Design and Configuration ......................................... 50  
- Managing and Automating the Fabric .................................... 52  
- Cisco UCS Director .............................................................. 52  
- Cisco Prime Data Center Network Manager ......................... 53  

© 2018 Cisco and/or its affiliates. All rights reserved. This document is Cisco Public Information.
Conclusion ........................................................................................................................................... 55

Addendum A: Automation and Programmability ............................................................................ 55
  Automation Tools .......................................................................................................................... 55
  Automation with Puppet ............................................................................................................... 56

Helpful Links .................................................................................................................................... 56
  Cisco ACI Mode ............................................................................................................................. 56
  External Resources ....................................................................................................................... 56
  General Guides ............................................................................................................................... 56
  Licensing ....................................................................................................................................... 57
  Cisco Nexus 9000 Data Sheets ....................................................................................................... 57
  Other Cisco Products ..................................................................................................................... 57
  Other Overlays and Protocols ......................................................................................................... 57
  Software Release Notes ............................................................................................................... 57
  Transceivers .................................................................................................................................. 58
  Unicast and Multicast Routing ....................................................................................................... 58
  VXLAN ........................................................................................................................................... 58
Introduction

What You Will Learn

The Cisco Nexus® 9000 Series Switches product family makes the next generation of data center switching accessible to customers of any size. This white paper is intended for commercial customers new to Cisco Nexus 9000 Series Switches who are curious how the switches might be feasibly deployed in their data centers.

This white paper highlights the benefits of Cisco Nexus 9000 Series Switches, outlines several designs suitable for small-to-midsize customer deployments, discusses integration with existing networks, and walks through a topology validated by Cisco for Cisco Nexus 9000 Series Switches, complete with configuration examples.

The featured designs are practical for both entry-level insertion of Cisco Nexus 9000 Series Switches, and for existing or growing organizations that are scaling out the data center. The configuration examples transform the logical designs into tangible easy-to-use templates, simplifying deployment and operations for organizations with a small or growing IT staff.

Optionally, you can learn how to get started with the many powerful programmability features of Cisco NX-OS from a beginner’s perspective by seeing Addendum A, “Automation and Programmability,” at the end of this document. This white paper also provides many valuable links for further reading about protocols, solutions, and designs discussed. A thorough discussion on the programmability features of Cisco Nexus 9000 Series Switches are out of the scope of this document.

Disclaimer

Always refer to the Cisco website for the most recent information about software versions, supported configuration maximums, and device specifications at https://www.cisco.com/go/nexus9000.

Why Cisco Nexus 9000 Series Switches

Cisco Nexus 9000 Series Switches (shown in Figure 1) are ideal for small-to-midsize data centers, offering five key benefits: price, performance, port-density, programmability, and power efficiency.

Figure 1. Cisco Nexus 9000 Series Switches Product Family
Cisco Nexus 9000 Series Switches are cost-effective by taking a merchant-plus approach to switch design. Both silicon developed by Cisco-plus merchant silicon application-specific integrated circuits (ASICs) Trident II, or T2, power Cisco Nexus 9000 Series Switches. The T2 ASICs also deliver power efficiency gains. Additionally, Cisco Nexus 9000 Series Switches lead the industry in 10 and 40 Gb price-per-port densities. The cost-effective design approach coupled with a rich feature set make Cisco Nexus 9000 Series Switches a great fit for commercial data centers.

Licensing is greatly simplified on the Cisco Nexus 9000 Series Switches. At the time of this writing, there are two licenses available: the Enterprise Services Package license enables dynamic routing protocol and VXLAN support, and the Cisco Data Center Network Manager (DCNM) license provides a single-pane-of-glass GUI management tool for the entire data center network. Future licenses may become available as new features are introduced.

Cisco Nexus 9000 Series Switches offer powerful programmability features to drive emerging networking models including automation and DevOps, using tools, such as the Cisco NX-OS API (NX-API), as well as Python, Chef, and Puppet.

For small-to-midsize commercial customers, the Cisco Nexus 9000 Series Switches product family is the best platform for 1-to-10 Gb migration, 10-to-40 Gb migration, and is an ideal replacement for aging Cisco Catalyst® switches in the data center. The Cisco Nexus 9000 Series Switches can easily be integrated with existing networks. This white paper introduces a design as small as two Cisco Nexus 9000 Series Switches and provides a path to scale out your data center as it grows, highlighting both access and aggregation designs, and spine-leaf designs.

About Cisco Nexus 9000 Series Switches

The Cisco Nexus 9000 Series consists of larger Cisco Nexus 9500 Series modular switches and smaller Cisco Nexus 9300 Series fixed-configuration switches. The product offerings are discussed in detail later in this white paper.

Cisco provides two modes of operation for the Cisco Nexus 9000 Series. Customers can use Cisco NX-OS Software to deploy the Cisco Nexus 9000 Series in standard Cisco Nexus switch environments. Alternately, customers can use the Cisco Application Centric Infrastructure (ACI), which is ready for the hardware infrastructure to take full advantage of an automated, policy-based, systems management approach.

In addition to traditional Cisco NX-OS features, such as Cisco virtual PortChannels (vPC), PowerOn Auto Provisioning (POAP), and Nexus 2000 Series Fabric Extenders support, the single-image Cisco NX-OS running on Cisco Nexus 9000 Series Switches introduces several key new features:

- The intelligent Cisco NX-API provides administrators a way to manage the switch through remote procedure calls (JSON or XML) over HTTP and/or HTTPS, instead of accessing the Cisco NX-OS command line directly.
- Linux shell access enables the switch to be configured through Linux shell scripts, helping automate the configuration of multiple switches and helping ensure consistency among multiple switches.
- Continuous operation is maintained through cold and hot patching, which provides fixes between regular maintenance releases or between the final maintenance release and the end-of-maintenance release, in a nondisruptive manner (for hot patches).
- Virtual Extensible LAN (VXLAN) bridging and routing in hardware at full line rate facilitates and accelerates communication between virtual and physical servers. VXLAN is designed to provide the same Layer 2 Ethernet services as VLANs, but with greater flexibility and at a massive scale.
For more information about upgrades, see the Cisco Nexus 9000 Series NX-OS Software Upgrade and Downgrade Guide.

Cisco Nexus 9000 Series Switches support the plug-in for OpenStack Networking, also known as Neutron. With the plug-in you can build an infrastructure as a service (IaaS) network and to deploy a cloud network. With OpenStack, you can build an on-demand, self-service, multitenant computing infrastructure. However, implementing the OpenStack VLAN networking model across virtual and physical infrastructures can be difficult.

The OpenStack Networking extensible architecture supports plugins to configure networks directly. However, when you choose a network plug-in, only that plug-in’s target technology is configured. When you are running OpenStack clusters across multiple hosts with VLANs, a typical plug-in configures either the virtual network infrastructure or the physical network, but not both.

The Cisco Nexus plug-in solves this difficult problem by including support for configuring both the physical and virtual networking infrastructure. The Cisco Nexus plug-in accepts OpenStack Networking API calls and uses the Network Configuration Protocol (NETCONF) to configure Cisco Nexus devices, as well as Open vSwitch (OVS), which runs on the hypervisor.

The Cisco Nexus plug-in configures VLANs on both the physical and virtual network. It also allocates scarce VLAN IDs by deprovisioning them when they are no longer needed and reassigning them to new tenants whenever possible. VLANs are configured so that virtual machines running on different virtualization (computing) hosts that belong to the same tenant network transparently communicate through the physical network. In addition, connectivity from the computing hosts to the physical network is trunked to allow traffic only from the VLANs that are configured on the host by the virtual switch.

For more information, see OpenStack at Cisco.

This white paper focuses on basic design, integration of features, such as vPC, VXLAN, access layer device connectivity, and Layer 4-7 service insertion. Please refer to the addendum For an introduction to the advanced programmability features of Cisco NX-OS on Cisco Nexus 9000 Series Switches, see Addendum A, “Automation and Programmability,” in this document. Other features are out of the scope of this white paper.

Cisco ACI Readiness

What Is Cisco ACI?

The future of networking with Cisco Application Centric Infrastructure (ACI) is about providing a network that is deployed, monitored, and managed in a way that supports rapid application change. ACI does this through the reduction of complexity and a common policy framework that can automate provisioning and managing resources.

Cisco ACI works to solve the business problem of slow application deployment due to focus on primarily technical network provisioning and change management problems by enabling rapid deployment of applications to meet changing business demands. Cisco ACI provides an integrated approach by providing application-centric end-to-end visibility from a software overlay down to the physical switching infrastructure while accelerating and optimizing Layer 4-7 service insertion to build a system that brings the language of applications to the network.

Cisco ACI delivers automation, programmability, and centralized provisioning by allowing networks to be automated and configured based on business-level application requirements. Cisco ACI provides accelerated, cohesive deployment of applications across network and Layer 4-7 infrastructure, and enables visibility and management at the application level. Advanced telemetry for visibility into network health and simplified day two operations also opens up troubleshooting to the application itself.
The Cisco ACI diverse and open ecosystem is designed to plug in to any upper-level management or orchestration system and attract a broad community of developers. With the integration and automation of both Cisco and third-party Layer 4-7 virtual and physical service devices, you can use a single tool to manage the entire application environment.

With the Cisco ACI mode, customers can deploy networks based on application requirements in the form of policies, removing the need to translate the complexity of current network constraints. In tandem, Cisco ACI ensures security and performance while maintaining complete visibility into application health on both virtual and physical resources.

Figure 2 highlights how the network communication might be defined for a three-tier application from the Cisco ACI GUI interface. The network is defined in terms of the needs of the application by mapping out who is allowed to talk to whom and what they are allowed to talk about by defining a set of policies, or contracts, inside an application profile, instead of configuring lines and lines of command line interface (CLI) code on multiple switches, routers, and appliances.

Figure 2. Example of Three-Tier Application Policy

Converting Cisco Nexus 9000 NX-OS Mode to ACI Mode
This white paper features the Cisco Nexus 9000 Series Switches in NX-OS (standalone) mode. However, the Cisco Nexus 9000 hardware is ready for Cisco ACI. Cisco Nexus 9300 switches and many of the Cisco Nexus 9500 line cards can be converted to ACI mode.

The Cisco Nexus 9000 Series Switches are the foundation of the Cisco ACI architecture and provide the network fabric. A new operating system is used by the Cisco Nexus 9000 switches running in ACI mode. The switches are then coupled with a centralized controller called the Cisco Application Policy Infrastructure Controller (APIC) and its open API. The Cisco APIC is the unifying point of automation, telemetry, and management for the Cisco ACI fabric, enabling an application policy model approach to the data center.
Conversion from standalone NX-OS mode to ACI mode on the Cisco Nexus 9000 switch is out of the scope of this white paper. For more information about the ACI mode on the Cisco Nexus 9000 Series Switches, see the Cisco ACI homepage.

Data Center Design Evolution

Cisco Nexus 9000 Series Switches can be used in a number of potential designs in a midsize commercial data center. Cisco Nexus 9000 Series Switches can be used as end-of-row or top-of-rack access layer switches, as aggregation or core switches in traditional hierarchical two- or three-tier network designs, or deployed in a modern leaf-spine architecture. This white paper discusses both access and aggregation designs and leaf-spine designs.

Traditional Data Center Design

Traditional data centers are built on a three-tier architecture with core, aggregation, and access layers, or a two-tier collapsed core with the aggregation and core layers combined in one layer, as shown in Figure 3. This architecture accommodates a north-south traffic pattern where client data comes in from a WAN or the Internet to be processed by a server in the data center and is then pushed back out of the data center. This is common for applications, such as web services, where most communication is between an external client and an internal server. The north-south traffic pattern permits hardware oversubscription, because most traffic is funneled in and out through the lower-bandwidth WAN or Internet bottleneck.

