AMD EPYC

TUNING GUIDE
AMD EPYC 9004

Linux® Network

Publication	58012
Revision	1.5
Issue Date	February, 2024

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

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

Trademarks

AMD, the AMD Arrow logo, AMD EPYC, 3D V-Cache, and combinations thereof are trademarks of Advanced Micro Devices, Inc. Linux is a registered trademark of Linus Torvalds. PCIe is a registered trademark of PCI-SIG Corporation. Other product names and links to external sites used in this publication are for identification purposes only and may be trademarks of their respective companies.

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

Date	Version	Changes
July, 2022	0.1	Initial NDA partner release
Sep, 2022	0.2	Updated BIOS and testing information
Nov, 2022	1.0	Initial public release
Dec, 2022	1.1	Added 128-core and AMD 3D V-Cache™ technology information
Mar, 2023	1.2	Added 97xx OPN and AMD 3D V-Cache™ technology information
Jun, 2023	1.3	Second public release
Jan, 2024	1.4	Added new network cards, minor errata corrections
Feb, 2024	1.5	Minor updates and errata fixes

Audience

This document is intended for a technical audience with a server configuration background who have:

- Admin access to the server's management interface (BMC).
- Familiarity with the server's management interface.
- Admin OS access.
- Familiarity with the OS-specific configuration, monitoring, and troubleshooting tools.

Author

Steve Rochefort

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 an overview of steps needed to tune your chosen network adapters for optimal performance in a platform powered by AMD EPYC™ 9004 Series Processors running Linux®, including the steps taken by AMD engineers to prepare the reference platform for maximum performance. If you are testing a system powered by AMD EPYC 9004 Series Processors that was designed by another company, then be sure to also review the vendor product documentation to achieve optimum results.

There is no single golden rule for tuning a network interface card (NIC) for all conditions. Different adapters have different parameters that can be changed. Operating systems also have settings that can be modified to help with overall network performance. Depending on the exact hardware topology, you may have to make different adjustments to optimize a specific workload. With Ethernet speeds going higher, up to 400 Gbps, and the number of ports being installed in servers growing, these tuning guidelines become even more important to achieve the best performance possible.

This guide does not provide exact settings for modifying every scenario. Rather, it includes parameters to check and modify for a given configuration. In this guide, the steps are focused on TCP/IP network performance. Table 5-1 provides tables of recommended tuning parameters used. Review the block diagrams and descriptions of the AMD EPYC™ 9004 processor NUMA architecture in the following sections before you begin tuning:

- "Memory and I/O" on page 8
- "NUMA Topology" on page 10
- "Dual-Socket Configurations" on page 12

All I/O uses data transfers into or out of memory, hence the I/O bandwidth can never exceed the capabilities of the memory subsystem. Therefore, before you start, verify that your memory subsystem is properly configured for maximum frequency. To reach maximum memory bandwidth on modern CPUs, you must populate one DIMM in every DDR channel. For AMD EPYC™ 9004 Series Processor-based servers, there are twelve DDR5 channels on each CPU socket. For a single-socket platform, populate all twelve memory channels. Likewise, on a dual-socket platform, you should populate twenty-four memory channels.

In addition, AMD recommends consulting the tuning guide available from your NIC vendor. Each vendor decides which standard commands they support and may have also created their own value-added commands to support. As examples: A vendor may support interrupt coalescing or not. Another vendor may support relaxed ordering of PCI transactions while another does not.

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.

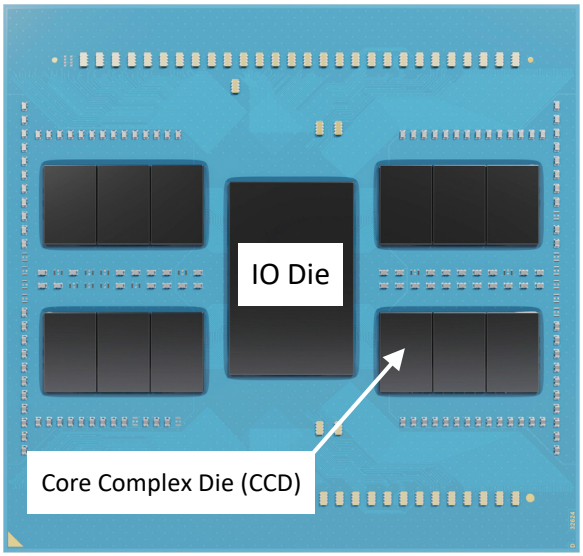

Figure 2-1: AMD EPYC 9004 configuration with 12 Core Complex Dies (CCD) surrounding a central I/O Die (IOD)

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.

