AMD EPYC™ 9005

BIOS & Workload Tuning Guide

PID: 58467
v1.0
October 2024

Muhammad Ashfaq, Anil Rajput, and Jesse Rangel

© 2024 Advanced Micro Devices, Inc. All rights reserved.

The information contained herein is for informational purposes only and is subject to change without notice. While every precaution has been taken in the preparation of this document, it may contain technical inaccuracies, omissions and typographical errors, and AMD is under no obligation to update or otherwise correct this information. Advanced Micro Devices, Inc. makes no representations or warranties with respect to the accuracy or completeness of the contents of this document, and assumes no liability of any kind, including the implied warranties of noninfringement, merchantability or fitness for particular purposes, with respect to the operation or use of AMD hardware, software or other products described herein. No license, including implied or arising by estoppel, to any intellectual property rights is granted by this document. Terms and limitations applicable to the purchase or use of AMD's products are as set forth in a signed agreement between the parties or in AMD's Standard Terms and Conditions of Sale.

Trademarks

AMD, the AMD Arrow logo, AMD EPYC, and combinations thereof are trademarks of Advanced Micro Devices, Inc. Other product names and links to external sites used in this publication are for identification purposes only and may be trademarks of their respective companies.

* Links to third party sites are provided for convenience and unless explicitly stated, AMD is not responsible for the contents of such linked sites and no endorsement is implied.

DATE	VERSION	CHANGES
June, 2024	0.1	Initial NDA release
June, 2024	0.2	Errata fixes, additional tuning information in Chapter 4.
October, 2024	1.0	Initial public release
April, 2025	2.0	Errata fixes, Page 5 SEV Control Disable to Enable

AUDIENCE

This document describes best practices for optimizing performance using AMD default BIOS settings. It is intended for a technical audience such as system architects, production deployment, and performance engineering teams with:

- A background in configuring servers.
- Administrator-level access to both the server management Interface (BMC) and the OS.
- Familiarity with both the BMC and OS-specific configuration, monitoring, and troubleshooting tools.

BIOS & WORKLOAD TUNING GUIDE

CONTENTS

Chapter 1 - Introduction ---------- 1

1.1 - Important Reading 1

Chapter 2 - BIOS Defaults Summary ---------- 3

2.1 - Processor Core Settings 4
2.2 - Power Efficiency Settings 5
2.3 - NUMA and Memory Settings 6
2.4 - Infinity Fabric Settings 7
2.5 - PCIe, I/O, Security, and Virtualization Settings 8
2.6 - Higher-Level Settings 9

Chapter 3 - BIOS Option Details ---------- 11

3.1 - Processor Core Settings 11
3.1.1 - Simultaneous Multithreading (SMT) Settings 11
3.1.2 - Cache Prefetchers 12
3.1.3 - Core Performance Boost 13
3.1.4 - Global C-States Control 13
3.2 - Power Management Settings 14
3.2.1 - Power Profile Selection 14
3.2.2 - Power vs. Performance Determinism Settings 14
3.2.3 - Processor Cooling and Power Dissipation Limit Settings 15
3.2.4 - ACPI—Collaborative Processor Performance Control (CPPC) 15
3.3 - NUMA and Memory Settings 16
3.3.1 - L3 Cache as NUMA Domain 16
3.3.2 - NUMA Nodes per Socket (NPS) 16
3.3.3 - Memory Target Speed 17
3.3.4 - Memory Interleaving 17
3.4 - Infinity Fabric Settings 18
3.4.1 - Link Speed 18
3.4.2 - xGMI Link Width Management 18
3.4.3 - Power States 19
3.4.4 - DF C-States 19

3.5 - PCIe, I/O, Security, and Virtualization Settings 20
3.5.1 - APIC Settings 20
3.5.2 - SR-IOV Settings 20
3.5.3 - PCIe Ten Bit Tag 20
3.5.4 - Input-Output Memory Management Unit (IOMMU) Settings 21
3.5.5 - Transparent Secure Memory Encryption (TSME) 21
3.5.6 - SEV, SEV-ES, and SEV-SNP 21
3.5.7 - AVIC& x2AVIC 22
3.6 - Options Eliminated in AMD EPYC 9005 Series Processors 22

Chapter 4 - Workload-Specific BIOS Settings 23

4.1 - General-Purpose Workloads 23
4.1.1 - Processor Core Settings 23
4.1.2 - Power Management Settings 24
4.1.3 - NUMA and Memory Settings 24
4.1.4 - Infinity Fabric Settings 24
4.1.5 - I/O Settings 25
4.2 - Memory and I/O Intensive Workloads 25
4.2.1 - Processor Core Settings 25
4.2.2 - Power Management Settings 26
4.2.3 - NUMA and Memory Settings 26
4.2.4 - Infinity Fabric Settings 26
4.2.5 - I/O Settings 27
4.3 - Virtualization and Containers 27
4.3.1 - Processor Core Settings 27
4.3.2 - Power Management Settings 28
4.3.3 - NUMA and Memory Settings 28
4.3.4 - Infinity Fabric Settings 28
4.3.5 - I/O Settings 29
4.4 - Database, Analytics and AI/ML 29
4.4.1 - Processor Core Settings 29
4.4.2 - Power Management Settings 30
4.4.3 - NUMA and Memory Settings 30
4.4.4 - Infinity Fabric Settings 31
4.4.5 - I/O Settings 31
4.5 - HPC, Telco, and EDA Settings 32
4.5.1 - Processor Core Settings 32
4.5.2 - Power Management Settings 32
4.5.3 - NUMA and Memory Settings 33
4.5.4 - Infinity Fabric Settings 33
4.5.5 - I/O Settings 34

Chapter 5 - Debugging BIOS Setting Updates 35

Chapter 1: Introduction

Default BIOS options generally produce the best overall performance for generic workloads, but these defaults may not be optimal for a specific workload. AMD continually tests various workloads; this tuning guide discusses BIOS options to deliver both maximum performance and performance-per- watt (power efficiency).

- "BIOS Defaults Summary" on page 3 provides a quick overview of default AMD EPYC 9005 BIOS settings.
- "BIOS Option Details" on page 11 provides detailed information about the AMD EPYC 9005 BIOS options and the potential benefit of modifying each one.
- "Workload-Specific BIOS Settings" on page 23 presents sample workloads and recommended BIOS settings. Keep in mind that these BIOS settings are not "one size fits all" because your specific workload(s) are not identical to synthetic benchmarks.

Note: Not all platforms support all of the BIOS settings described in this Tuning Guide. Please contact your platform vendor if you cannot see one or more needed settings.

Note: Not all operating systems support all of the BIOS settings described in this Tuning Guide. Verify that you have a thorough understanding of your platform BIOS settings and corresponding OS support before making any changes to your system.

AMD strongly recommends that customers download and install the latest BIOS update for your AMD EPYC 9005 Series Processor-based server from your platform vendor. BIOS updates often help customers by providing new and updated features, bug fixes, enhancements, security features, and other improvements. These improvements can help your system software stability and dependency modules (such as hardware, firmware, drivers, and software) by giving you a more robust environment to run your applications.

1.1 - Important Reading

Please be sure to read the following guides (available from the AMD Documentation Hub), which contain important foundational information about 5th Gen AMD EPYC processors:

- *AMD EPYC™ 9005 Processor Architecture Overview*
- *Memory Population Guidelines for AMD EPYC™ 9005 Series Processors*

THIS PAGE INTENTIONALLY LEFT BLANK.

Chapter 2: BIOS Defaults Summary

This chapter provides high-level lists of the default AMD EPYC 9005 BIOS settings and their default values. Please see "BIOS Option Details" on page 11 for detailed descriptions. Later chapters in this Tuning Guide discuss the BIOS options as they relate to a specific workload or set of workload(s).

Note: The default setting name and values described in this chapter are the AMD default names and values that serve as recommendations for OEMs. End users must confirm their OEM BIOS setting availability and options.

