AMD EPYC

TUNING GUIDE
AMD EPYC 9004

Virtual Desktop Infrastructure (VDI)

Publication	58001
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 release
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 document is intended for a technical audience with experience in deploying, configuring, and managing virtual desktop environments on physical and/or virtualized servers, such as:

- Datacenter architects
- Support staff
- Server/application planning specialists

Authors

John Kennedy and Muhammad Ashfaq

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

Virtual Desktop Infrastructure (VDI) is a method of allowing end users to access a desktop environment that is completely configured with a desktop OS, permanent storage, and exclusive CPU and memory resources on a remote computer that typically resides in a large, well-managed datacenter and is accessed via public or private networks. By contrast, a typical physical desktop infrastructure consists of locally allocated CPU, memory, and storage resources such as a desktop computer, workstation, or laptop.

VDI is one solution to the problem of maintaining centralized data control while utilizing remote workers. Local PCs and workstations can provide more than adequate processing, memory, and disk storage for remote workers, but having to maintain these systems is both costly and time consuming. Further, keeping information secure in remote locales is sometime impossible, which can make VDI a necessity.

This tuning guide provides parameters that can optimize VDI performance on servers built with AMD EPYC™ 9004 Series processors. The tuning recommendations are based on AMD's internal test results on AMD EPYC™ 9004 Series Processors in a typical OEM server.

1.1 VDI Design Overview

VDI systems typically combine products that may not necessarily originate from the same vendor. The architectures vary widely with different solutions, but the most common designs use virtualization to concentrate multiple desktops onto a single host.

1.1.1 Architecture

The most common architectures for VDI deployment are:

- **Traditional hypervisor deployment:** Uses a centralized Storage Area Network (SAN) commonly shared with other servers across the enterprise.
- **Converged infrastructure:** Composed of similarly-equipped servers with dedicated SAN and networking allocated as a tested unit. These predesigned and validated infrastructures do not share hardware access for users outside of their own system.
- **Hyper-converged infrastructure (HCI):** Composed of "blocks" of compute, memory, and storage connected by high-speed networking. Each block's resources are added to a virtual pool for allocation to individual tasks as needed. They are typically managed by a central management software component that makes the system appear as one logical pool of resources.

1.1.2 Software Components

VDI software components can be categorized into the functional groups shown in Table 1-1.

Component	Description
Hypervisor	The hypervisor resides directly on the "bare metal" server. It provides virtual machines in software and runs the guest OS software to create a virtual desktop environment.
Guest OS	Any operating system supported by the hypervisor can provide the desktop experience. The guest OS is typically graphical, but a command-line based OS can also be used. A graphical guest OS usually provides a full desktop with multiple apps and folders just as on a standalone PC or laptop. However, individual desktop applications can be delivered instead of an entire desktop, depending on the guest OS capabilities.
Remote KVM software	Compression algorithms help provide a desirable user experience by minimizing the bandwidth consumed while transmitting video. Most virtual desktops are based on Microsoft Windows. There are two types of remote KVM software: • Remote Desktop software provides access to an entire desktop environment. The experience is identical to logging into a local workstation or laptop. • Remote Application software provides only the UI of a particular application that is "published" by the remote system. Most Windows-based applications can be published in this manner, allowing the user to interact only with the applications they want. This provides a slightly lower bandwidth solution for accessing remote systems compared to a full remote desktop experience.

Table 1-1: VDI software components

1.2 End User Device

The end user device through which the user connects to the VDI is the final piece of the puzzle. The first end user devices were simple, less powerful desktops or laptops. The advent of the Bring Your Own Device (BYOD) movement has made endpoint devices take many forms. For example:

- **Thin client:** A small, inexpensive device with limited compute power but with graphical capability. Thin clients typically store no data and are stateless, providing only KVM access from the end user location to the virtual desktop.
- **Portable devices:** Tablet computers and even cellular phones can and are being used as endpoint devices for virtual apps and desktops.

1.3 Measuring VDI Performance

VDI performance measurements typically examine both of the following:

- **Capacity:** How many virtual desktop users an architecture can support while maintaining an acceptable level of responsiveness.
- **Response time:** How quickly the remote desktop responds to user commands.

The tuning recommendations in this Tuning Guide are based on tests conducted using the industry-standard Login VSI Pro® measurement software.

