AMD EPYC

TUNING GUIDE

AMD EPYC™ 9005

Data Plane Development Kit (DPDK) Tuning Guide

AMD

together we advance_data center computing

PID: 58468
v1.0
October 2024

Manish Kumar and Vaibhav Patankar

DATE	VERSION	CHANGES
June, 2024	0.1	Initial NDA release
October, 2024	1.0	Initial public release

AUDIENCE

This document 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.

DATA PLANE DEVELOPMENT KIT TUNING GUIDE

CONTENTS

THIS PAGE INTENTIONALLY LEFT BLANK.

Chapter 1: Introduction

This tuning guide provides various system configuration parameters that can optimize DPDK workload performance on servers based on AMD EPYC™ 9005 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 configurations
- 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.

1.1 - Important Reading

Please be sure to read the following guides (available from the AMD Documentation Hub), which contain important foundational information about 5th Gen AMD EPYC processors:

- *AMD EPYC™ 9005 Processor Architecture Overview*
- *BIOS & Workload Tuning Guide for AMD EPYC™ 9005 Series Processors*
- *Memory Population Guidelines for AMD EPYC™ 9005 Series Processors*

THIS PAGE INTENTIONALLY LEFT BLANK.

Chapter 2: System Configuration

2.1 - Recommended BIOS Settings

This section describes the recommended BIOS settings for optimal DPDK workload performance on AMD EPYC 9005 Series Processors.

Name	Recommended Value	Description
Enable ACPI Auto Configuration	Disabled	Allows the DPDK library to manage power.
SMT Control	Enabled	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.*
L1 Stream HW Prefetcher	Enabled	Enables the L1 Stream HW Prefetcher.
L2 Stream HW Prefetcher	Enabled	Enables the L2 Stream HW Prefetcher.
Core Performance Boost	Enabled	Enables the boost function on all cores.
Global C-state Control	Enabled	Enables CPU C-states for power management. Controls IO-based C-state generation and DF C-states.
Local APIC Mode	x2APIC	Sets the local APIC mode to **x2APIC=Enabled** to allow the operating system to work more efficiently on high core count configuration and improve performance over legacy APIC.
SMEE	Disabled	Disables Secure Memory Encryption (SMEE). 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 the AVX512 ISA for optimized performance.
Monitor and MWAIT Disable	Disabled	Allows DPDK libraries to leverage MWAITX and MONITORX ISA.
FP512	Enabled	Enables 512bit datapath for the AVX512 ISA. Disable this option when **AVX512** is **Disabled**.
NUMA Nodes per Socket (NPS)	NPS 1	• **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.

Table 2-1: Recommended common BIOS settings

ACPI SRAT L3 Cache as NUMA Domain	Disabled	Do not report each CCX/L3 cache as a NUMA domain to OS. • **Enabled:** Each CCX in the system will be declared as a separate NUMA domain. • **Disabled:** Memory Addressing/NUMA nodes per socket will be declared.
GMI Encryption Control	Disabled	Controls GMI Link encryption.
xGMI Encryption Control	Disabled	Controls xGMI Link encryption.
SDCI	Enabled	Enables the Smart Data Cache Injection (SCDI) feature, if supported by the NIC.
Determinism Control	Manual	Enable the **Determinism** control.
Determinism Enable	Performance	Ensures 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.
xGMI Force Link Width Control	Force	Forces the **xGMI Link Width** control.
xGMI Force Link Width	2	Forces the xGMI link width to x16.
APBDIS	1	Disables APB (Algorithm Performance Boost) and enables fixed Infinity Fabric P-state control. • **0 (Disable APBDIS):** Locks the fabric clock to the non-boosted speeds. • **1 (Enable APBDIS):** Unlocks the fabric clock to the boosted speeds.
DfPstate	0	[0-2]
Power Profile Selection	High Performance Mode	Select the power profile based on your workload requirements. Setting **High Performance Mode** ensures that the CPU and its subsystems will not be throttled. • **0:** High Performance Mode • **1:** Efficiency Mode • **2:** Maximum IO Performance Mode • **3:** Balanced Memory Performance Mode • **4:** Balanced Core Performance Mode • **5:** Balanced Core Memory Performance Mode • **6:** Auto
DF Cstates	Disabled	Disables Infinity Fabric C-States.
CPPC	Enabled	Enables the **CPPC** feature.
PCIe Ten Bit Tag	Enabled	Enables PCIe Ten Bit Tag for supported devices. This increases the number of non-posted requests from 256 to 768 for better performance. The increase in unique tags is required to maintain peak performance as latency increases.
PCIE Link Speed	Auto	You can force a speed by manually setting **Gen5** or **Gen4** based on NIC capability. Setting **Gen5** and using a PCIe Gen5 NIC is preferred for optimum system performance.
IOMMU	Enabled	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)