2.1 - Processor Core Settings

Name	Default	Description
SMT Control	Auto	• **Enabled/Auto:** Two hardware threads per core. • **Disabled:** Single hardware thread per core.
L1 Stream HW Prefetcher	Auto	• **Enabled/Auto:** Enables the prefetcher. • **Disabled:** Disables the prefetcher.
L1 Stride Prefetcher	Auto	• **Enabled/Auto:** Enables the prefetcher. • **Disabled:** Disables the prefetcher.
L1 Region Prefetcher	Auto	• **Enabled/Auto:** Enables the prefetcher. • **Disabled:** Disables the prefetcher.
L1 Burst Prefetch Mode	Auto	• **Enabled/Auto:** Enables the prefetcher. • **Disabled:** Disables the prefetcher.
L2 Stream HW Prefetcher	Auto	• **Enabled/Auto:** Enables the prefetcher. • **Disabled:** Disables the prefetcher.
L2 Up/Down Prefetcher	Auto	• **Enabled/Auto:** Enables the prefetcher. • **Disabled:** Disables the prefetcher.
Core Performance Boost	Auto	• **Enabled/Auto:** Enables Core Performance Boost. • **Disabled:** Disables Core Performance Boost.
BoostFmaxEn	Auto	• **Auto:** Use the default Fmax • **Manual:** User can set the boost Fmax
BoostFmax	Auto	Specify the boost Fmax frequency limit to apply to all cores (MHz in decimal)
Global C-State Control	Auto	• **Enabled/Auto:** Controls IO based C-state generation and DF C-states, including core processor C-States • **Disabled:** AMD strongly recommends not disabling this option because this also disables core processor C-States.

Table 2-1: Processor core BIOS settings

2.2 - Power Efficiency Settings

Name	Default	Description
Power Profile Selection	Auto	• **Auto/0:** High-performance mode • **1:** Efficiency mode • **2:** Maximum I/O performance mode • **3:** Balanced Memory Performance Mode • **4:** Balanced Core Performance Mode • **5:** Balanced Core Memory Performance Mode
Determinism Control	Auto	• **Auto:** Use default performance determinism settings. • **Manual:** Specify custom performance determinism settings.
Determinism Enable	Auto	• **Auto/0:** Power. • **1:** Performance.
TDP Control	Auto	• **Auto:** Use platform- and OPN-default TDP. • **Manual:** Set custom configurable TDP.
TDP	OPN Max	This option appears once the user sets the **TDP Control** to **Manual**. • **Values 100-500:** Set configurable TDP, in watts.
PPT Control	Auto	Enables or disables the **PPT** control. • **Auto:** Automatically set PPL in watts. • **Manual:** Specify a custom PPL.
PPT	OPN Max	This option appears once the user sets the **PPT Control** to **Manual**. • **Values 100-500:** Set configurable PPT, in watts.
CPPC	Auto	• **Enabled/Auto:** Allows the OS to make performance/power optimization requests using ACPI CPPC. • **Disabled:** Prevents the OS from making performance/power optimization requests using ACPI CPPC.

Table 2-2: Power efficiency BIOS settings

2.3 - NUMA and Memory Settings

Name	Default	Description
LLC as NUMA Domain (ACPI SRAT L3 Cache as NUMA Domain)	Disabled	• **Disabled (recommended):** Both NUMA nodes (`cpubind`) and memory interleaving (`membind`) are determined by the NPS setting. • **Enabled:** Overrides the NPS setting for # of NUMA nodes by mapping each LLC as a NUMA node. This does not impact the memory interleaving
Nodes Per Socket (NPS)	1	**Memory Interleaving:** The **NPS** setting always determines the memory interleaving regardless of whether **LLC as NUMA** is **Enabled** or **Disabled**. # of NUMA nodes (if **LLC as NUMA Domain** is **Disabled**): • **NPS1:** One NUMA node per socket (Most cloud providers use this as it provides consistent average memory latency to all the accesses within a socket). • **NPS2:** Two NUMA nodes per socket. • **NPS4:** Four NUMA nodes per socket • **NPS0 (not recommended):** Only applicable for dual-socket systems. A single NUMA node is created for the whole two-socket platform. AMD recommends either NPS1 or NPS4 depending on your use case. **Windows systems:** Make sure that the number of logical processors per NUMA node is <=64. You can do this by using NPS2 or NPS4 instead of the default NPS1.
Memory Target Speed	Auto	• **Auto:** Determine the maximum memory speed based on SPD information from populated DIMMs and platform memory speed support. Alternatively, you can select: • **Values 3200–6400 MT/s:** Run the DRAM memory target clock speed at the specified speed. The DRAM memory target is the DDR rate. Your OEM system default value may vary.
Memory Interleaving	Auto	• **Auto/Enable:** Enables memory interleaving. • **Disable:** Allows for disabling memory interleaving. The N**UMA Nodes per Socket** setting will be honored regardless of this setting. AMD strongly recommends not disabling this setting because most production deployments benefit from memory interleaving.
Periodic Directory Rinse (PDR) Tuning	Auto	Control PDR settings that may impact performance by workload and/or processor. • **Auto/Blended (Cache Load Based Floss with Background RefClock Based Floss):** Demand based Directory Rinse. • **Periodic (RefClock Based Floss Only):** Rate based Directory Rinse.

Table 2-3: NUMA and memory BIOS settings

2.4 - Infinity Fabric Settings

Name	Default	Description
3-4 xGMI Link Max Speed	Auto	• 20 Gbps • 25 Gbps • 32 Gbps/Auto Your OEM system default value may vary.
xGMI Link Width Control	Auto	• **Auto:** Use the default xGMI link width controller settings. • **Manual:** Specify a custom xGMI link width controller setting.
xGMI Force Link Width Control	Auto	• **Unforce:** Do not force the xGMI to a fixed width. • **Force:** Use the xGMI link to the user-specified width.
xGMI Force Link Width	Auto	• **0:** Force xGMI link width to x4. • **1:** Force xGMI link width to x8. • **2:** Force xGMI link width to x16.
xGMI Max Link Width Control	Auto	• **Auto:** Use the default xGMI link width controller settings. • **Manual:** Specify a custom xGMI link with controller setting.
xGMI Max Link Width	Auto	• **0:** Set max xGMI link width to x4. • **1:** Set max xGMI link width to x8. • **2:** Set max xGMI link width to x16.
APBDIS	Auto	• **0/Auto:** Dynamically switch the Infinity Fabric P-state based on link usage. • **1:** Enabled fixed Infinity Fabric P-state control.
DfPstate Range Support	Auto	• **Auto:** If this feature is enabled, the range value setting should follow the rule that MaxDfPstate<=MinDfPstate. Otherwise, it will not work. • **Enable:** Add the values MaxDfPstate & MinDfPstate. • **Disable:** No MaxDfPstate & MinDfPstate option.
DF C-States	Auto	Controls DF C-states. • **Disabled:** Prevents the AMD Infinity Fabric from entering a low-power state. • **Enabled/Auto:** Allows the AMD Infinity Fabric to enter a low-power state.