Figure 3. Traditional Three-Tier Design

More common in a midsized commercial data center is a collapsed two-tier design where there are two main design decisions:

- Layer 2/Layer 3 boundary is at the aggregation layer.
- Layer 3 is down to the access layer (routed access).
The factors used in making a decision about the Layer 2-Layer 3 boundary is discussed with the designs, but some of the major considerations of each design are shown in Table 1:

Table 1. Major Considerations of Each Design

<table>
<thead>
<tr>
<th>Consideration</th>
<th>Layer 3 at Core</th>
<th>Layer 3 at Access</th>
</tr>
</thead>
<tbody>
<tr>
<td>Multipathing</td>
<td>One active path per VLAN due to spanning tree mode between access and core switches</td>
<td>Equal cost multipathing through dynamic routing protocol between access and core switches</td>
</tr>
<tr>
<td>STP</td>
<td>More Layer 2 links, therefore more loops and links to block</td>
<td>No STP running north of access layer</td>
</tr>
<tr>
<td>Layer 2 reach</td>
<td>Greater Layer 2 reachability for workload mobility and clustering applications</td>
<td>Layer 2 adjacency limited to devices connected to the same access layer switch</td>
</tr>
<tr>
<td>Convergence time</td>
<td>Spanning tree mode generally has a slower convergence time than a dynamic routing protocol</td>
<td>Dynamic routing protocols generally have a faster convergence time than STP</td>
</tr>
</tbody>
</table>

Spanning Tree Support
The Cisco Nexus 9000 Series supports two spanning tree modes: Rapid per VLAN spanning tree plus (PVST+), which is the default mode, and MST (Multiple Spanning Tree).

The Rapid PVST+ protocol is the IEEE 802.1w standard, Rapid Spanning Tree Protocol (RSTP), implemented on a per-VLAN basis. Rapid PVST+ interoperates with the IEEE 802.1Q VLAN standard, which mandates a single STP instance for all VLANs, rather than per VLAN. Rapid PVST+ is enabled by default on the default VLAN (VLAN1) and on all newly created VLANs on the device. Rapid PVST+ interoperates with devices that run legacy IEEE 802.1D STP. RSTP is an improvement on the original STP standard, 802.1D, which allows faster convergence.

MST maps multiple VLANs onto a spanning tree instance, with each instance having a spanning tree topology independent of other instances. This architecture provides multiple forwarding paths for data traffic, enables load balancing, and reduces the number of STP instances required to support a large number of VLANs. MST improves the fault tolerance of the network because a failure in one instance does not affect other instances. The MST mode provides rapid convergence through explicit handshaking because each MST instance uses the IEEE 802.1w standard. MST mode improves spanning tree operation and maintains backward compatibility with the original 802.1D Spanning Tree and Rapid per-VLAN spanning tree (Rapid PVST+) protocol.

Commercial Collapsed Designs
The following Figures 4, 5, and 6 depict two-tier designs, with Layer 3 at the aggregation layer and Layer 2 at the access and up to the aggregation layer, found frequently in commercial data centers. Some customers may not require a complete two-tier design and often deploy a pair of switches as the collapsed data center architecture.

Figure 4. Traditional One-Tier Collapsed Design
Customers can also have the small collapsed design replicated to another rack or building with a Layer 3 connection between the pods (Figure 7).

**Figure 5.** Traditional Two-Tier Small Access-Aggregation Design

<table>
<thead>
<tr>
<th>Layer 3</th>
<th>Layer 2</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Collapsed Core Layer:</strong> Provides connectivity between access switches, external networks, policy enforcement, and Layer 4-7 service insertion</td>
<td></td>
</tr>
<tr>
<td><strong>Access Layer:</strong> Provides connectivity for servers</td>
<td></td>
</tr>
</tbody>
</table>

**Figure 6.** Traditional Two-Tier Medium Access-Aggregation Design

<table>
<thead>
<tr>
<th>Layer 3</th>
<th>Layer 2</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Collapsed Core Layer:</strong> Provides connectivity between access switches, external networks, policy enforcement, and Layer 4-7 service insertion</td>
<td></td>
</tr>
<tr>
<td><strong>Access Layer:</strong> Provides connectivity for servers</td>
<td></td>
</tr>
</tbody>
</table>

**Layer 2 versus Layer 3 Implications**

Data center traffic patterns are changing. Today, more traffic moves east-west from server to server through the access layer as servers need to talk to each other and consume services in the data center. This shift is primarily driven by the evolution of application design. Many modern applications need to talk to each other in the data center. Applications driving this shift include big data’s often-distributed processing design (for example, Hadoop), live virtual machine or workload migration (for example, VMware vMotion), server clustering (for example, Microsoft Cluster Services), and multitier applications. The oversubscribed hardware now isn’t sufficient for the east-west 10-to-10 Gb communications. Additionally, east-west traffic is often forced up through the core or aggregation layer, taking a suboptimal path.
Spanning tree is another hindrance in the traditional three-tier data center design. Spanning tree is required to block loops in flooding Ethernet networks so that frames are not forwarded endlessly. Blocking loops means blocking links, leaving only one active path (per VLAN). Blocked links severely impact available bandwidth and oversubscription. This can also force traffic to take a suboptimal path, because spanning tree can block a more desirable path.

**Figure 8.** Suboptimal Path between Servers in Different Pods Due to Spanning Tree Blocked Links

Addressing these issues can include upgrading hardware to support 40 or 100 Gb interfaces, bundling links into Cisco Port Channels to appear as one logical link to spanning tree or moving the Layer 2-Layer 3 boundary down to the access layer to limit the reach of spanning tree. Using a dynamic routing protocol between the two layers allows all links to be active, and have fast reconvergence and equal-cost multipathing (ECMP).

**Figure 9.** Two-Tier Routed Access Layer Design

The trade-off in moving Layer 3 routing to the access layer in a traditional Ethernet network is this limits Layer 2 reachability. Applications, such as virtual machine workload mobility and some clustering software, require Layer 2 adjacency between source and destination servers. By routing at the access layer, only servers connected to the same access switch with the same VLANs trunked down are Layer 2 adjacent. However, the alternative of spanning a VLAN across the entire data center for reachability is problematic due to Ethernet’s broadcast nature and spanning tree reconvergence events.
Cisco Layer 2 Design Evolution: Virtual PortChannels

Over the past few years many customers have sought ways to move past the limitations of spanning tree. The first step on the path to modern data centers based on Cisco Nexus solutions came in 2008 with the advent of Cisco virtual PortChannels (vPC). A vPC allows a device to connect to two different physical Cisco Nexus switches using a single logical Cisco PortChannel interface.

Prior to Cisco vPC, Cisco PortChannels generally had to terminate on a single physical switch. Cisco vPC gives the device active-active forwarding paths. Because of the special peering relationship between the two Cisco Nexus switches, spanning tree does not see any loops, leaving all links active. To the connected device, the connection appears as a normal Cisco PortChannel interface, requiring no special configuration. The industry standard term is called Multi-Chassis EtherChannel; the Cisco Nexus implementation is called vPC.

Cisco vPC deployed on a spanning tree Ethernet network is a very powerful way to curb the number of blocked links, and thereby increasing available bandwidth. Cisco vPC on Cisco Nexus 9000 switches is a great solution for commercial customers, and those satisfied with current bandwidth, oversubscription, and Layer 2 reachability requirements.
Two of the sample small-to-midsized traditional commercial topologies are depicted using vPCs in Figures 12 and 13. These designs leave the Layer 2-Layer 3 boundary at the aggregation to permit broader Layer 2 reachability, but all links are active because spanning tree does not see any loops to block. Details about the special vPC peering relationship between the Nexus switches is discussed later in this white paper.

**Figure 12.** Traditional One-Tier Collapsed Design with Cisco vPC

![Image of Traditional One-Tier Collapsed Design with Cisco vPC]

**Figure 13.** Traditional Two-Tier Design with Cisco vPC

![Image of Traditional Two-Tier Design with Cisco vPC]

**Virtual Overlays**

Customers with a two- or three-tier design looking to route to the access layer yet still maintain Layer 2 reachability between servers can take the next step in the data center evolution by implementing a virtual overlay fabric. In a Cisco Nexus 9000 fabric design, dynamic routing is configured between switches down to the access layer so that all links are active. This eliminates the need for spanning tree on the fabric and enables ECMP through the dynamic routing protocol.

A virtual overlay fabric called Virtual Extensible LAN (VXLAN) can be used to provide Layer 2 adjacencies over the Layer 3 fabric for servers and other devices that require Layer 2 reachability in the Cisco Nexus 9000 design. Combining VXLAN and a dynamic routing protocol offers the benefits of an intelligent Layer 3 routing protocol, yet it can also provide Layer 2 reachability across all access switches for applications such as virtual machine workload mobility and clustering.
The limitations of spanning tree in three-tier designs and the needs of modern applications are driving a shift in network design towards a spine-leaf or access and aggregation architecture. Two- and three-tier designs are still a valid and prevalent architecture; spine-leaf architecture provides another easily integrated option.

**Spine-Leaf Data Center Design**

Spine-leaf topologies are based on a Clos network architecture. The term originates from Charles Clos at Bell Laboratories, who published a paper in 1953 that describes a mathematical theory of a multipathing, nonblocking, multiple-stage network topology to switch telephone calls through.

Today, Clos’s original thoughts on design are applied to the modern spine-leaf topology. Spine-leaf architecture is typically deployed as two layers: spines (such as an aggregation layer), and leaves (such as an access layer). Spine-leaf topologies provide high-bandwidth, low-latency, nonblocking server-to-server connectivity.

Leaf (aggregation) switches are what provide devices access to the fabric (the network of spine and leaf switches) and are typically deployed at the top of the rack. Generally, devices connect to the leaf switches. Devices can include servers, Layer 4-7 services (firewalls and load balancers), and WAN or Internet routers. Leaf switches do not connect to other leaf switches (unless running vPC in standalone NX-OS mode). However, every leaf should connect to every spine in a full mesh. Some ports on the leaf are used for end devices (typically 10 Gb), and some ports are used for the spine connections (typically 40 Gb).

Spine (aggregation) switches are used to connect to all leaf switches and are typically deployed at the end or middle of the row. Spine switches do not connect to other spine switches. Spines serve as a backbone interconnect for leaf switches. Generally, spines only connect to leaves, but when integrating Cisco Nexus 9000 switches with an existing environment, it is acceptable to connect other switches, services, or devices to the spines.
All devices connected to the fabric are an equal number of hops away from one another. This delivers predictable latency and high bandwidth between servers. Figure 15 depicts a simple two-tier design.

**Figure 15.** Two-Tier Design and Connectivity

![Diagram of Two-Tier Design and Connectivity](image)

Another way to think about spine-leaf architecture is by thinking of the spines as a central backbone with all leaves branching off the spine like a star. Figure 16 depicts this logical representation, which uses identical components in an alternate visual mapping configuration.

**Figure 16.** Logical Representation of Two-Tier Design

![Diagram of Logical Representation of Two-Tier Design](image)

With Cisco Nexus 9000 Series Switches, small-to-midsize commercial customers can start with a few switches and implement a pay-as-you-grow model. When more access ports are needed, more leaves (access switches) can be added. When more bandwidth is needed, more spines (aggregation switches) can be added. The Cisco Nexus 9000 switches are well suited for traditional designs, spine-leaf, or any combination of existing topologies.

**Overlay Design**

Virtual network overlays partition a physical network infrastructure into multiple logically isolated networks that can be individually programmed and managed to deliver optimal network requirements.

Small-to-midsize commercial customers may require mobility between data centers, in different pods in a single data center, and across Layer 3 network boundaries. Virtual network overlays make mobility and Layer 2 reachability possible.

Cisco has had a role in developing multiple overlay networks, each solving a different problem by providing an innovative solution. For example, Cisco Overlay Transport Virtualization (OTV) provides cross-data center mobility over Layer 3 Cisco Data Center Interconnect (DCI) networks using MAC-in-IP encapsulation.
In Figure 17 two overlays in use are shown: Cisco OTV on a Cisco ASR 1000 Router for data center–to–data center connectivity, and VXLAN in the data center. Both provide Layer 2 reachability and extension. Cisco OTV is also available on the Cisco Nexus 7000 Series Switch.

**Figure 17. Virtual Overlay Networks Providing Dynamic Reachability for Applications**

For more information on overlays in general, see the "Data Center Overlay Technologies" white paper. For more information on Cisco OTV, see the Cisco OTV webpage.