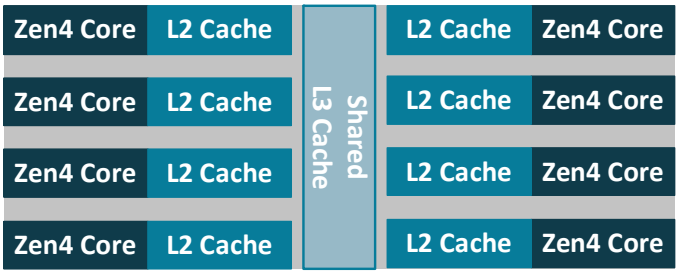

Figure 2-2: Top view of 8 compute cores sharing an L3 cache (91xx-96xx models)

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.

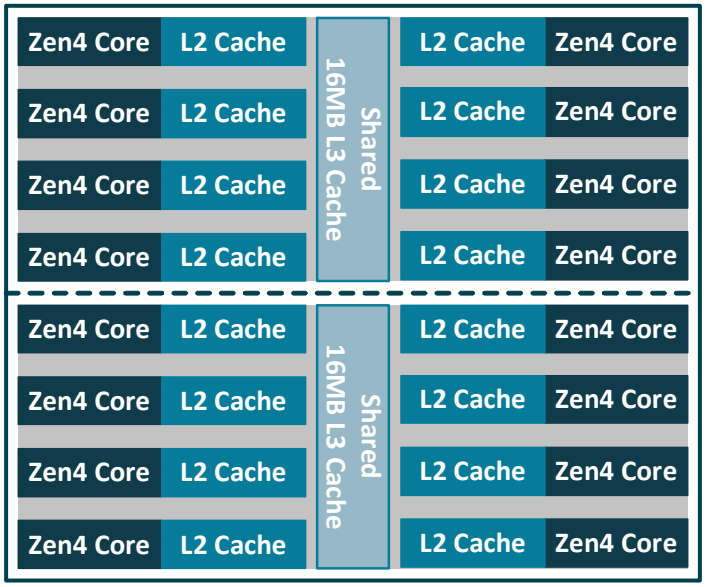

Figure 2-3: 2 CCXs in a single 4th Gen AMD EPYC 97xx CCD

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.

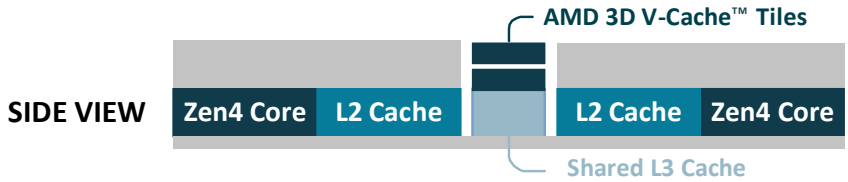

Figure 2-4: Side view of vertically-stacked central L3 SRAM tiles

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.

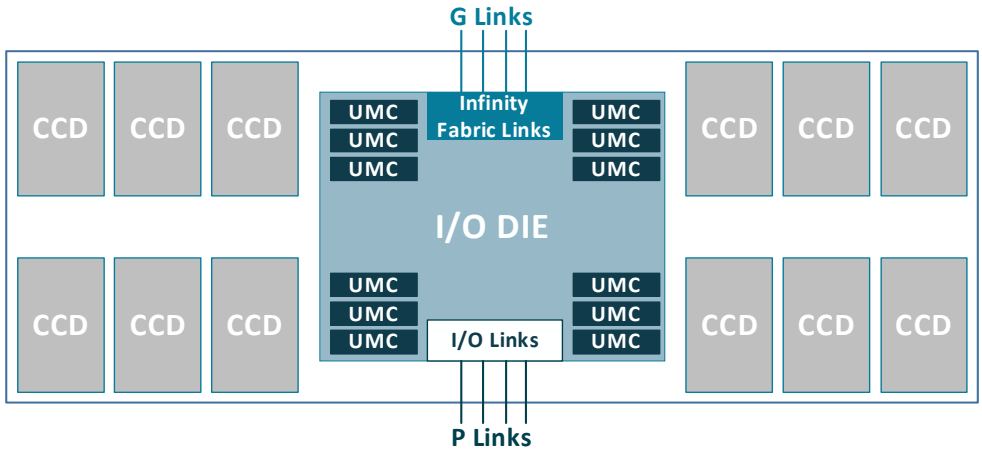

Figure 2-5: AMD EPYC 9654 processor internals interconnect via AMD Infinity Fabric (12 CCD processor shown)

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.