Table 2-4: Infinity Fabric BIOS settings

2.5 - PCIe, I/O, Security, and Virtualization Settings

Name	Default	Description
Local APIC Mode	Auto (0x02 x2APIC)	• **xAPIC:** Use xAPIC, supports up to 255 cores. • **x2APIC:** Supports more than 255 cores. • **Auto:** The system will choose the mode that best fits the number of active cores in the system.
PCIe ARI Support (SRIOV)	Auto	• **Enabled/Auto:** Enables Alternative Routing ID interpretation. • **Disabled:** Disables Alternative Routing ID interpretation.
PCIe Ten Bit Tag Support	Auto	• **Enabled/Auto:** Enables PCIe 10-bit tags for supported devices. • **Disabled:** Disables PCIe 10-bit tags for all devices.
IOMMU	Auto	• **Enabled/Auto:** Enables IOMMU. AMD recommends setting this to `pt:pass-through` in the Linux kernel settings. • **Disabled:** Disables IOMMU.
AVIC	Disabled	Advanced Virtual Interrupt Controller. • **Disabled:** Disables AVIC. • **Enabled:** Enables AVIC.
x2AVIC	Disabled	x2AVIC is an extension of the advanced virtual interrupt controller. This feature currently requires a custom AMD Linux kernel. • **Disabled:** Disables x2AVIC. • **Enabled:** Enables x2AVIC.
TSME	Auto	• **Auto/Disabled:** Disables transparent secure memory encryption. • **Enabled:** Enables transparent secure memory encryption.
SEV Control	Enabled	In a multi-tenant environment (such as a cloud), Secure Encrypted Virtualization (SEV) mode isolates virtual machines from each other and from the hypervisor. • **Disabled:** SEV is disabled. • **Enabled:** SEV is enabled. If you disable and then reenable SEV, then you will need to power cycle your system after changing this setting back to **Enabled**.
SEV-ES	Disabled	Secure Encrypted Virtualization-Encrypted State (SEV-ES) mode extends SEV protection to the contents of the CPU registers by encrypting them when a virtual machine stops running. Combining SEV and SEV-ES can reduce the attack surface of a VM by helping protect the confidentiality of data in memory. • **Disabled:** SEV-ES is disabled. • **Enabled:** SEV-ES is enabled.
SEV-SNP Support	Disabled	Secure Encrypted Virtualization-Secure Nested Paging (SEV-SNP) mode builds on SEV and SEV-ES by adding strong memory integrity protection to create an isolated execution environment that helps prevent malicious hypervisor-based attacks such as data replay and memory re-mapping. SEV-SNP also introduces several additional optional security enhancements that support additional VM use models, offer stronger protection around interrupt behavior, and increase protection against recently-disclosed side channel attacks. • **Disabled:** SEV-SNP is disabled. • **Enabled:** SEV-SNP is enabled.

Table 2-5: PCIe, I/O, security, and virtualization BIOS settings

2.6 - Higher-Level Settings

The system powers on to an initial state, after which succeeding software layers may affect system settings:

1. System firmware validates basic hardware functionality and allows users to change various settings via the BIOS Setup menus.
2. UEFI provides a shell environment that allows users to further interact with the system.
3. The operating system or hypervisor is the next software layer that provides control over system hardware.
4. Lastly, certain applications can also affect underlying hardware.

Each of the lines above may alter settings made by prior line, and some user changes require a reboot to take effect.

Please refer to your OEM documentation and/or applicable AMD Tuning Guide(s) for further guidance.

THIS PAGE INTENTIONALLY LEFT BLANK.

Chapter 3: BIOS Option Details

3.1 - Processor Core Settings

3.1.1 - Simultaneous Multithreading (SMT) Settings

Enabling SMT causes neutral to negative performance impacts on some workloads, especially HPC. Also, some application licenses count the number of hardware threads enabled instead of the physical core count. It may therefore be best to disable SMT on your AMD EPYC 9005 Series Processor.

Setting	Default	Options
SMT Control	Auto	• **Enable/Auto:** Two hardware threads per core. • **Disable:** Single hardware thread per core.

Table 3-1: SMT settings

3.1.2 - Cache Prefetchers

Most workloads and production deployments benefit from the L1 & L2 Stream Hardware prefetchers gathering data and keeping the core pipeline busy, but some workloads that stress the memory bandwidth to its maximum capacity may perform better when some or all prefetchers are disabled. All prefetchers are enabled by default. Be sure to evaluate the prefetchers for your deployments.

Setting	Default	Options
L1 Stream HW Prefetcher	Auto	This prefetcher uses the history of L1 cache memory access patterns to fetch additional sequential lines in ascending or descending order. • **Disable:** Disable prefetcher. • **Enable:** Enable prefetcher.
L1 Stride Prefetcher	Auto	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. • **Disable:** Disable prefetcher. • **Enable:** Enable prefetcher.
L1 Region Prefetcher	Auto	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. • **Disable:** Disable prefetcher. • **Enable:** Enable prefetcher.
L1 Burst Prefetch Mode	Auto	• **Disable:** Disable burst mode. • **Enable:** Enable burst mode.
L2 Stream HW Prefetcher	Auto	This prefetcher uses the history of L2 cache memory access patterns to fetch additional sequential lines in ascending or descending order. • **Disable:** Disable prefetcher. • **Enable:** Enable prefetcher.
L2 Up/Down Prefetcher	Auto	Uses the L2 cache memory access history to determine whether to fetch the next or previous line for all memory accesses. • **Disable:** Disable prefetcher. • **Enable:** Enable prefetcher.

Table 3-2: Cache prefetcher settings

3.1.3 - Core Performance Boost

Core Performance Boost can be enabled or disabled. Enabling this setting allows the processor to opportunistically increase a set of CPU cores to higher than the CPU's rated base clock speeds based on the number of active cores, power, and thermal headroom in a system.

Some workloads don't need maximum core frequency to achieve acceptable performance. Limiting the maximum core boost frequency can reduce power consumption. The BoostFmax setting limits the maximum boost frequency but does not set a fixed frequency. The SoC will not exceed the maximum algorithm-allowable frequency if BoostFmax is set too high. Actual boost performance depends on many factors, including the other settings discussed in this tuning guide.

Setting	Default	Options
Core Performance Boost	Auto	• **Enable/Auto:** Enables Core Performance Boost. • **Disable:** Disables Core Performance Boost.
BoostFmaxEn	Auto	• **Manual:** Use specified BoostFmax setting. • **Auto:** Use default BoostFmax setting.
BoostFmax	Auto	Specify the boost Fmax frequency limit to apply to all cores (MHz in decimal).

Table 3-3: Core boost settings

3.1.4 - Global C-States Control

The **Global C-States Control** enables and disables C-states on the server across all cores. Disabling this feature means that the CPU cores can only be in C0 (active) or C1 state because the C1 state cannot be disabled. A CPU core will be in C1 state if the core is halted by the OS. IO based C-state generation and DF C-states include core processor C-States. If you have a low latency or extremely low jitter use case, then consider disabling DF C-states as described in this Tuning Guide. AMD strongly recommends not disabling Global C-states except for debugging.

Setting	Default	Options
Global C-States Control	Auto	• **Enabled/Auto:** Controls IO based C-state generation and DF C-states, including core processor C-States • **Disabled:** AMD strongly recommends not disabling this option because this also disables core processor C-States.

Table 3-4: Global C-states control

3.2 - Power Management Settings

3.2.1 - Power Profile Selection

Setting	Default	Options
Power Profile Selection Control	Auto	• **0:** High Performance mode (Default) • **1:** Efficiency Mode • **2:** Maximum IO Performance Mode • **3:** Balanced Memory Performance Mode • **4:** Balanced Core Performance Mode • **5:** Balanced Core Memory Performance Mode

Table 3-5: Power profile selection

3.2.2 - Power vs. Performance Determinism Settings

The **Determinism Enable** selects between:

- **Performance (default for most OPNs):** Uniform performance across identically configured systems in a datacenter. Set TDP and PPL to the same value, as described in "Processor Cooling and Power Dissipation Limit Settings" on page 15.
- **Power:** Maximum performance of any individual system by leveraging the capabilities of a given CPU to the maximum, resulting in a varying performance range across the datacenter or larger deployments.

Setting	Default	Options
Determinism Control	Auto	• **Auto:** Use default performance determinism settings. • **Manual:** Specify custom power/performance determinism.
Determinism Enable	Auto	• **Auto:** This setting may be either **Power** or **Performance** based on OEM Platform and OPN selection. • **0:** Power (default) • **1:** Performance See "Processor Cooling and Power Dissipation Limit Settings" on page 15 for additional information.

