Cisco Unified Computing System

What you will learn

Cisco Unified Computing System™ (Cisco UCS®) is an integrated computing infrastructure with intent-based management to automate and accelerate deployment of all your applications, including virtualization and cloud computing, scale-out and bare-metal workloads, and in-memory analytics, as well as edge computing that supports remote and branch locations and massive amounts of data from the Internet of Things (IoT). This document provides an overview of the main components of Cisco UCS: unified fabric, unified management, and unified computing resources.

Since the system was first announced, we have joined the top tier of server vendors with more than 85 percent of Fortune 500 companies investing in Cisco UCS. The system has set more than 150 world performance records—all serving as testament to leadership and dedication to customer-centered innovation.

Cisco UCS with AMD EPYC™ processors

Cisco UCS with Intel® Xeon® Scalable processors

Contents

About this document

This document is designed to help you learn what you need to know about Cisco Unified Computing System (Cisco UCS®) to the level of depth that interests you.

- The challenge describes the stresses facing IT organizations today.
- The solution discusses the most important reasons why Cisco UCS is such a compelling solution for today's challenges.
- Cisco UCS anatomy gives a nuts-and-bolts description of Cisco UCS in terms of how our components—Cisco Unified Fabric, Cisco UCS management, and Cisco UCS servers—combine to create a single unified system.
- The next three sections, Unified I/O architecture, Cisco UCS management, and Cisco UCS servers take a deep dive into the technology behind Cisco UCS.

The challenge

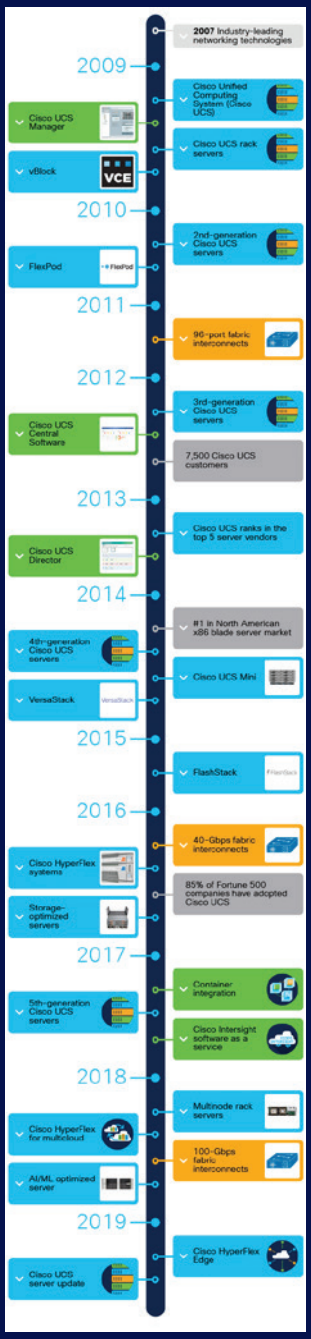
The digital business transformation is upon us, and it is creating new demands on IT organizations:

- **Applications:** Modern applications are becoming less monolithic and more like organic entities that grow and shrink through the use of modular, distributed microservices. This reduces dependence on traditional IT infrastructure and places new demands on IT organizations in terms of the large number of endpoints to manage and flexible infrastructure needed to support them. With agile development and DevOps deployment approaches becoming the norm, developers and administrators demand the capability to program their own infrastructure. This is necessary in order to quickly roll out new applications and updates to existing ones.
- **Management:** The main issue for IT organizations is managing infrastructure at scale and being able to match resources to application requirements—yet manage all of the different types of infrastructure in a simple, holistic fashion. Furthermore, the IT organization's clients—line-of-business managers, application developers, and DevOps teams—are asserting increasing influence over the adoption, purchase, and deployment of technology. IT organizations saddled with ponderous management tools are hard pressed to compete with the fluidity of self-service, multicloud environments that their clients can purchase on their own if the IT organization doesn't keep pace.
- **Location:** Once upon a time traditional IT organizations could support and manage their clients and applications safely within the glass walls of an on-premises data center. Now they must support clients, applications, and workloads in public, managed, edge, and private cloud environments—all while maintaining compliance with best practices, business and governmental regulations, and data sovereignty requirements.

As a modern IT organization, the challenge is to meet the needs of the digital business transformation while still supporting the traditional, monolithic applications that are the foundation of your business. You have to simultaneously embrace the traditional deployment model and the fluid nature of today's multicloud deployments.

Ten years of innovation

Explore our interactive timeline and learn how Cisco UCS has led the industry for more than a decade.



The solution

The straightforward solution to today's infrastructure challenges is Cisco Unified Computing System™ (Cisco UCS®). It's not a collection of servers. It's a fully self-aware, self-integrating system. Because it is 100 percent programmable, it has been the solution to computing challenges for more than 55,000 customers since 2009. The system is flexible, agile, and adaptable, and the portfolio of products supported by Cisco UCS includes blade, rack, multinode, and storage-intensive servers; converged infrastructure; hyperconverged infrastructure (Cisco HyperFlex™ systems); and solutions for the network edge such as Cisco UCS Mini and Cisco HyperFlex Edge.

There is a fundamental difference between vendors that sell servers and Cisco Unified Computing System. Servers arose as more powerful personal computers, taking many of their attributes, including time-consuming, manual, error-prone configuration of I/O, network, and storage subsystems. Traditional servers are monolithic, complex to deploy, and even more complex to adapt to new workload demands. In contrast, Cisco UCS is a single unified system, with six fundamental attributes that transformed the industry.

Fabric centric

We blend all of the system's I/O traffic into a single shared active-active network that carries all modes of communication from servers to the outside world. Our low-latency, high-bandwidth network fabric is a shared resource so networking can be allocated to interfaces based on policies rather than physical interface configuration and hard-wired cabling. The result is that you can provision and balance resources to meet your workload needs easily.

Endpoint aware

Cisco UCS was developed in the virtualization era, where the norm was multiple independent workloads running on the same server. Today the number of workloads has become practically unimaginable as containerized environments place hundreds of workloads in a single virtual machine. Whether you run virtualized, containerized, or bare-metal workloads, all I/O is virtualized. This gives you the capability to support a massive number of endpoints but with a level of control equivalent to each endpoint having its own dedicated (but virtual) cable to the outside world. This gives you the scale of virtualized, with the workload isolation and security of the physical world.

100 percent programmable

From the very beginning, Cisco UCS was designed with the entire state of each server—identity, configuration, and connectivity—abstracted into software. This makes our system fully composable: adaptable through software to meet the varying requirements of both modern workloads and traditional monolithic business applications. With a completely programmable system, you can give your clients the level of control they need to manage their workloads. For global organizations, Cisco Intersight™ software-as-a-service management gives you complete role- and policy-based control over all of your resources regardless of where they reside. Fine-grained infrastructure management can be handled in Agile development and DevOps shops with scripting languages that provide access to the Cisco UCS unified API.

Intent based

Intersight software helps you to more precisely align your infrastructure with the needs of your business. It enables administrators to automate configurations or tasks based on specific requirements that are tied to business objectives and application performance. Rather than having to be concerned with every detail of system configuration, intent-based management enables you to describe what you want to accomplish, with the cloud-based automation we provide translating your intent into action.

Delivers business benefits

Our policy-based approach to management gives you the simplicity, automation, and capabilities you need to increase productivity and support a fast-paced business environment.

- **Improved staff productivity:** By aligning your infrastructure with applications and the way teams work together, you can create synergies that can't be achieved with other architectures.
- **Better use of IT staff:** The common, simplified management of servers, storage, and fabric establishes best practices and eliminates the need to understand the nuances of specific components. You can use your subject-matter experts to develop policies and use lower-level administrators or operations personnel to implement policies.
- **Effective communication:** Cisco UCS management improves communication between roles through cross-visibility and role-based administrative access.
- **Faster time to value:** You can rapidly roll out new applications and business services at cloud-like speed and enhance the competitive strength of your enterprise.

- **Increased operational efficiency:** You can automate many routine tasks, improve resource utilization, and proactively prevent manual errors that typically keep your IT staff from working on high-value tasks.
- **Improved flexibility and agility:** The capability to automatically integrate additional resource capacity into larger, flexible pools helps ensure that your IT staff can achieve economies of scale and efficiency without greater complexity.

Analytics powered

What if your infrastructure talked directly with your support organization? The recommendation engine built into Intersight software integrates with the Cisco® Technical Assistance Center (TAC) to help you easily detect problems and initiate support requests. As the Cisco Intersight recommendation engine gains intelligence, our vision is for it to provide suggestions and recommendations for you to optimize your configurations to gain the most from your investment.

Cisco UCS anatomy

Cisco UCS is built using the hierarchy of components illustrated in Figure 1 and described in the sections that follow. Each Cisco UCS domain is established with a pair of Cisco UCS fabric interconnects, with a comprehensive set of options for connecting blade, rack, multinode, and storage servers to them either directly or indirectly.

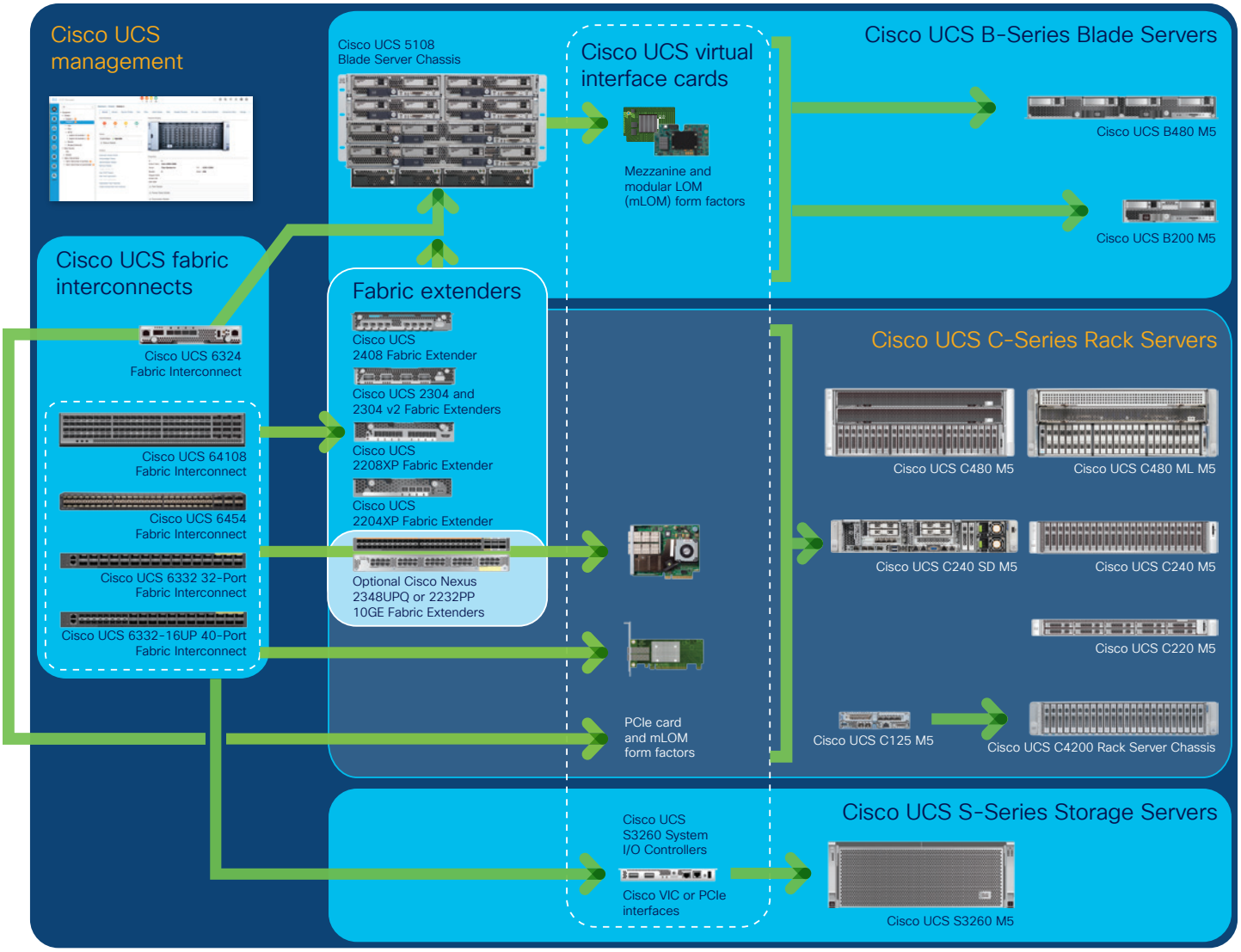


Figure 1 Cisco UCS component hierarchy

Cisco UCS management

While Cisco UCS is stateless, programmable infrastructure, the Cisco UCS unified API is how management tools program it. This enables the tools to help guarantee consistent, error-free, policy-based alignment of server personalities with workloads. Through automation, transforming the server and networking components of your infrastructure into a complete solution is fast and error-free because programmability eliminates the error-prone manual configuration of servers and integration into solutions. Server, network, and storage administrators are now free to focus on strategic initiatives rather than spending their time performing tedious tasks.

