White paper
Cisco public

BIOS Performance and workload: Tuning guide for Cisco UCS M8 Platforms

Using the Intel® Xeon® 6 Processor Family

Contents

Purpose and scope 4

What you will learn 4

BIOS-tuning scenarios 4

Tuning for general-purpose workloads 4

Tuning for enterprise workloads 5

Product overview 5

Cisco UCS X210c M8 Compute Node 5

Cisco UCS C220 M8 Rack Server 5

Cisco UCS C240M8 Rack Server 5

Intel Xeon 6 Processor Family 6

Cisco UCS BIOS options 7

Processor settings 7

Intel Hyper-Threading Technology 7

Enhanced Intel SpeedStep Technology 7

Intel Turbo Boost Technology 7

Processor Prefetchers 7

Hardware prefetcher 8

Adjacent-cache-line prefetcher 8

Data cache unit streamer prefetcher 8

Data cache unit–IP prefetcher 9

Last-level cache prefetch 9

Intel Virtualization Technology 9

Intel Ultra Path Interconnect link enablement 10

UPI power management 10

UPI Link Frequency Select 10

Sub-NUMA clustering 10

Extended prediction table prefetch 10

KTI prefetch 11

XPT remote prefetch 11

Last-level cache deadline 11

Memory settings 11

NUMA Optimized 11

Virtual NUMA 12

Memory RAS Configuration 12

Patrol scrub 12

Power and performance configuration 13
Enhanced CPU performance 13
Energy-efficient turbo mode 13
Intel Turbo Boost Technology 13
Processor C6 report 13
Processor C1E 14
Package C-state control 14
Power Performance tuning 14
Processor EPP Profile 15
Latency Optimized Mode 15
Workload configuration 15
Fan policy 15
BIOS settings for Cisco UCS M8 servers 16
BIOS recommendations for various general-purpose workloads 17
CPU-intensive workloads 17
Energy-efficient workloads 18
Low-latency workloads 18
Summary of BIOS settings optimized for general-purpose workloads 18
Additional BIOS recommendations for enterprise workloads 20
Virtualization and Container-base workloads 20
Virtualization workloads 20
Container workloads 20
Relational database workloads 21
Data analytics workloads 21
Analytical database systems workloads 21
High-Performance Computing (HPC) workloads 22
HPC workloads 22
Summary of BIOS settings recommended for enterprise workloads 22
Conclusion 24
For more information 25

Purpose and scope

The Basic Input and Output System (BIOS) tests and initializes the hardware components of a system and boots the operating system from a storage device. A typical computational system has several BIOS settings that control the system's behavior. Some of these settings are directly related to the performance of the system.

This document explains the BIOS settings that are valid for the Cisco Unified Computing System™ (Cisco UCS®) M8 server generation of the following servers: Cisco UCS C220 M8 Rack Server, Cisco UCS C240 M8 Rack Server, and Cisco UCS X210c M8 Compute Node. All servers use the Intel® Xeon® 6 Processor family. The document describes how to optimize the BIOS settings to meet requirements for the general-purpose workloads optimal for best performance and energy efficiency for the Cisco UCS M8 generation of blade and rack servers. This document also describes the BIOS recommendations for industry-standard Enterprise workloads.

With the release of the Intel Xeon® 6 Processor Family (architecture code-named Sapphire Rapids), Cisco released seventh-generation Cisco UCS servers to take advantage of the increased number of cores, higher memory speeds, and PCIe gen 5.0 features of the new processors, thus benefiting CPU-, memory-, and I/O-intensive workloads.

Understanding the BIOS options will help you select appropriate values to achieve optimal system performance. This document does not discuss the BIOS options for specific firmware releases of Cisco UCS M8 servers. The settings demonstrated here are generic.

What you will learn

The process of setting performance options in your system BIOS can be daunting and confusing, and some of the options you can choose are obscure. For most options, you must choose between optimizing a server for power savings or for performance. This document provides some general guidelines and suggestions to help you achieve optimal performance from your Cisco UCS blade M8 and rack M8 servers that use Intel Xeon 6 Processor Family CPUs.

BIOS-tuning scenarios

This document focuses on two main scenarios: how to tune the BIOS for general-purpose workloads and how to tune the BIOS for enterprise workloads.

Tuning for general-purpose workloads

With the latest multiprocessor, multicore, and multithreading technologies in conjunction with current operating systems and applications, the new Cisco UCS M8 servers based on the Intel Xeon 6 Processor Family deliver the highest levels of performance, as demonstrated in numerous industry-standard benchmark publications.

Cisco UCS servers with standard settings already provide an optimal ratio of performance to energy efficiency. However, through BIOS settings you can further optimize the system with higher performance and less energy efficiency. Basically, this optimization operates all the components in the system at the maximum speed possible and prevents the energy-saving options from slowing down the system. In general, optimization to achieve greater performance is associated with increased consumption of electrical power. This document explains how to configure the BIOS settings to achieve optimal computing performance.

Tuning for enterprise workloads

With the evolution of computer architecture, performance has reached results that were unimaginable a few years ago. However, the complexity of modern computer architectures requires end users and developers to know how to write code. It also requires them to know how to configure and deploy software for a specific architecture to get the most out of it.

Performance tuning is difficult and general recommendations are problematic. This document tries to provide insights into optimal BIOS settings and OS tunings that have an impact on overall system performance. This document does not provide generic rule-of-thumb (or values) to be used for performance tuning. The finest tuning of the parameters described requires a thorough understanding of the enterprise workloads and the Cisco UCS platform on which they run.

Product overview

Cisco UCS X210c M8 Compute Node

The Cisco UCS X210c M8 Compute Node two-socket server brings Intel Xeon 6 Processors to the Cisco UCS X-Series Modular System powered by Cisco Intersight®. It offers more performance, faster I/O, and more storage than the previous M6 and M7 generation servers.

New to the Cisco UCS X210c M8 Compute Node is Cisco UCS X10c Pass Through Controller for E3.S drives, supporting up to nine PCIe Gen5 NVMe drives. With all the benefits of a modular system and increased storage capacity, the X210c M8 is an ideal server for data-intensive applications, including hyperconverged infrastructure, AI, databases, and backup and disaster recovery (for example, Rubrik and Cohesity).

