White Paper
FUJITSU Server PRIMERGY
BIOS optimizations for Xeon E5-2600 v3 based systems

This document explains the BIOS settings that are valid for the Intel Xeon E5-2600 v3 based PRIMERGY server generation (PRIMERGY BX2560 M1, BX2580 M1, CX2550 M1, CX2570 M1, RX2510 M1, RX2530 M1, RX2540 M1, RX2560 M1, TX2560 M1).

Its purpose is to optimize BIOS settings according to requirements. The objectives here are to optimize PRIMERGY servers for best performance and maximum energy efficiency. In addition to optimization for maximum throughput, application scenarios are also taken into account, in which the shortest possible response time matters.

Version
1.2
2015-11-26

Contents

Document history

Version 1.0

First edition

Version 1.1

- Description for BIOS option “CPU C3/C6 Report” extended
- Minor correction

Version 1.2

- PRIMERGY RX2510 M1 added

Overview

When they leave the factory, Fujitsu PRIMERGY servers 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 deviate from standard settings and thus configure the server - depending on requirements - for the maximum possible throughput (performance), the minimum possible latency (low latency), or the maximum possible energy saving (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 pure BIOS settings, the entire system must also be considered when optimizing PRIMERGY servers. The following aspects should be given particular consideration when planning server systems:

- Server hardware
  - Processor: Processor type and frequency
  - Memory: Memory type and memory configuration
  - I/O cards: Optimal distribution of several cards over PCIe slots
- Operating system and application software
  - Power plan: Performance or energy efficiency
  - Tuning: Kernel, registry, interrupt binding, thread splitting
- Network
  - Network technology: 1/10/40 Gbit Ethernet, Fibre Channel, Infiniband, RDMA
  - Network architecture: Switches, multichannel
- Storage
  - Technology: RAID, Fibre Channel, Direct Attached
  - Disks: HDD, SSD, SATA, SAS

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 2-socket PRIMERGY servers based on the Intel Xeon E5-2600 v3 processor generation deliver the highest levels of performance, as proven by the numerous benchmark publications of the Standard Performance Evaluation Corporation (SPEC), SAP, or the Transaction Processing Performance Council (TPC). When you talk about 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 new parallel processor generation. Although PRIMERGY servers with standard settings already provide an optimal ratio between performance and energy efficiency, it is possible to further optimize the system as regards performance and to a lesser degree energy efficiency via the BIOS. Basically, this optimization is a matter of operating all the components in the system at the maximum speed possible and of preventing the energy-saving options from slowing down the system. This is why 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 more with increasing the speed of each individual transaction, i.e. of 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 - if e.g. 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 reduce the minimal fluctuations in performance of computing operations 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 generation, 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 higher such an 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 reduced performance to maximum performance. This time worsens the latency of the system, particularly if a burst of transactions is pending after an idle period, or if the system is utilized irregularly. This document explains how to configure the power saving modes for users from the low-latency segment in order to minimize system latency. The optimization of server latency, particularly in an idle state, always results in substantially 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 throughout 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 - in conjunction with optimally set I/O components - almost always provide the maximum possible values for these components. 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 Override" at "Enabled" 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 it is not pure performance that plays the greatest role, but energy consumption. Two different objectives are pursued in this respect.

On the one hand, it is possible 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 pursue the aim of reducing power consumption per rack and per server respectively with performance only playing a subordinate role. Optimization in this direction consists primarily of reducing the speed and thus the performance of the server.

On the other hand, it is possible 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 *SAP Server Power Benchmark* and *VMmark V2 Performance with Server Power* prove that PRIMERGY servers are the best choice when it comes to energy-efficient servers.