- **Cisco Intersight software-as-a-service** provides a consistent management interface for all of your Cisco UCS instances, Cisco HyperFlex clusters, edge deployments, and standalone rack servers, regardless of their location. You can access the Intersight platform through the cloud or through an optional management appliance. Intersight is designed to integrate management capabilities with a broader set of features including a recommendation engine, integration with Cisco TAC, contract management, inventory management, and alerts.
- **Cisco UCS Manager** is embedded in each fabric interconnect. Running in a redundant, high-availability configuration, it creates a single, self-aware, self-integrating unified system that recognizes and integrates components as they are added to the system. It quickly and accurately configures computing, network, storage, and storage-access resources to reduce the chance of errors that can cause downtime. Its role- and policy-based approach helps organizations more easily align policies and configurations with workloads. While Cisco UCS Manager requires an "always on" connection, our other tools are evolving to manage systems to which they are not continuously connected.
- **Cisco UCS Director** provides complete software lifecycle orchestration and management. Through a workflow-based approach, you can manage the allocation of physical and virtual infrastructure needed to deploy applications, including servers, networking, and third-party storage. Cisco UCS Director is an essential component of Cisco Integrated Infrastructure offerings that combine Cisco UCS with storage from industry-leading vendors to create effective and easily deployed solutions.
- **Cisco UCS Central Software** extends Cisco UCS Manager functions so that they can be uniformly implemented across your entire organization. For example, Cisco UCS service profiles that define how particular types of servers are centrally defined and can be applied uniformly across your entire organization, regardless of size. Cisco UCS Central Software is hosted in your own data centers.
- **An entire ecosystem** of third-party management tools integrate with Cisco UCS through the Cisco UCS management API.

Cisco SingleConnect technology

Cisco SingleConnect technology is based on a high-bandwidth, low-latency unified fabric that combines LAN, SAN, and management traffic on a single set of cables.

Cisco SingleConnect technology

SingleConnect technology provides an exceptionally easy, intelligent, and efficient way to connect and manage computing in the data center. An exclusive Cisco innovation, SingleConnect technology dramatically simplifies the way that data centers connect to rack and blade servers; physical servers and virtual machines; and LAN, SAN, and management networks.

Cisco UCS fabric interconnects

Cisco UCS fabric interconnects provide a single point of connectivity and management for an entire Cisco UCS or Cisco HyperFlex system. Deployed as an active-active pair, the system's fabric interconnects integrate all components into a single, highly available management domain. The fabric interconnects manage all I/O efficiently and securely at a single point, resulting in deterministic I/O latency regardless of a server or virtual machine's topological location in the system. Cisco fabric interconnects support low-latency, line-rate, lossless Ethernet and Fibre Channel over Ethernet (FCoE) connectivity. The Cisco UCS 6300 Series Fabric Interconnects support 10- and 40-Gbps connectivity, and the Cisco UCS 6400 Series supports 10-, 25-, 40-, and 100-Gbps. For remote and edge locations, blade-resident Cisco UCS 6324 Fabric Interconnects can be used to create a self-contained Cisco UCS Mini solution for branch offices and remote locations.

Cisco fabric extenders

Cisco fabric extenders are zero-management, low-cost, low-power-consuming devices that distribute the system's connectivity and management planes to rack servers and blade chassis to scale the system without adding complex switches or management points. Cisco fabric extenders eliminate the need for top-of-rack switches and blade-server-resident Ethernet and Fibre Channel switches and management modules, dramatically reducing the infrastructure cost per server. Rack and storage servers can be connected directly to Cisco fabric interconnects for outstanding dedicated network bandwidth. Rack servers can be connected through fabric extenders for increased scale. Regardless of connectivity method, all servers are integrated through single-wire management in which all network, storage-access, and management traffic is carried over a single set of cables.

Cisco UCS virtual interface cards

Cisco UCS virtual interface cards (VICs), extend the network fabric directly to both servers and virtual switches so that a single connectivity mechanism can be used to connect both physical and virtual servers with the same level of visibility and control. Cisco VICs provide complete programmability of the Cisco UCS I/O infrastructure, with the number and type of I/O interfaces configurable on demand with a zero-touch model.

Blade server chassis

The Cisco UCS 5108 Blade Server Chassis features flexible bay configurations for blade servers. It can support up to eight half-width blades, up to four full-width blades in a compact 6-rack-unit (6RU) form factor. The blade chassis is a highly simplified device, in contrast to traditional blade chassis that host multiple switches and management modules.

The chassis adds no points of management to the system because it is logically part of the fabric interconnects. The Cisco UCS 5100 Series Blade Server Chassis hosts two fabric extenders: low-power-consuming devices that leave the chassis with the power budget and sufficient airflow to support multiple future generations of blade servers and network connectivity options. A chassis can be deployed as a standalone Cisco UCS Mini solution by installing two Cisco UCS 6324 Fabric Interconnects in the slots that would normally be used for the fabric extenders.

Cisco UCS servers

Delivering performance, versatility, and density in servers designed without compromise, Cisco UCS blade, rack, multinode, and storage servers can power every workload, including (but not limited to) workloads for:

- Agile development environments requiring bare-metal servers
- Artificial intelligence and machine learning applications
- Big data
- Content delivery
- Cloud computing environments delivering virtual machines and bare-metal servers as a service
- Database management systems
- High-frequency trading
- Hyperconverged applications: Cisco HyperFlex™ nodes are based on Cisco UCS servers.
- Gaming applications
- Internet infrastructure applications
- Mission-critical enterprise applications
- Mobile application back-end services
- Virtualized environments

Cisco UCS is designed so that it is logically a single very large blade server chassis in which every server in our product line can connect and be managed as if it were part of the same single, unified system. In this sense, Cisco UCS is form-factor neutral, giving you more flexibility to choose the servers that best meet your needs without the penalty of having to use a different management approach for each type of server (Figure 2).

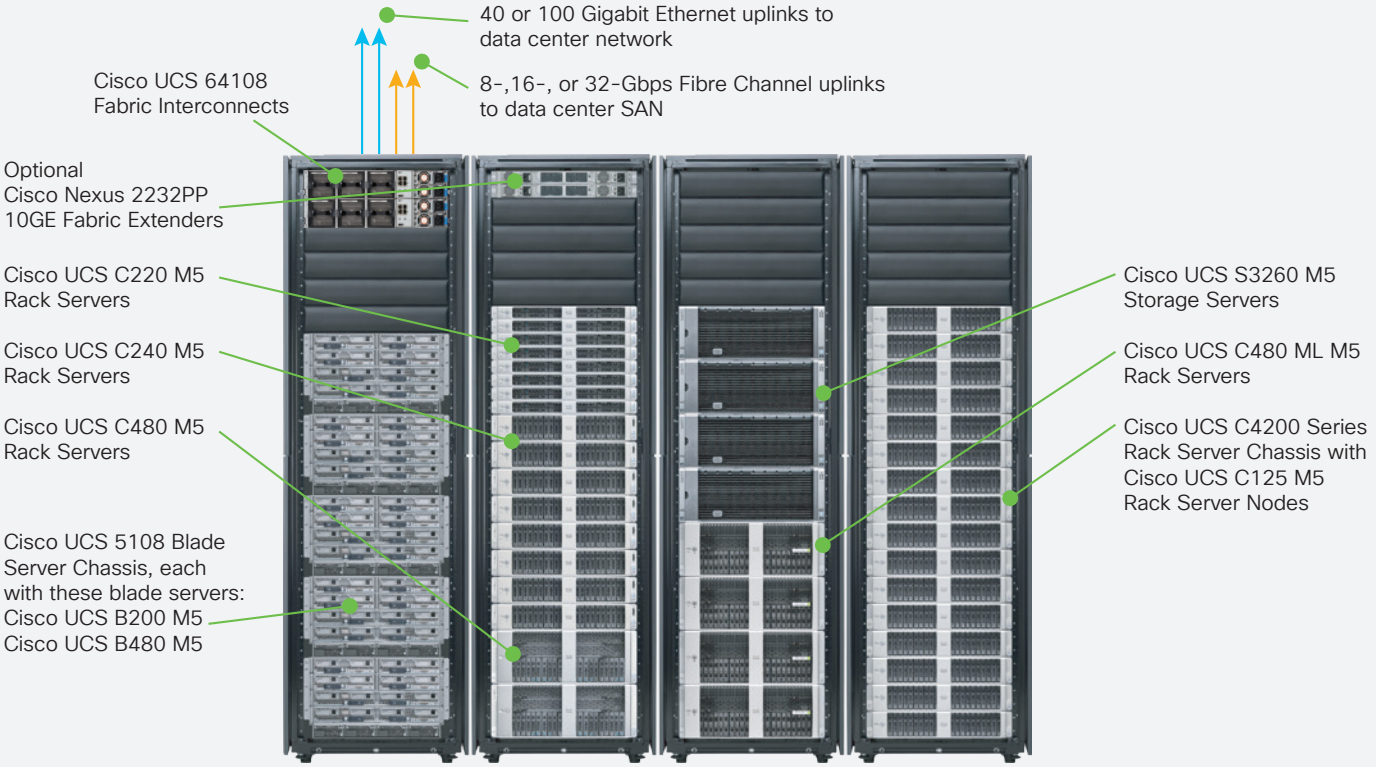


Figure 2 Cisco UCS supports blade, rack, multinode, and storage servers in a single domain of up to 160 servers

- **Cisco UCS B-Series Blade Servers** provide massive amounts of computing power in a compact form factor, helping increase density in computation-intensive and enterprise application environments. Our blade servers are available in two form factors (half- and full-width) with two or four 2nd Gen Intel® Xeon® Scalable processors. Blade servers use modular LAN on motherboard (mLOM) or mezzanine-form-factor Cisco VICs to increase I/O flexibility and accelerate deployment.
- **Cisco UCS C-Series Rack Servers** can integrate into Cisco UCS through Cisco UCS fabric interconnects or be used as standalone servers with Cisco or third-party switches. With world-record-setting performance for 2- and 4-socket servers, Cisco rack and storage servers can integrate into Cisco UCS through a single set of cables. These servers provide a wide range of I/O, memory, internal disk, solid-state disk (SSD) drive and NVMe storage device capacity, enabling you to easily match servers to workloads. The Cisco UCS C480 ML M5 server is designed for artificial intelligence and machine learning workloads. Its purpose-specific design incorporates eight NVIDIA V100 SMX2 32-GB graphics processing units (GPUs) to power compute-intensive deep learning applications.
- **Cisco UCS C4200 Series Multinode Rack Servers** are designed for clustered workloads where high core density is essential. They provide up

to 50 percent more servers per rack than our most dense blade servers, up to 242 percent more cores per rack, and up to 20 percent more storage per rack than our most dense rack servers. To achieve this density, we have equipped the Cisco UCS C125 M5 Rack Server Node with up to two AMD EPYC™ processors. With the highest core density in the industry, large memory capacity, superior memory bandwidth, massive I/O capacity, and dedicated disk storage, AMD EPYC processors combined with Cisco UCS C125 Rack Server Nodes brings more performance with less complexity to high-intensity compute clusters.