Table 2-1: Recommended common BIOS settings (Continued)

2.2 - Linux OS Recommended Settings

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

2.2.1 - 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.

2.2.1.1 - 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`.

2.2.1.2 - 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`

2.2.1.3 - IRQ Affinity

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

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

2.2.1.4 - 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`

2.2.1.5 - Additional Settings

- `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` configures full `dynticks`.
- `pci=norom` disables the default PCI expansion ROM memory resource allocation.
- `pcie_aspm=off` disables PCIe Active State Power Management.
- `amd_pstate=passive` enables passive `amd-pstate` for driver operation modes.
- `audit=0` means the audit will use no resources and starting the userspace `auditd` daemon will not cause the kernel audit system to activate.
- `nmi_watchdog=0` disables `nmi_watchdog`.
- `acpi_irq_nobalance` disables ACPI IRQ balance.
- `lapic` enables the local APIC even if BIOS disabled it.
- `tsc=reliable` sets the timestamp counter to reliable mode.
- `skew_tick=1` ensures that the ticks per CPU do not occur simultaneously by making their start times "skewed."
- `nopti` disables the Kernel Page Table Isolation (KPTI) feature.

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

2.2.1.6 - Set Power Governor

To set the power governor to performance mode:

1. Install `CPUFreq` on server (If not available):
   `apt-get install cpufrequtils -y` (For Ubuntu)
2. Check the current governor mode:
   `cat /sys/devices/system/cpu/cpu<CoreNum>/cpufreq/scaling_governor`
3. Set the CPU cores to Performance mode:
   `echo "performance" | sudo tee /sys/devices/system/cpu/cpu<CoreNum>/cpufreq/scaling_governor`

2.2.2 - Linux Configuration

2.2.2.1 - 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 L#15 (32MB)
      L2 L#120 (1024KB) + L1d L#120 (48KB) + L1i L#120 (32KB) + Core L#120
        PU L#240 (P#120)
        PU L#241 (P#376)
      L2 L#121 (1024KB) + L1d L#121 (48KB) + L1i L#121 (32KB) + Core L#121
        PU L#242 (P#121)
        PU L#243 (P#377)
      L2 L#122 (1024KB) + L1d L#122 (48KB) + L1i L#122 (32KB) + Core L#122
        PU L#244 (P#122)
        PU L#245 (P#378)
      L2 L#123 (1024KB) + L1d L#123 (48KB) + L1i L#123 (32KB) + Core L#123
        PU L#246 (P#123)
        PU L#247 (P#379)
      L2 L#124 (1024KB) + L1d L#124 (48KB) + L1i L#124 (32KB) + Core L#124
        PU L#248 (P#124)
        PU L#249 (P#380)
      L2 L#125 (1024KB) + L1d L#125 (48KB) + L1i L#125 (32KB) + Core L#125
        PU L#250 (P#125)
        PU L#251 (P#381)
      L2 L#126 (1024KB) + L1d L#126 (48KB) + L1i L#126 (32KB) + Core L#126
        PU L#252 (P#126)
        PU L#253 (P#382)
      L2 L#127 (1024KB) + L1d L#127 (48KB) + L1i L#127 (32KB) + Core L#127
        PU L#254 (P#127)
        PU L#255 (P#383)
    HostBridge
      PCIBridge
        PCI 71:00.0 (Ethernet)
          Net "enp113s0f0np0"
          OpenFabrics "mlx5_0"
        PCI 71:00.1 (Ethernet)
          Net "enp113s0f1np1"
          OpenFabrics "mlx5_1"
```

- For PCIe devices, you can also probe `sysfs` to determine the NUMA 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:    12
Node 0 HugePages_Free:     10
Node 0 HugePages_Surp:      0
Node 1 HugePages_total:    12
Node 1 HugePages_Free:     12
Node 1 HugePages_Surp:      0
```

