AMD EPYC

TUNING GUIDE
AMD EPYC 9004

Data Plane Development Kit (DPDK)

Publication	58017
Revision	1.3
Issue Date	June, 2023

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

Audience

This tuning guide describes best practices for optimizing performance using the Data Plane Development Kit (DPDK). It is intended for a technical audience such as DPDK application architects, production deployment, and performance engineering teams with:

- A background in configuring servers.
- Administrator-level access to both the server management Interface (BMC) and the OS.
- Familiarity with both the BMC and OS-specific configuration, monitoring, and troubleshooting tools. See the DPDK Debug & Troubleshoot Guide* for additional information.

Authors

Manish Kumar

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 various system configuration parameters that can optimize DPDK workload performance on servers based on AMD EPYC™ 9004 series processors. Default OEM system configurations may not provide the best possible performance for all DPDK workloads across different OS platforms. To optimize performance for a particular workload, this guide calls out:

- Hardware configuration (memory, PCIe) best practices
- BIOS settings that can impact performance
- Workload-specific settings in BIOS and OS parameters
- DPDK build optimization options
- Vendor-specific NIC configuration
- DPDK environment (EAL) options

Note: Do not use this tuning guide as a validation guide or a generic server optimization guide. It is only intended to help you optimize the performance of a specific AMD EPYC-based platform.

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.

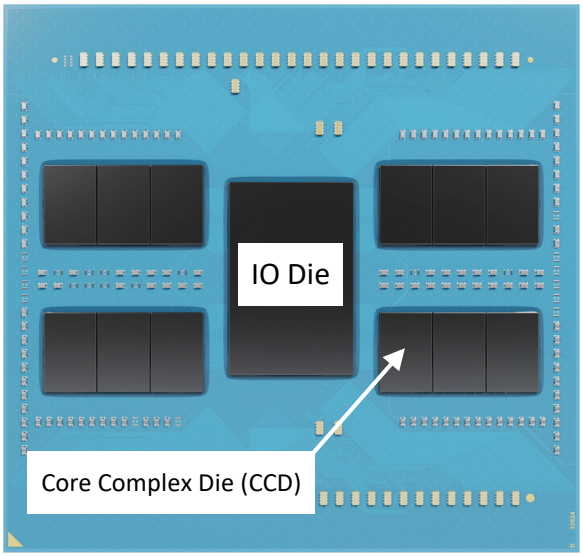

Figure 2-1: AMD EPYC 9004 configuration with 12 Core Complex Dies (CCD) surrounding a central I/O Die (IOD)

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.

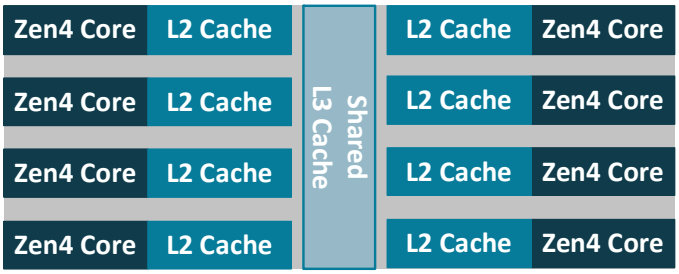

Figure 2-2: Top view of 8 compute cores sharing an L3 cache (91xx-96xx models)

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.

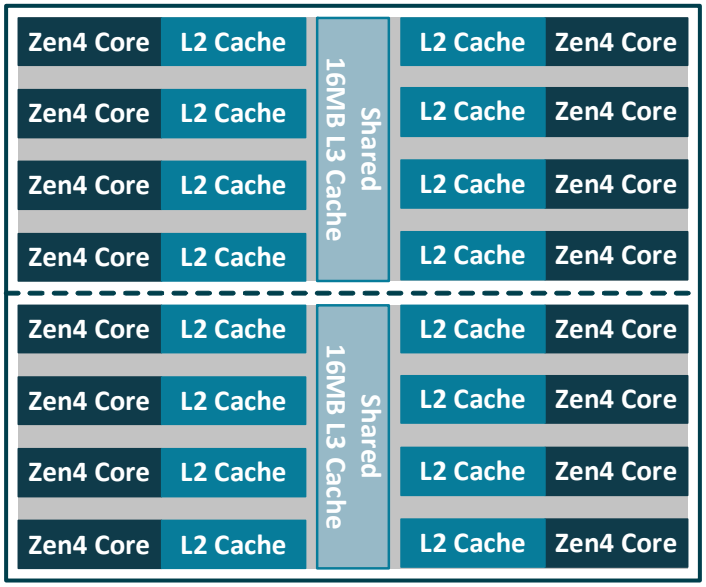

Figure 2-3: 2 CCXs in a single 4th Gen AMD EPYC 97xx CCD

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.

