

BIOS optimization for 3rd Generation Xeon Scalable Processor-based systems

This document explains the BIOS settings that can be modified for the 3rd generation Intel Xeon Scalable processor-based PRIMERGY server generation (PRIMERGY RX2530 M6, RX2540 M6, CX2550 M6, CX2560 M6, and RX4770 M6).

Its purpose is so that the user can optimize the BIOS settings according to their personal requirements. The objectives are to optimize PRIMERGY servers for either maximum performance or maximum energy efficiency. As far as performance is concerned, application scenarios that emphasize minimizing response time as much as possible are also taken into account in addition to optimization for maximum throughput.



Version
1.5
2023-10-03

Contents

Overview 3

Application scenarios 4

 Performance4

 Low Latency4

 Energy savings / Energy efficiency5

 Application Profile.....5

PRIMERGY BIOS options..... 7

 Recommendations for optimization7

 BIOS options details 11

Appendix..... 25

Literature 36

Overview

When Fujitsu PRIMERGY servers leave the factory, they are already configured with BIOS standard settings, which provide an optimal ratio between performance and energy efficiency for the most common application scenarios. And yet there are situations in which it may be necessary to modify the standard settings depending on requirements for the most throughput possible (performance), as little latency as possible (low latency), or emphasizing as much energy conservation as possible (energy efficiency). This document offers best-practice recommendations for optimal BIOS settings for these three scenarios, which are explained in more detail below. In addition to the BIOS settings, the entire system must also be considered when optimizing PRIMERGY servers. The following aspects should be given consideration when planning server systems:

- Server hardware
 - Processor: Number of cores and frequency
 - Memory: Memory type (3DS DIMM, LR DIMM, RDIMM, NVDIMM) and memory configuration
 - I/O cards: Optimal distribution of several cards over PCIe slots
- Operating system and application software
 - Hypervisor: vSphere, Hyper-V, KVM
 - Power plan: Performance or energy efficiency
 - Tuning: Kernel, registry, interrupt binding, thread splitting
- Network
 - Network technology: 1/10/25/40/100 Gbit Ethernet, Fiber Channel, InfiniBand, RDMA
 - Network architecture: Switches, multichannel
- Storage
 - Technology: RAID, Fiber Channel, Direct Attached, NVMe
 - Disks: HDD, SSD, SATA, SAS
- Accelerator
 - Architecture: GPU, GPGPU, FPGA

Application scenarios



Performance

Thanks to the latest multi-processor, multi-core, and multi-threading technology in conjunction with current operating systems and applications, today's PRIMERGY servers based on the Intel Xeon Scalable Processors deliver the highest levels of performance.

This is proven by the numerous benchmark publications of the Standard Performance Evaluation Corporation (SPEC), SAP, or VMware. When you emphasize server performance, you mostly mean throughput. Users, for whom maximum performance is essential, are interested in carrying out as many parallel computing operations as possible and utilizing if possible, all the resources of the parallel processor. Although PRIMERGY servers with the standard settings already provide an optimal ratio between performance and energy efficiency, it is possible to further optimize the system regarding performance and to a lesser degree energy efficiency via the BIOS. Performance optimization is a matter of operating all the components in the system at the fastest speed possible and preventing the energy-saving options from slowing down the system. Therefore, optimization toward maximum performance is in most cases also associated with an increase in electrical power consumption.



Low Latency

Minimum possible latency is a requirement that comes from the High Performance Computing (HPC) sector in particular and from finance market applications, where the object is to process millions of transactions per second and data in real time without any delay. Users in this segment are not primarily concerned with achieving the maximum

possible throughput through system optimization but emphasize more on increasing the speed of each individual transaction, i.e., reducing the time required to perform an individual transaction. In such cases, the focus is placed on the response time of a system, the so-called latency (typically measured in nanoseconds, microseconds, or milliseconds). The BIOS offers a variety of options to reduce latency. On the one hand, it is possible - such as when you know that the corresponding application does not make efficient use of all the threads available in the hardware - to disable threads that are not needed (Hyper-Threading) or even cores in the BIOS in order in this way to minimize fluctuations in computing speed that especially occur in a number of HPC applications. Furthermore, the disabling of cores that are not needed can improve the Turbo mode performance of the remaining cores under certain operating conditions. On the other hand, there are scenarios which require performance that is as constant as possible. In this case, it is necessary to keep the response time constant by avoiding configurations, in which changes in frequency occur, such as with Turbo mode. Although the current generation of Intel processors delivers a clearly better Turbo mode performance than the predecessor generations, the maximum Turbo mode frequency is not guaranteed under certain operating conditions. In such cases, disabling the Turbo mode can help avoid changes in frequency. Energy-saving functions, whose aim is to save energy whenever possible, through frequency/voltage reduction and through the disabling of certain function blocks and components, also have a negative impact on the response time. The stricter the energy-saving mode, the lower the performance. Furthermore, in each one of these energy-saving modes, the processor requires a certain time in order to change back from temporarily reduced performance to maximum performance. This time worsens the latency of the system, particularly after a transaction is pending and the system remains idle, or if the system load fluctuates irregularly. This document explains how to configure the power saving modes for users from the low-latency segment in order to minimize system latency. However, the optimization of server latency, particularly in an idle state, always results in higher electrical power consumption.

Note about "Performance" and "Low latency":

The maximum throughput or minimum latency of the I/O system can be of significance for I/O critical applications. These values have - in conjunction with the I/O system - a different meaning to the one associated with processors. For example, the I/O throughput means the amount of data transferred per time unit by the I/O system. In order to achieve maximum I/O throughput or minimum I/O latency, the BIOS optimization of the processors does not have to be set at maximum throughput of computing operations (i.e., "performance") or "low latency". In most situations, the BIOS standard settings are optimal and are in conjunction with optimally set I/O components. This almost always provides the highest possible optimization for these components. However, in certain rare situations, these target values can be missed with very high requirements (for SSDs). The solution can be either to set the BIOS option [Uncore Frequency Scaling] to [Maximum] or the BIOS option [Utilization Profile] (see the respective section for a more detailed description).

**Energy savings / Energy efficiency**

In addition to the scenarios for maximum throughput and minimum latency, there are also environments in which energy consumption is emphasized more than performance. Two different objectives are pursued regarding this.

One way is to select the BIOS options in such a way that the lowest possible electrical power consumption is achieved in each case. This is for example an option for data center operators, who only have a restricted budget of electrical power and are aiming to reduce power consumption for each rack and for each server respectively with performance only playing a subordinate role. Optimization in this direction consists primarily of modifying the settings to reduce the speed and thus the performance of the server.

The other way is to configure a server in such a way that it gives the best possible ratio between throughput and electrical power consumption. This is the only way to achieve the optimal energy efficiency of a server (measured in performance per watt). Such optimization is particularly targeted by data center operators, for whom the maximum performance of a server is of secondary importance and optimizing total cost of ownership is more significant.

Numerous publications of the Standard Performance Evaluation Corporation (SPEC) with the first industry-standard benchmark for measuring energy efficiency in servers, the SPECpower_ssj2008, as well as VMmark V3 Performance with Server Power prove that PRIMERGY servers are also the best choice when it comes to energy-efficient servers.

Application Profile

Though general applications can be categorized into above 3 types, it requires the user's effort of the setting individual BIOS options to achieve the best performance. So [Application Profile] option was added from PRIMERGY servers with 3rd generation Intel Xeon scalable processors for the convenience of users. Users can configure the optimized BIOS settings automatically by selecting a workload which is close to their actual operational environment. Please refer to Appendix for the detailed settings for each profile.

PRIMERGY servers provide the following 10 type of application profiles.

- **Total Throughput Performance**

The profile optimized for the workload which requires the maximum throughput.

- **Single Thread Performance**

The profile optimized for the workload which requires the peak performance of single core, rather than the throughput.

- **Energy Efficiency**

The profile to balance between the performance and the power

- **Virtualization Performance**

The profile optimized for the workload which requires the performance for virtualization host environment such as VMware vSphere.

- **Low Latency**

The profile to minimize the execution time of individual processing

- **Online Transaction Processing**

The profile optimized for the workload such as online transaction processing applications used in the database back-end.

- **Online Analytical Processing In-Memory DB**

The profile optimized for the workload of In-Memory Database as represented by SAP HANA

- **I/O Throughput**

The profile optimized for the workload which requires I/O throughput performance

- **CPU Intensive HPC**

The profile optimized for the workload in High performance computing (HPC) area, which mainly requires CPU performance, rather than Memory performance.

- **Memory Intensive HPC**

The profile optimized for the workload in High performance computing (HPC) area, which mainly requires Memory performance, rather than CPU performance.

Note :

Although the settings selected in this BIOS option have been validated in some typical workloads, the actual workload varies widely and cannot be uniformly categorized into the above 10 profiles. After selecting the profile which most closely matches your workload in this BIOS option, you can change the BIOS options individually as needed.

PRIMERGY BIOS options

This white paper contains information about BIOS options that are valid for Intel Xeon Scalable processor based PRIMERGY servers. These servers are:

- PRIMERGY RX2530 M6
- PRIMERGY RX2540 M6
- PRIMERGY CX2550 M6
- PRIMERGY CX2560 M6
- PRIMERGY RX4770 M6

The BIOS of the PRIMERGY servers is being continuously developed. Therefore, it is important to use the latest BIOS version in each case so as to have all the BIOS functions listed here available. The appropriate downloads are available on the Internet at <https://www.fujitsu.com/global/support>.

