Fujitsu Server PRIMERGY
BIOS optimization for AMD EPYC 9004 and 9005 Processor-based systems

This document explains the BIOS settings that can be modified for the AMD EPYC 9004 and 9005 series processor-based PRIMERGY server generation (PRIMERGY RX1440 M2 and RX2450 M2).

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

Version
1.1 2025-01-14

Contents

Overview **3**

Application scenarios **4**

Performance 4

Low Latency 4

Energy savings / Energy efficiency 5

Application Profile 6

PRIMERGY BIOS options **7**

Recommendations for optimization 7

BIOS options details 11

Appendix **28**

Literature **32**

Overview

When Fujitsu PRIMERGY servers leave the factory, they are already configured with BIOS standard settings, which provide an optimal ratio between performance and energy efficiency for the most common application scenarios. And yet there are situations in which it may be necessary to modify the standard settings depending on requirements for the most throughput possible (performance), as little latency as possible (low latency), or emphasizing as much energy conservation as possible (energy efficiency). This document offers best-practice recommendations for optimal BIOS settings for these three scenarios, which are explained in more detail below.

In addition to the BIOS settings, the entire system must also be considered when optimizing PRIMERGY servers. The following aspects should be given consideration when planning server systems:

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

Application scenarios

Performance

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

Low Latency

Minimum possible latency is a requirement that comes from the High Performance Computing (HPC) sector in particular and from finance market applications, where the object is to process millions of transactions per second and data in real time without any delay. Users in this segment are not primarily concerned with achieving the maximum possible throughput through system optimization but emphasize more on increasing the speed of each individual transaction, i.e., reducing the time required to perform an individual transaction. In such cases, the focus is placed on the response time of a system, the so-called latency (typically measured in nanoseconds, microseconds, or milliseconds). The BIOS offers a variety of options to reduce latency. On the one hand, it is possible - such as when you know that the corresponding application does not make efficient use of all the threads available in the hardware - to disable threads that are not needed (Simultaneous Multi Processing) or even cores in the BIOS in order in this way to minimize fluctuations in computing speed that especially occur in a number of HPC applications. Furthermore, the disabling of cores that are not needed can improve the Turbo mode performance of the remaining cores under certain operating conditions. On the other hand, there are scenarios which require performance that is as constant as possible. In this case, it is necessary to reduce frequency changes and keep the response time constant by changing [Determinism Slider] to [Performance] or [Core Performance Boost] to [Disabled]. Although the current generation of AMD processors deliver a clearly better Turbo mode performance than the predecessor generations, the maximum Turbo mode frequency is not guaranteed under certain operating conditions. In such cases, disabling the Turbo mode can help avoid changes in frequency. Energy-saving functions, whose aim is to save energy whenever possible, through frequency/voltage reduction and through the disabling of certain function blocks and components, also have a negative impact on the response time. The stricter the energy-saving mode, the lower the performance. Furthermore, in each one of these energy-saving modes, the processor requires a certain time in order to change back from temporarily reduced performance to maximum performance. This time worsens the latency of the system, particularly after a transaction is pending and the system remains idle, or if the system load fluctuates irregularly. This document explains how to configure the power saving modes for users from the low-latency segment in order to minimize system latency. However, the optimization of server latency, particularly in an idle state, always results in higher electrical power consumption.

Note about "Performance" and "Low latency":

The maximum throughput or minimum latency of the I/O system can be of significance for I/O critical applications. These values have - in conjunction with the I/O system - a different meaning to the one associated with processors. For example, the I/O throughput means the amount of data transferred per time unit by the I/O system. In order to achieve maximum I/O 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 are in conjunction with optimally set I/O components. This almost always provides the highest possible optimization for these components. However, in certain rare situations, these target values can be missed with very high requirements (for SSDs). The solution can be to set BIOS options below, which can suppress frequency degradation in non core parts of processor (see the respective section for a more detailed description).

- [Global C-state Control]
- [DF Cstates]
- [APBDIS] and [DfPstate]
- [Power Profile Selection]

Energy savings / Energy efficiency

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

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

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

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

Application Profile

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

PRIMERGY servers provide the following 10 type of application profiles.

- **Total Throughput Performance**
  The profile optimized for the workload which requires the maximum throughput.

- **Single Thread Performance**
  The profile optimized for the workload which requires the peak performance of single core, rather than the throughput.

- **Low Latency**
  The profile to minimize the execution time of individual processing

- **Energy Efficiency**
  The profile to balance between the performance and the power

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

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

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

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

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

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

Note :

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

PRIMERGY BIOS options

This white paper contains information about BIOS options that are valid for the AMD EPYC 9004 and 9005 series processor based PRIMERGY servers. These servers are:

- PRIMERGY RX1440 M2
- PRIMERGY RX2450 M2

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

Recommendations for optimization

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

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

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

Recommended BIOS settings

BIOS Setup Menu	Setting[1]	Performance	Low Latency	Energy Efficient
Advanced -> CPU Configuration				
SMT Control	Disabled / **Enabled**	Enabled	Disabled	Enabled
CCD Control	**All** / 2 / 4 / 6 / 8 / 10[2]	All	All	All
Core Control	**All** / 1 / 2 / 3 / 4 / 5 / 6 / 7 / 8 / 9 / 10 / 11 / 12 / 13 / 14 / 15[2]	All	All	All
Prefetcher • L1 Stream HW Prefetcher • L1 Stride Prefetcher • L1 Region Prefetcher • L2 Stream HW Prefetcher • L2 Up/Down Prefetcher	Disabled[3] / **Enabled**	 Enabled Enabled Enabled Enabled Enabled	 Enabled Enabled Enabled Enabled Enabled	 Disabled Disabled Enabled Disabled Enabled
• Core performance Boost	Disabled / **Enabled**	Enabled	Enabled	Enabled
• BoostFmaxEn	**Manual** / Auto	Manual	Manual	Manual
BoostFmaxLimit[4]	[**0** - 65535]	0	0	0
Determinism Slider	Power / **Performance**	Power	Performance	Performance
TDP Control	Manual / **Auto**	Manual	Manual	Manual
TDP Limit[5]	[**85** – 400]	Max. per each SKU's specification	Max. per each SKU's specification	Max. per each SKU's specification
Package Power Limit Control	Manual / **Auto**	Manual	Manual	Manual
Package Power Limit[6]	[**85** – 400]	Max. per each SKU's specification	Max. per each SKU's specification	Max. per each SKU's specification
Global C-state Control	Disabled / **Enabled**	Enabled	Enabled	Enabled

[1] The settings in bold print are the default values.

[2] The actual options available depend on CPU installed.

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

[4] This option is only available if [BoostFmaxEn] is [Manual]

[5] This option is only available if [TDP Control] is [Manual]

[6] This option is only available if [Package Power Limit Control] is [Manual]

BIOS Setup Menu	Settings[1]	Performance	Low Latency	Energy Efficient
DF Cstates	Disabled / **Enabled**	Enabled	Disabled	Enabled
ACPI CST C2 latency	[18 – 1000] (Default: **800**)	800	800	800
APBDIS [7]	**0** / 1	0	1	0
DfPstate[7,8]	[**0** – 2]	-	0	-
xGMI Max Link Speed	20Gbps / 25Gbps / 32Gbps / **Auto**	Auto	Auto	20Gbps
Advanced -> Memory Configuration				
Memory Clock[9]	**Auto** / DDR3200 / DDR3600 / DDR4000 / DDR4400 / DDR4800 / DDR5200 / DDR5600 / DDR6000 / DDR6400	Auto	Auto	DDR4000
DRAM Scrub Time	Disabled / 1 hour / 4 hours / 6 hours / 8 hours / 12 hours / 16 hours / **24 hours** / 48 hours	24 hours	Disabled	24 hours
Power Down Enable	**Disabled** / Enabled	Disabled	Disabled	Enabled
Power Profile Selection	High Performance Mode / **Efficiency Mode** / Maximum IO Performance Mode / Balanced Memory Performance Mode	High Performance Mode	High Performance Mode	Efficiency Mode

[7] Not available in EPYC 9005 series processor as of Jan-2025. Will be available with BIOS which is released in Mar-2025.

[8] This option is only available if [APBDIS] is [1].

[9] Settings of DDR5200, DR5600, DDR6000, and DDR6400 are only available in AMD EPYC 9005 series processors. The maximum memory transfer rate that can be operated depends on the server model.

BIOS Setup Menu	Settings[1]	Performance	Low Latency	Energy Efficient
NUMA nodes per socket[2]	NPS0 / **NPS1** / NPS2 / NPS4	NPS4	NPS4	NPS2
ACPI SRAT L3 Cache As NUMA Domain	**Disabled** / Enabled	Disabled	Disabled	Enabled

BIOS options details

This section provides details about each BIOS option.

