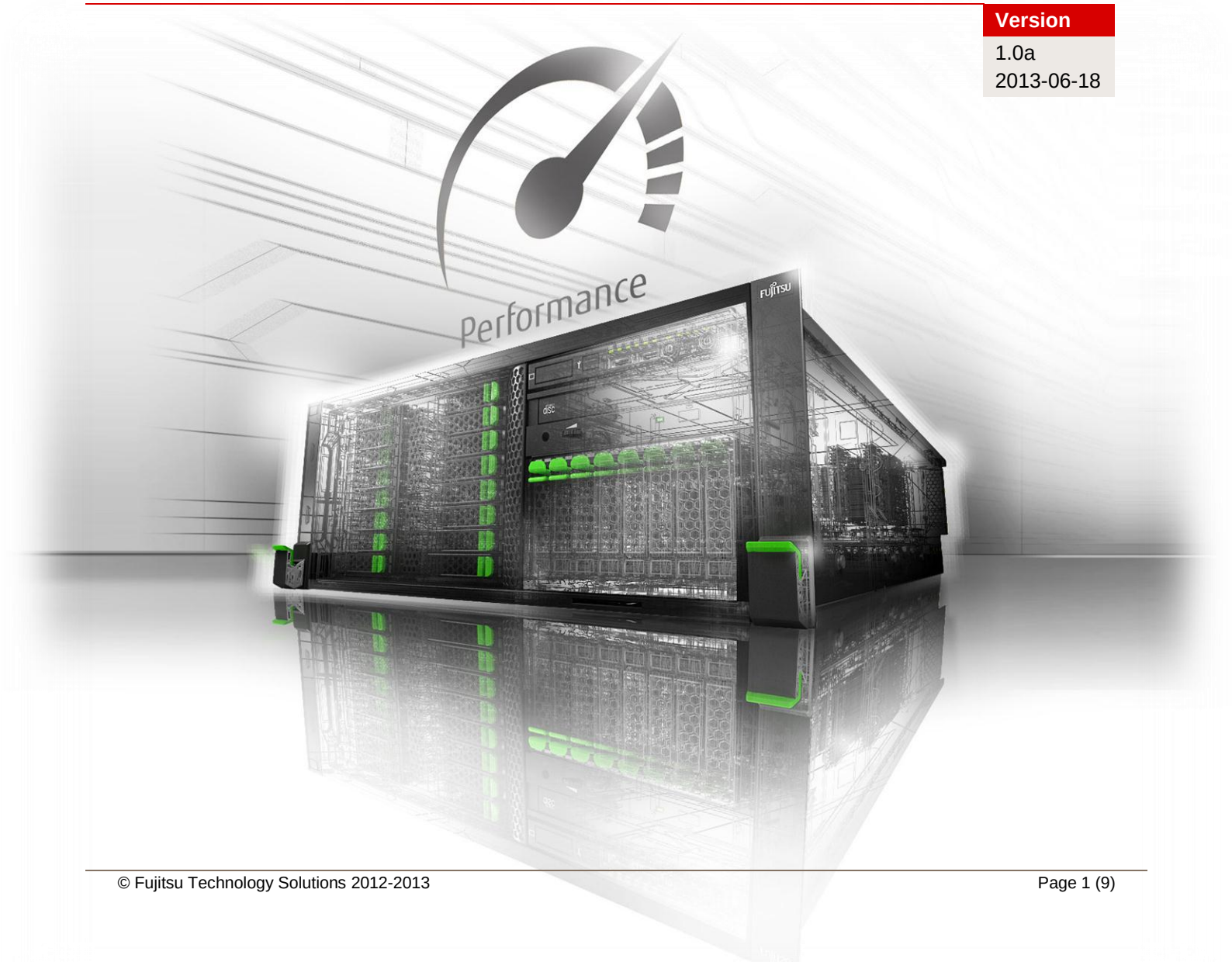
WHITE PAPER
FUJITSU PRIMERGY SERVER BIOS SETTINGS FOR PERFORMANCE, LOW-LATENCY AND ENERGY EFFICIENCY

This document explains the BIOS settings that are valid for the Intel Xeon E5-2400/2600/4600 based PRIMERGY server generation.

Recommendations are also provided on how the appropriate BIOS settings should be configured. The objectives here are to optimize PRIMERGY servers for best performance and maximum energy efficiency. In addition to optimization for pure performance, application scenarios are also taken into account, in which the shortest possible response time matters.

Version
1.0a
2013-06-18



Contents

Document history

Version 1.0

First edition

Version 1.0a

Minor change for general publication

Overview

When they leave the factory, Fujitsu PRIMERGY servers are already configured with BIOS standard settings, which provide an optimal relationship 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 maximum throughput, i.e. best performance with the lowest possible latency. This is particularly the case in the field of high performance computing (HPC) and for finance market applications, where it is a question of processing millions of transactions per second and data without any delay in real time if possible. As well as this scenario, there are also environments in which it is not pure performance that plays the major role, but maximum energy efficiency. This applies e.g. to data center operators, who only have a limited electrical power budget at their disposal. Here an attempt is made to optimize the servers in such a way that they provide as much throughput as possible, but with the lowest possible electrical power consumption. This white paper makes recommendations for optimal BIOS settings for these scenarios.

PRIMERGY BIOS options

This white paper only contains BIOS options that are valid for the Intel Xeon E5-2400/2600/4600 based PRIMERGY servers. 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 options 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 performance, a short response time and energy efficiency

The following table lists recommendations for BIOS options, which optimize the server either for best performance, a short response time 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.

Before changes are made to the BIOS options listed in the following table, it is advisable to observe the footnotes and subsequent description of the BIOS options.