Table 3-6: Power/performance settings

3.2.3 - Processor Cooling and Power Dissipation Limit Settings

The Thermal Design Power (TDP) of an OPN represents the power consumed by all processor components to operate for a given performance target. This performance target may be reached well below the TDP, which case the processor will have headroom to operate above the performance target. This headroom varies from part to part. You can tap into the additional performance above the target performance of a OPN via the BIOS Determinism setting.

- Setting Determinism to Performance makes the workload performance is the same regardless of variations in the environment or silicon.
- Setting Determinism to Power maximizes workload performance to part-specific power limits, thereby tapping the additional performance headroom based on the silicon. If you are using Power determinism, then you can obtain maximum performance by setting the TDP and Package Power Limit (PPL) to the maximum TDP value supported by the CPU. Setting the TDP and PPL to Auto sets both parameters to the CPU default TDP value for energy-efficient operation

The Package Power Tracker (PPT) controls the package power limit per processor that is applicable for all processors populated in the system. Please see Power/Performance Determinism for additional information.

Setting	Default	Options
TDP Control	Auto	• **Auto:** Use platform- and OPN-default TDP. • **Manual:** Set custom configurable TDP.
TDP	OPN Max	This option is available if the user sets the **TDP Control** to **Manual**. • **Values 100–500:** Set configurable TDP, in watts.
PPT Control	Auto	Enables or disables the **PPT** control. • **Auto:** Use platform- and OPN-default-PPL. • **Manual:** Set customized PPL.
PPT	OPN Max	This option is available if the user sets the **PPT Control** to **Manual**. • **Values 100–500:** Set PPT, in watts.

Table 3-7: TDP settings

3.2.4 - ACPI—Collaborative Processor Performance Control (CPPC)

Enabling CPCC allows the OS to help maintain energy efficiency by controlling when and how much turbo can be applied. ACPI 5.0 introduced this feature. Not all operating systems support CPCC. Microsoft began supporting CPCC with Windows® Server® 2016.

Setting	Default	Options
CPCC	Auto	• **Disabled:** Disabled. • **Enabled:** Allow the OS to make performance/power optimization requests using ACPI CPPC.

Table 3-8: CPCC settings

3.3 - NUMA and Memory Settings

This section describes NUMA- and memory-related BIOS settings.

3.3.1 - L3 Cache as NUMA Domain

This setting controls automatic or manual generation of distance information in the ACPI System Locality Information Table (SLIT) and NUMA proximity domains in the System Resource Affinity Table (SRAT). Some hypervisors and operating systems do not perform L3-aware scheduling, and some workloads will benefit from having the L3 declared as a NUMA domain. In dual-socket systems, the remote socket distance can affect memory allocation decisions. Setting this to a value of at least 32 (32 recommended) may improve scheduling of lightly-threaded workloads. Setting this to a value less than 32 (22 recommended) may improve scheduling of heavily-threaded workloads. In general:

- If a workload spans two sockets, then set the distance to < 32.
- If the workload can be confined to a socket, then set the distance to 32.

3.3.2 - NUMA Nodes per Socket (NPS)

This setting enables a trade-off between minimizing local memory latency for NUMA-aware or highly parallelizable workloads vs. maximizing per-core memory bandwidth for non-NUMA-friendly workloads. NPS2 and/or NPS4 may not be an option on certain OPNs or with certain memory populations.

- **NPS1:** Indicates a single NUMA node per socket. This setting configures all memory channels on the processor into a single NUMA domain. All of the processor cores, all attached memory, and all PCIe devices connected to the SoC are in that one NUMA domain. Memory accesses are interleaved across all 24 memory channels into a single address space. The default configuration (one NUMA domain per socket) is recommended for most workloads.
- **NPS2:** 2 NUMA domains per socket, which interleaves the corresponding six memory channels within the same 6 CCD NUMA domain. Half of the cores and half of the memory channels of the SoC are grouped together into one NUMA domain, with the remaining cores and memory channels grouped into the second NUMA domain. Memory is interleaved across the six memory channels of each NUMA domain.
- **NPS4:** 4 partitions the processor into four NUMA domains with each logical quadrant configured as its own NUMA domain. Memory is interleaved across the three memory channels within each quadrant. PCIe devices will be local to one of the four NUMA domains depending on the quadrant (of the I/O die) that has the PCIe root complex for that device. Every pair of memory channels is interleaved. This is recommended for HPC and other highly-parallel workloads. You must use NPS4 when booting Windows systems with SMT enabled for AMD EPYC processors with more than 64 cores because Windows limits the size of a CPU group to a maximum of 64 logical cores.

Note: For Windows systems, verify that the number of logical processors per NUMA node <=64 by using either NPS2 or NPS4 instead of the default NPS1.

Enabling **ACPI SRAT L3 Cache as NUMA Domain** (another name for **L3 as NUMA**) determines the number of NUMA nodes and overrides the number of NUMA nodes specified by the NPS setting while still using the NPS setting to determine the memory interleaving granularity.

Setting	Default	Options
L3 Cache as NUMA Domain (ACPI SRAT L3 Cache as NUMA Domain)	DIsabled	• **Disabled/Auto:** Do not report each L3 cache to the OS as a NUMA domain. • **Enabled:** Report each L3 cache to the OS as a NUMA domain.
NUMA Node Per Socket	1	• **NPS0:** Interleave memory accesses across all channels in both sockets (not recommended). • **NPS1:** Interleaves memory accesses across all channels in each socket and report one NUMA node per socket unless **L3 Cache as NUMA** is enabled. • **NPS2:** Interleaves memory accesses the channels associated with each half of a socket and reports two NUMA nodes per socket unless **L3 Cache as NUMA** is enabled. • **NPS4:** Interleaves memory accesses across the channels associated with a quadrant of each socket and reports four NUMA nodes per socket unless **L3 Cache as NUMA** is enabled.

Table 3-9: NPS settings

3.3.3 - Memory Target Speed

By default, the 5th Gen AMD EPYC processor BIOS runs at the maximum target frequency allowed by the platform and DIMM. This configuration allows maximum memory bandwidth and lowest latency for the processor. Lowering the memory clock speed reduces memory controller power consumption and allows the rest of the SoC to consume more power, thereby potentially boosting performance elsewhere for certain workloads.

Setting	Default	Options
Memory Target Speed	Auto	• **Auto:** Determine maximum memory speed based on SPD information from populated DIMMs and platform memory speed support. • **Values 3200–6400 MT/s:** Run the DRAM memory target speed at the specified speed (the DRAM memory target is the DDR rate.)

Table 3-10: Memory clock settings

3.3.4 - Memory Interleaving

This setting allows you to enable or disable memory interleaving within a NUMA node. The NUMA Nodes per Socket (NPS) setting will be honored regardless of this setting. This BIOS setting does not impact the number of NUMA nodes or how memory channels are mapped to the NUMA nodes.

Note: AMD strongly recommends not disabling this setting because most applications and deployments benefit from memory interleaving.

Setting	Default	Options
Memory Interleaving	Auto	• **Auto/Enable:** Enables memory interleaving. • **Disable:** Allows for disabling memory interleaving. The NUMA Nodes per Socket (NPS) setting will be honored regardless of this setting.

Table 3-11: Memory interleaving settings

3.4 - Infinity Fabric Settings

This section discusses BIOS settings related to AMD Infinity Fabric technology.

3.4.1 - Link Speed

Lowering the link speed decreases cross-socket bandwidth and increases cross-socket latency but can also save uncore power (CPU power not consumed by the cores) to either:

- Increase core frequency.
- Reduce overall power consumption.