Since the effect of changing BIOS options is affected by the hardware / software configuration and other BIOS / OS option settings, be sure to verify these settings in your operational environment.

SMT Control

BIOS Setup Menu	BIOS Option	Setting	Performance	Low Latency	Energy Efficiency
Advanced > CPU Configuration	SMT Control	Disabled **Enabled**	Enabled	Disabled	Enabled

This BIOS option enables or disables AMD Simultaneous Multi-Threading (SMT). Generally, Fujitsu recommends that you enable [SMT Control] ([Enabled]). Nevertheless, it can make sense to disable [SMT Control] 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 [SMT Control] can prevent the associated performance fluctuations of computing operations and thus improve latency. Applications that use a lot of AVX instructions, such as HPC, may also improve throughput performance by the setting of [Disabled].

CCD Control / Core Control

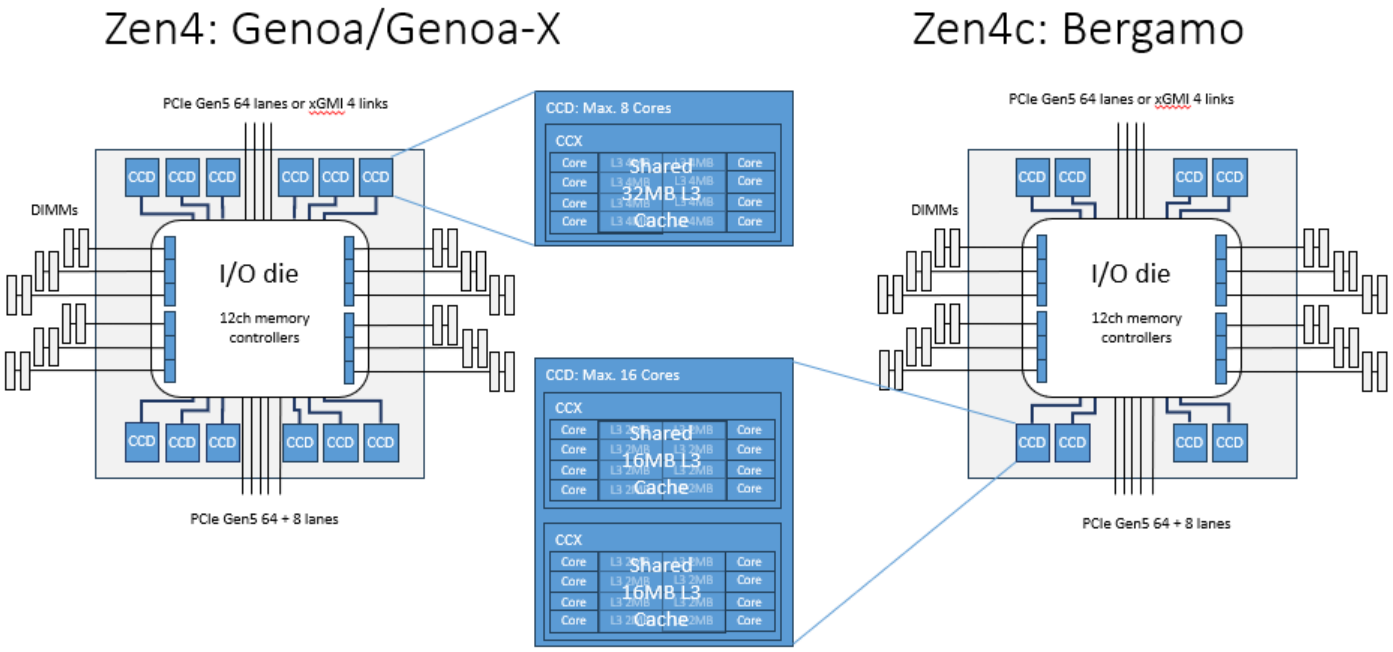
BIOS Setup Menu	BIOS Option	Setting	Performance	Low Latency	Energy Efficiency
Advanced • > CPU Configuration	CCD Control	**All** 2 4 6 8 10	All	All	All
	Core Control	**All** 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15	All	All	All

[CCD Control] option and [Core Control] option, alone or in combination, can limit the number of cores available in a processor.

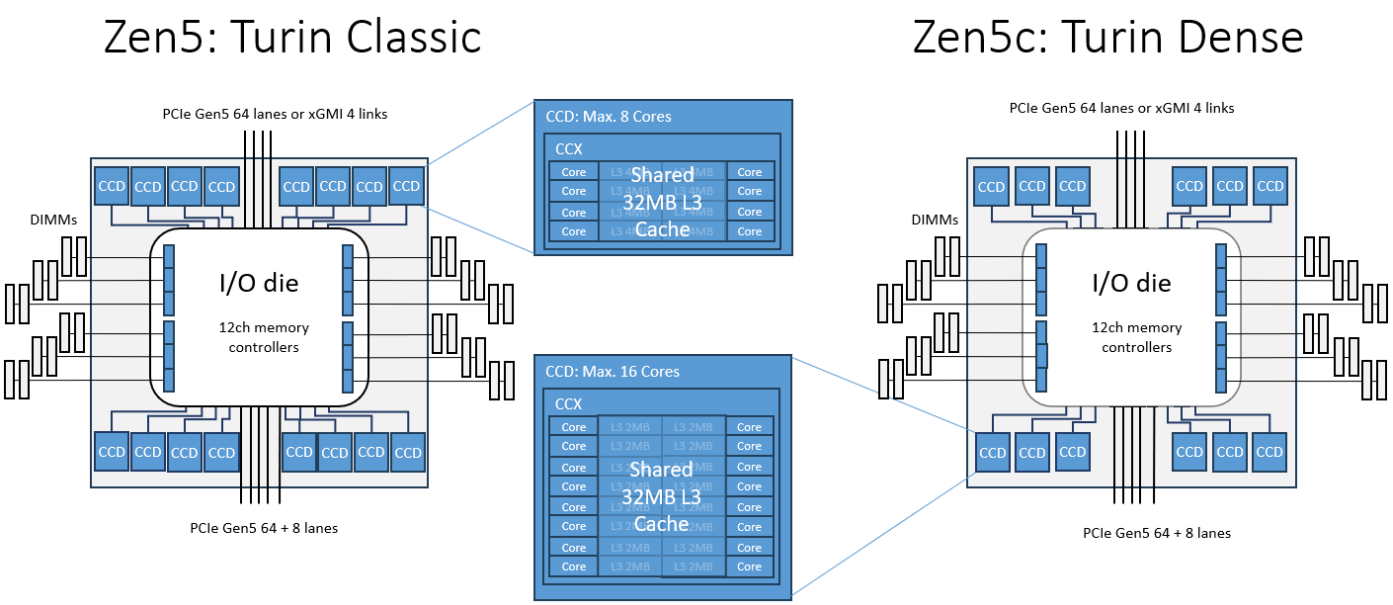
[CCD Control] option can specify the number of enabled CCDs (Core Complex Die) in a processor. The AMD EPYC 9004 series processors available in PRIMERGY RX1440 M2 and RX2450 M2 have two different architectures, the Zen4 (Genoa/Genoa-X) and the Zen4c (Bergamo). In the same way, the AMD EPYC 9005 series processors available in PRIMERGY RX1440 M2 and RX2450 M2 also have two different architectures, the Zen5 (Turin Classic) and Zen5c (Turin Dense). Each architecture has different maximum number of CCDs. If you specify a number that exceeds the maximum number of CCDs in the processor, the setting is ignored. Maximum number of CCDs per architecture are as follows.

- Zen4 (Genoa/Genoa-X): max. 12 CCDs
- Zen4c (Bergamo): max. 8 CCDs
- Zen5 (Turin Classic): max. 16 CCDs
- Zen5c (Turin Dense): max. 12 CCDs

The following figure shows the configuration of each architecture.



Architecture of AMD EPYC 9004 Series Processor



Architecture of AMD EPYC 9005 Series Processor

All CCDs are interconnected with memory, I/O, and other CCDs via the I/O die. Each CCD is connected to the I/O die by an interface called GMI (Global Memory Interconnect), and the I/O die has xGMI (External Global Memory Interconnect), which is an interface to other processors. These data transfer and control paths are called Infinity fabrics.

While [CCD Control] option specifies the number of CCDs, [Core Control] option can specify the number of enabled cores in each CCD. As shown above, Zen4 processors have a maximum of 8 cores in a CCD (Total 96 cores for a processor) and Zen4c processors have 16 cores in a CCD (Total 128 cores for a processor). Different processor models have different numbers of CCDs and different number of cores in a CCD. If you specify a number that exceeds the number of cores in the CCD, the setting is ignored.