PRIMERGY BIOS options

This white paper contains information about BIOS options that are valid for the Intel Xeon E5-2600 v3 based PRIMERGY servers. And these are:

- PRIMERGY BX2560 M1
- PRIMERGY BX2580 M1
- PRIMERGY CX2550 M1
- PRIMERGY CX2570 M1
- PRIMERGY RX2510 M1
- PRIMERGY RX2530 M1
- PRIMERGY RX2540 M1
- PRIMERGY RX2560 M1
- PRIMERGY TX2560 M1

The BIOS of the PRIMERGY servers is being continuously developed. This is why it is important to use the latest BIOS version in each case so as to have all the BIOS functions listed here available. Appropriate downloads are available in the Internet under http://www.fujitsu.com/fts/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 expressly recommended to observe the footnotes and subsequent description 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 under the following links.

Microsoft Windows:	http://msdn.microsoft.com/en-us/library/windows/hardware/dn529134
RedHat Linux:	https://access.redhat.com/articles/221153 https://access.redhat.com/documentation/en-US/Red_Hat_Enterprise_Linux/7/html/Performance_Tuning_Guide/
SUSE Linux:	https://www.suse.com/documentation/sles11/pdfdoc/book_sle_tuning/book_sle_tuning.pdf
VMware vSphere:	http://www.vmware.com/files/pdf/techpaper/VMW-Tuning-Latency-Sensitive-Workloads.pdf

Table 1: Overview BIOS options

BIOS Setup Menu	BIOS Option	Settings [1]	Performance	Low Latency	Energy Efficiency
Advanced > PCI Subsystem Settings	ASPM Support	**Disabled** L1 Only	Disabled	Disabled	L1 Only
Advanced > PCI Subsystem Settings	DMI Control	GEN 1 **GEN 2**	GEN 2	GEN 2	GEN 1 [2]
Advanced > CPU Configuration	Hyper-Threading	Disabled **Enabled**	Enabled	Disabled [3]	Enabled
Advanced > CPU Configuration	Active Processor Cores	**0 (All)** [1 – n]	0 (All)	1 – n [4]	0 (All)
Advanced > CPU Configuration	[Hardware] [Adjacent Cache Line] [DCU Streamer] [DCU Ip] Prefetcher	**Enabled** Disabled	Enabled	Enabled	Disabled [5]
Advanced > CPU Configuration	Intel Virtualization Technology	Disabled **Enabled**	Disabled [6]	Disabled	Disabled
Advanced > CPU Configuration	Power Technology	Disabled **Energy Efficient** Custom	Custom	Custom	Custom
Advanced > CPU Configuration	Enhanced SpeedStep [7]	Disabled **Enabled**	Enabled	Enabled	Enabled
Advanced > CPU Configuration	Turbo Mode [7]	Disabled **Enabled**	Enabled	Disabled [8]	Disabled
Advanced > CPU Configuration	Override OS Energy Performance [7]	**Disabled** Enabled	Enabled	Enabled	Disabled [9]
Advanced > CPU Configuration	Energy Performance [10]	Performance **Balanced Performance** Balanced Energy Energy Efficient	Performance	Performance	Energy Efficient
Advanced > CPU Configuration	Utilization Profile [10]	**Even** Unbalanced	Even	Unbalanced	Even
Advanced > CPU Configuration	CPU C1E Support [7]	Disabled **Enabled**	Enabled	Disabled	Enabled
Advanced > CPU Configuration	CPU C3 Report [7]	**Disabled** Enabled	Disabled	Disabled	Enabled

[1] The setting in bold print is the standard value.

[2] GEN 1 is recommended for low chipset I/O load (USB, onboard SATA and onboard LAN for servers of the CX model range); otherwise the setting should be GEN 2.

[3] Hyper-Threading doubles the number of logical cores, but can also result in performance fluctuations in computing operations. Disabling can improve latency.

[4] 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.

[5] The disabling of the prefetchers only increases energy efficiency if performance remains the same or improves. This should be checked in advance.

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

[7] This option is only visible if the setting for "Power Technology" is changed to "Custom".

[8] 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.

[9] If the operating system in use is able to set the "energy efficient policy" for the CPUs, 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 setting via the BIOS.

[10] This option can only be set if the setting for "Override OS Energy Performance" is changed to "Enabled".