- **Cisco UCS S-Series Storage Servers** are modular servers that support up to 60 large-form-factor internal drives to support storage-intensive workloads including big data, content streaming, online backup, and storage-as-a-service applications. The servers support one or two computing nodes with up to two CPUs each, connected to a system I/O controller that connects the server to the network. These servers offer the flexibility of compute processing to balance the needed storage for workloads like big data, data protections and software defined storage.

Cisco UCS physical connectivity options

A pair of Cisco UCS fabric interconnects forms the single point of connectivity for a Cisco UCS domain. Blade servers connect through fabric extenders, most rack servers can connect through optional fabric extenders, and rack servers, multinode rack servers, and storage servers can connect directly to the fabric interconnects as illustrated in Figure 3.

- **Blade server chassis** can be connected to the fabric interconnects through a pair of blade-chassis-resident fabric extenders. Cisco UCS 2200 Series Fabric Extenders can support up to eight 10-Gbps unified fabric uplinks. Cisco UCS 2300 Series Fabric Extenders can support up to four 40-Gbps unified fabric uplinks for up to 320 Gbps of connectivity per chassis. The

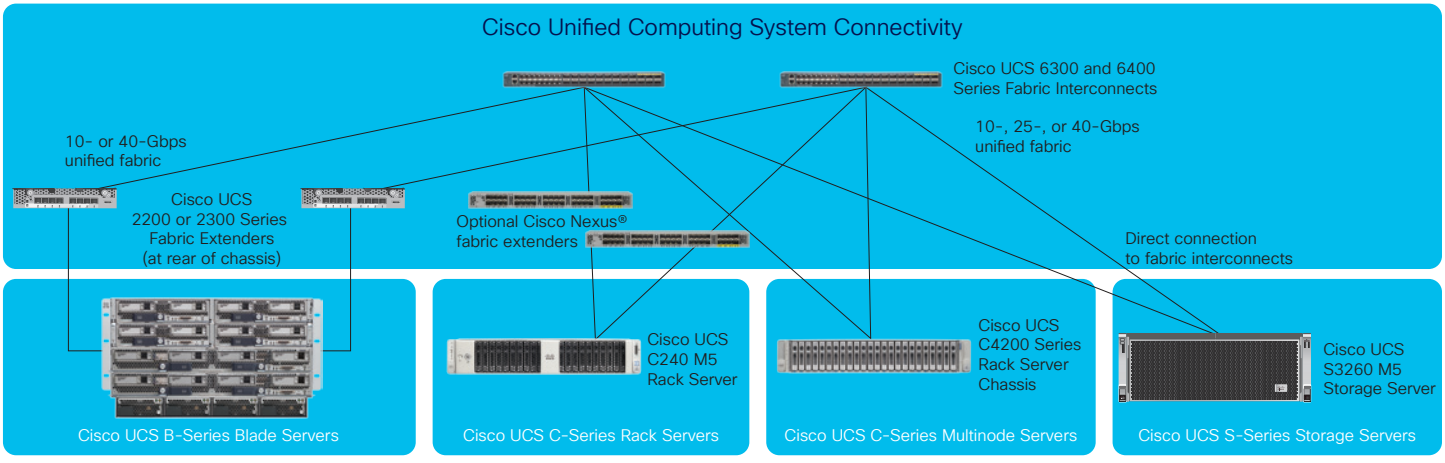


Figure 3 Cisco UCS connectivity options for blade, rack, and storage servers

Choose carefully between on-premises and cloud storage

On-premises storage with Cisco UCS S-Series Storage Servers can be less expensive over three years with only a 13-month break-even point.

Cisco UCS 2408 Fabric Extender can support up to eight 25-Gbps uplinks per fabric extender for a total of up to 400 Gbps of connectivity per chassis

- **Rack servers** can be connected directly to Cisco UCS fabric interconnects. Cisco UCS rack servers can also be connected indirectly through Cisco Nexus® 2200 or 2300 Series 10GE Fabric Extenders to achieve greater scale.
- **Multinode rack servers** are connected to Cisco UCS 6400 Series Fabric Interconnects through up to 4 10- or 25-Gbps connections, or indirectly through up to 4 10-Gbps unified fabric connections through Cisco Nexus 2200 Series Fabric Extenders.
- **Storage servers** can be connected directly fabric interconnects from the appropriate dual 10-, 40-, or 100-Gbps system I/O controller containing Cisco VIC technology.
- **Cisco UCS Mini solutions** can be created by using Cisco UCS 6234 Fabric Interconnects in the blade server chassis instead of rack-mount fabric extenders. This creates a highly versatile standalone Cisco UCS instance that can incorporate multiple blade chassis, rack servers, and Fibre Channel storage systems (Figure 4).

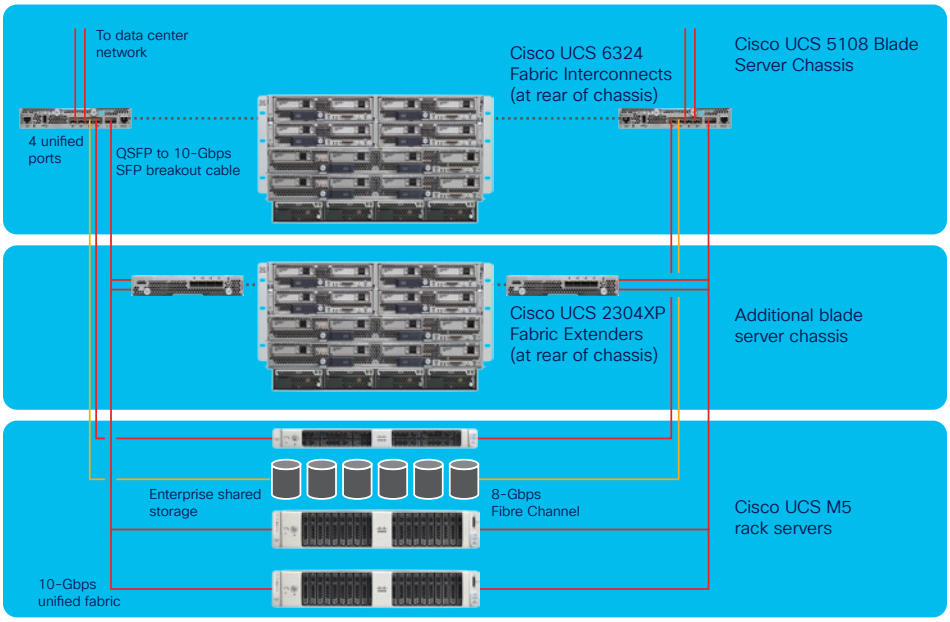


Figure 4 Example Cisco UCS Mini solution showing connectivity to additional blade chassis, rack servers, and Fibre Channel storage

Unified I/O architecture

Cisco UCS is organized around a low-latency, line-rate, lossless 10-, 25-, 40-, and 100-Gbps unified fabric that carries all I/O from servers (regardless of form factor) and virtual machines to the system's fabric interconnects. Using SingleConnect technology, the system is configured with A and B fabrics that are used in an active-active configuration to help provide high availability, increase resource utilization, and reduce costs.

Each Cisco UCS domain is wired for the desired bandwidth. Then all network features and capabilities are controlled through software settings. As a result, bandwidth is shared among all I/O modalities, so that bursts in one class of traffic can temporarily borrow bandwidth from other functions to achieve the best performance.

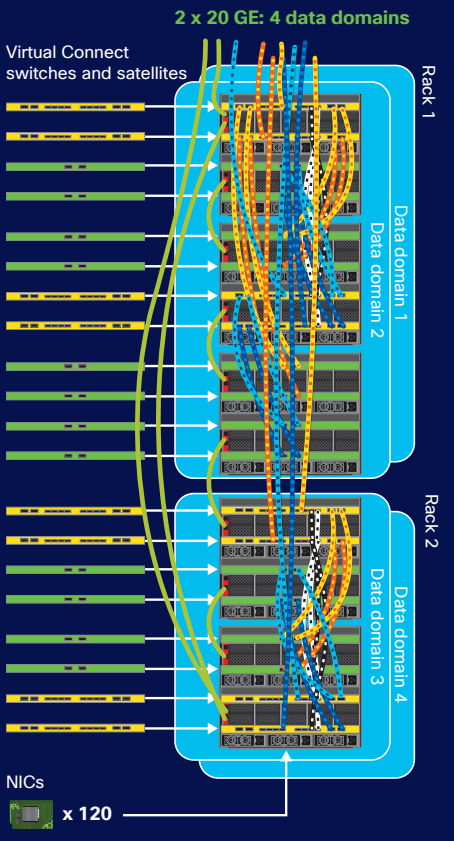
SingleConnect technology, in combination with unified management, makes the network fabric a strategic asset. The approach in which a single fabric is used to interconnect multiple systems to support all I/O modalities places Cisco UCS first in a class of solutions that Gartner Group refers to as fabric computing. The flexible, agile I/O infrastructure integrated into Cisco UCS lets you move more quickly with instant response to changing workload conditions and business priorities. With policy-based automation accelerating configuration and helping ensure consistency, the network becomes a strategic asset to business organizations.

- **Wire-once model:** The unified fabric uses a wire-once model in which you configure Cisco UCS for the level of desired capacity at deployment time. After configuration, all I/O resource allocation within that capacity is controlled through software, resulting in zero-touch, instant server and I/O configuration.
- **Flexible pool of resources:** Intelligent networking brings the server and I/O resources of Cisco UCS together as a flexible pool of resources that can be applied on demand to meet any workload challenge. Workload silos are a thing of the past because server power and I/O connectivity can be allocated instantly and accurately through software.
- **Application-centric configuration:** In Cisco UCS, the system adapts to the needs of applications, in contrast to systems in which applications run only on the servers that have been designed and deployed to support them. In virtualized environments, SingleConnect technology can create multiple Fibre Channel host bus adapters (HBAs) and separate network interface cards (NICs) to accommodate shared storage, a management network, another network for virtual machine movement, and multiple networks for LAN traffic—all moments before the hypervisor is booted.

Alleviate nightmares

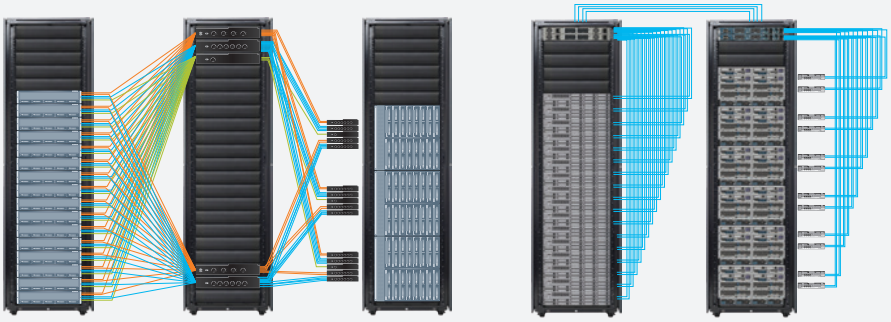
Compare Cisco UCS simplicity to the complexity of HPE Synergy. A 5-frame, 2-rack configuration scaled to support 40 Gbps of network bandwidth requires:

- 8 HPE Virtual Connect switches
- 4 network domains and uplinks for each
- Separate management network
- 12 satellites
- 120 NICs
- 10 frame link modules
- 2 Synergy composer appliances
- Fibre Channel adds even more complexity



Condense multiple parallel networks

Traditional environments must implement multiple parallel networks to support management, IP networking, and storage access (Figure 5). The result is a proliferation of NICs and HBAs in each server, and upstream switch ports. Each separate network must be sized to handle workload bursts without the capability to share bandwidth between the multiple networks. The complexity of maintaining so much physical infrastructure can lead to cabling errors that can cause downtime or result in security vulnerabilities. Server airflow can be obstructed by the massive number of cables, increasing server temperature and reducing performance.