Recommendations for optimization

The following tables list recommendations for BIOS options, which optimize the PRIMERGY servers either for best performance, low latency, or maximum energy efficiency. To change the BIOS options, it is first of all necessary to call up the BIOS setup during the system self-test (Power On Self Test = POST). More information about this can be found in the server manual.

Many of the BIOS options listed here have interdependencies. This can result in certain changes to specific options alone displaying undesirable system behavior and only having the desired effect when further options are also changed at the same time. Before changes are made to the BIOS options contained in the following tables, it is recommended that you look at the footnotes and subsequent descriptions of the BIOS options. Furthermore, any changes should first be examined in a test environment for the required effect, before transferring them to the production environment.

In addition to the recommendations for BIOS options, particular attention should also be paid to the selection and tuning of the operating system when planning a server system. Depending on the use, the selection of a specific operating system and its tuning can influence performance, latency, and energy efficiency. Additional information regarding the tuning for individual operating systems is available at the links in "Operating System Performance Tuning Guidelines" section of "Literature".

Recommended BIOS settings

BIOS Setup Menu	Setting ¹	Performance	Low Latency	Energy Efficient
Configuration -> CPU Configuration				
Hyper-Threading	Disabled / Enabled	Enabled	Disabled ²	Enabled
Active Processor Cores	All / [1 - n]	All	1 - n ³	All
Prefetcher • Hardware Prefetcher • Adjacent Cache Line Prefetch • DCU Streamer Prefetcher • DCU IP Prefetcher • LLC Prefetch • L2 RFO Prefetch • UPI Prefetch ¹²	Disabled / Enabled	Enabled	Enabled	Disabled ⁴
• XPT Prefetch	Disabled / Enabled	Disabled	Disabled	Disabled ⁴
• XPT Remote Prefetch ¹²	Disabled / Enabled / Auto	Auto	Auto	Disabled ⁴
Intel Virtualization Technology	Disabled / Enabled	Disabled ⁵	Disabled ⁵	Disabled ⁵
Enhanced SpeedStep	Disabled / Enabled	Enabled	Enabled	Enabled
Turbo Mode	Disabled / Enabled	Enabled	Disabled ⁶	Disabled
Override OS Energy Performance ⁷	Disabled / Enabled	Enabled	Enabled	Disabled ⁸

¹ The settings in bold print are the default values.

² Hyper-Threading doubles the number of logical cores but can also result in performance fluctuations. Disabling can improve latency.

³ By restricting the number of active cores for applications that are single-threaded or applications that do not use all the CPU threads, it is possible to improve Turbo Mode performance.

⁴ The disabling of the prefetchers increases energy efficiency if performance remains the same or improves. This should be verified in advance for the individual prefetchers.

⁵ If virtualization is not used, this option should be set to [Disabled].

⁶ Maximum Turbo Mode performance is not guaranteed under all operating conditions, which can result in fluctuations in performance. The turbo mode option should be set to [Disabled] for a stable and consistent response time.

⁷ If the option [HWPM Support] is set to [OOB Mode], this option is grayed out and the setting for it is automatically changed to [Enabled].

⁸ If the operating system in use is able to set the "energy efficient policy" for the CPUs, set [Override OS Energy Performance] to [Disabled] then the settings for the [Energy Performance] option should be made via the operating system's power plan. If the operating system is incapable of this, or you do not want to leave this up to the operating system, you can set the option to [Enabled] and make the [Energy Performance] setting via the BIOS.

BIOS Setup Menu	Settings ¹	Performance	Low Latency	Energy Efficient
Energy Performance	Performance / Balanced Performance / Balanced Energy / Energy Efficient	Performance	Performance	Energy Efficient
Utilization Profile	Even / Unbalanced	Even	Unbalanced	Even
P-State Coordination ⁹	HW_ALL / SW_ALL	HW_ALL	HW_ALL	HW_ALL
HWPM Support ¹⁰	Disabled / Native Mode / OOB Mode / Native Mode with no legacy	Native Mode	Disabled	Disabled
CPU C1E Support	Enabled / Disabled	Enabled	Disabled	Enabled
CPU C6 Report	Disabled / Enabled	Enabled	Disabled	Enabled
Package C State limit	C0 / C2 / C6 ¹¹ / C6 (Retention) / No Limit	C0	C0	No Limit
UPI Link Frequency Select	Auto / 9.6 GT/s / 10.4 GT/s / 11.2 GT/s ¹²	Auto	Auto	9.6 GT/s
UPI Link L0p	Disabled / Enabled	Enabled	Disabled	Enabled
UPI Link L1	Disabled / Enabled	Enabled	Disabled	Enabled
Uncore Frequency Scaling	Disabled / Maximum / Power Balanced	Disabled	Maximum ¹³	Power Balanced
LLC Dead Line Alloc	Disabled / Enabled	Disabled	Disabled	Disabled

⁹ Available for RX4770 M6.¹⁰ This option is only visible if [Enhanced SpeedStep] is [Enabled].¹¹ The default value for RX2530 M6 and RX2540 M6.¹² Available for RX2530 M6, RX2540 M6, CX2550 M6, and CX2560 M6.¹³ The [Maximum] setting for this option can be advantageous for applications with a high I/O utilization, but low core utilization.

BIOS Setup Menu	Settings ¹	Performance	Low Latency	Energy Efficient
Stale AtoS	Disabled / Enabled / Auto	Enabled	Enabled	Auto
Configuration -> Memory Configuration				
DDR Performance	Performance optimized / Energy optimized / Power balanced	Performance optimized	Performance optimized	Energy optimized
Patrol Scrub	Disabled / Enabled	Enabled	Disabled	Disabled
Virtual NUMA ^{12, 15}	Disabled / Enabled	Disabled	Disabled	Disabled
SNC(Sub NUMA)	Enabled ¹⁴ / Disabled	Enabled	Enabled	Enabled
UMA-Based Clustering ^{12, 15}	Disabled / Hemisphere	Disabled	Disabled	Disabled

¹⁴ The option name in RX2530M6, RX2540 M6, CX2550 M6, and CX2560 M6 is [Enable SNC2].

¹⁵ This option is only available if [SNC(Sub NUMA)] is [Disabled]

BIOS options details

This section provides details about each BIOS option.

Since the effect of changing BIOS options is also affected by the hardware / software configuration and other BIOS/OS option settings, be sure to verify these settings before actual operation.

Hyper-Threading

BIOS Setup Menu	BIOS Option	Setting	Performance	Low Latency	Energy Efficiency
Configuration >CPU Configuration	Hyper-Threading	Disabled Enabled	Enabled	Disabled	Enabled

Generally, Fujitsu recommends that you enable [Hyper-Threading] ([Enabled]). Nevertheless, it can make sense to disable [Hyper-Threading] for applications that especially attach importance to the shortest possible response times (e.g., for trading software from the finance market or HPC applications). Users from these fields are usually less interested in maximum system throughput, which is provided by the additional threads, than in the performance and stability of an individual thread. The disabling of [Hyper-Threading] can prevent the associated performance fluctuations of computing operations and thus improve latency.

Active Processor Cores

BIOS Setup Menu	BIOS Option	Setting	Performance	Low Latency	Energy Efficiency
Configuration >CPU Configuration	Active Processor Cores	All [1 – n]	All	1 - n	All

It is possible to disable individual cores of a processor in the BIOS (e.g., four cores on a 10-core processor can be disabled). In this case, you can now use the L3 cache which would be larger than when there were 10 cores for the remaining cores. Although maximum throughput is generally achieved with all the cores, you can utilize the Turbo Mode frequency which is higher than the remaining active cores by disabling the cores that are not needed. This is advantageous especially with latency-sensitive applications that do not utilize all the cores. This works because the disabled cores reduce the electrical power consumption of the processor and thereby allowing higher Turbo Mode frequencies on the remaining cores. This does not necessarily work with all the load profiles. Particularly, there is not much an effect on power-hungry AVX applications.

Prefetcher

BIOS Setup Menu	BIOS Option	Setting	Performance	Low Latency	Energy Efficiency
Configuration >CPU Configuration	Hardware Prefetcher	Disabled Enabled	Enabled	Enabled	Disabled
	Adjacent Cache Line Prefetch				
	DCU Streamer Prefetcher				
	DCU IP Prefetcher				
	LLC Prefetch				
	L2 RFO Prefetch				
	UPI Prefetch				
	XPT Prefetch	Disabled Enabled	Disabled	Disabled	Disabled
	XPT Remote Prefetch	Disabled Enabled Auto	Auto	Auto	Disabled

The PRIMERGY server BIOS has several prefetcher options as above.

The prefetchers are processor functions, which enable data to be loaded in advance according to specific patterns from the main memory to the L1 or L2 cache of the processor. Enabling the prefetchers usually ensures a higher cache hit rate and thus increases the overall performance of the system. Application scenarios, in which memory transfer is a performance bottleneck, are the exception to this. In these cases, it can be advantageous to set the prefetcher options to [Disabled] so the bandwidth that is otherwise used for the prefetching can be used. Furthermore, the power consumption of the server can be slightly reduced by disabling the prefetchers. Before the prefetcher options are changed on the active systems, the effects of the individual settings for the respective application scenario should first be examined in a test environment.

Details of the individual prefetchers:

Hardware Prefetcher

This prefetcher looks for data streams on the assumption that if the data is requested at address A and A+1, the data will also presumably be required at address A+2. This data is then prefetched into the L2 cache from the main memory.