BIOS Setup Menu	BIOS Option	Settings [1]	Performance	Low Latency	Energy Efficiency
Advanced > CPU Configuration	CPU C6 Report [7]	Disabled **Enabled**	Enabled	Disabled	Enabled
Advanced > CPU Configuration	Package C State limit [7]	C0 C2 C6 **C6(Retention)**	C0	C0	C6(Retention)
Advanced > CPU Configuration	QPI Link Frequency Select	6.4 GT/s 8.0 GT/s 9.6 GT/s **Auto**	Auto	Auto	6.4 GT/s
Advanced > CPU Configuration	Uncore Frequency Override	**Disabled** Enabled	Disabled [11]	Enabled	Disabled
Advanced > CPU Configuration	COD Enable [12]	Disabled Enabled **Auto**	Auto	Auto	Auto
Advanced > CPU Configuration	Early Snoop	Disabled Enabled **Auto**	Auto	Auto	Auto
Advanced > Memory Configuration	DDR Performance	**Performance optimized** Energy optimized	Performance optimized	Performance optimized	Energy optimized
Advanced > Memory Configuration	Patrol Scrub	Disabled **Enabled**	Enabled	Disabled	Enabled
Advanced > USB Configuration	Onboard USB Controllers	Disabled **Enabled**	Enabled	Enabled	Disabled [13]

[11] The enabling of this option can be advantageous for applications with a high I/O utilization, but low or no core utilization.

[12] The PRIMERGY RX2510 M1 does not support the snoop mode Cluster on Die (COD). In order to avoid losses in performance this option should not be enabled for the PRIMERGY RX2510 M1.

[13] Switching off this option prevents the use of internal or external USB devices.

BIOS options details

ASPM Support

BIOS Setup Menu	BIOS Option	Settings	Performance	Low Latency	Energy Efficiency
Advanced > PCI Subsystem Settings	ASPM Support	**Disabled** L1 Only	Disabled	Disabled	L1 Only

ASPM stands for "Active State Power Management" and allows putting the PCIe links to the PCIe devices in various power-saving modes so as to reduce power consumption. The setting "L1 Only" can be used by the system - depending on the activity of the PCIe link - to put the link into the most energy-efficient power saving mode. However, changing or exiting the power saving mode increases the latency. Full I/O performance of the PCIe devices is allowed with the setting "Disabled".

DMI Control

BIOS Setup Menu	BIOS Option	Settings	Performance	Low Latency	Energy Efficiency
Advanced > PCI Subsystem Settings	DMI Control	GEN 1 **GEN 2**	GEN 2	GEN 2	GEN 1

DMI stands for "Digital Media Interface" and is the connection between the Intel processors and the chipset. This link can be run with different speeds. Among other things the chipset provides the communication to the onboard LAN controllers (only for servers of the CX model range), USB controllers and onboard SAS/SATA controllers. In order to slightly reduce power consumption the speed of the DMI link from "GEN 2" to "GEN 1" can be reduced for environments, in which these components provided by the chipset are only little used.

Hyper-Threading

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

Generally Fujitsu always recommends you to 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	Settings	Performance	Low Latency	Energy Efficiency
Advanced > CPU Configuration	Active Processor Cores	**0 (All)** [1 – n]	0 (All)	1 – n	0 (All)

It is possible to disable individual cores of a processor in the BIOS (e.g.: 4 cores on a 10-core processor can be disabled). In this case, the L3 cache is retained in full for the remaining cores. Although maximum throughput is only achieved with the maximum number of cores, it is advantageous - especially with latency-sensitive applications that do not utilize all the cores - if you disable the cores that are not needed to allow maximum Turbo Mode frequency on the remaining, active cores. This works because the disabled cores reduce the electrical power consumption of the processor and in so doing allowing higher Turbo Mode frequencies on the remaining cores. This need not work with all the load profiles, power-hungry AVX applications in particular can be an exception here. It is nevertheless possible with this BIOS option to realize a configuration with the highest possible frequency and the highest possible cache usage.

