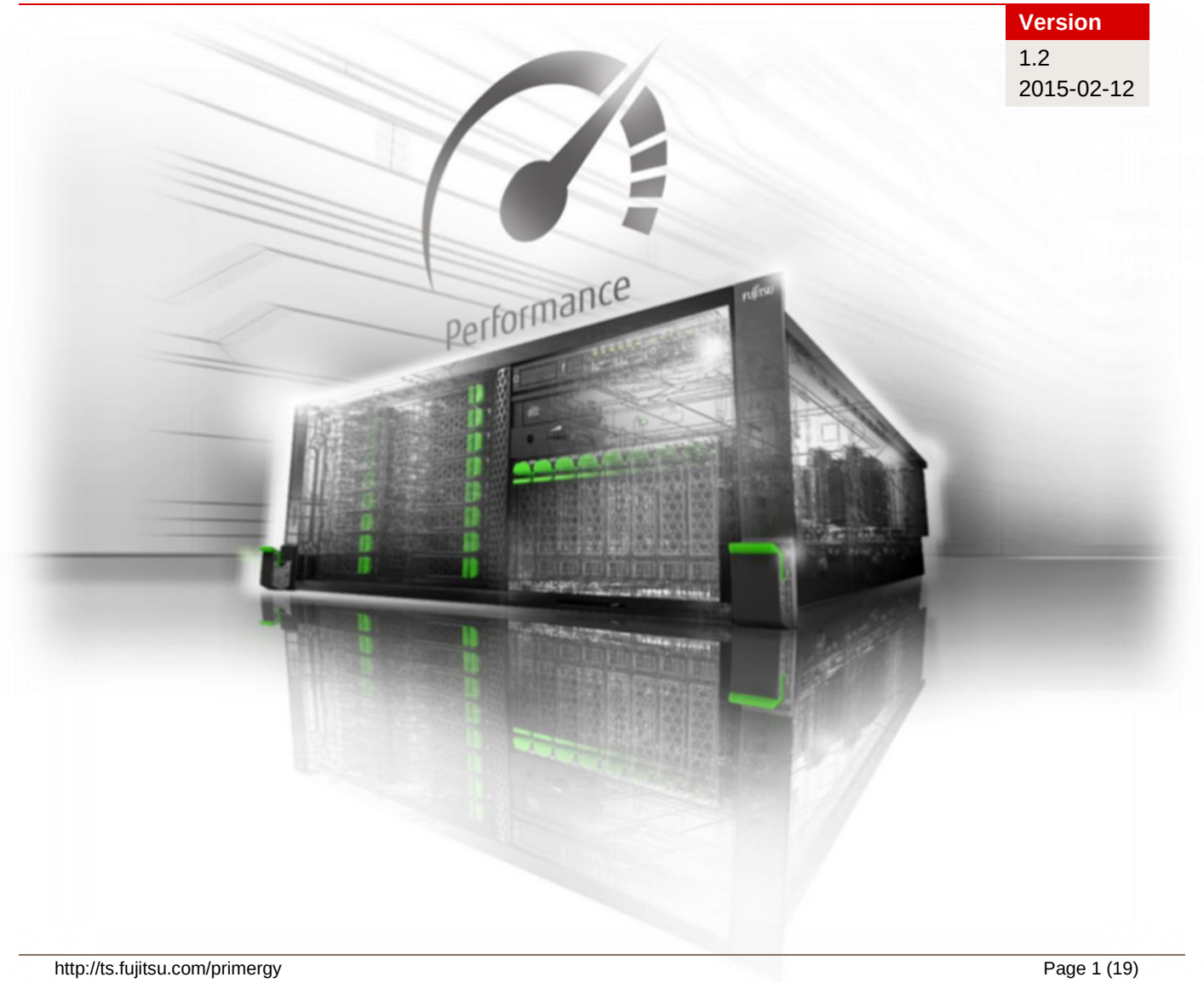
White Paper
FUJITSU Server PRIMERGY
BIOS optimizations for Xeon E5-2600 v2 based systems

This document explains the BIOS settings that are valid for the Intel Xeon E5-2600 v2 based PRIMERGY server generation (PRIMERGY BX924 S4, CX250 S2, CX270 S2, RX200 S8, RX300 S8, RX350 S8, TX300 S8).