Setting	Default	Options
3-4 Link xGMI Max Speed	Auto	• **32 Gbps/Auto** • Additional options depend on the OEM platform: 20, 25, 32 Gbps

Table 3-12: Link speed settings

3.4.2 - xGMI Link Width Management

xGMI Dynamic Link Width Management saves power during periods of low socket-to-socket data traffic by reducing the number of active xGMI lanes per link from 16 to 8, or x4 if the xGMI links have limited traffic. Latency may increase in some scenarios involving low-bandwidth, latency-sensitive traffic as the processor transitions from a low-power xGMI state to full-power xGMI state. Setting **xGMI Link Width Control** to **Manual** and specifying a **Force Link Width** eliminates any such latency jitter. Applications that are not sensitive to both socket-to-socket bandwidth and latency can use a forced link width of 8 (or 2 on certain platforms) to save power, which can divert more power to the cores for boost.

Setting	Default	Options
xGMI Link Max Speed	Auto	NUMA-unaware workloads may need maximum xGMI bandwidth because of extensive cross-socket communications. NUMA-aware workloads may want to minimize xGMI power because they do not have a lot of cross socket traffic and prefer to use the increased CPU boost. • 32 Gbps/Auto • 25 Gbps • 20 Gbps
xGMI Link Width Control	Auto	• **Auto:** Hide the **Max Link Width** and **Force Link Width** control options. • **Manual:** Show **Max Link Width** and **Force Link Width** control options.
xGMI Max Link Width	Auto	• **0:** Max width x8, min width x8 (x4 on certain platforms). • **1:** Max width x16, min width x8 (x4 on certain platforms).
xGMI Max Link Width Control	Auto	• **Auto:** Hide the **xGMI Max Link Width** control. • **Manual:** Show the **xGMI Max Link Width** control.
xGMI Force Link Width Control	Auto	• **Unforce:** Use automatic xGMI Link Width selection. • **Force:** Use the xGMI Force Link Width link width.
xGMI Force Link Width	Auto	• **0:** Use width x4 (not supported on all platforms). • **1:** Use width x8.

Table 3-13: DLWM settings

3.4.3 - Power States

Enable or disable Algorithm Performance Boost (APB). By default, the AMD Infinity Fabric selects between a full- and low-power fabric clock and memory clock based on usage. Latency may increase in some scenarios involving low-bandwidth, latency-sensitive traffic as the processor transitions from low to full power. Setting **APBDIS** to 1 (APB disabled) and specifying a fixed Infinity Fabric P-state of 0 forces the AMD Infinity Fabric and memory controllers into full-power mode and significantly reduces latency jitters.

Setting	Default	Options
APB Disable (APBDIS)	Auto	• **0:** Dynamically switch Infinity Fabric P-state based on link usage. • **1:** Enable fixed Infinity Fabric P-state control.
DfPstate	Auto	DfPstate index to set below when APBDIS [1]. • **Min Value:** 0 (default); highest-performing AMD Infinity Fabric P-state. • **Range Value:** 0-4
DfPstate Range Support	Auto	DF Pstate selection is overridden by the APB_DIS BIOS option if it is selected. If this feature is enabled, then range value setting should follow the rule that MaxDfP-state<=MinDfPstate. Otherwise, it will not work.

Table 3-14: Power state settings

3.4.4 - DF C-States

Much like CPU cores, the AMD Infinity Fabric can enter lower-power states while idle, but a delay occurs when transitioning back to full-power mode that causes some latency jitter. Disabling this feature for workloads requiring low latency and/or bursty I/O will increase both performance and power consumption.

Setting	Default	Options
DF C-states	Auto	• **Auto/Enabled:** Allow the AMD Infinity Fabric to enter a low-power state. • **Disabled:** Prevent the AMD Infinity Fabric from entering a low-power state.

Table 3-15: C-state settings

3.5 - PCIe, I/O, Security, and Virtualization Settings

3.5.1 - APIC Settings

Interrupt delivery is generally faster when using x2APIC compared to the legacy xAPIC mode, but not all operating systems include AMD x2APIC support. AMD recommends this mode if your OS supports it, including for configurations with fewer than 256 logical processors.

Setting	Default	Options
Local APIC Mode	Auto (0x02 x2APIC)	• **APIC:** Use xAPIC, which supports up to 255 cores. • **x2APIC:** Supports more than 255 cores. • **Auto:** The system will choose the mode that best fits the number of active cores in the system. • **Compatibility:** Threads below 255 run in xAPIC with xAPIC ACPI structures, and threads 255 and above run in x2 mode with x2 ACPI structures. • **XApicMode (0x01):** Force legacy xApic mode • **X2ApicMode (0x02):** F force x2Apic mode independent of thread count.

Table 3-16: APIC settings

3.5.2 - SR-IOV Settings

SR-IOV requires enabling PCIe Alternative Routing-ID interpretation (ARI) on both root complexes and endpoints. ARI devices interpret the PCI address as an 8-bit function number instead of a 3-bit function number, and the device number is implied to be 0.

Setting	Default	Options
PCIe ARI Support [SRIOV]	Auto	• **Disable:** Disable Alternative Routing ID interpretation. • **Enable:** Enable Alternative Routing ID interpretation.

Table 3-17: SR-IOV settings

3.5.3 - PCIe Ten Bit Tag

A PCIe adapter must support 10-bit extended tags to achieve maximum PCIe Gen 5 bandwidth. This boosts adapter performance by allowing a 3x increase over the previous number of non-posted requests. Not all PCIe Gen 5 devices support 10-bit extended tags, which can cause issues during boot. Disabling this feature allows the server to boot if the adapter is having issues.

Setting	Default	Options
PCIe Ten Bit Tag Support	Auto	• **Auto/Enable:** Enable PCIe 10-bit tags for supported devices. • **Disable:** Disable PCIe 10-bit tags for all devices.

Table 3-18: PCIe 10-bit settings

3.5.4 - Input-Output Memory Management Unit (IOMMU) Settings

Enabling the IOMMU allows devices such as the AMD EPYC processor-integrated SATA controller to present separate IRQs for each attached device instead of one IRQ for the subsystem. The IOMMU also allows operating systems to provide additional protection for DMA capable I/O devices. If you believe the IOMMU is limiting performance, then leave it enabled in BIOS and disable it via OS options (e.g., `iommu=pt` on the Linux® kernel command line). Enabling IOMMU is required when using x2APIC.

Setting	Default	Options
IOMMU	Auto	• **Enabled:** Enable IOMMU. • **Disabled:** Disable IOMMU.

Table 3-19: IOMMU settings

3.5.5 - Transparent Secure Memory Encryption (TSME)

This feature provides hardware memory encryption of all data stored on system DIMMs that is invisible to the OS and slightly increases memory latency.

Setting	Default	Options
TSME	Auto	• **Auto / Disabled:** Disable transparent secure memory encryption. • **Enabled:** Enable transparent secure memory encryption.

Table 3-20: TSME settings

3.5.6 - SEV, SEV-ES, and SEV-SNP

Please see the *AMD EPYC™ 9005 Cloud Infrastructure and Datacenter Design & Configuration Guide* (available from the AMD Documentation Hub) for information about AMD EPYC security features.

3.5.7 - AVIC& x2AVIC

Enabling Advanced Virtual Interrupt Controller (AVIC) may not improve system performance compared to systems with AVIC disabled. In some cases, performance with AVIC enabled may be significantly lower than system performance with AVIC disabled. AMD recommends not enabling AVIC on production systems.

x2AVIC is an extension of the advanced virtual interrupt controller (AVIC) that supports more than 255 virtual CPUs and offers better performance than AVIC. It can also be thought of as hardware virtualization of x2APIC. With x2AVIC, the guest's local APIC hardware-assisted virtualization extends to 512 virtual CPUs. Before AMD Socket SP5 Processors, the VM needed to disable x2APIC capabilities because using them would bypass hardware AVIC and use software-emulated x2APIC. However, with x2AVIC, the guest will be able to leverage x2APIC performance advantages.

