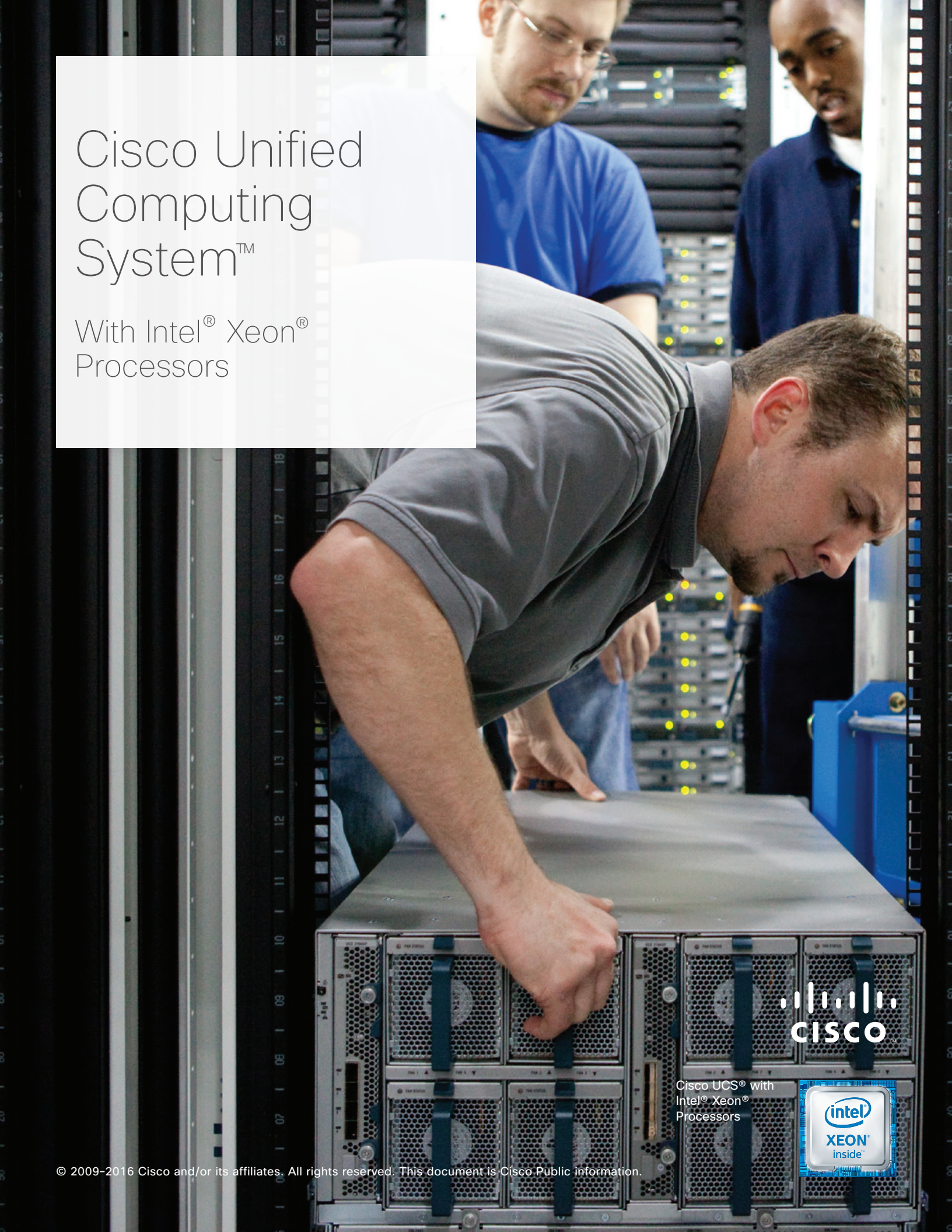
Cisco Unified Computing System™

With Intel® Xeon® Processors



Cisco UCS® with Intel® Xeon® Processors

© 2009–2016 Cisco and/or its affiliates. All rights reserved. This document is Cisco Public information.

Contents

© 2009–2016 Cisco and/or its affiliates. All rights reserved. This document is Cisco Public information.

© 2009-2016 Cisco and/or its affiliates. All rights reserved. This document is Cisco Public information.

What You Will Learn

Cisco Unified Computing System™ (Cisco UCS®) is an integrated computing infrastructure with embedded 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. Within just the seven years since the system was first announced, Cisco has joined the top tier of server vendors, attracted an enthusiastic customer base (greater than 46,500 as of December 2015), and set more than 100 world performance records—all serving as testament to Cisco's leadership and its dedication to customer-centered innovation.

The Challenge

Today's IT organizations are facing two significant challenges:

- **Increased workload diversity:** The workloads being placed on IT infrastructure is becoming increasingly diverse. Not long ago virtualization and cloud computing promised to support every business application, from Internet information services to enterprise applications such as database management systems and SAP. But then a new class of applications arose that needed massive scale-out capacity on bare-metal servers, including back-end support for mobile applications. Today, big data applications perform in-memory analytics and require local storage and massive memory capacity. And in the future, computing on the edge, or fog computing, will be more important for aggregating data and supporting the Internet of Things.
- **Increased speed of operational change:** The tempo at which IT organizations must roll out new applications and services has increased relentlessly. You need to deploy and redeploy IT assets rapidly to support agile, iterative development processes. Manual configuration has become so cumbersome that most organizations agree that zero-touch management is essential. Consistent policy and security is essential, but unless it can be supported as a transparent, integrated part of IT processes, your clients may balk and seek services from public cloud providers. And management must span clusters of local data centers, geographically distributed data centers, and systems at remote and branch locations.

These two challenges together present a set of difficult challenges for IT organizations. Workloads dictate application architecture, which affects server design, the relationship between servers and storage resources, and network architecture. Support for different types of workloads is creating technology silos within data centers, which creates additional challenges. Cloud computing environments, with disaggregated storage and dense blade server farms, are very

© 2009–2016 Cisco and/or its affiliates. All rights reserved. This document is Cisco Public information.

different from big data environments, with massive amounts of local disk storage and large memory capacities. Then factor in new applications that require massive amounts of scale-out capacity on bare-metal servers, in which applications themselves are programmed to handle individual server failures.

When each application requires its own dedicated infrastructure, you can end up with technology silos, each with its own dedicated, standalone infrastructure. Given the different architectural models these disparate systems require, IT staff struggles to deftly configure and deploy application infrastructure and faces a nearly impossible task in maintaining security and standards compliance across all environments.

What you need is an approach that can encompass all your applications in a single operating model. Cisco Unified Computing System™ (Cisco UCS®) provides that platform.

Introducing Cisco Unified Computing System

Cisco UCS is an integrated computing infrastructure with embedded management that automates and accelerates 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).

We designed Cisco UCS as a purpose-built system for the post-virtualization age. We didn't have the burden of supporting an existing traditional server product line, which gave us an immense amount of design freedom. At every level of the system's architecture, we employed innovations in silicon to achieve the best performance. We went beyond manual integration of components and chose a model in which policies guide configuration: a model that scales and allows routine tasks to be completely automated regardless of server type or application architecture. This philosophy allows Cisco UCS to deliver rapid scalability and deployment, increasing IT productivity and business agility.

Initially focused on providing a platform for server virtualization, Cisco UCS is designed fundamentally to evolve to support a wide range of application and server architectures. Today, Cisco UCS can support all your applications in your data centers around the world and in your smaller branch offices and remote offices. It supports a heterogeneous collection of servers–including blade and rack servers. It supports applications that scale up and servers that scale out–all with a single policy-based management and connectivity model. Because it can support all your applications under the same management model, it can help you overcome the challenges of workload diversity and support a rapid operational tempo.

Cisco UCS helps you standardize IT processes, allowing you to work at a higher level, defining policies and allowing automation to handle routine administrative tasks. It is the first self-aware, self-integrating, unified system that automates system configuration in a reproducible, scalable manner. As it recognizes new components added to the system, it places new server resources in pools that simplify the allocation and sharing of resources, allowing you to more rapidly move new servers into production and increase overall resource utilization. It incorporates

© 2009–2016 Cisco and/or its affiliates. All rights reserved. This document is Cisco Public information.

a simplified architecture that eliminates the need to have different network architecture for each application so that you now can scale without having to reevaluate whether your network resources are sufficient. It makes physical servers and virtual machines equivalent so that both can be managed with equivalent levels of visibility and control.

The system is intelligent infrastructure that is configured through integrated, model-based management. Server identity, personality, and I/O connectivity is abstracted so that the system accomplishes for physical environments what hypervisors accomplish for virtualized ones. Cisco UCS enables you to run any workload on any resource with dynamic provisioning, making it the ultimate platform for unifying your data center.

A hierarchy of Cisco® management tools extends this concept across all your data centers and remote locations. Cisco UCS Central Software uses the same model-based management to support up to 6000 servers regardless of location as if they were in a single Cisco UCS domain. Cisco UCS Director provides comprehensive infrastructure automation and orchestration, managing all the resources in Cisco Integrated Infrastructure solutions, including Cisco UCS, storage, and higher-level switching infrastructure. Cisco UCS Director automates your workflows, and it enables IT-as-a-service (ITaaS) offerings by providing a self-service portal through which administrators and clients can order infrastructure instances that are configured on demand. The fundamental, standards-based XML API that we expose to the outside world has been accepted and incorporated into third-party management tools from a large ecosystem of third-party vendors.

As the pressures to support increasing architectural diversity with an ever-increasing operational tempo continue unabated, Cisco UCS is the single platform that can adapt to support you at every step.

Cisco Unified Computing System Concepts

Cisco UCS is the first truly unified data center platform that combines industry-standard, x86-architecture servers with networking and storage access into a single system. The system is intelligent infrastructure that is automatically configured through integrated, model-based management to simplify and accelerate deployment of all your applications.

Powering Servers with Cisco Innovations

The system's x86-architecture rack and blade servers are powered exclusively by Intel® Xeon® processors and enhanced with Cisco innovations. These innovations include the capability to abstract and automatically configure the server state, built-in virtual interface cards (VICs), and leading memory capacity. Cisco's enterprise-class servers deliver world-record performance to power mission-critical workloads. Cisco's cloud-scale servers support a lower-performance, bare-metal deployment model in which massive numbers of servers support many instances of a single application. All our servers, in combination with simplified, unified architecture, increase IT productivity and provide superior price-to-performance ratios for lower total cost of ownership (TCO).

© 2009–2016 Cisco and/or its affiliates. All rights reserved. This document is Cisco Public information.

