Performance Tuning Guide for Cisco UCS M4 Servers

For Cisco UCS Blade and Rack Servers Using Intel Xeon Processor E5 v4 and E7 v4 Family CPUs
Contents

Purpose and Scope
What You Will Learn
BIOS Tuning Scenarios
  High Performance
  Low Latency
Cisco UCS BIOS Options
  Processor Configuration: Intel SpeedStep and Turbo Boost Technologies
  Processor C3 and C6 States
  CPU Hyperthreading
  Core Multiprocessing and Latency-Sensitive Single-Threaded Applications
  Energy or Performance Bias
  Power Technology Setting
  CPU Prefetcher Settings
  Memory Performance Settings
  CPU Snoop Settings

BIOS Settings for Various Workload Types
  Online Transaction Processing
  Virtualization
  High-Performance Computing
  Java Enterprise Edition Application Server
  Analytic Database Systems

Configuring the BIOS for Optimized CPU Hardware Power Management
  Recommended Power Management Settings to Optimize Power Efficiency
  Operating System Resources for Power Management
  Operating System Tuning Guidance for Maximum Performance
  Operating System Tuning Resources for Performance

Conclusion

For More Information
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®) M4 server generation (Cisco UCS B200 M4, B260 M4, B420 M4 and B460 M4 Blade Servers and C220 M4, C240 M4, and C460 M4 Rack Servers) using Intel® Xeon® processor E5 v4 and E7 v4 family CPUs. It describes how to optimize the BIOS settings to meet requirements for best performance and energy efficiency for the Cisco UCS M4 generation of blade and rack servers.

This document also discusses the BIOS settings that can be selected for various workload types on Cisco UCS M4 servers that use Intel Xeon processor E5 v4 and E7 v4 family CPUs. 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 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 servers that use Intel Xeon processor E5 v4 and E7 v4 family CPUs.

BIOS Tuning Scenarios
This document focuses on two main scenarios: how to tune the BIOS for high performance and how to tune it for low latency.

High Performance
With the latest multiprocessor, multicore, and multithreading technologies in conjunction with current operating systems and applications, today’s Cisco UCS servers based on the Intel Xeon processor E5 v4 and E7 v4 generations of CPUs deliver the highest levels of performance, as demonstrated by the numerous industry-standard benchmark publications from the Standard Performance Evaluation Corporation (SPEC), SAP, 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 in most cases associated with increased consumption of electrical power. This document explains how to configure the BIOS settings to achieve optimal computing performance.

Low Latency
The BIOS offers a variety of options to reduce latency. In some cases, the corresponding application does not make efficient use of all the threads available in the hardware. To improve performance, you can disable threads that are not needed (hyperthreading) or even cores in the BIOS to reduce the small fluctuations in the performance
of computing operations that especially occur in some high-performance computing (HPC) applications and analytical database applications. Furthermore, by disabling cores that are not needed, you can improve turbo-mode performance in the remaining cores under certain operating conditions.

However, other scenarios require performance that is as constant as possible. Although the current generation of Intel processors delivers better turbo-mode performance than the preceding generation, the maximum turbo-mode frequency is not guaranteed under certain operating conditions. In such cases, disabling the turbo mode can help prevent changes in frequency.

Energy-saving functions, whose aim is to save energy whenever possible through frequency and voltage reduction and through the disabling of certain function blocks and components, also have a negative impact on response time. The higher the settings for the energy saving modes, the lower the performance. Furthermore, in each energy-saving mode, the processor requires a certain amount of time to change back from reduced performance to maximum performance.

This document explains how to configure the power and energy saving modes for users to reduce system latency. The optimization of server latency, particularly in an idle state, results in substantially greater consumption of electrical power.

**Cisco UCS BIOS Options**

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

**Processor Configuration: Intel SpeedStep and Turbo Boost Technologies**

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.

One new feature in the Intel Xeon processor E5 v4 & E7 v4 CPUs is the capability for each core to run at a different speed, using Intel SpeedStep. In prior generations, all cores on a chip ran at the same speed.

Intel Turbo Boost 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 C3 and C6 States**

The C3 and C6 states are power-saving halt and sleep states 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 have the option, disable anything related to C-states.

**CPU Hyperthreading**

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.
Core Multiprocessing and Latency-Sensitive Single-Threaded Applications

The core multiprocessing option is designed to enable the user to disable cores. This option may affect the pricing of certain software packages that are licensed by the core. You should consult your software license and software vendor about whether disabling cores qualifies you for any particular pricing policies. Set core multiprocessing to All if pricing policy is not an issue for you.

