AMD EPYC

TUNING GUIDE
AMD EPYC 9004

Microsoft® SQL Server

Publication	58007
Revision	1.4
Issue Date	June, 2024

Date	Version	Changes
July, 2022	0.1	Initial NDA partner release.
Sep, 2022	0.2	Updated BIOS information
Nov, 2022	1.0	Initial public release
Dec, 2022	1.1	Minor errata corrections
Mar, 2022	1.2	Added 97xx OPN and AMD 3D V-Cache™ technology information
June, 2023	1.3	Second public release
June, 2024	1.4	Update to reflect MS SQL Server 2022

Audience

This tuning guide is intended for a technical audience such as MSSQL® 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

Bryon Georgson 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

Chapter

1

Introduction

This tuning guide provides detailed descriptions of parameters that can optimize performance on servers built with AMD EPYC™ 9004 Series processors. The default configurations on hardware and BIOS from different OEM vendors may not provide the best possible performance on all OS platforms and for all workloads. To enable optimization on a per platform and workload level, this guide calls out:

- BIOS settings that can impact performance
- Hardware configuration best practices
- Supported OS versions and configurations
- SQL Server performance tuning
- Workload-specific BIOS and OS settings for OLTP and DSS workloads.

This page intentionally left blank.

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.

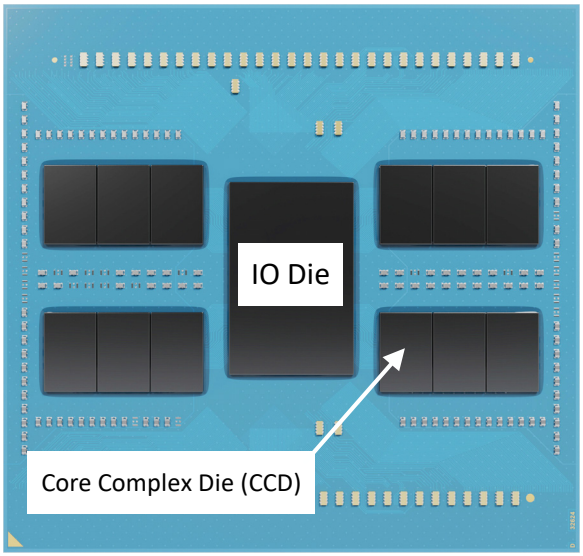

Figure 2-1: AMD EPYC 9004 configuration with 12 Core Complex Dies (CCD) surrounding a central I/O Die (IOD)

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.

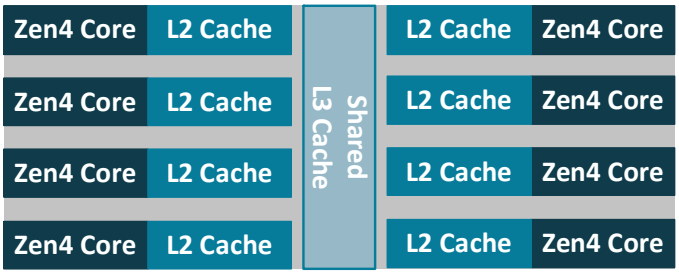

Figure 2-2: Top view of 8 compute cores sharing an L3 cache (91xx-96xx models)

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.

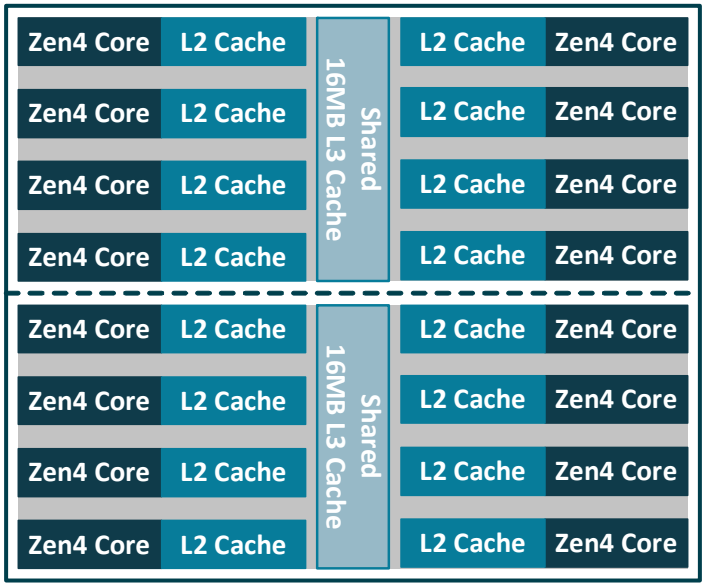

Figure 2-3: 2 CCXs in a single 4th Gen AMD EPYC 97xx CCD

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.

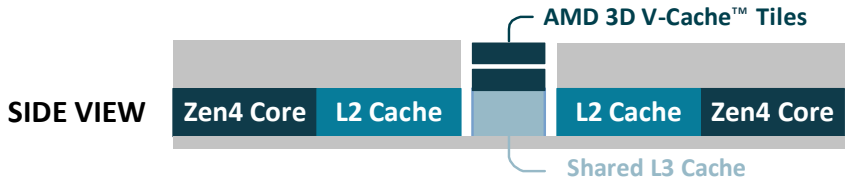

Figure 2-4: Side view of vertically-stacked central L3 SRAM tiles

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.

