White paper
Cisco public

The bridge to possible

Performance Tuning Guide for Cisco UCS M6 Servers

Using 3rd Gen Intel Xeon Scalable Processors

Contents

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®) M6 server generation of the following servers: Cisco UCS B200 M6 Blade Server, X210c M6 Compute Node, C220 M6 Rack Server, and C240 M6 Rack Server. All servers use third-generation (3rd Gen) Intel® Xeon® Scalable processors. The document describes how to optimize the BIOS settings to meet requirements for the best performance and energy efficiency for the Cisco UCS M6 generation of blade and rack servers.

With the release of the 3rd Gen Intel Xeon Scalable processor family (architecture code named Ice Lake), Cisco released sixth-generation Cisco UCS servers to take advantage of the increased number of cores, higher memory speeds, and PCIe 4.0 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 M6 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 and rack M6 servers that use 3rd Gen Intel Xeon Scalable 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 M6 servers based on the 3rd Gen Intel Xeon Scalable processor family deliver the highest levels of performance, as demonstrated in numerous industry-standard benchmark publications from the Standard Performance Evaluation Corporation (SPEC) and the Transaction Processing Performance Council (TPC).

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.

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

Intel SpeedStep Technology is designed to save energy by adjusting the CPU clock frequency up or down depending on how busy the system is. 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.

You can specify whether the processor uses 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 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.

CPU performance

Intel Xeon 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 (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 Scalable processor has shown that most applications run best with all prefetchers enabled. See Tables 2 and 3 for guidance.

Table 1. CPU performance and prefetch options from Intel

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

Table 2. Cisco UCS CPU performance options

Option	Description
CPU performance	Sets the CPU performance profile for the server. This can be one of the following: • Enterprise/HPC: All prefetchers are enabled. This is the platform-default setting for M6 servers. • High throughput: The DCU IP prefetcher is enabled, and all other prefetchers are disabled. • Custom: Allow users to choose the desired prefetcher settings depending on workloads. • Platform default: The BIOS uses the value for this attribute contained in the BIOS defaults for the server type and vendor.

Table 3. Cisco UCS CPU prefetcher options and target benchmarks and workloads

Prefetchers	Target benchmarks and workloads
All enabled	HPC benchmarks, web server, analytical database, virtualization, and relational database systems
DCU-IP enabled; all others disabled	SPECjbb2015 benchmark and certain server-side Java application-server applications

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.

By default, the LLC prefetch option is enabled.

Intel VT for Directed I/O

You can specify whether the processor uses Intel Virtualization Technology (VT) for Directed I/O (VT-d), which allows a platform to run multiple operating systems and applications in independent partitions.

The setting can be either of the following:

- Disabled: The processor does not permit virtualization.
- 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 UPI bandwidth performance in exchange for less power consumption.

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

Intel UPI power management

The Intel UPI power management is used to conserve power on a platform. Low power mode reduces 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 it 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.

Intel UPI link frequency

The Intel UPI link frequency 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

The Sub-NUMA Clustering (SNC) BIOS option provides localization benefits similar to the Cluster-on-Die (CoD) option, without some of the disadvantages of CoD. SNC is a replacement for the CoD feature found in previous processor families. SNC (two-way sub-NUMA) divides the LLC into two disjointed clusters called NUMA nodes, and it 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 to correspond with the SNC setting. OS support that recognizes each cluster and a separate NUMA node is 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.

Note: When SNC is selected, the operating system discovers each physical CPU socket as two NUMA nodes, except for 3rd Gen Intel Xeon Scalable processors with fewer than 12 cores, for which SNC is not supported. Refer to your OS documentation to determine whether SNC is supported.

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.

Intel UPI prefetch

Intel UPI prefetch is a new capability that is designed to reduce remote memory access latency. The UPI controller issues a UPI prefetch command, also in parallel with an LLC lookup, to the memory controller when a remote read request arrives in the home socket.

Extended prediction table remote prefetch

The XPT remote prefetch 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 dead line

With the Intel Xeon Scalable processors' noninclusive cache scheme, MLC evictions are filled into the 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 dead lines 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, dead lines 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.

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.

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.

Package C-state limit

When power technology is set to Custom, 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 Nonretention: 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.

Autonomous core C-state

When the operating system requests CPU core C1 state, system hardware automatically changes the request to the core C6 state.

The setting can be either of the following:

- Enabled: HALT and C1 requests are converted to C6 requests in hardware.
- Disabled: Only C0 and C1 are used by the OS. C1 is enabled automatically when an OS autohalts. By default, the autonomous core C-state setting is disabled.