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
2015-02-12



Contents

Document history

Version 1.0

First edition

Version 1.0a

Minor corrections

Version 1.0b

Minor corrections

Version 1.1

Hints regarding Monitor/MWAIT added
Note regarding “Performance” and “Low Latency” added
Description for BIOS option “Frequency Floor Override” extended

Version 1.1a

Minor correction

Version 1.2

Description for BIOS options “CPU C3/C6 Report” and “Monitor/MWAIT” extended

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 proposals and makes 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 v2 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 increasing 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 (Hyperthreading) or even cores in the BIOS in order in this way to reduce the minimal fluctuations in 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, wherever 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 the energy-saving mode, the lower the performance. Furthermore, each one of these energy-saving modes 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. For this reason it is recommended to configure the energy-saving functions, as explained in this document, according to requirements on systems, which are optimized for low latency. Users, who optimize their servers for low latency, should be aware that every optimization in this direction is at the expense of energy efficiency and such system optimization always entails significantly 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 to set the BIOS option "Frequency Floor 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 for example is one option for data center operators, who only have a limited budget of electrical power available, and who are consequently compelled to reduce the power consumption per rack/per server, etc. Performance only plays a subordinate role there, because optimization in this direction consists mainly of reducing the speed and with it 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. Data center operators, for whom maximum server performance is only of secondary importance, strive to achieve optimization of this kind in particular. In such scenarios it is more important to configure the servers as efficiently as possible so as to be able to operate the data centers in an optimal way as far as the overall operating costs are concerned. Numerous publications of the Standard Performance Evaluation Corporation (SPEC) with the first industry-standard benchmark for measuring energy efficiency in servers, the SPECpower_ssj2008, prove that PRIMERGY servers are the best choice when it comes to energy-efficient servers.

PRIMERGY BIOS options

This white paper only contains information about BIOS options that are valid for the Intel Xeon E5-2600 v2 based PRIMERGY servers. And these are:

- PRIMERGY BX924 S4
- PRIMERGY CX250 S2
- PRIMERGY CX270 S2
- PRIMERGY RX200 S8
- PRIMERGY RX300 S8
- PRIMERGY RX350 S8
- PRIMERGY TX300 S8

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 functionalities listed here available. The current BIOS version of the PRIMERGY servers can be downloaded in the Internet at http://www.fujitsu.com/fts/support.

Recommendations for optimization

The following tables list recommendations for BIOS options, which optimize the server 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 becoming apparent when further options are also changed at the same time. Before changes are made to the BIOS options listed 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.

Not all the BIOS options listed here are available on all servers. This is why there are two tables here. Table 1 contains BIOS options for the PRIMERGY BX, RX and TX servers and Table 2 has the BIOS options for the PRIMERGY CX servers.