Adjacent Cache Line Prefetch

This prefetcher always collects cache line pairs (128 bytes) from the main memory, providing that the data is not already contained in the cache. If this prefetcher is disabled, only one cache line (64 bytes) is collected, which contains the data required by the processor.

DCU Streamer Prefetcher

This prefetcher is a L1 data cache prefetcher, which detects multiple loads from the same cache line done within a time limit. Based on the assumption that the next cache line is also required, this is then loaded in advance to the L1 cache from the L2 cache or the main memory.

DCU Ip Prefetcher

This L1-cache prefetcher looks for sequential load history and attempts on this basis to determine the next data to be expected and, if necessary, to prefetch this data from the L2 cache or the main memory into the L1 cache.

LLC Prefetch

In the Xeon Scalable Processor family, L3 cache (LLC: Last Level Cache) is non-inclusive and data from main memory is loaded directly to L2 cache. This prefetcher enables cores to prefetch data from main memory to the LLC.

L2 RFO Prefetch

This prefetcher puts out an RFO (Read For Ownership) to get write privileges when prefetching data from memory to L2 cache.

UPI Prefetch

This prefetcher enables UPI controller to issue a speculative read request to the memory controller in parallel to an LLC lookup.

XPT Prefetch / Remote Prefetch

This prefetcher will issue a speculative read request to local / remote memory in parallel to an LLC lookup. This prefetcher improves the memory latency by using prefetched data in the case a cache miss occurred in LLC. This prefetcher makes a prediction based on the access history of Xtended Prediction Table (XPT).

Intel Virtualization Technology

BIOS Setup Menu	BIOS Option	Setting	Performance	Low Latency	Energy Efficiency
Configuration >CPU Configuration	Intel Virtualization Technology	Disabled Enabled	Disabled	Disabled	Disabled

This BIOS option enables or disables additional virtualization functions of the CPU. If the server is not used for virtualization, this option should be set to [Disabled]. This can result in energy savings.

Enhanced SpeedStep

BIOS Setup Menu	BIOS Option	Setting	Performance	Low Latency	Energy Efficiency
Configuration >CPU Configuration	Enhanced SpeedStep	Disabled Enabled	Enabled	Enabled	Enabled

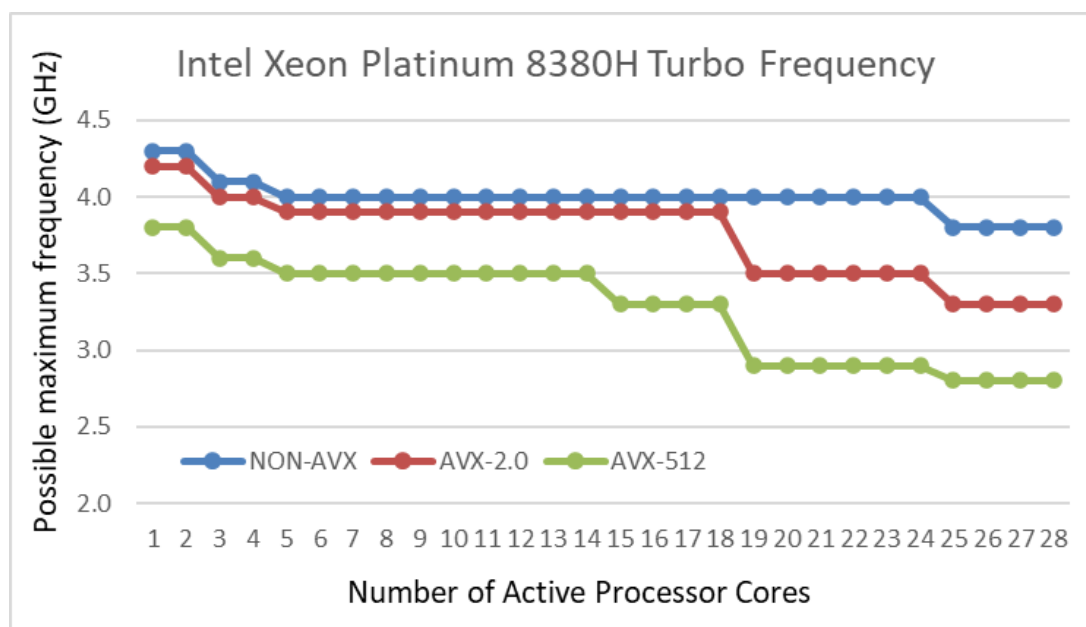
Enhanced Intel SpeedStep Technology (EIST) is a power saving function that allows individual cores or even the entire processor to adapt its performance to specific load profiles. This is achieved by reducing frequency and voltage when maximum computing performance is not required, which in turn considerably reduces energy requirements in part. Since the distribution of the computing performance is subject to the operating system and the therein implemented strategies (e.g., the power plan provided), Fujitsu recommends leaving the option [Enhanced SpeedStep] enabled. If this option is disabled, the Turbo Mode function, which allows more computing performance to be made available at short notice by increasing the frequency above nominal frequency, is also not available.

Turbo Mode

BIOS Setup Menu	BIOS Option	Setting	Performance	Low Latency	Energy Efficiency
Configuration >CPU Configuration	Turbo Mode	Disabled Enabled	Enabled	Disabled	Disabled

This BIOS option enables and disables the Intel Turbo Boost Technology function of the processor. The Turbo Boost technology function permits the processor to provide more computing performance at short notice by increasing the frequency above the nominal frequency. The maximum achievable frequency is influenced by numerous factors - processor type, number of active processor cores, power supply, current electrical power consumption, temperature, as well as the instructions that have to be executed (whether AVX512 instructions are used, AVX2.0 instructions are used, or Non-AVX instructions are used).

The following Figure shows Xeon 8380H maximum achievable core frequency per number of active processor cores. Here, active processor core means a core which is enabled by [Active processor core] and is not "C6 C-State". (See [Active processor cores] and [CPU C6 report] for details.)



In addition to these general conditions, the quality of the processors also plays a major role for the Turbo Mode performance, particularly with HPC applications. Thus, for example the production variance results in the individual processors of the same type having a different power consumption under the same load.

Generally, Fujitsu always recommends leaving the [Turbo Mode] option set at the standard setting [Enabled], as performance is substantially increased by the higher frequencies. However, as the higher frequencies depend on its operating conditions as mentioned above and are not always guaranteed, it can be advantageous for application scenarios, in which you want constant performance or to lower electrical power consumption, to disable the [Turbo Mode] option.

Override OS Energy Performance

BIOS Setup Menu	BIOS Option	Setting	Performance	Low Latency	Energy Efficiency
Configuration >CPU Configuration	Override OS Energy Performance	Disabled Enabled	Enabled	Enabled	Disabled

The new generation of Intel Xeon based processors comes with a large number of energy-saving options. The so-called power control unit (PCU) in the processors takes on the central role of controlling all these energy-saving options. The PCU can be parameterized in order to consequently control the settings more in the direction of energy efficiency or in the direction of maximum performance. This can be done in two ways. The standard setting allows you to control the [Energy Performance] option through the operating system. Depending on the selected power plan, which is set in the operating system, a specific value is written in a CPU register. This register is then evaluated by the PCU and the energy-saving functions of the CPU are controlled accordingly. The other option is to set the [Energy Performance] option directly via the BIOS and thus override the setting of the operating system. This makes particular sense if e.g., an older operating system is not able to write to this special CPU register, or if you want to set the energy-saving options centrally from the BIOS, i.e., independent of the operating system. In this case, the BIOS option [Override OS Energy Performance] must be enabled.

If hardware power management ([HWPM Support]) setting is [OOB Mode], then the option [Override OS Energy Performance] is enabled as the standard. Furthermore, in this case, the preference and PCU parameterization as regards energy efficiency or performance must be selected in this case via the BIOS option [Energy Performance].

Energy Performance

BIOS Setup Menu	BIOS Option	Setting	Performance	Low Latency	Energy Efficiency
Configuration >CPU Configuration	Energy Performance	Performance Balanced Performance Balanced Energy Energy Efficient	Performance	Performance	Energy Efficient

Depending on the setting, this BIOS option parameterizes the internal "Power Control Unit (PCU)" of the Intel processors and optimizes the power management functions of the processors between performance and energy efficiency. Possible settings are [Performance], [Balanced Performance], [Balanced Energy] and [Energy Efficient].

Energy Performance settings, also known as "Energy Performance Bias", can be set from the OS, but if the BIOS option [Override OS Energy Performance] is set to [Enabled], this setting specified in the BIOS is forced into effect. If [Override OS Energy Performance] is set to [Disabled], the operating system takes on the task of setting the [Energy Performance] via the power plan. However, this setting may affect the OS power policy depending on the OS type and settings.

Utilization Profile

BIOS Setup Menu	BIOS Option	Setting	Performance	Low Latency	Energy Efficiency
Configuration >CPU Configuration	Utilization Profile	Even Unbalanced	Even	Unbalanced	Even

The option [Utilization Profile] is used to parameterize an energy-saving option, which monitors both the UPI and the PCIe bandwidth, and attempts to adapt the processor frequency based on the utilization. The standard setting is [Even], because it is assumed that the CPU load is evenly distributed over all the processors, and in this way, the appropriate frequency is optimally adapted based on the CPU utilization. The [Even] setting accordingly ensures a less aggressive increase in the processor frequency. On the other hand, the [Unbalanced] setting targets application scenarios with high PCIe utilization for a low CPU load. Configurations with GPGPUs are a typical example of this. In such cases, the operating system could as a result of the rather lower utilization of the CPUs request accordingly lower frequencies, although in fact a high frequency is needed in order to achieve the maximum possible PCIe bandwidth. The [Unbalanced] setting ensures that in the case of high UPI or PCIe utilization the frequency of the processors is aggressively increased - even if CPU utilization is low. Fujitsu generally recommends working with the standard setting [Even], because this setting is clearly more energy efficient. However, if performance problems occur in application scenarios, in which a high PCIe bandwidth is required, the [Unbalanced] setting can counteract this.