1.3.1 Test Methodology

Login VSI Pro uses "Launcher" systems that make sequential connections to individual remote desktops. The desktops then execute a series of tasks that simulate a remote user performing routine tasks such as editing data in a Microsoft® Excel spreadsheet, editing a Microsoft Word document, using Java based applications, presenting with Microsoft PowerPoint®, and printing to PDF. The time taken to perform these tasks is measured and recorded.

The software waits for the default two minutes after launching the desired number of desktops before sending logoff signals to the desktops to log out each connection. It then calculates the baseline response time (VSIBase) and an arbitrary threshold response time. The threshold response time is the sum of the VSIBase plus 1,000 milliseconds. The number of desktops performing tasks at the time that the overall response time exceeds the threshold is called VSImax, which represents the maximum number of desktops that can be simultaneously executing on a given architecture while maintaining acceptable response times for end users.

The VSImax calculations were used to determine the tunable parameters that impact performance and response time. This Tuning Guide captures recommendations based on these test results without including the test parameters or results.

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.

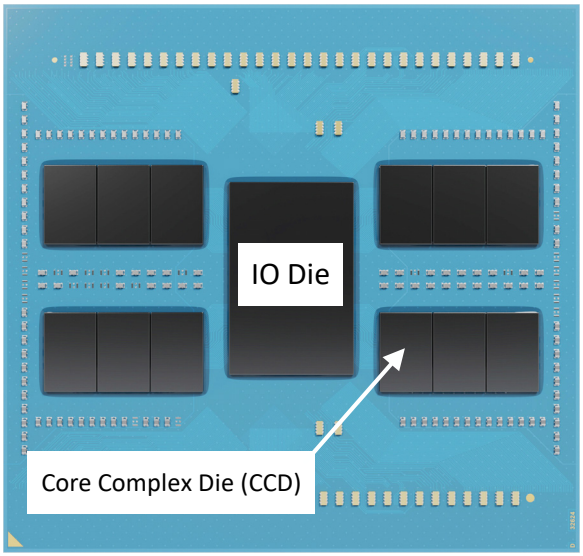

Figure 2-1: AMD EPYC 9004 configuration with 12 Core Complex Dies (CCD) surrounding a central I/O Die (IOD)

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.

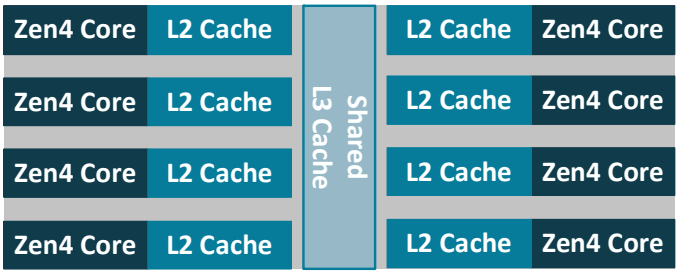

Figure 2-2: Top view of 8 compute cores sharing an L3 cache (91xx-96xx models)

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.

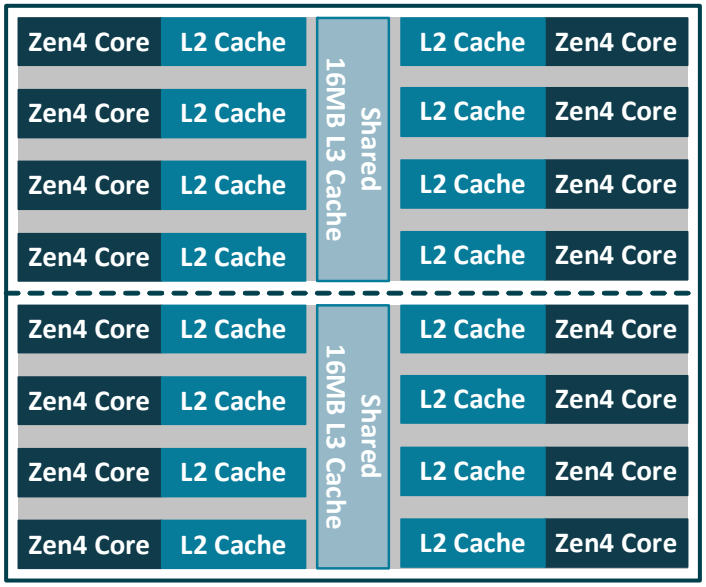

Figure 2-3: 2 CCXs in a single 4th Gen AMD EPYC 97xx CCD

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.