Simplifying Three Networks into One

Building on Cisco's strength in enterprise networking, Cisco UCS is integrated with a standards-based, high-bandwidth, low-latency, virtualization-aware 10-Gbps unified fabric, with a new generation of Cisco UCS fabric enabling an update to 40 Gbps. Cisco SingleConnect technology is implemented with an end-to-end system I/O architecture that uses Cisco Unified Fabric and Cisco Fabric Extender Technology (FEX Technology) to connect every Cisco UCS server within a single network and a single network layer. The system is wired once to support the desired bandwidth, and it carries all Internet protocol, storage, management, and virtual machine traffic with security isolation, visibility, and control equivalent to that of physical networks. The network fabric exceeds the bandwidth demands of today's multicore processors and eliminates the cost of separate networks for each type of traffic while increasing workload agility, reliability, and performance. As you expect from Cisco, the Cisco UCS I/O architecture is based on open standards and is reliable, available, and secure.

Bringing Automation to Information Technology

Cisco UCS is intelligent infrastructure that is self-aware and self-integrating. The system is built from the beginning so that every aspect of server identity, personality, and connectivity is abstracted and can be applied through software. With Cisco UCS, servers are configured automatically, eliminating the manual, time-consuming, error-prone assembly of components into systems. With Cisco VICs, even the number and type of I/O interfaces are programmed dynamically, making every server ready to power any workload at any time. With Cisco composable infrastructure, even storage devices are abstracted and included as part of dynamic server definition. This abstraction allows Cisco UCS to do what no hypervisor can do: adapt every aspect of the hardware configuration to propel applications with the exact balance of resources that make them perform best.

Using Policies to Align Configurations with Workloads

With integrated, model-based management, administrators manipulate a model of a desired system configuration and associate a model's service profile with hardware resources, and the system configures itself to match the model. This automation accelerates provisioning and workload migration with accurate and rapid scalability. For the first time, IT organizations have an automated, policy-based mechanism for aligning server configuration with workload. The process of establishing and maintaining configuration management databases (CMDBs) can be automated through the system's XML API, facilitating approaches based on Information Technology Infrastructure Library (ITIL) concepts. The result is increased IT staff productivity, improved compliance, and reduced risk of failures due to inconsistent configurations.

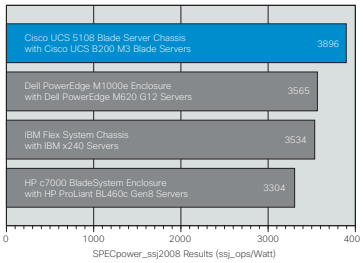
Directly Connecting Servers and Virtual Machines to the Network

With SingleConnect technology, the unified fabric requires fewer components and networks, and Cisco fabric extenders reduce the number of network layers by directly connecting physical and virtual servers to the system's fabric interconnects. This combination eliminates blade server, top-of-rack, and hypervisor-based switches by logically connecting fabric interconnect ports directly to individual servers and virtual machines. Virtual networks are now managed exactly the same way as physical networks, but have massive scalability. This capability represents a radical simplification compared to traditional systems, reducing capital and operating

© 2009–2016 Cisco and/or its affiliates. All rights reserved. This document is Cisco Public information.

Efficiency Quantified

True power and cooling savings come from new generations of silicon, architectural changes, and increased utilization. Cisco UCS was designed with these factors in mind to increase energy efficiency. Testing results show that Cisco UCS has better performance and more power efficiency than HP, IBM, and Dell solutions, with power savings that increase as the solution scales. For an 8-blade configuration, Cisco UCS demonstrated 10 to 18.7 percent greater performance-to-power ratios than equivalent servers–and the power savings can exceed 2 kW for a 160-blade solution.



For details, see Cisco UCS Power Efficiency Beats HP, IBM, and Dell Solutions, March 2014.

SPEC Fair Use Rule disclosure: At 100 percent target load, the Cisco UCS 5018 blade chassis with eight Cisco UCS B200 M3 Blade Servers installed achieved 11,252,295 ssj_ops using 2311W, HP BladeSystem c7000 enclosure with eight HP BL460c Gen8 servers installed achieved 11,349,890 ssj_ops using 2661W, IBM Flex System Enterprise Chassis with eight IBM x240 computing nodes installed achieved 11,173,554 ssj_ops using 2456W, and Dell PowerEdge M1000e blade enclosure with eight Dell M620 G12 servers installed achieved 11,269,813 ssj_ops using 2605W. SPEC and SPECpower are registered trademarks of Standard Performance Evaluation Corporation.

costs while increasing business agility, simplifying and accelerating deployment, and improving performance.

Scaling Without Complexity

The combination of unified fabric and Cisco fabric extenders in SingleConnect technology creates a system with one network layer and one point of management and connectivity for the entire system. As a result, Cisco UCS scales more gracefully, in smaller increments, and at lower cost than other systems. With low-cost and low-power-consuming fabric extenders supporting the system's growth, the infrastructure cost per server is dramatically lower than for traditional systems. Because Cisco fabric extenders act as distributed line cards and are implicitly managed by the system's fabric interconnects, the system can grow without the need to add a single management point. Fewer components and management points contribute to easier and more rapid scaling with lower capital and operating costs.

Increasing Energy Efficiency

Every aspect of Cisco UCS is designed for energy efficiency.

- Blade and rack servers are designed for minimal airflow obstruction, reducing the number of watts (W) used by cooling fans. In particular, the system's blade server chassis midplane is 63 percent open for easy front-to-back cooling. This design supports thermal conditions in which Intel Turbo Boost Technology can apply more processing power to handle workload peaks, helping Cisco UCS establish performance records for functions as basic as CPU performance.
- Power supplies are sourced to maintain high efficiency even at moderate power utilization levels.
- The elimination of blade chassis switching, top-of-rack switching, and blade-chassis management modules reduces the number of devices that need to be powered, reducing overall power consumption.
- The elimination of multiple parallel networks for IP, storage, and management traffic reduces the number of network interface cards (NICs) and host bus adapters (HBAs) and corresponding upstream ports that need to be powered and cooled.
- Large memory capacities help support large virtual machine footprints in 2-socket servers, eliminating the need to use 4-socket servers just to have a larger memory capacity and thereby also eliminating the energy cost. Cisco originally led the industry with Cisco Extended Memory Technology, and today offers one of the largest memory capacities available in a half-width blade server.
- Intel Xeon processors used in Cisco UCS servers adjust their energy consumption to the workload by scaling down energy use in small increments as workload conditions permit, and scaling up the processor clock rate when workload conditions demand and thermal conditions permit.

Investing in the Future

- Cisco UCS gives data centers room to scale while anticipating future technology developments, helping increase return on investment (ROI) today while protecting

© 2009–2016 Cisco and/or its affiliates. All rights reserved. This document is Cisco Public information.

that investment into the future. We understand the I/O and network bandwidth demands that high-performance, multicore processors can make and we have designed Cisco UCS to accommodate multiple generations of processor and network technologies without the need for replacement of entire systems. Proof of the investment protection offered by Cisco UCS is seen in the track record that the system has already established:

- The Cisco UCS blade chassis has already accommodated four generations of Intel Xeon processors and has the bandwidth capacity and the power budget to support future generations.
- The capacity of the system's fabric interconnects has grown from of 20 ports of 10-Gbps connectivity in the initial offering to 40 ports of 40-Gbps connectivity: an eightfold increase in overall bandwidth. The fabric interconnects have also increased in flexibility through universal ports that can connect to both native Fibre Channel networks and Fibre Channel over Ethernet (FCoE) devices.
- Three generations of fabric extenders have increased the blade chassis I/O capacity from 80 to 160 to 320 Gbps to better match bandwidth with processing capacity.
- Data center I/O connectivity can be managed efficiently at a single point, without the need to manage connectivity at each server, radically reducing the complexity involved in scaling up to meet growing I/O bandwidth needs.
- The system is modular, so servers and network components can be upgraded independently without forcing changes in other components. Earlier generations of servers interoperate with the newest network fabric, and newer generations of servers work with the first-generation fabric.
- Cisco UCS started with support for blade server form factors. It also supports rack servers with the same single-wire management as blade servers.

Beyond Efficiency: Making IT More Productive

Cisco UCS helps organizations go beyond efficiency: it helps them become more effective through technologies that engender simplicity rather than complexity. The result is flexible, agile, high-performance, self-integrating information technology; reduced staff costs with increased uptime through automation; and more rapid ROI.

Cisco Unified Computing System 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 various servers to them either directly or indirectly.

Cisco UCS Manager

Cisco UCS Manager integrates blade and rack servers into a single self-aware, self-integrating, unified system. It quickly and accurately configures computing, network, storage, and storage-access resources to increase compliance and reduce the chance of errors that can cause downtime. It uses a role- and policy-based approach that helps organizations more easily align policies and configurations with workloads, and it automates system configuration through a "create once, deploy many" approach. Cisco UCS Manager acts as a single point of management and

© 2009-2016 Cisco and/or its affiliates. All rights reserved. This document is Cisco Public information.