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. By default, this option is disabled.

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 performance, but which 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 only to Cisco UCS C-Series Rack Servers.

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.

Memory settings

You can use several settings to optimize memory performance.

Memory reliability, availability, and serviceability configuration

Always set the memory reliability, availability, and serviceability (RAS) configuration to Maximum Performance for systems that require the highest performance and do not require memory fault-tolerance options.

The following settings are available:

- Maximum Performance: System performance is optimized.
- Mirror Mode 1LM (one-level memory): System reliability is optimized by using half the system memory as backup.

Note: For the optimal balance of performance and system stability, you should use the platform default (adaptive double device data correction [ADDDC] sparing enabled) for the memory RAS configuration. ADDDC sparing will incur a small performance penalty. If maximum performance is desired independently of system stability the Maximum Performance memory RAS setting can be used.

Nonuniform memory access

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.

Integrated memory controller interleaving

The Integrated Memory Controller (IMC) BIOS option controls the interleaving between the integrated memory controllers. There are two integrated memory controllers per CPU socket in an x86 server running Intel Xeon Scalable processors. If integrated memory controller interleaving is set to 2-way, addresses will be interleaved between the two integrated memory controllers. If integrated memory controller interleaving is set to 1-way, there will be no interleaving.

The following settings are available:

- 1-way Interleave: There is no interleaving.
- 2-way Interleave: Addresses are interleaved between the two integrated memory controllers.
- Auto: The CPU determines the integrated memory controller interleaving mode.

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.

Adaptive double device data correction sparing

Adaptive Double Device Data Correction (ADDDC) is a memory RAS feature that enables dynamic mapping of failing DRAM by monitoring corrected errors and taking action before uncorrected errors can occur and cause an outage. It is now enabled by default.

After ADDDC sparing remaps a memory region, the system could incur marginal memory latency and bandwidth penalties on memory bandwidth intense workloads that target the affected region. Cisco recommends scheduling proactive maintenance to replace a failed DIMM after an ADDDC RAS fault is reported.

Partial cache line sparing

The Partial Cache Line Sparing (PCLS) BIOS option provides an error-prevention mechanism in memory controllers. PCLS statically encodes the locations of the faulty nibbles of bits into a sparing directory along with the corresponding data content for replacement during memory accesses. By default, this option is enabled.

Memory refresh rate

This BIOS option controls the refresh rate of the memory controller and may affect the performance and resiliency of the server memory. This option sets the memory refresh rate to either 1x Refresh or 2x Refresh. By default, 2X Refresh is enabled.

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.
- Enabled: 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.

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 the 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 this behavior. 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 results in short duration transient performance impacts.
  Acoustic mode is available only on the Cisco UCS C220 M6 Server, Cisco UCS C240 M6 Server, Cisco UCS C240 SD M6 Server.

Refer below steps to configure FAN policy on M6 servers:

- For standalone Cisco UCS C-Series M6 servers using the Cisco® Integrated Management Controller (IMC) console and the Cisco IMC supervisor. From the Cisco IMC web console, choose Compute > Power Policies > Configured Fan Policy.
- For Cisco UCS managed C-Series M6 servers, this policy is configurable using power control policies under Servers > Policies > root > Power Control Policies > Create Fan Power Control Policy > Fan Speed Policy.
- For Cisco Intersight™ managed C-Series M6 servers, the fan control policy is defined in Intersight via the Thermal policy using the Fan Control Mode object.
- For UCS B-series and X-series servers, the fan speeds are dynamically adjusted based on the resource usage.

BIOS settings for Cisco UCS M6 servers

Table 4 lists the BIOS token names, defaults, and supported values for the Cisco UCS M6 blade and rack servers for 3rd Gen Intel Xeon Scalable processors.

Table 4. BIOS token names and supported values

BIOS token	Platform default	Supported values
Processor configuration		
Intel Hyper-Threading Technology	Enabled	Enabled and Disabled
Intel Virtualization Technology	Enabled	Enabled and Disabled
CPU performance	Custom	Enterprise, High-Throughput, HPC, and Custom
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 VT for Directed I/O	Enabled	Enabled and Disabled