Table 1: PRIMERGY BX, RX, TX-servers

BIOS Setup Menu	BIOS Option	Settings [1]	Performance	Low Latency	Energy Efficiency
Advanced > PCI Subsystem Settings	ASPM Support	**Disabled** Auto Limit to L0s	Disabled	Disabled	Auto
Advanced > PCI Subsystem Settings	DMI Control	**GEN 2** GEN 1	GEN 2	GEN 2	GEN 1 [2]
Advanced > CPU Configuration	Hyper-threading	Disabled **Enabled**	Enabled	Disabled [3]	Enabled
Advanced > CPU Configuration	Active Processor Cores	**All** [1 – n]	All	1 – n [4]	All
Advanced > CPU Configuration	[Hardware] [Adjacent Cache Line] [DCU Streamer] [DCU IP] Prefetcher	Disabled **Enabled**	Enabled	Enabled	Disabled
Advanced > CPU Configuration	Intel Virtualization Technology	Disabled **Enabled**	Disabled [5]	Disabled	Disabled
Advanced > CPU Configuration	Power Technology	Disabled **Energy Efficient** Custom	Custom	Custom	Custom
Advanced > CPU Configuration	Turbo Mode [6]	Disabled **Enabled**	Enabled	Enabled [7]	Enabled
Advanced > CPU Configuration	Override OS Energy Performance [6]	**Disabled** Enabled	Enabled	Enabled	Disabled [8]
Advanced > CPU Configuration	Energy Performance [9]	Performance **Balanced Performance** Balanced Energy Energy Efficient	Performance	Performance	Energy Efficient
Advanced > CPU Configuration	Utilization Profile [9]	**Even** Unbalanced	Unbalanced	Unbalanced	Even
Advanced > CPU Configuration	CPU C1E Support [6]	Disabled **Enabled**	Enabled	Disabled	Enabled
Advanced > CPU Configuration	CPU C3 Report [6]	**Disabled** Enabled	Disabled	Disabled	Enabled
Advanced > CPU Configuration	CPU C6 Report [6]	Disabled **Enabled**	Disabled	Disabled	Enabled
Advanced > CPU Configuration	Monitor/MWAIT [6] [10]	Disabled **Enabled**	Enabled	Disabled	Enabled

1) The setting in bold print is the standard value.
2) GEN 1 is recommended for low chipset I/O load; otherwise the setting should be GEN 2.
3) Unless all threads of the CPU are used, switching off Hyper-Threading can improve the 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) If virtualization is not used, this option should be set to "Disabled".
6) This option is only visible if the setting for "Power Technology" is changed to "Custom".
7) Maximum Turbo mode performance is not guaranteed under all operating conditions, which can result in fluctuations in performance. Turbo mode can be set to "Disabled" for a stable and constant response time.
8) 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.
9) This option is only visible 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	Package C-State limit [6]	C0 C2 C6 C7 **No Limit**	C0	C0	No Limit
Advanced > CPU Configuration	QPI Link1 Disable [6]	**Disabled** Enabled	Disabled	Disabled	Enabled
Advanced > CPU Configuration	QPI Link Frequency Select	**Auto** 6.4 GT/s 7.2 GT/s 8.0 GT/s	Auto	Auto	6.4 GT/s
Advanced > CPU Configuration	Frequency Floor Override	**Disabled** Enabled	Disabled	Enabled [11]	Disabled
Advanced > Memory Configuration	DDR Performance	Low-Voltage optimized Energy optimized **Performance optimized**	Performance optimized	Performance optimized	Low-Voltage optimized
Advanced > Memory Configuration	DRAM Maintenance	Disabled **Auto**	Disabled	Disabled	Auto
Advanced > Memory Configuration	Patrol Scrub [12]	Disabled **Enabled**	Enabled	Disabled	Enabled
Advanced > Memory Configuration	Refresh Rate Multiplier [12]	1x **2x** 3x 4x	1x	1x	1x
Advanced > USB Configuration	Onboard USB Controllers	Disabled **Enabled**	Enabled	Enabled	Disabled [13]

Table 2: PRIMERGY CX servers

BIOS Setup Menu	BIOS Option	Settings [1]	Performance	Low Latency	Energy Efficiency
Advanced > PCI Subsystem Settings > PCI Express Settings	ASPM Support	**Disabled** Auto	Disabled	Disabled	Auto
Advanced > CPU Configuration	Hyper-threading	Disabled **Enabled**	Enabled	Disabled [3]	Enabled
Advanced > CPU Configuration	Active Processor Cores	**All** [1 – n]	All	1 – n [4]	All
Advanced > CPU Configuration	[Hardware] [Adjacent Cache Line] [DCU Streamer] [DCU IP] Prefetcher	Disabled **Enabled**	Enabled	Enabled	Disabled
Advanced	Intel Virtualization	Disabled	Disabled [5]	Disabled	Disabled

[10] When MONITOR/MWAIT is set to Disabled, the system may reboot under the following circumstances.
- Boot of Maintenance Mode from SVS DVD
- Boot of Rescue Mode from RHEL installation DVD
MONITOR/MWAIT should be set to Enabled at these times.