Now, the IETF-proposed VXLAN overlay on Cisco Nexus 9000 Series Switches provides Layer 2 connectivity extension across IP transport in the data center, and easy integration between VXLAN and non-VXLAN infrastructures. VXLAN enables scalable virtualized and multitenant data center designs over a shared common physical infrastructure.

Beyond a simple overlay sourced and terminated on the switches, the Cisco Nexus 9000 switch can act as a hardware-based VXLAN gateway. VXLAN is increasingly popular for virtual networking in the hypervisor for virtual-machine-to-virtual-machine communication, not just switch to switch. However, many devices are not capable of supporting VXLAN, such as legacy hypervisors, physical servers, and service appliances, such as firewalls, load balancers, and storage devices. Those devices need to continue to reside on classic VLAN segments. It is not uncommon that virtual machines in a VXLAN segment need to access services provided by devices in a classic VLAN segment. The Cisco Nexus 9000 switch, acting as a VXLAN gateway, can provide the necessary translation.
Today, a data center interconnect (DCI), such as Cisco OTV, must be provided by an outside device. In the next phase of VXLAN, the Cisco Nexus 9000 switch supports Border Gateway Protocol (BGP) Ethernet VPN (EVPN). EVPN is a next-generation Layer 2 VPN solution based on a BGP control plane, which can be used as a data center interconnect to distribute MAC address reachability information across the core or provider IP network.

For more information, see the paper “IETF BGP MPLS Based Ethernet VPN” (draft).

Basic VXLAN bridging for Layer 2 extension between leaf, or access, switches is the featured design in this white paper, and is discussed in more depth later.

**Additional Cisco Nexus 9000 Feature Support**

While many new and important features have been discussed, it is important to highlight the additional features support on the Cisco Nexus 9000 Series Switches, such as quality of service (QoS), multicast, and Simple Network Management Protocol (SNMP) and supported MIBs.

**Quality of Service**

New applications and evolving protocols are changing the QoS demands in modern data centers. High-frequency trading floor application are very sensitive to latency, while high-performance computing is typically characterized by bursty, many-to-one east-west traffic flows. Applications, such as storage, voice, and video, also require special and different treatment in any data center.

Like the other members of the Cisco Nexus family, the Cisco Nexus 9000 Series Switches support QoS configuration through Cisco Modular QoS CLI (MQC), which involves identifying traffic using class maps based on things such as protocol or packet header markings, defining how to treat different classes through policy maps by possibly marking packets, queuing, or scheduling, and lastly applying policy maps to either interfaces or the entire system using the service policy command. QoS is enabled by default and does not require a license.

Unlike Cisco IOS®, Cisco NX-OS on the Cisco Nexus 9000 switch uses three different types of policy maps depending on the QoS feature you are trying to implement. The three types of Cisco NX-OS QoS and their primary usages include:

- **Type QoS:**
  - Classification
  - Marking
  - Policing
• **Type Queuing:**
  - Buffering
  - Queuing
  - Scheduling

• **Type Network-QoS:**
  - Systemwide settings
  - Congestion control
  - Pause behavior

Type network-QoS policy-maps are applied only systemwide, whereas type QoS and type queuing can each be applied simultaneously on a single interface, one each ingress and egress for a total of four QoS policies per interface, if required.

On the Cisco Nexus 9500 switch, ingress traffic can be classified based on attributes such as header address fields, 802.1p CoS (class of service), IP precedence, differentiated services code point (DSCP) values, or matched based on access control lists (ACLs). After classified, traffic can be assigned to one of four QoS groups. The QoS groups serve as an internal identification of traffic classes that is used for subsequent QoS processes as packets go through the system. On ingress, the Cisco Nexus 9500 switch can also remark CoS, IP precedence, and DSCP, and perform class-based policing.

Egress, the Cisco Nexus 9500 switch uses a simple queuing architecture and does not use virtual output queues (VoQs) on ingress line card ports like other Cisco Nexus platforms. If egress ports are congested, packets are directly queued in the buffer of the egress line card, greatly simplifying system buffer management and queuing implementation. The Cisco Nexus 9500 switch can support up to six traffic classes on egress (four user-defined classes identified by QoS-group IDs, a CPU control traffic class, and a SPAN traffic class). Each user-defined class can have a unicast queue and a multicast queue per egress port. The 12 MB buffer on a network forwarding engine (NFE), the Trident II ASIC (application-specific integrated circuit), is shared among the local ports. The buffer's resources are dynamically shared by ingress and egress traffic. The switch software has a mechanism to meter and limit buffer utilization per egress port. This ensures that no single port can consume more than its fair share of the buffer memory leading to buffer starvation of other ports.

Additionally, egress, the Cisco Nexus 9500 switch provides up to three priority-based flow control no-drop queues, up to three strict priority queues, and the ability to enable weighted random early detection (WRED) or Explicit Congestion Notification (ECN) instead of the default tail drop queue management.

The Cisco ACI ready leaf line cards have an additional 40 MB buffer in each of their Cisco ACI leaf engines (ALEs). Allocated to fabric bound traffic is 10 MB of the buffer. The remaining 30 MB is allocated to egress traffic from fabric modules and locally switched traffic from higher-speed ingress port to a lesser-speed egress port.

This 30 MB buffer is used for extended output queues for unicast traffic. The NFE communicates the unicast queue status to the Cisco ALE through an out-of-band flow control (OOBFC) signaling channel. When an egress queue exceeds the configured threshold, the NFE sends an OOBFC signal to instruct the Cisco ALE to stop forwarding traffic for this queue and start queuing packets in its own buffer. After receiving this signal, the Cisco ALE starts to form the extended output queue for this traffic class on the given egress port. When the egress queue length is reduced to the configured restart threshold, the NFE sends another OOBFC signal to direct the Cisco ALE to resume transmitting traffic for this particular queue.
For more details on unicast and multicast packet forwarding through the ASICs, see the white paper, “Cisco Nexus 9500 Series Switches Architecture.”

For other QoS considerations in a mixed LAN and storage environment, for example, see the Cisco Nexus 9000 Series NX-OS Quality of Service Configuration Guide, Release 6.0.

**Multicast**

The Cisco Nexus 9500 platform provides high-performance, line rate Layer 2 and Layer 3 multicast throughput with low latency at scale at up to 8000 multicast routes with multicast routing enabled. The Cisco Nexus 9500 platform optimizes multicast lookup and replication by performing IP-based lookup for Layer 2 and Layer 3 multicast to avoid common aliasing problems in multicast IP-to-MAC encoding.

The Nexus 9500 platform supports Internet Group Management Protocol (IGMP) versions 1, 2, and 3, Protocol-Independent Multicast (PIM) sparse mode, anycast rendezvous point, and Multicast Source Discovery Protocol (MSDP).

**SNMP and Supported MIBs**

The Simple Network Management Protocol (SNMP) provides a standard framework and language to manage devices on the network, including the Cisco Nexus 9000 Series Switches. The SNMP manager controls and monitors the activities of network devices using SNMP. The SNMP agent lives in the managed device and reports the data to the managing system. The agent on the Cisco Nexus 9000 Series Switches must be configured to talk to the manager. MIBs are a collection of managed objects by the SNMP agent. The Cisco Nexus 9000 Series Switches supports SNMP v1, v2c, and v3. The Cisco Nexus 9000 Series Switches can also support SNMP over IPv6.

SNMP generates notifications about the Cisco Nexus 9000 Series Switches to send to the manager, for example, notifying the manager when a neighboring router is lost. A trap notification is sent from the agent on the Cisco Nexus 9000 switch to the manager, who does not acknowledge the message. A notification to inform is acknowledged by the manager. Table 2 provides the SNMP traps enabled by default.

**Table 2.** SNMP Traps Enabled by Default on Cisco Nexus 9000 Series Switches

<table>
<thead>
<tr>
<th>Trap Type</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>generic</td>
<td>: coldStart</td>
</tr>
<tr>
<td>generic</td>
<td>: warmStart</td>
</tr>
<tr>
<td>entity</td>
<td>: entity_mib_change</td>
</tr>
<tr>
<td>entity</td>
<td>: entity_module_status_change</td>
</tr>
<tr>
<td>entity</td>
<td>: entity_power_status_change</td>
</tr>
<tr>
<td>entity</td>
<td>: entity_module_inserted</td>
</tr>
<tr>
<td>entity</td>
<td>: entity_module_removed</td>
</tr>
<tr>
<td>entity</td>
<td>: entity_unrecognised_module</td>
</tr>
<tr>
<td>entity</td>
<td>: entity_fan_status_change</td>
</tr>
<tr>
<td>entity</td>
<td>: entity_power_out_change</td>
</tr>
<tr>
<td>link</td>
<td>: linkDown</td>
</tr>
<tr>
<td>link</td>
<td>: linkup</td>
</tr>
<tr>
<td>link</td>
<td>: extended-linkDown</td>
</tr>
<tr>
<td>link</td>
<td>: extended-linkUp</td>
</tr>
<tr>
<td>link</td>
<td>: cieLinkDown</td>
</tr>
<tr>
<td>link</td>
<td>: cieLinkUp</td>
</tr>
<tr>
<td>link</td>
<td>: delayed-link-state-change</td>
</tr>
<tr>
<td>Trap Type</td>
<td>Description</td>
</tr>
<tr>
<td>-----------</td>
<td>---------------------------------------</td>
</tr>
<tr>
<td>rf</td>
<td>redundancy_framework</td>
</tr>
<tr>
<td>license</td>
<td>notify-license-expiry</td>
</tr>
<tr>
<td>license</td>
<td>notify-no-license-for-feature</td>
</tr>
<tr>
<td>license</td>
<td>notify-licensefile-missing</td>
</tr>
<tr>
<td>license</td>
<td>notify-license-expiry-warning</td>
</tr>
<tr>
<td>upgrade</td>
<td>UpgradeOpNotifyOnCompletion</td>
</tr>
<tr>
<td>upgrade</td>
<td>UpgradeJobStatusNotify</td>
</tr>
<tr>
<td>rmon</td>
<td>risingAlarm</td>
</tr>
<tr>
<td>rmon</td>
<td>fallingAlarm</td>
</tr>
<tr>
<td>rmon</td>
<td>hcRisingAlarm</td>
</tr>
<tr>
<td>rmon</td>
<td>hcFallingAlarm</td>
</tr>
<tr>
<td>entity</td>
<td>entity_sensor</td>
</tr>
</tbody>
</table>

On a Cisco Nexus 9000 switch, Cisco NX-OS supports stateless restarts and is also aware of virtual routing and forwarding (VRF). Using SNMP does not require a license.

For a list of supported MIBs on the Cisco Nexus 9000 Series Switches, see this list.

For configuration help, see the section “Configuring SNMP” in the Cisco Nexus 9000 Series NX-OS System Management Configuration Guide, Release 6.0.

**Sample Commercial Topologies**

A spine-leaf or access-aggregation architecture using 10 and 40 Gigabit Ethernet is designed to handle the shift to east-west traffic patterns in the data center, limit the scope of spanning tree by routing between the layers, provide a loop-free all-links-active topology, perform ECMP, and optionally provide a virtual overlay for Layer 2 reachability using VXLAN.

A spine-leaf design is not required when deploying Cisco Nexus 9000 switches, and the topology can also be thought of as an access-aggregation architecture, as referred to interchangeably throughout this white paper.

The Cisco Nexus 9000 Series Switches portfolio provides a wide array of switches to choose from. Design considerations include desired throughput, oversubscription ratio, estimated growth, server traffic patterns, application requirements, and more.

This section is about the Cisco Nexus 9000 Series Switches portfolio and describes several designs suitable for small-to-midsize commercial data centers. This section also shows simple network topologies. Sample Commercial Topologies demonstrates how these design options can be integrated with an existing network later in this white paper.