P-State Coordination

BIOS Setup Menu	BIOS Option	Setting	Performance	Low Latency	Energy Efficiency
Configuration >CPU Configuration	P-State Coordination	HW_ALL SW_ALL	HW_ALL	HW_ALL	HW_ALL

The option [P-STATE Coordination] determines whether to select the processor hardware or OS to be responsible for the control of P-STATE. The standard [HW_ALL] setting selects the processor hardware to coordinate and performs the transition of P-STATE among all logical processors in a package. [SW_ALL] setting selects OS Power Management (OSPM) to do it.

HWPM Support

BIOS Setup Menu	BIOS Option	Setting	Performance	Low Latency	Energy Efficiency
Configuration >CPU Configuration	HWPM Support	Disabled Native Mode OOB Mode Native Mode with no legacy	Native Mode	Disabled	Disabled

HWPM stands for hardware power management and is a new power saving function that was introduced with the Intel Broadwell processor generation and enhanced with Intel Skylake processor generation. The option [HWPM support] can be used to configure two operating modes, which - depending on the respective utilization - assume control of the processor frequency in a similar way to legacy power management, which is based on enhanced Intel SpeedStep technology. In contrast to legacy power management, in which utilization evaluation and control of the P-states is regulated by the operating system, i.e., in the software, these tasks are in the case of hardware power management taken on in the hardware by the processor itself. Hardware power management can be the better choice for operating systems, which do not offer legacy power management support or offer inefficient legacy power management support.

The setting [Native Mode] provides the operating system with an interface, via which restrictions and information regarding power management can be passed on, and which are then considered by hardware power management for control. If on the other hand the setting [OOB Mode] is enabled, hardware power management then autonomously takes control of the processor frequency, i.e., completely independently of the operating system. If the setting [Native Mode with no legacy] is enabled, the BIOS provides the OS with only the interface which is used to inform power management control in HWPM [Native Mode]. This means that the BIOS doesn't provide legacy P-state information to the OS. The BIOS options [Enhanced SpeedStep] and [Turbo Mode] are still available in both [Native Mode] and in [OOB Mode] and are considered by hardware power management in Broadwell generations. If [HWPM Support] is [Disabled], legacy power management is enabled via [Enhanced SpeedStep].

CPU C1E Support

BIOS Setup Menu	BIOS Option	Setting	Performance	Low Latency	Energy Efficiency
Configuration >CPU Configuration	CPU C1E Support	Enabled Disabled	Enabled	Disabled	Enabled

Intel Xeon Scalable processor supports four C-States, C0, C1, C1E, and C6. The CPU C-states except for C0 are a type of sleep state. Power consumption becomes lower in the order of C0, C1, C1E, C6, but wake-up time becomes longer in the same order.

C-State transition is triggered by operating system request. If this option is enabled, request to C1 transition by the operating system is handled as a request to C1E transition by the processor and results in slightly lower power consumption. Some operating systems request direct transition to C1E and in this case this option has no effect.

C1E ensures that the frequency is clocked down to the lowest frequency supported by the processors. This takes place regardless of Intel SpeedStep technology. In other words, even if the setting that the processor is to run with maximum frequency is made via the power plan of the operating system, C1E would - if enabled - ensure that the processor in an idle state clocks down to the lowest frequency. This can be disadvantageous with low latency applications in particular, because the clocking down and back up again of the frequency increases the latency. In such cases, the setting can be changed to [Disabled]. Fujitsu recommends that you enable this option except for latency sensitive workloads.

CPU C6 Report

BIOS Setup Menu	BIOS Option	Setting	Performance	Low Latency	Energy Efficiency
Configuration >CPU Configuration	CPU C6 Report	Disabled Enabled	Enabled	Disabled	Enabled

This BIOS option is used to inform the operating system whether it can use the CPU C6 states ([Enabled]) or not ([Disabled]). C3 State is no longer supported in Xeon Scalable processor generation.

Since the wake-up from these C-states increases latency, it is advisable to change the setting to [Disabled] for the CPU C6 Report for applications where maximum performance with the lowest possible response time matters. It should be borne in mind that if CPU C6 C-state is disabled, the highest possible Turbo Mode frequency can no longer be achieved. In this case and regardless of the number of active cores, the highest Turbo Mode frequency would be limited to the maximum frequency that is possible if all the cores are active. Depending on the processor type, this is usually considerably lower. For maximum Turbo Mode frequency, it is necessary, unless all cores are enabled, to set [CPU C6 Report] to [Enabled]. Using the [Disabled] setting for the BIOS option [CPU C6 Report] only prevents the BIOS from transferring the appropriate CPU C-state via the ACPI to the operating system, which is then usually no longer in a position to use this state. CPU core C-state related BIOS settings will have no effect on some operating systems, notably on Linux distributions that use the "intel_idle" driver (as of 2021, all enterprise Linux distributions supported by Fujitsu). There are two ways to achieve C-State setting you want. The first way is to set the appropriate BIOS C-State options and to disable this driver by using the Linux kernel parameter "intel_idle.max_cstate=0". The Linux kernel will then instead use the acpi standard idle driver that corresponds to the BIOS settings. The second way is to use the Linux command "cpupower", which can set the C-State which the operating system uses regardless of BIOS options.

Reference : Processor Power States



Processor Performance Power State (P-State)

- Known as Enhanced Intel SpeedStep Technology (EIST) or Demand Based Switching (DBS)
- Based on CPU utilization the P-states reduce the electrical power consumption, whereas the processor executes code
- P-states are a combination of processor voltage and processor frequency
- P-states can be compared with various performance levels



Processor Idle Power State (C-State)

- C-states reduce the electrical power consumption if the processor is not executing code
- Parts of the processor can be disabled
- C-0 → Processor active
- C-6 → Processor in deep power down

Package C State limit

BIOS Setup Menu	BIOS Option	Setting	Performance	Low Latency	Energy Efficiency
Configuration >CPU Configuration	Package C State limit	C0 C2 C6 C6 (Retention) No Limit	C0	C0	No Limit

In addition to the CPU or core C-states, there are also so-called package C-states, which not only allow the individual cores of a processor, but the entire processor chip to be put into a type of sleep state. As a result, power consumption is even further reduced. The "waking-up time" that is required to change from the lower package C-states to the active C0 state is even longer in comparison with the CPU or core C-states. If the [C0] setting is made in the BIOS, the processor chip always remains active. However, if it is foreseeable that the server has longer idle periods during operating hours and that latency does not play a role when "waking up" from the package C-states, then the setting should be set to [C6 (Retention)], because this considerably reduces the power consumption of the server in an idle state. The difference between [C6] and [C6 (Retention)] is the voltage, with which the processor is operated in this package C-state. In the case of [C6 (Retention)] the voltage and thus also the power consumption are reduced even further.

UPI Link Frequency Select

BIOS Setup Menu	BIOS Option	Setting	Performance	Low Latency	Energy Efficiency
Configuration >CPU Configuration	UPI Link Frequency Select	Auto 9.6 GT/s 10.4 GT/s 11.2 GT/s	Auto	Auto	9.6 GT/s

Using this BIOS option makes it possible to reduce operating frequencies of the Ultra Path Interconnect (UPI) between the CPUs in a system in order to save power. This particularly makes sense if the available bandwidth is not necessary. However, if the specification is maximum performance and a short response time, the "Auto" setting which automatically sets the highest speed is left unchanged. Depending on which bandwidth is required, a selection can be made here between the speeds [9.6 GT/s], which brings the greatest energy savings, [11.2 GT/s] (10.4 GT/s in RX4770 M6), which is the maximum speed.

UPI Link L0p

BIOS Setup Menu	BIOS Option	Setting	Performance	Low Latency	Energy Efficiency
Configuration >CPU Configuration	UPI Link L0p	Disabled Enabled	Enabled	Disabled	Enabled

This BIOS option enables or disables Intel Ultra Path Interconnect (UPI) Link L0p power saving state. L0p state halts the half of lanes of UPI link when the processor load is low and can save the power.

UPI Link L1

BIOS Setup Menu	BIOS Option	Setting	Performance	Low Latency	Energy Efficiency
Configuration >CPU Configuration	UPI Link L1	Disabled Enabled	Enabled	Disabled	Enabled

This BIOS option enables or disables Intel Ultra Path Interconnect (UPI) Link L1 power saving state. L1 state, which is deeper power saving state than L0p, set all lanes of UPI link to standby mode and can save the power.

Uncore Frequency Scaling

BIOS Setup Menu	BIOS Option	Setting	Performance	Low Latency	Energy Efficiency
Configuration >CPU Configuration	Uncore Frequency Scaling	Disabled Maximum Power Balanced	Disabled	Maximum	Power Balanced

The Xeon Scalable processors work with independent frequencies for the individual cores and the so-called uncore area. Depending on the utilization, the frequencies are set accordingly for each area. This ensures that processors with a high utilization also achieve appropriate performance levels due to high frequencies. On the other hand, the frequencies can be reduced to a minimum if the processor or appropriate areas of a processor are not fully utilized in order to save energy.