[11] Switching on this for applications that do not use all CPUs of the system at full capacity can be advantageous. If the option is set to "Enabled", Turbo mode should be set to "Disabled" to ensure that the CPU is not operated out of specification.

[12] This option is only visible if the setting for "DRAM Maintenance" is changed to "Disabled".

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

BIOS Setup Menu	BIOS Option	Settings [1]	Performance	Low Latency	Energy Efficiency
> CPU Configuration	Technology	**Enabled**			
Advanced > CPU Configuration > CPU Power Management Configuration	Power Technology	Disabled **Energy Efficient** Custom	Custom	Custom	Custom
Advanced > CPU Configuration > CPU Power Management Configuration	Turbo Mode [6]	Disabled **Enabled**	Enabled	Enabled [7]	Enabled
Advanced > CPU Configuration > CPU Power Management Configuration	Energy Performance [6]	Performance **Balanced Performance** Balanced Energy Energy Efficient	Performance	Performance	Energy Efficient
Advanced > CPU Configuration > CPU Power Management Configuration	CPU C3 Report [6]	**Disabled** Enabled	Disabled	Disabled	Enabled
Advanced > CPU Configuration > CPU Power Management Configuration	CPU C6 Report [6]	Disabled **Enabled**	Disabled	Disabled	Enabled
Advanced > CPU Configuration > CPU Power Management Configuration	Package C-State limit [6]	C0 C2 C6 C7 **No Limit**	C0	C0	No Limit
Chipset > Northbridge > QPI Configuration	QPI Link Frequency Select	**Auto** 6.4 GT/s 7.2 GT/s 8.0 GT/s	Auto	Auto	6.4 GT/s
Chipset > Northbridge	DDR Voltage	Auto **Force 1.5V**	Force 1.5V	Force 1.5V	Auto
Chipset > Northbridge	DDR Speed	**Auto** Force DDR3 800 Force DDR3 1066 Force DDR3 1333 Force DDR3 1600 Force DDR3 1866	Force DDR3 1866	Force DDR3 1866	Auto
Chipset > Northbridge	Patrol Scrub	Disabled **Enabled**	Enabled	Disabled	Enabled
Chipset > Northbridge	Refresh Rate Multiplier	1x **2x** 3x 4x	1x	1x	1x
Chipset > Southbridge > USB Configuration	EHCI Controller 1	Disabled **Enabled**	Enabled	Enabled	Disabled [13]

BIOS options details

ASPM Support

BIOS Setup Menu	BIOS Option	Settings	Performance	Low Latency	Energy Efficiency
Advanced > PCI Subsystem Settings [> PCI Express Settings] [1)]	ASPM Support	**Disabled** Auto Limit to L0s [2)]	Disabled	Disabled	Auto

1) Different menu structure for PRIMERGY CX servers
2) Not available for PRIMERGY CX servers

ASPM stands for "Active State Power Management" and enables the PCIe links to the PCIe devices to be sent in various power-saving states so as to reduce power consumption. Depending on the activity of the PCIe link, the system uses the "Auto" setting to select the most energy-efficient power-saving state. However, changing the power-saving states and waking-up from the various power-saving states increases the latency. Thus, the "Disabled" setting should be selected for full I/O performance of the PCIe devices.

DMI Control

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

1) Not available for PRIMERGY CX servers

DMI stands for "Digital Media Interface" and is the connection between the Intel processor and the chipset. This link can be run with different speeds. Among other things the chipset provides the communication to the onboard LAN controllers, USB controllers, onboard SAS/SATA controllers and, if applicable, also to PCIe slots, etc. In order to reduce power consumption the speed of the DMI link from "GEN 2" to "GEN 1" can be reduced for environments, in which the communication provided by the chipset is 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, as is the case 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. In cases in which the additional threads are not used and the response time is to be minimized, Hyper-Threading should be disabled in the BIOS.