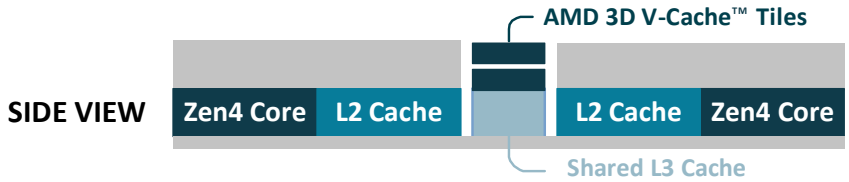

Figure 2-4: Side view of vertically-stacked central L3 SRAM tiles

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.

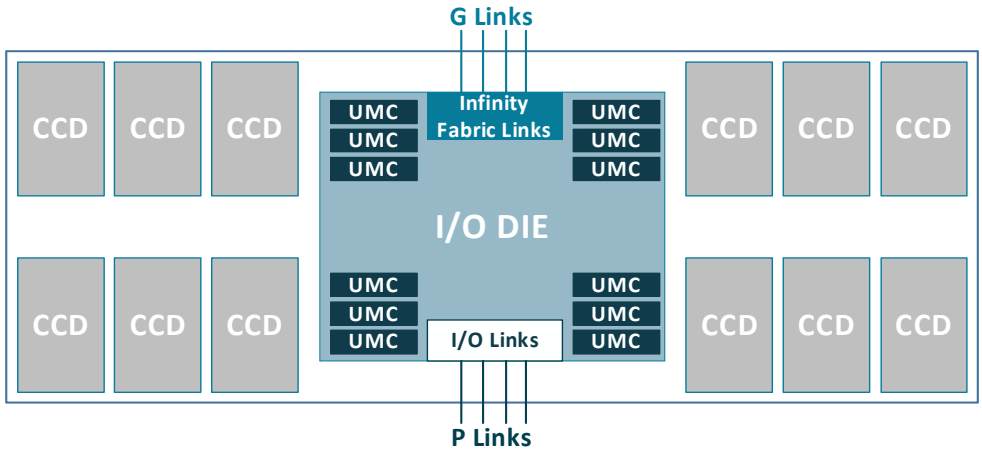

Figure 2-5: AMD EPYC 9654 processor internals interconnect via AMD Infinity Fabric (12 CCD processor shown)

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.

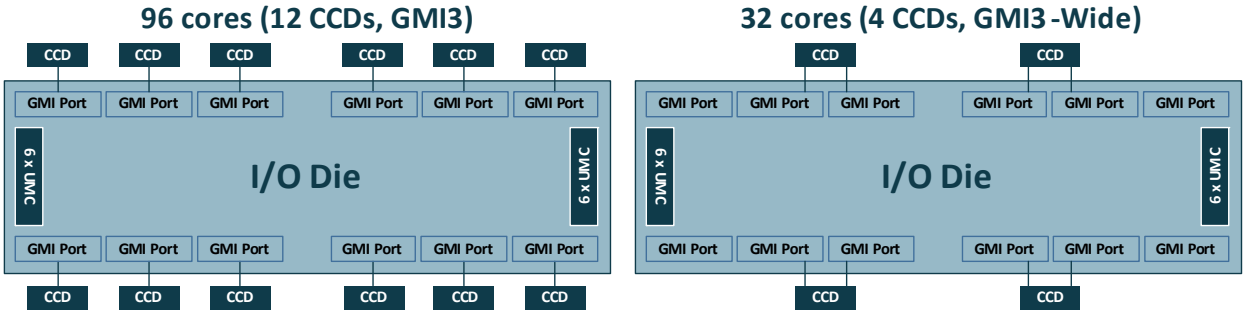

Figure 2-6: Standard vs. Wide GMI links

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 12 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.