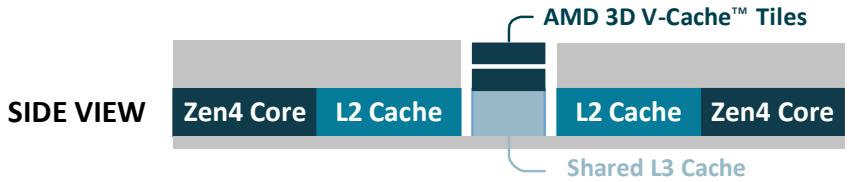

Figure 2-4: Side view of vertically-stacked central L3 SRAM tiles

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.

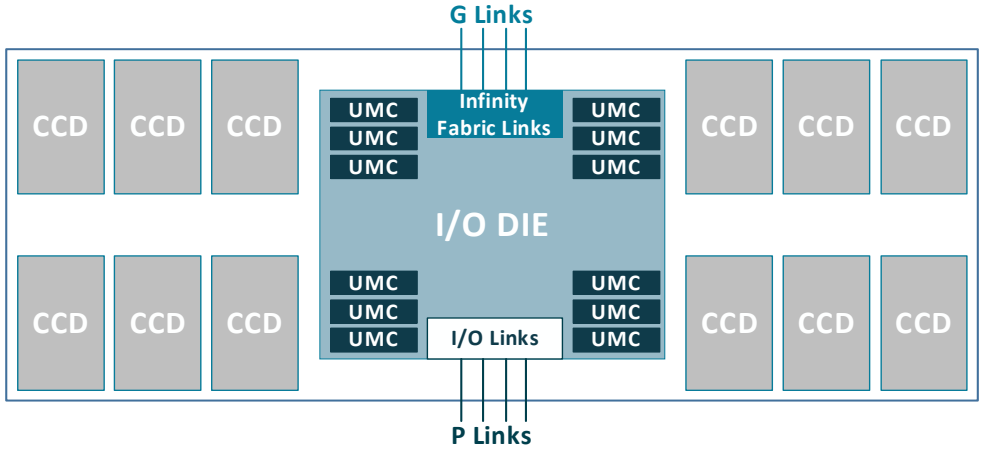

Figure 2-5: AMD EPYC 9654 processor internals interconnect via AMD Infinity Fabric (12 CCD processor shown)

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.

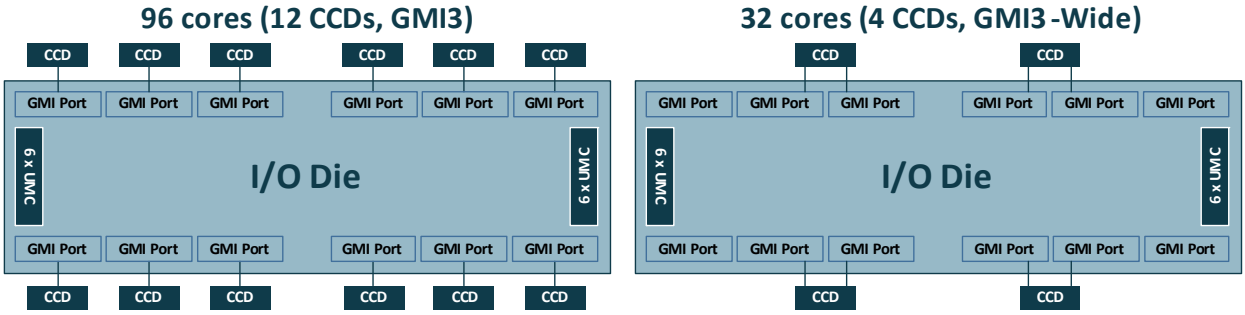

Figure 2-6: Standard vs. Wide GMI links

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.

