AMD EPYC

TUNING GUIDE
AMD EPYC 9004

RDBMS

Publication	57996
Revision	1.3
Issue Date	June, 2023

Date	Version	Changes
July, 2022	0.1	Initial NDA partner release
Sep, 2022	0.2	Updated BIOS information
Nov, 2022	1.0	Initial public version
Dec, 2022	1.1	Minor errata corrections
Mar, 2023	1.2	Added 97xx OPN and AMD 3D V-Cache™ technology information
Jun, 2023	1.3	Second public release

Audience

This tuning guide is intended for a technical audience such as RDBMS application architects, production deployment, and performance engineering teams who have:

- 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.

Authors

Muhammad Ashfaq and Sylvester Rajasekaran

Note: All of the settings described in this Tuning Guide apply to all AMD EPYC 9004 Series Processors of all core counts with or without AMD 3D V-Cache™ except where explicitly noted otherwise.

Table of Contents

Note: If you are running Microsoft® SQL Server, then please see the Microsoft® SQL Server Tuning Guide for AMD EPYC™ 9004 Series Processors.

Chapter 1 Introduction

Relational databases such as MySQL and MariaDB can deliver high performance with proper configuration and tuning. This Tuning Guide gives best practices for:

- CPU selection
- Memory sizing
- BIOS tuning
- Tuning OS- and database-specific parameters

Note: Storage and networking also affect performance.

Please review the *BIOS & Workload Tuning Guide for AMD EPYC™ 9004 Series Processors* (available from AMD EPYC Tuning Guides) before tuning your system for specific databases and workloads.

If you are running Microsoft® SQL Server, then please see the *Microsoft® SQL Server Tuning Guide for AMD EPYC™ 9004 Series Processors* (available from AMD EPYC Tuning Guides).

1.1 CPU Selection Guidelines for RDBMS

4th Gen AMD EPYC™ processors deliver high IOPS and throughput for all databases. Selecting the right CPU is important for archiving optimal database application performance. Table 1-1 recommends CPUs based on database size, concurrent users, and performance requirements.

Workload Characteristics	Workload Type	Size/# of Users	Cores	Processor	Memory
Small	DSS	• Up to 300 GB • Up to 5 concurrent users	8	1 x EPYC 16-core	128GB
Medium	DSS	• Up to 300GB - 1TB • Up to 10 concurrent users	16	1 x EPYC 16-core or 2 x EPYC 16-core	256GB
Large	DSS	• Up to >1TB • Up to 20 concurrent users	32	1 x EPYC 32-core or 2 x EPYC 16-core	512GB
Small	OLTP	• Up to 250GB • Up to 50 concurrent users	16	1 x EPYC 16-core or 2 x EPYC 16-core	256GB
Medium	OLTP	• Up to 250GB-500GB • Up to 100 con current users	32	1 x EPYC 32-core or 2 x EPYC 16-core	512GB
Large	OLTP	• Up to 500GB-1TBU • Up to 200 con current users	64	1 x EPYC 64-core or 2 x EPYC 32-core	1-2TB
Clouds	Each Cloud Service Provider (CSP) in your geographical region has different cloud instance/VM sizes to suit your choice of database workloads. Please refer to your selected CSP instance offerings to determine the instance with the best mix of vCPU cores, memory, storage, and networking for your particular workload.				

Table 1-1: Selecting the right 4th Gen AMD EPYC 9004 processors for various workloads

AMD recommends using 8, 16, or 32 vCPUs with 1:8 vCPU to memory ratio VMs with appropriate attached storage for optimal cloud IaaS performance.

Chapter 2

AMD EPYC™ 9004 Series Processors

AMD EPYC™ 9004 Series Processors represent the fourth generation of AMD EPYC server-class processors. This generation of AMD EPYC processors feature AMD's latest "Zen 4" based compute cores, next-generation Infinity Fabric, next-generation memory & I/O technology, and use the new SP5 socket/packaging.

2.1 General Specifications

AMD EPYC 9004 Series Processors offer a variety of configurations with varying numbers of cores, Thermal Design Points (TDPs), frequencies, cache sizes, etc. that complement AMD's existing server portfolio with further improvements to performance, power efficiency, and value. Table 1-1 lists the features common to all AMD EPYC 9004 Series Processors.

Common Features of all AMD EPYC 9004 Series Processors	
Compute cores	Zen4-based
Core process technology	5nm
Maximum cores per Core Complex (CCX)	8
Max memory per socket	6 TB
Max # of memory channels	12 DDR5
Max memory speed	4800 MT/s DDR5
Max lanes Compute eXpress Links	64 lanes CXL 1.1+
Max lanes Peripheral Component Interconnect	128 lanes PCIe® Gen 5

Table 2-1: Common features of all AMD EPYC 9004 Series Processors

2.2 Model-Specific Features

Different models of 4th Gen AMD EPYC processors have different feature sets, as shown in Table 1-2.

AMD EPYC 9004 Series Processor (Family 19h) Features by Model		
Codename	"Genoa"*	"Bergamo"*
Model #	91xx-96xx	97xx
Max number of Core Complex Dies (CCDs)	12	8
Number of Core Complexes (CCXs) per CCD	1	2
Max number of cores (threads)	96 (192)	128 (256)
Max L3 cache size (per CCX)	1,152 MB (96 MB)♦	256 MB (16 MB)
Max Processor Frequency	4.4 GHz♦♦	3.15 GHz
Includes ♦AMD 3D V-Cache (9xx4X) and ♦♦high-frequency (9xx4F) models.		
*GD-122: The information contained herein is for informational purposes only and is subject to change without notice. Timelines, roadmaps, and/or product release dates shown herein and plans only and subject to change. "Genoa" and "Bergamo" are codenames for AMD architectures and are not product names.		

Table 2-2: AMD EPYC 9004 Series Processors features by model

2.3 Operating Systems

AMD recommends using the latest available targeted OS version and updates. Please see AMD EPYC™ Processors Minimum Operating System (OS) Versions for detailed OS version information.

2.4 Processor Layout

AMD EPYC 9004 Series Processors incorporate compute cores, memory controllers, I/O controllers, RAS (Reliability, Availability, and Serviceability), and security features into an integrated System on a Chip (SoC). The AMD EPYC 9004 Series Processor retains the proven Multi-Chip Module (MCM) Chiplet architecture of prior successful AMD EPYC processors while making further improvements to the SoC components.

The SoC includes the Core Complex Dies (CCDs), which contain Core Complexes (CCXs), which contain the "Zen 4"-based cores. The CCDs surround the central high-speed I/O Die (and interconnect via the Infinity Fabric). The following sections describe each of these components.

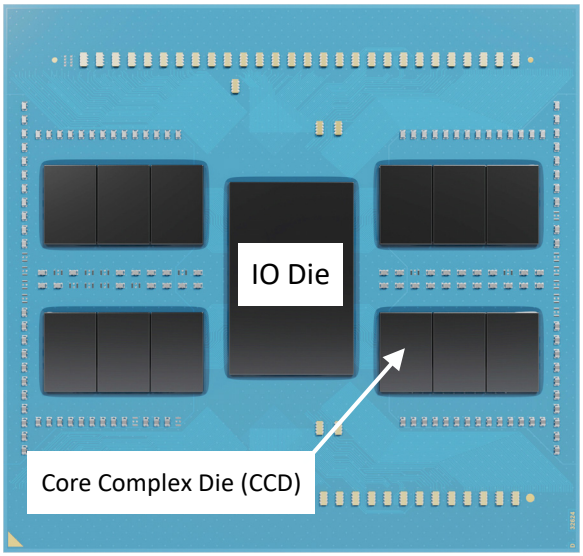

Figure 2-1: AMD EPYC 9004 configuration with 12 Core Complex Dies (CCD) surrounding a central I/O Die (IOD)

2.5 "Zen 4" Core

AMD EPYC 9004 Series Processors are based on the new "Zen 4" compute core. The "Zen 4" core is manufactured using a 5nm process and is designed to provide an Instructions per Cycle (IPC) uplift and frequency improvements over prior generation "Zen" cores. Each core has a larger L2 cache and improved cache effectiveness over the prior generation. Each "Zen 4" core includes:

- Up to 32 KB of 8-way L1 I-cache and 32 KB of 8-way of L1 D-cache
- Up to a 1 MB private unified (Instruction/Data) L2 cache.

Each core supports Simultaneous Multithreading (SMT), which allows 2 separate hardware threads to run independently, sharing the corresponding core's L2 cache.

2.6 Core Complex (CCX)

Figure 2-2 shows a Core Complex (CCX) where up to eight "Zen 4"-based cores share a L3 or Last Level Cache (LLC). Enabling Simultaneous Multithreading (SMT) allows a single CCX to support up to 16 concurrent hardware threads.