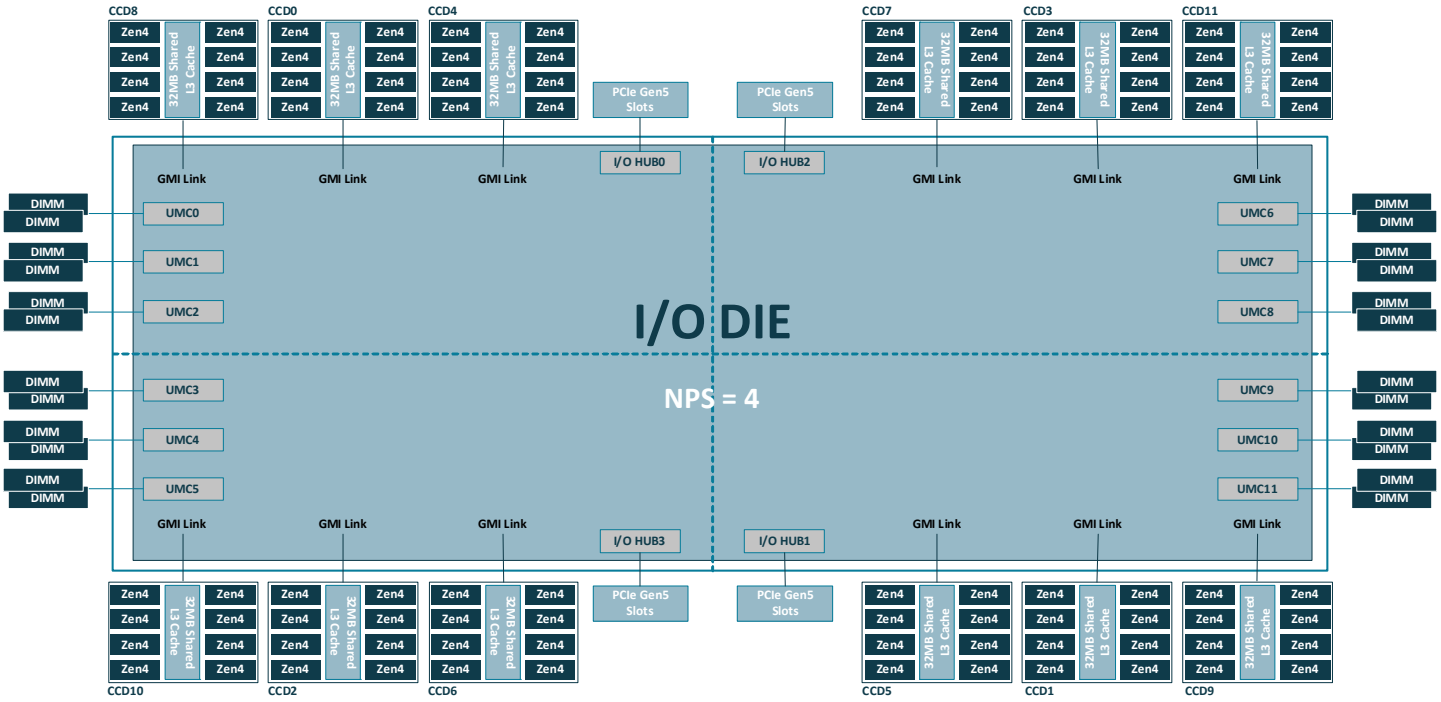

Figure 2-7: The AMD EPYC 9004 SoC consists of up to 12 CCDs and a central IOD for 91xx-96xx models, including "X" OPNs

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.