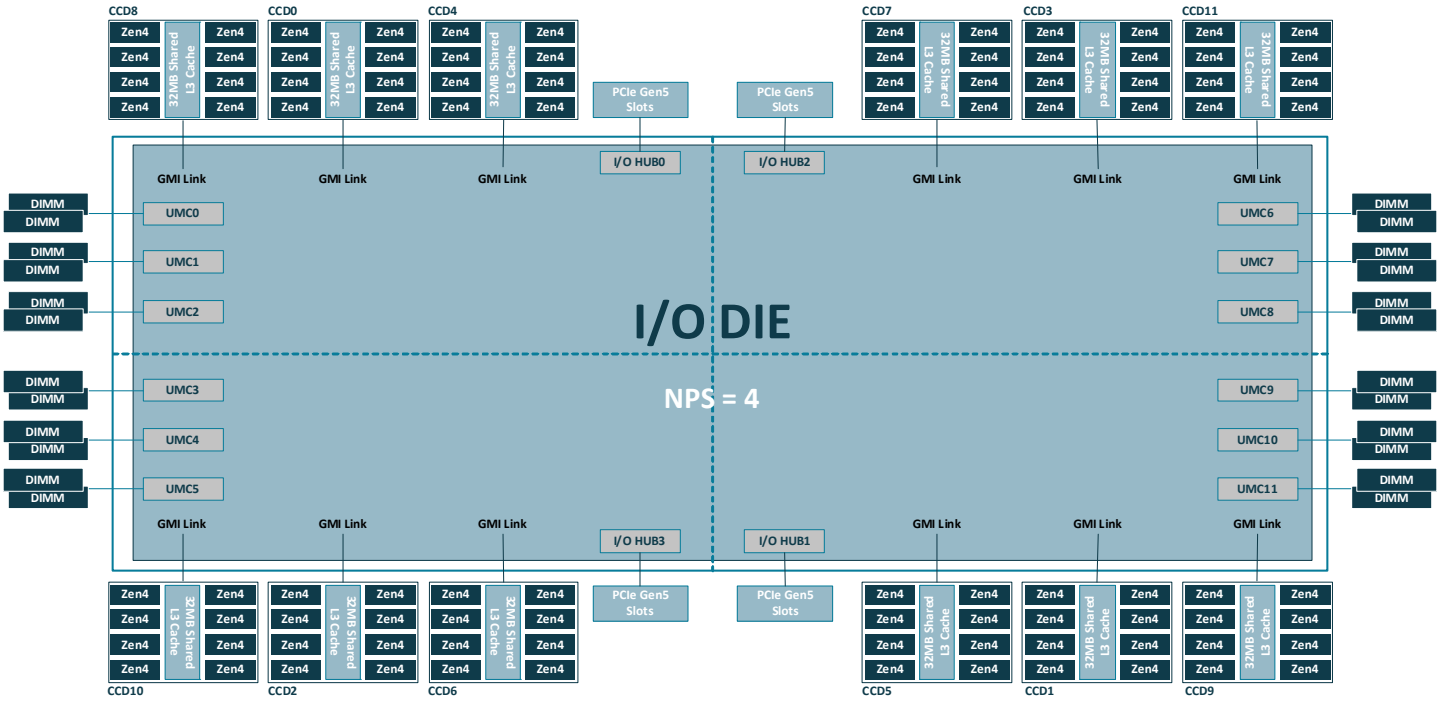

Figure 2-7: The AMD EPYC 9004 SoC consists of up to 12 CCDs and a central IOD for 91xx-96xx models, including "X" OPNs

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.