For latency-sensitive single-threaded applications, you can optimize performance by disabling unnecessary cores, disabling hyperthreading, enabling all C-states, enabling Intel SpeedStep, and enabling Intel Turbo Boost. With this configuration, the remaining cores often will benefit from higher turbo speeds and better use of the shared Layer 3 cache.

Energy or Performance Bias

You can use the power-saving mode to reduce system power consumption when the turbo mode is enabled. The mode can be set to Maximum Performance, Balanced Performance, Balanced Power, or Power Saver. Testing has shown that most applications run best with the Balanced Performance setting.

Power Technology Setting

For best performance, set the power technology option to Custom. If it is not set to Custom, the individual settings for Intel SpeedStep and Turbo Boost and the C6 power state are ignored.

CPU Prefetcher Settings

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 processor E5 v4 & E7 v4 CPUs has shown that most applications run best with all prefetchers enabled. See Table 2 for guidance.

<table>
<thead>
<tr>
<th>Performance Option</th>
<th>Cache Affected</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hardware prefetcher</td>
<td>Layer 2</td>
</tr>
<tr>
<td>Adjacent-cache-line prefetcher</td>
<td>Layer 2</td>
</tr>
<tr>
<td>DCU prefetcher</td>
<td>Layer 1</td>
</tr>
<tr>
<td>DCU instruction pointer (DCU-IP) prefetcher</td>
<td>Layer 1</td>
</tr>
</tbody>
</table>

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 through sequential data, such as database table scans and clustered index scans, or that run a tight loop in code.

Adjacent-Cache-Line Prefetcher (Buddy Fetch)

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 (or buddy) data or instructions.

**DCU Prefetcher**
Like the hardware prefetcher, the DCU 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.

**DCU-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.

**Data Reuse**
Enabling this option reduces the frequency of L3 cache updates from L1. This default setting is "Enabled". This may improve performance by reducing the internal bandwidth consumed by constantly updating L1 cache lines in L3. Since this results in more fetches to main memory, setting this option to Disabled may improve performance in some cases. Users should only disable this option after performing application benchmarking to verify improved performance in their environment. Data Reuse option is hidden not exposed in F2 BIOS.

**Table 2. Cisco UCS CPU Performance options**

<table>
<thead>
<tr>
<th>Name</th>
<th>Description</th>
</tr>
</thead>
</table>
| CPU Performance  | • Sets the CPU performance profile for the server. This can be one of the following:  
|                  |   • Enterprise - For M3 & M4 servers, all prefetchers and data reuse are enabled.  
|                  |   • high-throughput - Data reuse and the DCU IP prefetcher are enabled, and all other prefetchers are disabled.  
|                  |   • HPC - All prefetchers are enabled and data reuse is disabled. This setting is also known as high-performance computing.  
|                  |   • 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**

<table>
<thead>
<tr>
<th>Prefetchers</th>
<th>Target Benchmarks and Workloads</th>
</tr>
</thead>
<tbody>
<tr>
<td>All enabled</td>
<td>HPC benchmarks, webserver, SAP application server, virtualization, and TPC-E</td>
</tr>
<tr>
<td>DCU-IP enabled; all others disabled</td>
<td>SPECjbb2005 and certain server-side Java application-server applications</td>
</tr>
</tbody>
</table>

**Memory Performance 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.

**Nonuniform Memory Access**
Most modern operating systems, particularly virtualization hypervisors, support nonuniform memory access (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.
Isochronous Mode

Enabling the isochronous (ISOC) mode option reduces the credits available for memory traffic. For memory requests, this option reduces latency at the expense of throughput under heavy loads.

CPU Snoop Settings

When a system has more than one CPU, the chip set must work to maintain data consistency so that CPUs don’t simultaneously modify a data value and thus create inconsistent data. For instance, if a CPU is tasked with modifying a data value, it should first check with the other CPUs in the system to verify that they are not using the data value.

CPU snoop settings dictate the way that a system maintains this data consistency between two or more processors. Cisco UCS servers with E5 v4 & E7 v4 platform have three potential snoop settings are available: Early Snoop, Home Snoop, and Cluster on Die. The value for best performance depends on software support, the Intel Xeon processor E5 v4 & E7 CPU that is installed, the specific workload, and the NUMA settings.

Cluster-on-Die Snoop

Cluster-on-die (CoD) snoop is available on Intel Xeon processor E5 v4 & E7 v4 CPUs that have 10 or more cores. CoD snoop is the best setting to use when NUMA is enabled and the system is running a well-behaved NUMA application. A well-behaved NUMA application is one that generally accesses only memory attached to the local CPU.