For example, the EPYC 9654, which has the largest number of cores in the Zen4 architecture, has 12 CCDs, 8 cores per CCD, and 96 cores for the entire processor. If [CCD Control] option is set to [10] and [Core Control] option is set to [4], the number of cores per CPU is 40 (=10 CCDs x 4 cores).

Reducing the number of CCDs in [CCD Control] option disables all cores and L3 caches on the reduced CCD. For workloads where there are few active cores and multiple cores share data from the same L3 cache, it may be a good idea to disable some CCDs and allow processing on the same CCD to improve performance. On the other hand, if you reduce the number of cores on a CCD in [Core Control] option, the remaining cores benefit from using more L3 cache capacity than before the configuration change. Although maximum throughput is generally achieved with all the cores, you can utilize higher Turbo Mode frequency for remaining active cores by disabling the CCDs or cores that are not needed. This is advantageous especially with latency-sensitive applications that do not utilize all the cores. This works because the disabled cores reduce the electrical power consumption of the processor and thereby allowing higher Turbo Mode frequencies on the remaining cores. This does not necessarily work with all the load profiles.

Prefetcher

BIOS Setup Menu	BIOS Option	Setting	Performance	Low Latency	Energy Efficiency
Configuration >CPU Configuration	L1 Stream HW Prefetcher	Disabled **Enabled**	Enabled	Enabled	Disabled
	L1 Stride Prefetcher		Enabled	Enabled	Disabled
	L1 Region Prefetcher		Enabled	Enabled	Enabled
	L2 Stream HW Prefetcher		Enabled	Enabled	Disabled
	L2 Up/Down Prefetcher		Enabled	Enabled	Enabled

The PRIMERGY server BIOS has several prefetcher options as above.

The prefetchers are processor functions, which uses a history of memory accesses and 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. In application scenarios, in which memory transfer is a performance bottleneck, the prefetchers can degrade the performance. In these cases, it can be advantageous to set the prefetcher options to [Disabled] so the bandwidth that is otherwise used for the prefetching can be used. Furthermore, the power consumption of the server can be slightly reduced by disabling the prefetchers. Whether the prefetchers improve performance depends on the particular application. Before the prefetcher options are changed on the active systems, the effects of the individual settings for the respective application scenario should first be examined in a test environment.

Details of the individual prefetchers:

L1 Stream HW Prefetcher This prefetcher uses the history of L1 cache memory access patterns to fetch additional sequential lines in ascending or descending order.

L1 Stride Prefetcher The prefetcher uses the L1 cache memory access history of individual instructions to fetch additional lines when each access is a constant distance from the previous.

L1 Region Prefetcher This prefetcher uses the L1 cache memory access history to fetch additional lines when the data access for a given instruction that tends to be followed by a consistent pattern of other accesses within a localized region.

L2 Stream HW Prefetcher This prefetcher uses the history of L2 cache memory access patterns to fetch additional sequential lines in ascending or descending order.

L2 Up/Down Prefetcher Uses the L2 cache memory access history to determine whether to fetch the next or previous line for all memory accesses.

Core Performance Boost / BoostFmaxEn / BoostFmax

BIOS Setup Menu	BIOS Option	Setting	Performance	Low Latency	Energy Efficiency
Advanced > CPU Configuration	Core Performance Boost	Disabled **Enabled**	Enabled	Enabled	Enabled
	BoostFmaxEn	**Manual** Auto	Manual	Manual	Manual
	BoostFmax	[**0** – 65535]	0	0	0

[Core Performance Boost] option enables or disables the Core Performance Boost function of processor.

The AMD EPYC 9004 and 9005 series processors support 3 P-states: P0 (rated frequency), P1, and P2, and the voltage and frequency of each P-state are determined based on the CPU model. The Core Performance Boost function permits the processor to provide more computing performance by increasing the frequency above the rated frequency of P0. The maximum achievable frequency is influenced by numerous factors - processor type, number of active processor cores, power supply, current electrical power consumption, temperature. 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 [Core Performance Boost] option set at the standard setting [Enabled], as performance is substantially increased by the higher frequencies. However, as the higher frequencies depend on its operating conditions as mentioned above and are not always guaranteed, it can be advantageous for application scenarios, in which you want constant performance or to lower electrical power consumption, to disable the [Core Performance Boost] option.

When [BoostFmaxEn] option is set to [Manual], [BoostFmax] option can set the maximum core performance boost frequency in MHz. If you set [BoostFmaxEn] option to [Auto] or [BoostFmax] option to [0], the CPU will not change the maximum core performance boost frequency from specified by the CPU specification. Any setting for [BoostFmax] option that exceeds the frequency range defined by the CPU specifications will be ignored. If the maximum core performance boost frequency is not required, you can reduce power consumption by setting the frequency in the range between the maximum core performance boost frequency and the rated frequency.

Determinism Slider

BIOS Setup Menu	BIOS Option	Setting	Performance	Low Latency	Energy Efficiency
Advanced > CPU Configuration	Determinism Slider	Power **Performance**	Power	Performance	Performance

[Determinism Slider] option specifies the CPU performance level.

When [Performance] is selected, the processor operates to reduce performance variation among cores. While the processor provides stable performance, it may operate at a lower power level than the setting specified in TDP Control, reducing overall peak performance. If you select [Power], each core operates independently at its maximum performance level, as long as the power consumption of processor does not exceed the power set in TDP Control. In that case, overall performance will be higher, but performance may vary between cores.

TDP Control / TDP Limit

BIOS Setup Menu	BIOS Option	Setting	Performance	Low Latency	Energy Efficiency
Advanced > CPU Configuration	TDP Control	Manual **Auto**	Manual	Manual	Manual
	TDP Limit	[**85** – 400]	Max. per each SKU's specification	Max. per each SKU's specification	Max. per each SKU's specification

These BIOS options allow you to set the Thermal Design Power (TDP) level, which is the power limit that the processor can be cooled.

If [TDP Control] option is set to [Auto], the system operates with the default TDP defined in the specifications of the installed processor. If [TDP Control] option is set to [Manual], you can set a TDP value that is lower or higher than the default TDP in [TDP Limit] option, and the performance and power consumption will change according to the set value.

Package Power Limit Control / Package Power Limit

BIOS Setup Menu	BIOS Option	Setting	Performance	Low Latency	Energy Efficiency
Advanced > CPU Configuration	Package Power Limit Control	Manual **Auto**	Manual	Manual	Manual
	Package Power Limit	[**85** – 400]	Max. per each SKU's specification	Max. per each SKU's specification	Max. per each SKU's specification

These BIOS options allow you to set the Package Power Limit (PPL) level, which is the power limit of the processor.

If [Package Power Limit Control] option is set to [Auto], the system operates with the default PPL defined in the specifications of the installed processor. If [Package Power Limit Control] option is

set to [Manual], you can set a PPL that is lower or higher than the default PPL in [Package Power Limit] option.

Typically, [TDP Limit] and [Package Power Limit] are set to the same value, but you may choose to set [Package Power Limit] to a lower value than [TDP Limit] to control the core performance boost frequency and reduce processor power consumption below the default PPL.

Global C-state Control

BIOS Setup Menu	BIOS Option	Setting	Performance	Low Latency	Energy Efficiency
Advanced > CPU Configuration	Global C-state Control	Disabled **Enabled**	Enabled	Enabled	Enabled

This BIOS option enables or disables C-states on all cores and the Infinity fabric. Even when [Disabled] is set, core C1 state cannot be disabled and cores operate at C0 or C1 state.

C-states are inactive power states of the processor. Core C0 is the operating state in which instructions are executed and other C-states are idle. Higher numbered C-states have lower power. C-states are notified by the BIOS to the OS through ACPI, and the software can dynamically request C-state changes by executing a HALT instruction or by reading from a specific I/O address.

The AMD EPYC 9004 and 9005 series processors are designed to support the 3 AMD-defined C-states (C0, CC1, and CC6). There is no one-to-one correspondence between ACPI-defined C-states and AMD-defined C-states.

We recommend that this option should be set to [Enabled]. If you want to disable C-states for workloads which requires lower latency, consider setting [DF Cstates] option to [Disabled] instead of setting this option to [Disabled].