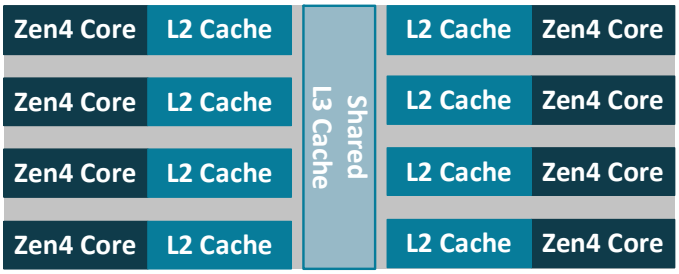

Figure 2-2: Top view of 8 compute cores sharing an L3 cache (91xx-96xx models)

2.7 Core Complex Dies (CCDs)

The Core Complex Die (CCD) in an AMD EPYC 9xx4 Series Processor may contain either one or two CCXs, depending on the processor (91xx-96xx "Genoa" vs. 97xx "Bergamo"), as shown in Figure 2-5.

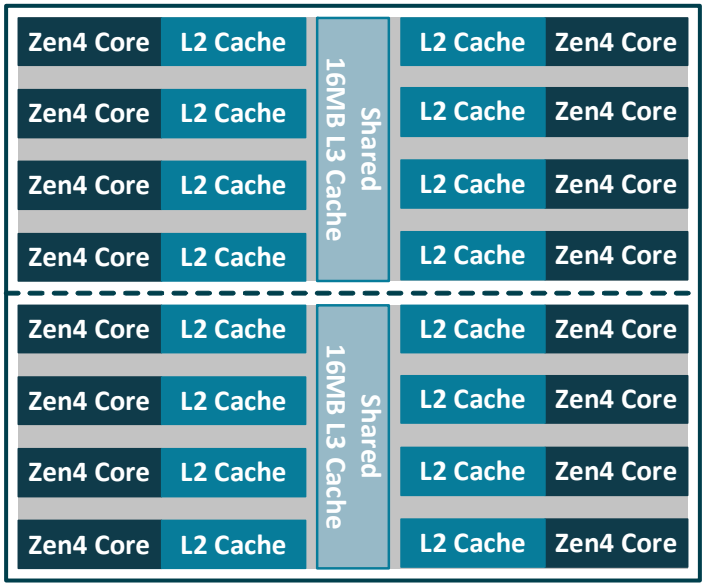

Figure 2-3: 2 CCXs in a single 4th Gen AMD EPYC 97xx CCD

Each of the Core Complex Dies (CCDs) in a 97xx model AMD EPYC 9004 Series Processor contains two CCXs (Figure 2-5):

AMD EPYC 9004 Series Processor	91xx-96xx	97xx
# of CCXs within a CCD	1	2

Table 2-3: CCXs per CCD by AMD EPYC model

You can disable cores in BIOS using one or both of the following approaches:

- Reduce the cores per L3 from 8 down to 7,6,5,4,3,2, or 1 while keeping the number of CCDs constant. This approach increases the effective cache per core ratio but reduces the number of cores sharing the cache.
- Reduce the number of active CCDs while keeping the cores per CCD constant. This approach maintains the advantages of cache sharing between the cores while maintaining the same cache per core ratio.

2.8 AMD 3D V-Cache™ Technology

AMD EPYC 9xx4X Series Processors include AMD 3D V-Cache™ die stacking technology that enables 97xx to achieve more efficient chiplet integration. AMD 3D Chiplet architecture stacks L3 cache tiles vertically to provide up to 96MB of L3 cache per die (and up to 1 GB L3 Cache per socket) while still providing socket compatibility with all AMD EPYC™ 9004 Series Processor models.

AMD EPYC 9004 Series Processors with AMD 3D V-Cache technology employ industry-leading logic stacking based on copper-to-copper hybrid bonding "bumpless" chip-on-wafer process to enable over 200X the interconnect densities of current 2D technologies (and over 15X the interconnect densities of other 3D technologies using solder bumps), which translates to lower latency, higher bandwidth, and greater power and thermal efficiencies.

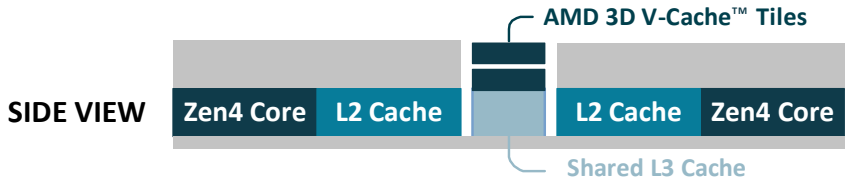

Figure 2-4: Side view of vertically-stacked central L3 SRAM tiles

AMD EPYC 9004 Series Processors	9xx4	9004X (with 3D V-Cache)
Max Shared L3 Cache per CCD	32 MB	96 MB

Table 2-4: L3 cache by processor model

Different OPNs also may have different numbers of cores within the CCX. However, for any given part, all CCXs will always contain the same number of cores.

2.9 I/O Die (Infinity Fabric™)

The CCDs connect to memory, I/O, and each other through an updated I/O Die (IOD). This central AMD Infinity Fabric™ provides the data path and control support to interconnect CCXs, memory, and I/O. Each CCD connects to the IOD via a dedicated high-speed Global Memory Interconnect (GMI) link. The IOD helps maintain cache coherency and additionally provides the interface to extend the data fabric to a potential second processor via its xGMI, or G-links. AMD EPYC 9004 Series Processors support up to 4 xGMI (or G-links) with speeds up to 32Gbps. The IOD exposes DDR5 memory channels, PCIe® Gen5, CXL 1.1+, and Infinity Fabric links.

All dies (chiplets) interconnect with each other via AMD Infinity Fabric technology. Figure 2-6 (which corresponds to Figure 2-2, above) shows the layout of a 96-core AMD EPYC 9654 processor. The AMD EPYC 9654 has 12 CCDs, with each CCD connecting to the IOD via its own GMI connection.

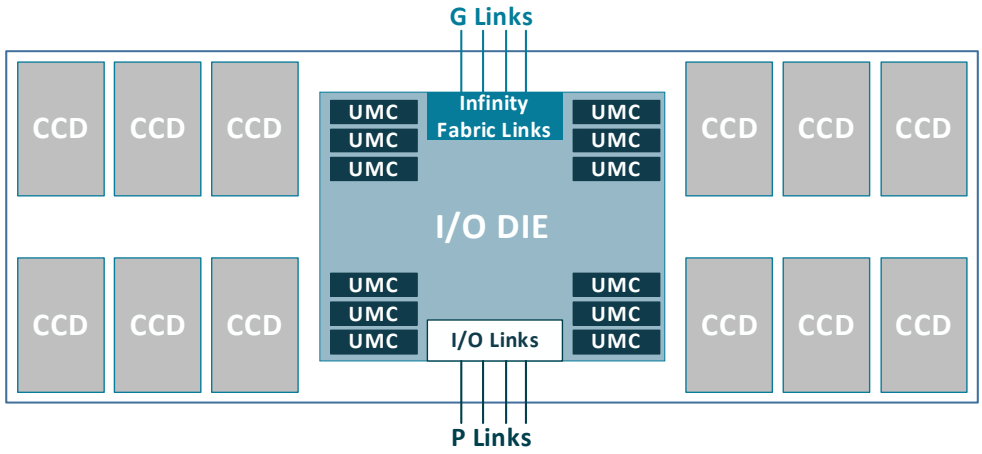

Figure 2-5: AMD EPYC 9654 processor internals interconnect via AMD Infinity Fabric (12 CCD processor shown)

AMD also provides "wide" OPNs (e.g. AMD EPYC 9334) where each CCD connects to two GMI3 interfaces, thereby allowing double the Core-to-I/O die bandwidth.

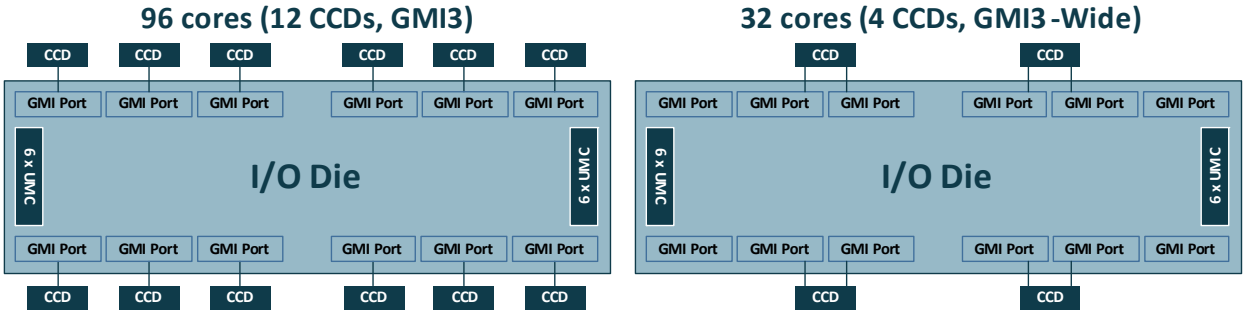

Figure 2-6: Standard vs. Wide GMI links

The IOD provides twelve Unified Memory Controllers (UMCs) that support DDR5 memory. The IOD also presents 4 'P-links' that the system OEM/designer can configure to support various I/O interfaces, such as PCIe Gen5, and/or CXL 1.1+.