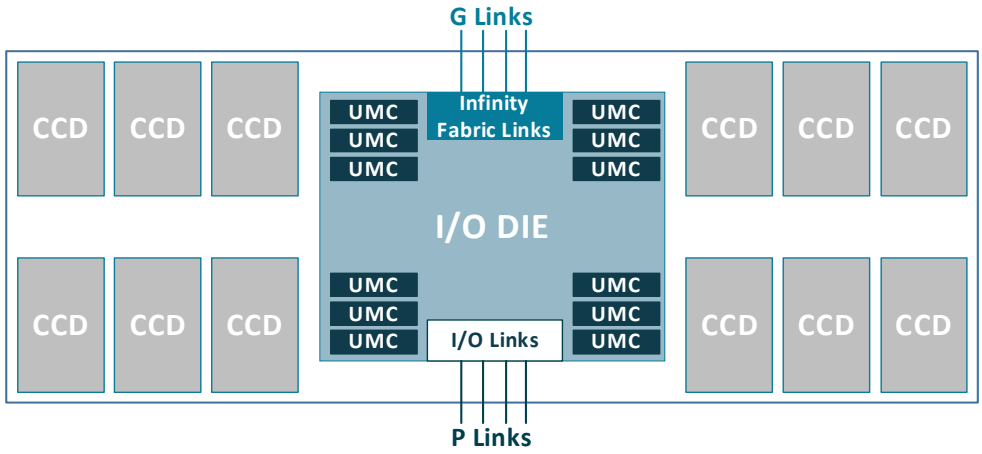

Figure 2-5: AMD EPYC 9654 processor internals interconnect via AMD Infinity Fabric (12 CCD processor shown)

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.

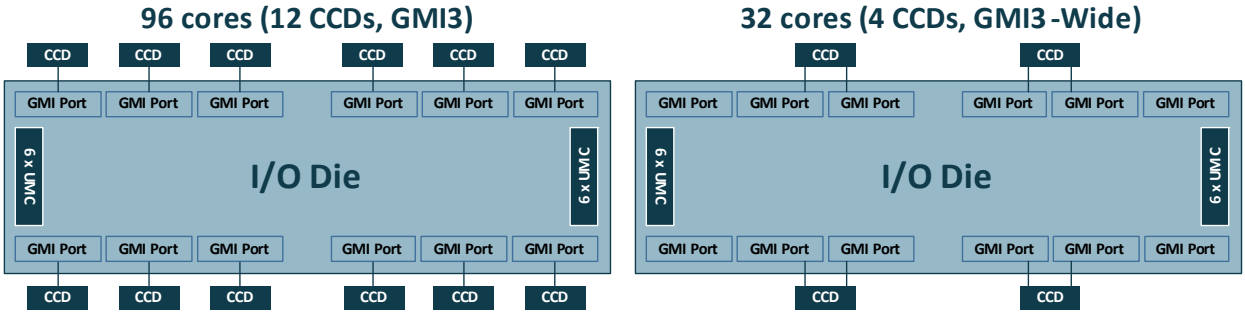

Figure 2-6: Standard vs. Wide GMI links

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.