Active Processor Cores

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

It is possible to disable individual cores 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 so as retain 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 offer the remaining cores greater scope as regards Turbo mode frequency. 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 a large cache.

Prefetcher

BIOS Setup Menu	BIOS Option	Settings	Performance	Low Latency	Energy Efficiency
Advanced > CPU Configuration	[Hardware] [Adjacent Cache Line] [DCU Streamer] [DCU IP] Prefetcher	Disabled **Enabled**	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 such cases it can be advantageous to set the prefetcher option 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 currently 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, in order to prefetch the next line from the L2 cache or the main memory into the L1 cache based on the assumption that the next cache line will also be needed.
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 [> CPU Power Management Configuration] [1)]	Power Technology	Disabled **Energy Efficient** Custom	Custom	Custom	Custom

[1)] Different menu structure for PRIMERGY CX servers

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". These BIOS options are:

- Turbo Mode
- Override OS Energy Performance [14)]
- Energy Performance [15)]
- CPU C1E Support [14)]
- CPU C3/C6 Report
- Monitor/MWAIT [14)]
- Package C-State limit
- QPI Link1 Disable [14)]

The "Disabled" setting deactivates the power management of the processors and at the same time limits the maximum processor frequency to the nominal frequency by disabling the "Turbo Mode" option.

Turbo Mode

BIOS Setup Menu	BIOS Option	Settings	Performance	Low Latency	Energy Efficiency
Advanced > CPU Configuration [> CPU Power Management Configuration] [1)]	Turbo Mode	Disabled **Enabled**	Enabled	Enabled	Enabled

[1)] Different menu structure for PRIMERGY CX servers

This BIOS option enables and disables the Intel Turbo Boost Technology function of the processor. The Turbo Boost Technology function allows the processor to operate with higher frequencies than the nominal frequency. Depending on the processor type, the maximum achievable frequency is different and also depends on the number of active cores, current consumption, electrical power consumption and the temperature of the processor. In addition to these general conditions, the quality of the processors also plays a role for the Turbo Mode performance, particularly with HPC applications.

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 [1)]	**Disabled** Enabled	Enabled	Enabled	Disabled

[1)] Not available for PRIMERGY CX servers

The new generation of Intel Xeon E5-2600 v2 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

[14)] Not available for PRIMERGY CX servers
[15)] Is always shown for PRIMERGY CX servers, regardless of the setting for the "Power Technology" option

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 [> CPU Power Management Configuration] [1)]	Energy Performance	Performance **Balanced Performance** Balanced Energy Energy Efficient	Performance	Performance	Energy Efficient

1) Different menu structure for PRIMERGY CX servers

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". In the case of the PRIMERGY BX, RX and TX servers the settings are only enabled 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. The settings are always enabled for the PRIMERGY CX servers, but are in some cases overridden by the operating system setting.

Utilization Profile

BIOS Setup Menu	BIOS Option	Settings	Performance	Low Latency	Energy Efficiency
Advanced > CPU Configuration	Utilization Profile [1)]	**Even** Unbalanced	Unbalanced	Unbalanced	Even

1) Not available for PRIMERGY CX servers

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 [1)]	Disabled **Enabled**	Enabled	Disabled	Enabled

[1)] Not available for PRIMERGY CX servers

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 50W 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 and the lifespan of the processors suffers. Operations over a longer period with the setting "Disabled" for the BIOS option "CPU C1E Support" can also result in warranty exclusion for the processors.

CPU C3/C6 Report

BIOS Setup Menu	BIOS Option	Settings	Performance	Low Latency	Energy Efficiency
Advanced > CPU Configuration [> CPU Power Management Configuration] [1)]	CPU C3 Report	**Disabled** Enabled	Disabled	Disabled	Enabled
Advanced > CPU Configuration [> CPU Power Management Configuration] [1)]	CPU C6 Report	Disabled **Enabled**	Disabled	Disabled	Enabled

1) Different menu structure for PRIMERGY CX servers

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.