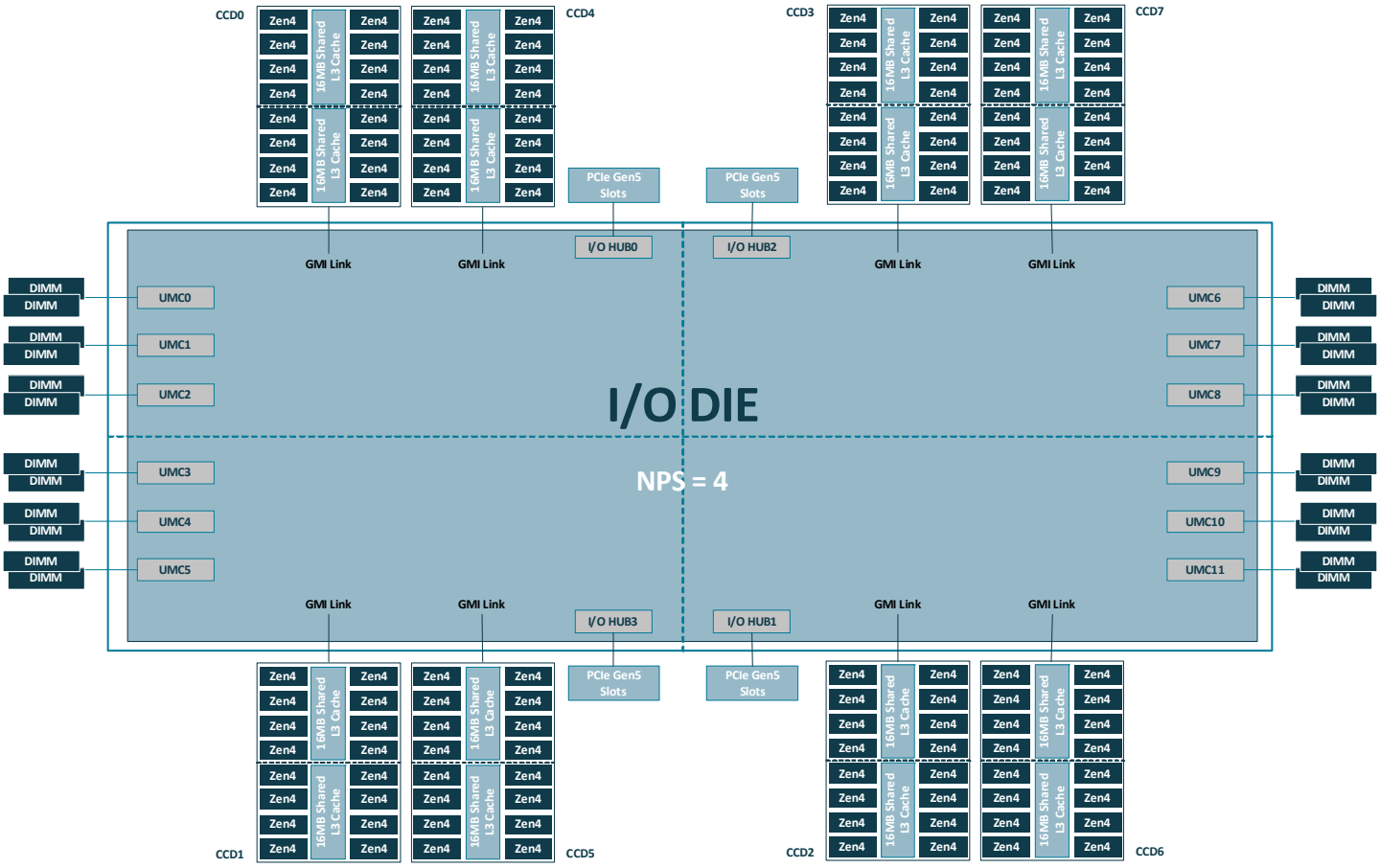

Figure 2-8: The AMD EPYC 9004 System on Chip (SoC) consists of up to 8 CCDs and a central IOD for 97xx models

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 10 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.