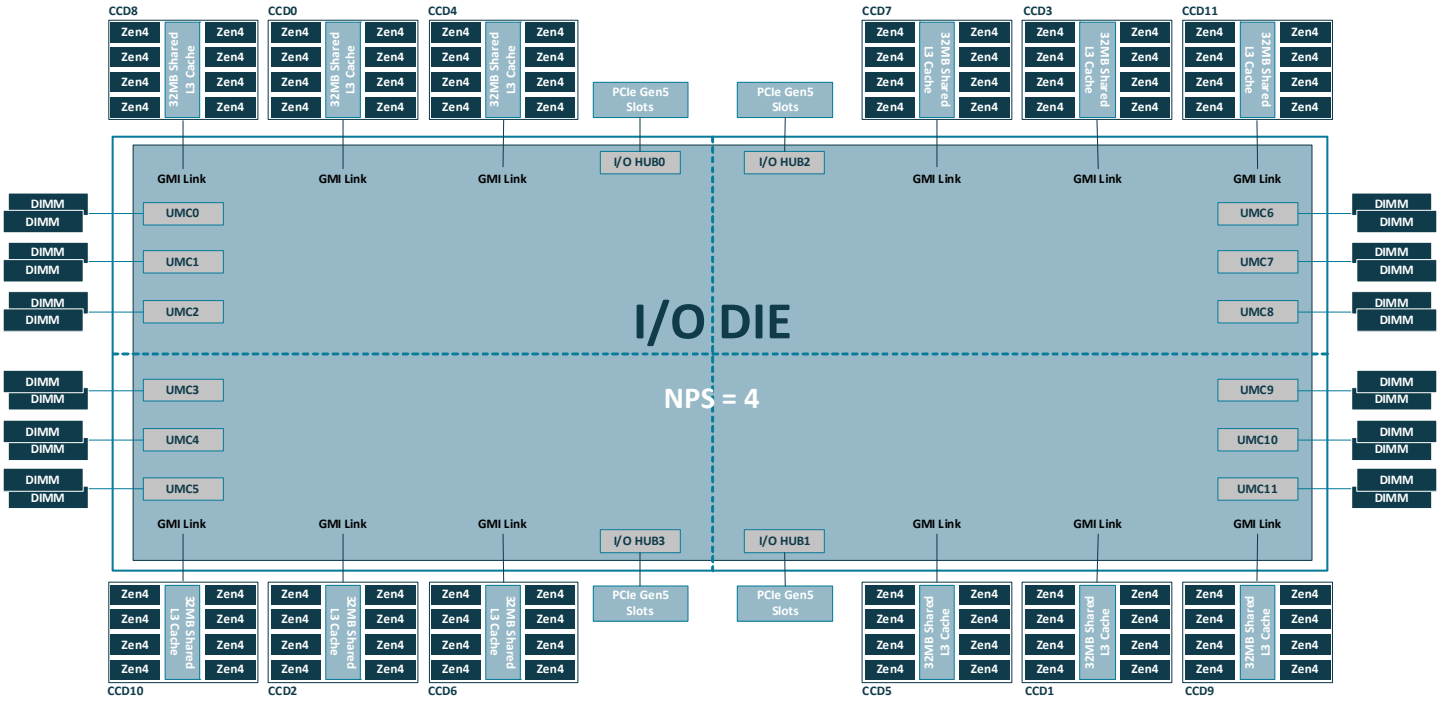

Figure 2-7: The AMD EPYC 9004 SoC consists of up to 12 CCDs and a central IOD for 91xx-96xx models, including "X" OPNs

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.

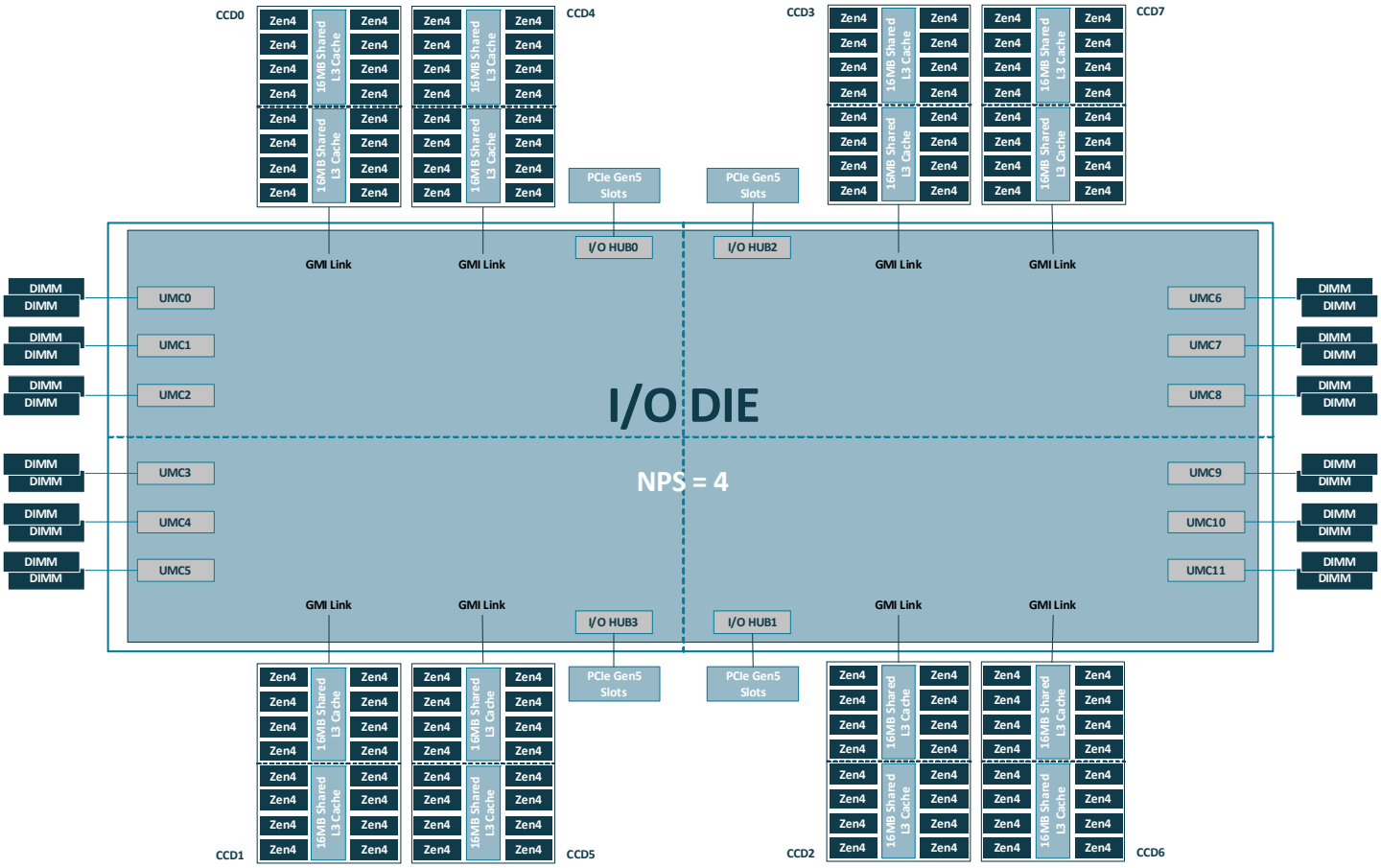

Figure 2-8: The AMD EPYC 9004 System on Chip (SoC) consists of up to 8 CCDs and a central IOD for 97xx models

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.