Reference: Processor Power States

	 **Performance Power State (P-State)** Based on CPU utilization, P-states reduce the electrical power consumption while the processor is executing code	 **Idle Power State (C-State)** C-states reduce the electrical power consumption while the processor is not executing code
Core	The larger P-State, the lower the performance - Core Performance Boost State[10]: Performance level higher than P0 - P0: Rated frequency - P1: Performance level lower than P0 - P2: Performance level lower than P1	The larger C-State, the lower the power - C0: Active state - CC1: Core is halted. BIOS cannot disable this state - CC6[11]: Equivalent to ACPI-defined C2, flush L2 cache and the core logic is power gated
Infinity fabric	The larger P-State, the lower the performance - DFP0[12]: The maximum performance level - DFP1[12]: Performance level lower than DFP0 - DFP2[12]: Performance level lower than DFP1	The larger C-State, the lower the power - DFC0: Active state - DFC2[11,13]: When all cores are in CC6, disconnect all links, place links in low power state, place all DRAM in self refresh mode, and clock gate most of the processor

[10] This state is disabled if [Core Performance Boost] is [Disabled]. The maximum frequency of this state can be set by [BoostFmax].

[11] This state is disabled if [Global C-state Control] is [Disabled].

[12] If [APBDIS] is [1], P-State of Infinity fabric is fixed by [DfPstate].

[13] This state is disabled if [DF Cstates] is [Disabled].

DF Cstates

BIOS Setup Menu	BIOS Option	Setting	Performance	Low Latency	Energy Efficiency
Advanced > CPU Configuration	DF Cstates	Disabled **Enabled**	Enabled	Disabled	Enabled

This BIOS option enables or disables the Infinity fabric C-state feature.

[Enabled] allows the Infinity fabric to enter a low-power C-state when all cores have entered CC6 state (= ACPI-defined C2 state). Idle power can be reduced, but there is overhead when resuming from CC6 state, which leads to latency fluctuations. [Disabled] prevents the Infinity fabric from transitioning to C-state, and reduces latency jitter.

If [Global C-state Control] is set to [Disabled], the setting of this option is ignored and the behavior of DF Cstates are the same as when [DF Cstates] is set to [Disabled].

ACPI CST C2 latency

BIOS Setup Menu	BIOS Option	Setting	Performance	Low Latency	Energy Efficiency
Advanced > CPU Configuration	ACPI CST C2 latency	[18 – 1000] (Default: **800)**	800	800	800

This BIOS option allows you to set the periods in microseconds before an idle core transitions to ACPI-defined C2 state (= AMD-defined CC6 state).

Larger values reduce the number of transitions to C2 state and consequently the residence time of C2 state. Larger values may be advantageous for workloads where low latency variation is important, because there is overhead when returning from C2 state to active C0 state. On the other hand, setting a smaller value increases the residence time of C2 state and reduces idle power. In addition, while inactive cores are in C2 state, active cores may be able to achieve a higher core performance boost frequency with more power available.

APBDIS / DfPstates

BIOS Setup Menu	BIOS Option	Setting	Performance	Low Latency	Energy Efficiency
Advanced • > CPU Configuration	APBDIS	**0** 1	0	1	0
	DfPstate	[**0** – 2]	-	0	-

In the AMD EPYC 9004 and 9005 series processors, individual cores and the Infinity fabric run on independent frequencies. The Infinity fabric increases the frequency as the utilization rate increases. On the other hand, when utilization decreases, the frequency is minimized to conserve energy.

[APBDIS] BIOS option enables or disables the Infinity fabric operating frequency variations through the Algorithm Performance Boost (APB) capability of the CPU. If [ABPDIS] is set to [0], the processor dynamically switches the P-state of the Infinity fabric based on the utilization of the

Infinity fabric excluding cores. If [APBDIS] is set to [1], P-state of the Infinity fabric is fixed at the state set by [DfPstate] option, regardless of the utilization. In general, [APBDIS] option should be set to [0]. For applications where I/O latency or throughput is important, or when it is not important at all, set [APBDIS] to [1] and change [DfPstate] option to lock P-state of the Infinity fabric to a higher or lower frequency.

[DfPstate] option specifies the P-state of the Infinity fabric that you want to maintain when [APBDIS] is set to [1]. I/O intensive applications typically have no or very low processor load, so the processor power management mechanisms will try to set the frequency to the lowest possible, but if you want to avoid I/O (PCIe, memory, xGMI, etc.) frequency degradation, you can set [DfPstate] option to [0] and the Infinity fabric will always operate at the maximum frequency, DFP0 state. The higher the number in this setting, the lower the Infinity fabric frequencies.

xGMI Max Link Speed

BIOS Setup Menu	BIOS Option	Setting	Performance	Low Latency	Energy Efficiency
Advanced > CPU Configuration	xGMI Max Link Speed	20Gbps 25Gbps 32Gbps **Auto**	Auto	Auto	20Gbps

This BIOS option sets the maximum frequency for the E**x**ternal **G**lobal **M**emory **I**nterconnect (xGMI), the link between processors. For NUMA-insensitive workloads which require more communication between processors, choose the maximum speed of [32Gbps]. For workloads with low memory traffic or NUMA-aware workloads, there is less communication between processors, so you can choose a lower speed to reduce xGMI power consumption. As a result, you may be able to increase the frequency of Core Performance Boost.

Memory Clock

BIOS Setup Menu	BIOS Option	Setting	Performance	Low Latency	Energy Efficiency
Advanced > Memory Configuration	Memory Clock	**Auto** DDR3200 DDR3600 DDR4000 DDR4400 DDR4800 DDR5200 DDR5600 DDR6000 DDR6400	Auto	Auto	DDR4000

This BIOS option specifies the memory clock frequency.

[Auto] sets the maximum possible frequency for the configuration. For workloads where memory bandwidth is not critical, you may be able to reduce system power consumption by setting the memory clock frequency to a lower frequency.

If you specify the setting that isn't supported in the processor, the setting is ignored.

DRAM Scrub Time

BIOS Setup Menu	BIOS Option	Setting	Performance	Low Latency	Energy Efficiency
Advanced > Memory Configuration	DRAM Scrub Time	Disabled 1 hour 4 hours 6 hours 8 hours 12 hours 16 hours **24 hours** 48 hours	24 hours	Disabled	24 hours

This BIOS option enables or disables the so-called memory scrubbing, which cyclically accesses the main memory of the system in the background, regardless of the operating system, to detect and correct memory errors in a preventive way. The disabling of the [DRAM Scrub Time] 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. If you want to enable memory scrubbing, select the amount of time between [1 hour] and [48 hours] to run through all the memory in your system. In general workloads, the performance impact is small even if [DRAM Scrub Time] is enabled. But under certain circumstances with high latency requirements, low latency may not be guaranteed. By setting [DRAM Scrub Time] to [Disabled] or a longer test duration like [48 hours], you can reduce the frequency of extra memory accesses caused by memory scrubbing and reduce the risk of performance fluctuations.

Power Down Enable

BIOS Setup Menu	BIOS Option	Setting	Performance	Low Latency	Energy Efficiency
Advanced > Memory Configuration	Power Down Enable	**Disabled** Enabled	Disabled	Disabled	Enabled

This BIOS option enables or disables Power Down Mode, a power saving feature of DDR.

When the Memory Controller has been idle for a programmable amount of clock cycles, it will drive the DRAM CKE low to enter power down mode. The DRAM will be periodically taken out of power down mode as required for refresh operations or for ECC scrubbing. The DRAM will be taken out of power down mode automatically when memory transactions are received.

If [Power Down Mode] is set to [Enabled], system power can be reduced but returning from power down mode will take a delay of about dozens of memory clock cycles. For workloads where memory latency is critical, leave the setting to [Disabled].

Power Profile Selection

BIOS Setup Menu	BIOS Option	Setting	Performance	Low Latency	Energy Efficiency
Advanced > Memory Configuration	Power Profile Selection	High Performance Mode **Efficiency Mode** Maximum IO Performance Mode Balanced Memory Performance Mode	High Performance Mode	High Performance Mode	Efficiency Mode

This BIOS option selects the power management policy for the processor.

[Efficiency Mode], which emphasizes power efficiency, improves performance per power by adjusting the operating frequency of the core and Infinity fabric. But if the operating frequency drops, the processor's original performance may not be fully achieved. [High Performance Mode] is a performance-sensitive policy that controls the processor cores to operate at a high frequency. [Maximum IO Performance Mode] is another performance-oriented policy, but it strives to keep the Infinity fabric operating at a high frequency for heavy I/O operations. [Balanced Memory Performance Mode] is a policy that emphasizes a balance between core performance and memory performance.