The setting of this BIOS option controls the frequency of the uncore area. The standard setting [Disabled] ensures that the uncore frequency is regulated by the processor itself. The frequency can vary between the minimum and the maximum possible uncore frequency according to the current CPU utilization. The possible uncore frequency depends on the processor type used and can accordingly be above or below the nominal frequency of the processor. The [Maximum] setting ensures that the uncore area of the processor always works at its maximum frequency, even if the cores are only slightly utilized or are even in an idle state. The power consumption is also accordingly higher. For this reason, the setting should normally always be set to Disabled for this option. Applications with high demands of I/O latency or generally I/O-intensive applications, which place no load or only a very small load on the processors, are the exceptions. In this situation, the processor's power management mechanisms attempt to reduce the frequency to a minimum. If this happens, the frequency of the so-called uncore area is also automatically lowered. As the entire I/O communication (PCIe, memory, UPI, etc.) is via the uncore area, this would have a negative effect on the I/O throughput. The [Uncore Frequency Scaling = Maximum] setting would prevent this, but the resulting increase in electrical power consumption cannot be avoided. The [Power balanced] setting behaves so that power consumption and performance is balanced.

LLC Dead Line Alloc

BIOS Setup Menu	BIOS Option	Setting	Performance	Low Latency	Energy Efficiency
Configuration >CPU Configuration	LLC Dead Line Alloc	Disabled Enabled	Disabled	Disabled	Disabled

In the Xeon Scalable processor cache scheme, L2 cache evictions (due to no space in L2) are filled into L3 cache. If a cache line is evicted from L2 cache, the core can flag the evicted L2 cache lines as "dead." This means that the lines are not likely to be read again.

If the [LLC Dead Line Alloc] is [Disabled], dead lines will never fill into the L3 cache. This can help save space in the L3 Cache and prevent it from evicting useful data. If the [LLC Dead Line Alloc] is [Enabled], the L3 cache can opportunistically fill dead lines if there is free space available.

Comparative measurements have shown that the disabling of the [LLC Dead Line Alloc] option has minor performance advantages for integer workload. However, the effect depends on application cache usage. Before this option is changed, the effect should first be examined in a test environment.

Stale AtoS

BIOS Setup Menu	BIOS Option	Setting	Performance	Low Latency	Energy Efficiency
Configuration >CPU Configuration	Stale AtoS	Disabled Enabled Auto	Enabled	Enabled	Auto

In the Xeon Scalable Processor, the in-memory directory has three states: I, A, and S. I (invalid) state means the data is clean and does not exist in any other socket's cache. The A (snoopAll) state means the data may exist in another socket in an exclusive or modified state. S (Shared) state means the data is clean and may be shared across one or more socket's caches.

When doing a read to memory, if the directory line is in the A state, we must snoop all the other sockets because another socket may have the line in modified state. If this is the case, the snoop will return the modified data. However, it may be the case that a line is read in A state and all the snoops come back a miss. This can happen if another socket reads the line earlier and then silently dropped it from its cache without modifying it. If the [Stale AtoS] feature is [Enabled], in the situation where a line in A state returns only snoop misses, the line will transition to S state. That way, subsequent reads to the line will encounter it in S state and not have to snoop, saving latency and snoop bandwidth. [Stale AtoS] may be beneficial in a workload where there are many cross-socket reads.

DDR Performance

BIOS Setup Menu	BIOS Option	Setting	Performance	Low Latency	Energy Efficiency
Configuration >Memory Configuration	DDR Performance	Performance optimized Energy optimized Power balanced	Performance optimized	Performance optimized	Energy optimized

This BIOS option controls the speed with which the memory modules are operated. In this respect, it is necessary to weigh up between performance and energy consumption. The "Performance optimized" setting operates the DIMMs with the maximum possible speed, depending on the CPU type used and the memory configuration, and as a result, it provides the highest possible memory performance. The [Energy optimized] setting always restricts the memory frequency to the lowest memory frequency supported. The [Power balanced] setting behaves so that power consumption and performance is balanced.

Patrol Scrub

BIOS Setup Menu	BIOS Option	Setting	Performance	Low Latency	Energy Efficiency
Configuration >Memory Configuration	Patrol Scrub	Disabled Enabled	Enabled	Disabled	Disabled

This BIOS option enables or disables the so-called memory scrubbing, which cyclically accesses the main memory of the system in the background, regardless of the operating system, to detect and correct memory errors in a preventive way. In general workloads, the performance impact is small even if [Patrol Scrub] is enabled. But since the time of this memory test cannot be influenced, it may cause the variability of the performance under certain circumstances. The disabling of the [Patrol Scrub] option increases the probability of discovering memory errors in case of active accesses by the operating system. Until these errors are correctable, the ECC technology of the memory modules ensures that the system continues to run in a stable way. However, too many correctable memory errors increase the risk of discovering non-correctable errors, which then result in a system standstill.

Virtual NUMA

BIOS Setup Menu	BIOS Option	Setting	Performance	Low Latency	Energy Efficiency
Configuration >Memory Configuration	Virtual NUMA	Disabled Enabled	Disabled	Disabled	Disabled

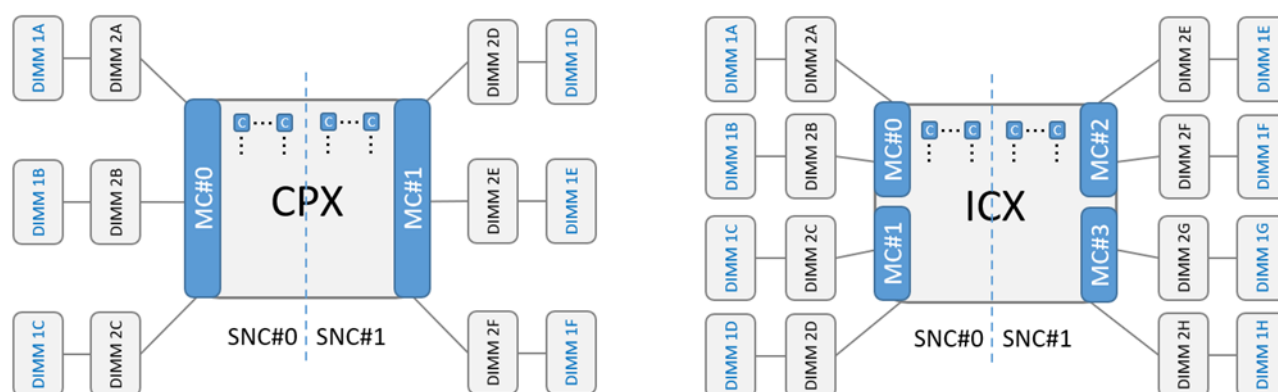
When [SNC(Stub NUMA)] is disabled, only one NUMA (Non-Uniform Memory Access) node is created per processor. In that case, if [Virtual NUMA] is enabled, BIOS divides the physical NUMA node into two virtual NUMA nodes in the ACPI table.

Since the current Windows server OS can handle up to 64 logical processors per NUMA node, the logical processors in excess of 64 are assigned to a NUMA node which doesn't have the memory resource and it results in the inefficient use of the memory. If you use a processor which has more than 64 logical processors with [SNC(Sub NUMA)] disabled, enabling this feature is useful to avoid performance degradation.

SNC(Sub NUMA)

BIOS Setup Menu	BIOS Option	Setting	Performance	Low Latency	Energy Efficiency
Configuration >Memory Configuration	SNC(Sub NUMA)	Enabled Disabled	Enabled	Enabled	Enabled

This BIOS option enables or disables Sub NUMA Clustering feature. [Enabled] breaks up L3 cache into two disjointed clusters based on address range, with each cluster bound to one memory controller and cores in a processor. Memory interleaving is not performed across the NUMA nodes.



Memory architecture and Cooper Lake (Left) and Ice Lake (Right)

Each cluster is seen as one NUMA domain from operating system and SNC improves average "local" L3 cache and memory latency within NUMA node. SNC is specially recommended for NUMA-optimized applications to achieve the lowest possible local memory latency and the highest possible local memory bandwidth.

On the other hand, DIMMs must be equally installed between all memory controllers to enable SNC. Depending on the DIMM installation, this feature may be disabled even when [SNC(Sub NUMA)] is enabled in BIOS menu. In the case that SNC is enabled, performance can get worse if the application is not NUMA optimized. Moreover, it should be noted that if some cores are disabled in [Active Processor Cores] option, NUMA nodes having less cores or no core may be created and cause performance degradation.

UMA-Based Clustering

BIOS Setup Menu	BIOS Option	Setting	Performance	Low Latency	Energy Efficiency
Configuration >Memory Configuration	UMA-Based Clustering	Disabled Hemisphere	Disabled	Disabled	Disabled

This BIOS option enables or disables UMA (Unified Memory Access) based clustering feature when [SNC(Sub NUMA)] is disabled. If [UMA-Based Clustering] was set to [Hemisphere], BIOS creates two clusters which the operating system isn't aware of. Unlike the SNC, the system address range and the cores in the processor are not divided into the clusters but the L3 cache and memory controllers are bound to either of clusters based on each affinity. Since this improves the access performance between L3 cache and memory controllers, this feature can improve the performance in UMA based workloads without the awareness of OS.