2.10 Memory and I/O

Each UMC can support up to 2 DIMMs per channel (DPC) for a maximum of 24 DIMMs per socket. OEM server configurations may allow either 1 DIMM per channel or 2 DIMMs per channel. 4th Gen AMD EPYC processors can support up to 6TB of DDR5 memory per socket. Having additional and faster memory channels compared to previous generations of AMD EPYC processors provides additional memory bandwidth to feed high-core-count processors. Memory interleaving on 2, 4, 6, 8, 10, and 12 channels helps optimize for a variety of workloads and memory configurations.

Each processor may have a set of 4 P-links and 4 G-links. An OEM motherboard design can use a G-link to either connect to a second 4th Gen AMD EPYC processor or to provide additional PCIe Gen5 lanes. 4th Gen AMD EPYC processors support up to eight sets of x16-bit I/O lanes, that is, 128 lanes of high-speed PCIe Gen5 in single-socket platforms and up to 160 lanes in dual-socket platforms. Further, OEMs may either configure 32 of these 128 lanes as SATA lanes and/or configure 64 lanes as CXL 1.1+. In summary, these links can support:

- Up to 4 G-links of AMD Infinity Fabric connectivity for 2P designs.
- Up to 8 x16 bit or 128 lanes of PCIe Gen 5 connectivity to peripherals in 1P designs (and up to 160 lanes in 2-socket designs).
- Up to 64 lanes (4 P-links) that can be dedicated to Compute Express Link (CXL) 1.1+ connectivity to extended memory.
- Up to 32 I/O lanes that can be configured as SATA disk controllers.

2.11 Visualizing AMD EPYC 9004 Series Processors (Family 19h)

This section depicts AMD EPYC 9004 Series Processors that have been set up with four nodes per socket (NPS=4). Please see "NUMA Topology" on page 10 for more information about nodes.

2.11.1 Models 91xx-96xx ("Genoa")

4th Gen AMD EPYC 9004 processors with model numbers 91xx-96xx have up to 12 CCDs that each contain a single CCX, as shown below.

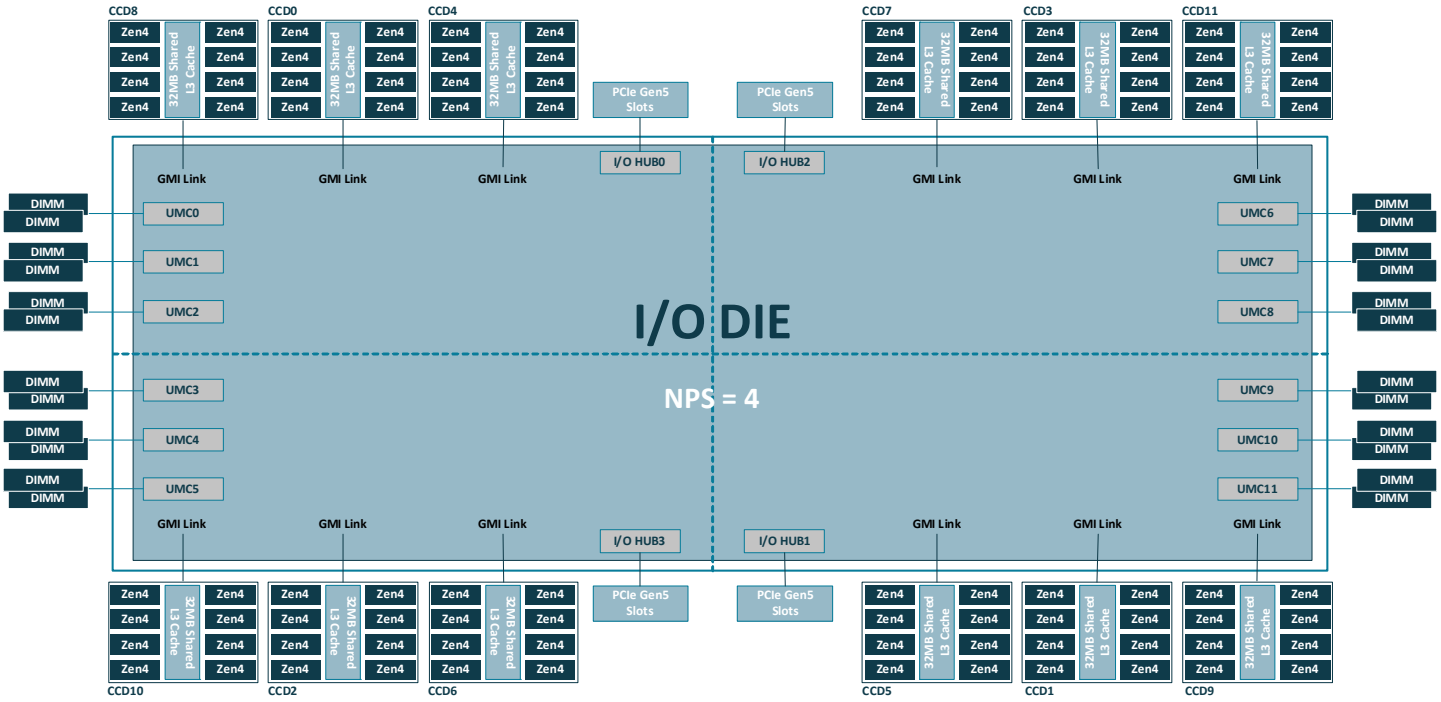

Figure 2-7: The AMD EPYC 9004 SoC consists of up to 12 CCDs and a central IOD for 91xx-96xx models, including "X" OPNs

2.11.2 Models 97xx (“Bergamo”)

97xx 4th Gen AMD EPYC 9004 Series Processors with model numbers 97xx have up to 8 CCDs that each contain two CCXs, as shown below.