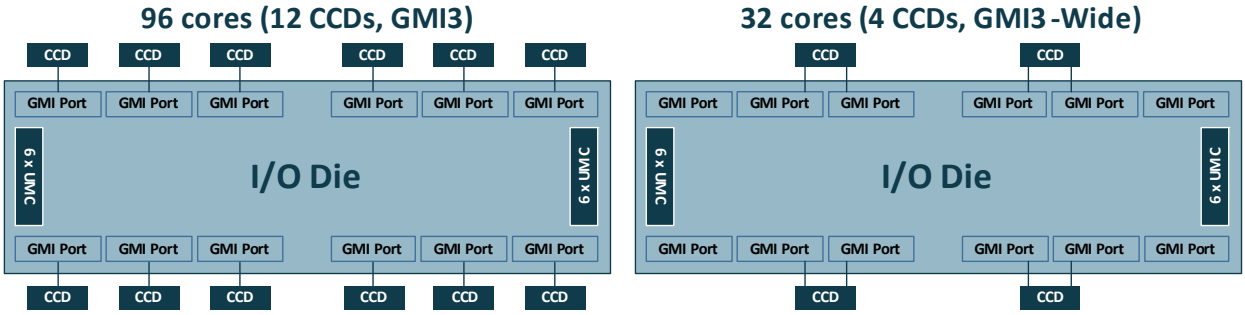

Figure 2-6: Standard vs. Wide GMI links

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.

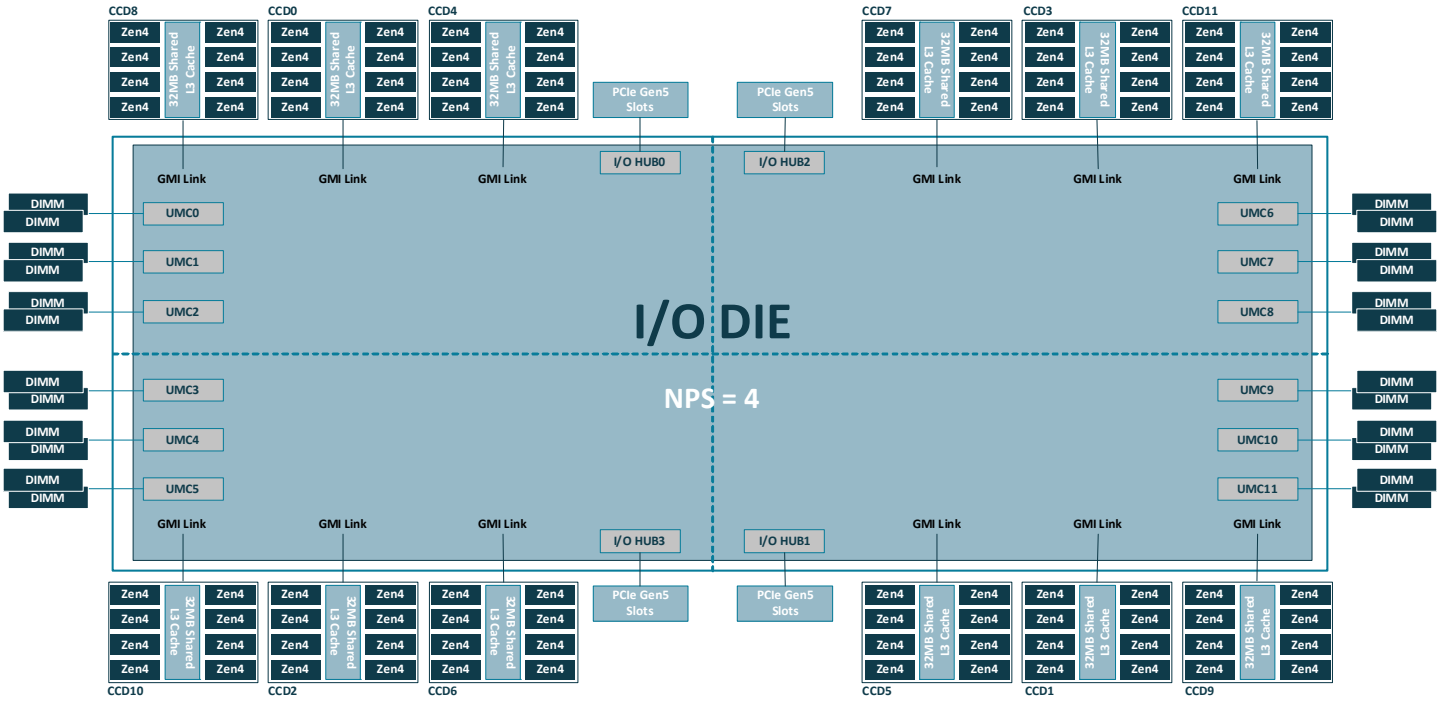

Figure 2-7: The AMD EPYC 9004 SoC consists of up to 12 CCDs and a central IOD for 91xx-96xx models, including "X" OPNs

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.