Traditional systems require multiple parallel networks for IP, management, and storage access traffic

Cisco UCS is wired with a single redundant network to support all I/O

Figure 5 Cisco Unified Fabric eliminates multiple parallel networks, simplifying blade, rack, and storage server environments

In contrast, SingleConnect technology allows a single unified network to bring LAN, SAN, and management connectivity to each server in a Cisco UCS domain using Cisco Unified Fabric. All three networks are carried over a single set of cables that securely carries production data (Ethernet) traffic, Fibre Channel traffic through FCoE, and management traffic. Every server—rack, blade, or storage—has equal access to all network resources, eliminating the need to support three physical interfaces, each with its own NICs, HBAs, transceivers, cables, top-of-rack switches, and upstream switch ports.

Infrastructure silos are eliminated because software—not cabling—determines the way that each server connects to the network. Every server is ready to support any workload at a moment's notice through automated configuration. Instead of requiring separate physical networks to be sized for each traffic class, the shared I/O resources in the unified fabric enable more flexible resource allocation: bursts of traffic in one resource class can borrow bandwidth from other classes subject to quality-of-service (QoS) limitations. Compare this to a 40-Gpbs configuration in HPE Synergy (see sidebar) which requires complex network cabling to interconnect various modules across the back and between racks.

Cisco Unified Fabric is based on open standards, some of which are based on Cisco innovations that were later approved by standards bodies:

- The lossless network fabric is implemented using the IEEE 802.3x PAUSE mechanism.
- Different classes of traffic are prioritized using the IEEE 802.1p Priority Flow Control capability; for example, management traffic has its own traffic class and is given the highest priority so that unified management can still function during even the most adverse traffic conditions.
- Network congestion is mitigated using the IEEE 802.1Qau Quantized Congestion Notification standard.
- Bandwidth management is handled through IEEE 802.1Qaz Enhanced Transmission Selection.
- FCoE is implemented following the International Committee for Information Technology Standards' T11 FC-BB-5 standard.

Reduce network layers

In traditional blade server environments, especially virtualized ones, the network access layer is fragmented into three layers, making visibility and control over network connectivity difficult to maintain. These layers add unnecessary and variable latency to virtual networks, they are complex because they usually have different feature sets, and they fragment access-layer management between network and server administrators.

A typical virtualized environment requires three network hops to reach the access-layer switch. As shown in Figure 6, one hop takes place between a virtual machine and the software switch, one between the software switch and the blade-chassis switch (or top-of-rack switch in rack server environments), and one between the blade-chassis switch and the access-layer switch.

SingleConnect technology brings the unified fabric to every blade chassis and server rack in Cisco UCS. This approach eliminates switches in the blade server chassis, reducing latency by one hop. Cisco Fabric Extender Technology, a prestandard implementation of the IEEE 802.1BR Bridge Port Extension standard, eliminates a switching layer with an architecture that is physically distributed but logically centralized. All network traffic passes through the system's fabric interconnects, establishing a single point of management and control.

Cisco Fabric Extender Technology brings the network fabric to blade chassis and to racks, passing all traffic to the fabric interconnects in a lossless manner. Cisco fabric extenders are low-cost, low-power-consuming devices that are physically distributed throughout a Cisco UCS deployment but remain logically part of the fabric interconnects, maintaining a single point of management for the entire system.

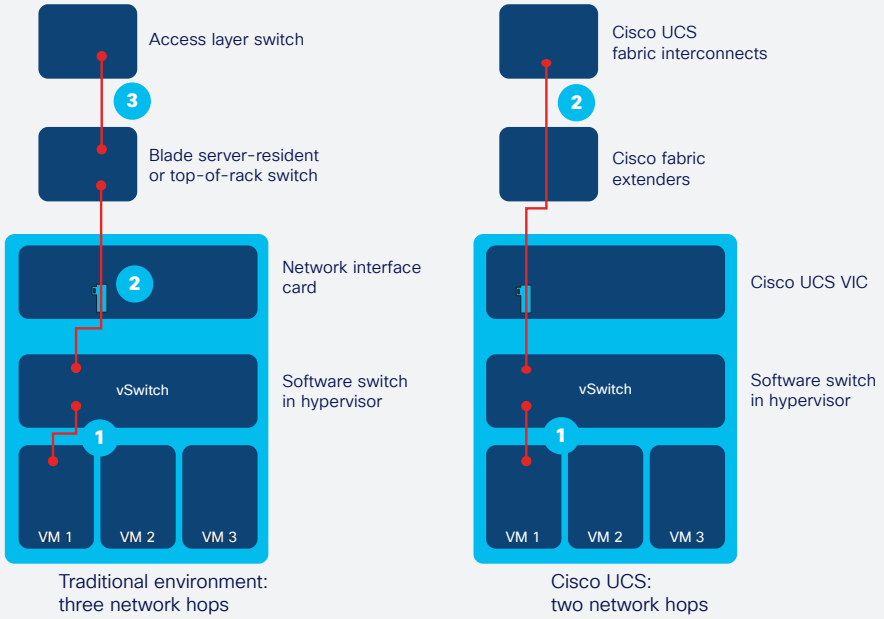


Figure 6 Cisco Unified Fabric eliminates multiple parallel networks, simplifying blade, rack, and storage server environments

- **In rack server environments,** Cisco Nexus fabric extenders bring the system's unified fabric to the top of every rack, simplifying rack-level cabling.
- **In blade server environments,** Cisco UCS fabric extenders bring the system's unified fabric to every blade server chassis, with the Cisco UCS 2408 Fabric Extenders supporting up to 400 Gbps of bandwidth for an 8-blade chassis.

The logical centralization of the I/O infrastructure means that after the system is established, it can scale without the need to reevaluate the infrastructure or configure additional switches to incorporate additional servers.

Virtualize networking

The unified fabric virtualizes I/O. Rather than requiring each server to be equipped with a set of physical I/O interfaces and cables for separate network functions, all I/O in the system is carried over a single set of cables and sent to separate physical networks at the system's fabric interconnects as necessary. For example, storage traffic destined for Fibre Channel storage systems is carried within the system using FCoE. At the fabric interconnects, storage traffic can transition to physical Fibre Channel networks through one or more of the fabric interconnect's unified ports.

I/O is further virtualized through the use of separate virtual network links for each class and each flow of traffic. For example, management, storage-access, and IP network traffic emanating from a server is carried to the

No special software required

Cisco VICs present physical PCIe-compliant devices to operating systems and hypervisors, supporting them without the need for them to implement Single Root I/O Virtualization (SR-IOV). This approach allows almost any hypervisor or operating system to be supported without additional complexity.

system's fabric interconnects with the same level of secure isolation as if it were carried over separate physical cables (Figure 7). These virtual network links originate within the server's virtual interface cards and terminate at virtual ports within the system's fabric interconnects.

Virtual network links are managed exactly as if they were physical. The only characteristic that distinguishes physical from virtual networks within the fabric interconnects is how the ports are named. This approach has a multitude of benefits: changing the way that servers are configured makes servers flexible, adaptable resources that can be configured through software to meet any workload requirement at any time. Servers are no longer tied to a specific function for their lifetime because of their physical configuration. Physical configurations are adaptable through software settings. The concept of virtual network links brings immense power and flexibility to support almost any workload requirement through flexible network configurations that bring complete visibility and control.

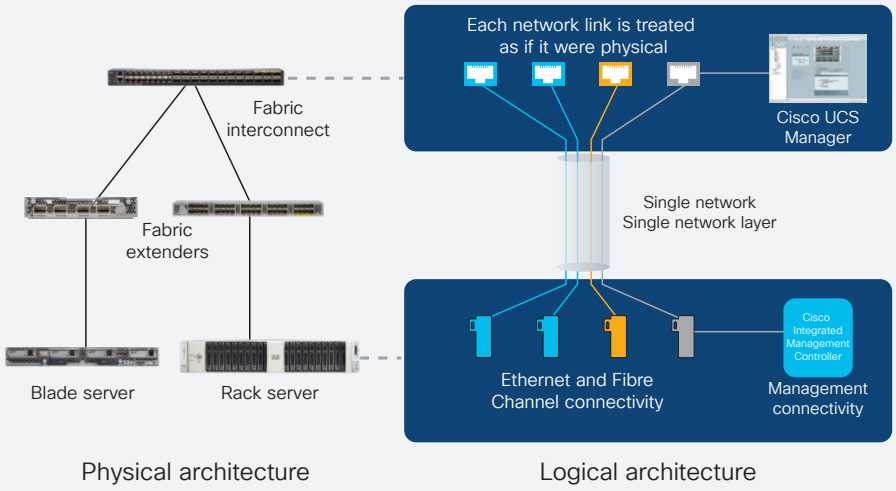


Figure 7 Cisco Unified Fabric eliminates multiple parallel networks, simplifying blade, rack, and storage server environments

Virtualize I/O interfaces

Cisco VICs are PCIe-compliant interfaces that support 256 or more PCIe devices with dynamically configured type (NIC or HBA), identity (MAC address or worldwide name [WWN]), fabric failover policy, bandwidth, and QoS policy settings. With Cisco VICs, server configuration—including I/O configuration—becomes configurable on demand, making servers stateless resources that can be deployed to meet any workload need at any time, without any physical reconfiguration or recabling required.

Cisco VICs support up to 200 Gbps of connectivity and are available in multiple form factors:

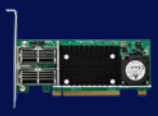
Cisco VICs

Cisco VICs are available in multiple form factors so that every Cisco UCS server can support virtualized I/O:

Modular LOM (blade servers)

Modular LOM (rack servers)

Mezzanine



PCIe

System I/O controller

- **mLOM:** Modular LAN-on-motherboard adapters are available for both blade and rack servers. These have the benefit of not occupying a mezzanine slot (in blade servers) or a PCIe slot (for rack servers), leaving the space for other devices. The Cisco UCS VICs 1340 and 1440 support up to 40 Gbps in blade servers, and up to 80 Gbps with an optional port expander card. In rack servers, the Cisco UCS VIC 1387 supports up to two 40-Gbps connections; the Cisco UCS VIC 1457 supports up to four 25-Gbps and the 1497 supports up to two 100-Gbps connections.
- **Mezzanine:** Standard Cisco VICs can be installed in any blade server's mezzanine slot: one for half-width blade servers and up to two for full-width blade servers. Each Cisco VIC supports up to 80 Gbps of connectivity per blade server.
- **PCIe:** these Cisco VICs integrate Cisco rack and storage servers into Cisco UCS through circuitry that passes the unified fabric's management traffic to the server's management network, enabling single-wire, unified management of rack servers (see "Integrated operation with single-wire management" on page 34
- **System I/O controller (SIOC):** Rack servers such as Cisco UCS S-Series Storage Servers support I/O through SIOCs. One model integrates Cisco VIC silicon for up to 80 Gbps of connectivity. Another provides a PCIe slot that can support the latest Cisco VICs and a range of third-party adapters.

Support best practices

Virtual interfaces can be used to support operating system requirements and best practices. Figure 8 shows devices supporting a hypervisor, with separate interfaces for the hypervisor console, the hypervisor management interface, virtual machine movement traffic, and two Fibre Channel HBAs to provide access to shared storage. Pairs of Ethernet NICs are shown connected to virtual switches in the hypervisor. Each interface connects to a virtual link that terminates at a virtual port in the fabric interconnect, giving the flexibility of virtual networking with the visibility and control of physical networks.

Integrate with data center networks and Cisco Application Centric Infrastructure

Condensing multiple network layers into a single physically distributed but logically centralized connectivity domain makes the process of integrating Cisco UCS into data center networks straightforward. Rather than appearing as an entire network with a collection of servers, Cisco UCS is integrated as a single system. This simplification is accomplished using both Ethernet and Fibre Channel end-host modes. It eliminates the need for Spanning Tree Protocol, and it pins the MAC addresses and WWNs for both physical and virtual servers at the uplink interfaces. This approach gives the fabric interconnects complete control over the unified fabric within Cisco UCS and allows greater utilization of uplink port bandwidth through the use of active-active Ethernet uplinks.