Name	Default	Description
AVIC	Disabled	Advanced Virtual Interrupt Controller. • **Disabled:** Disables AVIC. • **Enabled:** Enables AVIC.
x2AVIC	Enable	x2AVIC is an extension of the advanced virtual interrupt controller. This feature currently requires a custom AMD Linux kernel. • **Disabled:** Disables x2AVIC. • **Enabled:** Enables x2AVIC.

Table 3-21: AVIC and x2AVIC settings

3.6 - Options Eliminated in AMD EPYC 9005 Series Processors

The following BIOS options were available in 3rd Gen AMD EPYC processors but are not available for user customization with AMD EPYC 9005 Series Processors as part of architecture improvements:

- EfficiencyModeEn
- Root Complex LCLK Frequency
- Preferred I/O
- Preferred I/O Bus

Chapter 4: Workload-Specific BIOS Settings

Use these guidelines for general-purpose workloads. Some cases list the benchmarks used to better describe the workloads used to obtain the recommended settings. Default settings are used when labeled default.

4.1 - General-Purpose Workloads

4.1.1 - Processor Core Settings

Setting	CPU Intensive	Java Throughput	Java Latency	Power Efficiency
SMT Control	*default*	*default*	*default*	*default*
L1 Stream HW Prefetcher	*default*	*default*	*default*	*default*
L1 Stride Prefetcher	*default*	*default*	*default*	*default*
L1 Region Prefetcher	*default*	*default*	*default*	*default*
L1 Burst Prefetch Mode	*default*	*default*	*default*	*default*
L2 Stream HW Prefetcher	*default*	*default*	*default*	*default*
L2 Up/Down Prefetcher	*default*	*default*	*default*	*default*
BoostFmaxEn	*default*	*default*	*default*	*default*
BoostFmax	*default*	*default*	*default*	*default*

Table 4-1: Processor core settings

4.1.2 - Power Management Settings

Setting	CPU Intensive	Java Throughput	Java Latency	Power Efficiency
Power Profile Selection	High Performance	High Performance	High Performance	Efficiency Mode
Determinism Control	*default*	*default*	*default*	*default*
Determinism Enable	*default*	*default*	*default*	*default*
TDP Control	Manual	Manual	Manual	Manual
TDP	OPN Max	OPN Max	OPN Max	OPN Max
PPT Control	Manual	Manual	Manual	Manual
PPT	OPN Max	OPN Max	OPN Max	OPN Max
CPPC	*default*	*default*	*default*	*default*

Table 4-2: Power efficiency settings

4.1.3 - NUMA and Memory Settings

Setting	CPU Intensive	Java Throughput	Java Latency	Power Efficiency
ACPI SRAT L3 Cache as NUMA Domain	*default*	*default*	*default*	Enabled
NUMA Nodes per Socket (NPS)	*default*	4	2	4
Memory Target Speed	*default*	*default*	*default*	*default*
Memory Interleaving	*default*	*default*	*default*	*default*

Table 4-3: NUMA and memory settings

4.1.4 - Infinity Fabric Settings

Setting	CPU Intensive	Java Throughput	Java Latency	Power Efficiency
xGMI Link Max Speed	*default*	*default*	*default*	*default*
xGMI Link Width Control	*default*	*default*	*default*	*default*
xGMI Max Link Width	*default*	*default*	*default*	*default*
xGMI Max Link Width Control	*default*	*default*	*default*	*default*
xGMI Force Link Width Control	*default*	*default*	*default*	*default*
xGMI Force Link Width	*default*	*default*	*default*	*default*
APBDIS	*default*	*default*	*default*	1
DF C-States	*default*	*default*	*default*	*default*

Table 4-4: Infinity Fabric DP settings

Note: NUMA-unaware workloads may need maximum xGMI bandwidth because of extensive cross-socket communications. NUMA-aware workloads may want to minimize xGMI power because they do not have a lot of cross socket traffic and prefer to use the increased CPU boost.

4.1.5 - I/O Settings

Setting	CPU Intensive	Java Throughput	Java Latency	Power Efficiency
Local APIC Mode	*default*	*default*	*default*	*default*
PCIe ARI Support [SRIOV]	*default*	*default*	*default*	*default*
PCIe Ten Bit Tag Support	*default*	*default*	*default*	*default*
IOMMU	*default*	*default*	*default*	*default*
TSME	*default*	*default*	*default*	*default*

Table 4-5: I/O settings

4.2 - Memory and I/O Intensive Workloads

4.2.1 - Processor Core Settings

Setting	Memory Throughput	Storage I/O Throughput	NIC Throughput	NIC Latency
SMT Control	*default*	*default*	Disabled	Disabled
L1 Stream HW Prefetcher	*default*	*default*	*default*	*default*
L1 Stride HW Prefetcher	*default*	*default*	*default*	*default*
L1 Region Prefetcher	*default*	*default*	*default*	*default*
L2 Stream HW Prefetcher	*default*	*default*	*default*	*default*
L2 Stream HW Prefetcher	*default*	*default*	*default*	*default*
L2 Up/Down Prefetcher	*default*	*default*	*default*	*default*
BoostFmaxEn	*default*	*default*	*default*	*default*
BoostFmax	*default*	*default*	*default*	*default*

Table 4-6: Processor core settings

4.2.2 - Power Management Settings

Setting	Memory Throughput	Storage I/O Throughput	NIC Throughput	NIC Latency
Power Profile Selection	*default*	*default*	*default*	*default*
Determinism Control	*default*	Enabled	Manual	Manual
Determinism Enable	*default*	Power	Performance	Performance
TDP Control	*default*	*default*	*default*	*default*
TDP	OPN Max	*default*	*default*	*default*
PPT Control	*default*	*default*	*default*	*default*
PPT	OPN Max	*default*	*default*	*default*
CPPC	*default*	*default*	*default*	*default*

Table 4-7: Power efficiency settings

4.2.3 - NUMA and Memory Settings

Setting	Memory Throughput	Storage I/O Throughput	NIC Throughput	NIC Latency
ACPI SRAT L3 Cache as NUMA Domain	Enabled	*default*	*default*	*default*
NUMA Nodes per Socket (NPS)	*default*	*default*	*default*	*default*
Memory Target Speed	*default*	*default*	*default*	*default*
Memory Interleaving	*default*	*default*	*default*	*default*

Table 4-8: NUMA and memory settings

4.2.4 - Infinity Fabric Settings

Setting	Memory Throughput	Storage I/O Throughput	NIC Throughput	NIC Latency
xGMI Link Max Speed	*default*	*default*	*default*	*default*
xGMI Link Width Control	*default*	*default*	*default*	*default*
xGMI Max Link Width	*default*	*default*	*default*	*default*
xGMI Max Link Width Control	*default*	*default*	*default*	x16
xGMI Force Link Width Control	*default*	*default*	*default*	*default*
xGMI Force Link Width	*default*	*default*	*default*	*default*
APBDIS	*default*	1	Disabled	Disabled
DF C-States	*default*	*default*	*default*	*default*

Table 4-9: Infinity Fabric settings

Note: NUMA-unaware workloads may need maximum xGMI bandwidth because of extensive cross-socket communications. NUMA-aware workloads may want to minimize xGMI power because they do not have a lot of cross socket traffic and prefer to use the increased CPU boost.