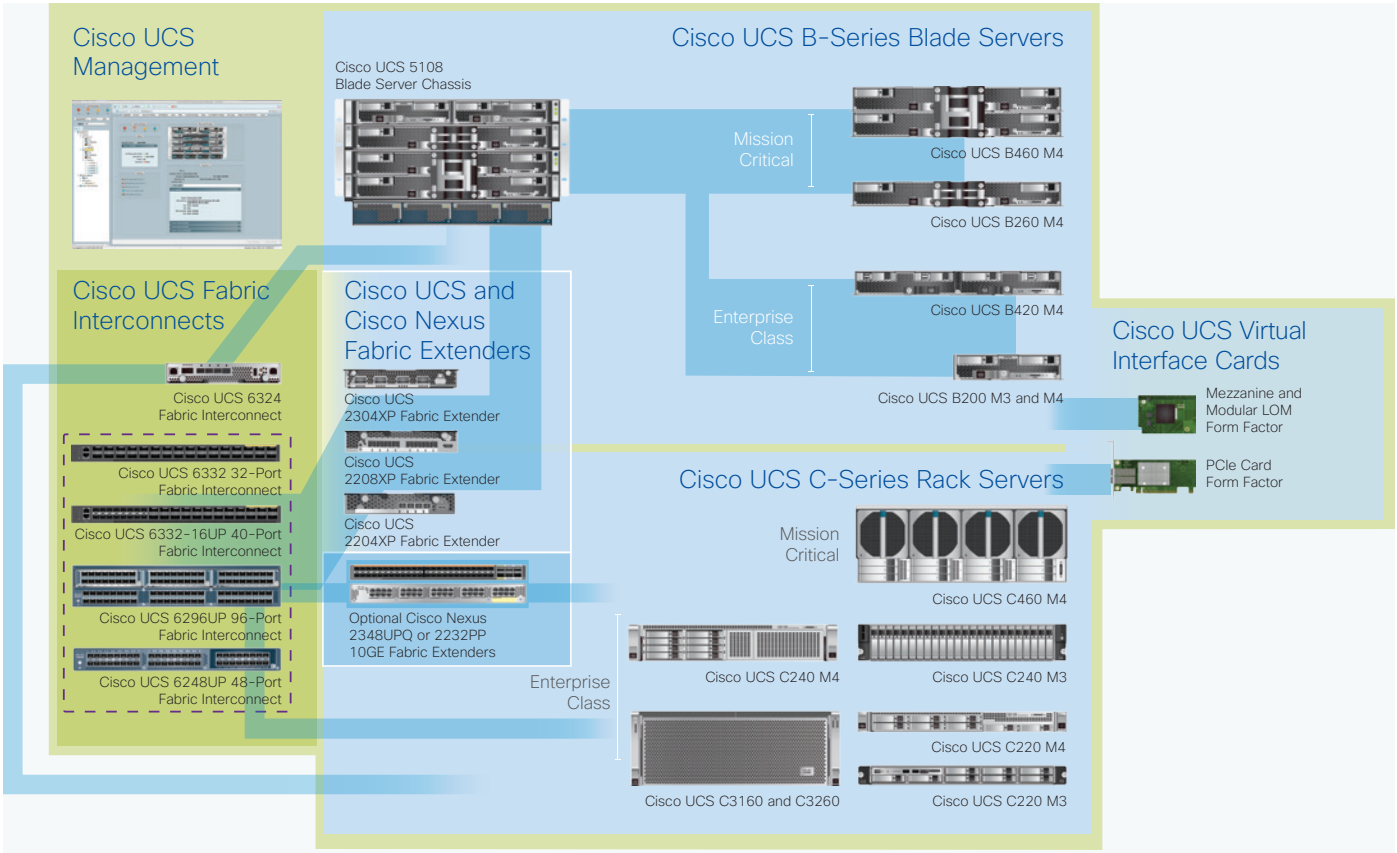


Figure 1 Cisco UCS Component Hierarchy

monitoring for the entire system. It is embedded software that runs on the system's fabric interconnects, typically in a redundant, high-availability configuration. It can be accessed through an intuitive GUI, command-line interface (CLI), or XML API. Cisco UCS Central Software and Cisco UCS Director access the XML API to provide higher-level management functions. More than 22 high-level management tools integrate with Cisco UCS through the Cisco UCS Manager's XML API.

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** provide a single point of connectivity and management for the entire system. Typically deployed as an active-active pair, the system's fabric interconnects integrate all components into a single, highly available management domain controlled by Cisco UCS Manager. The fabric interconnects manage all I/O efficiently and securely at a single point, resulting

© 2009-2016 Cisco and/or its affiliates. All rights reserved. This document is Cisco Public information.

in deterministic I/O latency regardless of a server or virtual machine's topological location in the system. Cisco UCS 6200 Series Fabric Interconnects support line-rate, lossless 10 Gigabit Ethernet and FCoE connectivity. The Cisco UCS 6200 Series can be used to create Cisco UCS domains containing blade or rack servers. Cisco UCS 6300 Series Fabric Interconnects support line-rate, lossless 40 Gigabit Ethernet and FCoE connectivity. 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** 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 complexity or new management points. Cisco fabric extenders eliminate the need for top-of-rack switches and blade-server-resident Ethernet and Fibre Channel switches or management modules, dramatically reducing the infrastructure cost per server. Rack 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**, or VICs, extend the network fabric directly to both servers and virtual machines 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.

Cisco UCS Chassis and Racks

Cisco UCS blade chassis and racks act as containers for the system components:

- 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, or up to two full-width double-height 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 up to 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 R-Series Racks** are standard 19-inch racks that are optimized to house both Cisco UCS blade chassis and rack servers in the same physical chassis, providing the flexibility to enhance a system using the server form factor most appropriate for the task.

Cisco UCS Servers

Delivering performance, versatility, and density in devices designed without compromise, Cisco UCS servers can power every workload, including workloads for:

© 2009–2016 Cisco and/or its affiliates. All rights reserved. This document is Cisco Public information.

- Agile development environments requiring bare-metal servers
- Big data
- Content delivery
- Cloud computing environments delivering virtual machines and bare-metal servers as a service
- Database management systems
- High-frequency trading
- High-performance computing
- Gaming applications
- Internet infrastructure applications
- Mission-critical enterprise applications
- Mobile application back-end services
- Virtualized environments

Powered by the latest Intel Xeon processors, Cisco UCS servers form the core of a flexible and efficient data center that meets diverse business needs with a balanced combination of performance, built-in capabilities, and cost effectiveness. Cisco UCS is platform neutral and can incorporate rack and blade servers in a single unified system with up to 160 blade and rack servers (Figure 2).

- **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

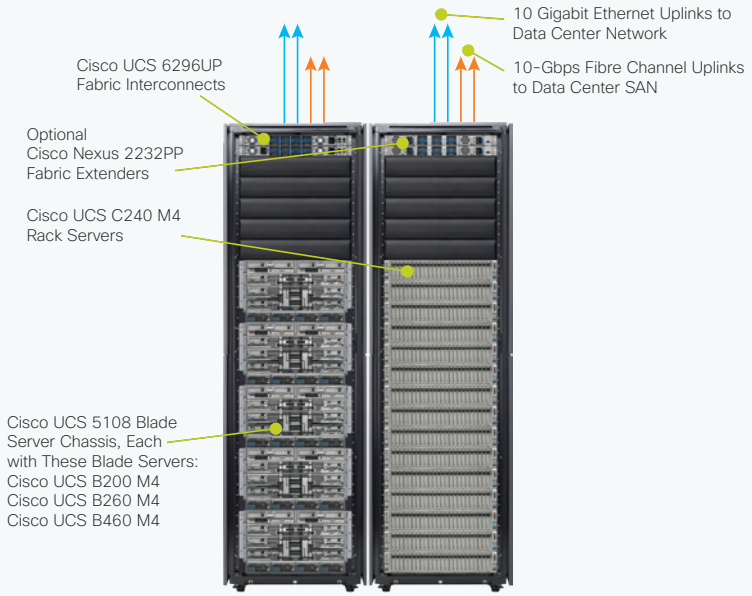


Figure 2 Cisco UCS Supports Blade and Rack Servers in a Domain

© 2009–2016 Cisco and/or its affiliates. All rights reserved. This document is Cisco Public information.

in three form factors (half-width, full-width, and full-width double-height) with two or four Intel Xeon processors. We continue to lead the industry with high memory capacity to better support virtualized environments, and our M3 and later blade servers can be ordered with built-in modular LAN on motherboard (mLOM) Cisco VICs to increase I/O flexibility and accelerate deployment.

- **Cisco UCS C-Series Rack Servers** provide a rack-server entry point to Cisco UCS. With world-record-setting performance for 2- and 4-socket servers, Cisco rack servers can integrate into Cisco UCS through a single set of cables. Cisco UCS rack servers provide a wide range of I/O, memory, internal disk, and solid-state disk (SSD) drive capacity, enabling you to easily match servers to workloads.

Cisco UCS Physical Connectivity Options

You can create a variety of Cisco UCS configurations based on Cisco UCS fabric extenders. Figure 3 shows how Cisco UCS 6200 Series Fabric Interconnects can connect directly to servers or indirectly through fabric extenders, depending on the server form factor and deployment choices.

Using Cisco UCS 6200 Series Fabric Interconnects, you can create topologies such as those illustrated in Figure 4, and even more bandwidth is available with Cisco UCS 6300 Series Fabric Interconnects.

- **Blade server chassis** can be connected to the fabric interconnects through a pair of Cisco UCS 2200 Series Fabric Extenders, which can support up to eight 10-Gbps unified fabric uplinks per fabric extender. With Cisco UCS 6300 Series Fabric Interconnects, Cisco UCS 2300 Series Fabric Extenders can support up to four 40-Gbps unified fabric uplinks per fabric extender.
- **Rack servers** can be connected directly to the fabric interconnects or indirectly through Cisco Nexus® 2232PP 10GE Fabric Extenders (for 10 Gigabit Ethernet using Cisco UCS 6200 Series Fabric Interconnects) and through Cisco Nexus 2348UPC 10GE Fabric Extenders (for 40 Gigabit Ethernet using Cisco UCS 6300 Series Fabric Interconnects) to achieve greater scale.
- **Cisco UCS Mini solutions** can be created by using Cisco UCS 6234 Fabric Interconnects in the blade server chassis instead of fabric extenders. This creates a standalone Cisco UCS instance that supports blade servers, rack servers, and external storage systems (Figure 5).

Unified I/O Architecture

Cisco UCS is organized across a low-latency, line-rate, lossless 10- or 40-Gbps unified fabric that carries all I/O from servers and virtual machines to the system's fabric interconnects. Using SingleConnect technology, the system is typically 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, with up to 320 Gbps from each blade chassis and with bandwidth shared between servers. All network features and capabilities are controlled through software settings. As a result, bandwidth is shared between all I/O modalities, so that bursts in one class of traffic can temporarily borrow bandwidth from other functions to achieve the best performance.

© 2009-2016 Cisco and/or its affiliates. All rights reserved. This document is Cisco Public information.

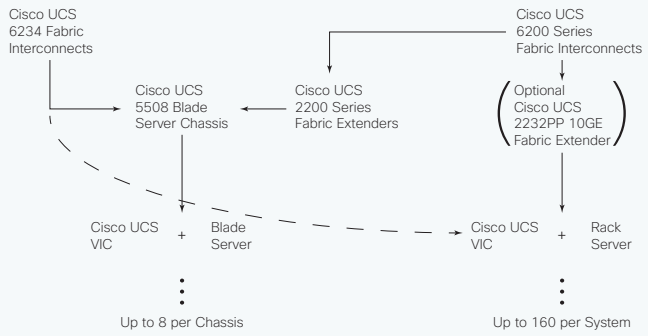


Figure 3 A Cisco UCS Domain Is Created by Connecting Servers Directly or Indirectly to Fabric Interconnects

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 as the 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 organizations move more quickly than their competitors 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 Cisco Unified Fabric uses a wire-once model in which IT departments 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.