Cisco UCS C220 M8 Rack Server

The 1RU, 2-socket Cisco UCS C220 M8 Rack Server is designed to meet the needs of customers that choose to deploy high-density rack-mount servers. Using the latest Intel processors, it is a versatile general-purpose application and infrastructure server delivering leading performance and efficiency for a wide range of workloads, including virtualization, collaboration, and bare-metal applications.

The Cisco UCS C220 M8 Rack Server extends the capabilities of the Cisco Unified Computing System (Cisco UCS) rack server portfolio by incorporating Intel Xeon 6 Processors. It improves security, performance, and efficiency while helping achieve sustainability goals with built-in accelerators such as Intel Trust Domain Extensions (TDX), Intel Data Streaming Accelerator (DSA), Intel QuickAssist Technology (QAT), Intel Advanced Matrix Extensions (AMX), and In-Memory Analytics Accelerator (IAA).

Cisco UCS C240M8 Rack Server

The 2RU, 2-socket Cisco UCS C240 M8 Rack Server is designed to meet the needs of customers who need I/O flexibility and larger storage capacity rack-mount servers. Using the fastest Intel processors, it is a versatile general-purpose application and infrastructure server delivering leading performance and efficiency for a wide range of workloads, including AI, big-data analytics, databases, collaboration, virtualization, and high-performance computing.

The Cisco UCS C240 M8 Rack Server extends the capabilities of the Cisco Unified Computing System (Cisco UCS) rack server portfolio by incorporating Intel Xeon 6 Processors. It improves security, performance, and efficiency while helping achieve sustainability goals with built-in accelerators such as Intel Trust Domain Extensions (TDX), Intel Data Streaming Accelerator (DSA), Intel QuickAssist Technology (QAT), Intel Advanced Matrix Extensions (AMX), and In-Memory Analytics Accelerator (IAA).

Intel Xeon 6 Processor Family

The Intel Xeon 6 Processor Family introduces a robust computing platform that excels at both performance and efficiency, which are crucial for meeting the evolving demands of modern data centers. From compute-intensive AI to scale-out microservices, the processor family provides versatility for diverse workload requirements.

The Intel Xeon 6700-series and Intel Xeon 6500-series processors are delivered in an updated server platform design featuring high performance with cost- and power-efficient solutions ideal for the widest array of data-center environments. These processors come in one-socket to four-socket options with enhanced I/O and memory within established data-center power and cooling footprints.

Performance and efficiency without compromise

The Intel Xeon 6 Processor Family introduces an innovative modular x86 architecture that allows data-center architects to configure and deploy infrastructures that are purpose-built for your unique needs and workloads across private, public, and hybrid clouds. Intel Xeon 6 Processors offer tiered capabilities from entry-level to demanding workloads through options for increased numbers of cores, larger caches, faster and higher-capacity memory, and improved I/O over previous generations.

Intel Xeon 6 Processors with Performance-cores (P-cores) are optimized for high performance per core. With more cores, double the memory bandwidth, and AI acceleration in every core, Intel Xeon 6 Processors provide twice the performance for the widest range of workloads, including AI and High-Performance Computing (HPC). Intel Xeon 6 Processors with P-cores excel at a wide range of workloads, delivering better performance than any other general-purpose CPU for compute-intensive workloads such as AI inference and Machine Learning (ML). Intel Xeon 6 Processors with P-cores are great for public-cloud workloads, with improved performance per vCPU for floating point operations, transactional databases, and HPC workloads. Through their leadership in AI inferencing, Intel Xeon Processors continue to be the host CPU of choice on the world's most powerful AI accelerator platforms for data-preprocessing support.

- Enable AI everywhere with AI acceleration in every core. Intel Advanced Matrix Extensions (Intel AMX) speeds up inferencing for INT8- and BF16-trained and offers new support for FP16-trained models with up to 2048 floating point operations per cycle per core for INT8 and 1024 floating point operations per cycle per core for BF16/FP16.
- Improve memory throughput with the fastest DDR5 memory available, MRDIMM. MRDIMMs can deliver more than 37 percent more memory bandwidth than RDIMMs, with an expected data-transfer rate of up to 8800 Megatransfers per second (MT/s). Intel Xeon 6 Processors with P-cores also support DDR5 6400 high-speed memory, providing memory bandwidth gains.
- Take advantage of up to 128 cores per socket with up to 504 MB L3 cache and exceptionally low latency at large L3 access sizes. Intel Advanced Vector Extensions 512 (Intel AVX-512) is only supported on Intel Xeon 6 Processors with P-cores and can be used out of the box, boosting the speed of vector math common to HPC and classical AI workloads.

Cisco UCS BIOS options

This section describes the options you can configure in the Cisco UCS BIOS.

Processor settings

This section describes processor options you can configure.

Intel Hyper-Threading Technology

You can specify whether the processor uses Intel Hyper-Threading Technology, which allows multithreaded software applications to process threads in parallel within each processor. You should test the CPU hyperthreading option both enabled and disabled in your specific environment. If you are running a single-threaded application, you should disable hyperthreading.

The setting can be either of the following:

- **Disabled:** the processor does not permit hyperthreading.
- **Enabled:** the processor allows parallel processing of multiple threads.
- **Platform default:** the BIOS uses the value for this attribute contained in the BIOS defaults for the server type and vendor.

Enhanced Intel SpeedStep Technology

Enhanced Intel SpeedStep Technology, which allows the system to dynamically adjust processor voltage and core frequency. This technology can result in decreased average power consumption and decreased average heat production.

The setting can be either of the following:

- **Disabled:** the processor never dynamically adjusts its voltage or frequency.
- **Enabled:** the processor uses Enhanced Intel SpeedStep Technology and enables all supported processor sleep states to further conserve power.
- **Platform default:** the BIOS uses the value for this attribute contained in the BIOS defaults for the server type and vendor.

Intel Turbo Boost Technology

Intel Turbo Boost Technology provides the capability for the CPU to adjust itself to run higher than its stated clock speed if it has enough power to do so. When the processor uses Intel Turbo Boost Technology, which allows the processor to automatically increase its frequency if it is running below power, temperature, or voltage specifications.

Intel Turbo Boost Technology depends on Enhanced Intel SpeedStep technology: If you want to enable Intel Turbo Boost, you must enable Intel SpeedStep first. If you disable Intel SpeedStep, you lose the capability to use Intel Turbo Boost.