4.2.5 - I/O Settings

Setting	Memory Throughput	Storage I/O Throughput	NIC Throughput	NIC Latency
Local APIC Mode	*default*	*default*	*default*	*default*
PCIe ARI Support [SRIOV]	*default*	*default*	*default*	*default*
PCIe Ten Bit Tag Support	*default*	*default*	*default*	*default*
IOMMU	*default*	*default*	*default*	*default*
TSME	*default*	*default*	*default*	*default*

Table 4-10: I/O settings

4.3 - Virtualization and Containers

4.3.1 - Processor Core Settings

Setting	VMware vSphere Optimized	Linux KVM Optimized	Containers
SMT Control	Enabled	Enabled	Enabled
L1 Stream HW Prefetcher	*default*	*default*	*default*
L1 Stride HW Prefetcher	*default*	*default*	*default*
L1 Region Prefetcher	*default*	*default*	*default*
L1 Burst Prefetch Mode	*default*	*default*	*default*
L2 Stream HW Prefetcher	*default*	*default*	*default*
L2 Up/Down Prefetcher	*default*	*default*	*default*
BoostFmaxEn	*default*	*default*	*default*
BoostFmax	*default*	*default*	*default*

Table 4-11: Processor core settings

4.3.2 - Power Management Settings

Setting	VMware vSphere Optimized	Linux KVM Optimized	Containers
Power Profile Selection	*default*	*default*	*default*
Determinism Control	*default*	Enabled	*default*
Determinism Enable	*default*	Performance	*default*
TDP Control	*default*	*default*	*default*
TDP	*default*	*default*	*default*
PPT Control	*default*	*default*	*default*
PPT	*default*	*default*	*default*
CPPC	*default*	*default*	*default*

Table 4-12: Power efficiency settings

4.3.3 - NUMA and Memory Settings

Setting	VMware vSphere Optimized	Linux KVM Optimized	Containers
ACPI SRAT L3 Cache as NUMA Domain	*default*	Enabled	*default*
NUMA Nodes per Socket (NPS)	*default*	*default*	*default*
Memory Target Speed	*default*	*default*	*default*
Memory Interleaving	*default*	*default*	*default*

Table 4-13: NUMA and memory settings

4.3.4 - Infinity Fabric Settings

Setting	VMware vSphere Optimized	Linux KVM Optimized	Containers
xGMI Link Max Speed	*default*	*default*	*default*
xGMI Link Width Control	*default*	*default*	*default*
xGMI Max Link Width	*default*	*default*	*default*
xGMI Max Link Width Control	*default*	*default*	*default*
xGMI Force Link Width Control	*default*	*default*	*default*
xGMI Force Link Width	*default*	*default*	*default*
APBDIS	*default*	*default*	*default*
DF C-states	*default*	*default*	*default*

Table 4-14: Infinity Fabric settings

Note: NUMA-unaware workloads may need maximum xGMI bandwidth because of extensive cross-socket communications. NUMA-aware workloads may want to minimize xGMI power because they do not have a lot of cross socket traffic and prefer to use the increased CPU boost.

4.3.5 - I/O Settings

Setting	VMware vSphere Optimized	Linux KVM Optimized	Containers
Local APIC Mode	*default*	*default*	*default*
PCIe ARI Support [SRIOV]	*default*	*default*	*default*
PCIe Ten Bit Tag Support	*default*	*default*	*default*
IOMMU	*default*	*default*	*default*
TSME	*default*	*default*	*default*

Table 4-15: I/O settings

4.4 - Database, Analytics and AI/ML

4.4.1 - Processor Core Settings

Setting	RDBMS Optimized	Big Data Analytics Optimized	IoT Gateway	AI/ML
SMT Control	*default*	*default*	*default*	Enabled
L1 Stream HW Prefetcher	*default**	*default*	*default*	*default*
L1 Stride HW Prefetcher	*default*	*default*	*default*	*default*
L1 Region Prefetcher	*default*	*default*	*default*	*default*
L1 Burst Prefetch Mode	*default*	*default*	*default*	*default*
L2 Stream HW Prefetcher	*default**	*default*	*default*	*default*
L2 Up/Down Prefetcher	*default*	*default*	*default*	*default*
BoostFmaxEn	*default*	*default*	*default*	*default*
BoostFmax	*default*	*default*	*default*	*default*

Table 4-16: Processor core settings

*Disable when workload is memory constrained.

4.4.2 - Power Management Settings

Setting	RDBMS Optimized	Big Data Analytics Optimized	IoT Gateway	AI/ML
Power Profile Selection	Maximum IO Performance	*default*	*default*	*default*
Determinism Control	Enabled	*default*	*default*	Enabled
Determinism Enable	Power	*default*	*default*	Power
TDP Control	Manual	*default*	*default*	Manual
TDP	OPN Max	*default*	*default*	OPN Max
PPT Control	Manual	*default*	*default*	Manual
PPT	OPN Max	*default*	*default*	OPN Max
CPPC	*default*	*default*	*default*	*default*

Table 4-17: Power efficiency settings

4.4.3 - NUMA and Memory Settings

Setting	RDBMS Optimized	Big Data Analytics Optimized	IoT Gateway	AI/ML
ACPI SRAT L3 Cache as NUMA Domain	*default*	Enabled	*default*	NPS1
NUMA Nodes per Socket (NPS)	• NPS4 (Windows)* • *default* (Linux)	*default*	*default*	*default*
Memory Target Speed	Auto	Auto	Auto	Auto
Memory Interleaving	*default*	*default*	*default*	*default*

Table 4-18: NUMA and memory settings

*Due to the Windows kgroup max of 64 logical processors:

- If (Cores/Socket <= 32 then set NPS1
- If (32 < Cores/Socket <= 64) then set NPS2.
- IF (64 > Cores/Socket) then set NPS4.

4.4.4 - Infinity Fabric Settings

Setting	RDBMS Optimized	Big Data Analytics Optimized	IoT Gateway	AI/ML
xGMI Link Max Speed	*default*	*default*	*default*	*default*
xGMI Link Width Control	*default*	*default*	*default*	*default*
xGMI Max Link Width	*default*	*default*	*default*	*default*
xGMI Max Link Width Control	*default*	*default*	*default*	*default*
xGMI Force Link Width Control	*default*	*default*	*default*	*default*
xGMI Force Link Width	*default*	*default*	*default*	*default*
APBDIS	1	*default*	*default*	*default*
DF C-States	*default*	*default*	*default*	*default*

Table 4-19: Infinity Fabric settings

Note: NUMA-unaware workloads may need maximum xGMI bandwidth because of extensive cross-socket communications. NUMA-aware workloads may want to minimize xGMI power because they do not have a lot of cross socket traffic and prefer to use the increased CPU boost.

4.4.5 - I/O Settings

Setting	RDBMS Optimized	Big Data Analytics Optimized	IoT Gateway	AI/ML
Local APIC Mode	*default*	*default*	*default*	*default*
PCIe ARI Support [SRIOV]	*default*	*default*	*default*	*default*
PCIe Ten Bit Tag Support	*default*	*default*	*default*	*default*
IOMMU	*default*	*default*	*default*	*default*
TSME	*default*	*default*	*default*	*default*

Table 4-20: I/O settings

4.5 - HPC, Telco, and EDA Settings

4.5.1 - Processor Core Settings

Setting	HPC	Telco (CORE)	Telco (Mgmt, Security, CDN)	EDA
SMT Control	Disabled	*default*	*default*	Disabled
L1 Stream HW Prefetcher	*default*	*default*	*default*	*default*
L1 Stride HW Prefetcher	*default*	*default*	*default*	*default*
L1 Region Prefetcher	*default*	*default*	*default*	*default*
L1 Burst Prefetch Mode	*default*	*default*	*default*	*default*
L2 Stream HW Prefetcher	*default*	*default*	*default*	*default*
L2 Up/Down Prefetcher	*default*	*default*	*default*	*default*
BoostFmaxEn	*default*	*default*	*default*	*default*
BoostFmax	*default*	*default*	*default*	*default*

Table 4-21: Processor core settings

4.5.2 - Power Management Settings

Setting	HPC	Telco (CORE)	Telco (Mgmt, Security, CDN)	EDA
Power Profile Selection	High Performance	*default*	*default*	*default*
Determinism Control	*default*	Manual	Manual	Manual
Determinism Enable	Performance	Performance	Performance	Power
TDP Control	Manual	*default*	*default*	Manual
TDP	OPN Max	*default*	*default*	OPN Max
PPL Control	Manual	*default*	*default*	Manual
PPL	OPN Max	*default*	*default*	OPN Max
CPPC	*default*	*default*	*default*	*default*