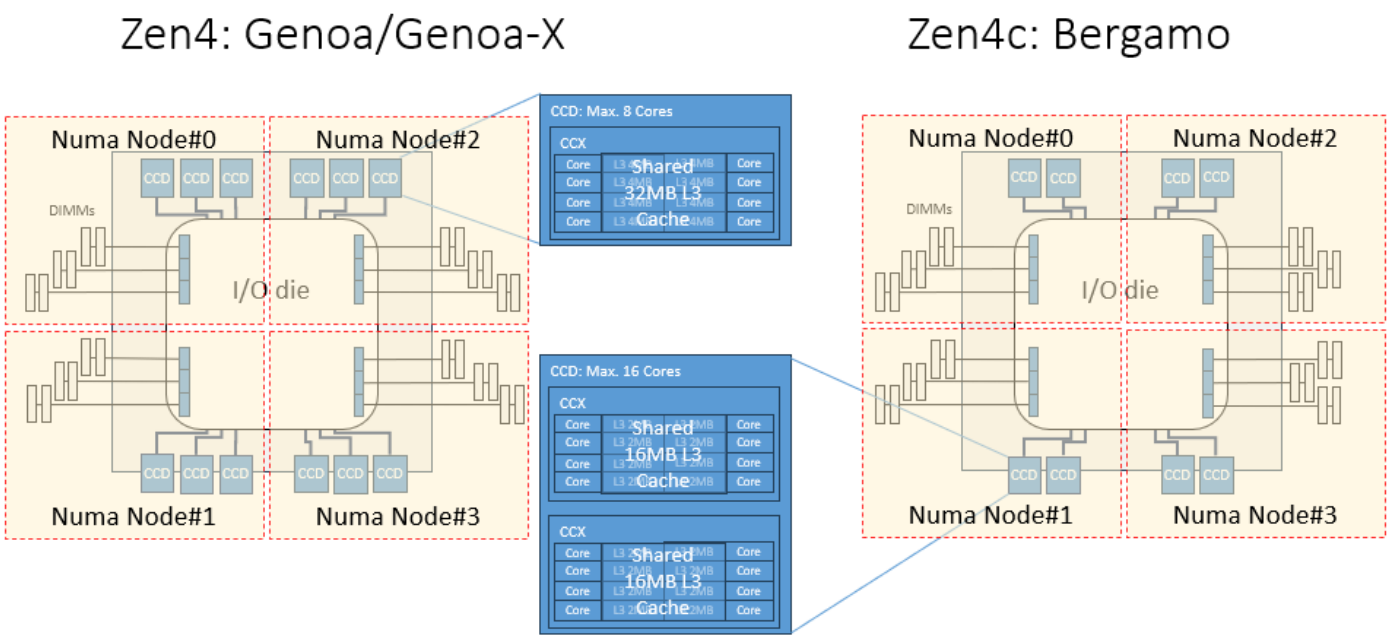
NUMA nodes per socket / ACPI SRAT L3 Cache As NUMA Domain

BIOS Setup Menu	BIOS Option	Setting	Performance	Low Latency	Energy Efficiency
Advanced > Memory Configuration	NUMA nodes per socket	NPS0 **NPS1** NPS2 NPS4	NPS4	NPS4	NPS2
	ACPI SRAT L3 Cache As NUMA Domain	**Disabled** Enabled	Disabled	Disabled	Enabled

[NUMA nodes per socket] option is a parameter that divides the cores, L3 cache, memory controller, and PCIe root complex into clusters within the processor. Splitting the NUMA domain improves access latency to L3 cache and memory from cores in a NUMA node. It is especially recommended for NUMA-optimized applications because it can minimize local memory latency and maximize bandwidth.

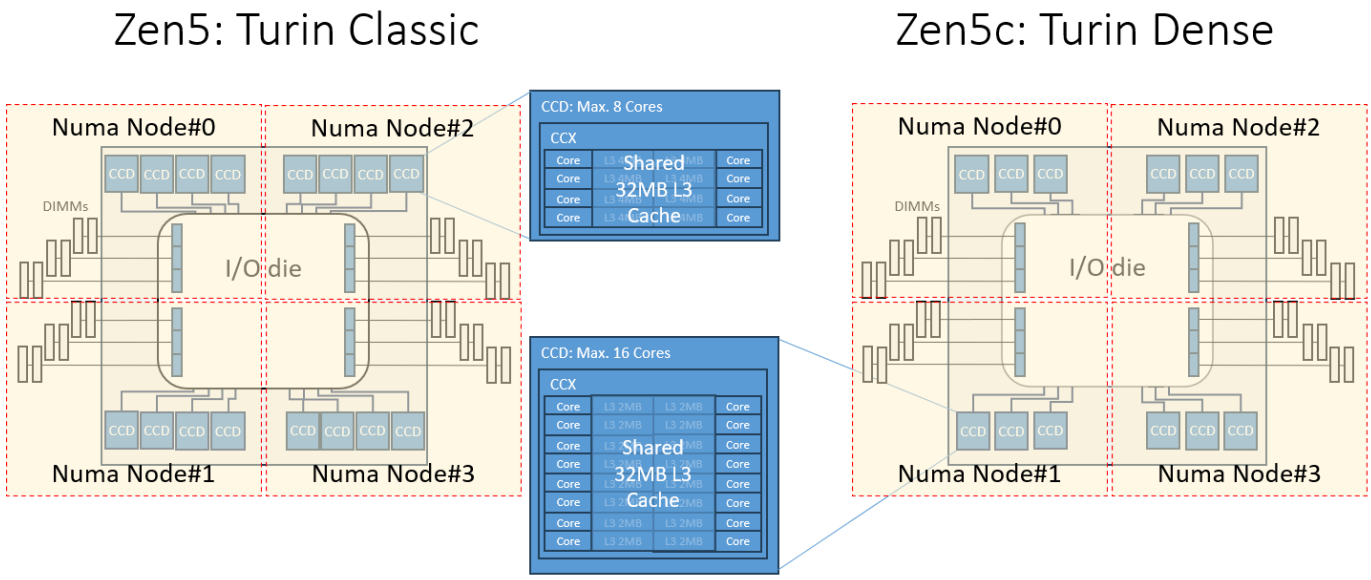
If [NUMA nodes per socket] is set to [NPS4], resources within a processor are associated with one of the four NUMA domains within the processor. When set to [NPS2], resources within a processor are associated with one of the NUMA domains in the two divided processors. When set to [NPS1], all resources in a processor are treated as a single NUMA domain. Unlike the other options, [NPS0] is available only on two-socket servers, which treats the two processors as a single NUMA domain. Memory is interleaved across the memory channels of the two processors. [NPS0] setting is generally not recommended unless you have special requirements, such as a non-NUMA OS.

If you use a processor with more than 64 logical CPUs in Windows, it is recommended that you change [NUMA nodes per socket] to [NPS2] or [NPS4] and limit the logical CPUs per NUMA node to no more than 64. The processor group that Windows uses to manage logical CPUs is capped at 64 logical CPUs, so any logical CPUs that exceed that limit are managed as a separate processor group. This results in uneven processor group size, which is a performance disadvantage. You can divide NUMA nodes into equally sized processor groups by the setting.



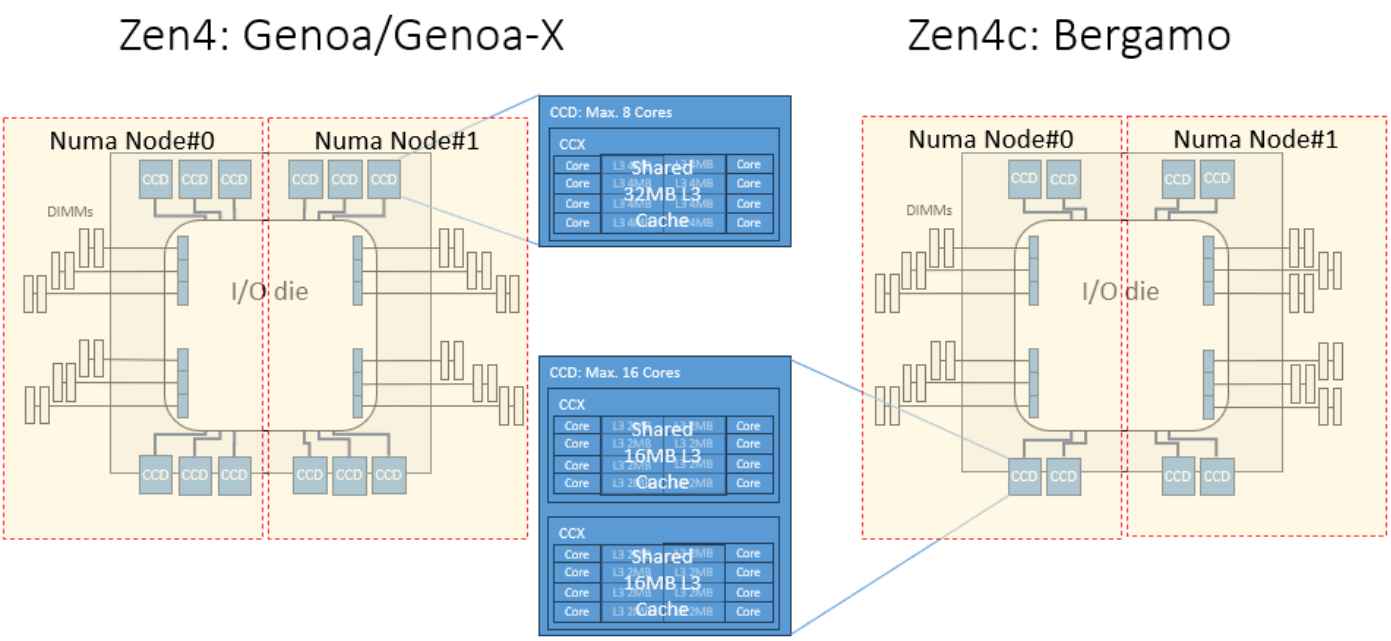
- Number of Numa nodes = Number of Memory Partitions = 4 per socket
- BIOS will attempt to interleave all memory channels (3-way) on each Numa node