**Cisco Nexus 9500 Platform Product Line**

The Cisco Nexus 9500 platform modular switches are typically deployed as spines (aggregation or core switches) in commercial data centers. Figure 19 lists the line cards available in the Cisco Nexus 9500 chassis in NX-OS standalone mode at the time of writing. (Line cards for only Cisco ACI have been removed.)
Cisco Nexus 9500 Line Card Types

<table>
<thead>
<tr>
<th>Line Cards</th>
<th>Ports</th>
<th>Mode</th>
<th>Fabric Modules</th>
<th>Chassis Support</th>
</tr>
</thead>
<tbody>
<tr>
<td>X9600</td>
<td>36p QSFP+</td>
<td>NX-OS</td>
<td>6</td>
<td>N9504, N9508</td>
</tr>
<tr>
<td>X9636PQ</td>
<td>48p 1/10G SFP+ and 4p QSFP+</td>
<td>NX-OS, ACI</td>
<td>3</td>
<td>N9504, N9508, N9516</td>
</tr>
<tr>
<td>X9556TX</td>
<td>48p 1/10G-T and 4p QSFP+</td>
<td>NX-OS, ACI</td>
<td>3</td>
<td>N9504, N9508, N9516</td>
</tr>
<tr>
<td>X9536PQ</td>
<td>36p QSFP+ (1:5:1)</td>
<td>NX-OS, ACI</td>
<td>3</td>
<td>N9504, N9508, N9516</td>
</tr>
<tr>
<td>X9400</td>
<td>48p 1/10G SFP+ and 4p QSFP+</td>
<td>NX-OS</td>
<td>4</td>
<td>N9504, N9508, N9516</td>
</tr>
<tr>
<td>X9432PQ</td>
<td>32p QSFP+</td>
<td>NX-OS</td>
<td>4</td>
<td>N9504, N9508, N9516</td>
</tr>
</tbody>
</table>

Note: For the latest product information, see the Cisco Nexus 9500 data sheets. Line card availability may have changed since the time of this writing.

Cisco Nexus 9300 Platform Product Line

The Cisco Nexus 9300 platform fixed-configuration switches are typically deployed as leaves (access switches). Some Cisco Nexus 9300 switches are semimodular by providing an uplink module slot for additional ports. The following figure lists the Cisco Nexus 9300 chassis that support NX-OS standalone mode available at the time of writing.

Cisco Nexus 9300 Chassis Offerings

<table>
<thead>
<tr>
<th>Line Cards</th>
<th>Ports</th>
<th>Mode</th>
<th>RU</th>
<th>Uplink Module</th>
</tr>
</thead>
<tbody>
<tr>
<td>N9396PX</td>
<td>48p 1/10G SFP+ and 12p QSFP+</td>
<td>NX-OS, ACI</td>
<td>2</td>
<td>Y</td>
</tr>
<tr>
<td>N9396TX</td>
<td>48p 1/10G-T and 12p QSFP+</td>
<td>NX-OS, ACI</td>
<td>2</td>
<td>Y</td>
</tr>
<tr>
<td>N93128TX</td>
<td>96p 1/10G-T and 8p QSFP+</td>
<td>NX-OS, ACI</td>
<td>3</td>
<td>Y</td>
</tr>
<tr>
<td>N9372PX</td>
<td>48p 1/10G SFP+ and 6p QSFP+</td>
<td>NX-OS, ACI</td>
<td>1</td>
<td>N</td>
</tr>
<tr>
<td>N9372TX</td>
<td>48p 1/10G-T and 6p QSFP+</td>
<td>NX-OS, ACI</td>
<td>1</td>
<td>N</td>
</tr>
<tr>
<td>N9332PQ</td>
<td>32p QSFP+</td>
<td>NX-OS, ACI</td>
<td>1</td>
<td>N</td>
</tr>
</tbody>
</table>

Note: For the latest product information, see the Cisco Nexus 9300 data sheets. Switch availability may have changed since the time of this writing.
Design A: Two Switches
Small commercial customers can deploy a single pair of Cisco Nexus 9300 switches. vPCs can be configured on the Cisco Nexus 9300 switches, enabling connected devices to have two active, redundant paths to each switch (Figure 21).

Figure 21. Two-Switch Small Business Deployment

The topology depicts a pair of fixed 1 rack unit (1RU) Cisco Nexus 9372PX Switches. Each Cisco Nexus 9372PX Switch provides 48 ports of 1/10 Gigabit Ethernet with enhanced small form-factor pluggable (SFP+) Transceivers and six Quad (4-channel) Small Form-factor Pluggable (QSFP+) 40 Gb ports. There is also a 10GBASE-T version of the switch available if twisted pair RJ-45 connectors are desired over SFP+ for 10 Gb. This design provides a total of 108 ports in a 2RU footprint, supporting 1 Gb, 10 Gb, and 40 Gb interfaces.

If the 40 Gb interfaces are not required, the ports can be converted to 10 Gb interfaces using Cisco QSFP to SFP or SFP+ Adapter (QSA) Module. The Cisco QSA module converts a 40 Gb QSFP port to a 1 Gb SFP or 10 Gb SFP+ port. Customers have the flexibility to use any SFP+ or SFP module or cable to connect to a lower-speed port. This flexibility allows a cost-effective transition to 40 Gb over time. The Cisco QSA module supports all SFP+ optics and cable reaches and several 1 Gb SFP modules.

Figure 22. Cisco QSA Module Empty (Left); SFP or SFP+ Inserted (Right)

For more information about Cisco QSA modules, see the Cisco QSA data sheet.

For information about savings on 10 Gb interfaces, Cisco SFP+ Copper Twinax direct-attach cables for distances up to 10 meters, click here Twinax cables (Figure 23) provide significant savings over traditional 10 GB fiber optic transceivers and cabling.

Figure 23. Cisco Direct-Attach Twinax Copper Cable Assembly with SFP+ Connectors
For more information on all 10 Gb SFP+ cable options, see the Cisco 10GBase SFP Modules data sheet.

Always check the Cisco Transceiver Modules Compatibility Information webpage for the most up-to-date information about chassis support.

With a two-switch design, small businesses can reap the benefits of the Cisco Nexus 9000 solution while taking advantage of a small, affordable footprint. Over time, as the data center grows and the need arises for more throughput, customers can expand the design by adding more leaves (access switches), or they can choose to include Cisco Nexus 9500 platform switches in the aggregation, or spine layer, which is the next featured topology.

**Design B: Two Aggregation and Two Access Switches**

A simple two-spine (aggregation) and two-leaf (access) design provides a satisfactory entry point for small-to-midsize customers requiring more bandwidth and room for future expansion as shown in Figure 24.

**Figure 24. Two-Aggregation Two-Access Commercial Design for Small-to-Medium-Sized Businesses**

In the leaf or access layer, the topology shown here depicts a pair of fixed 1RU Cisco Nexus 9372PX Switches. Each Cisco Nexus 9372PX Switch provides 48 ports of 1/10 Gigabit Ethernet with SFP+ transceivers and six QSFP+ 40 Gb ports. There is also a 10GBASE-T version of the switch available if twisted pair RJ-45 connectors are desired over SFP+ for 10 Gb.

In the spine or aggregation layer, the topology depicts two Cisco Nexus N9K-C9504 chassis with N9K-X9536PQ line cards. The line cards supply 36 QSFP+ 40 Gb each, used to connect to the leaves. The 9536PQ 36-port line cards open the door for future growth when more leaves are added for access ports, and leave the door open for transition to ACI mode.

For the latest comparison between line cards, check out the Cisco Nexus 9000 Series Switches Compare Models tool.
The Cisco QSA 40-to-10 Gb modules can also be used in this design as you transition from 10 Gb to 40 Gb. Alternately, Cisco provides a low-cost 40 Gb transceiver called a BiDi that eases the move from 10 Gb to 40 Gb. Existing short-reach 40 Gb transceivers use connectors that require 12 strands of fiber through a multifiber push-on (MPO) connector. Unfortunately existing 10 Gb fiber deployments and patch panels use Lucent connector (LC)-to-LC multimode fiber. Upgrading from 10 Gb to 40 Gb fiber can be an expensive endeavor if all transceivers, cabling, and patch panels had to be replaced.

Figure 25. Existing 40 Gb 12-Strand Fiber Cable and Connectors

As an alternative, the Cisco 40G QSFP BiDi transceiver (Figure 25) addresses the challenges of the fiber infrastructure by providing the ability to transmit full-duplex 40 Gb over a standard OM3 or OM4 multimode fiber with LC connectors. The Cisco BiDi transceiver can reuse existing 10 Gb fiber cabling, instead of deploying a new fiber infrastructure. The Cisco BiDi optic transceiver provides an affordable, simple upgrade path to 40 Gb at almost the same cost as 10 Gb fiber today.

Figure 26. Cisco QSFP BiDi Transceiver

Figure 27. Cisco BiDi Fiber Connectivity and Speed

For more information about Cisco BiDi optics, see the Cisco QSFP BiDi Technology white paper.
This design is showcased later in this white paper, and configuration examples are provided.

**Design C: Two Aggregation and Four Access Switches**

As your data center grows and more servers are added, the Cisco Nexus 9000 deployment can be scaled out by adding more leaves (access switches) to increase available server and device access ports as shown in Figure 28.

**Figure 28.** Two-Aggregation Four-Access Commercial Growth Design

In the leaf or access layer, the topology depicts four fixed 1RU Cisco Nexus 9372PX Switches. Each Cisco Nexus 9372PX Switch provides 48 ports of 1/10 Gigabit Ethernet with SFP+ transceivers and six QSFP+ 40 Gb ports. There is also a 10GBASE-T version of the switch available if twisted-pair RJ-45 connectors are desired over SFP+ for 10 Gb.

In the spine or aggregation layer, the topology depicts two Cisco Nexus N9K-C9504 chassis with Cisco N9K-X9536PQ line cards. The line cards supply 36 QSFP+ 40 GB, each used to connect to the leaves. The Cisco 9536PQ 36-port line cards open the door for future growth when more leaves are added for access ports. The line cards also leave the door open for transition to ACI mode.

**Design D: Four Aggregation and Four Access Switches with Spine-Leaf**

A design with at least four spines enters true spine-leaf territory where the function of the spine becomes pure backbone connectivity. In this design, the only devices connected to the spine are the leaves. Failure of a single spine has little impact on the system, leaving one less available path. All devices connect to the leaves.

Adding spines provides more cross-sectional bandwidth in the fabric, as well as additional paths to load balance traffic across. Like the previous design, as more access ports are required, more leaves can be added to the topology. More spines can also be added as bandwidth requirements grow. The spine-leaf topology allows for massive scale-out.
Adding Cisco Nexus 9000 Series Switches to your data center does not mean existing devices must be removed or reconfigured. The Cisco Nexus 9000 solution can be integrated easily with a commercial data center while still providing all the benefits of the platform.

**Access Layer Device Connectivity**

Depending on your existing network, devices, such as servers, storage, and Layer 4-7 service appliances can be connected to either the Cisco Nexus 9500 spine (aggregation) or the Cisco Nexus 9300 leaf (access), or left in place on existing data center switches.

In a true spine-leaf topology, all devices connect directly to leaves, and the leaves are the only devices that connect to the spines. This provides predictable latency and ensures each device is an equal number of hops away from any other device in the fabric.

As new pods of the Cisco Nexus 9300 leaves are added to the network, new servers can be attached to the leaves at the access layer. Other devices, such as the Cisco Nexus 5000 Series Switches, Cisco Unified Computing System™ (Cisco UCS®), or blade-server access-layer devices, can be connected to Cisco Nexus 9000 Series Switches.

Cisco Nexus 9000 Series Switches can be dropped in as a new data center aggregation pod when expanding for application growth or upgrading the network. This models a traditional insertion of a new pod. Existing servers or Layer 4-7 services do not need to be moved or changed. Sample integration with existing networks is provided for each design covered in the previous section.

**Design A Integration**

The two-switch Design A topology can easily be inserted as a new data center access and/or aggregation pod. The Cisco Nexus 9300 switches are shown connected to the existing data center aggregation switch. New servers can then be connected to the Cisco Nexus 9300 access switches as the data center expands.