If DIMM installation is not bilaterally symmetric in Figure in SNC(Sub NUMA), this feature cannot be enabled.

Appendix

Based on the profile you selected in [Application Profile] option, the following BIOS settings are automatically selected (in BIOS menu, modified settings are also visible), depending on the profile selected. Any BIOS settings that are not listed in the table and the blank spaces in the table will not be changed from the existing settings. After selecting the profile that most closely matches your workload in this BIOS option, you can override and change any BIOS options individually, including settings automatically changed in the [Application Profile] option, as needed. The settings take effect after you save and restart.

For BIOS options not included in this white paper, refer to the "BIOS Setup Utility" manual for your specific model from the support pages listed in the related publications.

Settings of Application Profile option for RX2530 M6 / RX2540 M6 (1/2)

Option Name	Default	Total Throughput Performance	Single Thread Performance	Energy Efficiency	Virtualization Performance	Low Latency
CPU Configuration						
Hyper-Threading	Enabled	Enabled	Enabled	Enabled	Enabled	Disabled
Hardware Prefetcher	Enabled	Enabled	Enabled			Enabled
Adjacent Cache Line Prefetch	Enabled	Enabled	Enabled			Enabled
DCU Streamer Prefetcher	Enabled	Disabled	Enabled	Disabled		Enabled
DCU Ip Prefetcher	Enabled	Enabled	Enabled	Enabled		Enabled
Intel Virtualization Technology	Enabled	Enabled	Enabled	Enabled	Enabled	Disabled
Intel(R) VT-d	Enabled	Enabled	Enabled	Enabled	Enabled	Enabled
Total Memory Encryption (TME)	Disabled	Disabled	Disabled	Disabled		Disabled
Enhanced SpeedStep	Enabled	Enabled	Enabled	Enabled	Enabled	Enabled
Turbo Mode	Enabled	Enabled	Enabled	Enabled	Enabled	Enabled
Energy Performance	Balanced Performance	Balanced Performance	Balanced Performance	Energy Efficient	Balanced Performance	Performance
Override OS Energy Performance	Disabled	Disabled	Disabled	Disabled	Disabled	Enabled
Utilization Profile	Even	Even	Even	Even		Unbalanced
HWPM Support	Native Mode	Native Mode	Native Mode	Native Mode		Disabled
CPU C1E Support	Enabled	Disabled	Disabled	Enabled		Disabled
CPU C6 Report	Enabled	Enabled	Enabled	Enabled		Disabled
Package C State limit	C0	C2	C0	C6	C0	C0
CPU C1 auto demotion	Enabled	Enabled	Enabled	Enabled		Enabled
CPU C1 auto undemotion	Enabled	Enabled	Enabled	Enabled		Enabled
UPI Link Frequency Select	Auto	Auto	Auto	9.6 GT/s	Auto	Auto
UPI Link L0p	Enabled	Enabled	Enabled	Enabled		Enabled
UPI Link L1	Enabled	Enabled	Enabled	Enabled		Enabled
Local x2APIC	Enabled	Enabled	Enabled	Enabled		Enabled
IODC Configuration	Auto	Auto	Auto	Auto		Auto
Uncore Frequency Scaling	Disabled	Disabled	Disabled	Power Balanced		Disabled
Stale AtoS	Auto	Auto	Auto	Auto		Enabled
LLC Dead Line Alloc	Enabled	Enabled	Enabled	Enabled		Disabled
AVX ICCP pre-grant level	no override	no override	no override	no override		no override
XPT Prefetch	Disabled	Enabled	Disabled	Disabled		Disabled
XPT Remote Prefetch	Auto	Auto	Auto	Auto		Auto
UPI Prefetch	Enabled	Disabled	Enabled	Enabled		Enabled
L2 RFO Prefetch	Enabled	Enabled	Enabled	Enabled		Enabled
Monitor MWAIT	Enabled	Enabled	Enabled	Enabled	Enabled	Enabled
LLC Prefetch	Disabled	Enabled	Disabled	Disabled		Disabled
Memory Configuration						
Memory Mode	Independent	Independent	Independent	Independent	Independent	Independent
Partial Cache Line Sparing (PCLS)	Disabled	Disabled	Disabled	Disabled	Disabled	Disabled
ADDC Sparing	Disabled	Disabled	Disabled	Disabled	Disabled	Disabled
NUMA	Enabled	Enabled	Enabled	Enabled	Enabled	Enabled
Virtual NUMA	Disabled	Disabled		Disabled		Disabled

	DDR Performance	Performance Optimized	Performance optimized	Performance optimized	Power balanced	Performance optimized	Performance optimized
	Patrol Scrub	Disabled	Disabled	Disabled	Disabled	Disabled	Disabled
	SNC(Sub NUMA)	Disabled	Enable SNC2	Disabled	Enable ENC2		Enable ENC2
	UMA-Based Clustering	Hemisphere	Hemisphere	Hemisphere	Hemisphere	Hemisphere	Hemisphere
PCI Subsystem Configuration							
	SR-IOV Support	Enabled	Enabled	Enabled	Enabled	Enabled	

Settings of Application Profile option for RX2530 M6 / RX2540 M6 (2/2)

Option Name	Default	Online Transaction Processing	Online Analytical Processing In-Memory DB	I/O Throughput	Memory Intensive HPC	CPU Intensive HPC
CPU Configuration						
Hyper-Threading	Enabled	Enabled	Enabled		Enabled	
Hardware Prefetcher	Enabled		Enabled		Enabled	Enabled
Adjacent Cache Line Prefetch	Enabled		Enabled		Enabled	Enabled
DCU Streamer Prefetcher	Enabled		Disabled		Enabled	Enabled
DCU Ip Prefetcher	Enabled		Enabled		Enabled	Enabled
Intel Virtualization Technology	Enabled		Enabled	Disabled	Disabled	Disabled
Intel(R) VT-d	Enabled		Enabled	Disabled	Enabled	Enabled
Total Memory Encryption (TME)	Disabled		Disabled		Disabled	Disabled
Enhanced SpeedStep	Enabled	Enabled	Enabled		Enabled	Enabled
Turbo Mode	Enabled	Enabled	Enabled		Enabled	Enabled
Energy Performance	Balanced Performance	Performance	Performance	Performance	Performance	Performance
Override OS Energy Performance	Disabled	Disabled	Disabled	Enabled	Enabled	Enabled
Utilization Profile	Even	Unbalanced	Even	Unbalanced	Even	Unbalanced
HWPM Support	Native Mode		Native Mode with no legacy		Disabled	Disabled
CPU C1E Support	Enabled		Disabled	Disabled	Enabled	Disabled
CPU C6 Report	Enabled		Enabled	Disabled	Enabled	Disabled
Package C State limit	C0	C0	C0	C0	C0	C0
CPU C1 auto demotion	Enabled		Enabled	Enabled	Enabled	Enabled
CPU C1 auto undemotion	Enabled		Enabled	Enabled	Enabled	Enabled
UPI Link Frequency Select	Auto	Auto	9.6 GT/s		Auto	Auto
UPI Link L0p	Enabled		Enabled		Enabled	Enabled
UPI Link L1	Enabled		Enabled		Enabled	Enabled
Local x2APIC	Enabled		Enabled		Enabled	Enabled
IODC Configuration	Auto		Auto		Auto	Auto
Uncore Frequency Scaling	Disabled	Maximum	Disabled	Maximum	Disabled	Disabled
Stale AtoS	Auto		Enabled		Enabled	Enabled
LLC Dead Line Alloc	Enabled		Enabled		Disabled	Disabled
AVX ICCP pre-grant level	no override		no override		no override	no override
XPT Prefetch	Disabled		Enabled		Disabled	Disabled
XPT Remote Prefetch	Auto		Auto		Auto	Auto
UPI Prefetch	Enabled		Enabled		Enabled	Enabled
L2 RFO Prefetch	Enabled		Enabled		Enabled	Enabled
Monitor MWAIT	Enabled	Enabled	Enabled		Enabled	Enabled
LLC Prefetch	Disabled		Disabled		Disabled	Disabled
Memory Configuration						
Memory Mode	Independent	Independent	Independent		Independent	Independent
Partial Cache Line Sparing (PCLS)	Disabled	Disabled	Disabled		Disabled	Disabled
ADDC Sparing	Disabled	Disabled	Disabled		Disabled	Disabled
NUMA	Enabled	Enabled	Enabled	Enabled	Enabled	Enabled

Virtual NUMA	Disabled	Disabled			Disabled	
DDR Performance	Performance Optimized	Performance optimized	Performance optimized		Performance optimized	Performance optimized
Patrol Scrub	Disabled	Disabled	Enabled		Disabled	Disabled
SNC(Sub NUMA)	Disabled	Enable ENC2	Disabled		Enable SNC2	Disabled
UMA-Based Clustering	Hemisphere	Hemisphere	Hemisphere		Hemisphere	Hemisphere

PCI Subsystem Configuration

SR-IOV Support	Enabled	Enabled	Disabled			
----------------	---------	---------	----------	--	--	--

Settings of Application Profile option for CX2550 M6 / CX2560 M6 (1/2)