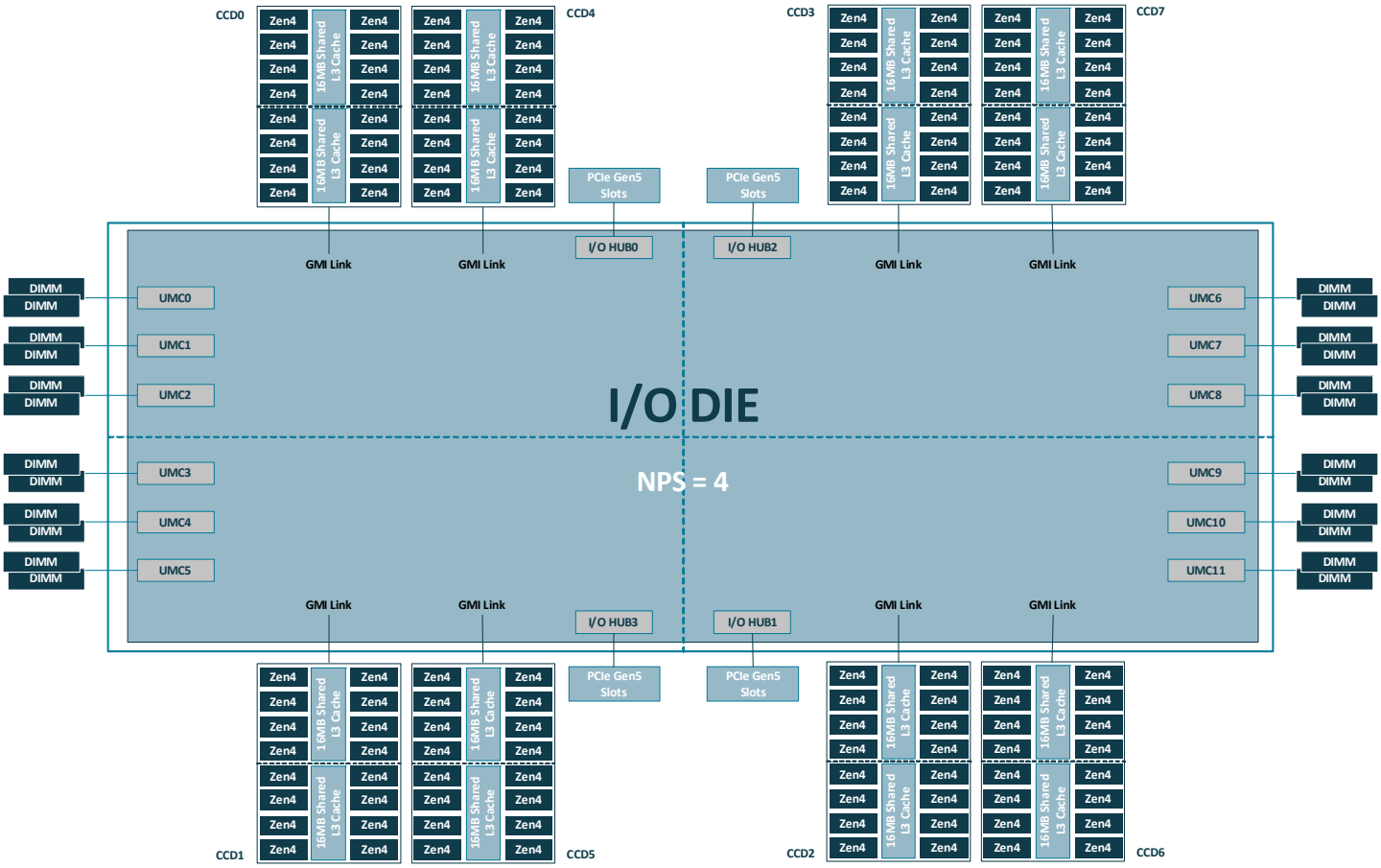

Figure 2-8: The AMD EPYC 9004 System on Chip (SoC) consists of up to 8 CCDs and a central IOD for 97xx models

2.12 NUMA Topology

AMD EPYC 9004 Series Processors use a Non-Uniform Memory Access (NUMA) architecture where different latencies may exist depending on the proximity of a processor core to memory and I/O controllers. Using resources within the same NUMA node provides uniform good performance, while using resources in differing nodes increases latencies.

2.12.1 NUMA Settings

A user can adjust the system **NUMA Nodes Per Socket** (NPS) BIOS setting to optimize this NUMA topology for their specific operating environment and workload. For example, setting NPS=4 as shown in “Memory and I/O” on page 8 divides the processor into quadrants, where each quadrant has 3 CCDs, 3 UMCs, and 1 I/O Hub. The closest processor-memory I/O distance is between the cores, memory, and I/O peripherals within the same quadrant. The furthest distance is between a core and memory controller or IO hub in cross- diagonal quadrants (or the other processor in a 2P configuration). The locality of cores, memory, and IO hub/devices in a NUMA-based system is an important factor when tuning for performance.

The NPS setting also controls the interleave pattern of the memory channels within the NUMA Node. Each memory channel within a given NUMA node is interleaved. The number of channels interleaved decreases as the NPS setting gets more granular. For example:

- A setting of NPS=4 partitions the processor into four NUMA nodes per socket with each logical quadrant configured as its own NUMA domain. Memory is interleaved across the memory channels associated with each quadrant. PCIe devices will be local to one of the four processor NUMA domains, depending on the IOD quadrant that has the corresponding PCIe root complex for that device.

- A setting of NPS=2 configures each processor into two NUMA domains that groups half of the cores and half of the memory channels into one NUMA domain, and the remaining cores and memory channels into a second NUMA domain. Memory is interleaved across the six memory channels in each NUMA domain. PCIe devices will be local to one of the two NUMA nodes depending on the half that has the PCIe root complex for that device.

- A setting of NPS=1 indicates a single NUMA node per socket. This setting configures all memory channels on the processor into a single NUMA node. All processor cores, all attached memory, and all PCIe devices connected to the SoC are in that one NUMA node. Memory is interleaved across all memory channels on the processor into a single address space.

- A setting of NPS=0 indicates a single NUMA domain of the entire system (across both sockets in a two-socket configuration). This setting configures all memory channels on the system into a single NUMA node. Memory is interleaved across all memory channels on the system into a single address space. All processor cores across all sockets, all attached memory, and all PCIe devices connected to either processor are in that single NUMA domain.

You may also be able to further improve the performance of certain environments by using the **LLC (L3 Cache) as NUMA** BIOS setting to associate workloads to compute cores that all share a single LLC. Enabling this setting equates each shared L3 or CCX to a separate NUMA node, as a unique L3 cache per CCD. A single AMD EPYC 9004 Series Processor with 12 CCDs can have up to 12 NUMA nodes when this setting is enabled.

Thus, a single EPYC 9004 Series Processor may support a variety of NUMA configurations ranging from one to twelve NUMA nodes per socket.

Note: If software needs to understand NUMA topology or core enumeration, it is imperative to use documented Operating System (OS) APIs, well-defined interfaces, and commands. Do not rely on past assumptions about settings such as APICID or CCX ordering.

2.13 Dual-Socket Configurations

AMD EPYC 9004 Series Processors support single- or dual-socket system configurations. Processors with a 'P' suffix in their name are optimized for single-socket configurations (see the "Processor Identification" chapter) only. Dual-socket configurations require both processors to be identical. You cannot use two different processor Ordering Part Numbers (OPNs) in a single dual-socket system.

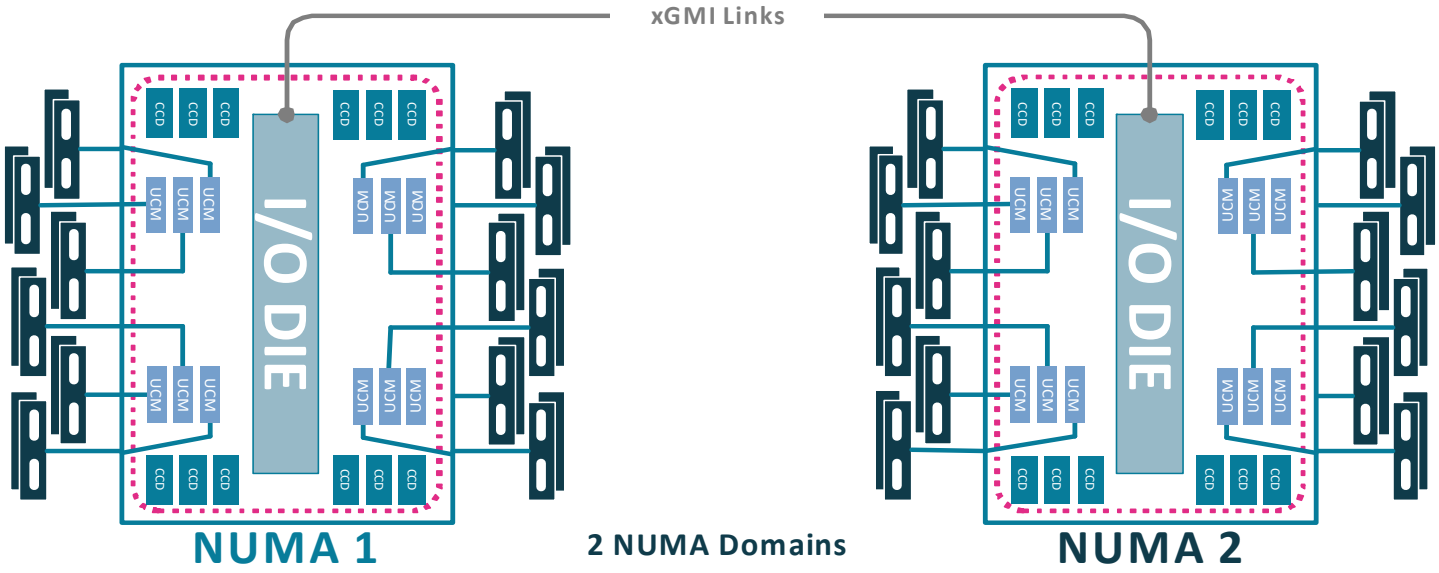

Figure 2-9: Two EPYC 9004 Processors connect through 4 xGMI links (NPS1)

In dual-socket systems, two identical EPYC 9004 series SoCs are connected via their corresponding External Global Memory Interconnect [xGMI] links. This creates a high bandwidth, low latency interconnect between the two processors. System manufacturers can elect to use either 3 or 4 of these Infinity Fabric links depending upon I/O and bandwidth system design objectives.

The Infinity Fabric links utilize the same physical connections as the PCIe lanes on the system. Each link uses up to 16 PCIe lanes. A typical dual socket system will reconfigure 64 PCIe lanes (4 links) from each socket for Infinity Fabric connections. This leaves each socket with 64 remaining PCIe lanes, meaning that the system has a total of 128 PCIe lanes. In some cases, a system designer may want to expose more PCIe lanes for the system by reducing the number of Infinity Fabric G-Links from 4 to 3. In these cases, the designer may allocate up to 160 lanes for PCIe (80 per socket) by utilizing only 48 lanes per socket for Infinity Fabric links instead of 64.