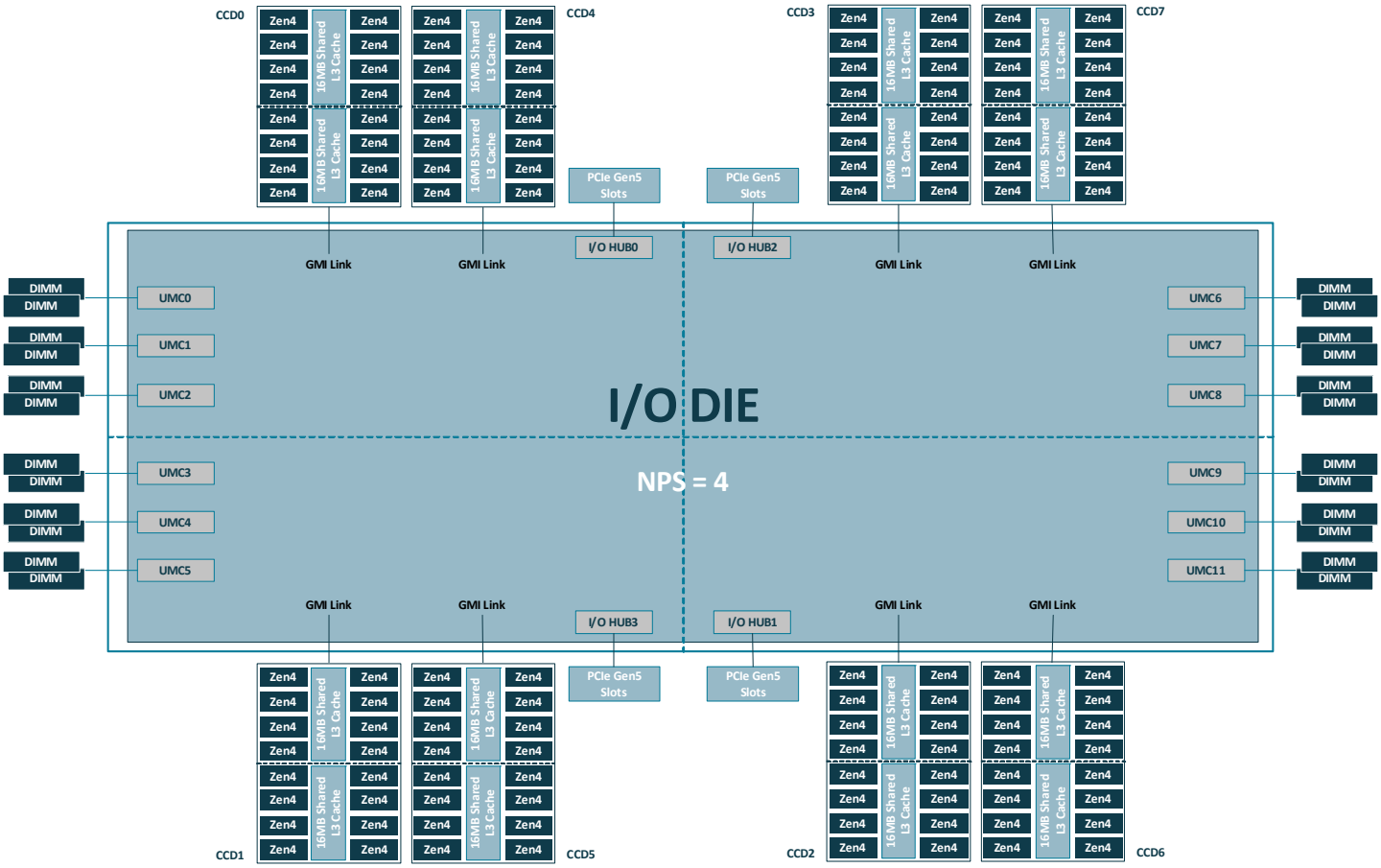

Figure 2-8: The AMD EPYC 9004 System on Chip (SoC) consists of up to 8 CCDs and a central IOD for 97xx models

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.