Processor Prefetchers

Intel Xeon 6 Processor Family processors have several layers of cache. Each core has a tiny Layer-1 cache, sometimes referred to as the Data-Cache Unit (DCU), that has 32 KB for instructions and 32 KB for data. Slightly bigger is the Layer-2 cache, with 256 KB shared between data and instructions for each core. In addition, all cores on a chip share a much larger Layer-3 cache, which is about 10 to 45 MB in size (depending on the processor model and number of cores).

The prefetcher settings provided by Intel primarily affect the Layer-1 and Layer-2 caches on a processor core (see Table 1). You will likely need to perform some testing with your individual workload to find the combination that works best for you. Testing on the Intel Xeon 6 Processor Family has shown that most applications run best with all prefetchers enabled. See Tables 2 and 3 for guidance.

Table 1. Processor prefetcher options

Processor prefetcher options	Cache affected
Hardware prefetcher	Layer 2
Adjacent-cache-line prefetcher	Layer 2
DCU prefetcher	Layer 1
DCU instruction pointer (DCU-IP) prefetcher	Layer 1

Hardware prefetcher

The hardware prefetcher prefetches additional streams of instructions and data into the Layer-2 cache upon detection of an access stride. This behavior is more likely to occur during operations that sort sequential data, such as database table scans and clustered index scans, or that run a tight loop in code.

You can specify whether the processor allows the Intel hardware prefetcher to fetch streams of data and instructions from memory into the unified second-level cache when necessary.

The setting can be either of the following:

- **Disabled:** the hardware prefetcher is not used.
- **Enabled:** the processor uses the hardware prefetcher when cache problems are detected.

Adjacent-cache-line prefetcher

The adjacent-cache-line prefetcher always prefetches the next cache line. Although this approach works well when data is accessed sequentially in memory, it can quickly litter the small Layer-2 cache with unneeded instructions and data if the system is not accessing data sequentially, causing frequently accessed instructions and code to leave the cache to make room for the adjacent-line data or instructions.

You can specify whether the processor fetches cache lines in even or odd pairs instead of fetching just the required line.

The setting can be either of the following:

- **Disabled:** the processor fetches only the required line.
- **Enabled:** the processor fetches both the required line and its paired line.

Data cache unit streamer prefetcher

Like the hardware prefetcher, the DCU streamer prefetcher prefetches additional streams of instructions or data upon detection of an access stride; however, it stores the streams in the tiny Layer-1 cache instead of the Layer-2 cache.

This prefetcher is a Layer-1 data cache prefetcher. It detects multiple loads from the same cache line that occur within a time limit. Making the assumption that the next cache line is also required, the prefetcher loads the next line in advance to the Layer-1 cache from the Layer-2 cache or the main memory.

The setting can be either of the following:

- **Disabled:** the processor does not try to anticipate cache read requirements and fetches only explicitly requested lines.
- **Enabled:** the DCU prefetcher analyzes the cache read pattern and prefetches the next line in the cache if it determines that it may be needed.

Data cache unit–IP prefetcher

The DCU-IP prefetcher predictably prefetches data into the Layer-1 cache on the basis of the recent instruction pointer load instruction history.

You can specify whether the processor uses the DCU-IP prefetch mechanism to analyze historical cache access patterns and preload the most relevant lines in the Layer-1 cache.

The setting can be either of the following:

- **Disabled:** the processor does not preload any cache data.
- **Enabled:** the DCU-IP prefetcher preloads the Layer-1 cache with the data it determines to be the most relevant.

Last-level cache prefetch

This BIOS option configures the processor's Last-Level Cache (LLC) prefetch feature as a result of the noninclusive cache architecture. The LLC prefetcher exists on top of other prefetchers that can prefetch data into the core DCU and Mid-Level Cache (MLC). In some cases, disabling this option can improve performance.

The setting for this BIOS option can be either of the following:

- **Disabled:** the LLC prefetcher is disabled. The other core prefetchers are not affected.
- **Enabled:** the core prefetcher can prefetch data directly to the LLC.
- **Platform default:** the LLC prefetch option is disabled.

Intel Virtualization Technology

Virtualization abstracts hardware that allows multiple workloads to share a common set of resources. On shared virtualized hardware, a variety of workloads can colocate while maintaining full isolation from each other, freely migrate across infrastructures, and scale as needed. Intel Virtualization Technology (Intel VT) allows a platform to run multiple operating systems and applications in independent partitions.

Intel VT represents a growing portfolio of technologies and features that make virtualization practical by eliminating performance overheads and improving security. Intel VT provides hardware assistance to the virtualization software, reducing its size, cost, and complexity. Special attention is also given to reduce the virtualization overheads occurring in cache, I/O, and memory.

- **Enabled:** the processor allows multiple operating systems in independent partitions.

Note: If you change this option, you must power the server off and on before the setting takes effect.

Intel Ultra Path Interconnect link enablement

The Intel Ultra Path Interconnect (UPI) BIOS option allows you to change the number of UPI links. Use this option to configure the UPI topology to use fewer links between processors, when available. Changing this option from the default can reduce Intel UPI bandwidth performance in exchange for less power consumption.

The values for this BIOS setting are 1, 2, and Auto.

UPI power management

The UPI power management is used to conserve power on a platform. Low-power mode reduces Intel UPI frequency and bandwidth. This option is recommended to save power; however, UPI power management is not recommended for high-frequency, low-latency, virtualization, and database workloads.

This BIOS option controls the link L0p Enable and link L1 Enable values.

L1 saves the most power but has the greatest impact on latency and bandwidth. L1 allows a UPI link to transition from the full-link-down state. L1 is the deepest power savings state.

L0p allows a partial-link-down state. A subset of all of the lanes will remain awake.

UPI Link Frequency Select

The UPI Link Frequency Select BIOS option allows you to set the UPI link speed. Running the UPI link speed (frequency) at a lower rate can reduce power consumption, but it can also affect system performance.

UPI link frequency determines the rate at which the UPI processor interconnect link operates. If a workload is highly nonuniform memory access (NUMA) aware, sometimes lowering the UPI link frequency can free more power for the cores and result in better overall performance.

Sub-NUMA clustering

SNC (two-way sub-NUMA) divides the LLC into two disjointed clusters called NUMA nodes and is based on address range, with each cluster bound to a subset of the memory controllers in the system. SNC improves average latency to the LLC and memory. For a multisocket system, all SNC clusters are mapped to unique NUMA domains. Integrated memory controller interleaving must be set to the correct value corresponding with the SNC setting. OS support that recognizes each cluster and a separate NUMA node are necessary to take advantage of SNC.