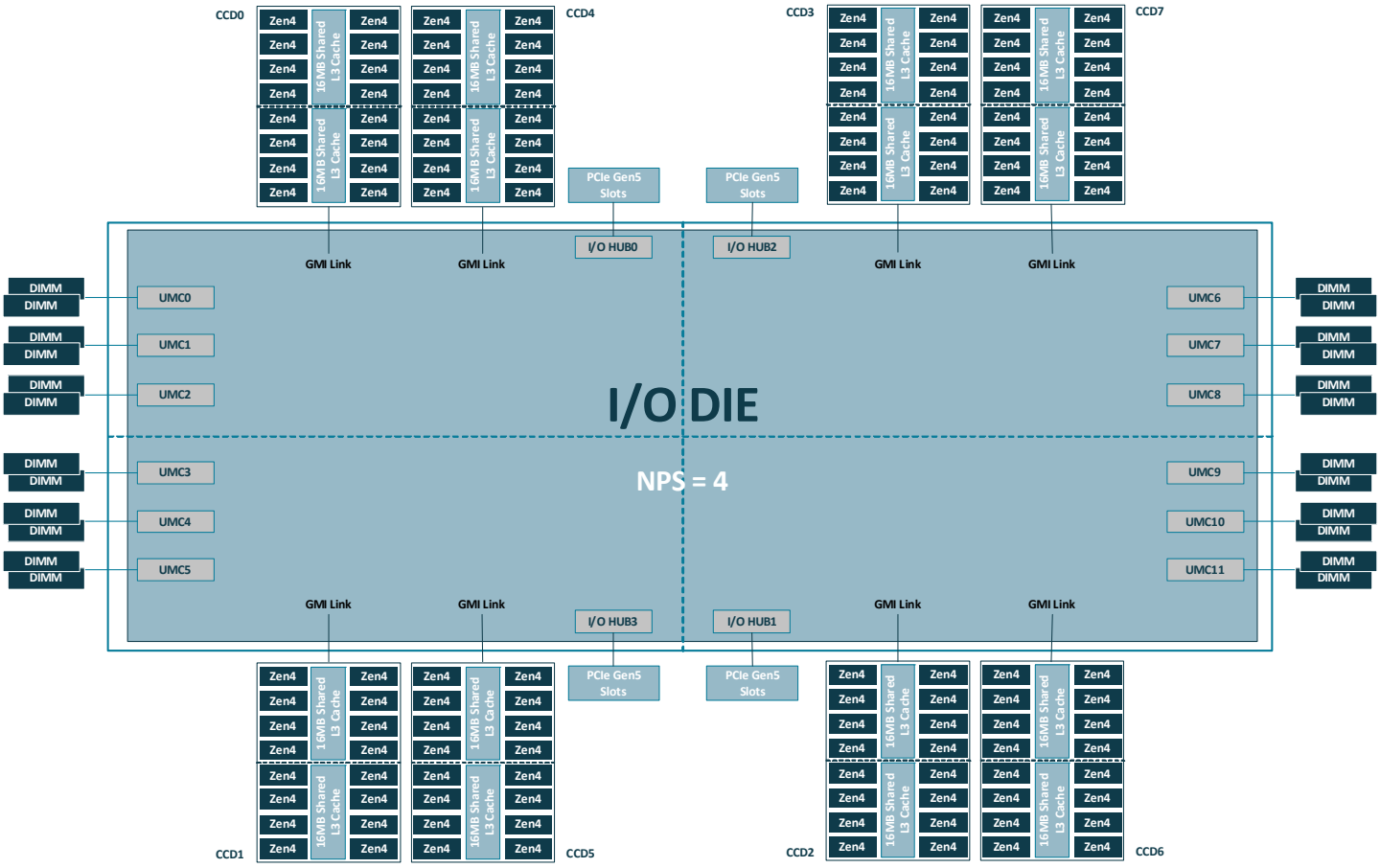

Figure 2-8: The AMD EPYC 9004 System on Chip (SoC) consists of up to 8 CCDs and a central IOD for 97xx models

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.

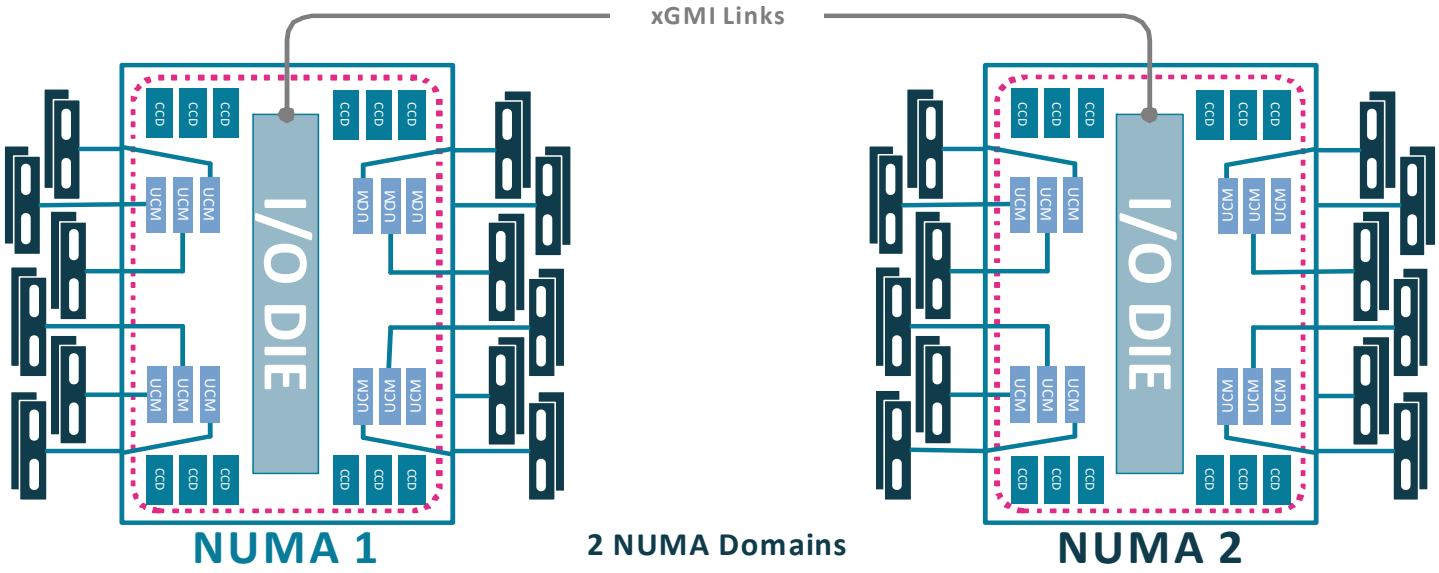

Figure 2-9: Two EPYC 9004 Processors connect through 4 xGMI links (NPS1)

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 System Configuration

4.1 Recommended BIOS Settings

This section describes the recommended BIOS settings for optimal DPDK workload performance on AMD EPYC 9004 Series Processors.

Name	Recommended Value	Description
ACPI Auto Configuration	Disabled	Allows the DPDK library to handle power management.
Local APIC Mode	X2APIC	Set the local APIC mode to **x2APIC = Enabled** to allow the operating system to work with 256 threads and improve performance over legacy APIC
SMT Control	Enable	Enables Symmetric Multithreading (SMT), which allows one hardware thread per core. *Note: Disable SMT if you are running an operating system that does not support X2APIC and has a dual-socket 64 core processor.*
NUMA Node per Socket (NPS)	NPS [1|2|4]	• **NPS1:** Maximum memory bandwidth without NUMA affinity. This is the recommended setting for monolithic application with complete resource provisioning flexibility. • **NPS2:** For multi-tenant VNF/CNF workloads and resource (compute/memory and IO) partitioning. • **NPS4:** This is preferred for low latency and IO throughput applications.
L1/L2 Stream HW Prefetcher	Enable	Enables the L1/L2 Stream HW Prefetcher.
IOMMU	Enable	IOMMU allows operating systems to provide additional protection for DMA capable I/O devices. If needed, you can disable IOMMU in BIOS and enable it via OS options (i.e., `amd_iommu=pt` in the `grub` configuration)
Determinism Control	Manual	Enable the **Determinism Slider** control.
Determinism Enable	Disable Performance Determinism	Ensure maximum performance levels for each CPU in a large population of identically-configured CPUs by only throttling CPUs when they reach the same cTDP. Please see Power/Performance Determinism for more details.