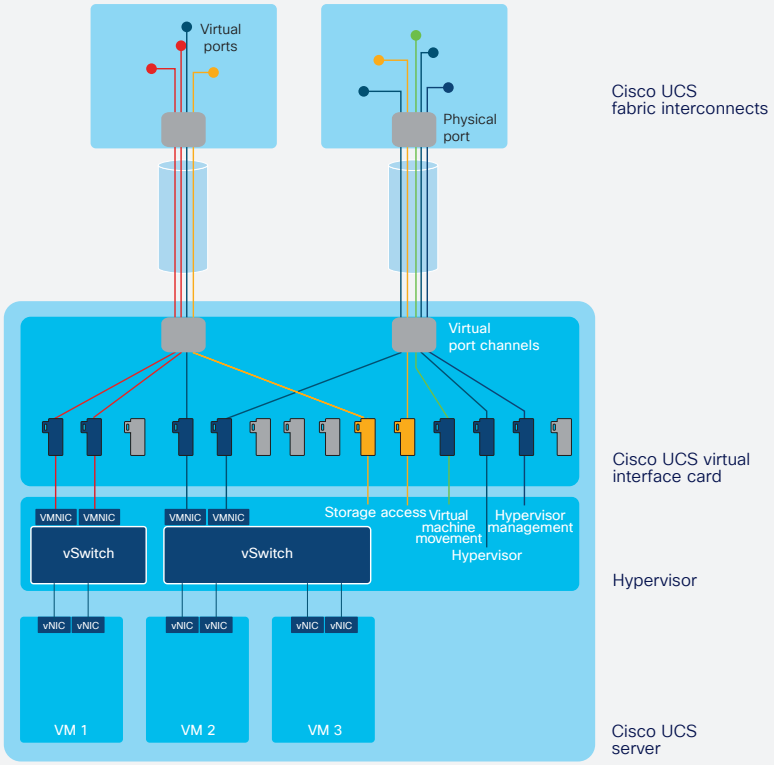


Figure 8 Cisco VIC with fabric failover enabled for Ethernet NICs and not for Fibre Channel HBAs, following best practices

With the fabric interconnects in end-host mode, Cisco Application Centric Infrastructure (Cisco ACI™) leaf switches see their connections to fabric interconnects as if they were connected the servers themselves.

With Cisco UCS servers and virtual machines grouped into specific VLANs, and with each VLAN associated with a specific Cisco ACI endpoint group (EPG), any server appearing in that VLAN is automatically associated with the correct EPG. The Cisco ACI encapsulation normalization mechanism automatically wraps any upstream traffic from these servers in Virtual Extensible LAN (VXLAN) tunnels so that the traffic is securely isolated as if it were on its own physical network segment.

If you wish to extend Cisco ACI VXLAN tunneling closer to virtual machines, the Cisco Application Virtual Switch acts as a virtual leaf switch, extending the policy-based networking of Cisco ACI into the virtual environment.

Fabric failover best practices

Cisco VICs support fabric failover at the device level, so losing one of the two fabrics can be transparent to the operating system or hypervisor. They can implement best practices for each type of network:

- **For LAN connections,** failover can be performed by the Cisco VIC so that the operating system or hypervisor can continue to operate without knowledge of the failure
- **For SAN connections,** failover is usually handled by the operating system drivers, so fabric failover is typically not used.

Architecture for high availability

Cisco UCS is designed for high availability, with no single point of failure in its network infrastructure. The fabric interconnects are designed to work in an active-active model, with automated failover of network management in the event of a failure. The system is designed so that if either fabric A or fabric B fails, the remaining fabric will take on the traffic from the failed fabric. Cisco VICs support fabric failover by moving traffic from one fabric to the other according to failover policies established on a per-NIC basis. This eliminates complicated operating system or hypervisor NIC teaming configuration.

Fabric failover makes the failure of a fabric transparent to the operating system or hypervisor. Figure 9 illustrates that different types of network modalities have different best practices for failover, and Cisco VICs support both. Fabric failover is typically configured for Ethernet because it avoids complex operating system and hypervisor configuration. In the example, vNIC 1 connects primarily to fabric A but fails over to fabric B in the event of a fabric failure. Fibre Channel is typically cabled with two networks with the driver software handling failover, so HBAs 3 and 4 connect to fabrics A and B, respectively, without fabric failover configured.

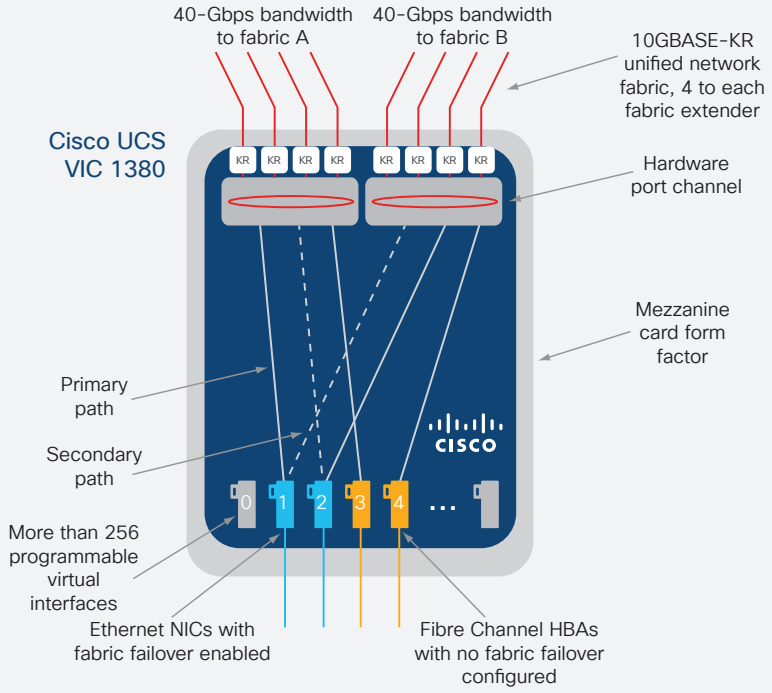


Figure 9 Cisco VIC with fabric failover enabled for Ethernet NICs and not for Fibre Channel HBAs, following best practices

Cisco UCS management

The foundation of Cisco UCS management is The Cisco UCS unified API. It enables an entire ecosystem of high-level tools that are discussed in the next section. The API supports the most fundamental management tool—Cisco UCS Manager—which is embedded in every system's fabric interconnects.

Programmable infrastructure

Cisco UCS is the only system designed from the beginning to abstract every aspect of system personality, configuration, and connectivity from the hardware so that it can be configured through software. This makes Cisco UCS fully composable infrastructure. Everything from firmware revisions to network profiles are abstracted into more than 125 configuration variables that fully specify each server. In this sense the system is stateless, and any of the management tools using the Cisco UCS unified API can set server state and thus automate the integration of servers into systems.

From one perspective, Cisco UCS does for physical servers what hypervisors do for virtual machines: it creates an environment in which any server can be configured through software to support almost any workload. This increases flexibility, utilization, and business agility.

The power of a unified API

The Cisco UCS unified API enables higher-level tools to turn the knobs and program the infrastructure by setting the abstracted state variables. This applies whether your Cisco UCS servers are deployed within a single UCS instance composed of fabric interconnects, and blade and/or rack servers, Cisco HyperFlex clusters, and even standalone servers not connected to fabric interconnects. The Cisco UCS unified API is uniquely positioned to adopt evolving standards while continuing to provide a premier framework for automation (Figure 10).

Cisco Intersight software-as-a-service management

This platform has the broadest scope of the Cisco UCS management tools. It enables programming the infrastructure by automating configuration and management, but it goes the farthest in integrating with outside services and tools.

Accessed from the cloud or through an optional local management appliance, Intersight provides a single interface from which you can undertake lifecycle management of your servers and Cisco Nexus® switches whether they are in a core data center or at the network edge. New features

are continually integrated and you can keep up to date on the most current enhancements by visiting cisco.com/go/intersight.

The Intersight platform enables you to configure the identity, personality, and connectivity of blade and rack servers. It provides access to automated deployment for entities such as Cisco HyperFlex clusters, and it can automate and support large-scale edge deployments. Intersight provides the following additional capabilities that are complementary to the basic deployment and configuration features:

- **Global dashboard and inventory.** When you manage your infrastructure with Cisco Intersight, you can view a global dashboard that gives you overall server status and enables you to drill down to view individual components (such as disk drives) With a global inventory of your devices, it's easy to track the location of each of your assets.
- **Cisco TAC.** With Intersight's integration with Cisco TAC, you can quickly remediate problems because expertise and information can flow seamlessly

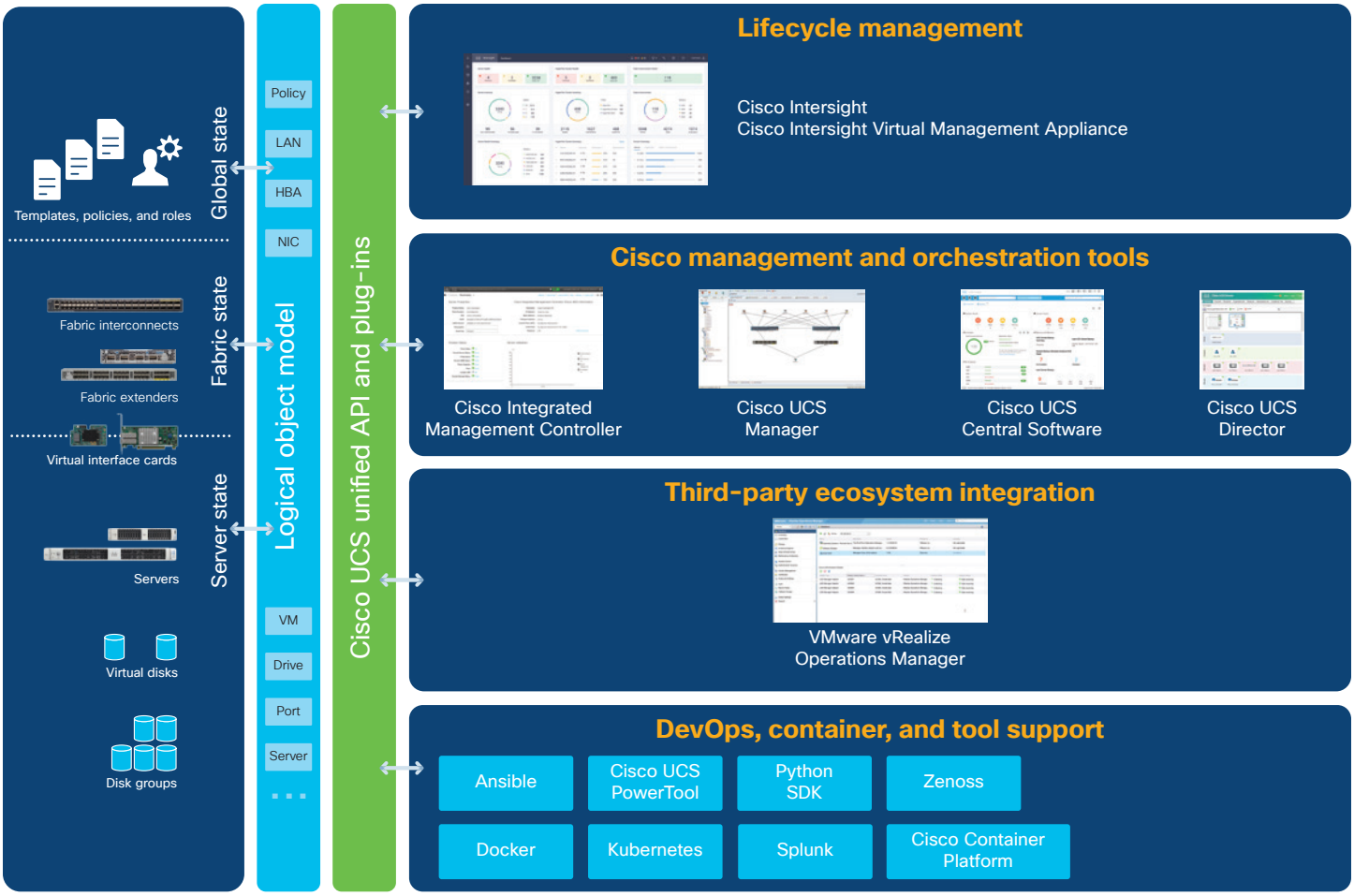


Figure 10 The Cisco UCS unified API enables a vibrant ecosystem of Cisco and third-party management tools

Ready for modern practices

Through a unified API, Cisco UCS adapts easily to support new practices. Cisco UCS infrastructure can be programmed on demand to support DevOps practices through a variety of tools and scripting languages.