CoD snoop provides the best overall latency and bandwidth performance for memory access to the local CPU. For access to remote CPUs, however, this setting results in higher latency and lower bandwidth. This snoop mode is not advised when NUMA is disabled.

Note: In the previous generation of Intel Xeon v2 processors, VMware vSphere does not support the CoD setting. With the current generation of Intel Xeon v3 and v4 processors, vSphere supports this setting.

Early Snoop

Early snoop (ES) is available on all Intel Xeon processor E5 v4 & E7 v4 CPUs. It provides good local memory latency and bandwidth, but with a severe penalty in bandwidth for remote CPU access. Early snoop should be used for latency-sensitive applications that do not require high remote bandwidth: for example, certain online transaction processing (OLTP) database workloads.

Home Snoop

Home snoop (HS) is also available on all Intel Xeon processor E5 v4 & E7 v4 CPUs and is excellent for NUMA applications that need to access a remote CPU on a regular basis. Of the snoop modes, home snoop provides the best remote CPU bandwidth and latency, but with the penalty of slightly higher local latency, and is the best choice for memory and bandwidth-intensive applications: for example, certain decision-support-system (DSS) database workloads.

Home Directory Snoop with Opportunistic Snoop Broadcast

Home directory snoop with opportunistic snoop broadcast (OSB) is available on all Intel Xeon processor E5 v4 & E7 v4 CPUs. In the OSB directory mode, the HA (Home Agent) could choose to do speculative home snoop broadcast under very lightly loaded conditions even before the directory information has been collected and checked.
BIOS Settings for Various Workload Types

This document discusses BIOS settings for the following types of workloads:

- Online transaction processing (OLTP)
- Virtualization
- High-performance computing (HPC)
- Java Enterprise Edition application server
- Analytical database systems

Table 4 summarizes the settings available.

**Table 4. BIOS Options for Different Workloads**

<table>
<thead>
<tr>
<th>BIOS Options</th>
<th>Online Transaction Processing</th>
<th>Virtualization</th>
<th>High-Performance Computing</th>
<th>Java Application Servers</th>
<th>Analytic Database Systems</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intel Turbo Technology</td>
<td>Enabled</td>
<td>Enabled</td>
<td>Enabled</td>
<td>Enabled</td>
<td>Enabled</td>
</tr>
<tr>
<td>Intel SpeedStep Technology</td>
<td>Enabled</td>
<td>Enabled</td>
<td>Enabled</td>
<td>Enabled</td>
<td>Enabled</td>
</tr>
<tr>
<td>Intel Hyper-Threading Technology</td>
<td>Enabled</td>
<td>Enabled</td>
<td>Disabled</td>
<td>Enabled</td>
<td>Enabled</td>
</tr>
<tr>
<td>Intel Virtualization Technology (VT)</td>
<td>Platform default</td>
<td>Enabled</td>
<td>Platform default</td>
<td>disabled</td>
<td>Platform default</td>
</tr>
<tr>
<td>Intel VT for Directed I/O (VT-d)</td>
<td>Enabled</td>
<td>Enabled</td>
<td>Platform default</td>
<td>Platform default</td>
<td>Platform default</td>
</tr>
<tr>
<td>CPU performance</td>
<td>Enterprise</td>
<td>Enterprise</td>
<td>HPC</td>
<td>Enterprise</td>
<td>HPC</td>
</tr>
<tr>
<td>Direct cache access</td>
<td>Enabled</td>
<td>Enabled</td>
<td>Enabled</td>
<td>Enabled</td>
<td>Enabled</td>
</tr>
<tr>
<td>Power technology</td>
<td>Performance</td>
<td>Performance</td>
<td>Performance</td>
<td>Custom</td>
<td>Custom</td>
</tr>
<tr>
<td>Processor C-states</td>
<td>All C-states disabled</td>
<td>All C-states disabled</td>
<td>All C-states enabled</td>
<td>All C-states disabled</td>
<td>All C-states disabled</td>
</tr>
<tr>
<td>Energy performance</td>
<td>Performance</td>
<td>Performance</td>
<td>Performance</td>
<td>Performance</td>
<td>Performance</td>
</tr>
<tr>
<td>Frequency-floor override</td>
<td>Enabled</td>
<td>Enabled</td>
<td>Enabled</td>
<td>Enabled</td>
<td>Enabled</td>
</tr>
<tr>
<td>DRAM clock throttling</td>
<td>Performance</td>
<td>Performance</td>
<td>Performance</td>
<td>Performance</td>
<td>Performance</td>
</tr>
<tr>
<td>Package C-state limit</td>
<td>Platform default</td>
<td>Platform default</td>
<td>Platform default</td>
<td>Platform default</td>
<td>Platform default</td>
</tr>
<tr>
<td>CPU hardware power management</td>
<td>Platform default</td>
<td>Platform default</td>
<td>Platform default</td>
<td>Platform default</td>
<td>Platform default</td>
</tr>
<tr>
<td>Memory RAS configuration</td>
<td>Maximum</td>
<td>Maximum</td>
<td>Maximum</td>
<td>Maximum</td>
<td>Maximum</td>
</tr>
<tr>
<td>Low-voltage double-data-rate (DDR) mode</td>
<td>Performance mode</td>
<td>Performance mode</td>
<td>Performance mode</td>
<td>Performance mode</td>
<td>Performance mode</td>
</tr>
<tr>
<td>DRAM refresh rate</td>
<td>Platform default</td>
<td>Platform default</td>
<td>Platform default</td>
<td>Platform default</td>
<td>Platform default</td>
</tr>
<tr>
<td>Intel QuickPath Interconnect (QPI) snoop mode</td>
<td>Early Snoop</td>
<td>Platform default</td>
<td>Platform default</td>
<td>Platform default</td>
<td>Home Snoop</td>
</tr>
</tbody>
</table>