Table 4-1: Recommended common BIOS settings

ACPI SRAT L3 Cache as NUMA Domain	Disable	Do not report each CCX/L3 cache as a NUMA domain to OS.
APB Disable (APBDIS)	1	Disables APB (Algorithm Performance Boost) and enables fixed Infinity Fabric P-state control. By default, the AMD Infinity Fabric selects between a full-power and low-power fabric clock and memory clock based on fabric and memory usage. This transition from low power to full power can increase latency in cases involving low bandwidth but latency-sensitive traffic (and memory latency checkers). Disabling APB by setting APBDIS to 1 and specifying a fixed Infinity Fabric SOC P-state of 0 forces the Infinity Fabric and memory controllers into full-power mode, thereby eliminating any latency jitter.
Power Profile Selection	High Performance Mode	DPDK applications can be heavy in I/O, memory, or compute. Allowing **High-Performance Mode** ensures the CPU and its subsystem will not be throttled.
Global C-state Control	Enabled	Enables CPU C-States for power management.
DF C-States	Disabled	Disables Infinity Fabric C-States.
GMI/XGMI encryption control	Disabled	Controls GMI/XGMI Link encryption
xGMI Force Link Width Control	Forced	Forces the **XGMI Link Width** control.
xGMI Force Link Width	x2	Forces the XGMI link width to the minimum width.
SME	Disable	Disables SME (Secure Memory Encryption). This feature provides hardware-based encryption of all data stored on system DIMMs at a slight increase in memory latency. AMD recommends disabling this feature for high-throughput applications with low security risks.
AVX512	Enabled	Allows DPDK to enable AVX512 ISA for optimized performance.
Monitor and MWAIT Disable	Disabled	This allows DPDK libraries to leverage MWAITX and MONITORX ISA.
Core Performance Boost	Auto (Enabled)	Turns the boost function ON or OFF on all cores
PCIe Speed PMM Control	Auto (0)	Enables the PCIe speed controller: • **0:** Enable the activity based PCIE PMM controller. • **1:** Limit to Gen4 (set the maximum idle link speed to 16 GT/s). • **2:** Limit to Gen5 (set the maximum idle link speed to 32 GT/s, thereby essentially disabling the feature)
Workload Profile		Depending on the workload, choose **Compute**, **Memory**, or **NIC** intensive throughput.
CPPC	Enabled	Enable the CPPC feature.

Table 4-1: Recommended common BIOS settings (Continued)

PCIE Link Speed	Auto	Gen5/Gen4.

Table 4-1: Recommended common BIOS settings (Continued)

4.2 Linux OS Recommended Settings

This section lists some parameters that you can configure for DPDK applications by editing the grub configuration (`/etc/default/grub`).

4.2.1 Huge Pages

The minimum recommended huge pages for DPDK applications running on a bare metal (host) OS are:

- **Single Socket:** `default_hugepagesz=1GB hugepagesz=1G hugepages=12`
- **Dual Socket:** `default_hugepagesz=1GB hugepagesz=1G hugepages=24`

Note: The recommended single-socket setting is multiples of 12. For dual sockets, the recommendation is multiples of 24.

4.2.1.1 IOMMU

Place the IOMMU in passthrough mode (`iommu=pt`) to improve host performance by disabling the DMAR to the memory `iommu=pt amd_iommu=on`.

4.2.1.2 ISOLCPUs

Linux kernel parameters to isolate CPUs from the Kernel scheduler to avoid context switches by preventing non-DPDK workloads from running on reserved cores. You can also specify `nohz_full` can be specified to avoid unbound timer callbacks execution to outside the `nohz_full` range. Similarly, you can specify `rcu_nocbs` to avoid RCU processing on the specified CPU cores. FOr a 96-core scenario:

- **Single Socket:**
  - **SMT (ON):** `isolcpus=8-95,104-191 nohz_full=8-95,104-191 rcu_nocbs=8-95,104-191`
  - **SMT (OFF):** `isolcpus=8-95 nohz_full=8-95 rcu_nocbs=8-95`
- **Dual Socket:**
  - **SMT (ON):** `isolcpus=8-191,200-383 nohz_full=8-191,200-383 rcu_nocbs=8-191,200-383`
  - **SMT (OFF):** `isolcpus=8-191 nohz_full=8-191 rcu_nocbs=8-191`

4.2.1.3 IRQ Affinity

Linux kernel parameter to set the default IRQ affinity mask / set of CPUs (non DPDK cores) that should process interrupts. For a 96-core scenario:

- **SMT (ON):** `irqaffinity=0-7,96-103`
- **SMT (OFF):** `irqaffinity=0-7`

4.2.1.4 C-States

The CPU can idle in several Core-States or C-States:

- **C0:** Active. This is the active state while running an application.
- **C1:** Idle
- **C2:** Idle and power gated. This is a deeper sleep state and will have a greater latency when moving back to the C0 state relative to when coming out of C1.