Prefetcher

BIOS Setup Menu	BIOS Option	Settings	Performance	Low Latency	Energy Efficiency
Advanced > CPU Configuration	[Hardware] [Adjacent Cache Line] [DCU Streamer] [DCU Ip] Prefetcher	**Enabled** Disabled	Enabled	Enabled	Disabled

The PRIMERGY server BIOS has several prefetcher options. These include:

- Hardware Prefetcher
- Adjacent Cache Line Prefetch
- DCU Streamer Prefetcher
- DCU Ip Prefetcher

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 the main memory is used to full capacity and the memory connection is a performance bottleneck, are the exception to this. In these cases it can be advantageous to set the prefetcher options to "Disabled" so as to also use the bandwidth that is otherwise used for the prefetching. Furthermore, the power consumption of the server can be slightly reduced by disabling the prefetchers. Before the prefetcher options are changed on productive 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 previous sequential accesses 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.

Intel Virtualization Technology

BIOS Setup Menu	BIOS Option	Settings	Performance	Low Latency	Energy Efficiency
Advanced > 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.

Power Technology

BIOS Setup Menu	BIOS Option	Settings	Performance	Low Latency	Energy Efficiency
Advanced > CPU Configuration	Power Technology	Disabled **Energy Efficient** Custom	Custom	Custom	Custom

The BIOS option "Power Technology" is a superset of different BIOS options, which control the performance and power management functions of the processors. The standard setting "Energy Efficient" regulates a good balance between electrical power consumption and compute power. In order to see and individually set the corresponding relevant options, select the setting "Custom". The standard settings of the individual options for the "Energy Efficient" setting are in bold print in the following sections. These BIOS options are:

- Enhanced SpeedStep
- Turbo Mode
- Override OS Energy Performance
  - Energy Performance
  - Utilization Profile
- CPU C1E Support
- CPU C3/C6 Report
- Package C State limit

The "Disabled" setting deactivates the power management of the processors (P-States → „Enhanced SpeedStep = Disabled“ and C-States are deactivated) and at the same time limits the maximum processor frequency to the nominal frequency by disabling the "Turbo Mode" option.

Enhanced SpeedStep

BIOS Setup Menu	BIOS Option	Settings	Performance	Low Latency	Energy Efficiency
Advanced > 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	Settings	Performance	Low Latency	Energy Efficiency
Advanced > 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 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 carried out (the key factor here is whether they are AVX or so-called non-AVX instructions). 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 general conditions and are not always guaranteed, it can be advantageous for application scenarios, in which constant performance or lower electrical power consumption is required, to disable the "Turbo Mode" option.

Override OS Energy Performance

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

The new generation of Intel Xeon E5-2600 v3 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 this option is enabled, it is also possible to make the settings for the BIOS option "Utilization Profile".

The BIOS option "Override OS Energy Performance" is not available for the PRIMERGY CX servers. This means that it is not possible to override the operating system setting from the BIOS.

Energy Performance

BIOS Setup Menu	BIOS Option	Settings	Performance	Low Latency	Energy Efficiency
Advanced > 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". The settings are only active if the BIOS option "Override OS Energy Performance" is set to "Enabled". In the other case, the operating system takes on the task of setting the "Energy Performance" option via the power plan.