The following tools, described in our management handbook, are designed to support programmable configuration of standalone rack servers:

- PowerTool
- Cisco IMC Python software development kit (SDK)
- IMC Ruby SDK
- IMC Ansible SDK
- IMC Nagios SDK

between Intersight and your Cisco support center. The system can open cases and upload supporting documentation for fast resolution. It maintains the status of your contracts and licenses so that you can administer them from the same interface.

- **Recommendation engine.** This feature gives you recommendations on configurations and help you implement best practices. Intersight has insight into your operating system and driver versions. It can use these to validate that your implementations are supported by Cisco's hardware configuration list (HCL).

Cisco management and orchestration tools

These give you access to the fundamental Cisco UCS management concepts described in the following sections, including inventory and resource pools, role- and policy-based management, Cisco UCS service profiles and templates, and logging and audit capabilities.

- **Cisco UCS Manager** manages a single domain through an intuitive HTML 5-based GUI. The system is self-aware and self-integrating. Because Cisco UCS integrates computing and networking, it is aware of any device connected to it. When the system is first powered on, or when a new component is added, the system discovers the components and adds them to its internal model of the system configuration. Cisco UCS Manager is model based, allowing administrators to create models of desired server configurations and then configure them simply by associating a model with the physical resources. The system helps guarantee consistent, error-free, policy-based alignment of server personalities with workloads, increasing standards compliance. Server configuration is essentially guaranteed to be correct because Cisco UCS Manager automatically sequences configuration steps and backs them out if an error occurs. Role- and policy-based management preserves current administrator roles (server, storage, and network), helps administrators be more effective in their jobs, helping to reduce overall cost. After configuration, Cisco UCS Manager aggregates element monitoring so that every aspect of the system can be monitored from the GUI or through higher-level tools.
- **Cisco UCS Central Software** extends the scope of Cisco UCS Manager to all of your domains worldwide (up to 10,000 servers). It does so by coordinating Cisco UCS Manager instances. It provides global awareness of inventory, automated standards compliance, increased business ability, and increased asset utilization, and it helps you to meet and exceed service-level agreements (SLAs).
- **Cisco UCS Director** provides complete application lifecycle orchestration and management, extending the reach beyond Cisco UCS server and storage access resources into higher-level switching and storage that is incorporated into Cisco integrated infrastructure solutions.
- **Cisco Integrated Management Controller** (IMC) gives you access to standalone rack server configuration. For more information see "Standalone operation with the Cisco IMC" on page 35.

Cisco UCS integrated infrastructure solutions

We offer solutions from an ecosystem of integrated infrastructure partners, including:

- Pure Storage (FlashStack)
- NetApp (FlexPod)
- Microsoft (AzureStack)
- IBM (VersaStack)
- Dell EMC (VxBlock Systems).
- Hitachi (Cisco and Hitachi Adaptive Solutions for Converged Infrastructure)

Third-party ecosystem integration

Tools illustrated in Figure 10 provide seamless integration between management tools. VMware vRealize Operations suites integrate with the Cisco UCS unified API. Within virtual machines, Cisco Container Platform can help deploy containerized environments in multicloud environments.

DevOps and tool support

The Cisco UCS unified API is of great benefit to developers who want to treat physical infrastructure the way they treat other application services, using processes that automatically provision or change IT resources. Similarly, your IT staff needs to provision, configure, and monitor physical and virtual resources; automate routine activities; and rapidly isolate and resolve problems. The Cisco UCS unified API integrates with DevOps management tools and processes, and enables you to easily adopt DevOps methodologies.

The Cisco UCS Platform Emulator facilitates the use of Cisco UCS Manager and the Cisco UCS API without requiring physical hardware. The emulator significantly shortens the development cycle for software that interfaces with the Cisco UCS API. You can create and test programs using the emulator installed on a PC or laptop computer.

Partner ecosystem and customization

The API provides a unified control plane that facilitates deep integration with third-party IT operations management tools. Building on this API, we collaborate with a broad set of ecosystem partners to integrate unique capabilities within their native functions and user interfaces. These integrations enhance performance monitoring of the operating system and higher layers of the application stack and enable consistent cross-system management, which is especially important in heterogeneous infrastructure environments.

Cisco UCS management concepts

All of the tools mentioned in the previous section use the same fundamental concepts that Cisco UCS Manager exposes though the unified API. The software runs in the system's fabric interconnects and is configured in a high-availability arrangement with two active fabric interconnects. Cisco UCS Manager supports management domains of up to 160 blade, rack, multinode, or storage servers in any combination.

Inventory and resource pools

When the system is powered on, or when new components are configured in the system, Cisco UCS Manager adds the components to a hierarchical model that represents all the objects in the system. This model acts as the single source of truth for all connected components and their configuration.

Physical components are configured by manipulating the model, not the actual devices.

Cisco UCS Manager can classify servers into resource pools based on criteria including physical attributes (such as processor, memory, and disk capacity) and location (for example, blade chassis slot). Server pools can help automate configuration by identifying servers that can be configured to assume a particular role (such as web server or database server) and automatically configuring them when they are added to a pool.

Resource pools are collections of logical resources that can be accessed when configuring a server. These resources include unique user IDs (UUIDs), MAC addresses, and WWNs.

Role- and policy-based management

In typical use cases, subject-matter experts define the way different classes of systems are to be configured by creating resource pools and policies that cover their specific domains of expertise. For example, network administrators can create policies that determine every aspect of the way that a Microsoft Exchange Server should connect to the network. These policies can indicate that certain aspects of identity should be drawn from a specific resource pool—for example, a pool of MAC addresses dedicated to Microsoft Exchange Server NICs.

Cisco UCS Manager supports typical divisions of responsibility in IT departments while giving each role visibility into the actions taken by other roles, enhancing communication and simplifying coordination between roles. Role- and policy-based management makes organizations more effective because administrators can configure policies once, and then administrators at any level of authority can use these policies to configure a server.

Cisco UCS service profiles and templates

Administrators can select among policies to create a Cisco UCS service profile, which is the complete specification detailing how a system should be identified, configured, and connected to IP and storage networks (Figure 11). For example, an administrator might select server, network, and storage access policies designed to support Oracle database servers.

Cisco UCS service profiles also can include storage profiles. Disk groups specify a group of disks characterized by the number and type of disks, RAID level, and number of spares. From a disk group, administrators can create virtual disks that are connected to servers as if they were physical drives. This capability is particularly useful for configuring disks in storage-intensive servers (such as the Cisco UCS C240 rack servers and S3260 Storage Servers).

Whereas a Cisco UCS service profile dictates how to configure a single server, Cisco UCS service profile templates dictate how to create multiple service profiles. These templates can be used to create Cisco UCS service

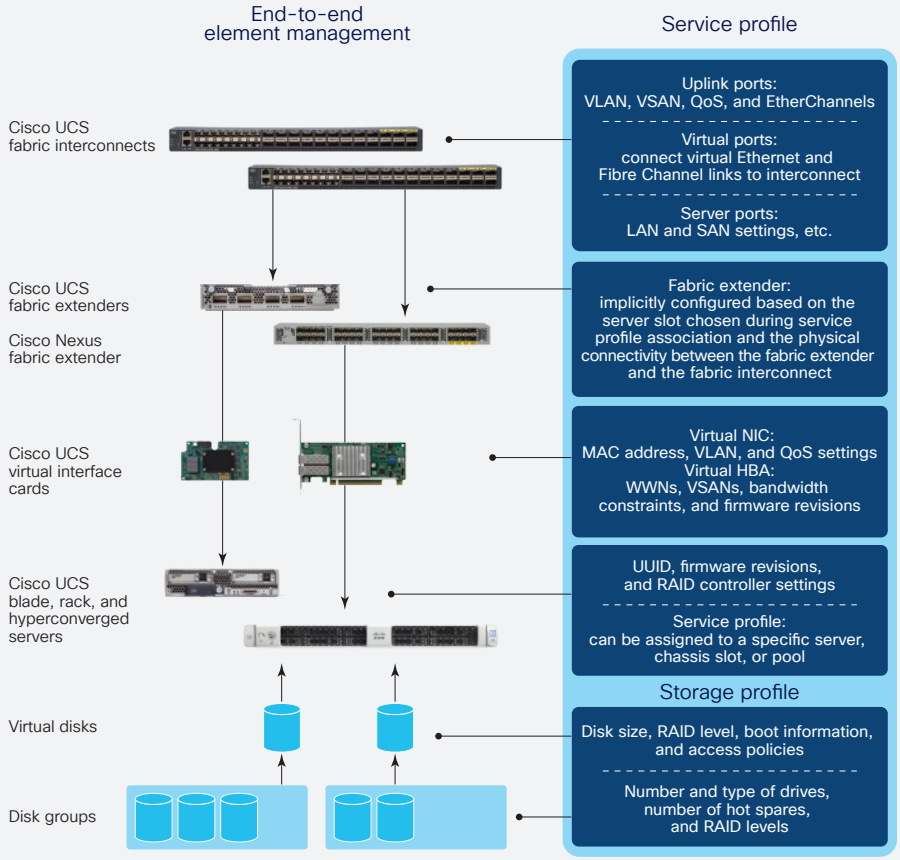


Figure 11 Cisco UCS service profiles fully specify a server's identity, configuration, storage, and connectivity

profiles to configure hundreds of servers as easily as you can configure one. Cisco UCS service profiles and templates allow a Cisco UCS domain to be treated as a flexible, composable pool of resources that can be configured rapidly and accurately to support changing workloads and business conditions:

- **Server configuration**, including changes in the number and type of I/O devices, is completely automated with a zero-touch model.
- **Applications** can quickly be scaled by adding new servers under the direction of service profiles, accelerating the movement of servers from the loading dock into production.
- **Servers** can be repurposed on demand to meet immediate workload requirements. For example, a server that supports a web server farm during the day can be repurposed to support a virtualization cluster at night simply by changing the service profile associated with the server.
- **Firmware** can be revised simply by changing the specification in a Cisco UCS service profile and applying it to a server. Changing versions in a Cisco

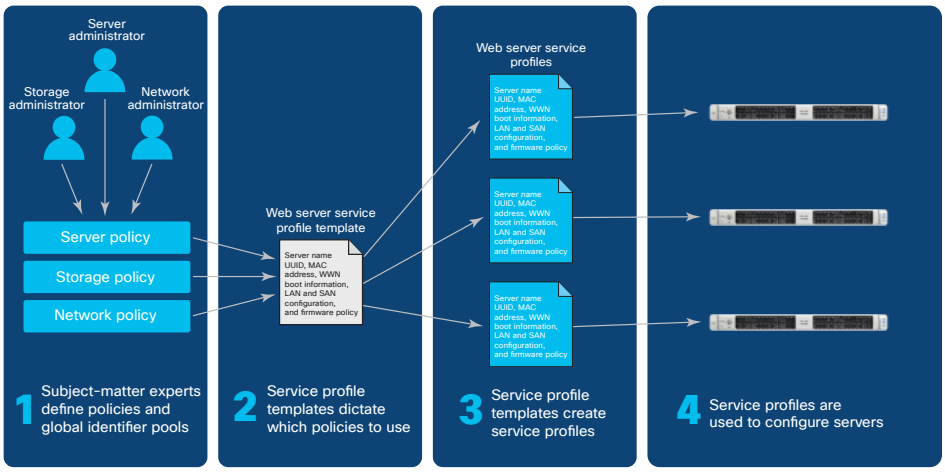


Figure 12 Workflow for role- and policy-based server configuration with Cisco UCS service profiles

UCS service profile template can cause all server configurations derived from it to be updated.