The setting for this BIOS option can be either of the following:

- **Disabled:** The LLC is treated as one cluster when this option is disabled.
- **Enabled:** The LLC capacity is used more efficiently, and latency is reduced as a result of the core and integrated memory controller proximity. This setting may improve performance on NUMA-aware operating systems.
- **Auto:** The CPU determines the SNC functionality.

Extended prediction table prefetch

Extended prediction table (XPT) prefetch is a new capability that is designed to reduce local memory access latency. This prefetcher exists on top of other prefetchers that can prefetch data in the core DCU, MLC, and LLC. The XPT prefetcher will issue a speculative DRAM read request in parallel with an LLC lookup. This prefetch bypasses the LLC, reducing latency. You can specify whether the processor uses the XPT prefetch mechanism to fetch the date into the XPT.

The setting can be either of the following:

- **Disabled:** the processor does not preload any cache data.
- **Enabled:** the XPT prefetcher preloads the Layer-1 cache with the data it determines to be the most relevant.

KTI prefetch

KTI prefetch is a mechanism to get the memory read started early on a DDR bus.

The settings can be one of the following:

- **Disabled:** The processor does not preload any cache data
- **Enabled:** The KTI prefetcher preloads the L1 cache with the data it determines to be the most relevant
- **Platform default:** The BIOS uses the value for this attribute container in the BIOS defaults for the server type.

XPT remote prefetch

The XPT (extended prediction table) remote prefetch (extended prediction table) BIOS option configures the XPT remote prefetcher processor performance option. When it is enabled, this feature can improve remote read request latency from a processor core by directly accessing the UPI. Values for this BIOS setting can be auto, enabled, or disabled.

Last-Level Cache deadline

With the Intel Xeon 6 Processor Family's noninclusive cache scheme, MLC evictions are filled into The Last-Level Cache (LLC) if the data is shared across processor cores. When cache lines are evicted from the MLC, the processor core can flag them as "dead," meaning that they are not likely to be read again. With this option, the LLC can be configured to drop deadlines and not fill them in the LLC.

Values for the LLC dead line BIOS option can be either of the following:

- **Disabled:** if this option is disabled, deadlines will be dropped from the LLC. This setting provides better utilization in the LLC and prevents the LLC from evicting useful data.
- **Enabled:** if this option is enabled, the processor determines whether to keep or drop dead lines. By default, this option is enabled.

Memory settings

You can use several settings to optimize memory performance.

NUMA Optimized

Most modern operating systems, particularly virtualization hypervisors, support NUMA because in the latest server designs a processor is attached to a memory controller: therefore, half the memory belongs to one processor, and half belongs to the other processor. If a core needs to access memory that resides in another processor, a longer latency period is needed to access that part of memory. Operating systems and hypervisors recognize this architecture and are designed to reduce such trips. For hypervisors such as those from VMware and for modern applications designed for NUMA, keep this option enabled.

Virtual NUMA

When virtual NUMA is enabled, two NUMA nodes are created per physical CPU socket without changing memory controller and channel interleaving and LLC grouping. Virtual NUMA mode provides a potential memory bandwidth advantage. The latency between these two virtual NUMA nodes is identical to its local latency. The BIOS options are enabled and disabled. By default, this option is disabled.

Memory RAS Configuration

Memory RAS (Reliability, Availability, and Serviceability) configuration refers to the settings and features within a computer system that enhance its ability to detect and correct memory errors, ensuring data integrity and system uptime. Key aspects include error detection and correction mechanisms like ECC, and redundancy features like memory mirroring and sparing, as well as features like Post Package Repair (PPR).

The setting can be either of the following:

- **ADDDC Sparing:** System reliability is optimized by holding memory in reserve so that it can be used in case other DIMMs fail. This mode provides some memory redundancy, but does not provide as much redundancy as mirroring.
- **Maximum Performance:** Optimizes the system performance and disables all the advanced RAS features.
- **Mirror Mode 1LM:** Mirror Mode 1LM will set the entire 1LM memory in the system to be mirrored, consequently reducing the memory capacity by half. This mode is used for UCS M5 and M6 blade servers.
- **Partial Mirror Mode 1LM:** Partial Mirror Mode 1LM will set a part of the 1LM memory in the system to be mirrored, consequently reducing the memory capacity by half. This mode is used for UCS M5 and M6 blade servers.

Note: For the optimal balance of performance and system stability, you should use the platform default. ADDDC sparing will incur a small performance penalty for memory-intensive workloads.

Patrol scrub

You can specify whether the system actively searches for, and corrects, single-bit memory errors even in unused portions of the memory on the server.

The setting can be either of the following:

- **Disabled:** the system checks for memory Error-Correcting Code (ECC) errors only when the CPU reads or writes a memory address.
- **Enable at End of POST:** the system periodically reads and writes memory searching for ECC errors. If any errors are found, the system attempts to fix them. This option may correct single-bit errors before they become multiple-bit errors, but it may adversely affect performance when the patrol-scrub process is running.

Power and performance configuration

Enhanced CPU performance

This BIOS option helps users modify the enhanced CPU performance settings. When it is enabled, this option adjusts the processor settings and enables the processor to run aggressively, which can improve overall CPU performance, but may result in higher power consumption. Values for this BIOS option can be auto or disabled. By default, the enhanced CPU performance option is disabled.

Note: This BIOS feature is applicable for benchmarks purposes only and is not recommended for any production workloads. When this option is enabled, we highly recommend setting the fan policy at Maximum Power.

Energy-efficient turbo mode

The energy-efficient turbo mode BIOS option allows you to control whether the processor uses an energy-efficiency based policy. In this operation mode, a processor's core frequency is adjusted within the turbo-mode range based on workload.

When energy efficient turbo is enabled, the CPU's optimal turbo frequency will be tuned dynamically based on CPU utilization. The power performance bias setting also influences energy-efficient turbo.

By default, this option is disabled.

Intel Turbo Boost Technology

Intel Turbo Boost Technology depends on Intel SpeedStep: if you want to enable Intel Turbo Boost, you must enable Intel SpeedStep first. If you disable Intel SpeedStep, you lose the capability to use Intel Turbo Boost.