Monitor/MWAIT

BIOS Setup Menu	BIOS Option	Settings	Performance	Low Latency	Energy Efficiency
Advanced > CPU Configuration	Monitor/MWAIT [1)]	Disabled **Enabled**	Enabled	Disabled	Enabled

1) Not available for PRIMERGY CX servers
When MONITOR/MWAIT is set to Disabled, the system may reboot under the following circumstances.
- Boot of Maintenance Mode from SVS DVD
- Boot of Rescue Mode from RHEL installation DVD
MONITOR/MWAIT should be set to Enabled at these times.

This BIOS option can be used to let operating systems not utilize its C-State capabilities. However, this setting can only be changed if the setting for the BIOS options "CPU C1E Support" and "CPU C3/C6 Report" was previously set to "Disabled". Since it modifies the hardware capabilities seen by the operating system, this option also works on operating systems that disregard the BIOS C-state settings (see previous paragraph). It is not recommended to change the Monitor/MWait setup option when using Linux operating systems[2)].

2) It is preferable to use the Linux kernel parameter "idle=X", as follows: "idle=nomwait" will have the same effect as the BIOS option “Monitor/MWAIT” set to “disabled”, and "idle=poll" will even use a busy loop in idle state, minimizing wake-up latency at the cost of much higher energy consumption. Both Linux kernel parameter options will also disable the Linux "intel_idle" driver (see paragraph "CPU C3/C6 Report").

Package C-State limit

BIOS Setup Menu	BIOS Option	Settings	Performance	Low Latency	Energy Efficiency
Advanced > CPU Configuration [> CPU Power Management Configuration] [1)]	Package C-State limit	C0 C2 C6 C7 **No Limit**	C0	C0	No Limit

[1)] Different menu structure for PRIMERGY CX servers

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.

QPI Link1 Disable

BIOS Setup Menu	BIOS Option	Settings	Performance	Low Latency	Energy Efficiency
Advanced > CPU Configuration	QPI Link1 Disable [1)]	**Disabled** Enabled	Disabled	Disabled	Enabled

[1)] Not available for PRIMERGY CX servers

The processors of the Intel Xeon E5-2600 v2 family each have 2 QuickPath Interconnects (QPI) for coupling purposes. These QPI links are mainly used for communication between the two CPUs (cache coherency), remote memory access and remote I/O access. In configurations with low I/O utilization and evenly configured memory configurations, which ensure that remote memory accesses are kept to a minimum, the bandwidth of only one QPI link may already be sufficient. In such cases, the disabling of one of the two links by setting the BIOS option "QPI Link1 Disable" to "Enabled" helps reduce the electrical power consumption of the server, depending on the processor type used.

QPI Link Frequency Select

BIOS Setup Menu	BIOS Option	Settings	Performance	Low Latency	Energy Efficiency
Advanced > CPU Configuration [Chipset > Northbridge > QPI Configuration] [1)]	QPI Link Frequency Select	**Auto** 6.4 GT/s 7.2 GT/s 8.0 GT/s	Auto	Auto	6.4 GT/s

[1)] Different menu structure for PRIMERGY CX servers

Using this BIOS option makes it possible to reduce the interconnect speed (QPI) 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, "7.2 GT/s" and "8.0 GT/s", which is the maximum speed.

Frequency Floor Override

BIOS Setup Menu	BIOS Option	Settings	Performance	Low Latency	Energy Efficiency
Advanced > CPU Configuration	Frequency Floor Override [1)]	**Disabled** Enabled	Disabled	Enabled	Disabled