Figure 12 illustrates a workflow in which subject-matter experts create policies that contribute to the definition of a Cisco UCS service profile template. Cisco UCS service profile templates are then used to create multiple Cisco UCS service profiles. When applied to servers, the Cisco UCS service profiles completely specify the server personality, configuration, and connectivity. Cisco UCS service profiles can be specified as "updating," so that changes to the template change the Cisco UCS service profiles derived from that template. Updating service profiles can allow you, for example, to change all firmware versions in a pool of servers at one time simply by changing the specification in the template.

Logging and audit capabilities

Cisco UCS Manager produces detailed logs that show how servers are configured and deployed. Because these logs are complete, accurate, and reflect all changes to any component in the entire system, they can be used to detect and automatically remediate any unauthorized change in server configuration, eliminating the configuration drift that can make a server become noncompliant and potentially cause downtime. In the event that a hardware failure does occur, the same information can be used to reboot the software that was running on a failed server onto a known good server. Backup servers can be recruited from a backup pool of servers.

Cisco UCS M5 servers

Designed to provide your computing infrastructure now and into the future, Cisco UCS M5 servers give you the benefits of the latest processors:

- **More cores** to accelerate parallelized virtualized and bare-metal workloads
- **Larger memory capacity** for better performance and larger in-memory databases
- **Higher memory bandwidth** to accelerate the flow of information to and from the CPU
- **Up to 8 GPU accelerators** for the most demanding machine learning workloads, or a smooth user experience in virtual desktop environments
- **Cloud management ready** and prepared to connect to the Cisco Intersight platform

Cisco UCS servers

Cisco UCS is based on industry-standard, x86-architecture servers with Cisco innovations. Although many vendors offer servers with the same processors, we integrate them into a system with a better balance of CPU, memory, and I/O resources. This balance brings processor power to life with more than 150 world-record-setting benchmark results that demonstrate our leadership in application areas including virtualization, cloud computing, enterprise applications, database management systems, enterprise middleware, high-performance computing, and basic CPU integer and floating-point performance metrics.

Match servers to workloads

The breadth of our server product line makes the process of matching servers to workloads straightforward, enabling you to achieve the best balance of CPU, memory, I/O, internal disk, and external storage-access resources using the blade, rack, multinode,or storage server form factor that best meets your organization's data center requirements and preferred purchasing model. Our server family is featured in the sidebars on this and the following pages.

Powered by AMD EPYC processors

Our multinode server is targeted to deliver maximum density, which is what drove our decision to power the Cisco UCS C125 M5 Rack Server Node with 2nd Gen AMD EPYC processors (see the sidebar: "Cisco UCS M5 multinode rack servers" on page 33). 2nd Gen AMD EPYC processors support up to 64 cores per CPU for up to 512 cores in a single 2RU chassis, delivering 242 percent more cores per rack than our most dense rack servers—the highest core density in the industry. These processors deliver more than just density: at launch, 2nd Gen AMD EPYC processors captured 80 world performance records across a wide range of workloads.

The 2nd Gen AMD EPYC processor's high core density is matched with more memory bandwidth to make compute-intensive applications run faster. The processor's 128 lanes of I/O capacity are harnessed to speed I/O from fourth-generation Cisco UCS VICs and direct connectivity to internal disk drives and NVMe storage. A silicon-embedded security processor manages up to 509 encryption keys that enable individual virtual machines to be encrypted, helping to protect against a malicious hypervisor or virtual machine. Full memory encryption can be used without any changes to your software for bare-metal workloads, helping repel cold-boot attacks against main memory. The security processor also helps verify the integrity of the boot process.

Cisco UCS M5 blade servers

Cisco UCS M5 blade servers are equipped with Intel Xeon Scalable processors:

- **The Cisco UCS B200 M5 Blade Server** delivers high-density computing in a blade server form factor with flexible configuration options.

- **The Cisco UCS B480 M5 Blade Server** delivers performance and versatility for a wide range of memory-intensive enterprise applications and bare-metal, virtual desktop, and virtualized workloads.

Powered by Intel Xeon Scalable processors

Cisco UCS blade, rack, and storage servers are powered by 2nd Gen Intel Xeon Scalable processors to deliver highly robust capabilities with outstanding performance, security, and agility. They offer up to 28 cores in 2- and 4-socket configurations for excellent performance and scalability. The CPUs provide excellent memory channel performance and include three Intel UltraPath Interconnect (Intel UPI) links across the sockets for scalability and intercore data flow. The processors also offer hardware-assisted security advancements that lower the performance overhead for data encryption and decryption, lowering the cost of securing data. These features further enhance the value of IT infrastructure in your enterprise.

Intel Optane DC persistent memory

Our Intel Xeon processor-powered blade and rack servers are enabled for Intel Optane™ DC persistent memory. This memory can be used as high-performance storage, or as RAM for applications. With modules supporting up to 512 GB of memory, Cisco UCS M5 servers can have up to 6 TB of persistent memory and 3 TB of regular DDR4 memory. Think of how you can load databases into persistent memory, access them at memory speeds, and use the stored image even across reboots. Or run more virtual machines per server with higher memory density.

Industry-leading bandwidth

Cisco UCS virtual interface cards have dramatically simplified the deployment of servers for specific applications. By making the number and type of I/O devices programmable on demand, we enable organizations to deploy and repurpose server I/O configurations without ever touching the hardware.

Blade server bandwidth

In blade servers, Cisco VICs provide access to up to 160 Gbps of bandwidth per server.

- In the half-width Cisco UCS B200 M5 Blade Server, you can use up to two Cisco VICs by populating both the mLOM and mezzanine card slot.
- In the full-width Cisco UCS B480 M5 Blade Server, you can install up to three Cisco VICs using the two mezzanine card slots and the mLOM slot.

This amount of I/O capacity, combined with the simplified I/O infrastructure of Cisco UCS, allows more total bandwidth per blade server compared to traditional systems. Without the complexity of stacking ports, separate Ethernet and Fibre Channel switches in each chassis, and the physical partitioning of bandwidth between I/O modalities, Cisco UCS delivers up to 640 Gbps of bandwidth for every eight blades, or up to 80 Gbps of dedicated bandwidth per blade server.

Cisco UCS M5 rack servers

Cisco UCS M5 rack servers are equipped with Intel Xeon Scalable processors:

- **The Cisco UCS C220 M5 Rack Server** is the most versatile general-purpose infrastructure and application server in the industry.

- **The Cisco UCS C240 M5 Rack Server** offers industry-leading performance as demonstrated in this document, and can support a wide range of storage, SSD, and NVMe options.

- **The Cisco UCS C240 SD M5 Rack Server** is a front-panel accessible, short-depth server

- **The Cisco UCS C480 M5 Rack Server** is designed for memory-intensive, mission-critical applications, it is our most flexible and customizable server.

- **The Cisco UCS C480 ML M5 Rack Server** propel machine learning and deep learning workloads with 8 integrated GPUs.

Rack and storage server bandwidth

In rack servers, I/O connectivity can be configured as needed using PCIe-format adapters, dedicated mLOM adapters, or system I/O controllers (SIOCs). Each adapter supports two 40-Gbps, four 25-Gbps, or two 100-Gbps unified fabric connections.

- Cisco UCS C125 Rack Server Nodes support 4th-generation Cisco VICs for 100 Gbps of connectivity per node.
- Cisco UCS C220 M5 and C240 M5 Rack Servers support the mLOM form factor as well as PCIe adapters.
- Cisco UCS C480 M5 and Cisco UCS C480 ML M5 servers support PCIe form-factor Cisco VICs.
- Cisco UCS S3260 Storage Servers support up to two SIOCs. One SIOC model integrates Cisco VIC 1300 technology for up to two 40-Gbps connections. A second model provides a PCIe slot that can support a the newest Cisco VICs and a wide range of third-party adapters, making it easy to swap in the latest technology or the interconnect technology you need.

Consistent and low latency

Consistent and low latency is important for managing performance. Traditional systems exhibit different latency characteristics depending on workload placement, which places unnecessary constraints on applications.

With a single logically centralized yet physically distributed network, SingleConnect technology requires only a single network hop and offers consistent latency. With Cisco UCS, traffic between any two rack or blade servers requires only one network hop (see path A in Figure 13). With traditional environments, intrachassis communication likewise requires only one hop (path X), but communication between chassis or between blade and rack servers requires three hops (see path Y in Figure 13).

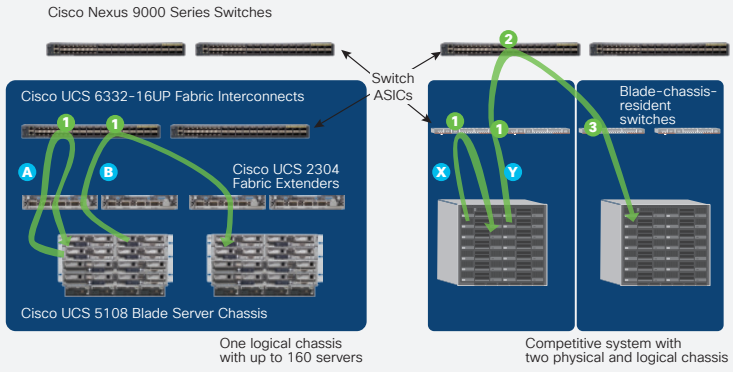
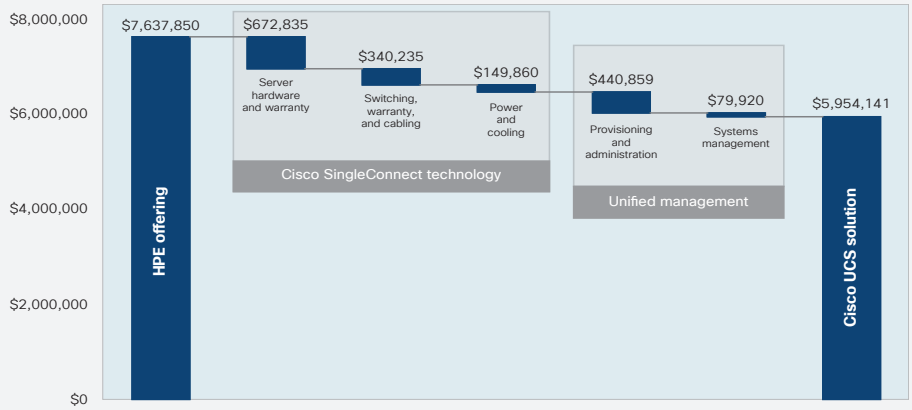


Figure 13 Every server is only one network hop away in Cisco UCS, compared to up to three hops in traditional blade server environments

Cisco UCS M5 multinode rack servers

Cisco UCS multinode rack servers are powered with the latest AMD EPYC™ processors for the highest core-per-rack density available from a commercially available multinode solution:

- **The Cisco UCS C125 M5 Rack Server Node** is a 2-socket server equipped with up to 2 AMD EPYC processors with up to 64 cores each. Up to four Cisco UCS C125s can be configured in a single **Cisco UCS C4200 Rack Server Chassis** shown at the left above. The chassis features up to 256 cores per chassis, shared power and cooling, and flexible drive bays that support up to 6 small-form-factor drives per server.



This graph compares the 3-year TCO for 80 HP ProLiant DL380 Gen10 Servers and 80 HPE Synergy SY480 Gen10 Servers with the 3-year TCO for 80 Cisco UCS C240 M5 Rack Servers and 80 Cisco UCS B200 M5 Blade Servers. Each server has two Intel Xeon Gold 6240Y CPUs and 384 GB of memory. HPE networking includes two 10 Gigabit Ethernet and two 16-Gbps Fibre Channel connections for the rack servers. The Synergy frames require a total of two HP Virtual Connect SE 40Gb F8 Modules and six Synergy 10Gb Interconnect Link Modules. The Cisco solution includes the Cisco UCS VIC 1455 quad-port 25-Gbps unified fabric PCIe adapter for Cisco rack servers and Cisco UCS VIC 1440 dual 40-Gbps unified fabric mLOM adapters for Cisco blade servers, plus corresponding switches. Pricing is as of August 7, 2020.