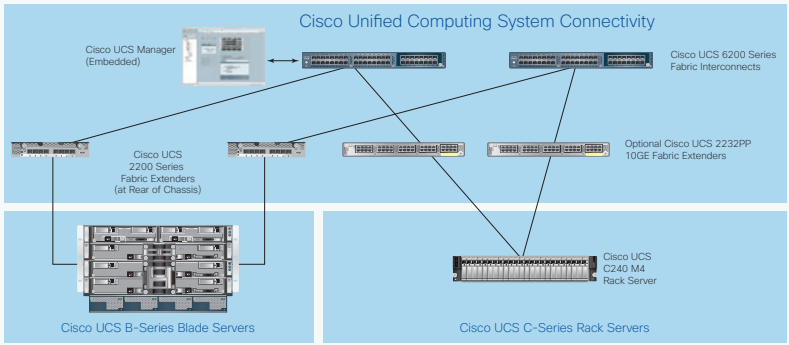


Figure 4 Cisco UCS Connectivity Options with Cisco UCS 6200 Series Fabric Interconnects

© 2009–2016 Cisco and/or its affiliates. All rights reserved. This document is Cisco Public information.

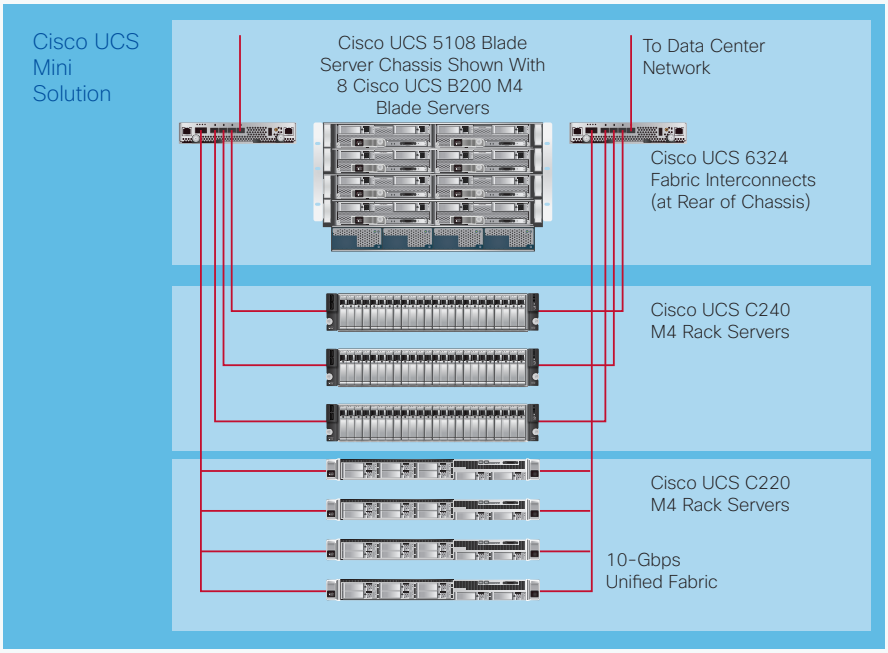


Figure 5 Cisco UCS Mini Solution Illustrated with Cisco UCS C-Series Rack Servers

- **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 HBAs and separate 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.

Condensing Multiple Parallel Networks into a Single Unified Network

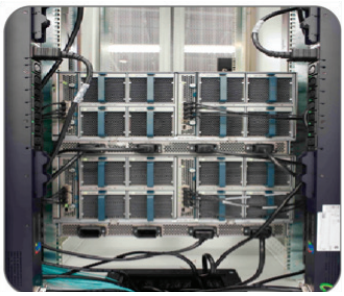
Traditional environments must implement multiple parallel networks to support management, IP networking, and storage access (Figure 6). The result is a proliferation of NICs and HBAs in each server, and upstream switch ports. Each separate network must be sized to handle workload bursts without the capability 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.

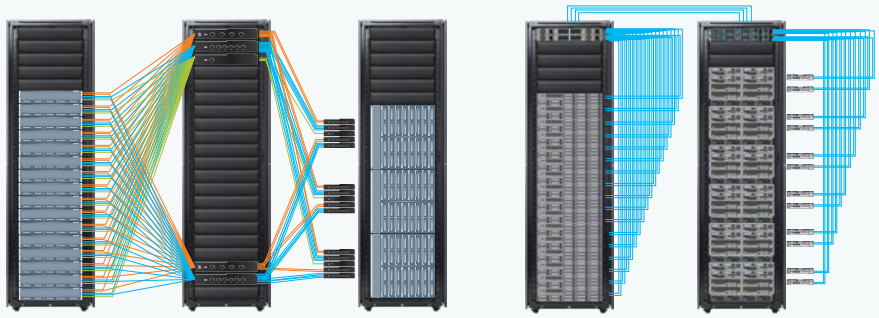
In contrast, SingleConnect technology allows a single unified network to bring LAN, SAN, and management connectivity to each rack and blade server in Cisco UCS using Cisco Unified Fabric. All three networks are carried over a single set of cables that securely carries production data (Ethernet) traffic, Fibre Channel traffic through FCoE, and management traffic. Every server—rack or blade—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.

© 2009-2016 Cisco and/or its affiliates. All rights reserved. This document is Cisco Public information.

Traditional Blade Server Environment

Simplified Cisco UCS Cabling





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 6 Cisco Unified Fabric Eliminates Multiple Parallel Networks, Simplifying Both Blade and Rack Server Environments

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.

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.

Eliminating Multiple Network Layers

In traditional 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.

© 2009–2016 Cisco and/or its affiliates. All rights reserved. This document is Cisco Public information.

A typical virtualized environment requires three network hops to reach the access-layer switch. As shown in Figure 7, one layer exists 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.

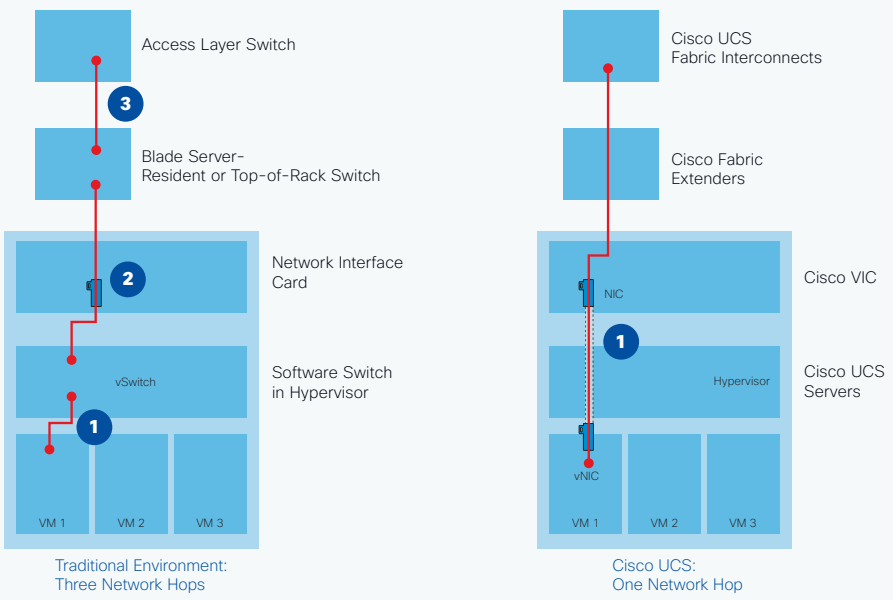


Figure 7 Cisco Fabric Extender Technology Eliminates Network Layers

SingleConnect technology brings the unified fabric to every blade chassis and server rack in Cisco UCS. This approach condenses layers of networking while eliminating hypervisor switches and blade-chassis-resident switches, replacing them with a single point of management and connectivity. Cisco Fabric Extender Technology, a prestandard implementation of the IEEE 802.1BR Bridge Port Extension standard, condenses three network layers into one 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 regardless of virtual machine location (see Figure 7).

Cisco Fabric Extender Technology brings the network fabric to blade chassis, to the top of the rack, and even to individual virtual machines, 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.

- In rack server environments, Cisco Nexus fabric extenders bring the system's unified fabric to the top of every rack, making each rack self-contained and easily moved within the data center.

© 2009-2016 Cisco and/or its affiliates. All rights reserved. This document is Cisco Public information.

- In blade server environments, Cisco UCS fabric extenders bring the system's unified fabric to every blade server chassis, with the current generation of Cisco UCS 2300 Series Fabric Extenders supporting up to 320 Gbps of bandwidth for an eight-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.

Virtualizing I/O Interfaces

The unified fabric virtualizes I/O so that rather than requiring each server to be equipped with a set of physical I/O interfaces to 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 in the system using FCoE. At the fabric interconnects, storage-access traffic can transition to physical Fibre Channel networks through a Fibre Channel transceiver installed in 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 system's fabric interconnects with the same level of secure isolation as if it were carried over separate physical cables. These virtual network links originate within the server's converged network adapters and terminate at virtual ports within the system's fabric interconnects, as illustrated earlier in Figure 7.

These virtual links are managed exactly as if they were physical networks. The only characteristic that distinguishes physical from virtual networks within the fabric interconnects is the naming of the ports. 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 for both physical servers and virtual machines.

Virtualized I/O with Converged Network Adapters

Virtual links can originate from converged network adapters that typically host a dual 10 Gigabit Ethernet NIC and a dual HBA from either Emulex or Q-Logic, along with circuitry to multiplex the four streams of traffic onto two 10-Gbps unified fabric links. Cisco innovations first brought this concept to market, with the first generation of converged network adapters (CNAs) supported by Cisco silicon that multiplexed multiple traffic flows onto the unified fabric (Figure 8).

With servers connected to Cisco UCS through CNAs, the traffic from each of the interface's four devices is passed over four virtual links that terminate at virtual ports within the fabric interconnects.

© 2009–2016 Cisco and/or its affiliates. All rights reserved. This document is Cisco Public information.

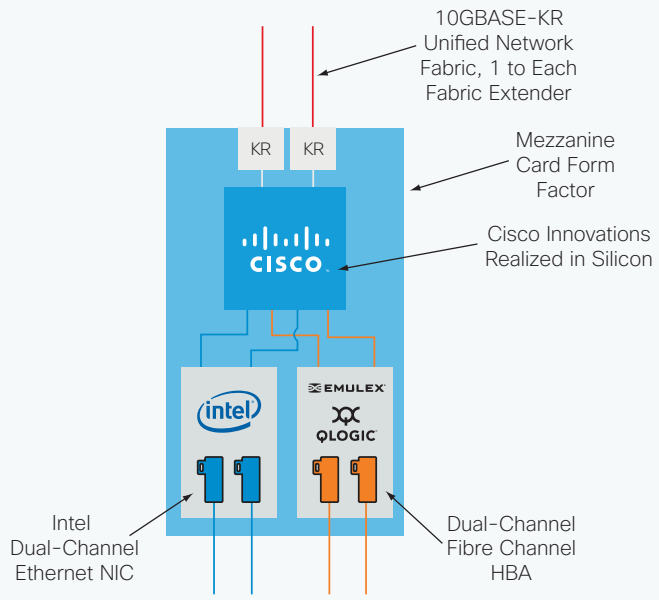


Figure 8 First-Generation CNAs Were Supported with Cisco Silicon

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.