The recommended settings for maximum power saving state when C-states is enabled in BIOS are:

- **Low latency:** `processor.max_cstate=0`
- **Power management:** `processor.max_cstate=1`

4.2.1.5 Additional Settings

- `nohz=on rcu_nocb_poll numa_balancing=disable transparent_hugepage=never nosoftlockup rcu_nocb_poll` will relieve each CPU from the responsibility of awakening their RCU offload threads.
- `nohz=on` can be used to configure full `dynticks`.

Note: Edit the distro-specific grub configuration file and then execute the `update-grub; reboot` *command.*

4.2.2 Linux Configuration

4.2.2.1 NUMA Awareness

- For optimal performance, AMD recommends placing logical cores, memory, and IO devices in the same NUMA node. Execute the `lstopo-no-graphics` command to both determine the NPS setting and obtain the list of logical cores along with the IO devices within each NUMA node. For example:

```
L3 L4S (32MB)
  L2 L#40 (1024KB) + Lid L140 (32KB) + L11 L140 (32KB) + Core L#40
    PU L#BO (P#64)
    PU L#81 (P#256)
  L2 L#41 (1024KB) + Lid L441 (32KB) + L11 1141 (32KB) + Core L41
    PU L#82 (#65)
    PU L#83 (P#257)
  L2 L#42 (1024KB) + Lid L442 (32KB) + L11 L442 (32KB) + Core L#42
    PU L#84 (P#66)
    PU L#85 (P#258)
  L2 L#43 (1024KB) + Lld L443 (32B) + Ll1 L143 (32KB) + Core L#43
    PU L#86 (P#67)
    PU 1#87 (P#259)
  L2 L#44 (1024KB) + Lld L444 (328) + L11 L44 (32KB) + Core L#44
    PU L#88 (D#68)
    PU L#89 (P#260)
   L2 L#45 (1024KB) + Lid L445 (32KB) + L11 1445 (32KB) + Core L#45
    PU L#90 (P#69)
    PU L#91 (P#261)
  L2 L#46 (1024KB) + Lid L446 (32KB) + Lli L146 (32KB) + Core L#46
    PU L#92 (P#70)
    PU L#93 (P#262)
  L2 L#47 (1024KB) + Lid L447 (32KB) + Lli L147 (32KB) + Core L#47
```

```
    PU L#94 (P#71)
    PU L#95 (P#263)
HostBridge
  PCIBridge PCI 41:00.0 (NVMExp)
    Block (Disk) "nvme0n1"
  PCIBridge PCI 42:00.0 (NVMExp)
    Block (Disk) "nvmeln1"
  PCIBridge PCI 43:00.0 (NVMExp)
    Block (Disk) "nvme2n1"
HostBridge
  PCIBridge
    2 x { PCI 61:00.0-1 (Ethernet) }
  PCIBridge PCI 63:00.0 (Ethernet)
    Net "enp9930"
  PCIBridge
    PCIBridge
      PCI 65:00.0 (VGA)
PCIBridge
  2 x { PCI 67:00.0-1 (SATA) }
```

- For PCIe devices, you can also probe `sysfs` to determine the NUME node:
  `cat /sys/bus/pci/devices/0000\:xx\:00.x/numa_node`

Use `cat /sys/devices/system/node/node*/meminfo | grep HugePages` to determine the hugepage memory per NUMA. For example:

```
Node 0 HugePages_Total:     8
Node 0 HugePages_Free:      4
Node 0 HugePages_Free:      0
Node 1 HugePages_Total:     8
Node 1 HugePages_Free:      7
Node 1 HugePages_Surp:      0
Node 2 HugePages_Total:     8
Node 2 HugePages_Free:      7
Node 2 HugePages_Surp:      0
Node 3 HugePages_Total:     8
Node 3 HugePages_Free:      8
Node 3 HugePages_Surp:      0
```

You can also use the `dpdk` utility `dpdk-hugepages.py -s` command to determine this.

4.2.2.2 Additional Tuning Options

You can also exercise the following tuning knobs to avoid kernel noise. These configurations indirectly affect DPDK workloads because `ISOLCPUs` only prevents user applications from using the dedicated cores, but all kernel and interrupt processing can be triggered on the DPDK dedicated cores. These additional settings allow further fine tuning for asynchronous events. You will need sudo privileges to run the following commands:

- `swapoff -a; ufw disable`
- `echo 0 | tee /proc/sys/vm/zone_reclaim_mode`
- `echo 1 | /proc/sys/vm/drop_caches`
- `systemctl disable irqbalance`
- `sysctl -w kernel.sched_migration_cost_ns=5000000;`