Utilization Profile

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

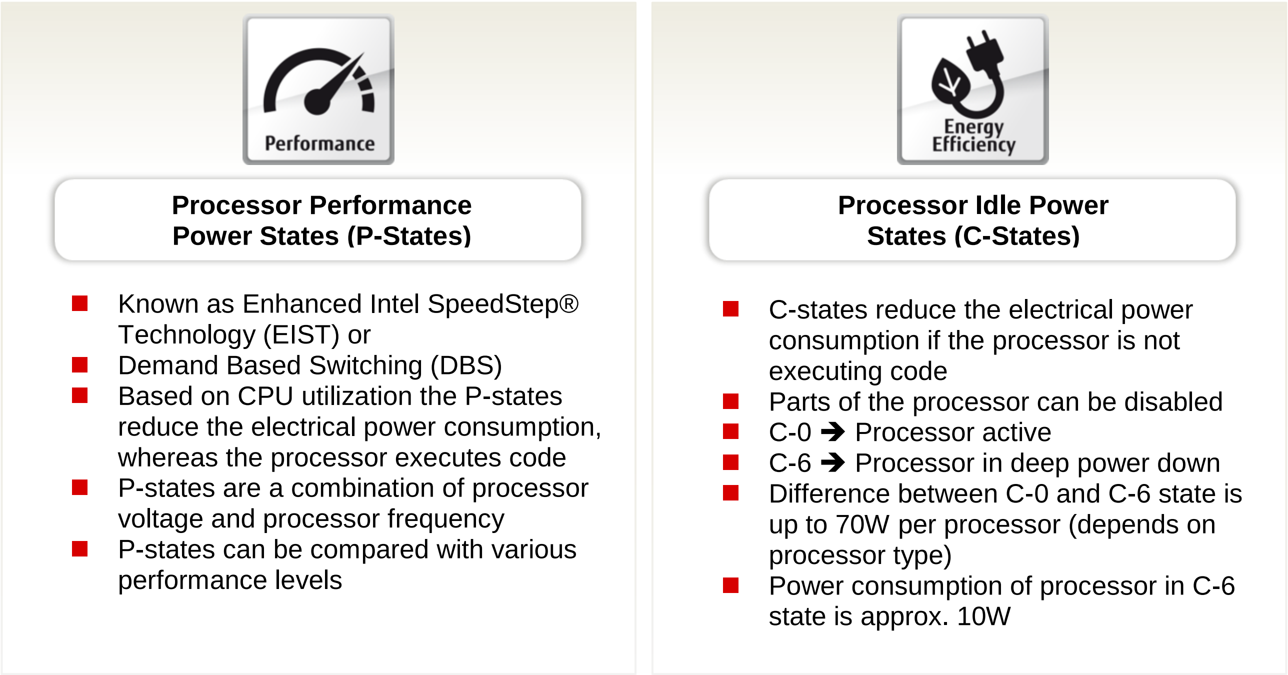
If the BIOS option "Override OS Energy Performance" is enabled, it is also possible to set a so-called "Utilization Profile". The option "Utilization Profile" is used to parameterize an energy-saving option, which monitors both the QPI and the PCIe bandwidth, and attempts to adapt the processor frequency 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 QPI 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.

CPU C1E Support

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

The C1E is a CPU C-state, which is enabled as soon as the operating system informs the CPU that it is idle. The CPU C-states are idle states, in which the core of a processor is put into a type of sleep state if it has no code to run. Consequently, power consumption is substantially reduced in an idle state. In an enabled state the P-states of a processor ensure energy-efficient implementation of the code by only making as much power available as is required.



Processor Performance Power States (P-States)

- 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 States (C-States)

- 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
- Difference between C-0 and C-6 state is up to 70W per processor (depends on processor type)
- Power consumption of processor in C-6 state is approx. 10W

C1E ensures that in an idle state the frequency is always clocked down to the minimum of 1.20 GHz. 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 1.20 GHz. 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". Here you should be aware that electrical power consumption in an idle state increases drastically.

CPU C3/C6 Report

BIOS Setup Menu	BIOS Option	Settings	Performance	Low Latency	Energy Efficiency
Advanced > CPU Configuration	CPU C3 Report	**Disabled** Enabled	Disabled	Disabled	Enabled
Advanced > CPU Configuration	CPU C6 Report	Disabled **Enabled**	Enabled	Disabled	Enabled

These BIOS options are used to inform the operating system whether it can use the CPU C3 or C6 states ("Enabled") or not ("Disabled"). Since the waking-up from these C-states increases latency, it is advisable to put the setting to "Disabled" for the CPU C-states for applications where maximum performance with the lowest possible response time matters. The following applies in this case - the higher the C-state, the longer the waking-up time. It should be borne in mind that if all the CPU C-states are 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 at least set "CPU C3 Report" to "Enabled". Using the "Disabled" setting for the BIOS option "CPU C3/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. 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 2015, all enterprise Linux distributions supported by Fujitsu). To force the operating system to respect the BIOS settings, disable this driver by using the Linux kernel parameter "intel_idle.max_cstate=0". The Linux kernel will then instead use the "processor" idle driver that respects the BIOS settings.