[1)] Not available for PRIMERGY CX servers

The enabling of this BIOS option ensures that the processor always works with its maximum nominal frequency, i.e. when it has little to do. Accordingly, power consumption is also higher and therefore the setting for this option should normally always be "Disabled". Applications, whose threads do not put all the CPUs of the system under full load, are an exception. In this case, accesses to the remote CPU, e.g. for cache coherency, and in particular accesses to the remote memory of the other CPU, or to PCIe devices, which are connected to the other CPU, are considerably slower. In order to keep latency as low as possible

in this case the BIOS option "Frequency Floor Override" can be set to "Enabled" if the consequently increased electrical power consumption is accepted. Depending on the processor type used, this can in an idle state mean electrical power consumption that is up to about 130 watts higher. One further disadvantage here is that when enabling this BIOS option it is mandatory to disable the Turbo mode to ensure that the processors are operated within specification. If you do not want to be without the performance advantage due to the Turbo mode, the BIOS option "Utilization profiles = Unbalanced" offers a possible alternative to "Frequency Floor Override = Enabled".

Another exception is I/O-intensive applications which do not place any load on the processors or at most a very low load. 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 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. This is avoided via the setting "Frequency Floor Override = Enabled".

DDR Performance

BIOS Setup Menu	BIOS Option	Settings	Performance	Low Latency	Energy Efficiency
Advanced > Memory Configuration	DDR Performance [1)]	Low-Voltage optimized Energy optimized **Performance optimized**	Performance optimized	Performance optimized	Low-Voltage optimized

[1)] Not available for PRIMERGY CX servers

This BIOS option for the PRIMERGY BX, RX and TX servers controls the speed and the voltage with which the memory modules are operated. Performance and energy consumption are weighed up against each other in this case. The "Performance optimized" setting operates the DIMMs with a voltage of 1.5 V, if the memory frequency can as a result be increased, and thus enables maximum speed. The "Low-Voltage optimized" setting is used to operate the DIMMs, if possible, with an energy-saving 1.35 V. This operation is only possible for certain memory configurations and can limit the memory frequency (information about this is provided in the Memory Performance white paper). The "Energy optimized" setting additionally limits the memory frequency to the minimum value (800 MHz) with the lowest electrical power consumption. The "Low-Voltage optimized" setting is recommended for optimal energy efficiency and the "Performance optimized" setting for maximum memory performance. This BIOS option is not available in this form for the PRIMERGY CX servers. In their case, the voltage and memory frequency are set using the BIOS options "DDR Voltage" and "DDR Speed".

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 Memory Performance white paper (see Literature at the end of the document).

DDR Voltage

BIOS Setup Menu	BIOS Option	Settings	Performance	Low Latency	Energy Efficiency
Chipset > Northbridge	DDR Voltage [1)]	Auto **Force 1.5V**	Force 1.5V	Force 1.5V	Auto

[1)] Not available for PRIMERGY BX, RX, TX servers

This BIOS option controls the voltage with which the memory modules are operated for the PRIMERGY CX servers. The standard setting "Force 1.5V" operates the DIMMs with a voltage of 1.5 V and thus always enables maximum speed. The "Auto" setting automatically sets the voltage depending on the memory modules used and the setting for the "DDR Speed" option. To enable the so-called "low-voltage" memory modules to be run with energy-saving 1.35 V, both the settings for the BIOS option "DDR Voltage" and for "DDR Speed" should be set to "Auto". This BIOS option is not available in this form for the PRIMERGY BX, RX and TX servers. In their cases, the voltage and memory frequency are set using the BIOS option "DDR Performance".

DDR Speed

BIOS Setup Menu	BIOS Option	Settings	Performance	Low Latency	Energy Efficiency
Chipset > Northbridge	DDR Speed [1)]	**Auto** Force DDR3 800 Force DDR3 1066 Force DDR3 1333 Force DDR3 1600 Force DDR3 1866	Force DDR3 1866	Force DDR3 1866	Auto

1) Not available for PRIMERGY BX, RX, TX servers

This BIOS option controls the speed with which the memory modules are operated for the PRIMERGY CX servers. The standard setting "Auto" operates the DIMMs with the maximum possible speed, according to the Intel memory reference code, for the voltage that is in each case set using the "DDR Voltage" option. The further settings are used to set the speed manually. Information about possible configurations and the resulting voltages and frequencies is provided in the Memory Performance white paper. This BIOS option is not available in this form for the PRIMERGY BX, RX and TX servers. In their cases, the voltage and memory frequency are set using the BIOS option "DDR Performance".