- `sysctl -w kernel.sched_min_granularity_ns=10000000`
- `echo 0 > /sys/kernel/mm/ksm/run`
- `echo "never"> /sys/kernel/mm/transparent_hugepage/enabled`
- `echo "never"> /sys/kernel/mm/transparent_hugepage/defrag`
- `echo 0 > /sys/kernel/mm/transparent_hugepage/khugepaged/defrag`
- `echo -1 > /proc/sys/kernel/sched_rt_period_us`
- `echo -1 > /proc/sys/kernel/sched_rt_runtime_us`
- `echo 10 > /proc/sys/vm/stat_interval`

Along with this, disable the watchdog to reduce overhead:

- `echo 0 > /proc/sys/kernel/watchdog`
- `echo 0 > /proc/sys/kernel/watchdog_thresh`
- `echo 0 > /proc/sys/kernel/nmi_watchdog`

4.2.2.3 Disable Services

You may stop the following optional services if they are not needed:

- `service cryptdisks stop`
- `service cups stop`
- `service mdadm stop`
- `service whoopsie stop`
- `service ufw stop`
- `service speech-dispatcher stop`
- `service ModemManager stop`
- `service lightdm stop`
- `service gdm3 stop`
- `systemctl stop irqbalance`
- `systemctl disable irqbalance`
- `systemctl –w vm.zone reclaim`

Chapter 5 DPDK

This chapter explains how to configure DPDK on servers powered by AMD EPYC 9004 Series Processors.

5.1 Prerequisites

Please see Getting Started Guide for Linux* for information on building DPDK libraries along with the Linux prerequisites . Install all the required packages to compile DPDK.

5.2 Compilation

- **Library mode:** static (recommended for best performance)
- **For a native build:**
  - **Single Socket:** `CC=gcc meson --default-library=static amd_zen4_linuxapp_gcc -Dmax_lcores=192 -Dc_args="-march=znver4 -Ofast"; ninja –C amd_linuxapp_gcc install; ldconfig`
  - **Dual Socket:** `CC=gcc meson --default-library=static amd_zen4_linuxapp_gcc -Dmax_lcores=384 -Dc_args="-march=znver4 -Ofast"; ninja –C amd_linuxapp_gcc install; ldconfig`

Note: AMD EPYC 9004 Series Processors require GCC versions 12.3 onwards or clang 16 onwards to fully leverage the micro-architecture features and optimizations.

Note: Update the number of cores in the `Dmax_lcores` *argument based on the number of available CPU threads.*

- **For building applications with DPDK libraries:**
  - **Static Library mode:**
    ```
    gcc test.c $(pkg-config --static --cflags --libs libdpdk) -o test.exe
    ```
  - **Shared Library mode:**
    ```
    gcc test.c $(pkg-config --cflags --libs libdpdk) -o test.exe
    ```

5.3 Environmental Abstraction Layer (EAL) Options

Select following EAL parameters should be selected based on AMD EPYC platform. See EAL Parameters* for more information.

Function	Command	Description
Logical core number	`-l <core list>` (or) `-c <core mask>`	List of cores to run on. (or) Set the hexadecimal bitmask of the cores to run on.
Number of memory channels	`-n <number of channels>`	Set the number of memory channels to use.
Master logical core number	`-- main-lcore`	Initialize EAL and load environment parameters.
Socket NUMA Huge page memory allocation	`--socket-mem`	Allows fine grain control of huge page allocation for a given NUMA node.
Force maximum SIMD bitwidth	`--force-max-simd-bitwidth=512`	The internal default setting for libraries and PMD is to use 256b SIMD operation. Force 512bit mode improves the application performance by making use of AVX512 optimization in both the libraries and PMDs

Table 5-1: EAL parameter options

5.4 NIC-Specific Tunable Settings

Please see the DPDK NIC performance reports from your NIC vendor(s) for firmware and driver version requirements and for recommended settings. Please also refer to the vendor-specific DPDK Driver* documentation for specific functionalities (e.g., SR-IOV).

5.4.1 Broadcom P2100G

None; please see BNXT PMD* for more details.

5.4.2 Intel E810

Per the DPDK Intel performance reports, AMD recommends building the DPDK PMD with 16B descriptors for optimal performance by passing the `-DRTE_LIBRTE_ICE_16BYTE_RX_DESC` option. Also, force the SIMD bit width to 512 using the EAL option `--force-max-simd-bitwidth`. You can optimize the following PMD options for:

- Low RX latency: `rx_low_latency=1`
- Normal operation: none

5.4.3 Mellanox Cx-6:

The Mellanox PMD makes use of `libibverbs` for device control and configuration via PMD. Thus, the device need not be bound to `vfio-pci` or `igb_uio`. The network Interface is control and maintained by Linux control tools. To use the network interface for DPDK applications, enable specific configurations for high-throughput packet transfer, as described in NVIDIA MLX5*:

- For a 100Gbps port:
  - `ethtool -A $netdev1 rx off tx off`
  - `ethtool -s $netdev1 autoneg off speed 100000`
  - `ethtool -G $netdev1 rx 8192 tx 8192`
  - `ifconfig $netdev1 up`
  - `mlxconfig -d $PORT_PCI_ADDRESS1 set CQE_COMPRESSION=1`
  - `mlxconfig -d $PORT_PCI_ADDRESS1 s PCI_WR_ORDERING=1`
- For 64B throughput (max MPPs) using `testpmd` in IO forward mode with MLX PMD in vector mode (100Gbps)
  - 90 to 95 Mpps can be achieved with 1 logical core and 3 RX/TX queues.
  - 145Mpps can be achieved with 8 logical cores with 8 RX/TX queues.
  - 149Mpps can be achieved with 10 logical cores with 13 RX/TX queue.s

This page intentionally left blank.