A dual-socket system has a total of 24 memory channels, or 12 per socket. Different OPNs can be configured to support a variety of NUMA domains.

Chapter 3

BIOS Defaults Summary

This chapter provides high-level lists of the default AMD EPYC 9004 BIOS settings and their default values. Please see Chapter 4 of the *BIOS & Workload Tuning Guide for AMD EPYC™ 9004 Series Processors* (available from AMD EPYC Tuning Guides) for detailed descriptions. Later chapters in this Tuning Guide discuss the BIOS options as they relate to a specific workload or set of workloads.

Note: The default setting names 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.

AMD strongly recommends that customers download and install the latest BIOS update for your AMD EPYC 9004 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.

3.1 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 3-1: Processor core BIOS settings

X3D	Auto	Enables or disables AMD 3D V-Cache™ technology on Cache Optimized (9004X) processors. • **Auto:** Enabled on an AMD EPYC 9004 Series processor with AMD 3D V-Cache™ technology, enabling this option enables the AMD 3D V-Cache module in the CCD to increase the total size of the L3 cache memory to 96MB • **Disabled:** Disabling this option reduces the L3 cache in the CCD to 32MB. *Note: This option only applies to AMD EPYC 9004 Series Processors with AMD 3D V-Cache technology.* *Note: AMD engineers performed extensive internal testing and validation for various applications using the X3D BIOS option found in AMD EPYC 9xx4X processors with AMD 3D V-Cache technology. This testing and validation cannot cover all applications or use cases. Testing has shown AMD 3D V-Cache to be beneficial for most workloads, however AMD recommends that you test and evaluate the benefits of enabling or disabling the X3D BIOS option for your application workloads in your environment and proceeding based on those results.*

Table 3-1: Processor core BIOS settings

3.2 Power Efficiency Settings

Name	Default	Description
Power Profile Selection	Auto	• **Auto/0:** High-performance mode • **1:** Efficiency mode • **2:** Maximum I/O performance mode
Determinism Control	Auto	• **Auto:** Use default performance determinism settings. • **Manual:** Specify custom performance determinism settings.
Determinism Enable	Auto	• **Auto:** Performance. • **1:** Power.
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 85-400:** 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 85-400:** 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 3-2: Power efficiency BIOS settings

3.3 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/Auto:** 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 3200–5600 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.

Table 3-3: NUMA and memory BIOS settings

3.4 Infinity Fabric Settings

Name	Default	Description
3-4 xGMI Link Max Speed	Auto	• 12 Gbps • 16 Gbps • 17 Gbps • 18 Gbps • 20 Gbps • 22 Gbps • 23 Gbps • 24 Gbps • 25 Gbps/Auto • 26 Gbps • 27 Gbps • 28 Gbps • 30 Gbps • 32 Gbps 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 x8. • **1:** 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.

Table 3-4: Infinity Fabric BIOS settings

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 3-4: Infinity Fabric BIOS settings

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

Name	Default	Description
Local APIC Mode	Auto(0x02)	• **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. • **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):** Forces legacy xAPIC mode. • **X2ApicMode (0x02):** Forces x2APIC mode independent of thread count.
PCIe Speed PMM Control	Auto	• **0:** Dynamic link speed determined by power management functionality. • **1:** Static Target Link Speed (Gen4); sets the maximum idle link speed to 16 GT/s. • **Auto/2:** Static Target Link Speed (Gen5); sets the maximum idle link speed to 32 GT/s, thereby disabling the feature).
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.

Table 3-5: PCIe, I/O, security, and virtualization BIOS settings

TSME	Auto	• **Auto/Disabled:** Disables transparent secure memory encryption. • **Enabled:** Enables transparent secure memory encryption.
SEV	Disabled	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.
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	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 3-5: PCIe, I/O, security, and virtualization BIOS settings

3.6 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.

Chapter 4

BIOS Settings for RDBMS

Table 4-1 describes the BIOS options that most impact performance for common RDBMS systems. Please see the *BIOS & Workload Tuning Guide for AMD EPYC™ 9004 Series Processors* (available from AMD EPYC Tuning Guides) for additional BIOS setting information.

The **NUMA Nodes per Socket** setting is a trade-off between minimizing local memory latency for NUMA- aware or highly parallelizable workloads versus maximizing per-core memory bandwidth for non-NUMA friendly workloads. The greater the number of memory channels interleaved, the higher the memory throughput is for certain memory operations. Increasing the number of NUMA nodes per socket will decrease the memory channels per NUMA node and decrease both throughput and latency. There are multiple L3 caches in the NUMA domain, even when NPS=4. The **ACPI SRAT L3 Cache as NUMA Domain** setting further splits the NUMA domains such that each L3 cache has its own domain. Enabling this option uses the **NUMA Nodes per Socket** setting to determine memory interleaving granularity.

4.1 BIOS Settings for Maximizing Performance

Name	Value	Description
TSME	Disabled	Transparent Secure Memory Encryption (TSME) provides hardware memory encryption of all data stored on system DIMMs and increases memory latency by 5-7ns.
Determinism Control	Manual	Enables the **Determinism Slider** control.
Determinism Enables	Disable performance determinism	Ensures maximum performance for each CPU in a large population of identically-configured CPUs by only throttling CPUs when they reach the same TDP.
cTDP Control	Manual	Setting **Configurable Thermal Design Power** (TDP) to **Manual** allows you to modify the platform CPU cooling limit.
cTDP	OPN Max	Set TDP in watts.
PPT Control	Manual	Setting **PPL** to **Manual** allows you to modify the CPU Power Dissipation Limit.
PPT	OPN Max	Set PPT in watts.
SMT Control	Auto	Enables Symmetric Multithreading (SMT), which allows two hardware threads per core. You must enable AMD x2APIC with support more than 383 threads if you are using a system with dual 96-core AMD EPYC 9004 Series Processors. If you are running dual 96-core processors and your OS does not support AMD x2APIC, then you must disable SMT.

Table 4-1: Recommended BIOS settings

NUMA Node per Socket (NPS)	Varies by RDBMS and workload	Determines the number of NUMA nodes between which to split the memory channels. Higher numbers reduce memory channels per NUMA node and lower both memory throughput and latency.
ACPI SRAT L3 Cache as NUMA Domain	Varies by RDBMS and workload	Specifies whether or not to report each L3 cache to the OS as a NUMA domain. Allowing processes that use the same data to be scheduled on the set of CPUs that share an L3 cache increases the L3 cache hit rate.
Power Profile Selection	Maximum IO Performance Mode	Select **Maximum IO Performance Mode** to enable the maximum I/O performance if the number of NUMA nodes per socket is set to either 1 or 4.
X3D	Enabled	In an AMD EPYC 9004 processor with AMD 3D V-Cache technology, this option enables the AMD 3D V-Cache module in the CCD to increase the total size of the L3 cache memory to 96MB. Disabling this option reduces the L3 cache in the CCD to 32MB. AMD recommends enabling this option, if available and if doing so benefits your workload after testing. This option is only available on AMD EPYC 9004 Series Processors with AMD 3D V-Cache.

Table 4-1: Recommended BIOS settings (Continued)

4.2 Memory

RDBMS systems generally consume large amounts of memory. AMD EPYC 9004 Series Processors have 12 memory channels per CPU socket to increase both the total addressable memory per socket and the memory throughput per socket. Enabling large pages in the OS can improve system performance by reducing the amount of system resources required to access page table entries. The size of the large pages varies from platform to platform. You can enable large pages on Linux based systems in one of the following two ways:

- Explicitly set the `vm.nr_hugepages` parameter in `/etc/sysctl.conf`, as described in "Linux Tuning Considerations" on page 27.
- Implicitly use transparent huge pages.

Relational databases on Linux prefer transparent huge pages, while most other vendors prefer explicitly defined huge pages.

4.3 Network

Tuning and configuring the RDBMS system's network is crucial for quickly getting information in and out of the RDBMS. Tuning also prevents the network from consuming all system resources. This is less critical in smaller systems that run transactions and queries directly on the local server. Please see the *Linux® Network Tuning Guide for AMD EPYC™ 9004 Series Processors* or *Windows® Network Tuning Guide for AMD EPYC™ 9004 Series Processors* (available from AMD EPYC Tuning Guides), as appropriate, for OS-specific network tuning instructions.

4.4 Storage

AMD EPYC 9004 Series Processors support PCIe® Gen 5 connections with double the I/O bandwidth of PCIe Gen 4. All normal RDMBS I/O tuning rules still apply, such as separating the log devices from data devices and splitting I/O across the remaining devices, controllers, PCIe bus, etc. Please see the I/O-related information in the *MongoDB® Tuning Guide for AMD EPYC™ 9004 Series Processors* (available from AMD EPYC Tuning Guides) for additional instructions.