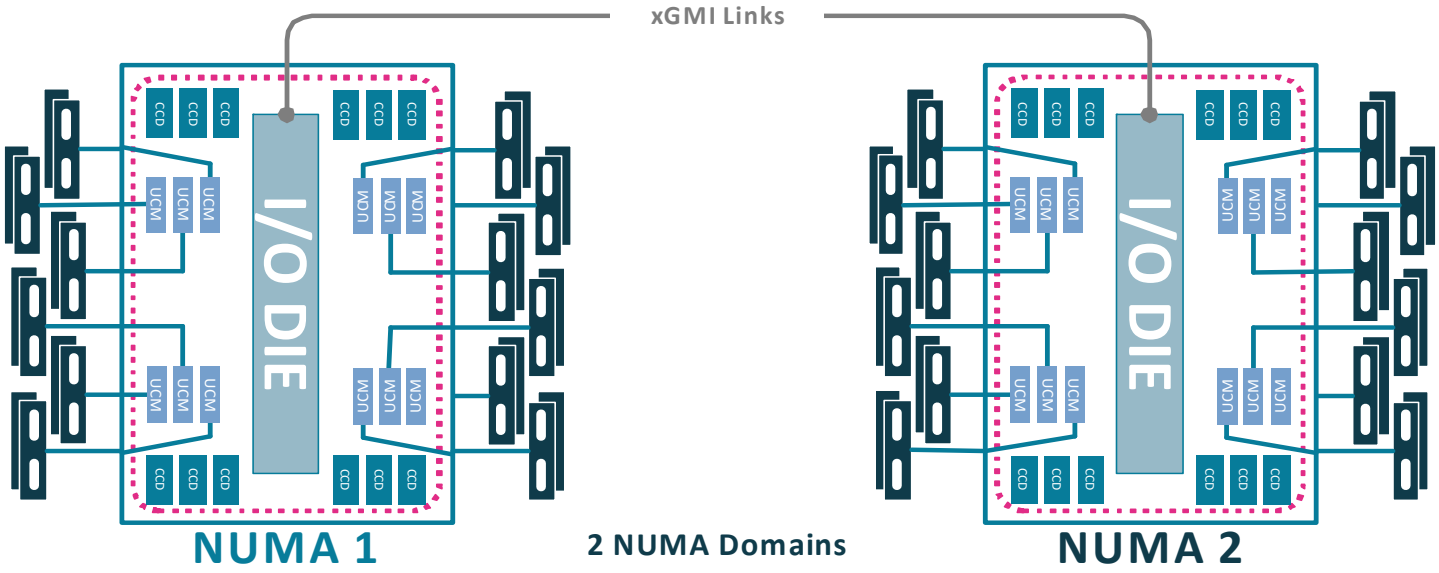

Figure 2-9: Two EPYC 9004 Processors connect through 4 xGMI links (NPS1)

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 TCP Performance Tuning

This chapter addresses test configuration, BIOS tuning, network adapter tuning and OS tuning.

4.1 Test Configuration

The testing performed when creating this Tuning Guide used two reference systems powered by AMD EPYC 9004 Series Processors and equipped with very high-speed network ports that are faster than the switches available in some labs. AMD engineers connected these two systems to each other to directly pass data between them, as shown in Figure 4-1.

Figure 4-1: Direct connections between the AMD reference systems used when creating this Tuning Guide

Note: It is important to use two identically-configured systems.

4.2 Single- and Dual-Socket Systems

AMD generally measures traffic passing between two identical network adapters plugged into PCIe slots in reference boards. The network adapter should only use local resources regardless of whether the processor it is connected to is installed in a single- or dual-socket system.

Determining which socket the adapter is connected to and which NUMA node the adapter is in is a standard first step when preparing to run tests. This ensures that the adapter is only using local cores and memory when passing traffic. Performance will be sub-par if the adapter uses cores and/or memory on the distant socket.

Dual-socket systems use xGMI links between the sockets. You can dynamically reduce the xGMI width to save power. It is a good idea to force the XGMI link to the maximum width supported by your board, but the most important safeguard is to verify that your test script is NUMA aware and uses local resources.

4.3 BIOS Tuning

It is a good practice to start fresh by loading the optimized default BIOS settings before beginning the tuning process. This is especially true if you are sharing a system with other users. Resetting the BIOS to default can also be faster than manually changing BIOS settings, especially if you are not certain what those defaults are. It can also be a time saver to manually set BIOS settings if you are not sure what the default settings are. For example, your BIOS may default to an **Auto** memory speed instead of the maximum available speed.

Note: This section presents BIOS settings as they appear in the default AMD BIOS. Different OEMs may modify the names and/or locations of these settings.

4.3.1 Numa Nodes Per Socket (NPS)

The BIOS **NPS** setting allows you to make a trade-off between minimizing local memory latency for NUMA-aware or highly parallel workloads versus maximizing per-core memory bandwidth for non-NUMA friendly workloads. Setting NPS=1 interleaves all 12 memory channels on a socket.

AMD standard BIOS defaults to NPS=1 with **LLC as NUMA** disabled. This setting reports one NUMA node per socket to the operating system. However, using NPS=1 with LLC as NUMA enabled means the OS will see one NUMA node per L3 cache and still use 12-channel interleaving. The combination of NPS=1 and **LLC as NUMA** enabled is usually the best combination for NIC tuning. If you are concerned about latency, then set NPS=2 to interleave 6 memory channels.