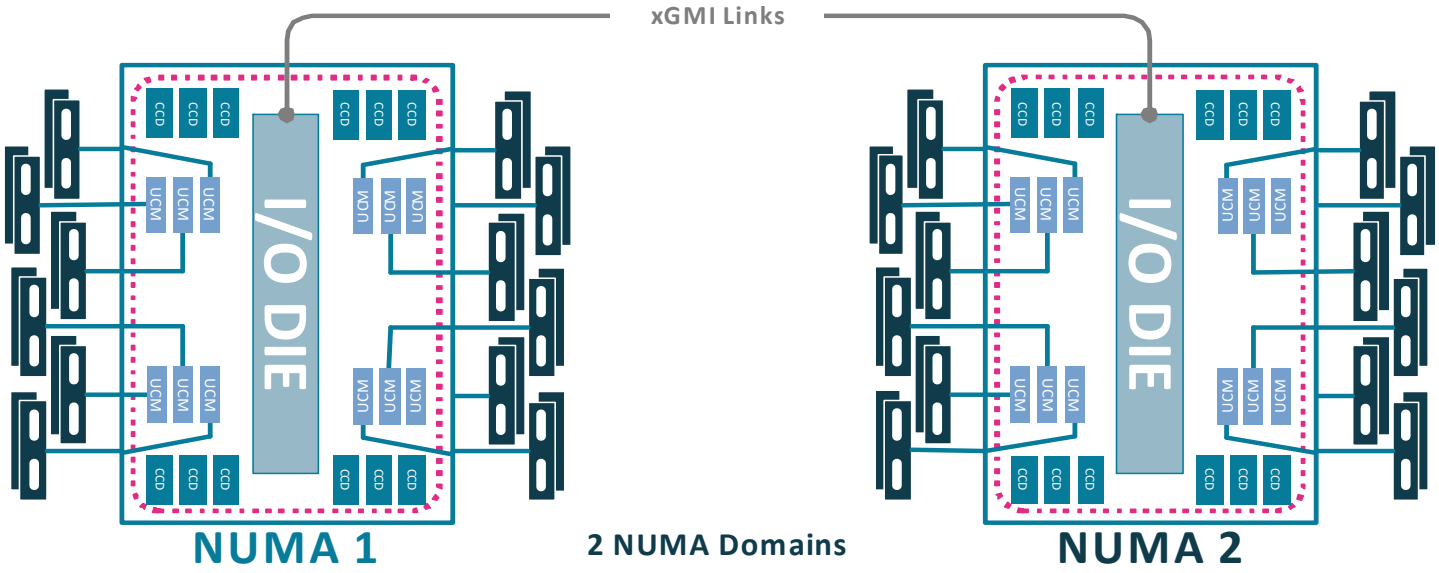

Figure 2-9: Two EPYC 9004 Processors connect through 4 xGMI links (NPS1)

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 Best Practices

This chapter describes best practices for tuning the hardware and OS for VDI deployments.

4.1 Hardware

The best practices recommendations listed in this chapter are based on an individual server with local storage, high speed networking, and varying processor types. These components provide a baseline for determining the impact of BIOS setting changes, such as adjusting CPU features, that will positively impact virtual desktop performance.

4.1.1 Core Count per Socket

Core count per socket impacts VDI in multiple ways. Processors with higher core counts can cost more and not provide a linear increase in desktops per server. Testing on processors with more than 32 cores shows less-than-linear scaling because the higher power requirements for each core bring the overall processor closer to thermal limits (TDP) as cores are added. This prevents high-core-count processors from running every core at the higher clock speeds achieved by processors with fewer cores, resulting in less-than-linear gains in the number of desktops that can run concurrently. Thus:

- If you need high performance, then use processors with 32 or fewer cores per socket.
- If you need high desktop density per server, then use processors with up to 96 cores per socket.

4.1.2 Single vs. Dual Processor

The choice of single or dual processors in a system does not impact VDI performance. However, a dual-processor system can increase the core count while staying under thermal limits. This allows all cores to run at higher frequencies for significantly enhanced performance. It is appropriate to select dual-socket servers for VDI systems. In general:

- Use single-socket systems for small deployment.
- Use dual-docket systems for medium and larger deployments.

4.1.3 Memory

High density desktops require populating the server with the maximum possible amount of memory. Further, the memory clock frequency provides marginally significant performance improvement in VDI environments, with memory operating at 4800 MHz or higher providing the most optimal performance.

Most systems powered by AMD EPYC 9004 Series Processors automatically lower memory frequency when equipped with 2 DIMMs per channel (DPC) due to power constraints. 1 DPC provides slightly faster speed while only providing half of the memory capacity. This means either that virtual desktops must use less memory per desktop or that the number of concurrent virtual desktops is reduced by half. The best practice recommendation therefore depends on your architecture goal:

- Use 1 DPC for the fastest possible response times.
- Use 2 DPC for maximum desktop density.
- Use memory with 64-256GB per DIMM, or higher if available and not cost prohibitive.

4.2 Operating System

AMD EPYC processors can accept OS commands for power management and core frequency adjustments. Enabling ACPI power management in the OS allows the below BIOS settings to operate properly and is therefore recommended as a best practice.

Chapter 5 Recommended BIOS Settings

This chapter focuses on tuning BIOS parameters.

5.1 SMT

Symmetric Multi-threading (SMT) is a core configuration parameter that allows two simultaneous processes to use a single processor core without interfering with each other. Disabling SMT means that only one virtual processor can be scheduled on a core at a given time. Enabling SMT means that two virtual processors can be scheduled to use the same core in different threads simultaneously, thereby doubling desktop density with little performance loss. Enabling SMT is therefore a recommended best practice for VDI.

5.2 Core Boost Frequency

AMD EPYC processors are equipped with a **Core Boost** control that allows the processor to increase core frequencies as needed to ensure performance while maintaining power control. VDI performance increases linearly with core frequency. Many server vendors enable this control by default, and enabling this feature is therefore a best practice for VDI.

5.3 P States

Most server vendors provide BIOS options that allow the host OS to control processor core frequencies via P-states. P-states are different core frequencies that can be set for the OS to choose during operation. Putting cores at a lower P-state frequency can conserve power when that core is not being heavily utilized. Frequency determines performance and density in VDI deployments, with higher frequencies yielding better performance. Allowing the OS to control the frequency is essential, and enabling P-States is therefore a best practice for VDI.