Chapter 5 MySQL - Additional Configuration Settings

This chapter contains MySQL tuning recommendations. You can download MySQL from https://support.oracle.com/* and install it as instructed in https://dev.mysql.com/doc/refman/8.0/en/installing.html*. Be sure to download and install the latest version. MySQL database configurations are defined in `/etc/my.cnf`. Table 5-1 lists the tunable parameters. Please also see "Linux Tuning Considerations" on page 27 for additional Linux tuning considerations.

5.1 Tuning my.cnf Parameters

Name	Value	Description
datadir	Directory to store database data and log files	This is the database directory mounted on the RAID volume created during initial setup.
innodb_buffer_pool_size	70%-80% of total RAM	Based on the amount of RAM available
Innodb_buffer_pool_instances	64	Number of regions the InnoDB buffer pool is divided into. Reduces contention because different threads write to cached pages.
max_connections	4000	Maximum number of allowed connections (plus 1 for SUPER account).
Innodb_log_buffer_size	1G	Enables large transactions to run without needing to write the log to disk before the transactions commit.
Innodb_log_file_size	4G	Controls the checkpoint activity. Larger log files have fewer checkpoints but also make crash recovery take longer.
table_open_cache	8000	Number of open tables for all threads.
innodb_flush_method	O_DIRECT	Defines the method used to flush data to the InnoDB log and data files.
Innodb_doublewrite	0	Disables the doublewrite buffer, which boosts performance.
large_pages	ON	Enables huge pages in the database.

Table 5-1: MySQL my.cnf parameter settings

```
[mysqld]
datadir=/var/lib/mysql
pid-file=/var/run/mysqld/mysqld.pid
datadir = /data/mysql
default_authentication_plugin=mysql_native_password
user            = mysql
bind-address            = [Server IP Address]
```

```
key_buffer_size         = 16M
myisam-recover-options  = BACKUP
log_error = /log/error.log
max_binlog_size   = 100M
innodb_buffer_pool_size=500G
innodb_buffer_pool_instances=64
innodb_log_buffer_size=1G
innodb_log_file_size=4G
max_connections=4000
table_open_cache=8000
table_open_cache_instances=16
back_log=1500
default_password_lifetime=0
ssl=0
performance_schema=OFF
max_prepared_stmt_count=128000
skip_log_bin=1
character_set_server=latin1
collation_server=latin1_swedish_ci
transaction_isolation=REPEATABLE-READ
innodb_flush_method=O_DIRECT
innodb_doublewrite=0
innodb_thread_concurrency=0
innodb_flush_log_at_trx_commit=0
innodb_max_dirty_pages_pct=90
innodb_max_dirty_pages_pct_lwm=10
innodb_use_native_aio=1
innodb_stats_persistent=1
innodb_spin_wait_delay=6
innodb_max_purge_lag_delay=300000
innodb_max_purge_lag=0
innodb_checksum_algorithm=none
innodb_io_capacity=4000
innodb_io_capacity_max=20000
innodb_lru_scan_depth=9000
innodb_change_buffering=none
innodb_read_only=0
innodb_page_cleaners=4
innodb_undo_log_truncate=off
innodb_file_per_table
join_buffer_size = 4096M
innodb_log_files_in_group=32
innodb_open_files=4000
innodb_adaptive_flushing=1
innodb_flush_neighbors=0
innodb_read_io_threads=16
innodb_write_io_threads=16
innodb_purge_threads=4
innodb_adaptive_hash_index=0
```

Chapter 6 MariaDB - Additional Configuration Settings

This chapter contain MariaDB tuning recommendations. You can download MariaDB from https://downloads.mariadb.org/mariadb/10.5.8/* and install it as instructed in https://mariadb.com/docs/release-notes/mariadb- enterprise-server-10-5-8-5-release-notes/#installation*. Be sure to download and install the latest version. MariaDB and MySQL have several installation and configuration similarities. MariaDB database configurations are defined in /etc/my.cnf. Table 6-1 lists the tunable parameters. Please also see "Linux Tuning Considerations" on page 27 for additional Linux tuning considerations.

6.1 Tuning my.cnf Parameters

Name	Value	Description
datadir	Directory to store database data and log files	This is the database directory mounted on the RAID volume created during initial setup.
innodb_buffer_pool_size	70%-80% of total RAM	Based on the amount of RAM available
Innodb_buffer_pool_instances	64	Number of regions the InnoDB buffer pool is divided into. Reduces contention because different threads write to cached pages.
max_connections	4000	Maximum number of allowed connections (plus 1 for SUPER account).
Innodb_log_buffer_size	1G	Enables large transactions to run without needing to write the log to disk before the transactions commit.
Innodb_log_file_size	4G	Controls the checkpoint activity. Larger log files have fewer checkpoints but also make crash recovery take longer.
table_open_cache	8000	Number of open tables for all threads.
innodb_flush_method	O_DIRECT	Defines the method used to flush data to the InnoDB log and data files.
Innodb_doublewrite	0	Disables the doublewrite buffer, which boosts performance.
large_pages	ON	Enables huge pages in the database.

Table 6-1: MySQL my.cnf parameter settings

Note: Consider looking at the MariaDB Comparison Table to verify that you are using the correct MariaDB (not MySQL) values.

```
[mysqld]
datadir=/var/lib/mysql
pid-file=/var/run/mysqld/mysqld.pid
datadir = /data/mysql
default_authentication_plugin=mysql_native_password
user            = mysql
bind-address            = [Server IP Address]
key_buffer_size         = 16M
myisam-recover-options  = BACKUP
log_error = /log/error.log
max_binlog_size   = 100M
innodb_buffer_pool_size=500G
innodb_buffer_pool_instances=64
innodb_log_buffer_size=1G
innodb_log_file_size=4G
max_connections=4000
table_open_cache=8000
table_open_cache_instances=16
back_log=1500
default_password_lifetime=0
ssl=0
performance_schema=OFF
max_prepared_stmt_count=128000
skip_log_bin=1
character_set_server=latin1
collation_server=latin1_swedish_ci
transaction_isolation=REPEATABLE-READ
innodb_flush_method=O_DIRECT
innodb_doublewrite=0
innodb_thread_concurrency=0
innodb_flush_log_at_trx_commit=0
innodb_max_dirty_pages_pct=90
innodb_max_dirty_pages_pct_lwm=10
innodb_use_native_aio=1
innodb_stats_persistent=1
innodb_spin_wait_delay=6
innodb_max_purge_lag_delay=300000
innodb_max_purge_lag=0
innodb_checksum_algorithm=none
innodb_io_capacity=4000
innodb_io_capacity_max=20000
innodb_lru_scan_depth=9000
innodb_change_buffering=none
innodb_read_only=0
innodb_page_cleaners=4
innodb_undo_log_truncate=off
innodb_file_per_table
join_buffer_size = 4096M
innodb_log_files_in_group=32
innodb_open_files=4000
innodb_adaptive_flushing=1
innodb_flush_neighbors=0
innodb_read_io_threads=16
innodb_write_io_threads=16
innodb_purge_threads=4
innodb_adaptive_hash_index=0
```

Chapter 7 Resources

7.1 Linux Tuning Considerations

7.1.1 Storage

Storage technology has evolved rapidly. Today's NVMe drives offer significant performance boosts compared to previous technologies such as SATA. HDDs are cheaper and offer more storage space, but SSDs are faster, lighter, more durable, and use less energy. Your needs will dictate the correct type(s) of storage. In general:

- **Local storage:** Use HDD, SATA/SAS SSDs, or NVMe for:
  - Log, data, and index requirements
  - RAID (HW and SW)
- **Shared Storage:** Use NAS or SAN for:
  - Log, Data, Index requirements
  - RAID (HW and SW)

7.1.1.1 MDADM Striped Partitions

This example shows you how to use `mdadm` to create RAID o striped drives.

Note: Note: Change the device name from `/dev/nvme[0-n]n1` *to* `/dev/sdb` *or* `/dev/sdc` *sequences accordingly if you are not using NVMe drives. Execute the command* `lsblk -t` *to see what devices you're using. If you are using NVMe drives, then you may execute the command* `nvme list` *to get the NVMe device names.*

```
# Enter the command below to create RAID with n drives as md0
# mdadm --create /dev/md0 --level=0 --raid-devices=n /dev/nvme[0-{n-1}]n1p1
```

1. After the command completes,

```
# mkfs.xfs /dev/md0
# UUID=`blkid -s UUID -o value /dev/md0`
# echo "UUID=$UUID /YourMountPoint xfs discard,defaults,nofail 0 2" >> /etc/fstab
```

2. Mount the created `/dev/md0` onto the data directory.

```
# mkdir /YourMountPoint # mount -a
# df -hT (command will show you the mount point)
/dev/md0 xfs      1.8T 1.2T 519G 69% /YourMountPoint
```

7.1.1.2 Linux Logical Volume Manager (LVM) Striped Partitions

This example shows you how to use Linux LVM to stripe storage devices for the MySQL database.