The Cisco Nexus 9300 platform can be integrated easily with existing infrastructure using standard protocols. The in-place collapsed core (aggregation) switches can still be used to connect to the Layer 4-7 services, external networks, and host the default gateways. While the existing network depicts Cisco Nexus 7000 and 5000 Series Switches, the existing network can be Cisco Catalyst®, or other vendor, infrastructure.
Similar to Design A, the two-aggregation two-access Design B topology can easily be inserted as a new data center access-aggregation pod. The Cisco Nexus 9300 switches are shown connected to the existing data center aggregation switches. New servers can then be connected to the Cisco Nexus 9300 Series access switches as the data center expands.

The Cisco Nexus 9300 platform can be integrated easily with existing infrastructure using standard protocols. The in-place collapsed-core (aggregation) switches can still be used to connect to the Layer 4-7 services, external networks, and host the default gateways. While the existing network depicts Cisco Nexus 7000 and 5000 Series Switches, the existing network can be Cisco Catalyst, or other vendor, infrastructure.

As older switches are phased out, eventually the Layer 4-7 services and default gateways can be moved to the new Nexus 9000 Series Switches.
Additional Cisco Nexus 9300 leaf access switches can be added to support a growing application environment and data center. Adding more Cisco Nexus 9500 spine aggregation switches accommodates growing bandwidth requirements and the ability to phase out aging switches. Services, gateways, and external connections can be migrated to the Cisco Nexus 9000 Series Switches as part of the growth strategy.

Figure 32 depicts the addition of two Cisco Nexus 9300 leaf and/or access switches to meet application and server growth demands.
When the existing gear can be phased out, the existing servers, services, and external connections can be migrated to the new Cisco Nexus 9300 leaf and/or access switches as depicted in Figure 33.

**Figure 33.** Removal of Legacy Gear and Migration of Servers and Services to the Cisco Nexus 9000 Series Switches

As applications and the business grow, more bandwidth may be required. Figure 34v depicts the addition of two more Cisco Nexus 9500 spine and/or aggregation switches.
For more details about specific Cisco Catalyst migration details, see the Migrate from Cisco Catalyst 6500 Series Switches to Cisco Nexus 9000 Series Switches white paper.

**Fabric Extender Support**

To use existing hardware, increase access port density and increase 1 Gigabit Ethernet port availability, Cisco Nexus 2000 Series Fabric Extenders can be attached to the Cisco Nexus 9300 platform in a single-homed straight-through configuration. Host uplinks that are connected to the fabric extenders can either be Active/Standby, or Active/Active, if configured in a vPC.

The following Cisco Nexus 2000 Series Fabric Extenders are supported at the time of this writing:

- N2224TP
- N2248TP
- N2248TP-E
- N2232TM
- N2232PP
- B22HP

For the most up-to-date feature support, refer to the Nexus 9000 Software release notes.
Fabric extender transceivers (FETs) also are supported to provide a cost-effective connectivity solution (FET-10 Gb) between Cisco Nexus 2000 Fabric Extenders and their parent Cisco Nexus 9300 switches.

For more information about FET-10 Gb transceivers, see the Nexus 2000 Series Fabric Extenders data sheet. The supported Cisco Nexus 9000 to Nexus 2000 Fabric Extender topologies are shown in Figures. As with other Nexus platforms, think of the Cisco Fabric Extender Technology (FEX Technology) like a logical remote line card of the parent Cisco Nexus 9000 switch. Each Cisco FEX Technology <noun> connects to one parent switch. Servers should be dual-homed to two different fabric extenders. The server uplinks can be in an Active/Standby NIC team, or they can be in a vPC if the parent Cisco Nexus 9000 switches are set up in a vPC domain.

**Figure 36.** Supported Cisco Nexus 9000 Series Switches Plus Fabric Extender Design with Active/Standby Server
Figure 37. Supported Cisco Nexus 9000 Series Switches Plus Fabric Extender Design with server vPC


Storage Design
Existing IP-based storage, such as network-attached storage (NAS) or Internet Small Computer Interface (iSCSI) storage, can be integrated with a Cisco Nexus 9000 fabric. Currently, Fibre Channel and Fibre Channel over Ethernet (FCoE) are not supported on the Cisco Nexus 9000 platform; however, this section show how they can be designed alongside the Cisco Nexus 9000 platform and evolve with added features. Refer to the Cisco website for updates on future support for FCoE N port virtualization (NPV). Customers who prefer to have separate, dedicated storage connections from each server can cable the storage connections to their physical storage network, and cable the production IP LAN connections to the Nexus 9000s.

The dedicated, physical storage network can use IP-based storage, such as iSCSI or NAS, or can consist of a Fibre Channel or FCoE network. Regardless of the protocol, this design does not require any change in existing storage cabling. Figure 38 provides an example of a separate, dedicated storage network for SAN traffic, with LAN traffic passing through the Nexus 9000 fabric.
The new Cisco MDS 9148S Multilayer Fabric Switch is featured in Figure 38. The Cisco MDS 9148 fabric switch is a highly affordable, versatile, easy-to-manage multiprotocol storage networking switch for entry-level and departmental SANs. The Cisco MDS 9148S fabric switch, based on the switch-on-a-chip Cisco storage networking ASIC, delivers 16 Gbps dedicated Fibre Channel bandwidth per port. The Cisco MDS 9148S fabric switch is available in a base configuration of 12 ports with on-demand pay-as-you-grow 12-port activation license to scale up to 48-ports. A fully licensed 48-port configuration is also available. The Cisco 9148S fabric switch enables scalable and flexible SAN architectures for a converged fabric, including support for FCoE, FCoE-FEX, and FCoE over Cisco FabricPath.

For more information about Cisco MDS 9148S Multilayer Fabric Switch, see the Cisco Storage Networking webpage.

A converged storage fabric design is possible through the Cisco Nexus 9000 Series Switches if using IP-based storage, such as iSCSI or NAS to reduce cabling, switchports, number of switches, and number of adapters required, as well as gaining significant power savings. All servers connect to the Cisco Nexus 9300 platform leaves (access switches), carrying both LAN and IP-based storage traffic. The storage devices can also connect to the leaves or can be left connected to existing infrastructure. Figures 39 and 40 highlight both options.
The Cisco Nexus 9372 leaf access switches split out the LAN traffic to send to the spine aggregation switches and send the IP-based storage traffic to switches dedicated to storage traffic. While many switches can be used, the Cisco Nexus 5672UP switches are featured in Figure 40.

The Cisco Nexus 5600 Series Switches are the third generation of the leading data center server access Cisco Nexus 5000 Series Switches. The Cisco Nexus 5600 platform is the successor of the industry's most widely adopted Cisco Nexus 5500 switches that maintain all the existing Cisco Nexus 5500 features including LAN/SAN convergence (unified ports, FCoE), fabric extenders and fabric path but in addition brings integrated line-rate Layer 2 and 3 with true 40 Gigabit Ethernet support, Cisco Dynamic Fabric Automation innovation, network virtualization using Generic Routing Encapsulation (NVGRE), VXLAN bridging and routing capability, network programmability and visibility, deep buffers, and significantly higher scale and performance for highly virtualized, automated, and cloud environments.

Figure 40.  Cisco Nexus 5672UP Switch

For more information, see the Cisco Nexus 5600 Platform Switches data sheet.

Figure 41 illustrates a design where both servers and IP-based storage are directly connected to the leaf access switches, reducing the number of hops and devices.
End-State Topology

Figure 42 is a simplified diagram of an example of an end-state design integrated with an existing data center. The design features two Cisco Nexus 9372 Switches, two Cisco Nexus 9504 Switches, IP-based storage, a Cisco ASA firewall appliance, connections to the campus Cisco Catalyst LAN environment, and connectivity to the WAN router. Notice that all devices are connected to the Cisco Nexus 9300 leaf access switches.
Example: Cisco Nexus 9000 Series Design and Configuration

This section walks through an example of the Cisco Nexus 9000 Series design using real equipment, including Cisco Nexus 9508 switches, Nexus 9396 switches, both VMware ESXi servers and standalone bare metal servers, and a Cisco ASA firewall. This topology has been tested in a Cisco lab to validate and demo the Cisco Nexus 9000 solution.

Your exact configuration may vary. The intent of this section is to give a sample network design and configuration including the features discussed earlier in this white paper. Always refer to the Cisco configuration guides for the latest information.

The terminology and switch names used reference “leaf” and “spine,” but the same configuration can be applied to an access-aggregation design.

This section includes explanations and end-state configurations for the following features:

- Basic network setup:
  - VLANs and trunking
  - IP addressing
- ECMP routing infrastructure:
  - EIGRP
- Server NIC teaming:
  - Virtual PortChannel (vPC)
- Overlay:
  - Virtual Extensible LAN (VXLAN)
  - Multicast
- Services:
  - Cisco ASA firewall
- Management:
  - Cisco Data Center Network Manager (DCNM)
- Cloud readiness

Hardware and Software Specifications

- Two Cisco Nexus N9K-C9508 Switches (8-slot) with the following components in each switch:
  - One N9K-X9636PQ (36p QSFP 40G) Ethernet module
  - Six N9K-C9508-FM fabric modules
  - Two N9K-SUP-A supervisor modules
  - Two N9K-SC-A system controllers
- Two Cisco Nexus 9396PX Series Switches (48p SFP+ 1/10G) with the following expansion module in each switch:
  - One N9K-M12PQ (12p QSFP 40G) Ethernet module
- Three Cisco UCS C-Series servers (can be replaced with Cisco UCS Express servers for smaller deployments). For more information about server options, see Cisco Unified Computing System Express.
- One Cisco ASA 5510 firewall appliance
The Cisco Nexus 9000 switches used in this design run Cisco NX-OS version 6.1(2)I2(3). The VMware ESXi servers used run ESXi 5.1.0 build 799733. The Cisco ASA runs version 8.4(7).

Basic Network Setup

In this simple topology, VLAN 50 is the DMZ, VLAN 70 is used for vMotion, VLAN 80 is the internal production web server VLAN, and VLAN 90 is for external traffic. The physical interfaces eth2/1 and eth2/2 connect to the spines. Interface eth1/1 connects to a VMware ESXi server, and interface eth1/48 connects to Cisco ASA 5510 firewall appliance (configuration shown later). The configuration is highlighted in the following logical topology diagram as it is built upon (Figure 43).

Figure 43. VLANs Used in Sample Topology

![VLANs Used in Sample Topology](image)

Figure 44. Lab Topology Diagram and VLAN Usage

![Lab Topology Diagram and VLAN Usage](image)

Sample interface configuration from a leaf switch in the lab-tested topology is provided. The other leaf and spines are configured similarly. Explanatory comments are provided inline, italicized, and prefaced by an exclamation point (!). Cisco NX-OS commands are shown in boldface.
Figure 45. Basic VLAN and Interface Configuration

! Create and (optionally) name VLANs

Leaf1# configure terminal
Leaf1(config)# vlan 50,70,80,90
Leaf1(config-vlan)# vlan 50
Leaf1(config-vlan)# name DMZ
Leaf1(config-vlan)# vlan 70
Leaf1(config-vlan)# name vmotion
Leaf1(config-vlan)# vlan 80
Leaf1(config-vlan)# name internal
Leaf1(config-vlan)# vlan 90
Leaf1(config-vlan)# name external

! Configure Layer 3 spine-facing interfaces

Leaf1(config-vlan)# interface Ethernet2/1
Leaf1(config-if)# description to Spine1
Leaf1(config-if)# no switchport
Leaf1(config-if)# ip address 10.1.1.2/30
Leaf1(config-if)# no shutdown

Leaf1(config-if)# interface Ethernet2/2
Leaf1(config-if)# description to Spine2
Leaf1(config-if)# no switchport
Leaf1(config-if)# ip address 10.1.1.10/30
Leaf1(config-if)# no shutdown

! Configure Layer 2 server and service-facing interfaces
! (Optionally) Limit VLANs and use STP best practices

Leaf1(config-if)# interface Ethernet1/1-2
Leaf1(config-if)# switchport
Leaf1(config-if)# switchport mode trunk
Leaf1(config-if)# switchport trunk allowed vlan 50,70,80,90
Leaf1(config-if)# spanning-tree bpduguard enable
Leaf1(config-if)# spanning-tree port type edge trunk

Warning: Edge port type (portfast) should only be enabled on ports connected to a single host. Connecting hubs, concentrators, switches, bridges, etc... to this interface when edge port type (portfast) is enabled, can cause temporary bridging loops. Use with CAUTION.