Chapter 6 Resources

- DPDK Debug & Troubleshoot Guide*
- Getting Started Guide for Linux*
- DPDK EAL Parameters*
- 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.
- From the AMD Documentation Hub:
  - *BIOS & Workload Tuning Guide for AMD EPYC™ 9004 Series Processors*
  - *Linux® Network Tuning Guide for AMD EPYC™ 9004 Series Processor Based Servers*
  - *Windows® Network Tuning Guide for AMD EPYC™ 9004 Series Processor Based Servers*
  - *VMware® Network Tuning Guide for AMD EPYC™ 9004 Series Processor Based Servers*
  - *Ubuntu® Tuning Guide for AMD EPYC™ 9004 Series Processor Based Servers*
- NIC and ecosystem vendor-specific tuning guided for the AMD EPYC platform:
  - NVIDIA (Mellanox)*
  - Broadcom*
  - RedHat Network Guide*

This page intentionally left blank.

Chapter 7 Glossary

- **ACPI** - Advanced Configuration and Power Interface
- **AVX -** Advanced Vector Extensions
- **BIOS** - Basic Input/Output System
- **BMC -** Baseboard Management Controller
- **CCD** – Core Complex Die
- **CCX** – Core Complexes
- **cTDP** - Configurable Thermal Design Power
- **DIMM** – Dual In-line Memory Module
- **DPC** – DIMMs Per Channel
- **DPDK -** Data Plane Development Kit
- **DRAM** – Dynamic Random-Access Memory
- **IOMMU -** Input-Output Memory Management Unit
- **IRQ -** Interrupt Request
- **LLC** - Last Level Cache
- **NDA -** Non-Disclosure Agreement
- **NIC** – Network Interface Card
- **NUMA** - Non-Uniform Memory Access
- **PPL** - Package Power Limit
- **OEM -** Original Equipment Manufacturer
- **OPN** – Orderable Part Number
- **OS** - Operating System
- **SLIT** - System Locality Information Table
- **SMT** - Symmetric Multithreading

- **SRAT** - System Resource Affinity Table
- **TCO** – Total Cost of Ownership
- **TDP** - Thermal Design Power

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:

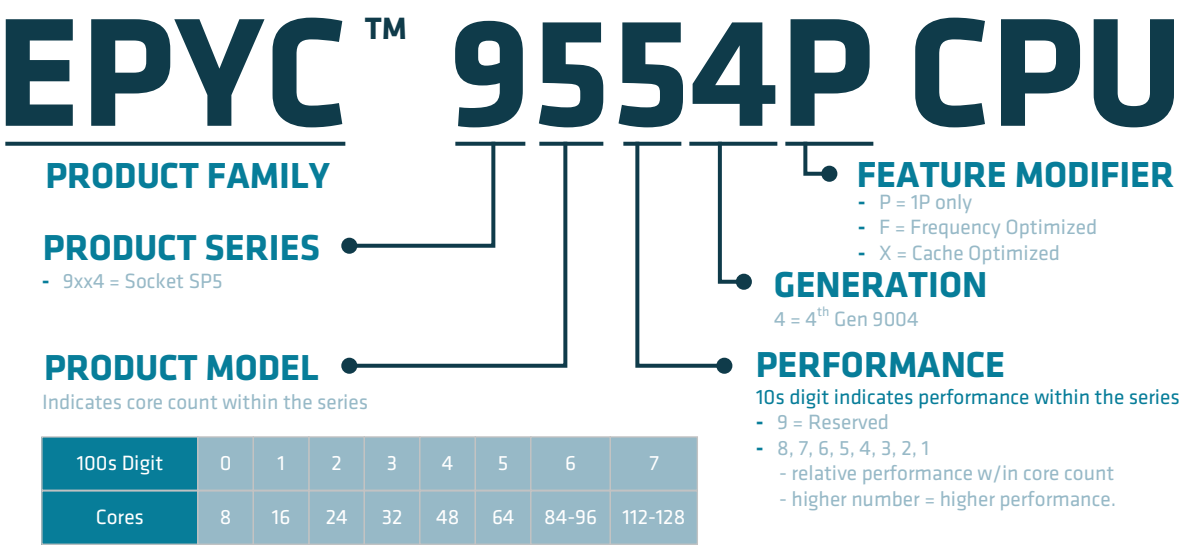

Figure 8-1: AMD EPYC SoC naming convention

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.