5.4 Determinism Control

The **Determinism** control helps manage workload performance.

- Setting the **Determinism** control to **Performance** maintains all processor core frequencies at stable, uniform levels, thereby ensuring that the performance of each core is close to or the same as every other core.
- Setting the **Determinism** control to **Power** allows the processor to manage each core frequency individually, thereby allowing each core to run at as high a frequency as possible without encountering thermal limits. Allowing some cores to run at low speeds allows other cores can run at higher speeds without generating excessive heat.

As VDI is a highly compute-variable workload where some desktops may be operating at maximum compute capacity while others are idle. Leveraging performance-per-watt mode is therefore critical to achieving fast response times and maximum desktop density. Setting the **Determinism** control to **Power** is therefore a best practice for VDI.

5.5 C States

Idle processor cores can be placed in very low power states to save power. These are referred to as C-states and are offered in multiple levels from C0 to C6. Placing these cores in low power states saves power at the cost of a time penalty when recovering the core to a functioning state. Testing shows that C-states have very little effect on performance and can be left at C0 if desired.

Most modern operating system account for NUMA and can even leverage it for higher performance by placing data in memory "nearest" to its cores to allow faster access. However, some workloads such as databases that rely on consistent memory access speeds desire all cores to access memory at the same speed.

AMD EPYC processors can be configured into 1, 2, or 4 NUMA zones. Dual processor systems can also be configured to force all cores to access all memory in the system at the same speed (NUMA 0). NUMA zones below 4 reduce all core access times to memory to the lowest common denominator, thereby impacting individual core performance but creating uniformity. This allows workloads to use the NUMA geometry that is most beneficial to them.

VDI is a highly individualistic workload made up of many different VMs running or sleeping independently of each other. Most virtual desktops also have a low number of virtual CPUs accessing small amounts of memory. The virtual desktop VMs can be contained within one NUMA zone in almost all cases. Setting **NUMA Nodes Per Socket** to 4 is the recommended best practice for VDI because it shows a significant increase in performance compared NPS 1, as follows:

- **vSphere 7.0U1 and below:** NPS 4
- **vSphere 7.0U2 and above:** NPS 1

Please see "NUMA Topology" on page 12 for additional information.

5.6 LLC as NUMA Zone

In the AMD EPYC processor architecture referenced above and in "NUMA Topology" on page 12, each processor has up to 12 CCDs with eight cores residing on each CCD. In the 9004 architecture, each CCD has up to 32 MB of Last Level Cache (LLC). Setting NPS=4 on a 12-CCD processor gives three CCD the same access time to a given part of memory. However, if that memory is cached in LLC, then one CCD has a speed advantage when accessing that LLC over the other CCDs. Thus, despite the dynamic RAM access times being the same, the effective time to load memory into a core is not. This creates a "virtual" NUMA zone where effective access times are different between CCDs.

Operating systems and workloads can take advantage of this effect by placing processes on cores that have memory access within their LLC area. This is referred to as **LLC as NUMA**, or ACPI SLIT. The processor uses the SLIT to report which memory is within an LLC zone of which cores to the operating system. This boosts performance because of the faster memory access. VDI performance can improve when **LLC as NUMA** is **Enabled**.

Note: As of this writing, VMware ESXi only allows 16 NUMA nodes per server. Thus, using LLC as NUMA in 2P 96-core systems is prohibited because this would create 24 NUMA nodes. See VMware configuration maximums for more details.

- **vSphere 7.0U1 and below:** Enabled
- **vSphere 7.0U2 and above:** Disabled

5.7 Power Profile Selection Control

AMD EPYC 9004 Series Processors provide a **Power Profile Selection Control** that can be set to either **High Performance**, **Efficiency Mode**, or **Maximum IO Performance Mode**. AMD recommends **High Performance Mode** as best practice for VDI.

5.8 VDI Sizing Recommendations

The following table shows VDI sizing recommendations when various generations of Gen AMD EPYC processors. Testing was based on Windows 10 virtual desktops running Microsoft Office 2016 or 2019 with 2 vCPUs and 4GB of RAM on VMware ESXi versions 6, 7, and 8.