BIOS token	Platform default	Supported values
Power and performance configuration		
Enhanced CPU performance	Disabled	Auto and Disabled
Intel Turbo Boost Technology	Enabled	Enabled and Disabled
Processor C1E	Disabled	Enabled and Disabled
Processor C6 Report	Disabled	Enabled and Disabled
Energy Efficient turbo	Disabled	Enabled and Disabled
Energy performance tuning	OS	OS, BIOS, and PECI
Package C-state limit	C0/C1 State	No Limit, Auto, C0/C1 State, C2, C6 Retention, and C6 Nonretention
Autonomous core C-state	Disabled	Enabled and Disabled
Energy and performance BIOS configuration	Balanced Performance	Performance, Balanced Performance, Balanced Power, and Power
Workload configuration	I/O Sensitive	Balanced and I/O Sensitive
UPI prefetch	Auto	Auto, Enabled, and Disabled
XPT prefetch	Auto	Auto, Enabled, and Disabled
XPT remote prefetch	Auto	Auto, Enabled, and Disabled
UPI link enablement*	Auto	Auto, 1, and 2
UPI power management*	Disabled	Enabled and Disabled
UPI link speed*	Auto	Auto, 9.6 GTs, 10.4 GTs, and 11.2 GTs
Sub-NUMA clustering	Disabled	Auto, Enabled, and Disabled
Uncore frequency scaling	Enabled	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
IMC interleaving	Auto	Auto, 1-way Interleave, and 2-way Interleave
Memory RAS configuration	ADDDC Sparing	Mirror Mode 1LM, ADDDC Sparing, Partial Mirror Mode 1LM, and Maximum Performance
Virtual NUMA*	Disabled	Enabled and Disabled
Memory refresh rate	2x Refresh	1x Refresh and 2x Refresh
Patrol scrub	Enabled	Enabled and Disabled

Note: BIOS tokens with an asterisk* marked are available and supported only on M6 servers with 3rd Gen Intel Xeon Scalable processors.

BIOS recommendations for various general-purpose workloads

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

- CPU-intensive workloads
- I/O-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.

I/O-intensive workloads

I/O-intensive optimizations are configurations that depend on maximum throughput between I/O and memory. Processor utilization-based power management features affect performance on the links between I/O and memory are disabled.

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 the general application performances rather than power efficiency. Processor power management settings can affect performance when virtualization operating systems are used. Hence, these settings are recommended for customers that do not typically tune the BIOS for their workloads.

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 5 summarizes BIOS settings optimized for general-purpose workloads.

Table 5. BIOS recommendations for CPU-intensive, I/O-intensive, energy-efficient, and low-latency workloads

BIOS tokens	BIOS values (platform defaults)	CPU intensive	I/O intensive	Energy efficient	Low latency
Processor configuration					
Intel Hyper-Threading Technology	Enabled	Platform default	Platform default	Platform default	Platform default
Intel Virtualization Technology	Enabled	Platform default	Platform default	Platform default	Disabled
CPU performance	Custom	Platform default	Platform default	Platform default	Enterprise
Hardware prefetcher	Enabled	Platform default	**Disabled**	**Disabled**	Platform default
Adjacent cache line prefetcher	Enabled	**Disabled**	**Disabled**	**Disabled**	Platform default
DCU IP prefetcher	Enabled	Platform default	Platform default	**Disabled**	Platform default
DCU streamer prefetch	Enabled	**Disabled**	**Disabled**	**Disabled**	Platform default
LLC prefetch	Enabled	**Disabled**	**Disabled**	Platform default	Platform default
Intel VT for Directed I/O	Enabled	Platform default	Platform default	Platform default	Disabled

BIOS tokens	BIOS values (platform defaults)	CPU intensive	I/O intensive	Energy efficient	Low latency
Power and performance configuration					
Enhanced CPU performance*	Disabled	**Auto**	Platform default	Platform default	Platform default
Intel Turbo Boost Technology	Enabled	Platform default	Platform default	Platform default	**Disabled**
Processor C1E	Disabled	**Enabled**	Platform default	**Enabled**	Platform default
Processor C6	Disabled	**Enabled**	Platform default	**Enabled**	Platform default
Energy Efficient turbo	Disabled	Platform default	Platform default	**Enabled**	**Enabled**
Energy-performance tuning	OS	Platform default	Platform default	Platform default	Platform default
Package C-state limit	C0/C1 State	Platform default	Platform default	**C6 Non-Retention**	**Platform default**
Workload configuration	I/O Sensitive	**Balanced**	Platform default	**Balanced**	**Balanced**
UPI prefetch	Auto	**Enabled**	Platform default	Platform default	Platform default
XPT prefetch	Auto	**Enabled**	Platform default	Platform default	Platform default
UPI link enablement	Auto	**1**	Platform default	Platform default	Platform default
UPI power management	Disabled	**Enabled**	Platform default	Platform default	Platform default
UPI link speed	Auto	Platform default	Platform default	Platform default	Platform default
Sub-NUMA clustering	Disabled	**Enabled**	Platform default	Platform default	Platform default
Uncore frequency scaling	Enabled	**Disabled**	Platform default	Platform default	Platform default
LLC dead line	Enabled	**Disabled**	Platform default	Platform default	Platform default
Memory configuration					
NUMA optimized	Enabled	Platform default	Platform default	Platform default	Platform default
IMC interleaving	Auto	**1-way Interleave**	Platform default	Platform default	Platform default
Memory RAS configuration	ADDDC Sparing	**Maximum Performance**	Platform default	Platform default	Platform default
Memory refresh rate	2x Refresh	**1x Refresh**	Platform default	Platform default	1x Refresh
Patrol scrub	Enabled	**Disabled**	Platform default	Platform default	Disabled