Intel Turbo Boost is especially useful for latency-sensitive applications and for scenarios in which the system is nearing saturation and would benefit from a temporary increase in the CPU speed. If your system is not running at this saturation level and you want the best performance at a utilization rate of less than 90 percent, you should disable Intel SpeedStep to help ensure that the system is running at its stated clock speed at all times.

Processor C6 report

The C6 state is a power-saving halt-and-sleep state that a CPU can enter when it is not busy. Unfortunately, it can take some time for the CPU to leave these states and return to a running condition. If you are concerned about performance (for all but latency-sensitive single-threaded applications), and if you can do so, disable anything related to C-states.

You can specify whether the BIOS sends the C6 report to the operating system. When the OS receives the report, it can transition the processor into the lower C6 power state to decrease energy use while maintaining optimal processor performance.

The setting can be either of the following:

- **Disabled:** the BIOS does not send the C6 report.
- **Enabled:** the BIOS sends the C6 report, allowing the OS to transition the processor to the C6 low-power state.
- **Auto:** The CPU determines the functionality.

Processor C1E

Enabling the C1E option allows the processor to transition to its minimum frequency upon entering the C1 state. This setting does not take effect until after you have rebooted the server. When this option is disabled, the CPU continues to run at its maximum frequency in the C1 state. Users should disable this option to perform application benchmarking.

You can specify whether the CPU transitions to its minimum frequency when entering the C1 state.

The setting can be either of the following:

- **Disabled:** the CPU continues to run at its maximum frequency in the C1 state.
- **Enabled:** the CPU transitions to its minimum frequency. This option saves the maximum amount of power in the C1 state.

Package C-state control

Use this option to configure the lowest processor idle power state (C-state). The processor automatically transitions into package C-states based on the core C-states to which cores on the processor have transitioned. The higher the package C-state, the lower the power use of that idle package state. The default setting, Package C6 (nonretention), is the lowest power idle package state supported by the processor.

You can specify the amount of power available to the server components when they are idle.

The possible settings are as follows:

- **C0/C1 State:** when the CPU is idle, the system slightly reduces power consumption. This option requires less power than C0 and allows the server to return quickly to high-performance mode.
- **C2 State:** when the CPU is idle, the system reduces power consumption more than with the C1 option. This option requires less power than C1 or C0, but the server takes slightly longer to return to high-performance mode.
- **C6 Non-retention:** when the CPU is idle, the system reduces power consumption more than with the C3 option. This option saves more power than C0, C1, or C3, but the system may experience performance problems until the server returns to full power.
- **C6 Retention:** when the CPU is idle, the system reduces power consumption more than with the C3 option. This option consumes slightly more power than the C6 Nonretention option, because the processor is operating at Pn voltage to reduce the package's C-state exit latency.

Power Performance tuning

This BIOS option determines how aggressively the CPU is power-managed and placed into turbo mode. If you select BIOS Control, the system controls the setting. If you select OS Control, the operating system controls the setting.

By default, OS Control is enabled.

Processor EPP Profile

This BIOS option allows you to determine whether system performance or energy efficiency is more important on this server.

These settings are: Performance, Balanced Performance, Balanced power, power

By default, EPP profile set to “Balanced Performance”

Latency Optimized Mode

A set of BIOS settings that prioritize low latency and consistent performance over energy efficiency or other optimizations.

These settings are: Enabled, Disabled

By default, this option set to Disabled.

Workload configuration

You can tune the system’s I/O bandwidth between balanced and I/O sensitive by adjusting the processor’s core and uncore frequencies. This configuration allows users to set a parameter to optimize workload characterization.

This setting can be either of the following:

- **Balanced:** the balanced setting is used for optimization.
- **I/O Sensitive:** the I/O-sensitive setting is used for optimization. By default, I/O Sensitive is enabled.

Fan policy

Fan policy enables you to control the fan speed to reduce server-power consumption and noise levels. Prior to fan policy, the fan speed increased automatically when the temperature of any server component exceeded the set threshold. To help ensure that fan speeds were low, the threshold temperatures of components were usually set to high values. Although this behavior suited most server configurations, it did not address the following situations:

- **Maximum CPU performance:** for high performance, certain CPUs must be cooled substantially below the set threshold temperature. This cooling requires very high fan speeds, which results in increased power consumption and noise levels.
- **Low power consumption:** to help ensure the lowest power consumption, fans must run very slowly and, in some cases, stop completely on servers that allow fans to stop. But slow fan speeds can cause servers to overheat. To avoid this situation, you need to run fans at a speed that is moderately faster than the lowest possible speed.

You can choose the following fan policies:

- **Balanced:** this is the default policy. This setting can cool almost any server configuration, but it may not be suitable for servers with PCI Express (PCIe) cards, because these cards overheat easily.
- **Low Power:** this setting is well suited for minimal-configuration servers that do not contain any PCIe cards.
- **High Power:** this setting can be used for server configurations that require fan speeds ranging from 60 to 85 percent. This policy is well suited for servers that contain PCIe cards that easily overheat and have high temperatures. The minimum fan speed set with this policy varies for each server platform, but it is approximately in the range of 60 to 85 percent.

- **Maximum Power:** this setting can be used for server configurations that require extremely high fan speeds ranging between 70 and 100 percent. This policy is well suited for servers that contain PCIe cards that easily overheat and have extremely high temperatures. The minimum fan speed set with this policy varies for each server platform, but it is approximately in the range of 70 to 100 percent.
- **Acoustic:** the fan speed is reduced to reduce noise levels in acoustic-sensitive environments. Rather than regulating energy consumption and preventing component throttling as in other modes, the Acoustic option could result in short-term throttling to achieve a lowered noise level. Applying this fan control policy might result in short-duration, transient performance impacts.

BIOS settings for Cisco UCS M8 servers

Table 2 lists the BIOS token names, defaults, and supported values for the Cisco UCS M8 blade and rack servers for the Intel Xeon 6 Processor Family.