Virtualized I/O with Cisco Virtual Interface Cards

Cisco VICs are PCIe-compliant interfaces that support up to 256 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 80 Gbps of connectivity and are available in multiple form factors:

- **mLOM:** These Cisco VICs can be ordered preinstalled in Cisco UCS M3 and M4 blade servers, occupying a dedicated slot for the device (Figure 9). If more than 40 Gbps of bandwidth is needed, a port expander card can be installed in the server's mezzanine slot to give the card access to an additional 40 Gbps of bandwidth. When Cisco UCS 2304XP Fabric Extenders are installed in the blade chassis, the Cisco UCS VIC 1340 detects the availability of 40 Gigabit Ethernet and disables the port channel for greater efficiency.
- **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, for a total of up to 320 Gbps of aggregate bandwidth for double-width, double-height servers such as the Cisco UCS C460 M4 Rack Server.
- **PCIe:** PCIe form-factor cards can be installed in Cisco rack servers. Cisco VICs are required when you integrate these servers into Cisco UCS because they have the circuitry to pass 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 32).

© 2009–2016 Cisco and/or its affiliates. All rights reserved. This document is Cisco Public information.

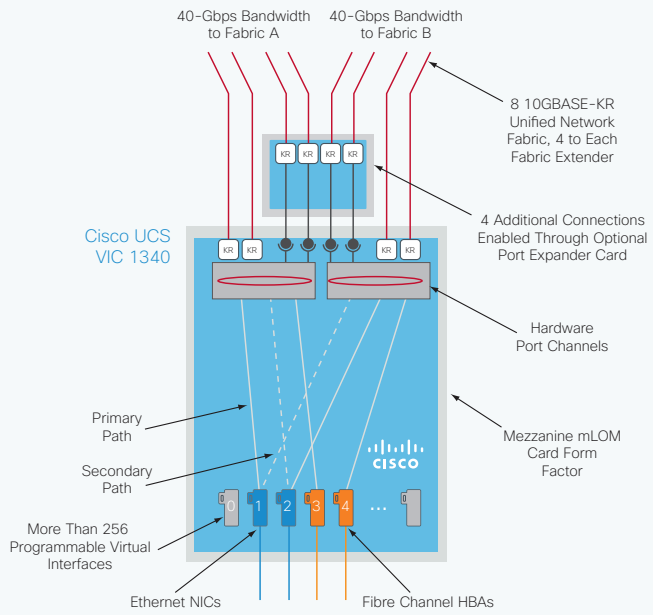


Figure 9 Cisco UCS VIC 1340 with Port Expander Card

Cisco UCS VIC Static Interfaces

Static interfaces can be used to support operating system requirements and best practices. Figure 10 shows static 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. Static devices also can be used to connect traditional software switches to the network. In the sense that the Cisco VIC brings the network directly to these different devices, the VIC acts as an adapter fabric extender.

Cisco UCS VIC Dynamic Interfaces

Dynamic interfaces connect the network directly to virtual machines, bringing exceptional visibility and control to virtualized environments, increasing network performance by up to 38 percent by bypassing hypervisor switches, and freeing the host CPU to deliver better application performance (See VM 1 in Figure 10).

This Cisco Data Center Virtual Machine Fabric Extender (VM-FEX) technology is a significant innovation that benefits virtualized environments by allowing virtual machines to connect to the network exactly like physical servers do, bringing together the scalability of virtual networks and the manageability of physical networks. Within Cisco UCS, physical and virtual networks are managed in the same way, with physical and virtual network links terminated by physical and virtual ports within the system's fabric interconnects. Now, for example, a rogue virtual machine can be identified by the traffic on its port, and the port can easily be disabled, using the same process as a network administrator would use to handle such an incident on a physical server.

The impact of this innovation on security is equally profound. Cisco UCS Manager coordinates with the major hypervisors to maintain a virtual machine's network

© 2009-2016 Cisco and/or its affiliates. All rights reserved. This document is Cisco Public information.

profile regardless of the virtual machine's location. Security does not have to be compromised to allow virtual machines to migrate freely from server to server. Now when virtual machines migrate, their network links migrate with them (Figure 11). In the fabric interconnects, the virtual port associated with a specific virtual link simply changes the physical port to which it is associated.

Data Center VM-FEX helps organizations maintain existing administrator roles across bare-metal applications and virtualized and cloud-computing environments. Because the network remains always in the domain of network administrators, this technology eliminates the overlap of server and network administrator roles that often occurs in virtualized environments.

Integrating 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

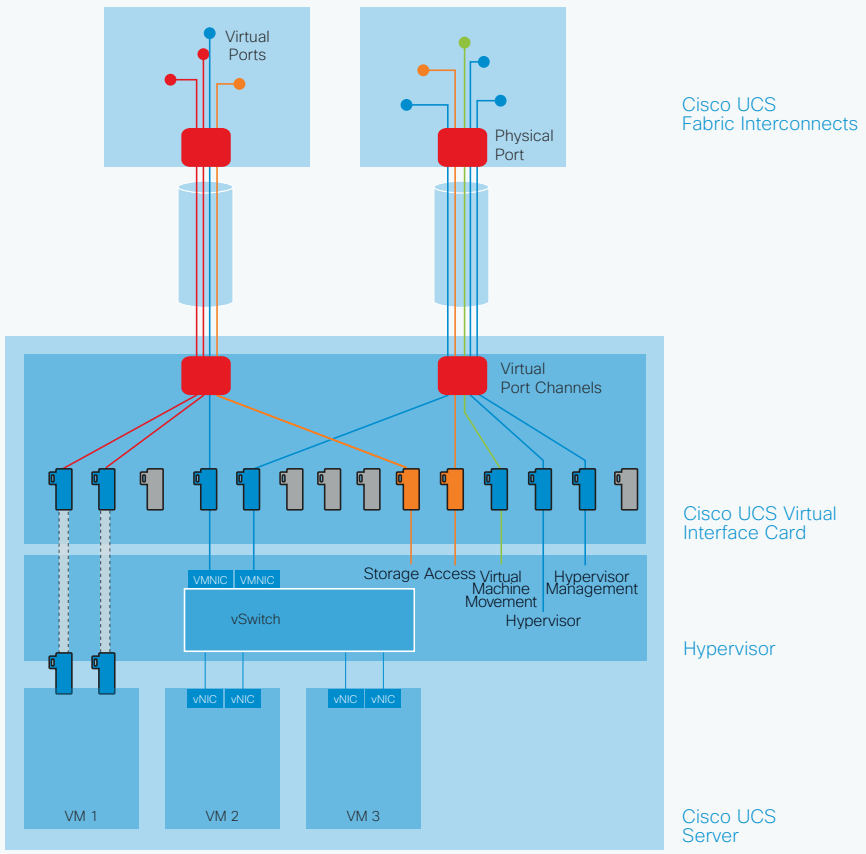


Figure 10 Cisco UCS VICs Support Static and Dynamic PCIe Interfaces

© 2009-2016 Cisco and/or its affiliates. All rights reserved. This document is Cisco Public information.

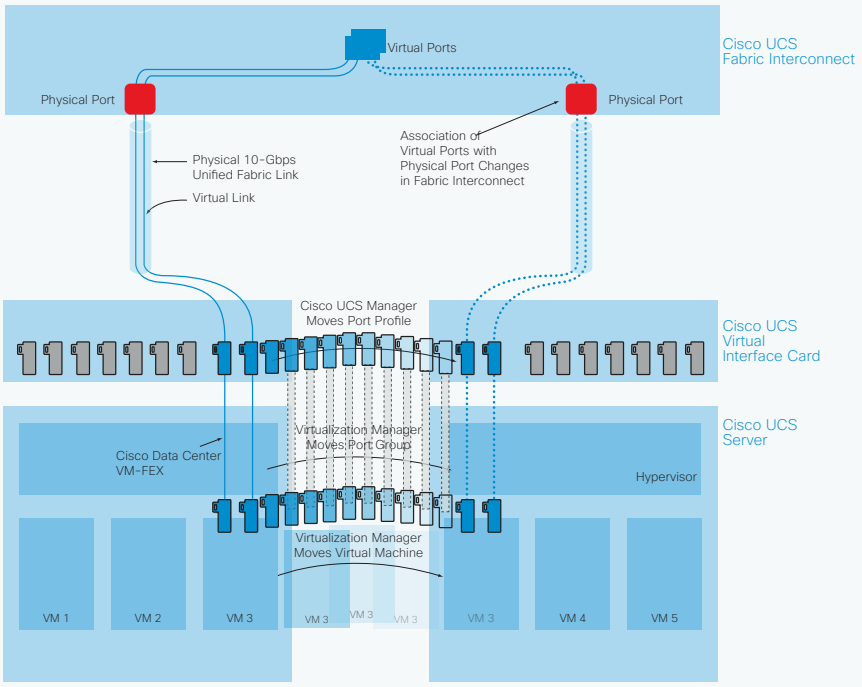


Figure 11 Cisco Data Center VM-FEX Directly Connects Virtual Machines to the Network, Maintaining the Connection as the Virtual Machines Move Between Servers

modes. This approach 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.

With the fabric interconnects in end-host mode, and with virtual network interfaces connecting both servers and virtual machines to the fabric extenders with a single network hop, Cisco Application Centric Infrastructure (Cisco ACI™) leaf switches see their connections to fabric interconnects as if they were connected to a large number of servers. 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.

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

© 2009-2016 Cisco and/or its affiliates. All rights reserved. This document is Cisco Public information.

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.

Cisco UCS Management

Ever since its introduction in 2009, Cisco UCS has remained unique in the marketplace with a combination of intelligent infrastructure and integrated, model-based management. Cisco UCS management takes the form of Cisco UCS Manager, which is embedded in every Cisco UCS instance, and Cisco UCS Central Software, which uses the exact same concepts to manage multiple Cisco UCS instances with global control over policies and resources.

- **Intelligent infrastructure:** Cisco UCS is the only system designed from the beginning so that every aspect of system personality, configuration, and connectivity is abstracted from the hardware and can be configured through software. With everything from firmware revisions to network profiles abstracted into more than 100 configuration variables that fully specify each server, the system is in this sense stateless. 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, increasing flexibility, utilization, and business agility. The system is self-aware and self-integrating. 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.
- **Integrated, model-based management:** Whereas the system's intelligent infrastructure provides the "knobs" that allow every aspect of the system's configuration to be controlled through software, Cisco UCS Manager is the integrated, model-based manager that turns those knobs to automate the setting of server identity, configuration, and connectivity. Cisco UCS Manager allows administrators to create a model of a desired server configuration and then configure the server simply by associating the 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 administrators), helps administrators be more effective in their jobs, and reduces overall cost. After configuration, Cisco UCS Manager aggregates element monitoring so that every aspect of the system can be monitored from a single interface.