The following sections describe the BIOS tuning recommendations for these workloads in detail.
Online Transaction Processing

OLTP systems contain the operational data 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.

OLTP 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.

Processor Settings for OLTP Workloads

Obtaining peak performance requires some system-level tuning. Figures 1 and 2 illustrate the selections that are recommended to optimize OLTP workloads on Cisco UCS B-Series and managed Cisco UCS C-Series M4 servers.

Figure 1. Processor Settings for OLTP Workloads
Figure 2. Processor Energy Settings for OLTP Workloads

The Intel Turbo Boost and SpeedStep technologies are powerful management features that adjust the CPU voltage and frequency settings to optimize performance and power consumption dynamically. During periods of low CPU consumption, Intel SpeedStep can reduce the CPU frequency by reducing power consumption. Intel Turbo Boost increases the processing speed to accommodate higher demand in situations in which CPU utilization is extremely high. Each core has 20 to 30 percent more processing capability when Intel Turbo Boost is enabled. For example, the Cisco UCS CPU with Intel Xeon processor E5 2697 v4 operates at a base frequency of 2.3 GHz. If Intel Turbo Boost is enabled, the system can achieve frequencies as high as 3.6 GHz.

When you tune for consistent performance for OLTP applications on a system that does not run at close to 100 percent CPU utilization, you should enable Intel SpeedStep and Turbo Boost and disable C-states. Although this configuration foregoes power savings during idle times, it keeps all CPU cores running at a consistent speed and delivers the most consistent and predictable performance.

Enabling Intel Hyper-Threading Technology helps OLTP systems handle I/O-intensive workloads by allowing the processing of multiple threads per CPU core. OLTP applications typically are multithreaded, with each thread performing a small amount of work that may include I/O operations. A large number of threads results in a considerable amount of context switching, but with Intel Hyper-Threading, the effect of context switching is reduced. When Intel Direct Cache Access is enabled (Figures 1 and 3), the I/O controller places data directly into the CPU cache to reduce the cache misses while processing OLTP workloads. This approach results in improved application performance.
If you are deploying the system in a virtualized environment and the OLTP application uses a directed I/O path, make sure to enable the VT for Directed IO option. With Cisco® Data Center Virtual Machine Fabric Extender (VM-FEX) technology, virtual machines can now directly write to the virtual network interface cards (vNICs) when directed I/O is enabled at the BIOS level.

**Note:** This feature is applicable only if the OLTP system is running in a virtualized environment.

Figure 4 shows the selections that are recommended for OLTP workloads on “Standalone” Cisco UCS C-Series M4 servers.
Figure 4. Processor and Energy Settings for OLTP Workloads

Memory Settings for OLTP Workloads

Figures 5 and 6 show memory settings for OLTP workloads.
OLTP applications have a random memory access pattern and benefit greatly from larger and faster memory. Therefore, Cisco recommends setting memory RAS features to maximum performance for optimal system performance. Also, DDR mode should be set to performance mode so that DIMMs run at the highest frequency for the installed memory and CPU combination. In OLTP transactions, if these modes are enabled, I/O operations will be serviced at the highest frequency and will have reduced memory latency.

**Note:** If the DIMM pairs in the server have the same type, size, and organization and are populated across the Scalable Memory Interconnect (SMI) channels, you can enable the lockstep mode, an option on the Select Memory RAS menu, to reduce memory-access latency and achieve better performance.

The DRAM refresh rate feature allows you to set the refresh interval of the memory chips. For OLTP applications, the platform default setting is recommended.
Virtualization

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.