Configuration of NPS4 (Zen4 / Zen4c)



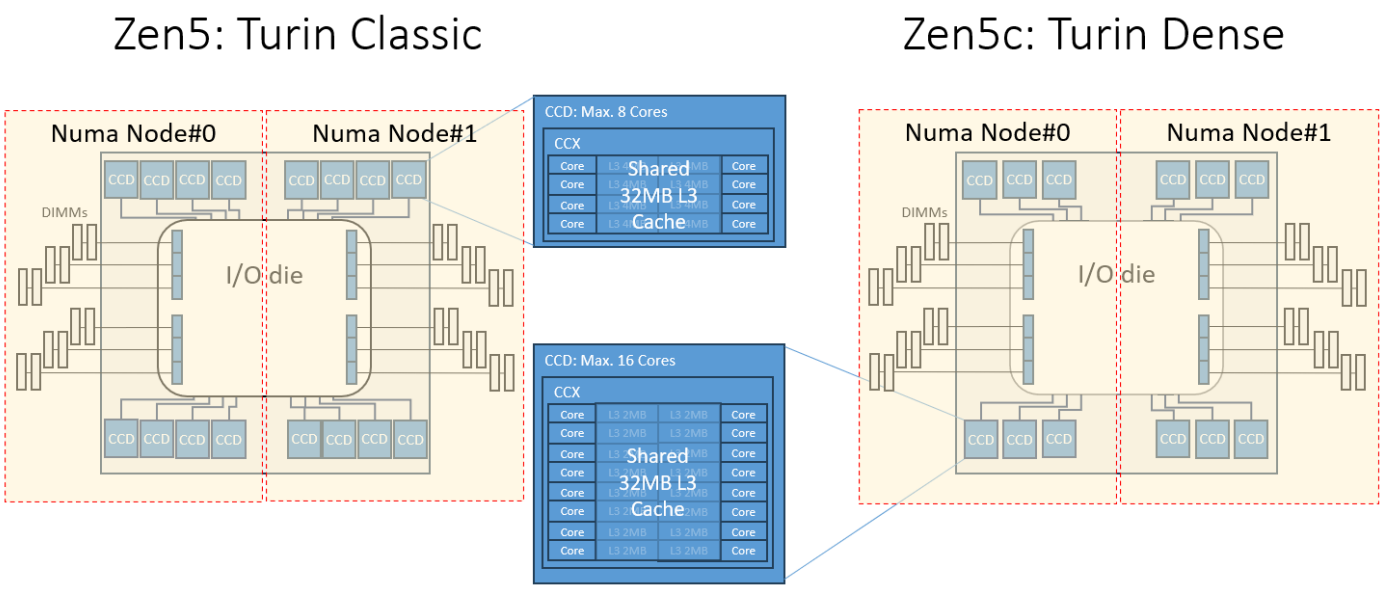
- Number of Numa nodes = Number of Memory Partitions = 4 per socket
- BIOS will attempt to interleave all memory channels (3-way) on each Numa node

Configuration of NPS4 (Zen5 / Zen5c)



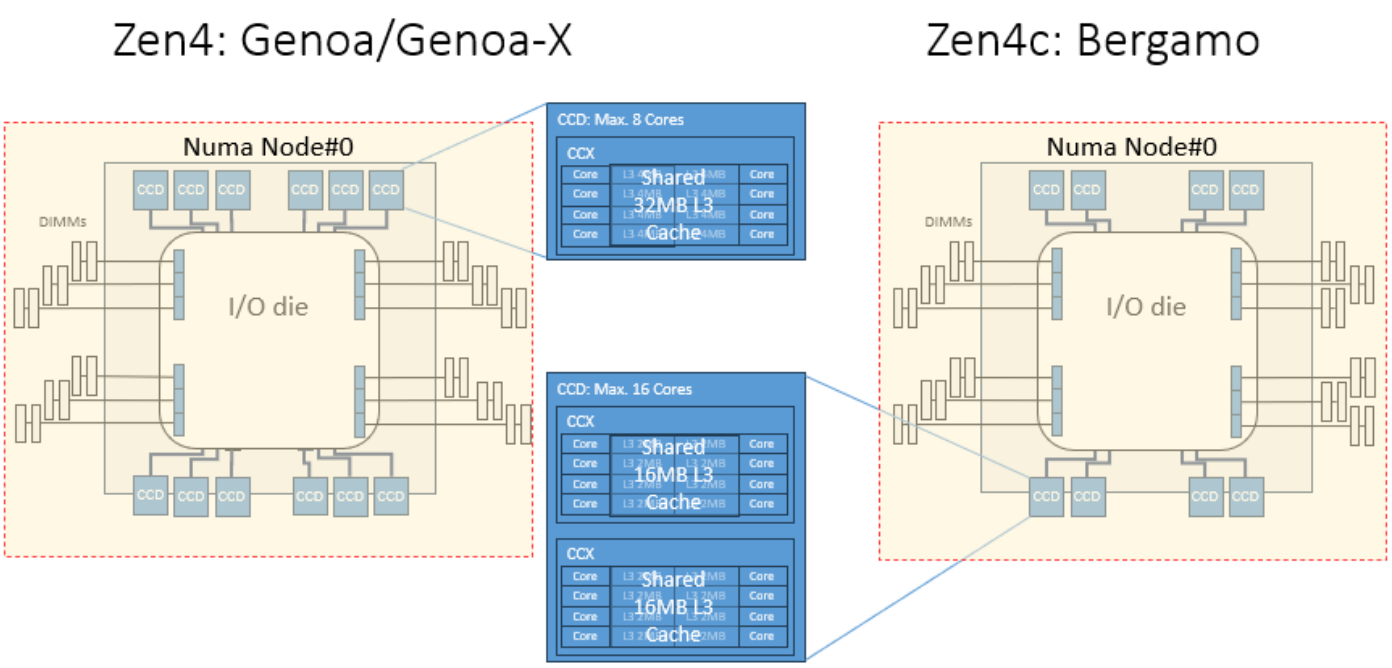
- Number of Numa nodes = Number of Memory Partitions = 2 per socket
- BIOS will attempt to interleave all memory channels (6-way) on each Numa node

Configuration of NPS2 (Zen4 / Zen4c)



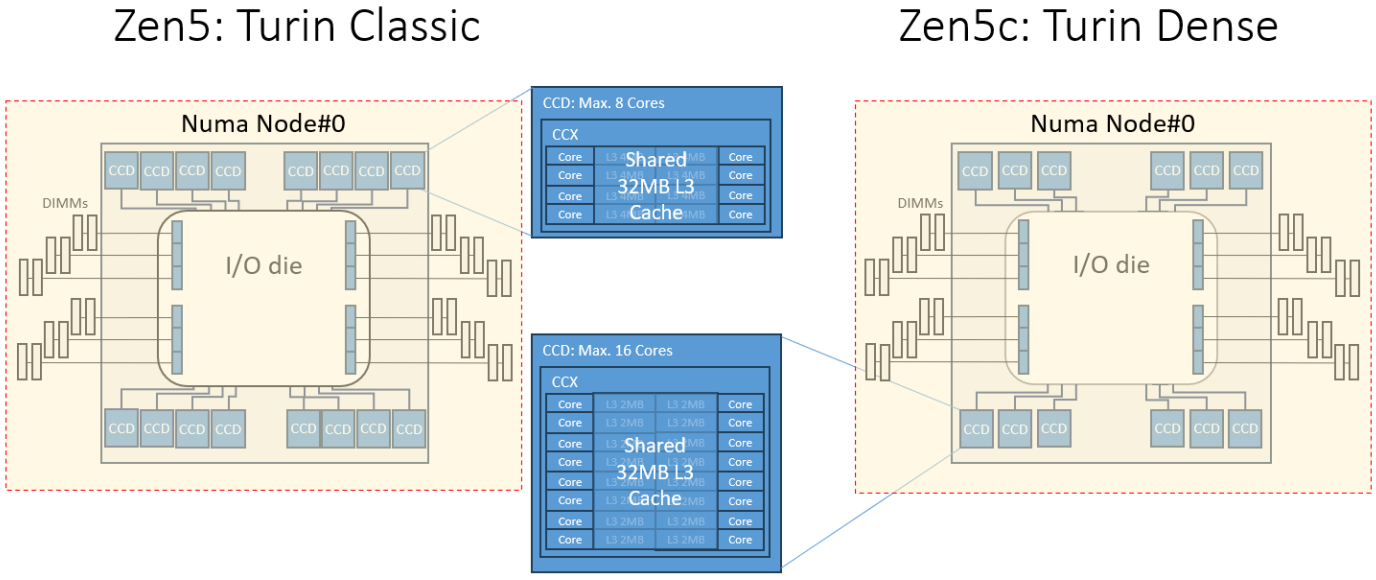
- Number of Numa nodes = Number of Memory Partitions = 2 per socket
- BIOS will attempt to interleave all memory channels (6-way) on each Numa node