Note:

- From Table 5. Enhanced CPU Performance* - This feature is currently available on C-series only. However, it is being extended to B- and X-series. It may be better to state, that this BIOS token will be extended to these platforms at a later firmware release. This way you do NOT have to update this document when that happens.
- 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 choices for your specific workloads.

Additional BIOS recommendations for enterprise workloads

This section summarizes optimal BIOS settings for enterprise workloads:

- Relational database (Oracle and SQL) workloads
- Virtualization (virtual desktop infrastructure [VDI] and virtual server infrastructure [VSI]) workloads
- Data analytics (big data) workloads
- Analytical database systems (SAP HANA) workloads
- High-Performance Computing (HPC) workloads

The following sections describe each enterprise workload.

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.

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 can be partitioned and can be projected as several independent servers, allowing the 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.

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 is the use of 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, which are used to process transactions, such as order entry 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) on 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.

For more information about SAP HANA, see the SAP help portal: http://help.sap.com/hana/.

High-performance computing 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 that 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 6 summarizes the BIOS tokens and settings recommended for various enterprise workloads.

Table 6. BIOS options recommended for enterprise workloads

BIOS tokens	BIOS values (platform defaults)	Relational database systems	Virtualization	Analytical database systems	Data analytics	High-performance computing
Processor configuration						
Intel Hyper-Threading Technology	Enabled	Platform default	Platform default	Platform default	Platform default	Platform default
Intel Virtualization Technology	Enabled	Platform default	Platform default	Disabled	Platform default	Disabled
CPU performance	Custom	Platform default	Platform default	Platform default	Platform default	Platform default
Hardware prefetcher	Enabled	Platform default	Platform default	Platform default	Platform default	Platform default
Adjacent cache line prefetcher	Enabled	Platform default	Platform default	Platform default	Platform default	Platform default
DCU IP prefetcher	Enabled	Platform default	Platform default	Platform default	Platform default	Platform default
DCU streamer prefetch	Enabled	Platform default	Platform default	Platform default	Platform default	Platform default
LLC prefetch	Enabled	Disabled	Platform default	Platform default	Platform default	Disabled
Intel VT for Directed I/O	Enabled	Platform default	Platform default	Disabled	Disabled	Disabled
Power and performance configuration						
Enhanced CPU performance*	Disabled	Auto	Auto	Auto	Auto	Auto
Intel Turbo Boost Technology	Enabled	Platform default	Platform default	Platform default	Platform default	Platform default
Processor C1E	Disabled	Platform default	Enabled	Platform default	Enabled	Platform default
Processor C6	Disabled	Enabled	Enabled	Enabled	Enabled	Platform default
Energy Efficient turbo	Disabled	Platform default	Enabled	Enabled	Enabled	Platform default
Energy performance tuning	OS	Platform default	Platform default	Platform default	Platform default	Platform default
Package C-state limit	C0/C1 State	Platform	Platform	Platform	Platform	Platform default

BIOS tokens	BIOS values (platform defaults)	Relational database systems	Virtualization	Analytical database systems	Data analytics	High-performance computing
		default	default	default	default	
Workload configuration	I/O Sensitive	Platform default	Platform default	Balanced	Balanced	Balanced
UPI prefetch	Auto	Platform default	Platform default	Platform default	Platform default	Platform default
XPT prefetch	Auto	Enabled	Platform default	Platform default	Platform default	Enabled
UPI link enablement	Auto	Platform default	Platform default	Platform default	Platform default	Platform default
UPI power management	Disabled	Enabled	Platform default	Platform default	Platform default	Enabled
UPI link speed	Auto	Platform default	Platform default	Platform default	Platform default	Platform default
Sub-NUMA clustering	Disabled	Enabled	Platform default	Platform default	Platform default	Enabled
Uncore frequency scaling	Enabled	Platform default	Platform default	Platform default	Platform default	Platform default
LLC dead line	Enabled	Disabled	Platform default	Platform default	Platform default	Disabled
Memory configuration						
NUMA optimized	Enabled	Platform default	Platform default	Platform default	Platform default	Platform default
IMC interleaving	Auto	1-way Interleave	Platform default	Platform default	Platform default	1-way Interleave
Memory RAS configuration	ADDDC Sparing	Platform default	Platform default	Platform default	Platform default	Platform default
Memory refresh rate	2x Refresh	1x Refresh	Platform default	1x Refresh	Platform default	1x Refresh
Patrol scrub	Enabled	Disabled	Platform default	Disabled	Platform default	Disabled