Table 2. BIOS token names and supported values

BIOS token	Platform default	Supported values
Processor configuration		
Intel Hyper-Threading Technology	Enabled	Enabled and Disabled
CPU performance	Custom	Custom, Enterprise, High-throughput, HPC
Hardware prefetcher	Enabled	Enabled and Disabled
Adjacent cache line prefetcher	Enabled	Enabled and Disabled
DCU IP prefetcher	Enabled	Enabled and Disabled
DCU streamer prefetch	Enabled	Enabled and Disabled
LLC prefetch	Enabled	Enabled and Disabled
Intel Virtualization Technology	Enabled	Enabled and Disabled
Uncore configuration		
UPI link enablement	Auto	Auto, 1, 2, and 3
UPI power management	Disabled	Enabled and Disabled
UPI Link Frequency Select	Auto	Auto, 16 GTs, 20 GTs, and 24 GTs, Use Per Link setting
XPT remote prefetch	Auto	Auto, Enabled, and Disabled
KTI prefetch	Done	Auto, Enabled, and Disabled
XPT prefetch	Auto	Auto, Enabled, and Disabled
Sub-NUMA clustering	Disabled	Auto, Enabled and Disabled
LLC dead line	Enabled	Auto, Enabled, and Disabled

BIOS token	Platform default	Supported values
Memory configuration		
NUMA Optimized	Enabled	Enabled and Disabled
Virtual NUMA	Disabled	Enabled and Disabled
Memory RAS configuration setup		
Select Memory RAS Configuration	ADDDC Sparing	ADDDC Sparing, Maximum Performance, Mirror Mode 1LM, Partial Mirror Mode 1LM
Patrol scrub	Enable at End of POST	Enable at End of POST and Disabled
Power and performance configuration		
Enhanced CPU performance	Disabled	Auto and Disabled
Energy-efficient turbo mode	Disabled	Enabled and Disabled
Intel Turbo Boost Technology	Enabled	Enabled and Disabled
Processor C6 Report	Auto	Auto, Enabled and Disabled
Processor C1E	Enabled	Enabled and Disabled
Package C-state control	C0/C1 State	No Limit, Auto, C0/C1 State, C2, C6 Retention, and C6 Nonretention
Power performance tuning	OS	OS, BIOS, PECI
Processor EPP Profile	Balanced Performrmance	Performance, Balanced Performrmance, Power, Balanced Power
Latency optimized mode	Disabled	Enabled and Disabled
Workload configuration	I/O Sensitive	Balanced and I/O Sensitive

BIOS recommendations for various general-purpose workloads

This section summarizes the BIOS settings recommended to optimize general-purpose workloads:

- CPU-intensive workloads
- Energy-efficient workloads
- Low-latency workloads

The following sections describe each workload.

CPU-intensive workloads

For CPU-intensive workloads, the goal is to distribute the work for a single job across multiple CPUs to reduce the processing time as much as possible. To do this, you need to run portions of the job in parallel. Each process¬, or thread, handles a portion of the work and performs the computations concurrently. The CPUs typically need to exchange information rapidly, requiring specialized communication hardware.

CPU-intensive workloads generally benefit from processors that achieve the maximum turbo frequency for any individual core at any time. Processor power management settings can be applied to help ensure that any component frequency increase can be readily achieved.

This BIOS option helps users modify the enhanced CPU performance settings. When it is enabled, this option adjusts the processor settings and enables the processor to run aggressively, which can improve overall CPU performance, but may result in higher power consumption.

Energy-efficient workloads

Energy-efficient optimizations are the most common balanced performance settings. They benefit most application workloads while also enabling power management settings that have little impact on overall performance. The settings that are applied for energy-efficient workloads increase general application performance rather than power efficiency. Processor power management settings can affect performance when virtualization operating systems are used.

Low-latency workloads

Workloads that require low latency, such as financial trading and real-time processing, require servers to provide a consistent system response. Low-latency workloads are for customers who demand the least amount of computational latency for their workloads. Maximum speed and throughput are often sacrificed to lower overall computational latency. Processor power management and other management features that might introduce computational latency are disabled.

To achieve low latency, you need to understand the hardware configuration of the system under test. Important factors affecting response times include the number of cores, the processing threads per core, the number of NUMA nodes, the CPU and memory arrangements in the NUMA topology, and the cache topology in a NUMA node. BIOS options are generally independent of the OS, and a properly tuned low-latency operating system is also required to achieve deterministic performance.

Summary of BIOS settings optimized for general-purpose workloads

Table 3. BIOS recommendations for CPU-intensive, energy-efficient, and low-latency workloads

BIOS token	Platform default	CPU-intensive	Energy-efficient	Low-latency
Processor configuration				
Intel Hyper-Threading Technology	Enabled	Platform default	Platform default	Disabled
CPU performance	Custom	Platform default	Platform default	Platform default
Hardware prefetcher	Enabled	Platform default	Platform default	Platform default
Adjacent cache line prefetcher	Enabled	Disabled	Disabled	Disabled
DCU IP prefetcher	Enabled	Platform default	Platform default	Platform default
DCU streamer prefetch	Enabled	Disabled	Disabled	Disabled
LLC prefetch	Enabled	Platform default	Platform default	Platform default

BIOS token	Platform default	CPU-intensive	Energy-efficient	Low-latency
Intel Virtualization Technology	Enabled	Platform default	Platform default	Disabled
Uncore configuration				
UPI link enablement	Auto	Platform default	Platform default	Platform default
UPI power management	Disabled	Platform default	Platform default	Platform default
UPI Link Frequency Select	Auto	Platform default	Platform default	Platform default
XPT remote prefetch	Auto	Disabled	Platform default	Platform default
KTI prefetch	Auto	Disabled	Platform default	Platform default
XPT prefetch	Auto	Disabled	Platform default	Platform default
Sub-NUMA clustering	Disabled	Enabled	Platform default	Platform default
LLC dead line	Enabled	Enabled	Platform default	Platform default
Memory configuration				
NUMA Optimized	Enabled	Platform default	Platform default	Platform default
Virtual NUMA	Disabled	Platform default	Platform default	Platform default
Memory RAS configuration setup				
Select Memory RAS Configuration	ADDDC Sparing	Maximum Performance	Platform default	Platform default
Patrol scrub	Enable at End of POST	Disabled	Platform default	Platform default
Power and performance configuration				
Enhanced CPU performance*	Disabled	Auto	Platform default	Platform default
Energy-efficient turbo mode	Disabled	Enabled	Enabled	Platform default
Intel Turbo Boost Technology	Enabled	Platform default	Platform default	Disabled
Processor C6 Report	Auto	Disabled	Enabled	Enabled
Processor C1E	Enabled	Platform default	Platform default	Platform default
Package C-state control	C0/C1 State	Platform default	C6 non-retention	Platform default
Power-performance tuning	OS	BIOS	BIOS	BIOS

BIOS token	Platform default	CPU-intensive	Energy-efficient	Low-latency
Processor EPP profile	Balanced Perforrmance	Performance	Power	Balanced Power
Latency optimized mode	Disabled	Enabled	Platform default	Platform default
Workload configuration	I/O Sensitive	Balanced	Balanced	Balanced

Note:

- From Table 3. Enhanced CPU performance*: this BIOS feature is mostly applicable for CPU-intensive workloads and benchmarks. When this option is enabled, we highly recommend setting the fan policy at Maximum Power. This BIOS feature is not applicable for the Cisco UCS C220 M8 Rack Server.
- Default BIOS options are generally selected to produce the best overall performance for typical workloads. However, typical workloads differ from end user to end user; therefore, the default settings may not be the best choice for your specific workloads.