Package C State limit

BIOS Setup Menu	BIOS Option	Settings	Performance	Low Latency	Energy Efficiency
Advanced > CPU Configuration	Package C State limit	C0 C2 C6 **C6(Retention)**	C0	C0	C6(Retention)

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 left at "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.

QPI Link Frequency Select

BIOS Setup Menu	BIOS Option	Settings	Performance	Low Latency	Energy Efficiency
Advanced > CPU Configuration	QPI Link Frequency Select	6.4 GT/s 8.0 GT/s 9.6 GT/s **Auto**	Auto	Auto	6.4 GT/s

Using this BIOS option makes it possible to reduce the QuickPath interconnect (QPI) speed 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 "6.4 GT/s", which brings the greatest savings, "8.0 GT/s" and "9.6 GT/s", which is the maximum speed.

Uncore Frequency Override

BIOS Setup Menu	BIOS Option	Settings	Performance	Low Latency	Energy Efficiency
Advanced > CPU Configuration	Uncore Frequency Override	**Disabled** Enabled	Disabled	Enabled	Disabled

The new generation of Intel Xeon E5-2600 v3 based processors works 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 enabling of this BIOS option 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, form the exceptions. In this situation, the processor's power management mechanisms attempt to reduce the frequency to a minimum (see "CPU C1E Support“). If this happens, the frequency of the so-called uncore area is also automatically lowered. As the entire I/O communication (PCIe, memory, QPI, etc.) is via the uncore area, this would have a negative effect on the I/O throughput. The "Uncore Frequency Override = Enabled" setting would prevent this, but the resulting increase in electrical power consumption must be accepted.

QPI Snoop Modes

BIOS Setup Menu	BIOS Option	Settings	Performance	Low Latency	Energy Efficiency
Advanced > CPU Configuration	COD Enable	Disabled Enabled **Auto**	Auto	Auto	Auto
Advanced > CPU Configuration	Early Snoop	Disabled Enabled **Auto**	Auto	Auto	Auto

The Intel Xeon E5-2600 v3 processor generation supports three different snoop modes, which regulate how cache coherency is implemented via the Intel QPI links. The BIOS of each of these PRIMERGY servers has two BIOS options to offer, which can be used to help set the three different snoop modes. Before the required settings are made, you need to ensure that the required snoop mode is supported by the used application and the used operating system. This white paper consciously does not make a specific snoop mode recommendation for the three application scenarios described, because the optimal setting always depends on the individual application. For example, there can be two different applications that originate from the same segment, e.g. "Low Latency", but have different requirements for the snoop mode. This means that the different snoop modes must be tested with the individual application in order to decide which setting is optimal from a performance viewpoint. Help in this respect is provided by the table "Relative Snoop Mode Performance" later in this section.

The following table shows how the two available BIOS options "COD Enable" and "Early Snoop" must be configured in order to set the snoop mode required in each case.

BIOS option Snoop Mode	COD Enable	Early Snoop
Early Snoop (ES)	Disabled	Enabled
Cluster on Die (COD)	Enabled	Disabled
Home Snoop (HS)	Disabled	Disabled

The available snoop modes depend on the CPUs used and the further hardware configuration. Thus, the standard setting for the two options in the BIOS is in each case "Auto". Depending on the hardware configuration used, this ensures that the best possible snoop mode is always set. In the case of a 2-socket configuration the standard setting "Auto" for both BIOS options results in the “Early Snoop” mode.