Processor Settings for Virtualization Workloads

Figures 7 and 8 show the selections that are recommended for virtualized workloads on Cisco UCS B-Series and managed Cisco UCS C-Series m4 servers.

Figure 7. Processor Settings for Virtualized Workloads
Figure 8. Processor Energy Settings for Virtualized Workloads

![Unified Computing System Manager](image)

Figure 9 shows the selections that are recommended for virtualized workloads on “Standalone” Cisco UCS C-Series M4 servers.
Figure 9. Processor and Energy Settings for Virtualized Workloads

Most of the CPU and memory settings for virtualized workloads are the same as for OLTP workloads. It is important to enable Intel Virtualization Technology in the BIOS to support virtualization workloads. Make sure that the Intel VT-d options are enabled.
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. To enhance performance, be sure to enable Intel Turbo Boost and Hyper-Threading for the processors.

The cache prefetching mechanisms (data-prefetch-logic [DPL] prefetch, hardware prefetch, Layer 2 streaming prefetch, and adjacent-cache-line prefetch) usually help increase system performance, especially when memory-access patterns are regular. Because virtualized workloads have random memory access, you should disable the prefetchers.

Intel Directed I/O for Virtualized Workloads
Figure 10 shows the recommended Intel Directed I/O settings for virtualized workloads.

**Figure 10. Intel Directed I/O Settings for Virtualized Workloads in Cisco UCS B-Series Blade and Managed Cisco UCS C-Series Rack M4 Servers**

With Cisco’s VM-FEX technology, virtual machines can now directly write to the vNICs when the directed I/O option is enabled at the BIOS level.

Memory Settings for Virtualized Workloads
Figures 11 and 12 show the recommended memory settings for virtualized workloads.
When running applications that access memory randomly, set the Select Memory RAS option to Maximum Performance. This setting helps achieve optimal system performance. In virtualized environments, run the memory at the highest frequency to reduce memory latency.

**High-Performance Computing**

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.

Processor Settings for HPC Workloads

Figures 13 and 14 show the selections that are recommended for HPC workloads on Cisco UCS B-Series and managed Cisco UCS C-Series M4 servers.

**Figure 13.** Processor Settings for HPC Workloads
Figure 14. Processor Energy Settings for HPC Workloads

Figures 15 and 16 show the selections that are recommended for HPC workloads on “standalone” Cisco UCS C-Series M4 servers.

Figure 15. Processor and Energy Settings for HPC Workloads on “Standalone” Cisco UCS C-Series M4 Servers
You should enable Intel Turbo Boost technology for HPC workloads to increase the computing power. When Intel Turbo Boost is enabled, each core provides higher computing frequency potential, allowing a greater number of parallel requests to be processed efficiently.
Intel SpeedStep is enabled because it is required for Intel Turbo Boost to function.

HPC workloads typically do not benefit from Intel Hyper-Threading. Additional threads only serve to create resource contention within the microarchitecture of the CPU. Generally, Intel Hyper-Threading has the greatest impact on workloads in which threads are forced to wait for completion of back-end I/O requests, to reduce thread contention for CPU resources.

Enabling the processor power state C6 helps save power when the CPU is idle. Because HPC is computing intensive, the CPU will likely seldom go into an idle state. However, enabling C-states saves CPU power in the event that there are any inactive requests.

You should set CPU performance to HPC mode to handle more random, parallel requests by HPC applications. If HPC performs more in-memory processing (for example, for video data), you should enable the prefetcher options so that they can handle multiple parallel requests. This configuration also helps retain some hot data in the Layer 2 cache, and it improves HPC performance (CPU performance).

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.

You can set the Energy Performance option to Maximum Performance, Balanced Performance, Balanced Power, or Power Saver (see Figure 16). Test results demonstrate that most applications run best with the Balanced Performance setting. Applications that are highly I/O sensitive perform best when the Energy Performance option is set to Maximum Performance.

Memory Settings for HPC Workloads

Figures 17 and 18 show the memory settings for HPC workloads.

**Figure 17.** Memory Settings for HPC Workloads in Cisco UCS B-Series Blade and Managed Cisco UCS C-Series Rack M4 Servers
Figure 18. Memory Settings for OLTP Workloads

![Memory Configuration](image)

The NUMA option should be enabled for HPC workloads so that NUMA can determine the memory allocation for each thread run by the HPC applications.

Because HPC workloads perform mostly in-memory processing, you should set DIMMs to run at the highest available frequency to process the data more quickly.

Java Enterprise Edition Application Server

Java Platform, Enterprise Edition (Java EE; previously referred to as the J2EE), defines the core set of APIs and features of Java application servers. Usually, Java EE applications are client-server or server-side applications, which require a Java EE application server.