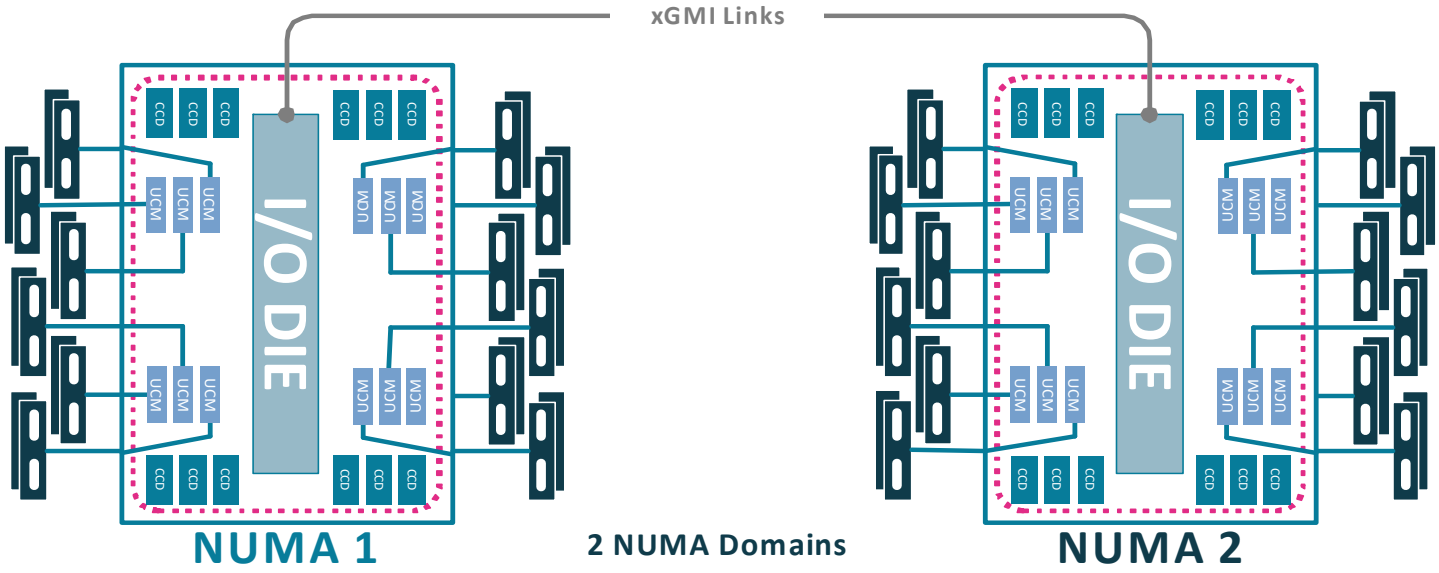

Figure 2-9: Two EPYC 9004 Processors connect through 4 xGMI links (NPS1)

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 Hardware Configuration

This chapter discusses the hardware components to consider when planning a SQL Server system.

4.1 CPU Selection Guidelines

Selecting the right processor depends on the following factors:

- **Licensing costs:** Microsoft SQL Server is licensed by the core. Processors with lower core counts and higher frequencies are therefore ideal.
- **Database size, IOPS, and throughput:** Single-socket systems are better for memory latency dependent applications. Dual-socket systems with 8 memory channels per socket are better for applications that depend on memory bandwidth.
- **Inter-process data sharing:** It may be better to use AMD EPYC 9004 Series Processors with larger caches when the SQL Server threads either have little data sharing between them or those that do share data can be isolated to a single Core Complex.

4.2 Recommended CPUs

Selecting the correct CPU is an important part of achieving optimal database application performance. Table 4-1 provides CPU selection guidelines along with the minimum memory requirements for SQL Server running Decision Support Systems (DSS) and Online Transaction Processing (OLTP) applications. You can choose other AMD EPYC processors not listed in Table 4-1 if needed for your specific requirements.

Workload					
Size	**Type**	**Size/# of Users**	**Cores**	**Processor**	**Memory**
Small	DSS	• Up to 300GB • Up to 5 concurrent users	16	AMD EPYC 9004 16-core	128GB
Medium	DSS	• 300GB-1TB • Up to 10 concurrent users	32	AMD EPYC 9004 32-core	256GB
Large	DSS	• 1TB to 3TB • Up to 20 concurrent users	96	AMD EPYC 9004 96-core	512GB
Very Large	DSS	• >3TB • >20 concurrent users	192	2 x AMD EPYC 9004 96-core	1TB-4TB

Table 4-1: Recommended AMD EPYC CPUs for various SQL Server applications

Small	OLTP	• Up to 100GB • Up to 50 concurrent users	16	AMD EPYC 9004 16-core	256GB
Medium	OLTP	• 100GB-300GB • Up to 100 concurrent users	32	AMD EPYC 9004 32-core	512GB
Large	OLTP	• 300GB-1TB • Up to 200 concurrent users	64	AMD EPYC 9004 64-core	1TB-2TB
Very Large	OLTP	• >1TB • >200 concurrent users	96	AMD EPYC 9004 96-core	1TB-4TB
Clouds	Mixed	• Private clouds • PaaS • SaaS	192	2 c AMD EPYC 9004 96-core	Size based on unique requirements

Table 4-1: Recommended AMD EPYC CPUs for various SQL Server applications (Continued)

For cloud deployments, use 8, 16, 32, or 64 vCPU cores with at least 8 GB memory per vCPU instance VM with appropriate storage attached for optimum performance.

4.3 Memory

SQL Server systems consume a lot of memory. AMD EPYC 9004 Series Processors include 12 memory channels per CPU socket, which increases both the total addressable memory and memory throughput per socket. Using large pages in the OS system can improve system performance by reducing the amount of system resources to access page table entries. The size of large pages varies from platform to platform. Most Linux distributions enable transparent hugepages by default, and SQL Server will use them if enabled.