4.3.2 Last Level Cache (LLC) as NUMA Domain

AMD EPYC processors use multiple Last Level Caches (LLCs, or L3 caches). Operating systems can handle multiple LLCs and schedule jobs accordingly; however, the AMD BIOS **LLC as NUMA** setting allows creating a single NUMA domain per LLC. This can help the operating system schedulers maintain locality to the LLC without causing unnecessary cache-to-cache transactions. Please see the latest versions of the Socket SP5 Platform NUMA Topology for AMD Family 19h Models 10h–1Fh (login required) and the *BIOS & Workload Tuning Guide for AMD EPYC™ 9004 Series Processors* (available from AMD EPYC Tuning Guides) for additional information.

Advanced > AMD CBS > DF Common Options > ACPI > ACPI SRAT L3 Cache as NUMA Domain > Enable

4.3.3 SMT

Symmetric Multithreading (SMT) is enabled by default. To take measurements with SMT disabled:

Advanced > AMD CBS > CPU Common Options > Performance > SMT Control > Disable

4.3.4 X2APIC

AMD EPYC 9004 Series Processors include an x2APIC controller. This has two benefits:

- Allows operating systems to work with the 384 CPU threads now available on AMD platforms.
- Provides improved performance over the legacy APIC AMD recommends without requiring you to enable the x2APIC mode in BIOS, even for lower core counts.

This option should be selected by default. To set it manually:

Advanced -> AMD CBS -> CPU Common Options -> Local APIC -> x2APIC

4.3.5 Determinism Control and Slider

The **Determinism** BIOS setting can affect throughput,

Advanced > AMD CBS > NBIO Common Options > SMU Common Options > Determinism Control > Manual > Determinism Enable > Enable Performance

4.3.6 10-Bit Tag

The **PCIE Ten Bit Tag Support** setting increases the maximum number of non-posted requests from 256 to 768 and sometimes helps with high bandwidth port throughput. AMD CRBs implement the 10-bit tag by default.

Advanced > AMD CBS > NBIO Common Options > PCIE Ten Bit Tag Support > Enabled

4.3.7 Memory Clock Speed

Setting the memory clock speed to **Auto** should select the maximum memory speed using default BIOS settings; however, you are welcome to manually set the speed in BIOS. Either way, be sure to check the memory speed from the operating system before starting performance tests. The `dmidecode` and `lsmem` commands are very useful.

Advanced > AMD CBS > UMC Common Options > DDR Timing Configuration > Accept > Memory Target Speed > 4800

4.3.8 Slot Bifurcation

The Intel® E810-2CQDA2 requires slot bifurcation for full performance. This is the only card tested that has this requirement. AMD Quartz systems have all four PCIE slots connected to Socket 0. You can change the bifurcation for slot1 from 16x1 to 8x2 as follows:

Advanced > AMD CBS > CRB Board > Socket 0 Slot Info Override > BFFF instead of FFFF

4.4 Network Adapter Tuning

AMD strongly recommends that you disable firewalls and install a fresh copy of the operating system on your EPYC platform, being sure to install the latest NIC vendor firmware and drivers before proceeding. Be sure to review the installation for errors. Be sure that the OS has access to the network during installation and can download anything needed. Lastly, if you are adopting a script that someone else wrote or that worked on another CPU platform or with another NIC vendor's product, then be prepared to debug it. These are common mistakes.

4.4.1 Local NUMA Node Usage

Ensure maximum performance by using cores and memory that are in the same NUMA node as your network adapter. There are different ways to check to see the core and NUMA node assignments, such as by executing the `lscpu` command. Other useful commands are:

- `cat /sys/class/net/(ethernet interface)/device/numa_node`
- `cat /sys/class/net/(ethernet interface)/device/local_cpulist`

4.4.2 Controlling IRQ

Controlling the number of interrupt queues used within a NUMA node is a key element in network performance because you do not want to have more interrupt queues than you have CPU cores. Assigning multiple queues to a single core risks thrashing the interrupt handler swapping between queues. It is more efficient to have a single queue per core to eliminate that thrashing.

Before making assignments, stop and then disable `irqbalance` altogether.

There are three types of interrupt queues:

- Receive (RX)
- Transmit (TX)
- Combined. This uses a single queue to handle both RX and TX interrupts. Some vendors still have separate RX and TX queues while others implement only combined queues. For example, if you have a one-port NIC that combines the RX and TX interrupts, then you could execute the following command:

  ```
  ethtool -L enp33s0f0 combined 16 tx 0 rx 0
  ```

Make sure to review the default settings before you begin making changes. In the above example, you can read the default setting by executing the `ethtool -I enp33s0f0` command.

4.4.3 TX/RX Flow Steering

Advanced Receive Flow Steering (aRFS) is an effective way to direct incoming packets to the cores where the application is running. Use this If supported by your NIC.