In the industry, Java EE application servers are distinguished by the following characteristics:

- They are fully compliant application servers that implement the full Java EE stack specification with features such as JBoss Enterprise. Examples of fully compliant application servers are Apache Geronimo and JBoss Application Server.
- They are web application servers that support only the web-tier of Java EE, including the servlet. Examples of fully compliant application servers are Apache Tomcat and Jetty.

Processor Settings for Java Application Servers

Figures 19 and 20 show the selections that are recommended for Java EE application servers on Cisco UCS B-Series and managed Cisco UCS C-Series M4 servers.
Figure 19. Processor Settings for Java EE Applications
Figure 20. Processor Energy Settings for Java EE Applications

Figure 21 shows the selections that are recommended for Java EE applications on “standalone” Cisco UCS C-Series M4 servers.
**Figure 21.** Processor and Energy Settings for Java EE Applications on “Standalone” Cisco UCS C-Series M4 Servers

![Configure BIOS Parameters](image)

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Setting</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intel(R) Hyper-Threading Technology</td>
<td>Enabled</td>
</tr>
<tr>
<td>Number of Enabled Cores</td>
<td>All</td>
</tr>
<tr>
<td>Execute Disable</td>
<td>Enabled</td>
</tr>
<tr>
<td>Intel(R) VT</td>
<td>Disabled</td>
</tr>
<tr>
<td>Intel(R) VT-d</td>
<td>Enabled</td>
</tr>
<tr>
<td>Intel(R) Interrupt Remapping</td>
<td>Disabled</td>
</tr>
<tr>
<td>Intel(R) Pass Through DMA</td>
<td>Disabled</td>
</tr>
<tr>
<td>Intel(R) VT-d Coherency Support</td>
<td>Disabled</td>
</tr>
<tr>
<td>Intel(R) VT-d ATS Support</td>
<td>Disabled</td>
</tr>
<tr>
<td>CPU Performance</td>
<td>Custom</td>
</tr>
<tr>
<td>Hardware Prefetcher</td>
<td>Enabled</td>
</tr>
<tr>
<td>Adjacent Cache Line Prefetcher</td>
<td>Enabled</td>
</tr>
<tr>
<td>DCU Streamer Prefetch</td>
<td>Enabled</td>
</tr>
<tr>
<td>DCU IP Prefetcher</td>
<td>Enabled</td>
</tr>
<tr>
<td>Direct Cache Access Support</td>
<td>Enabled</td>
</tr>
<tr>
<td>Power Technology</td>
<td>Custom</td>
</tr>
<tr>
<td>Enhanced Intel Speedstep(R) Technology</td>
<td>Enabled</td>
</tr>
<tr>
<td>Intel(R) Turbo Boost Technology</td>
<td>Enabled</td>
</tr>
<tr>
<td>Processor C3 Report</td>
<td>Disabled</td>
</tr>
<tr>
<td>Processor C6 Report</td>
<td>Disabled</td>
</tr>
<tr>
<td>Processor Power state C1 Enhanced</td>
<td>Disabled</td>
</tr>
<tr>
<td>P-STATE Coordination</td>
<td>HW ALL</td>
</tr>
<tr>
<td>Energy Performance Tuning</td>
<td>OS</td>
</tr>
<tr>
<td>Energy Performance</td>
<td>Performance</td>
</tr>
<tr>
<td>Package C State Limit</td>
<td>CG Retention</td>
</tr>
<tr>
<td>Extended APIC</td>
<td>XAPIC</td>
</tr>
<tr>
<td>Workload Configuration</td>
<td>Balanced</td>
</tr>
<tr>
<td>CPU HWPM</td>
<td>Enabled</td>
</tr>
<tr>
<td>CPU Autonomous Cstate</td>
<td>Disabled</td>
</tr>
</tbody>
</table>
Intel Turbo Boost Technology enables higher CPU frequency, which helps accelerate processing of application requests. This feature helps reduce end-user response time.

Business scenarios such as batch processes run at a certain time of the day benefit from Intel Turbo Boost. It enables CPU cores to achieve at higher frequency clock speeds, which helps lower batch processing time, thereby helping the business complete and generate business reports more quickly.

You should enable all the C-states. This configuration helps reduce power consumption because only active cores will process requests during nonpeak hours. If the application demands more CPU cores, the inactive cores will become active, which helps increase throughput.

The CPU Performance option should be set to Enterprise. When a web server needs to process a large amount of data in a system, the data-access pattern is predictable (mostly sequential or adjacent lines are accessed). In this situation, it is desirable to enable the prefetchers (mid-level cache [MLC] and data cache unit [DCU]) by setting CPU Performance to Enterprise, to reduce access latency for memory-bound operations.