Cisco UCS Management Concepts

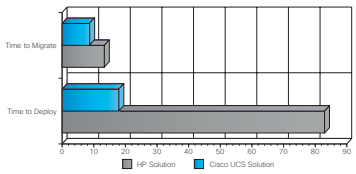
Cisco UCS Manager is software that runs in the system's fabric interconnects. It is usually configured in a high-availability arrangement with two active fabric interconnects. A single system, or management domain, can consist of up to 160 blade or rack servers in any combination.

Cisco UCS Manager can be accessed through an intuitive GUI, a CLI, and a standards-based XML API. More than 22 partners in the management ecosystem

IT Agility with Model-Based Management

Cisco UCS blade servers can be deployed 77 percent faster with 67 percent fewer steps than HP blade servers.

Migration of identities from server to server also takes less time. The process is 34 percent faster with Cisco UCS, with 83 percent fewer steps, than with HP.



Source: Principled Technologies reports from July 2013:

"Migrate Server Identities and Data 34 Percent Faster with Cisco UCS and with 83 Percent Fewer Steps than HP," and "Cisco UCS Blades Deploy 77 Percent Faster with 67 Percent Fewer Steps than HP Blades."

© 2009-2016 Cisco and/or its affiliates. All rights reserved. This document is Cisco Public information.

use the XML API to incorporate Cisco UCS into higher-level processes. For example, orchestration tools now can set every aspect of server configuration before software is installed on a server, eliminating time-consuming and error-prone manual server provisioning processes. For cloud computing environments, where elasticity is of paramount concern, Cisco UCS scales rapidly through automation, providing an excellent match in hardware for what cloud computing delivers in software.

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.

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 that 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 one 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 the way that a system should be identified, configured, and connected to IP and storage networks (Figure 12). For example, an administrator might select server, network, and storage access policies designed to support Oracle database servers.

For storage-intensive servers (such as the Cisco UCS C240 and C3260 Rack Servers), Cisco UCS service profiles can also include storage profiles. Disk groups

© 2009–2016 Cisco and/or its affiliates. All rights reserved. This document is Cisco Public information.

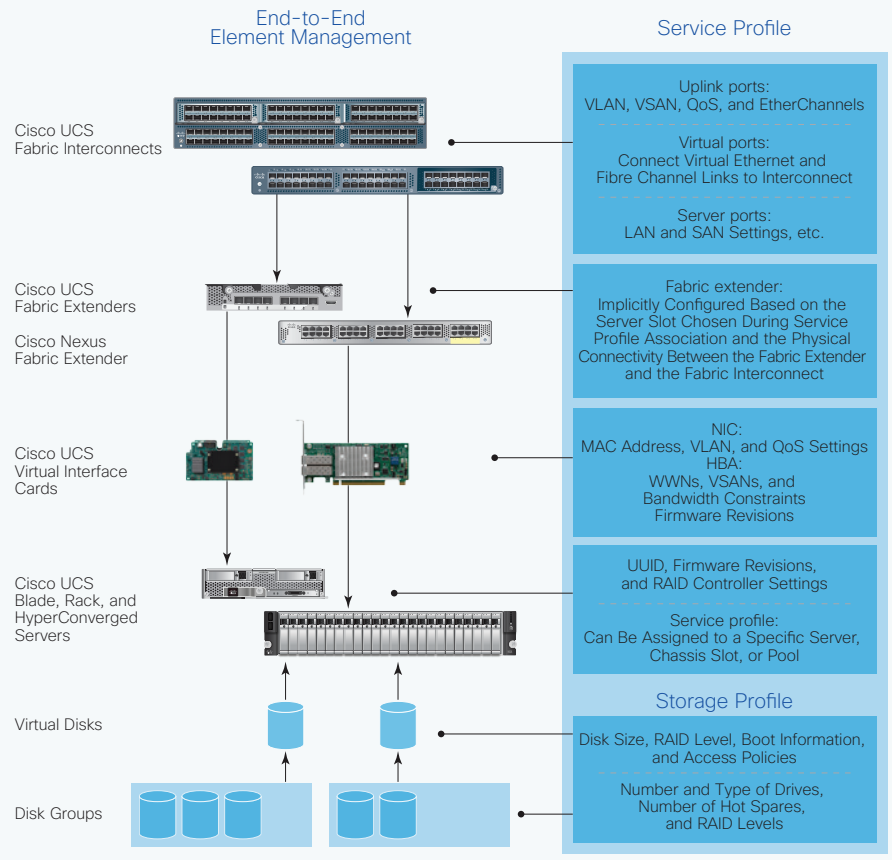


Figure 12 Cisco UCS Service Profiles Fully Specify a Server's Identity, Configuration, and Connectivity

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.

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 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, malleable 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.

© 2009-2016 Cisco and/or its affiliates. All rights reserved. This document is Cisco Public information.

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

Figure 13 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.

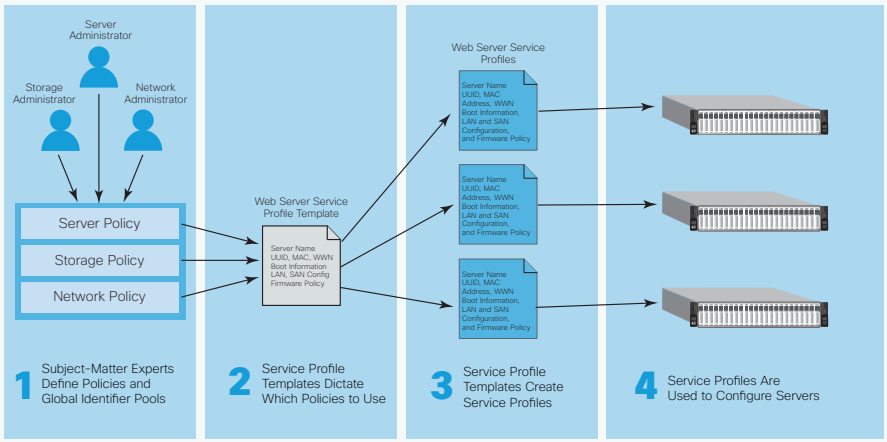


Figure 13 Workflow for Role- and Policy-Based Server Configuration with Cisco UCS Service Profiles

Logging and Audit Capabilities

Cisco UCS Manager produces detailed logs that show how servers are configured and deployed. Because these logs are complete and 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.

© 2009–2016 Cisco and/or its affiliates. All rights reserved. This document is Cisco Public information.

Managing Multiple Cisco UCS Domains with Cisco UCS Central Software
Cisco UCS customers managing growth within a single data center or growth across multiple sites, or both, can use Cisco UCS Central Software to manage multiple Cisco UCS domains using the same concepts that Cisco UCS Manager uses to support a single domain (Figure 14). The software provides benefits including:

- Global awareness of inventory
- Automated standards compliance
- Increased business agility
- Increased asset utilization
- Capability to meet and exceed service-level agreements (SLAs)

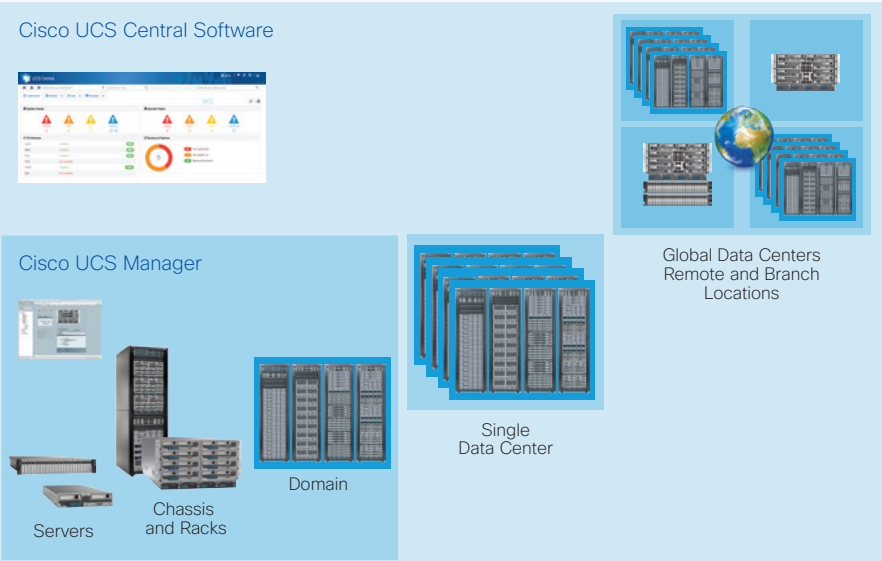


Figure 14 Cisco UCS Central Software Manages Multiple Cisco UCS Domains in a Single Data Center or Multiple Distributed Data Centers

Capable of managing up to 6000 servers, Cisco UCS Central Software provides role- and policy-based management; manages global server identity, configuration, and connectivity; maintains audit logs; and supports access through an intuitive GUI, CLI, and XML API. After a Cisco UCS domain is registered with Cisco UCS Central Software, the domain can inherit policies such as service profiles and templates. Cisco UCS Central Software can direct the policies to be applied to specific servers in specific domains, and the local Cisco UCS Manager applies them (Figure 15).

Cisco UCS Servers

Cisco UCS is based on industry-standard, x86-architecture servers with Cisco innovations and Intel Xeon processors. 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

© 2009–2016 Cisco and/or its affiliates. All rights reserved. This document is Cisco Public information.

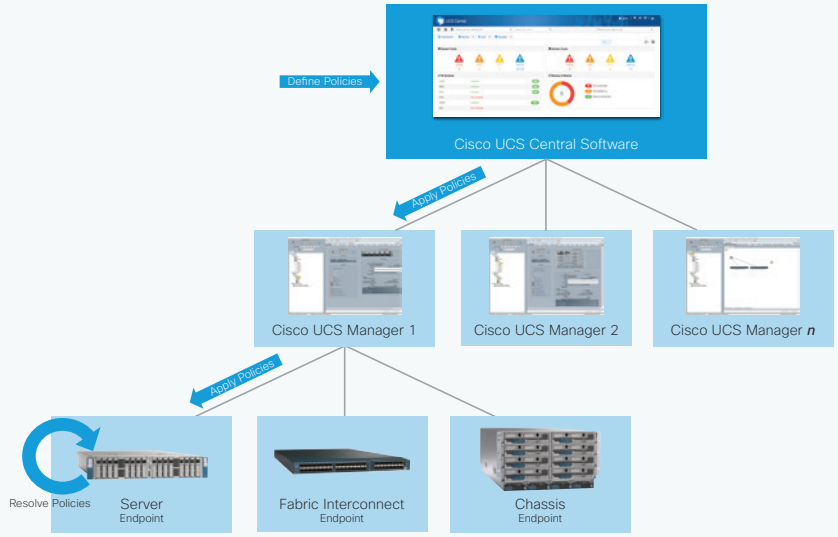


Figure 15 Cisco UCS Central Software Extends Role- and Policy-Based Management Across Multiple Cisco UCS Domains

than 100 world-record-setting benchmark results that demonstrate our leadership in application areas including virtualization, cloud computing, enterprise applications, database management systems, enterprise middleware, and high-performance computing and in basic CPU performance metrics.