4.4 Network

Tuning and configuring an RDBMS network critical for rapidly transferring information in and out of the database. Tuning also ensures that the network does not overtake all system resources. This is less critical on smaller systems that run transactions and queries directly on the local server this will be a less critical component in the tuning process. Review the *Windows® Network Tuning Guide for AMD EPYC™ 9004 Series Processors* or *Linux® Network Tuning Guide for AMD EPYC™ 9004 Series Processors* (available from AMD EPYC Tuning Guides) for OS-specific network tuning information.

Configure the number of RSS queues and set the processor affinity of each network port to its own set of CPUs that are within a Core/Cache Die (CCD). Enable jumbo frames in and OLTP environment for better network throughput. Adjust flow control settings to be identical on the switch and server ports.

On Linux systems, enable Rx Tx buffer size as follows:

```
#To check the pre-set maximums, run the command, example NIC name used here is:
"eth0"
     ethtool -g eth0
     #command to set both the rx(recieve) and tx (transmit) buffer size to 4 KB.
     ethtool -G eth0 rx 4096 tx 4096
     #command to check the value is properly configured is:
     ethtool -g eth0
```

To reduce number of RSS queues to #cores in a CCD:

```
ethtool -l eth0
ethtool -L eth0 <queue type> 16
ethtool -l eth0
```

On Windows systems, use PowerShell cmdlets to set Rx Tx buffer size:

```
# list all the NET adapters
Get-NetAdapter
# Using $rssNumaNode socket
Set-NetAdapterRss -Name <network adapter> -Enable 1 -BaseProcessorGroup <processor
group that nic is attached to> -MaxProcessorGroup <processor group that nic is
attached to> -NumberOfReceiveQueues 16 -BaseProcessorNumber 0 -MaxProcessorNumber 15
-MaxProcessors 16 -NumaNode -Profile NUMAStatic
```

4.5 Storage

AMD EPYC 9004 Series Processors support PCIe® 5.0 with double the I/O bandwidth of PCIe 4.0. All normal RDBMS I/O layer tuning rules still apply, such as:

- Using RAID devices when possible.
- Separating log and data devices.
- Splitting I/O across the remaining devices, controllers, PCIe bus, etc.

Either mirror database logs or use RAID 10. You may also want to RAID the data, depending on your fault tolerance requirements. Hardware RAID is always more efficient than software RAID.

Storage technology has evolved rapidly over the years. NVMe drives offer significantly faster performance than SATA/SAS. HDDs are cheaper for systems that require substantial amounts of storage, but SSDs are faster, lighter, more durable, and use less energy. Your needs will dictate the best storage for your system.

- Local storage (HDD, SATA/SAS SSDs, NVMe):
  - Log, data, index requirements
  - RAID (either HW or SW)
- Shared Storage (NAS, SAN):
  - Log, data, index requirements
  - RAID (either HW or SW)

This page intentionally left blank.

Chapter 5

BIOS Settings

Table 5-1 describes 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 tuning information.

Name	Value	Description
TSME	Disabled	Transparent Secure Memory Encryption (TSME) provides hardware memory encryption of all data stored on system DIMMs at a cost of 5-7ns of memory latency.
Determinism Control	Enabled	Enables the **Determinism Slider** control.
Determinism Slider	Power	Ensures maximum CPU performance by only throttling the CPUs when they reach the cTDP.
TDP Control	Manual	Enable setting a custom TDP.
TDP	OPN Max	Set the TDP in watts to the maximum supported by OPN.
Package Power Limit Control	Manual	Enable setting a custom Package Power Limit (PPL)
Package Power Limit	OPN Max	Set the PPL in watts to maximum supported by OPN.
SMT Control	Auto	Enables Symmetric Multithreading (SMT), which allows two hardware threads per core.
NUMA Node per Socket (NPS)	1, 2, or 4	Determines the number of NUMA nodes to split the memory channels between. A higher number indicates fewer memory channels per NUMA node, which lowers both memory throughput and latency for a NUMA node. If your configuration has more than 64 logical CPU's (2* # of cores when SMT is either **Enabled** or **Auto**) per socket, then set NPS to 2. If your configuration has more than 128 logical CPU cores per socket, then set NPS to 4.
ACPI SRAT LLC as NUMA Domain	Varies by workload	Dictates whether to report each LLC as NUMA domain to the OS. Allowing processes that use the same data to be scheduled on the set of CPUs that share an L3 cache increases the chance of L3 Cache hits.
Local APIC Mode	X2APIC	In general, interrupt delivery is faster when using x2APIC mode over the legacy xAPIC mode. Windows 2019 and later have support for x2APIC mode.

Table 5-1: Recommended BIOS settings for most RDBMS applications

This page intentionally left blank.

Chapter 6 OS Settings

Microsoft SQL Server can run on both Linux and Windows operating systems. AMD recommends using the latest available OS version. Please see AMD EPYC™ Processors Minimum Operating System (OS) Versions for a list of operating systems supported by AMD EPYC 9004 Series Processors. SQL server supports all of the Windows versions listed therein. For Linus support, please cross-reference the AMD EPYC Operating Systems page with the Installation guidance for SQL Server on Linux found at SQL Server | Microsoft Learn* to determine OS requirements for both the processors and SQL Server.

6.1 Windows Configuration

Most of the default Windows OS configuration is already optimized for SQL Server. However, enabling large pages on Windows Server requires modifying the group policy for the database user to enable the **Lock Pages in Memory** setting.

Name	Value	Description
Lock Pages in Memory	<Enable>	Enable SQL Server user to lock pages in memory through the group policy editor.