Option Name	Default	Total Throughput Performance	Single Thread Performance	Energy Efficiency	Virtualization Performance	Low Latency
CPU Configuration						
Hyper-Threading	Enabled	Enabled	Enabled	Enabled	Enabled	Disabled
Hardware Prefetcher	Enabled	Enabled	Enabled			Enabled
Adjacent Cache Line Prefetch	Enabled	Enabled	Enabled			Enabled
DCU Streamer Prefetcher	Enabled	Disabled	Enabled	Disabled		Enabled
DCU Ip Prefetcher	Enabled	Enabled	Enabled	Enabled		Enabled
Intel Virtualization Technology	Enabled	Enabled	Enabled	Enabled	Enabled	Disabled
Intel(R) VT-d	Enabled	Enabled	Enabled	Enabled	Enabled	Enabled
Total Memory Encryption (TME)	Disabled	Disabled	Disabled	Disabled		Disabled
Enhanced SpeedStep	Enabled	Enabled	Enabled	Enabled	Enabled	Enabled
Turbo Mode	Enabled	Enabled	Enabled	Enabled	Enabled	Enabled
Energy Performance	Balanced Performance	Balanced Performance	Balanced Performance	Energy Efficient	Balanced Performance	Performance
Override OS Energy Performance	Disabled	Disabled	Disabled	Disabled	Disabled	Enabled
Utilization Profile	Even	Even	Even	Even		Unbalanced
HWPM Support	Native Mode	Native Mode	Native Mode	Native Mode		Disabled
CPU C1E Support	Enabled	Enabled	Disabled	Enabled		Disabled
CPU C6 Report	Enabled	Enabled	Enabled	Enabled		Disabled
Package C State limit	C0	C0	C0	C6	C0	C0
CPU C1 auto demotion	Enabled	Enabled	Enabled	Enabled		Enabled
CPU C1 auto undemotion	Enabled	Enabled	Enabled	Enabled		Enabled
UPI Link Frequency Select	Auto	Auto	Auto	9.6 GT/s		Auto
UPI Link L0p	Enabled	Enabled	Enabled	Enabled		Enabled
UPI Link L1	Enabled	Enabled	Enabled	Enabled		Enabled
Local x2APIC	Enabled	Enabled	Enabled	Enabled		Enabled
IODC Configuration	Auto	Auto	Auto	Auto		Auto
Uncore Frequency Scaling	Disabled	Disabled	Disabled	Power Balanced		Disabled
Stale AtoS	Auto	Auto	Auto	Auto		Enabled
LLC Dead Line Alloc	Enabled	Enabled	Enabled	Enabled		Disabled
AVX ICCP pre-grant level	no override	no override	no override	no override		no override
XPT Prefetch	Disabled	Disabled	Disabled	Disabled		Disabled
XPT Remote Prefetch	Auto	Auto	Auto	Auto		Auto
UPI Prefetch	Enabled	Enabled	Enabled	Enabled		Enabled
L2 RFO Prefetch	Enabled	Enabled	Enabled	Enabled		Enabled
Monitor MWAIT	Enabled	Enabled	Enabled	Enabled	Enabled	Enabled
LLC Prefetch	Disabled	Enabled	Disabled	Disabled		Disabled
Memory Configuration						
Memory Mode	Independent	Independent	Independent	Independent	Independent	Independent
Partial Cache Line Sparing (PCLS)	Disabled	Disabled	Disabled	Disabled	Disabled	Disabled
ADDC Sparing	Disabled	Disabled	Disabled	Disabled	Disabled	Disabled
NUMA	Enabled	Enabled	Enabled	Enabled	Enabled	Enabled
Virtual NUMA	Disabled	Disabled		Disabled		Disabled

	DDR Performance	Performance Optimized	Performance optimized	Performance optimized	Energy Optimized	Performance optimized	Performance optimized
	Patrol Scrub	Disabled	Disabled	Disabled	Disabled	Disabled	Disabled
	SNC(Sub NUMA)	Disabled	Enable SNC2	Disabled	Enable ENC2		Enable ENC2
	UMA-Based Clustering	Hemisphere	Hemisphere	Hemisphere	Hemisphere	Hemisphere	Hemisphere
PCI Subsystem Configuration							
	SR-IOV Support	Enabled	Enabled	Enabled	Enabled	Enabled	

Settings of Application Profile option for CX2550 M6 / CX2560 M6 (2/2)

Option Name	Default	Online Transaction Processing	Online Analytical Processing In-Memory DB	I/O Throughput	Memory Intensive HPC	CPU Intensive HPC
CPU Configuration						
Hyper-Threading	Enabled	Enabled	Enabled		Enabled	
Hardware Prefetcher	Enabled		Enabled		Enabled	Enabled
Adjacent Cache Line Prefetch	Enabled		Enabled		Enabled	Enabled
DCU Streamer Prefetcher	Enabled		Disabled		Enabled	Enabled
DCU Ip Prefetcher	Enabled		Enabled		Enabled	Enabled
Intel Virtualization Technology	Enabled		Enabled	Disabled	Disabled	Disabled
Intel(R) VT-d	Enabled		Enabled	Disabled	Enabled	Enabled
Total Memory Encryption (TME)	Disabled		Disabled		Disabled	Disabled
Enhanced SpeedStep	Enabled	Enabled	Enabled		Enabled	Enabled
Turbo Mode	Enabled	Enabled	Enabled		Enabled	Enabled
Energy Performance	Balanced Performance	Performance	Performance	Performance	Balanced Performance	Performance
Override OS Energy Performance	Disabled	Disabled	Disabled	Enabled	Enabled	Enabled
Utilization Profile	Even	Unbalanced	Even	Unbalanced	Even	Unbalanced
HWPM Support	Native Mode		Native Mode with no legacy		Disabled	Disabled
CPU C1E Support	Enabled		Disabled	Disabled	Enabled	Disabled
CPU C6 Report	Enabled		Enabled	Disabled	Enabled	Disabled
Package C State limit	C0	C0	C0	C0	C0	C0
CPU C1 auto demotion	Enabled		Enabled	Enabled	Enabled	Enabled
CPU C1 auto undemotion	Enabled		Enabled	Enabled	Enabled	Enabled
UPI Link Frequency Select	Auto	Auto	9.6GT/s		Auto	Auto
UPI Link L0p	Enabled		Enabled		Enabled	Enabled
UPI Link L1	Enabled		Enabled		Enabled	Enabled
Local x2APIC	Enabled		Enabled		Enabled	Enabled
IODC Configuration	Auto		Auto		Auto	Auto
Uncore Frequency Scaling	Disabled	Maximum	Disabled	Maximum	Disabled	Disabled
Stale AtoS	Auto		Enabled		Enabled	Enabled
LLC Dead Line Alloc	Enabled		Enabled		Disabled	Disabled
AVX ICCP pre-grant level	no override		no override		no override	no override
XPT Prefetch	Disabled		Enabled		Enabled	Disabled
XPT Remote Prefetch	Auto		Auto		Auto	Auto
UPI Prefetch	Enabled		Enabled		Enabled	Enabled
L2 RFO Prefetch	Enabled		Enabled		Enabled	Enabled
Monitor MWAIT	Enabled	Enabled	Enabled		Enabled	Enabled
LLC Prefetch	Disabled		Disabled		Disabled	Disabled
Memory Configuration						
Memory Mode	Independent	Independent	Independent		Independent	Independent
Partial Cache Line Sparing (PCLS)	Disabled	Disabled	Disabled		Disabled	Disabled
ADDC Sparing	Disabled	Disabled	Disabled		Disabled	Disabled
NUMA	Enabled	Enabled	Enabled	Enabled	Enabled	Enabled

Virtual NUMA	Disabled	Disabled			Disabled	
DDR Performance	Performance Optimized	Performance optimized	Performance optimized		Performance optimized	Performance optimized
Patrol Scrub	Disabled	Disabled	Enabled		Disabled	Disabled
SNC(Sub NUMA)	Disabled	Enable ENC2	Disabled		Enable SNC2	Disabled
UMA-Based Clustering	Hemisphere	Hemisphere	Hemisphere		Hemisphere	Hemisphere
PCI Subsystem Configuration						
SR-IOV Support	Enabled	Enabled	Disabled		Enabled	

Settings of Application Profile option for RX4770 M6 (1/2)

Option Name	Default	Total Throughput Performance	Single Thread Performance	Energy Efficiency	Virtualization Performance	Low Latency
-------------	---------	------------------------------	---------------------------	-------------------	----------------------------	-------------