Matching 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 rack or blade form factor that best meets your organization's data center requirements and preferred purchasing model. We classify our servers into four categories:

- **Mission-critical servers:** These servers are optimized for high performance and high reliability, availability, and serviceability (RAS). This category includes 2- and 4-socket servers with up to 6 TB of memory, up to 10 PCIe slots for massive I/O bandwidth and flexibility, and up to 12 internal disk drives, with optional high-performance SSD drive and RAID options. Equipped with the Intel Xeon processor E7 family, these servers are for mission-critical applications, including enterprise databases, enterprise resource planning (ERP) applications, and consolidation and virtualization workloads.
- **Enterprise-class servers:** These servers are optimized for performance and offer a wide variety of processing, memory, I/O, and internal disk capacity options. Powered by the Intel Xeon processor E5 family, these servers support up to 1.5 TB of main memory, up to 5 PCIe slots for I/O bandwidth and flexibility, and up to 56 internal disk drives with built-in and additional RAID options. These servers are optimized for database and data warehouse workloads; big data applications; enterprise application middleware; and collaboration, web, and IT infrastructure functions.

© 2009-2016 Cisco and/or its affiliates. All rights reserved. This document is Cisco Public information.

- **Scale-out servers:** These servers are optimized for good performance with excellent value. They offer slightly more limited processor, memory, and I/O expansion capabilities and are suited for scale-out applications, including big data, and web workloads and IT infrastructure functions including proxy and caching servers.

Powered by Intel Xeon Processors

Our servers are equipped with three advanced microprocessor families from Intel:

- **Intel Xeon processor E7 family:** The Intel Xeon processor E7 family is designed to meet the mission-critical IT challenge of managing and keeping business-critical data secure. Powerful, reliable servers such as the Cisco UCS C460 M2 High-Performance Rack Server are equipped with the top-of-the-line Intel Xeon processor E7 family to deliver performance that is excellent for the most data-demanding workloads, with improved scalability and increased memory and I/O capacity. These features help businesses quickly adapt to short-term changes in business needs while addressing requirements for long-term business growth. Advanced reliability and security features help maintain data integrity, accelerate encrypted transactions, and increase the availability of mission-critical applications. The powerful and reliable Intel Xeon processor E7 product family delivers flexibility for business-critical solutions.
- **Intel Xeon processor E5 family:** The Intel Xeon processor E5 family is at the core of a flexible and efficient data center that meets diverse business needs and is used in Cisco scale-out and enterprise-class servers. This family of processors is designed to deliver versatility, with the best combination of performance, built-in capabilities, and cost effectiveness. The Intel Xeon processor E5 family delivers exceptional performance to a broad range of data center environments and applications: from virtualization and cloud computing to design automation and real-time financial transactions. With these processors, I/O latency is dramatically reduced with Intel Integrated I/O, which helps eliminate data bottlenecks, simplify operations, and increase agility.

Industry-Leading Memory Density

When Cisco first entered the server market, virtualization was gaining widespread use but was forcing IT organizations into making costly compromises. The bottleneck in most environments was not CPU capacity, but memory footprint. This challenge left organizations with the choice of moving from 2- to 4-socket servers simply to gain the increased memory capacity that 4-socket servers offered. Cisco was the first to recognize this limitation and to develop and patent extended memory technology that would allow the state-of-the-art Intel Xeon processors of the time to access twice the amount of memory. Cisco Extended Memory Technology gave IT organizations the choice of using a 2-socket server to achieve greater memory capacity than the capacities that other vendors could achieve, or of populating the additional DIMM slots with lower-cost, lower-density memory. Both options helped increase virtual machine density and reduce TCO, including by helping customers make more effective use of software licenses by increasing utilization and performance on 2-socket servers.

Today, we continue to lead by delivering servers with the highest memory capacities offered by today's Intel Xeon processors. The Cisco UCS B200 M4 is remarkable

© 2009–2016 Cisco and/or its affiliates. All rights reserved. This document is Cisco Public information.

in its support for up to 1.6 TB of memory in a half-width blade server. Similarly, the Cisco UCS C460 M4 is remarkable in its support for 6 TB of memory: enough memory so that only a few servers can support an entire ERP system.

Regardless of the trajectory of the memory market, Cisco customers can be assured that Cisco has their best interests in mind and is ready to offer options that lower TCO regardless of the moves that other vendors choose to make.

Industry-Leading Bandwidth

Cisco UCS rack servers are designed to host up to 10 industry-standard PCIe form-factor I/O cards, giving organizations freedom and flexibility to use the I/O configuration that best meets their needs. Cisco UCS blade servers are designed to accommodate up to two mezzanine form-factor I/O cards made by Cisco and third parties to provide access to massive amounts of backplane capacity.

Cisco UCS virtual interface cards, available only in Cisco servers, 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 servers.

Cisco UCS VICs provide access to more blade server midplane bandwidth than any other mezzanine card. With access to up to 80 Gbps of bandwidth from a half-width blade server and up to 160 Gbps of bandwidth from a full-width blade server (equipped with two cards), we are poised to keep I/O bottlenecks in abeyance even as future multicore processors demand even more I/O bandwidth.

This massive 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 switching in each chassis, and the physical partitioning of bandwidth between I/O modalities, Cisco UCS delivers up to 320 Gbps of bandwidth for every eight blades compared to only 92 Gbps for a traditional 16-server blade system, an improvement of almost six times (using Cisco UCS 6300 Series Fabric Interconnects; see Figure 16).

Lower Latency to Accelerate Transactional 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 17). 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 17).

Servers with 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.

© 2009–2016 Cisco and/or its affiliates. All rights reserved. This document is Cisco Public information.

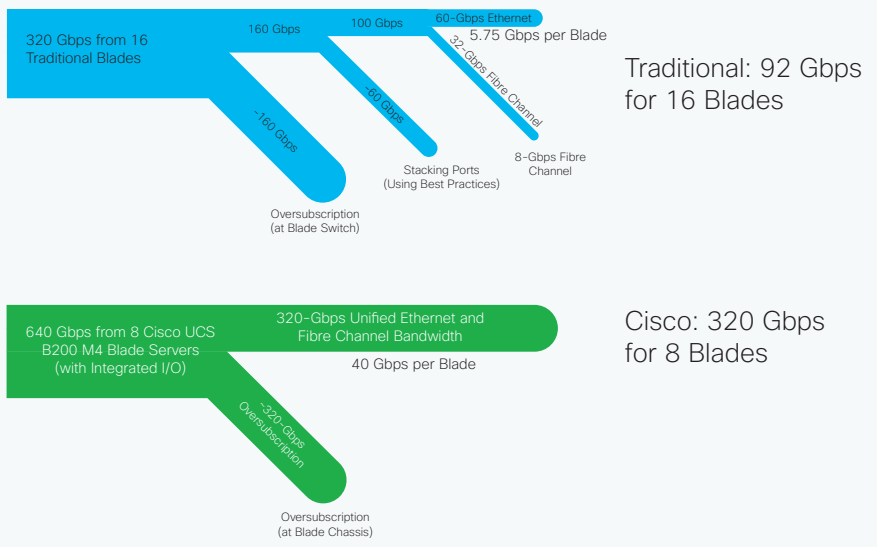


Figure 16 Cisco SingleConnect Technology Delivers More than Three Times the Bandwidth to Each Blade Than Traditional Blade Server Chassis

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 modular 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.

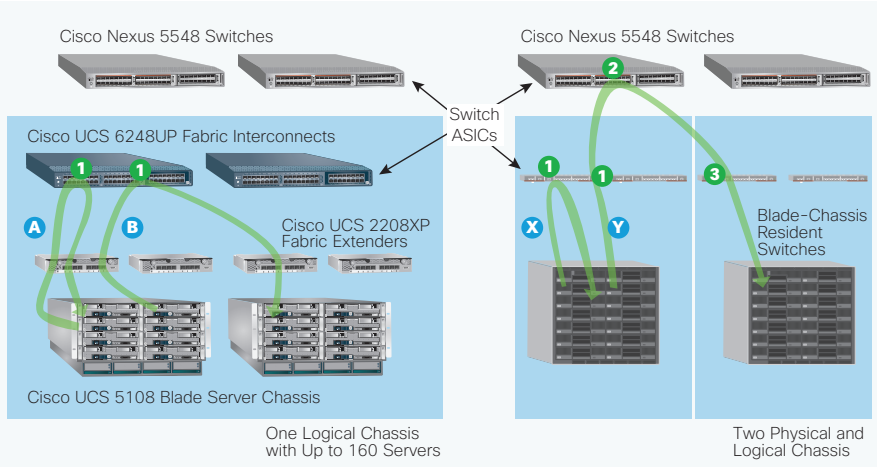


Figure 17 Every Server Is Only One Network Hop Away in Cisco UCS Compared to Up to Three Hops in Traditional Blade Server Environments

© 2009–2016 Cisco and/or its affiliates. All rights reserved. This document is Cisco Public information.

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.

Our rack servers are similarly integrated into Cisco UCS with lower infrastructure cost per server. Instead of requiring up to five active switching components at the top of every rack (two Ethernet, two Fibre Channel, and one management network switch), Cisco UCS requires only two low-cost, low-power-consuming Cisco Nexus fabric extenders at the top of every other rack. This arrangement dramatically simplifies the network and physical infrastructure needed to support a large server farm. The example in Figure 18 demonstrates how the simplified infrastructure of Cisco UCS contributes to 38 percent lower TCO for a 160-server installation.

This graph compares the 3-year TCO for 80 HP ProLiant DL380 Gen9 Servers and 80 HP BL460c G9 servers with the 3-year TCO for 80 Cisco UCS C240 M4 Rack Servers and 80 Cisco UCS B200 M4 Blade Servers. Each server has two Intel Xeon processor E5-2640 v3 CPUs and 64 GB of memory. HP networking includes two 10 Gigabit Ethernet and two 8-Gbps Fibre Channel connections for the HP rack servers, plus HP FlexFabric modules in the HP blade chassis. The Cisco solution includes the Cisco UCS VIC 1227 dual-port 10-Gbps unified fabric adapter for Cisco rack servers and Cisco UCS VIC 1340 for Cisco blade servers, plus corresponding switches. Pricing is as of March 7, 2016.