Figure 14 Three-year TCO for Cisco UCS is 22 percent less than for traditional environments for a combination of 80 blade and 80 rack servers

Lower infrastructure cost

We designed Cisco UCS for lower infrastructure cost per server, a choice that makes scaling fast, easy, and inexpensive in comparison to manually configured approaches. This choice is evident in the design of the Cisco UCS 5108 Blade Server Chassis.

The blade server chassis is designed to be low cost, and therefore it is little more than sheet metal, a passive midplane, sensors, and slots for power supplies, fans, and blade servers. The chassis intelligence is contained in the modular Cisco UCS fabric extenders that plug into the rear of the chassis. These devices separate the management plane from the data plane and provide access to the chassis temperature and power sensors and to each server's integrated management controller. Because the fabric extenders are logically part of the Cisco UCS fabric interconnects, the entire blade chassis is part of a single centrally managed but physically distributed system.

The lower infrastructure cost that characterizes Cisco UCS also derives from the use of low-cost, low-power-consuming Cisco fabric extenders to bring all three networks—data, storage access, and management—to each blade server chassis without the need for three pairs of redundant management, Ethernet, and Fibre Channel modules.

Cisco UCS M5 storage servers

Cisco UCS S-Series Storage Servers deliver high-density solutions with the flexibility to independently scale the ratio of CPU, network, and storage resources to best match your application needs.

- **The Cisco S3260 M5 Storage Server** supports up to 60 large-form-factor disk drives plus additional boot drives. The server can support up to two 2-socket server nodes so that you can install the right amount of CPU power for your workloads.

Our rack servers are similarly integrated into Cisco UCS with lower infrastructure cost per server. Instead of requiring up to five switch ports at the top of every rack (two Ethernet, two Fibre Channel, and one management network switch), Cisco UCS requires only two.

The example in Figure 14 demonstrates how the simplified infrastructure of Cisco UCS contributes to 22 percent lower TCO for a 160-server installation. The lower cost is due to lower cost for servers, switching, power, cooling, provisioning, administration, and systems management. In this comparison with HPE servers, it's important to note that while Cisco UCS offers unified management for all servers in the product line, HPE requires different managers for different products. Synergy requires a pair of x86 server-based management appliances called composers. But these can't manage the rack servers. So a different version of HPE OneView must be hosted on additional servers to manage the rack-form-factor servers.

Rack server deployment flexibility

Cisco UCS C-Series Rack Servers are unique in the industry because they can be integrated with Cisco UCS connectivity and management or used as standalone servers.

Cisco Integrated Management Controller

The Cisco IMC runs in the system's baseboard management controller (BMC). When a Cisco UCS C-Series Rack Server is integrated into a Cisco UCS domain, the fabric interconnects interface with the Cisco IMC to make the server part of a single unified management domain. When a server is used as a standalone server, direct access to the Cisco IMC through the servers's management port allows a range of software tools (including Cisco Intersight) to configure the server through its API. With either approach you have access to a wide range of management capabilities (Figure 15)

Integrated operation with single-wire management

Single-wire management uses the unified fabric to carry management traffic between the server and the fabric interconnects. It is enabled with through Cisco VICs, which separate management traffic from production data and storage traffic, passing it to an internal switch that connects to the Cisco IMC. The internal switch also makes the controller accessible for standalone management through the server's network management ports (Figure 16).

When single-wire management is used, the unified fabric securely isolates management traffic by connecting the fabric interconnect's management network directly to the controller using the IEEE 802.1BR standard.
To prevent any high-traffic condition on the network from impeding management traffic, Cisco UCS gives management traffic the highest priority using the IEEE 802.1Qbb Priority Flow Control standard.

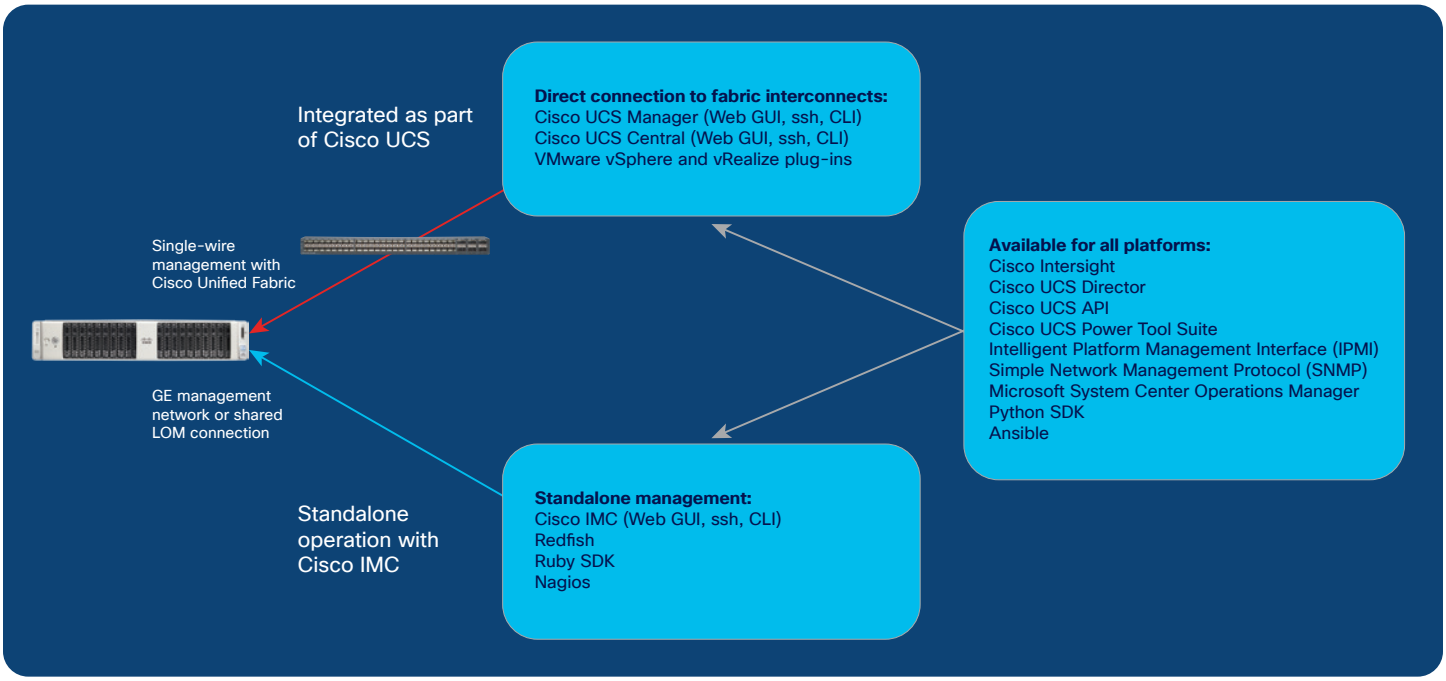


Figure 15 Whether integrated with Cisco UCS through fabric interconnects, or managed through Cisco IMC, a wide variety of tools can be used to manage the servers.

Standalone operation with the Cisco IMC

As standalone servers, Cisco UCS C-Series Rack Servers are managed through the dedicated management network interface using any of the approaches illustrated in Figure 15. Cisco Intersight connects through SSL to this interface for secure management from the cloud or an optional locally-hosted management appliance.

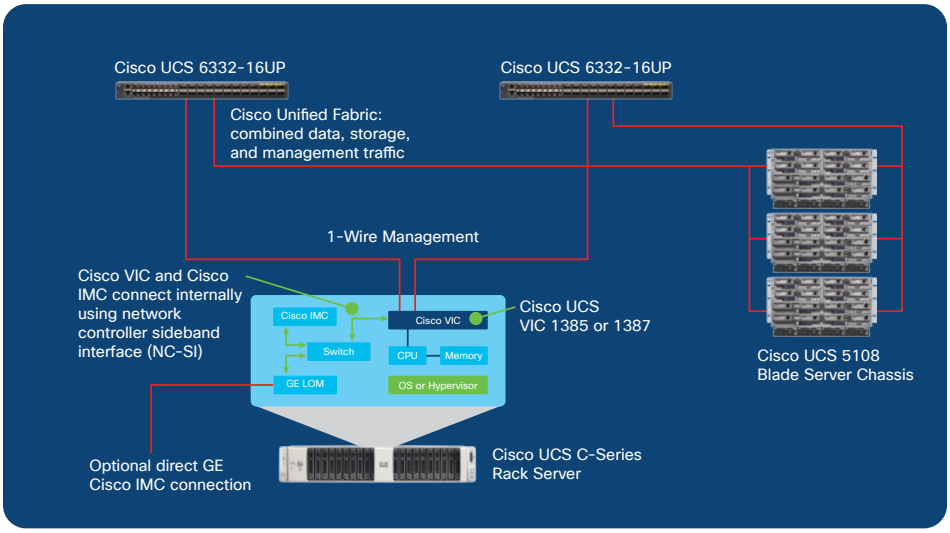


Figure 16 Rack servers integrate into Cisco UCS with single-wire management

Conclusion

The Cisco Unified Computing System is the first integrated data center platform that combines industry-standard, x86-architecture servers with networking and storage access into a single unified system. The system is intelligent infrastructure that uses integrated, model-based management to simplify and accelerate deployment of applications and services running in bare-metal, virtualized, and cloud-computing environments. Employing Cisco's innovative SingleConnect technology, the system's unified I/O infrastructure uses a unified fabric to support both network and storage I/O. The Cisco fabric extender architecture extends the fabric directly to servers and virtual machines for increased performance, security, and manageability.

Cisco UCS helps change the way that IT organizations do business, with benefits including the following:

- **Increased IT staff productivity** and business agility through just-in-time provisioning and equal support for both virtualized and bare-metal environments
- **Reduced TCO** at the platform, site, and organization levels through infrastructure consolidation
- **A unified, integrated system** that is managed, serviced, and tested as a whole
- **A comprehensive management ecosystem** that supports complete infrastructure provisioning and management that can make your Cisco UCS instance anything from a bare-metal enterprise application engine to a multlicloud containerized environment. Locally hosted tools give you a range of options and Intersight software as a service is emerging as the solution to help you manage all of your assets worldwide.
- **Scalability** with the Intersight software-as-a-service platform that can manage all of your infrastructure wherever it resides. Many of the most tedious administration tasks, such as setting up clusters in edge locations, can be automated completely so that new locations can be rolled out quickly, easily, and accurately.
- **Open industry standards** supported by a partner ecosystem of industry leaders
- **A system that scales** to meet future data center needs for computing power, memory footprint, and I/O bandwidth; it has hosted five generations of servers and three generations of network fabric in its highly simplified blade server chassis—and is poised to continue to support future generations of servers and networks.

For more information

- Cisco UCS
- Cisco UCS performance
- Cisco converged infrastructure solutions
- Cisco hyperconverged infrastructure solutions
- Cisco solutions for hybrid clouds
- Cisco Intersight
- Cisco Workload Optimization Manager
- Cisco UCS management

Ten years of innovation

When we first entered the server market in 2009, the challenge was to prove to our customers that we could out-innovate our competitors, and that we were committed to the market for the long term. Today, we are still the only vendor to offer a unified system that eliminates the tedious, manual, error-prone assembly of components into systems, providing instead a system that is self-aware and self-integrating and that brings true automation to IT operations.

Our commitment to the marketplace has been demonstrated by a rise to join the top tier of server manufacturers in just three years, with more than 55,000 customers and more than 150 world-record performance benchmarks. We have continued to innovate and demonstrate our commitment to customers and to the server market. With four generations of fabric technology supporting modular upgrades to the system's connectivity, and with even more generations of Intel Xeon and AMD EPYC processors incorporated into our products, we are demonstrating our strong support for customer investments and our readiness to take our customers into the future.

LE-49401-05 08/20