CPU Configuration

Hyper-Threading	Enabled	Enabled	Disabled	Enabled	Enabled	Disabled
Hardware Prefetcher	Enabled	Enabled	Enabled	Disabled		Enabled
Adjacent Cache Line Prefetch	Enabled	Enabled	Enabled	Disabled		Enabled
DCU Streamer Prefetcher	Enabled	Disabled	Enabled	Disabled		Enabled
DCU Ip Prefetcher	Enabled	Enabled	Enabled	Enabled		Enabled
Intel Virtualization Technology	Enabled	Disabled	Enabled	Disabled	Enabled	Disabled
Intel(R) VT-d	Enabled	Enabled	Enabled	Enabled	Enabled	Enabled
Enhanced SpeedStep	Enabled	Enabled	Enabled	Enabled	Enabled	Enabled
Turbo Mode	Enabled	Enabled	Enabled	Disabled	Enabled	Disabled
Energy Performance	Balanced Performance	Balanced Performance	Performance	Energy Efficient	Performance	Performance
Override OS Energy Performance	Disabled	Disabled	Enabled	Enabled	Disabled	Enabled
Utilization Profile	Even	Unbalanced	Even	Even		Unbalanced
P-State Coordination	HW_ALL	HW_ALL	HW_ALL	HW_ALL		HW_ALL
HWPM Support	Native Mode	Native Mode	OOB Mode	Disabled		Disabled
CPU C1E Support	Enabled	Enabled	Enabled	Enabled		Disabled
CPU C6 Report	Enabled	Enabled	Enabled	Enabled		Disabled
Package C State limit	C0	C0	C0	No limit	C0	C0
UPI Link Frequency Select	Auto	Auto	Auto	9.6GT/s		Auto
UPI Link L0p	Enabled	Enabled	Enabled	Enabled		Enabled
UPI Link L1	Enabled	Enabled	Enabled	Enabled		Enabled
Local x2APIC	Enabled	Enabled	Enabled	Disabled		Enabled
IODC Configuration	Auto	Auto	Auto	Auto		Auto
Uncore Frequency Scaling	Disabled	Disabled	Disabled	Power balanced		Maximum
Stale AtoS	Auto	Disabled	Disabled	Auto		Enabled
LLC Dead Line Alloc	Enabled	Disabled	Enabled	Enabled		Disabled
AVX ICCP pre-grant level	no override	no override	no override	no override		no override
XPT Prefetch	Disabled	Enabled	Disabled	Disabled		Disabled
L2 RFO Prefetch	Enabled	Enabled	Enabled	Enabled		Enabled
Monitor MWAIT	Enabled	Enabled	Enabled	Enabled		Enabled
LLC Prefetch	Disabled	Disabled	Disabled	Disabled		Disabled

Memory Configuration

Memory Mode	Independent	Independent	Independent	Independent	Independent	Independent
ADDC Sparing	Disabled	Disabled	Disabled	Disabled	Disabled	Disabled
NUMA	Enabled	Enabled	Enabled	Enabled	Enabled	Enabled
DDR Performance	Performance optimized	Performance optimized	Performance optimized	Power Balanced	Performance optimized	Performance optimized
Patrol Scrub	Disabled	Enabled	Disabled	Disabled	Disabled	Disabled
SNC (Sub NUMA)	Disabled	Enabled	Enabled	Enabled	Enabled	Enabled

PCI Subsystem Configuration

SR-IOV Support	Enabled				Enabled	
----------------	---------	--	--	--	---------	--

Settings of Application Profile option for RX4770 M6 (2/2)

Option Name	Default	Online Transaction Processing	Online Analytical Processing In-Memory DB	I/O Throughput	Memory Intensive HPC	CPU Intensive HPC
-------------	---------	-------------------------------	---	----------------	----------------------	-------------------

CPU Configuration

Hyper-Threading	Enabled	Enabled	Enabled		Disabled	
Hardware Prefetcher	Enabled		Enabled		Enabled	Enabled
Adjacent Cache Line Prefetch	Enabled		Enabled		Enabled	Enabled
DCU Streamer Prefetcher	Enabled		Enabled		Enabled	Enabled
DCU Ip Prefetcher	Enabled		Enabled		Enabled	Enabled
Intel Virtualization Technology	Enabled		Enabled	Disabled	Disabled	Disabled
Intel(R) VT-d	Enabled		Enabled	Disabled	Enabled	Enabled
Enhanced SpeedStep	Enabled	Enabled	Enabled		Enabled	Enabled
Turbo Mode	Enabled	Enabled	Enabled		Disabled	Disabled
Energy Performance	Balanced Performance	Performance	Performance	Performance	Performance	Performance
Override OS Energy Performance	Disabled	Enabled	Disabled	Enabled	Enabled	Enabled
Utilization Profile	Even	Unbalanced	Even	Unbalanced	Unbalanced	Unbalanced
P-State Coordination	HW_ALL		HW_ALL		HW_ALL	HW_ALL
HWPM Support	Native Mode	Disabled	Native Mode with no legacy		Disabled	Disabled
CPU C1E Support	Enabled	Disabled	Disabled		Disabled	Disabled
CPU C6 Report	Enabled	Disabled	Enabled	Disabled	Disabled	Disabled
Package C State limit	C0	C0	C0		C0	C0
UPI Link Frequency Select	Auto	Auto	Auto		Auto	Auto
UPI Link L0p	Enabled		Enabled		Enabled	Enabled
UPI Link L1	Enabled		Enabled		Enabled	Enabled
Local x2APIC	Enabled		Enabled		Enabled	Enabled
IODC Configuration	Auto		Auto		Auto	Auto
Uncore Frequency Scaling	Disabled	Maximum	Disabled		Disabled	Disabled
Stale AtoS	Auto		Enabled		Enabled	Enabled
LLC Dead Line Alloc	Enabled		Enabled		Disabled	Disabled
AVX ICCP pre-grant level	no override		no override		no override	no override
XPT Prefetch	Disabled		Enabled		Disabled	Disabled
L2 RFO Prefetch	Enabled		Enabled		Enabled	Enabled
Monitor MWAIT	Enabled		Enabled		Enabled	Enabled
LLC Prefetch	Disabled		Disabled		Disabled	Disabled

Memory Configuration

Memory Mode	Independent	Independent	Independent		Independent	Independent
ADDC Sparing	Disabled	Disabled	Disabled		Disabled	Disabled
NUMA	Enabled	Enabled	Enabled	Enabled	Enabled	Enabled
DDR Performance	Performance optimized	Performance optimized	Performance optimized		Performance optimized	Performance optimized
Patrol Scrub	Disabled	Disabled	Enabled		Disabled	Disabled
SN(C)Sub NUMA	Disabled	Enabled	Disabled		Enabled	Disabled

PCI Subsystem Configuration

SR-IOV Support	Enabled					
----------------	---------	--	--	--	--	--


Literature

PRIMERGY Servers

<https://www.fujitsu.com/global/products/computing/servers/primergy/>

BIOS optimization for 3rd Generation Xeon Scalable Processor-based systems

This Whitepaper

 <https://docs.ts.fujitsu.com/dl.aspx?id=3cf2d5b2-f58a-4308-ac63-d33adf7230d2>

 <https://docs.ts.fujitsu.com/dl.aspx?id=696984c2-7a49-4b64-ba34-77888c8a68d6>

PRIMERGY Performance

<https://www.fujitsu.com/global/products/computing/servers/primergy/benchmarks/>

PRIMERGY Manuals

Support Site:

<https://support.ts.fujitsu.com/>

You can download "BIOS Setup Utility" by searching the following document name per model.

- RX2530 M6 BIOS Setup Utility: "D3890 BIOS Setup Utility"
- RX2540 M6 BIOS Setup Utility: "D3891 BIOS Setup Utility"
- CX2550 M6/ CX2560 M6 BIOS Setup Utility: "D3893/D3894 BIOS Setup Utility"
- TX2550 M6 BIOS Setup Utility: "D3892 BIOS Setup Utility"

Operating System Performance Tuning Guidelines

- Microsoft Windows:

<https://docs.microsoft.com/en-us/windows-server/administration/performance-tuning/>

- RedHat Linux:

https://access.redhat.com/documentation/en-us/red_hat_enterprise_linux/7/html/performance_tuning_guide/index

https://access.redhat.com/documentation/en-us/red_hat_enterprise_linux/8/html/monitoring_and_managing_system_status_and_performance/index

https://access.redhat.com/documentation/en-us/red_hat_enterprise_linux/9/html/monitoring_and_managing_system_status_and_performance/index

- SUSE Linux:

- <https://documentation.suse.com/sbp/all/html/SBP-performance-tuning/index.html>
- <https://documentation.suse.com/sles/15-SP3/html/SLES-all/book-tuning.html>
- <https://documentation.suse.com/sles/15-SP4/html/SLES-all/book-tuning.html>
- <https://documentation.suse.com/sles/15-SP5/html/SLES-all/book-tuning.html>

- VMware vSphere:

<https://www.vmware.com/files/pdf/techpaper/VMW-Tuning-Latency-Sensitive-Workloads.pdf>

<https://www.vmware.com/techpapers/2019/vsphere-esxi-vcenter-server-67U2-performance-best-practices.html>

<https://www.vmware.com/techpapers/2022/tagging-vsphere70u1-perf.html>

<https://www.vmware.com/techpapers/2021/vsphere-esxi-vcenter-server-70U2-performance-best-practices.html>

<https://www.vmware.com/techpapers/2022/vsphere-esxi-vcenter-server-70U3-performance-best-practices.html>

<https://www.vmware.com/techpapers/2022/vsphere-esxi-vcenter-server-80-performance-best-practices.html>

<https://www.vmware.com/techpapers/2023/vsphere-esxi-vcenter-server-80U1-performance-best-practices.html>

Document change history

Version	Date	Description
1.5	2023-10-03	New Visual Identity format Minor correction
1.4	2022-06-07	Fixed some wrong option names of Application Profile menu
1.3	2022-04-06	Add the list of the settings of Application Profile menu
1.2	2021-11-01	Add the description for 2 socket models
1.1	2021-07-28	Update contact information and URLs Minor correction
1.0	2021-02-05	First edition

Contact

Fujitsu

Web site: <https://www.fujitsu.com>

PRIMERGY Performance and Benchmarks

<mailto:fj-benchmark@dl.jp.fujitsu.com>

© Fujitsu 2023. All rights reserved. Fujitsu and Fujitsu logo are trademarks of Fujitsu Limited registered in many jurisdictions worldwide. Other product, service and company names mentioned herein may be trademarks of Fujitsu or other companies. This document is current as of the initial date of publication and subject to be changed by Fujitsu without notice. This material is provided for information purposes only and Fujitsu assumes no liability related to its use.