Leaf1(config-if)# no shutdown
Leaf1(config-if)# interface Ethernet1/1
Leaf1(config-if)# description to Server1
Leaf1(config-if)# interface Ethernet1/2
Leaf1(config-if)# description to Server2

Leaf1(config-if)# interface Ethernet1/48
Leaf1(config-if)# description to ASA
Leaf1(config-if)# switchport
Leaf1(config-if)# switchport mode trunk
Leaf1(config-if)# switchport trunk allowed vlan 50,70,80,90
Leaf1(config-if)# no shutdown

For more information about basic configuration, see the Nexus 9000 Series NX-OS Interfaces Configuration Guide, Release 6.0.

For the latest Cisco NX-OS release guides, see the Cisco Nexus 9000 Series Switches Configuration Guides webpage.

ECMP Routing Design

A Layer 3 dynamic routing protocol used in the control plane enables optimal routing of traffic and equal cost multipathing, and alleviates the need for spanning tree in the fabric between switches. Spanning tree still needs to run on other nonfabric interfaces to block loops in a classical Ethernet environment. All paths between each switch actively forward traffic.

The Cisco Enhanced Interior Gateway Routing Protocol (EIGRP) is used in the sample topology. Open Shortest Path First (OSPF) BGP, Intermediate System to Intermediate System (IS-IS), and static routing are also supported with the Enhanced Layer 3 license on both the Cisco Nexus 9300 and 9500 platforms. EIGRP provides fast convergence and excellent scalability. Because every leaf and/or access switch is connected to every spine and/or aggregation switch, EIGRP load balances across all equal cost paths.

Sample EIGRP configuration from a leaf switch in the lab-tested topology is shown here in Figure 46. The other leaf and spines are configured similarly.

Figure 46. EIGRP Configuration

! Enable the EIGRP feature
! Configure global EIGRP process and (optional) router ID

Leaf1(config)# feature eigrp
Leaf1(config)# router eigrp 1
Leaf1(config-router)# router-id 1.1.1.33

! Enable the EIGRP process on spine-facing interfaces

Leaf1(config-router)# interface Ethernet2/1-2
Leaf1(config-if)# ip router eigrp 1

For more information about configuring EIGRP, see the Cisco Nexus 9000 Series NX-OS Unicast Routing Configuration Guide, Release 6.0.
Server NIC Teaming Design and Configuration

Modern applications and increasing virtual machine density due to advances in CPU and memory footprints drive server bandwidth demand, with many servers requiring 10 Gb connections. Ideally, every server is dual-homed to two different physical switches. Ordinarily, one of these connections actively forwards traffic, while the other connection stands by. While this design provides redundancy in case of a switch failure, a standby or blocked connection wastes potential bandwidth.

Cisco vPC allows connections from a device in a Cisco PortChannel to terminate on a pair of two different Cisco Nexus switches set up in a special peering relationship. Cisco vPC provides Layer 2 multipathing, increasing bandwidth while maintaining redundancy. Any device that supports Cisco PortChannels can be set up in a Cisco vPC connected to a pair of switches in a Cisco vPC domain and is unaware it is configured as a special type of PortChannel.

**Figure 47.** Traditional Ethernet Forwarding Paths Compared to Cisco vPC Active/Active Forwarding Paths

Cisco vPC provides the following benefits:

- Allows a single device to use a Cisco PortChannel connected to two upstream Cisco Nexus switches.
- Eliminates spanning tree blocked ports.
- Provides a loop-free topology.
- Uses all available uplink bandwidth.
- Provides fast convergence if either a link or a switch fails.
- Provides link-level resiliency through standard Cisco PortChannel mechanisms.
- Helps ensure high availability by connecting to two different physical switches.

A Cisco vPC domain includes two Cisco Nexus peer switches (Cisco Nexus 9000 Series Switches in the sample design), the Cisco vPC peer keepalive link used as a heartbeat between the peers, the vPC peer link (for exchange of control traffic, orphan traffic, and multidestination traffic), and all the PortChannels in the Cisco vPC domain connected to the downstream device. You can have only one vPC domain ID on each switch and only two switches per Cisco vPC domain.

In the sample topology, the two leaf (access) switches are set up in a Cisco vPC domain. The Cisco ASA firewall appliance and most servers are connected to the leaves (access switches) through a Cisco vPC. Some legacy servers are single-homed to one switch and are not a part of a Cisco vPC. As a best practice, all devices are connected to both switches through Cisco vPC, but this is not a requirement.
Cisco vPC peer switches in the same domain need to use the same domain number. Additionally, the Cisco vPC number for a given device must match on both sides. The Cisco PortChannel number is locally significant and does not have to match on both sides.

Sample Cisco vPC configuration from a leaf switch in the lab-tested topology is provided. The other leaf switch is configured similarly. Cisco vPCs are Layer 2 PortChannels, and therefore, there is no need to configure a Cisco vPC domain on the spine switches as they connect through Layer 3 interfaces to the leaves.

Following Cisco vPC design best practices is critical. Cisco strongly recommends reading the vPC Best Practices Design Guide, but the crucial best practices that should be implemented include:

- **Domain, peer link, peer keepalive, and member interface best practices**
  - Different Cisco vPC domains must use different Cisco vPC domain ID numbers.
  - All Cisco vPC type 1 consistency checks on both peers in a domain must match.
  - All Cisco vPC type 2 consistency checks on both peers in a domain should match.
  - Cisco vPC member interfaces should be configured identically on both peer switches.
  - Use Link Aggregation Control Protocol (LACP) on all Cisco PortChannels related to Cisco vPC.
  - Enable Unidirectional Link Detection (UDLD) in normal mode on Cisco vPC peer link and member ports.
  - Always enable PIM prebuild shortest-path tree (SPT) when using multicast with Cisco vPC.
  - Ideally, the peer-keepalive heartbeat link should be a pair of directly connected Layer 3 interfaces in a dedicated (VRF).
  - Use Cisco vPC to interconnect a maximum of two data centers. Use Overlay transport virtualization (OTV) when more than two data centers need to be interconnected.
  - Build the peer link on two different line cards to increase redundancy. Do not insert any devices between the peer switches for the peer link connections.
  - Split Cisco vPC VLANs and VLANS that are not Cisco vPC onto different Cisco PortChannels.
  - Use the same Cisco vPC ID and PortChannel number for ease of configuration, monitoring, and troubleshooting.
  - When possible, dual-attach every device connected to the Cisco vPC domain.
  - Always enable the vPC peer-gateway in the Cisco vPC domain (that is, configure peer-gateway on both Cisco vPC peer devices), even if there is no end device (devices that don’t perform standard ARP request for their default IP gateway) using this feature. There are no side effects enabling it.
  - Always enable Cisco vPC ARP Sync on both vPC peer devices.
  - Use Cisco vPC orphan port suspend when single-attached devices connected to Cisco vPC domain need to be disconnected from a network when the vPC peer link fails.

- **HSRP-VRRP best practices:**
  - When running Hot Standby Router Protocol (HSRP) and Virtual Router Redundancy Protocol (VRRP) in active-active mode (data plane standpoint), aggressive timers can be relaxed: use the default HSRP-VRRP timers.
  - Define the switch virtual interface (SVI) associated with First Hop Redundancy Protocol (FHRP) and VRRP as a passive routing interface to avoid forming routing adjacency over Cisco vPC peer link.
  - Define Cisco vPC primary peer device as the active HSRP-VRRP instance and Cisco vPC secondary peer device as standby HSRP-VRRP (from control plane standpoint) for ease of operations.
● Disable IP redirect on the interface VLAN where HSRP-VRRP is configured (command is no IP redirect). This is a general best practice related to HSRP-VRRP.

● Do not use HSRP-VRRP object tracking in a Cisco vPC domain.

- **Spanning tree best practices:**
  
  ◦ Do not disable spanning tree.
  
  ◦ For large Layer 2 domains, use MST.
  
  ◦ Use the same version of spanning tree on all devices in the same Layer 2 domain.
  
  ◦ Configure VLAN pruning on Cisco vPC member ports.
  
  ◦ Keep STP root function on the aggregation layer of the network (aggregation Cisco vPC domain).
  
  ◦ For each Cisco vPC peer device, configure root guard on ports connected to access devices.
  
  ◦ Bridge Assurance is enabled by default when configuring the Cisco vPC peer link. Do not disable it on the Cisco vPC peer link.
  
  ◦ Configure Cisco vPC member ports as SPT port type normal (not using Bridge Assurance on the link).
  
  ◦ Configure Cisco Port Fast (edge port type) products on the host-facing interfaces to avoid slow STP convergence (30 seconds or more) when port transitions to up state.
  
  ◦ Configure Bridge Protocol Data Unit (BPDU) guard on host-facing interfaces to block any BPDU sent from the host (access switch port receiving the BPDU is put in errdisable mode).
  
  ◦ Always define the Cisco vPC domain as STP root for all VLAN in that domain (configure aggregation vPC peer devices as STP root primary and STP root secondary). Enforce this rule by implementing STP root guard on Cisco vPC peer devices ports connected to another Layer 2 switch.

To upgrade the access layer without a disruption to hosts that are dual-homed in a vPC, complete the following tasks:

- Upgrade the vPC Primary peer switch first. During the upgrade the switch will be reloaded. When the switch is reloaded, the downstream devices in vPCs detect loss of connectivity to the vPC Primary peer switch and will start forwarding traffic on the links to the other peer switch.

- Verify that the upgrade of the vPC Primary peer switch has completed successfully. At the completion of the upgrade, the switch will restore vPC peering and all the vPC links.

- Repeat the same process to upgrade the other peer switch after you have verified the vPC domain has fully reconverged. The switch will reload during the upgrade process. During the reload, the first (upgraded) switch will forward all the traffic to/from the downstream devices.

- Verify that the upgrade of the second switch has completed successfully. At the end of this upgrade, complete vPC peering is established and the upgrade is complete.
Figure 48 illustrates the Cisco vPC configuration. Figure 49 walks through the Cisco vPC configuration on a leaf switch.

**Figure 48. Sample Cisco vPC Configuration**

![Sample Cisco vPC Configuration](image)

**Figure 49. Cisco vPC Domain and Device Configuration**

```
! Configure ULD in normal mode per best practices
Leaf1(config)# feature uld

! Prepare peer-keepalive heartbeat interface and VRF
Leaf1(config)# vrf context KEPPALIVE
Leaf1(config)# interface Ethernet1/20
Leaf1(config)# vrf member KEPPALIVE
Warning: Deleted all 13 config on interface Ethernet1/20
Leaf1(config)# ip address 192.168.99.1
Leaf1(config)# no shutdown

! Enable feature and create a vPC domain
Leaf1(config)# feature vpc
Leaf1(config)# vpc domain 1

! Set peer-keepalive heartbeat source and destination
Leaf1(config-vpc-domain)# peer-keepalive destination 192.168.99.2 source 192.168.99.1 vrf KEPPALIVE

! Automatically and simultaneously enable the following
! best practice commands using the mode auto command; peer-
! gateway, auto-recovery, ip arp synchronize, and ipv6 nd
! synchronize.

Leaf1(config-vpc-domain)# mode auto

! Create the peer-link Port Channel and add to domain
Leaf1(config-vpc-domain)# feature lacp
Leaf1(config)# interface Ethernet1/31-32
Leaf1(config-if-range)# channel-group 1 mode active
Leaf1(config-if-range)# interface port-channel 1
Leaf1(config)# switchport
Leaf1(config)# switchport mode trunk
Leaf1(config)# switchport trunk allowed vlan 50,70,80,90
Leaf1(config)# vpc peer-link
```

Please note that spanning tree port type is changed to "network" port type on vPC peer-link. This will enable spanning tree Bridge Assurance on vPC peer-link provided the STP Bridge Assurance (which is enabled by default) is not disabled.
Leaf1(config) # no shutdown

! Configure best practice Spanning Tree features on vPC
! Create a Port Channel and vPC for a server
! The vPC #11 must match on the other peer for the server

Leaf1(config) # interface Ethernet1/1
Leaf1(config) # channel-group 11 mode active
Leaf1(config) # interface port-channel 11
Leaf1(config) # spanning-tree port type edge
Leaf1(config) # spanning-tree bpdu guard enable
Leaf1(config) # vpc 11

! Repeat vPC configuration for any additional devices
! connected to the vPC domain

For more information, see the Configuring vPCs chapter of the Cisco Nexus 9000 Series NX-OS Interfaces Configuration Guide, Release 6.0.