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	[Hardware] [Adjacent Sector] [DCU Streamer] [DCU Ip] Prefetcher	Disabled **Enabled**	Enabled	Enabled	Disabled
Advanced > CPU Configuration	Intel Virtualization Technology	Disabled **Enabled**	Disabled [4]	Disabled	Disabled
Advanced > CPU Configuration	Power Technology	Disabled **Energy Efficient** Custom	Custom	Custom	Custom
Advanced > CPU Configuration	Turbo Mode	Disabled **Enabled**	Enabled	Enabled	Enabled
Advanced > CPU Configuration	Energy Performance [5]	Performance **Balanced Performance** Balanced Energy Energy Efficient	Performance	Performance	Energy Efficient
Advanced > CPU Configuration	CPU C3 Report [5]	**Disabled** Enabled	Disabled	Disabled	Enabled
Advanced > CPU Configuration	CPU C6 Report [5]	Disabled **Enabled**	Disabled	Disabled	Enabled
Advanced > CPU Configuration	CPU C7 Report [5]	**Disabled** Enabled	Disabled	Disabled	Enabled
Advanced > CPU Configuration	Package C-State limit [5]	C0 C2 C6 C7 **No Limit**	C0	C0	No Limit
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	QPI Link Power Management	Disabled **Enabled**	Disabled	Disabled	Enabled
Advanced > CPU Configuration	Frequency Floor Override	**Disabled** Enabled	Disabled [6]	Enabled	Disabled
Advanced > Memory Configuration	DDR Performance	Low-Voltage optimized Energy optimized **Performance optimized**	Performance optimized	Performance optimized	Low-Voltage optimized
Advanced > Memory Configuration	Patrol Scrub	Disabled **Enabled**	Disabled	Disabled	Disabled
Advanced > USB Configuration	Onboard USB Controllers	Disabled **Enabled**	Enabled	Enabled	Disabled [7]

[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] If virtualization is not used, this option should be set to "Disabled".
[5] This option is only visible if the setting for "Power Technology" is changed to "Custom".
[6] Switching on this for applications that do not use all CPUs of the system at full capacity can be advantageous.
[7] Switching off this option prevents the use of internal or external USB devices.

Description of the BIOS options

ASPM Support

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

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

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.

Prefetcher

The PRIMERGY server BIOS has several prefetcher options. These include "Hardware Prefetcher", "Adjacent Cache Line Prefetch", "DCU Streamer Prefetcher" and "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 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 then 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

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

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 performance. In order to see and individually set the corresponding options, select the setting "Custom" to enable you to make further settings for the BIOS options "Energy Performance", "CPU C3/C6/C7 Report" and "Package C-State limit". In addition to these settings, there are other options that are not listed here, because it is not advisable to deviate from the standard value. 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

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 is required, to disable the "Turbo Mode" option.

Energy Performance

Depending on the setting, this BIOS option manages the internal "Power Control Unit" 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". Some operating systems overwrite this setting, depending on how the appropriate energy-saving options are configured.

CPU C3/C6/C7 Report

These BIOS options are used to inform the operating system whether it can use the CPU C3, C6 or C7 states or not. 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. Since the waking-up from these sleep 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.

Using the "Disabled" setting for the BIOS option "CPU C3/C6/C7 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. Operating systems (this includes e.g. Red Hat Enterprise Linux 6.2) that determine the possible CPU C-states via drivers and not via ACPI are the exception. In such cases the use of CPU C-states must be prevented in the operating system (e.g. via kernel parameters).

Package C-State limit

In addition to the CPU or core C-states, there are also so-called package C-states, which not only allow the individual cores of a processor, but the entire processor chip to be put into a type of sleep state. As a result, power consumption is even further reduced. The "waking-up time" that is required to change from the lower package C-states to the active C0 state is even longer in comparison with the CPU or core C-states. If the "C0" setting is made in the BIOS, the processor chip always remains active.

QPI Link Frequency Select

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.

QPI Link Power Management

This BIOS option makes it possible to enable and disable the power-saving functions of the QPI links. Similar to the way in which CPUs have idle C-states, it is also possible with the QPI links to put them into a type of sleep state if one or both processors are idle. The savings through "QPI Link Power Management" in an idle state is particularly large. However, power management increases the latency in precisely the same way as with the CPU C-states. This is why "Disabled" should be set for applications, for which maximum performance with the shortest possible response time matters.

Frequency Floor Override

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.

DDR Performance

This BIOS option controls the speed and the voltage with which the memory modules that are inserted in the system 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 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 memory configurations with one or two DIMMs per memory channel 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). The "Low-Voltage optimized" setting is recommended for optimal energy efficiency and the "Performance optimized" setting for maximum memory performance.

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 can be found in the Memory Performance white paper (see Literature at the end of the document).

Patrol Scrub

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

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 Systems

http://primergy.com/

PRIMERGY Manuals

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

PRIMERGY BIOS download

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

White Paper: Memory performance of Xeon E5-2400 (Sandy Bridge-EN) based systems

http://docs.ts.fujitsu.com/dl.aspx?id=4bf85b37-fd9c-44a2-8593-3a913e007bd5

White Paper: Memory performance of Xeon E5-2600/4600 (Sandy Bridge-EP) based systems

http://docs.ts.fujitsu.com/dl.aspx?id=a17dbb55-c43f-4ac8-886a-7950cb27ec2a

PRIMERGY Performance

http://www.fujitsu.com/fts/products/computing/servers/primergy/benchmarks/

Contact

FUJITSU

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

PRIMERGY Product Marketing

mailto:Primergy-PM@ts.fujitsu.com

PRIMERGY Performance und Benchmarks

mailto:primergy.benchmark@ts.fujitsu.com