Each snoop mode has different features and effects on the memory bandwidth and memory latency, depending on how the snoop traffic is created.

Details of the individual QPI snoop modes:

Cluster on Die (COD)	The "Cluster on Die" (COD) mode logically divides every processor into two equal clusters, each with half of all the cores available on the processor and half of the available L3 caches. Thus, an individual physical processor in "Cluster on Die" mode is for the operating system like two separate NUMA nodes with - depending on the allocation - the appropriate number of cores and size of the L3 cache. Based on an "on-die directory cache" and on "in-memory directory bits" an assessment is made in this mode as to whether a snoop must be sent or not. The "Cluster on Die" (COD) mode is specially recommended for NUMA-optimized applications in order to achieve the lowest possible local memory latency and the highest possible local memory bandwidth. This mode is only supported with Intel Xeon E5-2600 v3 processors with 10 or more cores and only if the BIOS option "NUMA = Enabled" is configured (standard setting).
Home Snoop (HS)	Snoops are always sent in home snoop mode. The origin of the snoop is the difference to the early snoop mode. In home snoop mode the snoop is sent by the home agent. This mode is recommended for NUMA applications, which react sensitively to both local and remote memory bandwidths.
Early Snoop (ES)	As with home snoop mode, snoops are always sent in early snoop mode. In early snoop mode the snoops are sent by the caching agent. This mode is recommended for applications that require an as low as possible memory latency or small cache-to-cache transfer latency of the remote socket. The memory latency is reduced as a result of the fact that the snoops in this mode are sent earlier.

Exception - PRIMERGY RX2510 M1:

The PRIMERGY RX2510 M1 is only configured with half the possible memory channels and for this reason does not support the snoop mode Cluster on Die (COD). Enabling the BIOS option "COD Enable" for the PRIMERGY RX2510 M1 results in losses in performance and should therefore be avoided.

Relative Snoop Mode Performance Intel Xeon E5-2600 v3 processors with 10 or more cores; BIOS option „NUMA = **Enabled**“			
Performance Metric*	**Early Snoop (ES)**	**Cluster on Die (COD)**	**Home Snoop (HS)**
L3 cache Hit Latency	Low	**Lowest**	Low
Local Memory Latency	Medium	**Low**	High
Remote Memory Latency	**Lowest**	Low - High	Low
Local Memory Bandwidth	High	**Highest**	High
Remote Memory Bandwidth	Medium	Medium	**High**
Intel Xeon E5-2600 v3 processors with 10 or more cores; BIOS option „NUMA = **Disabled**“			
Memory Latency	Low	Not supported	Low
Memory Bandwidth	High		High
Intel Xeon E5-2600 v3 processors with less than 10 cores; BIOS option „NUMA = **Enabled**“			
L3 cache Hit Latency	Low	Not supported	Low
Local Memory Latency	**Lowest**		Low
Remote Memory Latency	**Lowest**		Low
Local Memory Bandwidth	High		High
Remote Memory Bandwidth	Low		**High**
Intel Xeon E5-2600 v3 processors with less than 10 cores; BIOS option „NUMA = **Disabled**“			
Memory Latency	**Lowest**	Not supported	Low
Memory Bandwidth	High		High

For latency, lower is better. For bandwidth, higher is better.

DDR Performance

BIOS Setup Menu	BIOS Option	Settings	Performance	Low Latency	Energy Efficiency
Advanced > Memory Configuration	DDR Performance	**Performance optimized** Energy optimized	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 (information about this is provided in the white paper Memory performance of Xeon E5-2600 v3 (Haswell-EP) based systems), and as a result provides the highest possible memory performance. The "Energy optimized" setting limits the memory frequency - at all times and regardless of the processor model and memory configuration - to the minimum value (1066 MHz) with the lowest electrical power consumption.

In addition to the BIOS options for memory performance, the memory type used and the optimal configuration of the DIMMs play a far greater role. A detailed description about this and the topic NUMA (Non-Uniform Memory Access) can be found in the white paper Memory performance of Xeon E5-2600 v3 (Haswell-EP)-based systems (see Literature at the end of the document).