Additional BIOS recommendations for enterprise workloads

This section summarizes optimal BIOS settings for enterprise workloads:

- Virtualization and container-based workloads
- Relational database (Oracle and SQL) workloads
- Data analytics (big data) workloads
- Analytical database systems (SAP HANA) workloads
- AI/ML and High-Performance Computing (HPC) workloads

The following sections describe each enterprise workload.

Virtualization and Container-base workloads

Virtualization workloads

Intel Virtualization Technology provides manageability, security, and flexibility in IT environments that use software-based virtualization solutions. With this technology, a single server (the base server) can be partitioned, and each partition can then be projected as an independent server; this allows the base server to run different applications on the operating system simultaneously. It is important to enable Intel Virtualization Technology in the BIOS to support virtualization workloads.

The CPUs that support hardware virtualization allow the processor to run multiple operating systems in the virtual machines. This feature involves some overhead because the performance of a virtual operating system is comparatively slower than that of the native OS.

Container workloads

Containerizing an application platform and its associated dependencies abstracts the underlying infrastructure and OS differences for efficiency. Each container is bundled into one package containing an entire runtime environment, including an application with all its dependencies, libraries and other binaries, and configuration files needed to run that application. Containers running applications in a production environment need management to ensure consistent uptime. If a container goes down, then another container needs to start automatically.

Workloads that scale and perform well on bare metal should see a similar scaling curve in a container environment with minimal performance overhead. Some containerized workloads can even see close to 0% performance variance compared to bare metal. Large overhead generally means that application settings and/or container configuration are not optimally set. These topics are beyond the scope of this tuning guide. However, the CPU load balancing behavior of Kubernetes or other container orchestration platform scheduler may assign or load balance containerized applications differently than in a bare metal environment.

Relational database workloads

Relational database systems, also known as Online Transaction Processing (OLTP) systems, contain the operational data needed to control and run important transactional business tasks. These systems are characterized by their ability to complete various concurrent database transactions and process real-time data. They are designed to provide optimal data-processing speed.

These database systems are often decentralized to avoid single points of failure. Spreading the work over multiple servers can also support greater transaction-processing volume and reduce response time. In a virtualized environment, when the OLTP application uses a direct I/O path, make sure that the Intel VT for Directed I/O option is enabled. By default, this option is enabled.

Data analytics workloads

Data analytics applications are important because they help businesses optimize their performance. Implementing data analytics in the business model can help organizations reduce costs by identifying more efficient ways of doing business and by storing large amounts of data. A company can also use data analytics to make better business decisions and help analyze customer trends and satisfaction, which can lead to new—and better—products and services.

Big-data analytics use advanced analytics techniques on very large, diverse big-data sets that include structured, semistructured, and unstructured data, from any source. These data sets can be defined as ones whose size or type is beyond the ability of traditional relational databases to capture, manage, and process with low latency. In addition, new capabilities include real-time streaming analytics and impromptu, iterative analytics on enormous data sets.

Analytical database systems workloads

An analytical database, also called an analytics database, is a read-only system that stores historical data about business metrics such as sales performance and inventory levels. Business analysts, corporate executives, and other workers run queries and reports against analytics databases. The information is regularly updated to include recent transaction data from an organization's operational systems.

An analytics database is specifically designed to support Business Intelligence (BI) and analytics applications, typically as part of a data warehouse or data mart. This feature differentiates it from an operational, transactional, or OLTP database; the latter databases are used to process transactions such as order entries and other business applications.

The SAP HANA platform is a flexible data-source–independent in-memory data platform that allows you to analyze large volumes of data in real time. Using the database services of the SAP HANA platform, you can store and access data in memory and using columns SAP HANA allows OLTP and online analytical processing (OLAP) in one system, without the need for redundant data storage or aggregates. Using the application services of the SAP HANA platform, you can develop applications, run your custom applications built on SAP HANA, and manage your application lifecycles.

High-Performance Computing (HPC) workloads

AI/ML workloads

AI workloads commonly involve intensive mathematical operations such as matrix multiplications, vector calculations, and optimization routines, which are integral to machine-learning algorithms and neural networks. These tasks are highly parallelizable, making them well-suited for acceleration through specialized hardware such as GPUs and multi-core CPUs. Leveraging such hardware allows for significant reductions in training and inference times, enabling more complex models and larger datasets to be processed efficiently.

For AI inference tasks, achieving low latency and high throughput is crucial, especially in real-time applications. Adjusting BIOS settings can help optimize hardware performance. For example:

- **Disabling hyperthreading:** this can sometimes improve inference performance by reducing resource contention, especially if hyperthreading causes context-switching overhead or cache thrashing.
- **Using dedicated AI optimization profiles:** many modern CPUs and GPUs offer profiles or features (for example, Intel's Deep Learning Boost or NVIDIA's Tensor Cores) that enhance AI workload efficiency. Enabling these features or using BIOS/firmware settings optimized for AI tasks can lead to substantial speedups.

Additionally, configuring system-power settings for maximum performance, ensuring proper thermal management, and using optimized inference frameworks (such as NVIDIA's TensorRT, OpenVINO, or ONNX Runtime) can further boost inference speed.

HPC workloads

HPC refers to cluster-based computing that uses multiple individual nodes that are connected and that work in parallel to reduce the amount of time required to process large data-sets that would otherwise take exponentially longer to run on any one system. HPC workloads are computation-intensive and typically also network-I/O intensive. HPC workloads require high-quality CPU components and high-speed, low-latency network fabrics for their Message Passing Interface (MPI) connections.