Note: Change the device name from `/dev/nvme[0-n]n1` *to* `/dev/sdb or /dev/sdc` *sequences accordingly if you are not using NVMe drives. Execute the command* `lsblk -t` *to see what devices you're using. If you are using NVMe drives, then you may execute the commnad* `nvme list` *to get the NVMe device names.*

```
===== Start =====
#!/bin/bash
# This script takes 4 NVMe drives as storage input and creates LVM. You could # modify the
number of devices based on your requirements.
#
for i in {1..4} do

     echo -e "Parting Labeling of /dev/nvme${i}n1"
     parted /dev/nvme${i}n1 mklabel gpt
     parted -a optimal /dev/nvme${i}n1 mkpart primary 0% 100%

done

pvcreate /dev/nvme1n1p1 /dev/nvme2n1p1 /dev/nvme3n1p1 /dev/nvme4n1p1
pvdisplay
vgcreate mysqldb-vol-vg /dev/nvme1n1p1 /dev/nvme2n1p1 /dev/nvme3n1p1 /dev/nvme4n1p1
vgdisplay
sleep 10
lvcreate -L 100G -i2 -I64 -n mysqlmysqldb-wal-strpd-lv mysqlmysqldb-
vol-vg lvcreate -L 100G -i2 -I64 -n mysqlmysqldb-bkp-strpd-lv
mysqldb-vol-vg lvcreate -L 1450G -i2 -I64 -n mysqldb-data-strpd-lv
mysqldb-vol-vg lvdisplay
ls -l /dev/mysqldb-bkp-strpd-lv
mkfs -t xfs -f -i size=2048 /dev/mysqldb-vol-vg/mysqldb-bkp-strpd-lv
mkfs -t xfs -f -i size=2048 /dev/mysqldb-vol-vg/mysqldb-wal-strpd-lv
mkfs -t xfs -f -i size=2048 /dev/mysqldb-vol-vg/mysqldb-data-strpd-lv
blkid /dev/mysqldb-vol-vg/mysqldb-data-strpd-lv
# blkid /dev/mysqldb-vol-vg/mysqldb-wal-strpd-lv
blkid /dev/mysqldb-vol-vg/mysqldb-bkp-strpd-lv

===== End =====

# Check and validate LVM Creation # vgs

VG               #PV   #LV   #SN     Attr       VSize      VFree
mysqldb-vol-vg     2     3     0     wz--n-     <1.64T     <26.38G

#lvs

LV                     VG         Attr        LSize     Pool Origin Data% Meta%
Move Log Cpy%Sync Convert
mysqldb-bkp-strpd-lv mysqldb-vol-vg  -wi-ao----
100.00g mysqldb-data-strpd-lv mysqldb-vol-vg -wi-ao
----------------------------------------------<1.42t
mysqldb-wal-strpd-lv mysqldb-vol-vg -wi-ao---- 100.00g

# pvs

PV             VG            Fmt  Attr PSize    Free
/dev/nvme3n1p1 mysqldb-vol-vg lvm2 a--  <838.19g <13.19g
/dev/nvme4n1p1 mysqldb-vol-vg lvm2 a--  <838.19g <13.19g

Create Directories
# mkdir -p /mysqldata /mysqllogs /mysqlbkup chown -R mysqlql:mysqlql
```

```
/mysql*
# Get the UUID for all 3 logical volumes so that it could be mounted accordingly
# blkid /dev/msqldb-vol-vg/msqldb-bkp-strpd-lv
/dev/msqldb-vol-vg/msqldb-bkp-strpd-lv: UUID="9003eda5-9ded-418d-a4a8-a4f506506a6f"
BLOCK_SIZE="512" TYPE="xfs"
# blkid /dev/msqldb-vol-vg/msqldb-data-strpd-lv
/dev/msqldb-vol-vg/msqldb-data-strpd-lv: UUID="8000adf7-6e7d-92ea-f497-bfc5089841d9"
BLOCK_SIZE="512" TYPE="xfs"
# blkid /dev/msqldb-vol-vg/msqldb-wal-strpd-lv
/dev/msqldb-vol-vg/msqldb-data-strpd-lv: UUID="6784ce87-e77a-83af-a3bb-a9d8054187ab"
BLOCK_SIZE="512" TYPE="xfs"
```

Add the following UUID to `/etc/fstab` for automatic mounting from every boot. You must add the UUID values from the previous 3 logical volumes to `/etc/fstab`.

```
UUID="9003eda5-9ded-418d-a4a8-a4f506506a6f"     /mysqldata      xfs     noatime     0
0
UUID="8000adf7-6e7d-92ea-f497-bfc5089841d9"     /mysqllogs      xfs     noatime     0
0
UUID="6784ce87-e77a-83af-a3bb-a9d8054187ab"     /mysqlbkp       xfs     noatime     0
0
```

7.1.2 Linux Configuration

7.1.2.1 Linux Huge Pages

Many modern Linux distributions enable Transparent Huge Pages (THP) by default. With THP enabled, the kernel tries to allocate memory in large chunks (usually 2MB) instead of 4K, which can improve performance by reducing the number of pages the CPU must track. MySQL performance benefits from enabling huge pages.

AMD EPYC architecture supports HugePages up to 1GB. This sample `lscpu` output shows the current `pse` and `pdpe1gb` supporting Huge Pages from 2MB to 1GB, respectively.

```
Architecture:                    x86_64
CPU op-mode(s):                  32-bit, 64-bit
Address sizes:                   52 bits physical, 57 bits virtual
Byte Order:                      Little Endian
CPU(s):                          384
On-line CPU(s) list:             0-383
Vendor ID:                       AuthenticAMD
Model name:                      AMD EPYC 9654 96-Core Processor
CPU family:                      25
Model:                           17
Thread(s) per core:              2
Core(s) per socket:              96
Socket(s):                       2
Stepping:                        1
Frequency boost:                 enabled
CPU max MHz:                     2400.0000
CPU min MHz:                     1500.0000
BogoMIPS:                        4800.15
Flags:                           fpu vme de pse tsc msr pae mce cx8 apic sep mtrr pge mca
cmov pat pse36 clflush mmx fxsr sse sse2 ht syscall nx mmxext fxsr_opt pdpe1gb rdtscp lm
constant_tsc rep_good nopl nonstop_tsc cpuid extd_apicid aperfmperf rapl pni pclmulqdq
monitor ssse3 fma cx16 pcid sse4_1 sse4_2 x2apic movbe popcnt aes xsave avx f16c rdrand
lahf_lm cmp_legacy svm extapic cr8_legacy abm sse4a misalignsse 3dnowprefetch osvw ibs
skinit wdt tce topoext perfctr_core perfctr_nb bpext perfctr_llc mwaitx cpb cat_l3 cdp_l3
invpcid_single hw_pstate ssbd mba ibrs ibpb stibp vmmcall fsgsbase bmi1 avx2 smep bmi2 erms
```

```
invpcid cqm rdt_a avx512f avx512dq rdseed adx smap avx512ifma clflushopt clwb avx512cd
sha_ni avx512bw avx512vl xsaveopt xsavec xgetbv1 xsaves cqm_llc cqm_occup_llc
cqm_mbm_total cqm_mbm_local avx512_bf16 clzero irperf xsaveerptr rdpru wbnoinvd amd_ppin
arat npt lbrv svm_lock nrip_save tsc_scale vmcb_clean flushbyasid decodeassists
pausefilter pfthreshold avic v_vmsave_vmload vgif v_spec_ctrl avx512vbmi umip pku ospke
avx512_vbmi2 gfni vaes vpclmulqdq avx512_vnni avx512_bitalg avx512_vpopcntdq la57 rdpid
overflow_recov succor smca fsrm flush_l1d
Virtualization:                  AMD-V
L1d cache:                       6 MiB (192 instances)
L1i cache:                       6 MiB (192 instances)
L2 cache:                        192 MiB (192 instances)
L3 cache:                        768 MiB (24 instances)
NUMA node(s):                    24
NUMA node0 CPU(s):               0-7,192-199
NUMA node1 CPU(s):               24-31,216-223
NUMA node2 CPU(s):               48-55,240-247
NUMA node3 CPU(s):               72-79,264-271
NUMA node4 CPU(s):               8-15,200-207
NUMA node5 CPU(s):               32-39,224-231
NUMA node6 CPU(s):               56-63,248-255
NUMA node7 CPU(s):               80-87,272-279
NUMA node8 CPU(s):               16-23,208-215
NUMA node9 CPU(s):               40-47,232-239
NUMA node10 CPU(s):              64-71,256-263
NUMA node11 CPU(s):              88-95,280-287
NUMA node12 CPU(s):              96-103,288-295
NUMA node13 CPU(s):              120-127,312-319
NUMA node14 CPU(s):              144-151,336-343
NUMA node15 CPU(s):              168-175,360-367
NUMA node16 CPU(s):              104-111,296-303
NUMA node17 CPU(s):              128-135,320-327
NUMA node18 CPU(s):              152-159,344-351
NUMA node19 CPU(s):              176-183,368-375
NUMA node20 CPU(s):              112-119,304-311
NUMA node21 CPU(s):              136-143,328-335
NUMA node22 CPU(s):              160-167,352-359
NUMA node23 CPU(s):              184-191,376-383
Vulnerability Itlb multihit:     Not affected
Vulnerability L1tf:              Not affected
Vulnerability Mds:               Not affected
Vulnerability Meltdown:          Not affected
Vulnerability Mmio stale data:   Not affected
Vulnerability Retbleed:          Not affected
Vulnerability Spec store bypass: Mitigation; Speculative Store Bypass disabled via prctl
and seccomp
Vulnerability Spectre v1:        Mitigation; usercopy/swapgs barriers and __user pointer
sanitization
Vulnerability Spectre v2:        Mitigation; Retpolines, IBPB conditional, IBRS_FW, STIBP
always-on, RSB filling, PBRSB-eIBRS Not affected
Vulnerability Srbds:             Not affected
Vulnerability Tsx async abort:   Not affected
```