Configuration of NPS2 (Zen5 / Zen5c)



- Number of Numa nodes = Number of Memory Partitions = 1 per socket
- BIOS will attempt to interleave all memory channels (12-way) on each Numa node

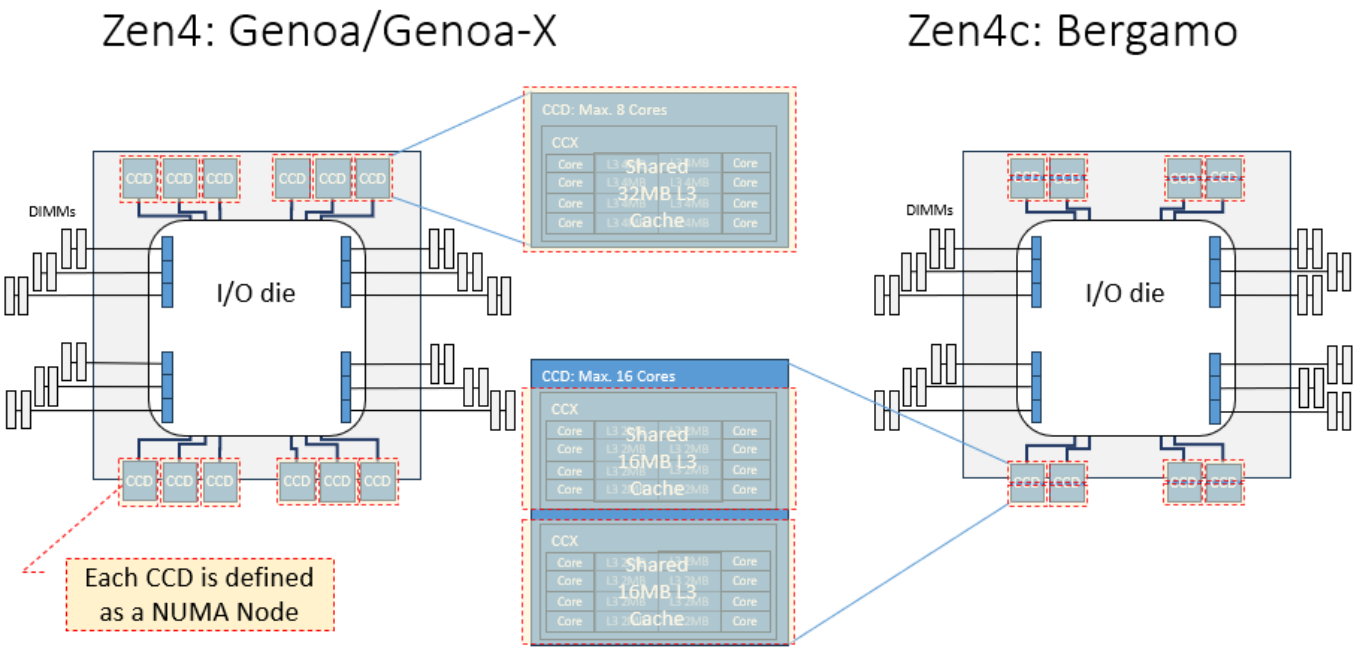
Configuration of NPS1 (Zen4 / Zen4c)



- Number of Numa nodes = Number of Memory Partitions = 1 per socket
- BIOS will attempt to interleave all memory channels (12-way) on each Numa node

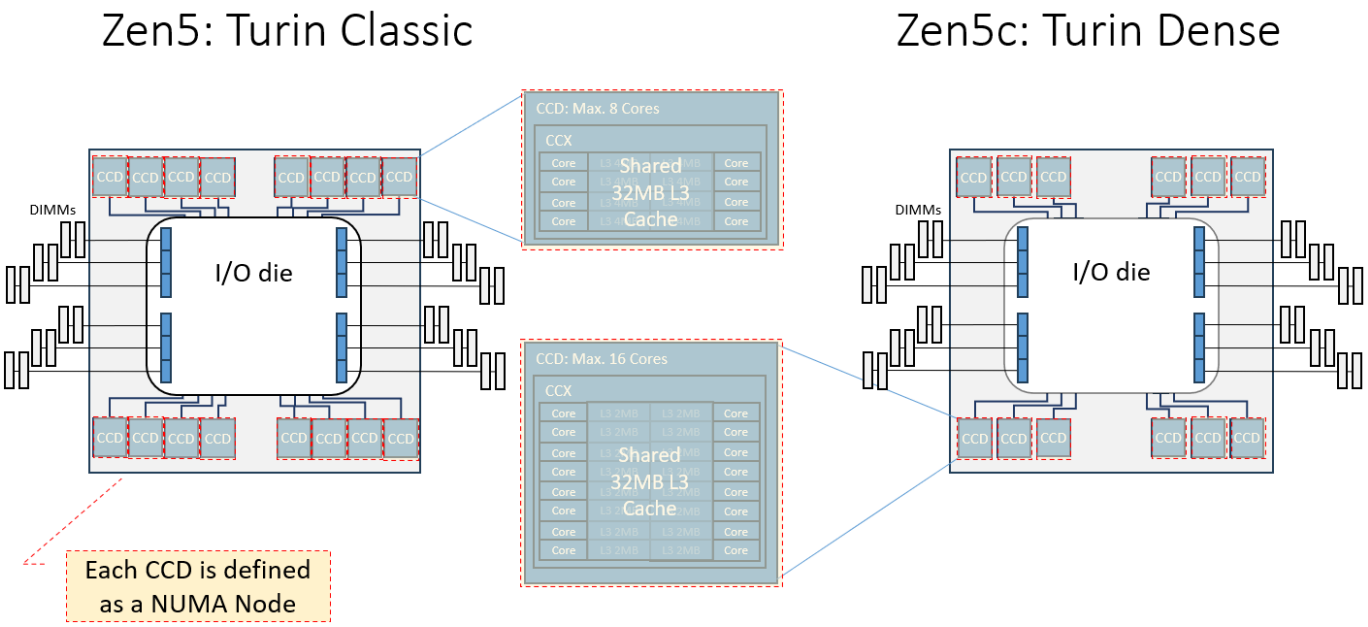
Configuration of NPS1 (Zen5 / Zen5c)

[ACPI SRAT L3 Cache As NUMA Domain] (=L3AsNumaNode) option can divide NUMA domains per CCX. For workloads that fall within the range of a L3 cache and the cores that share it, [Enabled] setting may increase the performance.



L3AsNumaNode:
- Number of Numa nodes = 12 (Zen4) or 8 (Zen4c) per socket
- Number of Memory Partitions = based on NPS setting
- BIOS will attempt to interleave all memory channels based on NPS setting

Configuration of L3AsNumaNode (Zen4 / Zen4c)



L3AsNumaNode:
- Number of Numa nodes = 16 (Zen5) or 12 (Zen5c) per socket
- Number of Memory Partitions = based on NPS setting
- BIOS will attempt to interleave all memory channels based on NPS setting

Configuration of L3AsNumaNode (Zen5 / Zen5c)

To use [NUMA nodes per socket] and [ACPI SRAT L3 Cache As NUMA Domain], memory must be equally populated in each memory controller, or the settings may not take effect. Moreover, if you change [CCD Control] option to limit the number of CCDs, the setting may not take effect. In addition, even if the settings are enabled, performance may degrade if the application is not NUMA optimized.

Appendix

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

For BIOS options not included in this white paper, refer to the “BIOS Setup Utility” manual for your specific model from the support pages listed in the Literature section.