Memory Settings for Java EE Application Servers

Figures 22 and 23 show the recommended memory settings.

**Figure 22.** Memory Settings for Java EE Applications in Cisco UCS B-Series Blade and Managed Cisco UCS C-Series Rack M4 Servers
Set the DDR mode to performance mode so that the DIMMs work at the highest available frequency for the installed memory and CPU combination. In-memory enterprise applications such as Terracotta Ehcache benefit from the high memory speed. If this mode is enabled in web server workloads, I/O operations will be serviced at highest frequency and memory latency will be reduced.

Analytic Database Systems
An analytic database is a read-only system that stores historical data about business metrics such as sales performance and inventory levels.

An analytic database is specifically designed to support business intelligence (BI) and analytic applications, typically as part of a data warehouse or data mart. This feature differentiates it from operational, transactional, and OLTP databases, which are used for transaction processing: order entry and other "run the business" applications.

Processor Settings for Analytic Database Systems
Figures 24 and 25 show the selections that are recommended for analytic database systems on Cisco UCS B-Series and managed Cisco UCS C-Series M4 servers.
Figure 24. Processor Settings for Analytic Database Systems
Figure 25. Processor Energy Settings for Analytic Database Systems

Figure 26 shows the selections that are recommended for analytic database systems on “standalone” Cisco UCS C-Series M4 servers.
Figure 26. Processor and Energy Settings for “Standalone” Cisco UCS C-Series M4 Servers
Memory Settings for Analytic Database Systems

Figures 27 and 28 show the recommended memory settings.

**Figure 27.** Memory Settings for Analytic Database Systems in Cisco UCS B-Series Blade and Managed Cisco C-Series Rack M4 Servers

**Figure 28.** Memory Settings for Analytic Database Systems in “Standalone” Cisco UCS C-Series M4 Servers
Configuring the BIOS for Optimized CPU Hardware Power Management

This section summarizes the BIOS settings you can configure to optimize CPU power management and provides additional resources for choosing the best settings for your environment.

Recommended Power Management Settings to Optimize Power Efficiency

This section presents the settings you can configure to optimize the power efficiency of your system.

**Power Technology: Custom**

You can configure CPU power management settings for the following options:

- Enhanced Intel SpeedStep Technology
- Intel Turbo Boost Technology
- Processor Power State C6

Power Technology can be set to one of the following:

- **Custom**: The server uses the individual settings for the BIOS parameters in the preceding list. You must select this option if you want to change any of these BIOS parameters.
- **Disabled**: The server does not perform any CPU power management, and any settings for the BIOS parameters in the preceding list are ignored.
- **Energy Efficient**: The server determines the best settings for the BIOS parameters in the preceding list and ignores the individual settings for these parameters.

**Enhanced Intel SpeedStep Technology: Enabled**

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 possible settings are as follows:

- **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.

You should contact your operating system vendor to make sure that your operating system supports this feature.

**Note**: Power Technology must be set to Custom; otherwise, the server ignores the setting for this parameter.

**Intel Turbo Boost Technology: Disabled**

You can specify whether 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. The possible settings are as follows:

- **Disabled**: The processor does not increase its frequency automatically.
- **Enabled**: The processor uses Intel Turbo Boost Technology if required.

**Note**: Power Technology must be set to Custom; otherwise, the server ignores the setting for this parameter.
Processor C3 Report: Disabled
You can specify whether the BIOS sends the C3 report to the operating system. When the OS receives the report, it can transition the processor into the lower C3 power state to decrease energy use while maintaining optimal processor performance. The possible settings are as follows:

- **Disabled**: The BIOS does not send the C3 report.
- **Enabled**: The BIOS sends the C3 report, allowing the OS to transition the processor to the C3 low-power state.

**Note**: Power Technology must be set to Custom; otherwise, the server ignores the setting for this parameter.

Processor C6 Report: Enabled
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 possible settings are as follows:

- **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.

**Note**: Power Technology must be set to Custom; otherwise, the server ignores the setting for this parameter.

Processor Power State C1 Enhanced: Enabled
You can specify whether the CPU transitions to its minimum frequency when entering the C1 state. The possible settings are as follows:

- **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.