4.4.4 TX/RX Queue Size

The NIC Queue size or ring size represents the number of buffers that a NIC uses to DMA data into system memory. Increasing the TX and RX queue size will help prevent dropped packets. The `ethtool` utility allows you to find the current ring size and maximum allowed ring size. For example:

```
root@testsystem:~# ethtool -g enp33s0f0
Ring parameters for enp33s0f0:
Pre-set maximums:
RX:             8192
RX Mini:        0
RX Jumbo:       0
TX:             8192
Current hardware settings:
RX:             512
RX Mini:        0
RX Jumbo:       0
TX:             512
Ethtool -G (device interface) tx 2047 rx 2047
```

You can then set the buffer sizes to the maximum value, a value recommended by your NIC vendor, or one that you find works best. For example:

```
ethtool -G (device interface) tx 2047 rx 2047
```

After issuing the `ethtool` command with a `-G` to set ring size, you should reissue the command with `-g` to verify that the change was accepted.

4.4.5 Relaxed Ordering

3rd Gen and prior AMD EPYC processors included the **Preferred I/O** and **Relaxed Ordering** settings that helped optimize network and disk I/O performance. 4th Gen AMD EPYC processors (9xx4 models) include architectural enhancements that deliver optimal network and disk I/O performance by default without the need for either of these features.

4.4.6 LRO

Many NIC vendors have tuning guides to help end users optimize for specific use cases. Those use cases usually involve optimizing for either the highest possible throughput or the lowest possible latency possible but can only rarely achieve both at the same time. Enabling Large Receive Offload (LRO) is a common way to improve network throughput performance. Consult your NIC documentation for information on enabling LRO on your adapter. Some providers use standard commands available via `ethtool`, while others might have extra parameters that are needed to fully enable LRO. Be aware that enabling LRO helps improve your throughput but could negatively impact network latency. Make sure to tune as appropriate for your workload.

4.4.7 TX-No Cache Copy

Either check the default value for your device driver setting or proactively shut it off as shown here:

```
Ethtool -K $[local interface name] tx-nocache-copy off
```

4.5 OS Tuning

AMD recommends that you perform testing using a fresh copy of an operating system that understands and support the AMD EPYC 9004 Series Processor.

4.5.1 IOMMU Settings

The Linux kernel is constantly updated. The official mainline releases from The Linux Foundation are available via http://kernel.org*. However, common enterprise level Linux distributions rarely use a mainline kernel.

AMD has contributed code into the Linux kernel for years. The most recent focus has been enabling the "Zen," "Zen 2," "Zen 3," and "Zen 4" architectures contained in AMD EPYC processors. One area of code contribution focuses on optimizing the input-output memory management unit (IOMMU) code for AMD EPYC processors. These IOMMU patches can have a direct impact on TCP/IP performance, even in a bare metal (non-virtualized) environment.

Either disabling the IOMMU or setting it to pass through mode (which disables DMAR to memory) can sometimes benefit the highest bandwidth adapters, such as 200 or 400 Gbps Ethernet adapters. To set the IOMMU to pass-through mode, the following kernel parameter must be passed in during boot time in the `grub` command line:

```
iommu=pt
```

After booting the system, check the setting by executing the following command:

```
Cat /proc/cmdline | grep -I iommu=pt
```

4.5.2 Nohz

Nohz is another boot time parameter that can be included in the `grub` command line to disables `dyntick idlemode`.

```
nohz=off
```

4.5.3 IRQ Balancing

CPUs generally automatically share interrupts but this can cause delayed interrupt processing. To disable this:

```
systemctl disable irqbalance
```

4.5.4 TCP Memory Configuration

Increasing the memory buffer size for TCP sockets can help eliminate transmission gaps when lots of data is in flight and device buffers are full. The following examples define minimum, nominal, and maximum TCP socket buffer values:

```
echo "4096 131072 268435456" > /proc/sys/net/core/tcp_rmax
```

```
echo "4096 131072 268435456" > /proc/sys/net/core/tcp_wmax
```

4.5.5 Scaling Governor

Set the CPU scaling governor to Performance mode by executing the following command:

```
Echo performance  | sudo tee /sys/devices/system/cpu/cpu*/cpufreq/scaling_governor
```

Chapter 5 Additional Information

5.1 Recommendations and Results

Table 5-1 provides recommended values for each of the options described in this Tuning Guide. As shown, not all adapters require modifying the default BIOS, OS, or adapter settings. You may also find that some drivers already enable a feature (such as relaxed ordering) by default. You should verify all the settings listed before making any assumptions about default settings.