Table 6-1: General Windows settings

6.2 Linux Configuration

Modify `sysctl.conf`, `mssql.conf`, and `rc.local` as shown in Table 6-2, Table 6-3, and Table 6-4 to improve SQL Server performance on Linux.

6.2.1 /etc/sysctl.conf

Name	Value	Description
kernel.numa_balancing	0	Disable Linux kernel auto NUMA balancing.
vm.max_map_count	262144	Increase settings for Virtual Address Space.
vm.swappiness	1	Sets a strong preference to keeping process memory in physical memory at the expense of filesystem cache.

Table 6-2: General sysctl.conf settings

6.2.2 /var/opt/mssql/mssql.conf

Name	Value	Description
memory.memorylimitmb	85-90% of system memory	Limits the max memory visible to SQL Server.
control.writethough	1	For NVMe and any drives that support force unit access (FUA). Use `O_DSYNC` for file flag write through requests.
control.alternatewritethough	0	Controls write, flush behavior takes place.

Table 6-3: MSSQL 2022 mssql.conf settings

6.2.3 /etc/rc.local

Name	Value	Description
/sys/block/<device>/queue/scheduler	deadline	Modify elevator scheduler for all SQL Server storage devices.
/sys/block/<device>/queue/nr_requests	1024	Number of concurrent requests to the I/O device.

Table 6-4: /etc/rc.local entries

6.2.4 tuned-adm profile

The throughput-performance tuned-adm profile produces optimal performance. This profile sets up system I/O and memory throughput by configuring 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 profile-specific settings. Please also see the Kernel and CPU settings for high performance section in Performance Best Practices and Configuration Guidelines for SQL Server on Linux* for additional information.

6.2.5 Filesystem Mount Options

The EXT4 filesystem handles smaller file sizes (16TGB) than XFS (500TB). Table 6-5 shows optimal mount options for both filesystems.

Name	Description
Noatime	This option disables updating the metadata associated with files in the filesystem with an updated access time. This tracking is redundant because the database logs its own accesses.
Nobarrier	Disables the filesystem write barrier. Using a write barrier degrades I/O performance by requiring more frequent data flushes.

Table 6-5: XFS filesystem mount options

Chapter 7 SQL Server Configuration Settings

SQL Server has many options and settings that can impact database functionality and performance. Test these options in a non-production environment to verify that they meet your requirements.

Note: Some settings alter and disable certain database features that reduce performance, such as checkpointing, logging, and database recoverability. Review and test your settings before implementing them in a production environment.

7.1 SQL Server Soft NUMA Configuration

By default, soft-NUMA nodes are created automatically whenever the SQL Server engine detects more than eight physical cores per NUMA node or socket at startup. Multi-threaded cores are not differentiated when counting physical cores either 1n a node. SQL Server Database Engine then creates soft-NUMA nodes that ideally contain eight cores but which can go down to five or up to nine logical cores per node.

You can disable soft-NUMA by using the `ALTER SERVER CONFIGURATION SET SOFTNUMA OFF` statement and then restarting the database engine. The SQL Server error log file will display something similar to the following:

```
19 09:21:46.78 Server      Automatic soft-NUMA was enabled because SQL Server has detected
hardware NUMA nodes with greater than 8 physical cores.

2022-01-19 09:21:46.78 Server      Processor affinity turned on: node 0, processor mask
0x00000000000000ff. Threads will execute on CPUs per affinity settings. This is an
informational message; no user action is required.

2022-01-19 09:21:46.78 Server      Processor affinity turned on: node 1, processor mask
0x000000000000ff00. Threads will execute on CPUs per affinity settings. This is an
informational message; no user action is required

…

2022-01-19 09:21:46.78 Server      Processor affinity turned on: node 30, processor mask
0x00ff000000000000. Threads will execute on CPUs per affinity settings. This is an
informational message; no user action is required.

2022-01-19 09:21:46.78 Server      Processor affinity turned on: node 31, processor mask
0xff00000000000000. Threads will execute on CPUs per affinity settings. This is an
informational message; no user action is required.
```

Note: This example is from a system with two 7763 processors and Symmetric Multi-Threading enabled.

Things to look for in this output include the first line showing both that Soft-NUMA was enabled and the processor mask for each node. Verify that the mask includes consecutive bits and that they are associated with the CPUs within a core complex. This ensures that the set of processor threads for a given node are all sharing the same last level cache. If the Automatic Soft-NUMA mask does not meet these requirements, then the cross Level 3 cache communication will increase and hurt performance. If this happens, then you must manually configure the Soft-NUMA nodes via the