Patrol Scrub

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

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 in order to detect and correct memory errors in a preventive way. The time of this memory test cannot be influenced and can under certain circumstances result in losses in performance. 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.

Onboard USB Controllers

BIOS Setup Menu	BIOS Option	Settings	Performance	Low Latency	Energy Efficiency
Advanced > USB Configuration	Onboard USB Controllers	Disabled **Enabled**	Enabled	Enabled	Disabled

The chipset of the PRIMERGY servers has several USB controllers. If you can completely do without the use of USB devices (this also includes mouse and keyboard), the setting for this BIOS option should be "Disabled". This saves power and increases the security against unauthorized third-party access. Regardless of the setting, the USB controllers remain active during system start (disabling only takes place after the POST) so that you also have the option with the "Disabled" setting of accessing the BIOS setup via the USB keyboard in order to change the setting again.

Literature

PRIMERGY Servers

http://primergy.com/

Performance of Server Components

http://www.fujitsu.com/fts/products/computing/servers/mission-critical/benchmarks/x86-components.html

This White Paper:
🇺🇸 http://docs.ts.fujitsu.com/dl.aspx?id=f154aca6-d799-487c-8411-e5b4e558c88b
🇩🇪 http://docs.ts.fujitsu.com/dl.aspx?id=b0877217-e9ef-4c7b-943d-299c0a4c4637
🇯🇵 http://docs.ts.fujitsu.com/dl.aspx?id=2009eb5b-f273-4f1f-94ef-07f1d0304255

BIOS settings for performance, low-latency and energy efficiency (for Xeon E5-2400/2600/4600 based systems)
http://docs.ts.fujitsu.com/dl.aspx?id=e5f29616-130e-47c7-8fa0-be230670edab

BIOS optimizations for Xeon E5-2600 v2 based systems
http://docs.ts.fujitsu.com/dl.aspx?id=84dc1adf-adb8-419f-8d08-b226eb077e46

Memory performance of Xeon E5-2600 v3 (Haswell-EP)-based systems
http://docs.ts.fujitsu.com/dl.aspx?id=74eb62e6-4487-4d93-be34-5c05c3b528a6

PRIMERGY Manuals

http://support.ts.fujitsu.com/Manuals/Index.asp

PRIMERGY BIOS downloads

http://support.ts.fujitsu.com/Download/Index.asp

SPECpower_ssj2008

http://www.spec.org/power_ssj2008

Benchmark Overview SPECpower_ssj2008
http://docs.ts.fujitsu.com/dl.aspx?id=166f8497-4bf0-4190-91a1-884b90850ee0

SAP Server Power

http://global.sap.com/campaigns/benchmark/appbm_benchmarks.epx

VMmark V2

http://www.vmmark.com

Benchmark Overview VMmark V2
http://docs.ts.fujitsu.com/dl.aspx?id=2b61a08f-52f4-4067-bbbf-dc0b58bee1bd

Operating System Performance Tuning Guidelines

Microsoft Windows
http://msdn.microsoft.com/en-us/library/windows/hardware/dn529134

RedHat Linux
https://access.redhat.com/articles/221153

https://access.redhat.com/documentation/en-US/Red_Hat_Enterprise_Linux/7/html/Performance_Tuning_Guide/

SUSE Linux
https://www.suse.com/documentation/sles11/pdfdoc/book_sle_tuning/book_sle_tuning.pdf

VMware vSphere
http://www.vmware.com/files/pdf/techpaper/VMW-Tuning-Latency-Sensitive-Workloads.pdf

Contact

FUJITSU

Website: http://www.fujitsu.com/

PRIMERGY Product Marketing

mailto:Primergy-PM@ts.fujitsu.com

PRIMERGY Performance and Benchmarks

mailto:primergy.benchmark@ts.fujitsu.com

2015-11-26 WW EN