Computing clusters include a head node that provides a single point for administering, deploying, monitoring, and managing the cluster. Clusters also have an internal workload management component, known as the scheduler, that manages all incoming work items (referred to as "jobs"). Typically, HPC workloads require large numbers of nodes with nonblocking MPI networks so they can scale. Scalability of nodes is the single most important factor in determining the achieved usable performance of a cluster.

HPC requires a high-bandwidth I/O network. When you enable DCA support, network packets go directly into the Layer-3 processor cache instead of the main memory. This approach reduces the number of HPC I/O cycles generated by HPC workloads when certain Ethernet adapters are used, which in turn increases system performance.

Summary of BIOS settings recommended for enterprise workloads

Table 4 summarizes the BIOS tokens and settings recommended for various enterprise workloads.

Table 4. BIOS options recommended for enterprise workloads

BIOS token	Platform default	Virtualization and container	RDBMS	Analytical database systems	AI/ML and HPC
Processor configuration					
Intel Hyper-Threading Technology	Enabled	Platform default	Platform default	Platform default	Disabled
CPU performance	Custom	Enterprise	Enterprise	Enterprise	HPC
Hardware prefetcher	Enabled	Platform default	Platform default	Platform default	Platform default
Adjacent cache line prefetcher	Enabled	Platform default	Platform default	Platform default	Platform default
DCU IP prefetcher	Enabled	Platform default	Platform default	Platform default	Platform default
DCU streamer prefetch	Enabled	Platform default	Platform default	Platform default	Platform default
LLC prefetch	Enabled	Platform default	Disabled	Disabled	Platform default
Intel Virtualization Technology	Enabled	Platform default	Disabled	Disabled	Disabled
Uncore configuration					
UPI link enablement	Auto	Platform default	Platform default	Platform default	Platform default
UPI power management	Disabled	Platform default	Platform default	Platform default	Platform default
UPI link speed	Auto	Platform default	Platform default	Platform default	Platform default
XPT remote prefetch	Auto	Platform default	Platform default	Platform default	Platform default
KTI prefetch	Auto	Platform default	Platform default	Platform default	Platform default
XPT prefetch	Auto	Platform default	Platform default	Platform default	Platform default
Sub-NUMA clustering	Disabled	Platform default	Enabled	Enabled	Platform default
LLC dead line	Enabled	Platform default	Disabled	Disabled	Disabled
Memory configuration					
NUMA Optimized	Enabled	Platform default	Platform default	Platform default	Platform default
Virtual NUMA	Disabled	Platform default	Platform default	Platform default	Platform default
Memory RAS configuration setup					

BIOS token	Platform default	Virtualization and container	RDBMS	Analytical database systems	AI/ML and HPC
Select Memory RAS Configuration	ADDDC Sparing	Platform default	Platform default	Platform default	Platform default
Patrol scrub	Enable at End of POST	Platform default	Disabled	Disabled	Platform default
Power and performance configuration					
Enhanced CPU performance*	Disabled	Platform default	Auto	Auto	Auto
Energy-efficient turbo mode	Disabled	Enabled	Enabled	Platform default	Platform default
Intel Turbo Boost Technology	Enabled	Platform default	Platform default	Platform default	Platform default
Processor C6 Report	Auto	Platform default	Platform default	Platform default	Platform default
Processor C1E	Enabled	Disabled	Disabled	Platform default	Platform default
Package C-state control	C0/C1 State	Platform default	Platform default	Platform default	Platform default
Power performance tuning	OS	BIOS	BIOS	BIOS	BIOS
Processor EPP Profile	Balanced Perforrmance	Performance	Performance	Performance	Performance
Latency-optimized mode	Disabled	Enabled	Enabled	Enabled	Enabled
Workload configuration	I/O Sensitive	Balanced	Platform default	Platform default	Platform default

Note:

- From Table 4. Enhanced CPU performance*: this BIOS feature is mostly applicable for CPU-intensive workloads & benchmarks. When this option is enabled, we highly recommend setting the Fan policy as Maximum power. This BIOS feature is not applicable for UCS C220 M8 rack server.
- Default BIOS options are generally selected to produce the best overall performance for typical workloads. However, typical workloads differ from end user to end user; therefore, the default settings may not be the best choice for your specific workloads.

Conclusion

When tuning system BIOS settings for performance, you need to consider a number of processor and memory options. If the best performance is your goal, be sure to choose options that optimize for performance in preference to power savings, and experiment with other options such as CPU prefetchers, CPU power management, and CPU hyperthreading.

For more information

For more information about the Intel Xeon 6 Processor Family, Cisco UCS BIOS tokens, Cisco UCS C-Series Rack M8 Servers, and Cisco UCS X-Series Modular M8 Servers, see the following resources:

- Intel Xeon 6 Processor Family brief: https://www.intel.com/content/www/us/en/products/details/processors/xeon/xeon6-product-brief.html.
- Cisco UCS BIOS token guide: https://www.cisco.com/c/en/us/td/docs/unified_computing/ucs/Intersight/IMM_BIOS_Tokens_Guide/b_IMM_Server_BIOS_Tokens_Guide/b_UCS_BIOS_Tokens_Guide_chapter_01.html.
- Cisco UCS C220 M8 Rack Server: https://www.cisco.com/c/en/us/products/collateral/servers-unified-computing/ucs-c-series-rack-servers/ucs-c220-m8-rack-server-aag.html.
- https://www.cisco.com/c/dam/en/us/products/collateral/servers-unified-computing/ucs-c-series-rack-servers/ucs-c220-m8-sff-rack-server.pdf.
- Cisco UCS C240 M8 Rack Server: https://www.cisco.com/c/en/us/products/collateral/servers-unified-computing/ucs-c-series-rack-servers/ucs-c240-m8-rack-server-aag.html.
- https://www.cisco.com/c/dam/en/us/products/collateral/servers-unified-computing/ucs-c-series-rack-servers/ucs-c240-m8-sff-rack-server.pdf.
- Cisco UCS X210c M8 Compute Node: https://www.cisco.com/c/en/us/products/collateral/servers-unified-computing/ucs-x-series-modular-system/ucs-x210c-m8-compute-node-aag.html.
- https://www.cisco.com/c/dam/en/us/products/collateral/servers-unified-computing/ucs-x-series-modular-system/x210cm8-specsheet.pdf.

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 **https://www.cisco.com/go/offices**.

Printed in USA C11-5169901-01 07/25