registry. The following example uses a 2-socket system with dual 64-core AMD EPYC 7763 CPUs and SMT enabled, with 128 cores / 256 threads split across 16 Core Complexes running MSSQL-Server 2022. This example will create 16 soft-NUMA nodes and then group the cores and SMT threads to match the Core Complex boundaries.

```
Windows Registry Editor Version 5.00

[HKEY_LOCAL_MACHINE\SOFTWARE\Microsoft\Microsoft SQL Server\160\NodeConfiguration]

[HKEY_LOCAL_MACHINE\SOFTWARE\Microsoft\Microsoft SQL Server\160\NodeConfiguration\Node0]
"CPUMask"=hex:00,00,00,00,00,00,ff,ff
"Group"=dword:00000000

[HKEY_LOCAL_MACHINE\SOFTWARE\Microsoft\Microsoft SQL Server\160\NodeConfiguration\Node1]
"CPUMask"=hex:00,00,00,00,ff,ff,00,00
"Group"=dword:00000000

[HKEY_LOCAL_MACHINE\SOFTWARE\Microsoft\Microsoft SQL Server\160\NodeConfiguration\Node2]
"CPUMask"=hex:00,00,ff,ff,00,00,00,00
"Group"=dword:00000000

[HKEY_LOCAL_MACHINE\SOFTWARE\Microsoft\Microsoft SQL Server\160\NodeConfiguration\Node3]
"CPUMask"=hex:ff,ff,00,00,00,00,00,00
"Group"=dword:00000000

[HKEY_LOCAL_MACHINE\SOFTWARE\Microsoft\Microsoft SQL Server\160\NodeConfiguration\Node4]
"CPUMask"=hex:00,00,00,00,00,00,ff,ff
"Group"=dword:00000001

[HKEY_LOCAL_MACHINE\SOFTWARE\Microsoft\Microsoft SQL Server\160\NodeConfiguration\Node5]
"CPUMask"=hex:00,00,00,00,ff,ff,00,00
"Group"=dword:00000001

[HKEY_LOCAL_MACHINE\SOFTWARE\Microsoft\Microsoft SQL Server\160\NodeConfiguration\Node6]
"CPUMask"=hex:00,00,ff,ff,00,00,00,00
"Group"=dword:00000001

[HKEY_LOCAL_MACHINE\SOFTWARE\Microsoft\Microsoft SQL Server\160\NodeConfiguration\Node7]
"CPUMask"=hex:ff,ff,00,00,00,00,00,00
"Group"=dword:00000001

[HKEY_LOCAL_MACHINE\SOFTWARE\Microsoft\Microsoft SQL Server\160\NodeConfiguration\Node8]
"CPUMask"=hex:00,00,00,00,00,00,ff,ff
"Group"=dword:00000002

[HKEY_LOCAL_MACHINE\SOFTWARE\Microsoft\Microsoft SQL Server\160\NodeConfiguration\Node9]
"CPUMask"=hex:00,00,00,00,ff,ff,00,00
"Group"=dword:00000002

[HKEY_LOCAL_MACHINE\SOFTWARE\Microsoft\Microsoft SQL Server\160\NodeConfiguration\Node10]
"CPUMask"=hex:00,00,ff,ff,00,00,00,00
"Group"=dword:00000002

[HKEY_LOCAL_MACHINE\SOFTWARE\Microsoft\Microsoft SQL Server\160\NodeConfiguration\Node11]
"CPUMask"=hex:ff,ff,00,00,00,00,00,00
"Group"=dword:00000002

[HKEY_LOCAL_MACHINE\SOFTWARE\Microsoft\Microsoft SQL Server\160\NodeConfiguration\Node12]
"CPUMask"=hex:00,00,00,00,00,00,ff,ff
"Group"=dword:00000003

[HKEY_LOCAL_MACHINE\SOFTWARE\Microsoft\Microsoft SQL Server\160\NodeConfiguration\Node13]
"CPUMask"=hex:00,00,00,00,ff,ff,00,00
"Group"=dword:00000003
```

```
[HKEY_LOCAL_MACHINE\SOFTWARE\Microsoft\Microsoft SQL Server\160\NodeConfiguration\Node14]
"CPUMask"=hex:00,00,ff,ff,00,00,00,00
"Group"=dword:00000003

[HKEY_LOCAL_MACHINE\SOFTWARE\Microsoft\Microsoft SQL Server\160\NodeConfiguration\Node15]
"CPUMask"=hex:ff,ff,00,00,00,00,00,00
"Group"=dword:00000003
```

The mask sometimes comes up as alternating between `*,aa, aa, aa, aa,*` and `*, 55, 55, 55, 55, *`. These CPU masks are undesirable. This is especially true when SMT is enabled because this places the sibling threads for a physical core on two separate Soft-NUMA nodes. If your configuration does this by default, then you will get better performance by switching off the Automatic Soft NUMA and using the manual configuration described above to obtain optimal Soft-NUMA alignment to the CCDs.

7.2 SQL Server 2022 Instance Trace Flags and Settings

Microsoft SQL Server 2022 has many trace flags that tweak how the database interacts and behaves. Table 7-1 lists a few trace flags that will help in most environments.

Name	Value	Description
traceflag	3979	Disables the WAL (`FlushFileBuffers`) forced flush feature.
traceflag	834	Use large pages for buffer pool.
traceflag	652	Disable read ahead.

Table 7-1: MS SQL Server 2022 trace flags

Set recovery mode to simple:

```
ALTER DATABASE <db_name> set recovery simple;
```

Set target recovery interval to 0:

```
ALTER DATABASE <db_name> set TARGET_RECOVERY_TIME = 0 seconds;
```

Set page verify to ‘none’:

```
ALTER DATABASE <db_name> set PAGE_VERIFY None;
```

Turn off auto-growth on db files and tempdb files:

```
ALTER DATABASE <db_name> MODIFY FILE (NAME=<dbfile_name>, FILEGROWTH=0)
```

Increase the tempdb file count to match the number of cpu threads. A default SQL Server installation will only create 8 tempdb data files, which is not enough for AMD EPYC processors. Remember to keep all tempdb files the same size to help ensure even usage. Further, tempdb should be stored in the fastest/lowest-latency storage available and distributed across multiple storage devices (either by hardware striping technologies or individual file placements):

```
ALTER DATABASE <db_name> ADD FILE (NAME=<dbfile_name>, FILENAME=<filepath\filename,
SIZE=<size of tempdb file>, FILEGROWTH=0);
```

7.3 sp_configure

Instance-level SQL Server configuration settings can increase performance across all workloads. Use the `sp_configure` command to set the parameters shown in Table 7-2.