Overlay Design and Configuration

A virtual overlay deployed in the sample commercial design provides Layer 2 reachability over the Layer 3 EIGRP network. This section discusses VXLAN basics, how to set up multicast routing to support VXLAN, and how to configure VXLAN on Cisco Nexus 9000 Series Switches. VXLAN is not required on Cisco Nexus 9000 switches, but it provides an optional overlay across a routed infrastructure. Virtual PortChannels and a routed aggregation is an equally valid design.

Why VXLAN?

VXLAN is featured as the choice overlay. VXLAN is a Layer 2 overlay running over a Layer 3 network and addresses several pain points that exist in traditional networks. Many common data center applications require Layer 2 adjacency to communicate. For example, virtual machine workload mobility and some clustering applications must communicate over a dedicated VLAN or subnet, and cannot be routed. However, building large Layer 2 segments creates large broadcast domains and increases the scope of undesirable events, such as an SPT reconvergence. Spanning Tree Protocol blocks links for loop prevention, which wastes bandwidth and thereby increases oversubscription. Additionally, VLANs can scale to only a maximum of 4094 segments with many VLANs reserved by the switch for internal use.

VXLAN is designed to provide the same Layer 2 reachability, segmentation, and services as traditional VLANs, but it addresses the many limitations. Compared to VLAN, VXLAN offers the following benefits:

- Flexible placement of multitenant segments throughout the data center: It provides a solution to extend Layer 2 segments over the underlying shared network infrastructure so that tenant workload can be placed, and moved, across physical pods in a data center.
- Higher scalability to address more Layer 2 segments: VLANs use a 12-bit VLAN ID to address Layer 2 segments, which results in limiting scalability of only 4094 VLANs. VXLAN uses a 24-bit segment ID known as the VXLAN network identifier (VNID), which enables up to 16 million VXLAN segments to coexist in the same administrative domain.
Better utilization of available network paths in the underlying infrastructure: VLAN uses SPT for loop prevention, which results in not using half of the network links in a network by blocking redundant paths. In contrast, VXLAN packets are transferred through the underlying network based on its Layer 3 header and can take complete advantage of Layer 3 routing, ECMP routing, and link aggregation protocols to use all available paths.

VXLAN can be implemented both on hypervisor-based virtual switches to allow scalable, repeatable virtual machine deployments and on physical switches acting as gateways to bridge VXLAN segments back to VLAN segments.

For more information, see the VXLAN Overview: Cisco Nexus 9000 Series Switches.

**How Does VXLAN Work?**

To extend Layer 2 networks across a Layer 3 infrastructure, VXLAN uses MAC-in-UDP (User Datagram Protocol) encapsulation and stateless tunneling over IP transport. Devices that host the tunnels, and perform encapsulation and decapsulation of packets are called VXLAN tunnel endpoints (VTEPs). When acting as a hardware VXLAN gateway, the Cisco Nexus 9000 VTEPs combine a VXLAN segment and a classic VLAN segment into one common Layer 2 domain.

Cisco Nexus 9000 VTEPs have two types of interfaces involved in VXLAN: classical Layer 2 Ethernet interfaces connected to servers, appliances, or other switches and a Layer 3 IP interface, which is called a Network Virtualization Edge (NVE) and is bound to a loopback interface. The NVE interface has several functions:

- Uniquely identifies the VTEP (switch) on the transport network.
- Encapsulates and decapsulates VXLAN frames.
- Discovers remote VTEPs (switches) for its configured VXLAN segments.
- Learns remote MAC address-to-VTEP mappings to be used for forwarding lookups.

The source and destination IPs of the VTEPs are used to forward VXLAN-encapsulated traffic through the underlying transport network, routing traffic from ingress source leaf switch to egress destination leaf switch. The VXLAN segments are independent of the underlying network topology, and the underlying network topology between VTEPs is independent of the VXLAN overlay.

On encapsulation, a VTEP takes the original Layer 2 frame and adds an 8-byte VXLAN header. Additionally UDP, IP, and MAC headers are added. The primary field in the VXLAN header is a 24-bit VXLAN network identifier (VNID). Each VNID represents a unique VXLAN segment ID.

**Multicast Considerations**

VXLAN uses multicast to reduce flooding scope for hosts in the same VXLAN segment. Multicast groups are used to transport broadcast, unknown unicast, and multicast traffic, discover remote VTEPs (switches), and learn remote host MAC addresses and MAC-to-VTEP mappings for each VXLAN segment. Each VLAN-to-VXLAN segment is mapped to a multicast group as part of the configuration.

Therefore, multicast routing must be configured in the transport network to support VXLAN. Enabling multicast support requires PIM and rendezvous point configuration. Each VXLAN segment is mapped to a multicast group in the transport network, which each VTEP joins through IGMP.
In the sample topology, PIM sparse mode using any-source multicast (ASM) is used for multicast routing to advertise group membership and construct multicast distribution trees. The Multicast Source Discovery Protocol (MSDP) is used to create an anycast-rendezvous point configuration to provide rendezvous-point redundancy and load sharing. Use of MSDP is not required for VXLAN, but the presence of a rendezvous point is required.

**Spine (Aggregation) Multicast Configuration**

A sample multicast configuration from a spine switch in the lab-tested topology is shown here in Figure 50. The other spine switch is configured similarly. The leaves will be configured differently and shown later.

**Figure 50.** Spine PIM and MSDP Multicast Configuration

! Enable PIM and MSDP to turn on multicast routing

```
Spine(config)# feature pim
Spine(config)# feature msdp
```

! Enable PIM on the spine-facing interfaces
! EIGRP should already be enabled on the interfaces

```
Spine(config)# interface Ethernet1/1-2
Spine(config-if-range)# ip pim sparse-mode
```

! Configure the Rendezvous Point IP address for the ASM
! group range 239.0.0.0/8 to be used by VXLAN segments

```
Spine(config-if-range)# ip pim rp-address 10.2.2.12 group-list 239.0.0.0/8
```

! Configure loopbacks to be used as redundant RPs with
! the other spine switch. L0 is assigned a shared RP IP
! used on both spines. L1 has a unique IP address.
! Enable PIM and EIGRP routing on both loopback interfaces.

```
Spine(config)# interface loopback0
Spine(config-if)# ip address 10.2.2.12/32
Spine(config-if)# ip router eigrp 1
Spine(config-if)# ip pim sparse-mode
```

```
Spine(config-if)# interface loopback1
Spine(config-if)# ip address 10.2.2.1/32
Spine(config-if)# ip router eigrp 1
Spine(config-if)# ip pim sparse-mode
```
! Configure MSDP sourced from loopback 1 and the peer spine's loopback IP address to enable RP redundancy

Spine1(config-if)# ip msdp originator-id loopback1
Spine1(config)# ip msdp peer 10.2.2.2 connect-source loopback1

For more information about PIM, see the chapter “Configuring PIM” in the Cisco Nexus 9000 Series NX-OS Multicast Routing Configuration Guide, Release 6.0.

For more information about MSDP, see the chapter “Configuring MSDP” in the Cisco Nexus 9000 Series NX-OS Multicast Routing Configuration Guide, Release 6.0.

**Leaf (Access) Multicast Configuration**

The leaf switches use the multicast groups and rendezvous point configured on the spine aggregation. The leaf switches must have multicast configured on the spine-facing interfaces, and point to the rendezvous point IP address.

Figure 51 illustrates the multicast setup on the spine and leaf. Sample multicast configuration from a leaf switch in the lab-tested topology is shown here. The other leaf switch (Figure 52) is configured similarly.

**Figure 51.** Spine and Leaf Multicast Illustration
Using VXLAN in Commercial Topology

In a commercial design, each VXLAN segment can be mapped to an application or to a related tier of applications, such as a web tier, an application tier, and a database tier. VXLAN is an optional feature, and one of the many differentiators of the Cisco Nexus 9000 Series Switches. VXLAN is not required on a Cisco Nexus 9000 switch; it is shown here to highlight its possible usage and configuration in a commercial design.

In this design VXLAN is configured for VLANs 70, 80, and 90 to create tunnels for leaf-to-leaf communication through what appears to be Layer 2 switching, although the underlying infrastructure is routed. For example, by creating a VXLAN segment for the vMotion VLAN 70, a virtual machine can be migrated nondisruptively between any servers connected to the VXLAN-enabled leaves. This benefit becomes increasingly pertinent as your infrastructure grows and leaves are added, allowing a virtual machine to migrate anywhere across a data center without worrying about SPT scope.

VXLAN is configured on the leaf switches. It is not necessary to configure VXLAN on the spine switches in this topology. The leaves serve as both source and destination endpoints of the VXLAN tunnels and act as gateways to translate VLANs to VXLAN segments on the tunnel ingress and back to VLANs on the tunnel egress. The servers and devices never see the VXLAN tags. The spines do have to support multicast routing for the group ranges configured on the leaves.

Configuring VXLAN requires several steps:

- Enable the VXLAN features.
- Create and assign a /32 IP to a dedicated loopback interface.
- Advertise the loopback /32 address through the transport routing protocol.
- Create the VXLAN NVE interface and attach it to the loopback interface.
- Map VXLAN segments to multicast groups.
- Map VLANs to VXLAN segments.

When configuring VXLAN on switches that also belong to a Cisco vPC domain, additional configuration is required. The loopback tied to the NVE interface must be assigned an identical secondary IP address on both Cisco vPC peer switches. The primary IP address should be different on both Cisco vPC peer switches.
Figure 53 illustrates the VXLAN configuration. Sample VXLAN configuration from a leaf switch in the lab-tested topology is provided in Figure 53. The other leaf switch is configured similarly.

**Figure 53. VXLAN Setup**

```
L1=192.168.1.1
Secondary IP =192.168.2.1
L1=192.168.1.2

Leaf1  Cisco vPC Domain 1  Leaf2
```

**Figure 54. VXLAN Configuration**

```bash
! Enable VXLAN features

Leaf1(config)# feature nv overlay
Leaf1(config)# feature vn-segment-vlan-based

! Configure loopback used for VXLAN tunnels
! Configure secondary IP address for vPC support

Leaf1(config)# interface loopback1
Leaf1(config-if)# ip address 192.168.1.1/32
Leaf1(config-if)# ip address 192.168.2.1/32 secondary
Leaf1(config-if)# ip router eigrp 1
Leaf1(config-if)# ip pim sparse-mode

! Create NVE interface, using loopback as the source
! Bind VXLAN segments to an ASM multicast group

Leaf1(config-if)# interface nve1
Leaf1(config-if)# source-interface loopback1
Leaf1(config-if)# member vni 5000 mcast-group 239.1.1.50
Leaf1(config-if)# member vni 7000 mcast-group 239.1.1.70
Leaf1(config-if)# member vni 8000 mcast-group 239.1.1.80
Leaf1(config-if)# member vni 9000 mcast-group 239.1.1.90
Leaf1(config-if)# no shutdown
```
For more information, see the chapter “Configuring VXLAN” in the Cisco Nexus 9000 Series NX-OS VXLAN Configuration Guide, Release 6.0.

Service Design and Configuration

As discussed earlier, Layer 4-7 service appliances can be connected to either the Cisco Nexus 9300 or 9500 switches, depending on existing infrastructure and traffic patterns. In the sample design, a Cisco ASA 5510 firewall appliance is dual-homed to a pair of Cisco Nexus 9300 switches.