7.1.2.2 Required /etc/sysctl.conf Kernel Parameters

Name	Value	Description
kernel.numa_balancing	0	Disable Linux kernel auto NUMA balancing
vm.swappiness	User defined	Depends on database type and user requirements.

vm.nr_hugepages	User defined	Depends on database type and user requirements.
		Example:
		If `vm.nr_hugepages=131072`
		then
		
		Hugepagesize: The size of a hugepage (usually 2MB on the modern x86_64 based system). To check your system's Hugepagesize:
		`# cat /proc/meminfo | grep -i Hugepagesize`
		`Hugepagesize: 2048 kB`
		`vm.nr_hugepages=(131072 x 2MB) / 1024 = 256 GB`

Table 7-1: General sysctl.conf settings

Set `vm.nr_hugepages set`, then add the following configuration settings in the respective database config files, and then restart the database.

- **MySQL:** `large_pages=ON`

7.1.2.3 tuned-adm Profile

Setting the `tuned-adm` profile to `throughput-performance` normally generates optimal system I/O and memory throughput by configuring system I/O and memory throughput via the CPU governor, kernel scheduler granularity, disk read ahead, swappiness behavior, and dirty cache write back settings. See `usr/lib/tuned/throughput-performance/tuned.conf` for specific `tuned-adm` profile settings.

7.1.2.4 XFS Filesystem Mount Options

- **Noatime:** This option disables updating the metadata associated with files in the filesystem with an updated access time. This tracking is unnecessary because databases log their own accesses in their logs, making filesystem tracking redundant.
- **Nobarrier:** Disables the filesystem write barrier. Using a write barrier degrades I/O performance by flushing data more frequently.

7.2 For Additional Reading

- From AMD EPYC Tuning Guides:
  - *Overview of AMD EPYC™ 9004 Series Processors Architecture*
  - *Microsoft SQL Server® Tuning Guide for AMD EPYC™ 9004 Series Processors*
  - *Microsoft® Windows Tuning Guide for AMD EPYC™ 9004 Series Processors*
  - *RedHat Enterprise Linux® Tuning Guide AMD EPYC™ 9004 Series Processors*

- MariaDB 10.5 documentation*
- MySQL 8.0 documentation.

This page intentionally left blank.

Chapter 8

Processor Identification

Figure 8-1 shows the processor naming convention for AMD EPYC 9004 Series Processors and how to use this convention to identify particular processors models:

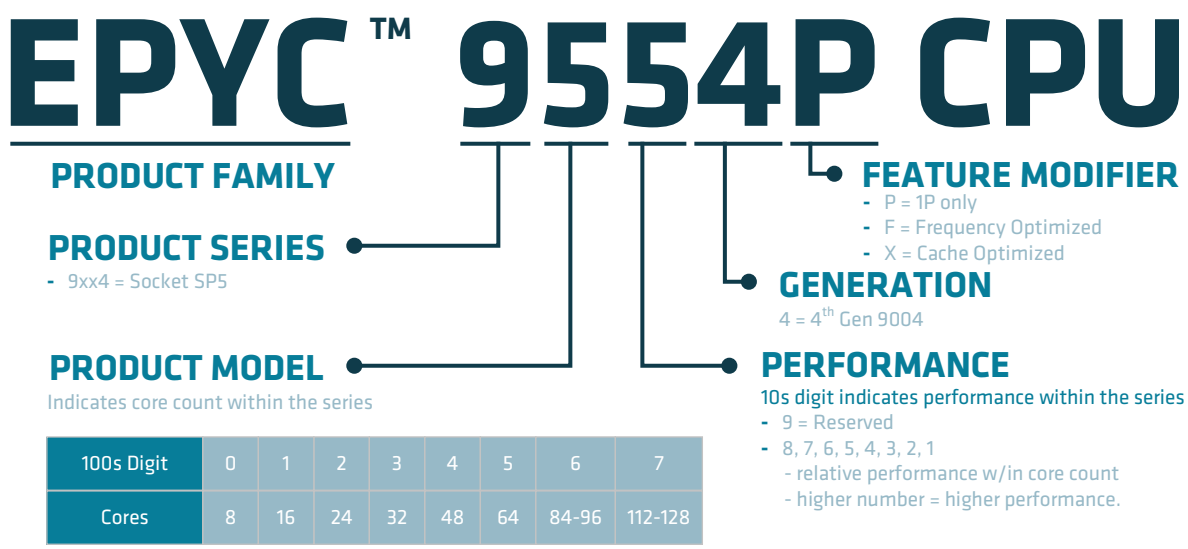

Figure 8-1: AMD EPYC SoC naming convention

8.1 CPUID Instruction

Software uses the `CPUID` instruction (`Fn0000_0001_EAX`) to identify the processor and will return the following values:

- **Family:** 19h identifies the "Zen 4" architecture
- **Model:** Varies with product. For example, EPYC Family 19h, Model 10h corresponds to an "A" part "Zen 4" CPU.
  - **91xx-96xx (including "X" OPNs):** Family 19h, Model 10-1F
  - **97xx:** Family 19h, Model A0-AF
- **Stepping:** May be used to further identify minor design changes

For example, `CPUID` values for Family, Model, and Stepping (decimal) of 25, 17, 1 correspond to a "B1" part "Zen 4" CPU.

8.2 New Software-Visible Features

AMD EPYC 9004 Series Processors introduce several new features that enhance performance, ISA updates, provide additional security features, and improve system reliability and availability. Some of the new features include:

- 5-level Paging
- AVX-512 instructions on a 256-byte datapath, including BFLOAT16 and VNNI support.
- Fast Short Rep STOSB and Rep CMPSB

Not all operating systems or hypervisors support all features. Please refer to your OS or hypervisor documentation for specific releases to identify support for these features.

Please also see the latest version of the *AMD64 Architecture Programmer's Manuals* or *Processor Programming Reference (PPR) for AMD Family 19h*.

8.2.1 AVX-512

AVX-512 is a set of individual instructions supporting 512-bit register-width data (i.e., single instruction, multiple data [SIMD]) operations. AMD EPYC 9004 Series Processors implement AVX 512 by "double-pumping" 256-bit-wide registers. AMD's AVX-512 design uses the same 256-bit data path that exists throughout the Zen4 core and enables the two parts to execute on sequential clock cycles. This means that running AVX-512 instructions on AMD EPYC 9004 Series will cause neither drops on effective frequencies nor increased power consumption. On the contrary, many workloads run more energy-efficiently on AVX-512 than on AVX-256P.

Other AVX-512 support includes:

- Vectorized Neural Network Instruction (VNNI) instructions that are used in deep learning models and accelerate neural network inferences by providing hardware support for convolution operations.
- Brain Floating Point 16-bit (BFLOAT16) numeric format. This format is used in Machine Learning applications that require high performance but must also conserve memory and bandwidth. BFLOAT16 support doubles the number of SIMD operands over 32-bit single precision FP, allowing twice the amount of data to be processed using the same memory bandwidth. BFLOAT16 values mantissa dynamic range at the expense of one radix point.