Name	Value	Description
Show advanced options	1	Enables View/Modify advanced `sp_configure` options.
Automatic soft-NUMA disabled	1‘	Disables automatic Soft-NUMA configuration for SQL Server. Instead, use Windows Registry settings for Soft-NUMA described in “SQL Server Soft NUMA Configuration” on page 29.
Lightweight pooling	1	Lightweight threaded pools for user connections.
Recovery interval	32767	Set recovery interval to highest allowed (32767 minutes).
Priority Boost Enabled	1	Boost priority for SQL Server process threads (Windows® only).
Max server memory (MB)	80-90% of system/VM memory	Maximum memory for SQL Server to use.

Table 7-2: MS SQL Server 2022 configuration settings

Chapter 8 Workload-Specific Considerations

SQL Server databases are generally deployed for one of two primary workloads that each have unique considerations. The Data Warehouse workload is more query oriented, while the Online Transaction Processing workload is more transaction oriented. Some databases run both workloads, which requires either deciding which workload has higher priority or determining the right tuning balance for each of them.

8.1 Data Warehouse

SQL Server data warehouses consume a high amount of CPU, memory, and I/O resources. SQL Server provides parallel queries to optimize query execution and index operations on systems with more than one processor. SQL Server can perform query or index operations in parallel by using several OS worker threads for both speed and efficiency. In an environment where only 1 query or operation is running at a time (batched queries/updates etc.), SQL Server can set the max degree of parallelism to twice the number of cores to fully utilize all system threads. In other cases where several query threads are running simultaneously, it may help to reduce the max degree of parallelism such that a single query does not consume all resources and force other queries to run single-threaded.

Setting NPS to 1 normally generates best query throughput performance. However, single socket systems with 32 or more cores show reliable performance with higher NUMA node per socket (NPS) BIOS setting. In those cases, setting this to 2 or 4 may boost performance boost. AMD therefore recommends setting NPS2 for systems with between 32-cores and 48-cores and NPS4 for systems with more than 48 cores. Test this setting with your specific workloads to evaluate performance.

8.2 Online Transaction Processing

OLTP workloads require high transaction log write performance because of their high insert and/or update rates. AMD strongly recommends SSD and NVMe drives for both data areas and redo logs to match CPU performance. The next level of contention issues arises around transactional latencies in memory, the network latencies, or the data disk(s).

This page intentionally left blank.

Chapter 9 Resources

- Database Tuning on Linux® OS: Reference Guide for AMD EPYC™ 7002 Series Processors; this Tuning Guide is also valid for AMD EPYC 9004 Series Processors.
- From AMD EPYC Tuning Guides:
  - *Microsoft Windows® Tuning Guide for AMD EPYC™ 9004 Series Processors*
  - *NVMe Tuning Guide for AMD EPYC™ 9004 Series Processors*
- Performance Best Practices and Configuration Guidelines for SQL Server on Linux*
- Soft-NUMA (SQL Server)*

This page intentionally left blank.

Chapter 10

Processor Identification

Figure 10-1 shows the processor naming convention for AMD EPYC 9004 Series Processors and how to use this convention to identify particular processors models:

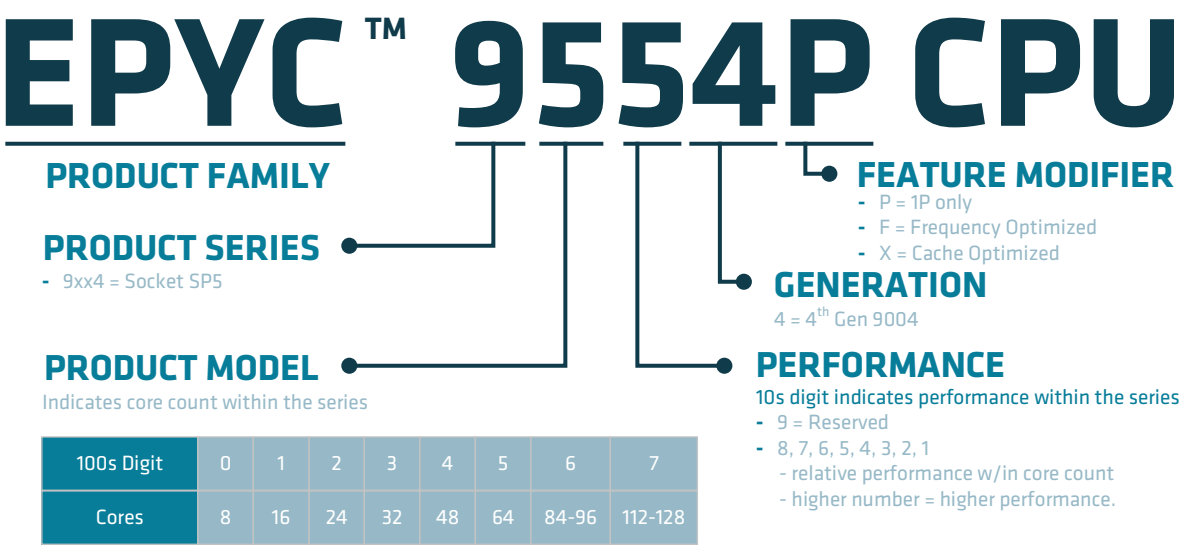

Figure 10-1: AMD EPYC SoC naming convention

10.1 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.

10.2 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*.

10.2.1 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.