Settings of Application Profile option (1/2)

Option Name	Default	Total Throughput Performance	Single Thread Performance	Energy efficient	Virtualization Performance	Low Latency
CPU Configuration						
SMT Control	Enabled	Enabled	Disabled	Enabled	Enabled	Disabled
L1 Stream HW Prefetcher	Enabled	Enabled	Enabled	Disabled (EPYC 9004 series) N/A (EPYC 9005 series)	N/A	Enabled
L1 Stride Prefetcher	Enabled	N/A	N/A	Disabled	N/A	N/A
L2 Stream HW Prefetcher	Enabled	Enabled	Enabled	Disabled	N/A	Enabled
SVM Mode	Enabled	N/A	N/A	Disabled	N/A	N/A
Core Performance Boost	Enabled	Enabled	Enabled	Enabled	Enabled	Enabled
Determinism Slider	Performance	Power	N/A	N/A	Power	N/A
TDP Control	Auto	Manual	N/A	Manual	Manual	Manual
TDP Limit	85	400	N/A	400	400	400
IOMMU	Enabled	N/A	N/A	N/A	Enabled	Disabled
Package Power Limit Control	Auto	Manual	N/A	N/A	Manual	Manual
Package Power Limit	85	400	N/A	400	400	400
APBDIS	0	N/A	1	N/A	N/A	1
DfPstate	0	N/A	0	N/A	N/A	0
DF Cstates	Enabled	N/A	Disabled	N/A	Disabled	Disabled
PCIE Speed PMM Control	Dynamic link speed determined by Power Management functionality	N/A	N/A	N/A	N/A	N/A
xGMI Link Width Control	Auto	N/A	N/A	N/A	N/A	Manual
xGMI Force Link Width Control	Force	N/A	N/A	N/A	N/A	Force
xGMI Force Link Width	x16	N/A	N/A	N/A	N/A	x16
xGMI Max Link Speed	32Gbps	N/A	N/A	20Gbps	N/A	N/A
ACPI CST C2 Latency	800	N/A	18 (EPYC 9004 series) N/A (EPYC 9005 series)	N/A	N/A	N/A
Memory Configuration						
Memory Clock	Auto	N/A	N/A	DDR4000 (EPYC 9004 series) DDR4800 (EPYC 9005 series)	N/A	N/A
Memory Interleaving	Auto	Auto	Auto	Auto	Auto	Auto
DRAM Scrub Time	24 hours	N/A	N/A	N/A	N/A	Disabled
Power Down Enable	Disabled	N/A	N/A	Enabled	N/A	N/A
Power Profile Selection	Efficiency Mode	High Performance	N/A	N/A	High Performance	N/A
NUMA nodes per socket	NPS1	NPS4	N/A	NPS2	NPS1	NPS4
ACPI SRAT L3 Cache As NUMA Domain	Disabled	N/A	N/A	Enabled	Disabled	N/A

Option Name	Default	Total Throughput Performance	Single Thread Performance	Energy efficient	Virtualization Performance	Low Latency
Memory Configuration (Cont.)						
Probe Filter Organization	Dedicated	Shared	N/A	N/A	N/A	N/A
Periodic Directory Rinse (PDR) Tuning	Auto	Cache-Bound (EPYC 9004 series) N/A (EPYC 9005 series)	Neutral (EPYC 9004 series) N/A (EPYC 9005 series)	N/A	N/A	N/A
PCI Subsystem Configuration						
SR-IOV Support	Enabled	N/A	N/A	N/A	Enabled	Disabled

Settings of Application Profile option (2/2)

Option Name	Default	Online Transaction Processing	Decision Supoprt	I/O Throughput	Memory Intensive HPC	CPU Intensive HPC
CPU Configuration						
SMT Control	Enabled	Enabled	Enabled	N/A	Disabled	Disabled
L1 Stream HW Prefetcher	Enabled	Enabled	Enabled	Enabled	N/A	Enabled
L1 Stride Prefetcher	Enabled	N/A	N/A	N/A	N/A	N/A
L2 Stream HW Prefetcher	Enabled	Enabled	Enabled	Enabled	N/A	Enabled
SVM Mode	Enabled	N/A	N/A	N/A	N/A	N/A
Core Performance Boost	Enabled	Enabled	N/A	N/A	N/A	Enabled
Determinism Slider	Performance	Power	Power	Power	N/A	Power
TDP Control	Auto	Manual	Manual	N/A	Manual	Manual
TDP Limit	85	400	400	N/A	400	400
IOMMU	Enabled	N/A	N/A	N/A	N/A	Disabled
Package Power Limit Control	Auto	Manual	Manual	N/A	Manual	Manual
Package Power Limit	85	400	400	N/A	400	400
APBDIS	0	1	1	1	1	N/A
DfPstate	0	0	0	0	0	N/A
DF Cstates	Enabled	N/A	N/A	N/A	N/A	N/A
PCIE Speed PMM Control	Dynamic link speed determined by Power Management functionality	N/A	N/A	Static Target Link Speed (GEN5)	N/A	N/A
xGMI Link Width Control	Auto	N/A	N/A	Manual	N/A	Manual
xGMI Force Link Width Control	Force	N/A	N/A	Force	N/A	Force
xGMI Force Link Width	x16	N/A	N/A	x16	N/A	x16
xGMI Max Link Speed	32Gbps	N/A	N/A	N/A	N/A	N/A
ACPI CST C2 Latency	800	N/A	N/A	N/A	N/A	N/A
Memory Configuration						
Memory Clock	Auto	N/A	N/A	N/A	N/A	N/A
Memory Interleaving	Auto	Auto	Auto	Auto	Auto	Auto
DRAM Scrub Time	24 hours	Disabled	N/A	N/A	N/A	N/A
Power Down Enable	Disabled	N/A	N/A	N/A	N/A	N/A
Power Profile Selection	Efficiency Mode	Maximum IO Performance Mode	N/A	N/A	N/A	High Performance Mode
NUMA nodes per socket	NPS1	NPS4	NPS4	NPS2	NPS4	NPS4
ACPI SRAT L3 Cache As NUMA Domain	Disabled	N/A	N/A	N/A	N/A	N/A
Probe Filter Organization	Dedicated	N/A	N/A	N/A	N/A	N/A
Periodic Directory Rinse (PDR) Tuning	Auto	N/A	N/A	N/A	N/A	N/A
PCI Subsystem Configuration						
SR-IOV Support	Enabled	Disabled	Disabled	Disabled	N/A	Disabled

Literature

PRIMERGY / PRIMEQUEST Servers

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

BIOS optimization for AMD EPYC 9004 and 9005 Processor-based systems

This Whitepaper

https://docs.ts.fujitsu.com/dl.aspx?id=24cacc7c-b128-4674-91d1-23bbc185bc89

https://docs.ts.fujitsu.com/dl.aspx?id=04aaa370-9b49-4b7d-8076-21215690885f

PRIMERGY Performance

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

PRIMERGY Manuals

Support Site:

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

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

- RX1440 M2 BIOS Setup Utility: "D4130 BIOS Setup Utility"
- RX2450 M2 BIOS Setup Utility: "D4129 BIOS Setup Utility"

Operating System Performance Tuning Guidelines

- Microsoft Windows:
  https://docs.microsoft.com/en-us/windows-server/administration/performance-tuning/

- RedHat Linux:
  https://access.redhat.com/documentation/en-us/red_hat_enterprise_linux/8/html/monitoring_and_managing_system_status_and_performance/index
  https://access.redhat.com/documentation/en-us/red_hat_enterprise_linux/9/html/monitoring_and_managing_system_status_and_performance/index

- SUSE Linux:
  https://documentation.suse.com/sles/15-SP4/html/SLES-all/book-tuning.html
  https://documentation.suse.com/sles/15-SP5/html/SLES-all/book-tuning.html
  https://documentation.suse.com/sles/15-SP6/html/SLES-all/book-tuning.html

- VMware vSphere:
  https://www.vmware.com/docs/vmw-tuning-latency-sensitive-workloads-white-paper
  https://www.vmware.com/docs/vsphere-esxi-vcenter-server-70u3-performance-best-practices
  https://www.vmware.com/docs/vsphere-esxi-vcenter-server-80u1-performance-best-practices
  https://www.vmware.com/docs/vsphere-esxi-vcenter-server-80U2-performance-best-practices
  https://www.vmware.com/docs/vsphere-esxi-vcenter-server-80U3-performance-best-practices

Document change history

Version	Date	Description
1.1	2025-01-14	Add items for the AMD EPYC 9005 series processors
1.0	2024-07-02	New

Contact
Fujitsu
Web site: https://www.fujitsu.com
PRIMERGY Performance and Benchmarks
mailto:fj-benchmark@dl.jp.fujitsu.com