	Dual Port 25 Gbps Ethernet	Single Port 100 Gbps Ethernet	Dual Port EDR InfiniBand	Dual Port 100 Gbps Ethernet	Single Port NDR InfiniBand
BIOS Options					
Local APIC Mode	default	x2apic	x2apic	x2apic	x2apic
Determinism mode	default	performance	performance	performance	performance
LLC as NUMA	default	enabled	enabled	enabled	enabled
10-bit tag	default	enabled	enabled	enabled	enabled
Adapter Options					
Relaxed Ordering	default	enabled	enabled	enabled	enabled
OS Options					
Ring Buffers	default	maximum	maximum	maximum	maximum
Large Receive Offload (lro)	default	enabled	enabled	enabled	enabled
Interrupts	default	combined 16	combined 16	combined 16	combined 16
MTU (default 1500)	default	default	default	default	default

Table 5-1: NIC configuration recommendations

Table 5-2 lists several adapters that have been tested with RHEL 9.0. Following the guidelines contained in this Tuning Guide should yield near-line rate performance with any NIC you choose.

Note: All adapters were tested using the AMD "Quartz" reference design with BIOS version RQZ100BD.

Tested Adapter	Port Speed	Product Description
Broadcom BCM957414A4142CC	25 Gbps	Dual-Port 25 Gbps Network Interface Card
Broadcom BCM957508-P2100G	100 Gbps	Dual-Port 100 Gbps Network Interface Card
Broadcom BCM957508-P2200G	200 Gbps	Dual-Port Network Interface Card (one port x 200Gbps or two ports x 100Gbps)
Broadcom BCM957608-P2200G	200 Gbps	Dual-Port 200 Gbps Network Interface Card
Broadcom BCM957608-P1400G	400 Gbps	Single-Port 400 Gbps Network Interface Card
Cornelis Networks 100HFA016LS	100 Gbps	Single-Port 100 Gbps Omni-Path Host Fabric Adapter
Intel E810-2CQDA2	100 Gbps	Intel Dual-Port 100 Gbps Ethernet Adapter
NVIDIA MCX512A-ACAT	25 Gbps	ConnectX-5 EN Dual Port Ethernet Adapter
NVIDIA MCX653106A-HDAT	200 Gbps	ConnectX-6 VPI Dual Port InfiniBand & Ethernet Adapter
NVIDIA MCX75310AAS-NEAT	400 Gbps	ConnextX-7 Single Port IB Adaptor
AMD X2541	40 Gbps	Single Port 40 Gbps Ultra Low Latency Ethernet Adapter
Testing completed with RHEL 9.0.		

Table 5-2: Tested network adapters

Chapter 6

Processor Identification

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

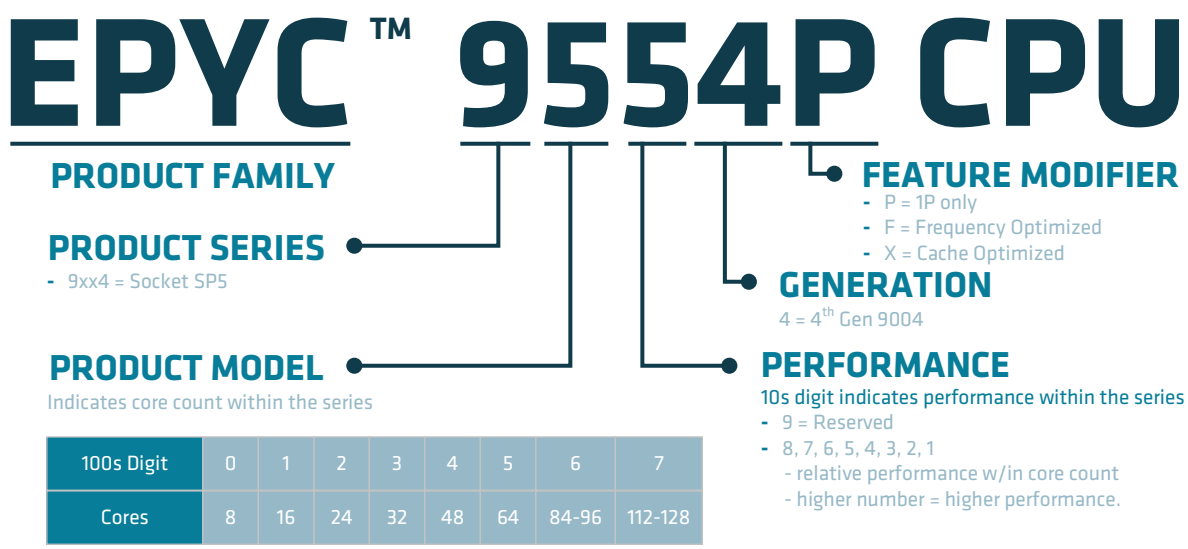

Figure 6-1: AMD EPYC SoC naming convention

6.1 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.

6.2 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*.

6.2.1 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.