Note:

- From Table 6. Enhanced CPU Performance* - This feature is currently available on C-series only. However, it is being extended to B- and X-series. It may be better to state, that this BIOS token will be extended to these platforms at a later firmware release. This way you do NOT have to update this document when that happens.

- 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 choices for your specific workloads.

Operating system tuning guidance for best performance

With OS tuning, the operating system controls power management to achieve the best performance. In Linux, these optimizations apply to the cpuspeed utility and cpufreq governor. The performance profile optimizes only for performance. Most CPU power management options are turned off.

For Linux, set the following:

- **cpupower frequency-set -governor performance**

  The performance governor forces the CPU to use the highest possible clock frequency. This frequency is statically set and will not change. Therefore, this particular governor offers no power-savings benefit. It is suitable only for hours of heavy workload, and even then, only during times in which the CPU is rarely (or never) idle. The default setting is On Demand, which allows the CPU to achieve maximum clock frequency when the system load is high, and the minimum clock frequency when the system is idle. Although this setting allows the system to adjust power consumption according to system load, it does so at the expense of latency from frequency switching.

Use the following Linux tools to measure maximum turbo frequency and power states:

- CPUpower monitor: The CPUpower monitor reports the processor topology and frequency and idle power state statistics. The command is forked, and statistics are printed upon the command's completion, or statistics are printed periodically. The command implements independent processor sleep state and frequency counters. Some are retrieved from kernel statistics, and some are read directly from hardware registers. Use this setting:

  cpupower monitor -l

Refer the following resources for more information about OS and Hypervisor performance tuning recommendations:

- Microsoft Windows and Hyper-V server platform tuning is straightforward: "Set the power policy to High Performance". See: https://docs.microsoft.com/en-us/windows-server/administration/performance-tuning/additional-resources
- VMware ESXi tuning is straightforward as well: "Set the power policy to High Performance". See: https://www.vmware.com/content/dam/digitalmarketing/vmware/en/pdf/techpaper/performance/vsphere-esxi-vcenter-server-70-performance-best-practices.pdf
- Citrix XenServer, set xenpm set-scaling-governor performance. See: https://support.citrix.com/article/CTX200390
- Red Hat Enterprise Linux, set "cpupower to Performance". See: https://access.redhat.com/documentation/en-us/red_hat_enterprise_linux/7/html/performance_tuning_guide/index
- SUSE Enterprise Linux, set "cpupower to Performance". See: https://documentation.suse.com/sles/15-SP2/pdf/book-sle-tuning_color_en.pdf

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 Cisco UCS B-Series, C-Series, and X-Series M6 servers, see the following resources:

- Cisco UCS B200 M6 Blade Server: https://www.cisco.com/c/dam/en/us/products/collateral/servers-unified-computing/ucs-b-series-blade-servers/b200m6-specsheet.pdf
- Cisco UCS C220 M6 Rack Server: https://www.cisco.com/c/dam/en/us/products/collateral/servers-unified-computing/ucs-c-series-rack-servers/c220m6-sff-specsheet.pdf
- Cisco UCS C240 M6 Rack Server: https://www.cisco.com/c/dam/en/us/products/collateral/servers-unified-computing/ucs-c-series-rack-servers/c240m6-lff-specsheet.pdf
- Cisco UCS X210c M6 Compute Node: https://www.cisco.com/c/dam/en/us/products/collateral/servers-unified-computing/ucs-x-series-modular-system/x210c-specsheet.pdf
- 3rd Gen Intel Xeon Scalable Processors Brief: https://www.intel.com/content/www/us/en/products/docs/processors/xeon/3rd-gen-xeon-scalable-processors-brief.html

Americas Headquarters
Cisco Systems, Inc.
San Jose, CA

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

Europe Headquarters
Cisco Systems International BV Amsterdam,
The Netherlands

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

Printed in USA C11-2950263-00 03/22