VDI Sizing Recommendations	2nd Gen AMD EPYC processors (7xx2)	3rd Gen AMD EPYC processors (7xx3)	4th Gen AMD EPYC processors (9xx4)
Small (~100 desktops/server)	• 1 x 7532 (32 core) • 2 x 7302 (32 core)	• 1 x 7343 (16 cores) • 2 x 72F3 (16 cores)	• 1 x 16-core (200 W)
Medium (~200-250 desktops/ server)	• 2 x 7542 (64 cores) • 2 x 7662 (64 cores)	• 2 x 7343 (32 cores) • 1 x 7713 (64 cores)	• 1 x 24-core (200 W)
Large (~350 desktops/server)	N/A	• 2 x 74F3 (48 cores) • 2 x 7543 (64 cores)	• 1 x 48-core (290 W) • 2 x 24-core (200 W)
X-Large (~500 desktops/server)	N/A	• 2 x 7713 (128 cores) • 2 x 7763 (128 cores)	• 1 x 96-core (360 W) • 2 x 48-core (290 W)
XX-Large (~700 desktops/server)	N/A	N/A	• 2 x 96-core (360 W)

Table 5-1: VDI sizing recommendations

This page intentionally left blank.

Chapter 6

Resources

- Memory Population Guidelines for AMD Family 19h Models 10h–1Fh - Login required; please review the latest version if multiple versions are present.
- Socket SP5 Platform NUMA Topology for AMD Family 19h Models 10h–1Fh - Login required; please review the latest version if multiple versions are present.
- *BIOS & Workload Tuning Guide for AMD EPYC™ 9004 Series Processors* (available from AMD EPYC Tuning Guides)
- LoginVSI Documentation*

This page intentionally left blank.

Chapter 7 Glossary

- **CCD:** Core Complex Die
- **CCX:** Core Complex
- **DIMM:** Dual In-line Memory Module
- **DPC:** DIMMs Per Channel
- **LLC:** Last Level Cache
- **NPS:** NUMA Per Socket
- **NUMA:** Non-Uniform Memory Access
- **OS:** Operating System
- **SMT:** Simultaneous Multi-Threading
- **VDI:** Virtual Desktop Infrastructure
- **VSImax:** Maximum number of virtual desktops that can run on a platform before exceeding a usability threshold
- **VSIBase:** Baseline response time of a virtual desktop running under minimal load
- **VM:** Virtual Machine

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:

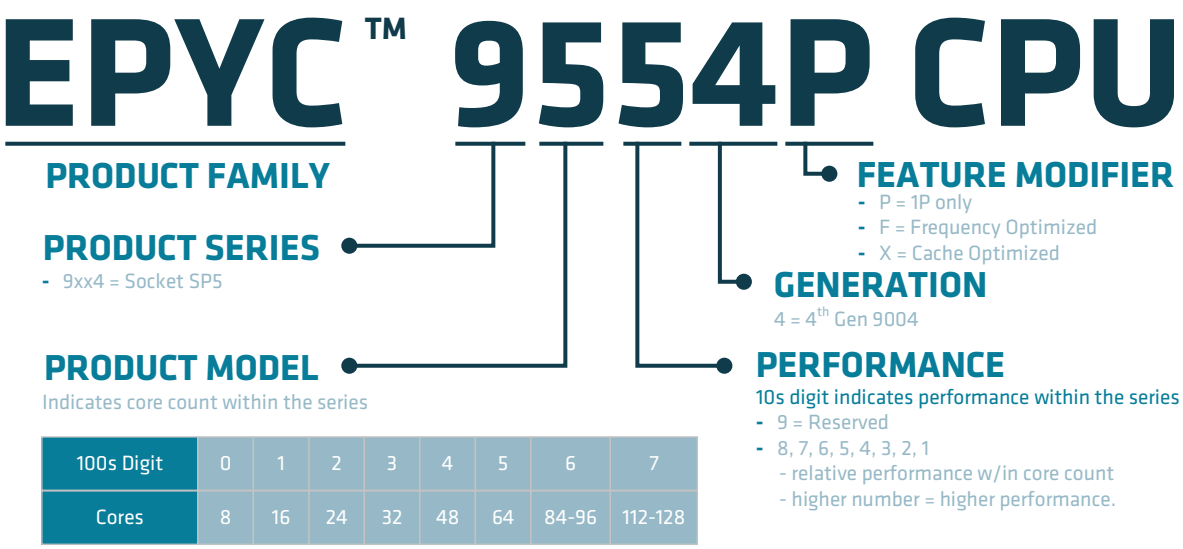

Figure 8-1: AMD EPYC SoC naming convention

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.