P-STATE Coordination: SW_ANY
You can define the way that the BIOS communicates the P-state support model to the operating system. Three models are available, as defined by the Advanced Configuration and Power Interface (ACPI) specification:

- **HW_ALL**: The processor hardware is responsible for coordinating the P-state among logical processors with dependencies (all the logical processors in a package).
- **SW_ALL**: The OS power manager (OSPM) is responsible for coordinating the P-state among logical processors with dependencies (all the logical processors in a physical package) and must initiate the transition on all the logical processors.
- **SW_ANY**: The OSPM is responsible for coordinating the P-state among logical processors with dependencies (all the logical processors in a package) and can initiate the transition on any of the logical processors in the domain.

**Note**: Power Technology must be set to Custom; otherwise, the server ignores the setting for this parameter.
**Energy Performance Tuning: OS or BIOS**
You can choose the BIOS or the operating system for energy performance bias tuning. The possible settings are as follows:

- **OS:** This setting chooses the OS for energy performance tuning.
- **BIOS:** This setting chooses the BIOS for energy performance tuning.

**Energy Performance: Energy Efficient (Power)**
You can specify whether system performance or energy efficiency is more important on this server. The possible settings are as follows:

- Balanced Energy
- Balanced Performance
- Energy Efficient
- Performance

**Package C State Limit: C6 Retention**
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 the 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 the 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 non-Retention option, because the processor is operating at Pn voltage to reduce the package’s C-state exit latency.

**QPI Link Frequency Select: 6.4 GT/s**
You can specify the Intel QuickPath Interconnect (QPI) link frequency, in gigatransfers per second (GT/s). The possible settings are as follows:

- Auto: The CPU determines the QPI link frequency.
- 6.4 GT/s
- 7.2 GT/s
- 8.0 GT/s
PCIe Slot: n Link Speed: GEN1
You can restrict the maximum speed of an adapter card installed in PCIe slot n. The possible settings are as follows:

- **GEN1**: 2.5 GT/s is the maximum speed allowed.
- **GEN2**: 5 GT/s is the maximum speed allowed.
- **GEN3**: 8 GT/s is the maximum speed allowed.
- **Disabled**: The maximum speed is not restricted.

For example, if you have a third-generation adapter card in PCIe slot 2 that you want to run at a maximum of 5 GT/s instead of the 8 GT/s that the card supports, set the PCIe slot-2 link speed to GEN2. The system then ignores the card's supported maximum speed of 8 GT/s and forces it to run at a maximum of 5 GT/s.

Fan Control Policy: Low Power
Fan control policies enable you to control the fan speed to reduce server power consumption and noise levels. Prior to these fan policies, 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 resulted 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 support it. 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.

Following are the fan policies that you can choose from:

- **Balanced**: This is the default policy. This setting can cool almost any server configuration, but it may not be suitable for servers with PCIe cards, because these cards overheat easily.
- **Performance**: This setting can be used for server configurations in which maximum fan speed is required for high performance. With this setting, the fan speeds will run at the same speed or a higher speed than with the Balanced fan policy.
- **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 to 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.
Operating System Resources for Power Management
Consult the following resources for more information about OS power management:

- Red Hat Enterprise Linux Power Management Guide—Managing Power Consumption:
- SUSE Enterprise Linux—Power Management (section 11, page 119):
- Power Policy Configuration and Deployment in Microsoft Windows:
- Host Power Management in VMware vSphere 5.5 and 6.0:

Operating System Tuning Guidance for Maximum Performance
For Linux, set the following:

```
x86_energy_perf_policy performance
```

When Energy Performance Tuning is set to OS, the OS controls the Energy Performance Bias (EPB) policy. The EPB features controlled by the policy are Intel Turbo Boost Override, Memory CKE, Memory OSR, Intel QPI L0p, C-state Demotion, and I/O Bandwidth P-Limit. The default OSPM profile is Performance, which will not sacrifice performance to save energy.

```
cpuset.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.

Edit /etc/init.d/grub.conf to set intel_pstate=disable

Operating System Tuning Resources for Performance
Consult the following resources for more information about OS performance tuning:

- Red Hat Enterprise Linux—CPU Performance:
- SUSE Enterprise Linux—CPU Performance (section: 11, page 119):
- Microsoft Windows and Hyper-V tuning is straightforward; set Power Policy to High Performance:
  Performance Tuning Guidelines for Windows Server 2012 R2 - Microsoft
VMware ESXi tuning is straightforward as well; set Power Policy to High Performance:

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, snoop settings, and CPU hyperthreading.

For More Information

For more information about Cisco UCS B-Series and C-Series M4 servers, see:

- Cisco UCS B200 M4 Blade Server:
- Cisco UCS C220 M4 Rack Server:
- Cisco UCS C240 M4 Rack Server:
- Cisco UCS B460 M4 Blade Server:
- Cisco UCS B420 M4 Blade Server:
- Cisco UCS C460 M4 Rack Server:

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