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

2.2.2.2 - 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`
- `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`
- `sysctl -w vm.zone_reclaim_mode=1`
- `sudo sysctl kernel.timer_migration=0`
- `echo 0 > /proc/sys/kernel/timer_migration`

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`

2.2.2.3 - 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 disable irqbalance`

Chapter 3: DPDK

This chapter explains how to configure DPDK on servers powered by AMD EPYC 9005 Series Processors.

3.1 - 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.

3.2 - Compilation

- **Library mode:** static (recommended for best performance)
- **For a native build: using a 96-core OPN**
  - **Single Socket:** `CC=gcc meson setup --default-library=static amd_zen5_linuxapp_gcc -Dmax_lcores=192 -Dc_args="-march=znver5 -Ofast"; ninja -C amd_zen5_linuxapp_gcc install; ldconfig`
  - **Dual Socket:** `CC=gcc meson setup --default-library=static amd_zen5_linuxapp_gcc -Dmax_lcores=384 -Dc_args="-march=znver5 -Ofast"; ninja -C amd_zen5_linuxapp_gcc install; ldconfig`

Note: AMD EPYC 9005 Series Processors require GCC versions 14.1 and 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 cores.*

- **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`

3.3 - Environmental Abstraction Layer (EAL) Options

The following table lists a few EAL parameters. Please see EAL Parameters* for more information on parameters used by the DPDK application..

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`	Core ID used as main.
Socket NUMA Huge page memory allocation	`--socket-mem`	Memory to allocate on sockets (comma-separated values). Allows fine grain control of huge page allocation for a given NUMA node.
Using AVX512	`--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 in both the libraries and PMDs

Table 3-1: EAL parameter options

3.4 - 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).

3.4.1 - Broadcom P2100G

None; please see BNXT PMD* for more details.

3.4.2 - 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

Please see ICE Poll Mode Driver* for additional information.

3.4.3 - NVIDIA Cx-7

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 64B packet size using DPDK `testpmd` in IO forward mode with MLX PMD in vector mode using CX7 2x200G:

- 100 MPPS can be achieved with 1 core and its sibling threads using 2 Rx and 2 Tx queues.
- 170 MPPS can be achieved with 2 core and its sibling threads using 4 Rx and 4 Tx queues.
- 190 MPPS can be achieved with 3 core and its sibling threads using 6 Rx and 6 Tx queues.

THIS PAGE INTENTIONALLY LEFT BLANK.

Chapter 4: Resources

- DPDK Debug & Troubleshoot Guide*
- Getting Started Guide for Linux*
- DPDK EAL Parameters*
- Memory Population Guidelines for AMD Family 1Ah Models 00h–0Fh and Models 10h–1Fh Socket SP5 Processors - Login required; please review the latest version if multiple versions are present.
- Socket SP5/SP6 Platform NUMA Topology for AMD Family 1Ah Models 00h–0Fh and 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™ 9005 Series Processors*
  - *Linux® Network Tuning Guide for AMD EPYC™ 9005 Series Processor Based Servers*
  - *Windows® Network Tuning Guide for AMD EPYC™ 9005 Series Processor Based Servers*
  - *VMware® Network Tuning Guide for AMD EPYC™ 9005 Series Processor Based Servers*
  - Ubuntu® Tuning Guide f*or AMD EPYC™ 9005 Series Processor Based Servers*
- NIC vendor specific tuning guide for AMD EPYC platform
  - NVIDIA (Mellanox)*
  - Broadcom*
  - RedHat Network Guide*

THIS PAGE INTENTIONALLY LEFT BLANK.

Chapter 5: 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

Data Plane Development Kit (DPDK) Tuning Guide for AMD EPYC™ 9005 Processors

PID: 58468

Manish Kumar and Vaibhav Pantakar