Any number of different vendors or service appliances can be connected to a Nexus 9000 commercial design. In this topology, there are several virtual machines in the DMZ, and several servers on the inside and the outside of the data center network. The DMZ is VLAN 50, internal servers are in VLAN 80, and external servers and users are in VLAN 90. The default gateways for the web servers live on the Cisco ASA 5510 firewall appliance. The firewall interfaces are divided into subinterfaces for each type of traffic, with the correct VLAN associated to the subinterface. Additionally, there is an access list to dictate which zones and devices can talk to each other.

The Cisco ASA configuration example from a lab-tested device is shown in Figure 55. Default and/or irrelevant ASA configuration is not shown. ASA version 8.4(7) is used in the sample configuration.

Figure 55. Example: Cisco ASA Configuration

ciscoasa(config)# interface Ethernet0/0
ciscoasa(config)# no nameif
ciscoasa(config)# no security-level
ciscoasa(config)# no ip address

! Create DMZ subinterface, using VLAN 50
! Assign security level 50 to DMZ
! Assign an IP address to act as a gateway for VLAN 50

ciscoasa(config)# interface Ethernet0/0.50
ciscoasa(config)# description DMZ
ciscoasa(config)# vlan 50
ciscoasa(config)# nameif DMZ
ciscoasa(config)# security-level 50

ciscoasa(config)# ip address 10.1.50.254 255.255.255.0
! Create internal subinterface, using VLAN 80
! Assign security level 100 to internal zone
! Assign an IP address to act as a gateway for VLAN 80

ciscoasa(config)# interface Ethernet0/0.80
ciscoasa(config)# description internal
ciscoasa(config)# vlan 80
ciscoasa(config)# nameif inside
ciscoasa(config)# security-level 100
ciscoasa(config)# ip address 10.1.80.254 255.255.255.0

! Create DMZ subinterface, using VLAN 90
! Assign security level zero to outside zone

ciscoasa(config)# interface Ethernet0/0.90
ciscoasa(config)# description outside
ciscoasa(config)# vlan 90
ciscoasa(config)# nameif outside
ciscoasa(config)# security-level 0
ciscoasa(config)# no ip address

! Create object to include DMZ subnet 10.1.50.0/24

ciscoasa(config)# object network DMZ
ciscoasa(config)# subnet 10.1.50.0 255.255.255.0

! Create object group to match web and HTTP traffic

ciscoasa(config)# object-group service WEB
ciscoasa(config)# service-object tcp destination eq www
ciscoasa(config)# service-object tcp destination eq https

! Create access list to permit DMZ 10.1.50.0/24 web and
! HTTP traffic

ciscoasa(config)# access-list 101 extended permit object-group WEB any object DMZ

! Apply access list to ingress traffic on the outside
! interface

ciscoasa(config)# access-group 101 in interface outside

For more information about Cisco ASA solutions, see the Cisco ASA 5500-X Series Next-Generation Firewalls webpage.
Managing and Automating the Fabric
Cisco UCS Director

Cisco UCS Director is an extremely powerful centralized management and orchestration tool that can make the day-to-day operations of a small commercial IT staff palatable. By using the automation Cisco UCS Director offers, a small IT staff can speed up the delivery of new services and applications. Cisco UCS Director abstracts hardware and software into programmable tasks and uses a workflow designer to allow the system administrator to drag and drop tasks to a workflow to deliver the necessary resources. A sample Cisco UCS Director workflow is depicted in Figure 56.

**Figure 56.** Example: Cisco UCS Director Workflow

Cisco UCS Director gives you the power to automate many common tasks you normally perform manually from the Cisco UCS Manager GUI. Many of the tasks that can be automated with Cisco UCS Director are listed in the following table.
Table 3. Common UCS Director Automation Tasks

<table>
<thead>
<tr>
<th>Task</th>
<th>Task</th>
</tr>
</thead>
<tbody>
<tr>
<td>2. Create UCS Server Pool</td>
<td>22. Create VLAN Group</td>
</tr>
<tr>
<td>3. Delete UCS Server Pool</td>
<td>23. Delete UCS VLAN Group</td>
</tr>
<tr>
<td>4. Add Servers to UCS Server Pool</td>
<td>24. Modify UCS VLAN/VLAN Group Org Permissions</td>
</tr>
<tr>
<td>5. Delete Servers from UCS Server Pool</td>
<td>25. Server Maintenance</td>
</tr>
<tr>
<td>6. Associate UCS Service Profile Template</td>
<td>26. Reacknowledge Server Slot</td>
</tr>
<tr>
<td>7. Reset UCS Server</td>
<td>27. Add VLAN</td>
</tr>
<tr>
<td>10. Create UCS Service Profile from Template</td>
<td>30. Add VLAN to Service Profile</td>
</tr>
<tr>
<td>11. Create UCS Service Profile</td>
<td>31. Delete VLAN from Service Profile</td>
</tr>
<tr>
<td>12. Select UCS Service Profile</td>
<td>32. Add iSCSI vNIC from Service Profile</td>
</tr>
<tr>
<td>13. Modify UCS Service Profile Boot Policy</td>
<td>33. Delete iSCSI vNIC from Service Profile</td>
</tr>
<tr>
<td>14. Delete UCS Service Profile</td>
<td>34. Add vNIC from Service Profile</td>
</tr>
<tr>
<td>15. Associate UCS Service Profile</td>
<td>35. Delete vNIC from Service Profile</td>
</tr>
<tr>
<td>16. Disassociate UCS Server</td>
<td>36. Create Service Profile iSCSI Boot Policy</td>
</tr>
<tr>
<td>17. Disassociate UCS Service Profile</td>
<td>37. Modify Service Profile Boot Policy to Boot from iSCSI</td>
</tr>
<tr>
<td>18. Create UCS Boot Policy</td>
<td>38. Delete VLAN from Service Profile vNIC</td>
</tr>
<tr>
<td>19. Modify UCS Boot Policy LUN ID</td>
<td>39. Add VLAN to vNIC Template</td>
</tr>
<tr>
<td>20. Clone UCS Boot Policy LUN ID</td>
<td>40. Delete VLAN from vNIC Template</td>
</tr>
</tbody>
</table>

For more information, see the Cisco UCS Director solution overview.

Cisco Prime Data Center Network Manager

Cisco Prime™ Data Center Network Manager (DCNM) is a very powerful tool for centralized data center monitoring, managing, and automation of Cisco data center computing, networking, and storage infrastructure. A basic version of Cisco Prime DCNM is available for free. More advanced features require a license. Cisco Prime DCNM allows centralized management of all Cisco Nexus switches, and Cisco UCS and MDS devices.

The Cisco Prime DCNM LAN Version 6.0 was used in the sample design for management of the infrastructure. One powerful feature of Cisco Prime DCNM is the ability to have a visual, autoupdating topology diagram. The following figure depicts the sample lab topology.
Cisco Prime DCNM can also be used to manage advanced features, such as Cisco virtual PortChannels (vPC) and inventory control. An example of the visual inventory from one of the spine switches is shown in Figure 58.

Figure 58. Cisco Prime DCNM Switch Inventory
Conclusion
Cisco Nexus 9000 Series Switches are affordable, powerful switches that can easily fit in a small-to-midsize commercial data center. A two-switch deployment provides a path from 1- to 10- to 40 Gb with affordable optics as your applications grow, has a rich feature set capable of providing redundancy, enables using all links, performs equal cost multipathing, and has an optional overlay, depending on the needs of the applications. The Cisco Nexus 9000 Series Switches are also easily integrated with exiting environments without forklifting existing infrastructure.

Addendum A: Automation and Programmability
Following is information about automation tools, including Puppet.

Automation Tools
When new applications are deployed to the network, the infrastructure resources need to be provisioned to enable the applications. From a network perspective, various configuration tasks may need to be automated, such as VLAN and VXLAN provisioning, QoS, and the addition of new routes. As applications grow and change, network configuration may also need to be changed to allow or restrict communications between tiers or to tune the network based on the needs of the applications.

Network devices need to provide comprehensive automation capabilities and APIs to support the new configuration change and operation model. Enhanced Cisco NX-OS on the Cisco Nexus 9000 switch is integrated with multiple open source and commercial ecosystem partners for automation, orchestration, programmability, monitoring, and compliance support.

In addition to the Cisco NX-OS command line interface, there are multiple northbound APIs the Cisco Nexus 9000 switches offer to expose all features available on the CLI and open the door for powerful automation. Figure 59 highlights the supported automation tools.

Figure 59. Cisco Nexus 9000 API Integration
Automation with Puppet

Puppet is one of many tool options for automation on Cisco Nexus 9000 Series Switches, but it is used as the primary example in this solutions guide. Puppet is a piece of automation software that allows a system administrator to declaratively enforce the state of a device running a Puppet agent in a consistent, reusable fashion, reducing the amount of manual configuration or one-off scripts required. Puppet is widely deployed and celebrated amongst the DevOps community.

Puppet uses declarative language, meaning the administrator tells Puppet what they want the end result or configuration to be, but not the details of how it is accomplished. With a Puppet agent installed on a Cisco Nexus 9000 switch, the administrator tells Puppet the intent (a reusable set of configuration or management tasks called a manifest) and the manifest is deployed to a Nexus 9000 and converted to NX-OS configuration. An overview of Cisco Nexus 9000 Series Switches and Puppet integration is provided in Figure 60.

**Figure 60.** Puppet Workflow Overview

Helpful Links

This section provides a comprehensive list of all links included in this white paper, plus additional helpful references.

**Cisco ACI Mode**

- ACI Homepage: Cisco Application Centric Infrastructure

**External Resources**

- OpenStack at Cisco
- What Is Puppet?

**General Guides**

- All Nexus 9000 Configuration Guides
- Cisco Nexus 9000 Series NX-OS Interfaces Configuration Guide, Release 6.0
- Cisco Nexus 9000 Series NX-OS Quality of Service Configuration Guide, Release 6.0
• Cisco Nexus 9000 Series NX-OS System Management Configuration Guide, Release 6.0
• Cisco Nexus 9000 Series Switches Compare Models tool
• Cisco Nexus 9500 Series Switches Architecture
• Configuring Cisco vPCs on Cisco Nexus 9000 NX-OS
• MIBs Supported from Cisco NX-OS Release 6.1(2)I2
• Migrate from Cisco Catalyst 6500 Series Switches to Cisco Nexus 9000 Series Switches
• Cisco vPC Best Practices Design Guide

Licensing
• Licensing Cisco NX-OS Software Features

Cisco Nexus 9000 Data Sheets
• All Cisco Nexus 9000 Data Sheets
• Cisco Nexus 9500 Platform Switches Data Sheet
• Cisco Nexus 9500 Platform Switches Data Sheet
• Cisco Nexus 9500 and 9300 Series Switches NX-OS Software Data Sheet

Other Cisco Products
• Cisco ASA 5500-X Series Next-Generation Firewalls
• Cisco Nexus 5600 Platform Switches Data Sheet
• Cisco UCS Director solution overview
• Cisco Unified Computing System Express
• Storage Networking webpage

Other Overlays and Protocols
• BGP MPLS Based Ethernet VPN IETF Draft
• Cisco FabricPath webpage
• Cisco Locator/ID Separation Protocol (LISP) webpage
• Cisco NX-OS/IOS Comparison
• Cisco Overlay Transport Virtualization Technology Introduction and Deployment Considerations
• Data Center Overlay Technologies white paper
• Cisco Nexus 7000 FabricPath white paper
• Overlay Transport Virtualization webpage

Software Release Notes
• All Cisco Nexus 9000 Release Notes
• Cisco Nexus 9000 Series NX-OS Software Upgrade and Downgrade Guide, Release 6.0
• Cisco NX-OS Version 6.1(2)I2(3) release notes
Transceivers
- Cisco 10GBase SFP Modules data sheet
- Cisco Transceiver Modules compatibility information
- Cisco QSA data sheet
- Cisco QSFP BiDi Technology white paper

Unicast and Multicast Routing
- Cisco Nexus 9000 Series NX-OS Unicast Routing Configuration Guide, Release 6.0
- Configuring MSDP
- Configuring PIM

VXLAN
- Cisco Nexus 9000 Series NX-OS VXLAN Configuration Guide, Release 6.0
- Configuring VXLAN
- VXLAN Overview: Cisco Nexus 9000 Series Switches