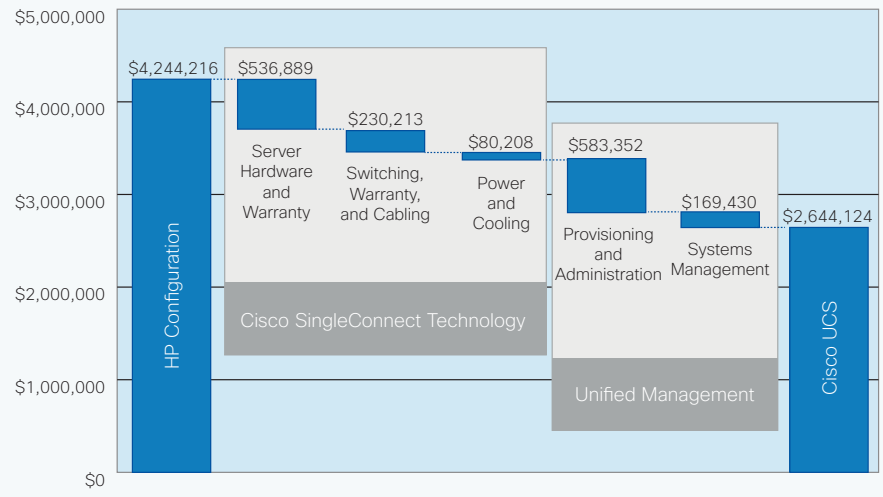


Figure 18 Three-Year TCO for Cisco UCS Is 38 Percent Less Than for Traditional Environments for a Combination of 80 Blade and 80 Rack Servers (US$)

Flexibility for Unified or Standalone Operation

Cisco UCS C-Series Rack Servers are unique in the industry in their capability to be integrated with Cisco UCS or used as standalone servers. This capability offers IT organizations the flexibility to purchase Cisco rack servers today and integrate them as part of a single unified system as their needs dictate.

Integrated Operation with Single-Wire Management

When integrated as part of Cisco UCS, Cisco rack servers gain the same benefits of Cisco unified fabric and unified management as Cisco blade servers. Just as blade servers integrate into the system through a single set of cables that carry network, storage, and management traffic, Cisco rack servers connect through a single set of cables (Figure 19).

Single-wire management is enabled with Cisco rack servers through the Cisco UCS VIC 1225 or 1385, which separates management traffic from production data and storage traffic, passing it to an internal switch that connects to the Cisco Integrated

© 2009-2016 Cisco and/or its affiliates. All rights reserved. This document is Cisco Public information.

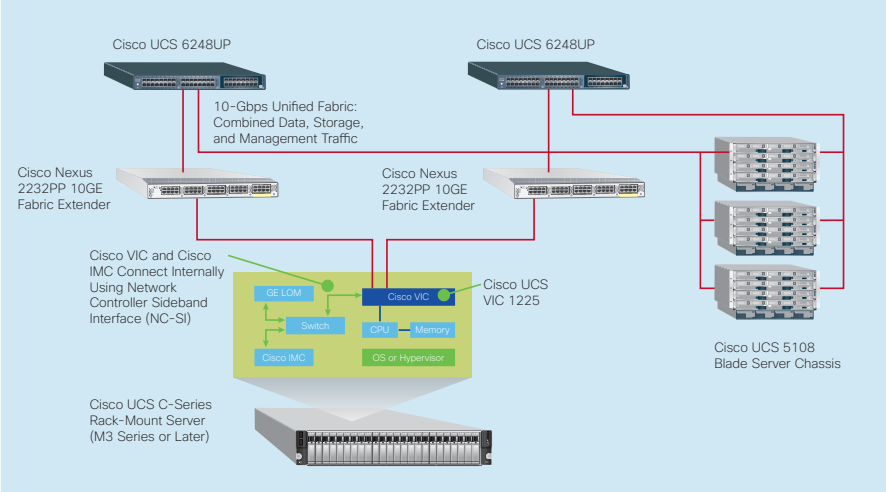


Figure 19 Cisco UCS C-Series Rack Servers Integrate into Cisco UCS with Single-Wire Management

Management Controller (IMC). The internal switch also makes the controller accessible for standalone management through the server's network management ports.

When single-wire management is configured, the unified fabric carries management traffic that is securely separated 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.

Standalone Operation with Cisco Integrated Management Controller
When operated as standalone servers, Cisco UCS C-Series Rack Servers provide up to three management interfaces that can be accessed by in-band or out-of-band tools and techniques (Figure 20):

- Ethernet network access to the Integrated Management Controller
- Agent and agentless management with third-party tools through in-band data-plane connections
- Front- or back-panel access for video, USB (with the capability to boot from a USB CD/DVD drive), and serial console access

The Cisco IMC runs in the system's baseboard management controller (BMC) and can be accessed through the server network management ports for standalone operation. The Cisco controller provides out-of-band management that can be accessed through standard management protocols, CLIs, and web-based interfaces.

© 2009-2016 Cisco and/or its affiliates. All rights reserved. This document is Cisco Public information.

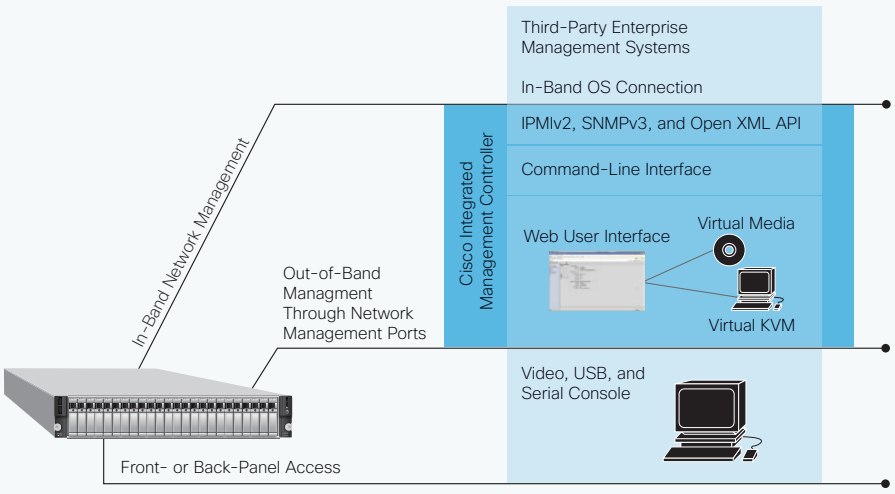


Figure 20 Cisco UCS C-Series Rack Servers Can Be Managed Through a Range of Physical and Logical Management Interfaces

Intelligent Platform Management Interface Version 2
Intelligent Platform Management Interface (IPMI) supports out-of-band management through third-party tools, including commercial enterprise management systems and open-source tools such as ipmitool. IPMI allows these tools to manage server power states and monitor operation parameters available through temperature, fan-speed, power-supply voltage, and power sensors.

Simple Network Management Protocol Version 3
Simple Network Management Protocol (SNMP) supports out-of-band management with third-party tools, including network management tools that use SNMP to monitor system status variables and receive SNMP traps in the event that the status falls outside predetermined ranges.

Open XML API
The Integrated Management Controller supports an open XML API that enables third-party software to access all the system's features and capabilities.

Command-Line Interface
The Cisco UCS Integrated Management Controller CLI can be accessed through a Secure Shell (SSH) connection to the controller. Through this interface, administrators can perform server control and administration tasks, and they can write scripts for configuration tasks so that these tasks can be reliably reproduced on a number of servers without errors.

Web User Interface
The web user interface supports out-of-band management through a standard web browser. It includes server management; virtual media; and remote keyboard, video, and mouse (KVM) capabilities:

- **Server management** includes power management, server reset, component inventory, and event logging.

© 2009-2016 Cisco and/or its affiliates. All rights reserved. This document is Cisco Public information.

- **Virtual media** capabilities enable peripherals such as CD and DVD drives to appear as if they were connected directly to the server, facilitating remote OS and application software installation.
- **Remote KVM** capabilities gives remote administrators the same level of control, including console video control, as when they are physically connected to the server.

Enterprise Management Tools
Third-party management tools typically use a combination of in-band and out-of-band management techniques, both of which are supported by Cisco UCS C-Series servers:

- **In-band management** is performed through the server's data network connection. Different tools use different techniques, including interaction with the host operating system with and without the use of agents. In-band management can interact with OS-based management tools to accomplish tasks including inventory, performance management, troubleshooting, and OS and interface provisioning.
- **Out-of-band management** tools such as Altiris Deployment Solution, BMC BladeLogic, CA Spectrum, HP IT Performance Suite, IBM Tivoli, and Microsoft System Center use Integrated Management Controller interfaces available through the network management port. These tools typically interact with servers through IPMI, SNMP, or the open XML API.

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 smart infrastructure that uses integrated, model-based management to simplify and accelerate deployment of enterprise-class 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
- Scalability through a design for up to 160 discrete servers and thousands of virtual machines, the capability to scale I/O bandwidth to match demand, the low infrastructure cost per server, and the capability to manage up to 6000 servers with Cisco UCS Central Software

© 2009–2016 Cisco and/or its affiliates. All rights reserved. This document is Cisco Public information.

- **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 is poised to help you move to 40 Gigabit Ethernet with the new Cisco UCS 6300 Series Fabric Interconnects

When we first entered the server market in 2009, its 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 46,500 customers and more than 100 world-record performance benchmarks. We have continued to innovate and demonstrate our commitment to customers and to the server market. With three generations of fabric technology supporting modular upgrades to the system's connectivity, and with even more generations of Intel Xeon processors incorporated into our products, We demonstrates the degree to which we support customer investments and how we are ready to take our customers well into the future.

For More Information

For more information about Cisco UCS, visit http://www.cisco.com/go/ucs.

For more information about Cisco UCS performance, visit http://www.cisco.com/go/ucsatwork.

Americas Headquarters
Cisco Systems, Inc.
San Jose, CA

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

Europe Headquarters
Cisco Systems International BV Amsterdam,
The Netherlands

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

Cisco and the Cisco logo are trademarks or registered trademarks of Cisco and/or its affiliates in the U.S. and other countries. To view a list of Cisco trademarks, go to this URL: www.cisco.com/go/trademarks. Third party trademarks mentioned are the property of their respective owners. The use of the word partner does not imply a partnership relationship between Cisco and any other company. (1110R)

LE-49401-00 07/16