Table 4-22: Power efficiency settings

4.5.3 - NUMA and Memory Settings

Setting	HPC	Telco (CORE)	Telco (Mgmt, Security, CDN)	EDA
ACPI SRAT L3 Cache as NUMA Domain	*default*	*default*	*default*	*default*
NUMA Nodes per Socket (NPS)	4	*default*	*default*	2
Memory Target Speed	*default*	*default*	*default*	*default*
Memory Interleaving	*default*	*default*	*default*	*default*

Table 4-23: NUMA and memory settings

4.5.4 - Infinity Fabric Settings

Setting	HPC	Telco (CORE)	Telco (Mgmt, Security, CDN)	EDA
xGMI Link Max Speed	*default*	*default*	*default*	*default*
xGMI Link Width Control	*default*	*default*	*default*	Manual
xGMI Max Link Width	*default*	*default*	*default*	*default*
xGMI Max Link Width Control	*default*	*default*	*default*	Manual
xGMI Force Link Width Control	*default*	*default*	*default*	Force
xGMI Force Link Width	*default*	*default*	*default*	*default*
APBDIS	*default*	*default*	*default*	1
DF C-States	*default*	*default*	*default*	Disabled

Table 4-24: Infinity Fabric settings

Note: NUMA-unaware workloads may need maximum xGMI bandwidth because of extensive cross-socket communications. NUMA-aware workloads may want to minimize xGMI power because they do not have a lot of cross socket traffic and prefer to use the increased CPU boost.

4.5.5 - I/O Settings

Setting	HPC	Telco (CORE)	Telco (Mgmt, Security, CDN)	EDA
Local APIC Mode	*default*	*default*	*default*	*default*
PCIe ARI Support [SRIOV]	*default*	Enabled	*default*	*default*
PCIe 10-Bit Tag Support	*default*	Enabled	*default*	*default*
IOMMU	Enabled*	*default*	*default*	*default* (Enabled with 2P 64-core CPUs with SMT=Enabled)
TSME	*default*	*default*	*default*	*default*

Table 4-25: I/O settings

* = For HPC, enable IOMMU in BIOS. Within Linux, add the boot command `iommu=pt` to set the IOMMU to passthrough mode.

Chapter 5: Debugging BIOS Setting Updates

Some BIOS default settings can be changed from the command line by a user with adequate privileges. Some of these settings take effect immediately while others may require rebooting. Some settings may not be available in the BIOS but can be set at the OS level. Some OS-level settings can either override or modify the expected BIOS default behavior. This section lists some of the settings that could have a significant impact on system performance and response times.

Note: For each of the following examples, if you are running Windows, then please see the Microsoft® Windows® Server Tuning Guide for AMD EPYC™ 9005 Series Processors (available from the AMD Documentation Hub).

NUMA Node Configurations:

The number of NUMA nodes, association of nodes to the memory channels, and binding of processes to specific NUMA nodes and memory nodes plays a vital role in many deployments. The **NPS** and **L3 as NUMA** settings control the number of NUMA nodes and memory channels associated with the NUMA nodes. To do this, a user leverages the NUMA command options as defined for a given OS and platforms and expects the processes to bind to certain CPUs and memory as per default BIOS settings. However, many OS-level commands and daemons may, if enabled, alter this expected behavior. The following methodology is recommended to understand and debug NUMA related issues:

Understand NUMA topology:

Verify that the number of NUMA nodes and topology are correct.

Linux has many commands such as `lstopo`, `hwloc`, and `numactl`. Here is some sample `numactl` output from a single socket AMD EPYC OPN with 12CCDs per Socket where NPS=4 and L3 as NUMA=Disabled:

```
$ numactl -H
available: 4 nodes (0-3)
node 0 cpus: 0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 96 97 98 99 100 101 102 103 104
105 106 107 108 109 110 111 112 113 114 115 116 117 118 119
node 0 size: 32225 MB
node 0 free: 31374 MB
node 1 cpus: 24 25 26 27 28 29 30 31 32 33 34 35 36 37 38 39 40 41 42 43 44 45 46 47 120 121 122 123 124
125 126 127 128 129 130 131 132 133 134 135 136 137 138 139 140 141 142 143
node 1 size: 64448 MB
node 1 free: 63633 MB
node 2 cpus: 48 49 50 51 52 53 54 55 56 57 58 59 60 61 62 63 64 65 66 67 68 69 70 71 144 145 146 147 148
149 150 151 152 153 154 155 156 157 158 159 160 161 162 163 164 165 166 167
node 2 size: 96407 MB
node 2 free: 95509 MB
node 3 cpus: 72 73 74 75 76 77 78 79 80 81 82 83 84 85 86 87 88 89 90 91 92 93 94 95 168 169 170 171 172
173 174 175 176 177 178 179 180 181 182 183 184 185 186 187 188 189 190 191
node 3 size: 32164 MB
node 3 free: 31221 MB
node distances:
node   0   1   2   3
   0:  10  12  12  12
   1:  12  10  12  12
   2:  12  12  10  12
   3:  12  12  12  10
```

This sample output shows that logical CPU ids are contiguous within nodes. For example, the cores on Node 0 are 0-23, which are contiguous. This convention is expected from most systems but is not guaranteed and could be different based on a given OEM platform. Depending on your operating system, you should confirm these mappings when using CPU affinity for important performance decisions.

Check daemons and services that could alter default behavior:

Each OS and platform provides various tools and services that attempt to optimize a given system. There are daemons and services intended to change or allocate CPU assignments and memory to leverage locality. This can sometimes result in unintended behavior where NUMA-assigned CPUs and memory may migrate after the initial application launch. For example, `numad` is a Linux system daemon that monitors NUMA topology and resource usage. It will attempt to locate processes for efficient NUMA locality and affinity by dynamically adjusting to changing system conditions. Enabling this feature may interfere with and override the initial launch commands of a deployment like `physcpubind`, `membind`, etc. If user is observing unexpected process threads and memory migrations, check if this daemon is enabled.

Check NUMA optimization policies:

Most platform attempt to allocate CPU and memory resources optimal NUMA leverage. Some OS and virtualization platform may have many available policy options. For example Linux automatic NUMA balancing moves tasks (threads or processes) closer to the memory they are accessing. Most applications and deployments benefit from being close to memory, meaning that this feature is often be enabled by default. NUMA balancing can have undesired effects. Further, a user's ability to bind the process CPUs and memory to different NUMA nodes may cause this setting to interfere with expected behavior. Execute the following command to disable NUMA balancing:

```
echo 0 > /proc/sys/kernel/numa_balancing
```

OS-level settings that are not available in BIOS:

An OEM platform may not provide BIOS options to change certain settings, but users may have been able to change them at the OS level. **AMD EPYC Core C6 (CC6) States** (alternately named C2 at the OS level) is one such example.

5th Gen AMD EPYC processors have C-States associated to cores and the Infinity Data fabric (DF). Disabling processor core C- States is highly discouraged. The system BIOS includes options to disable DF-C States for low latency and jitter-sensitive use cases. You can execute the following command to disable the Core C6 (CC6) state for the for all of the CPUs in a given system:

```
cpupower idle-set –d 2
```

You can also selectively disable Core C6. For a dual-socket system with 96-core processors, use 0-191 in the command to disable C2 for all 192 cores by executing the following command:

```
cpupower -c 0-191 idle-set -d 2.
```

THIS PAGE INTENTIONALLY LEFT BLANK.

BIOS & Workload Tuning Guide for AMD EPYC™ 9005 Processors

PID: 58467

Muhammad Ashfaq, Anil Rajput, and Jesse Rangel

READY TO CONNECT?

m/epyc