DRAM Maintenance

BIOS Setup Menu	BIOS Option	Settings	Performance	Low Latency	Energy Efficiency
Advanced > Memory Configuration	DRAM Maintenance [1)]	Disabled **Auto**	Disabled	Disabled	Auto

1) Not available for PRIMERGY CX servers

The standard setting of the "DRAM Maintenance" option is "Auto" and it should not be changed without a convincing reason. If you use the memory types that are available for the Xeon E5-2600 v2 based PRIMERGY servers, the "Auto" setting always provides the optimal memory performance. Setting the BIOS option "DRAM Maintenance" to "Disabled" also causes the BIOS options for "Patrol Scrub" and "Refresh Rate Multiplier" to appear. The latter is automatically set in the "DRAM Maintenance = Auto" setting based on the memory modules used.

The BIOS option "DRAM Maintenance" is not available for the PRIMERGY CX servers. The settings for the "Patrol Scrub" and "Refresh Rate Multiplier" options can also be made there, independent of the "DRAM Maintenance" option.

Patrol Scrub

BIOS Setup Menu	BIOS Option	Settings	Performance	Low Latency	Energy Efficiency
Advanced > Memory Configuration [Chipset > Northbridge] [1)]	Patrol Scrub	Disabled **Enabled**	Enabled	Disabled	Enabled

1) Different menu structure for PRIMERGY CX servers

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.

Refresh Rate Multiplier

BIOS Setup Menu	BIOS Option	Settings	Performance	Low Latency	Energy Efficiency
Advanced > Memory Configuration [Chipset > Northbridge] [1)]	Refresh Rate Multiplier	1x **2x** 3x 4x	1x	1x	1x

[1)] Different menu structure for PRIMERGY CX servers

To enable the "Refresh Rate Multiplier" to be changed it is necessary for the BIOS option "DRAM Maintenance" to be set to "Disabled" for the PRIMERGY BX, RX and TX servers. Changing the DRAM Refresh rate is only relevant for older memory types, for example the Xeon E5-2600 based predecessor generation, which with certain access patterns can show an accumulation of correctable memory errors that is known as the pass-gate effect. To eliminate this effect the default setting "DRAM Maintenance = Auto" sets the DRAM Refresh Rate for such memory types to 2x (double), otherwise to 1x (single) for the PRIMERGY BX, RX and TX servers. While accepting the option of frequent, correctable memory errors, the refresh rate can also be set to 1x for older memory modules. If the "DRAM Maintenance" setting is changed to "Disabled", the refresh rate multiplier is preset to 2x regardless of the configured memory types. In this case, the user assumes responsibility for the correct setting of the option. The standard setting for the PRIMERGY CX servers is 2x and the user always assumes responsibility for required changes to the option. Memory configurations with older memory modules, which have the pass-gate effect, are the exception. In such cases, the memory modules are automatically operated with a refresh rate of 2x regardless of the user setting in the BIOS.

Onboard USB Controllers / EHCI Controller 1

PRIMERGY BX, RX, TX servers

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

PRIMERGY CX servers

BIOS Setup Menu	BIOS Option	Settings	Performance	Low Latency	Energy Efficiency
Chipset > Southbridge > USB Configuration	EHCI Controller 1	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=84dc1adf-adb8-419f-8d08-b226eb077e46

http://docs.ts.fujitsu.com/dl.aspx?id=0a62a369-241b-4de2-a970-41437aa4901b

http://docs.ts.fujitsu.com/dl.aspx?id=c9c496f2-3cc9-4d8d-8e34-64d997bd0a86

Memory performance of Xeon E5-2600 v2 (Ivy Bridge-EP)-based systems
http://docs.ts.fujitsu.com/dl.aspx?id=a344b05e-2e9d-481b-8c9b-c6542defd839

PRIMERGY Manuals

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

PRIMERGY BIOS downloads

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

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

2015-02-12 WW EN