AMD EPYC

TUNING GUIDE
AMD EPYC 8004

Microsoft® Windows® Network

Publication	58316
Revision	1.0
Issue Date	September, 2023

Date	Version	Changes
Jul, 2023	0.1	Initial NDA release
Sep, 2023	1.0	Initial public version

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 created this document with support and input from Chris Karamatas, David Tanaka, and Tai Tse.

Table of Contents

This page intentionally left blank.

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™ 8004 Series Processors running Microsoft® Windows® Server, 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 8004 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. "Recommendations" on page 11 provides tables of recommended tuning parameters used in AMD labs. Review the block diagram of the AMD EPYC™ 8004 processor NUMA architecture in the following sections of the *AMD EPYC™ 8004 Series Architecture Overview* (available here) before you begin tuning:

- *Memory and I/O*
- *NUMA Topology*

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™ 8004 Series Processor-based servers, there are six DDR5 channels in the single CPU socket. Populate all six 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 Resources

2.1 Essential Reading

From AMD EPYC Tuning Guides:

- *BIOS and Workload Tuning Guide for AMD EPYC™ 8004 Series Processors*
- *AMD EPYC™ 8004 Series Architecture Overview*

This page intentionally left blank.

Chapter 3 TCP Performance Tuning

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

3.1 Test Configuration

The testing performed when creating this Tuning Guide used two reference systems powered by AMD EPYC 8004 Series Processors and equipped with very high speed network ports that are faster than some switches available in some labs. AMD engineers connected these two systems directly to each other so as to directly pass data between them, as shown in Figure 3-1.

Figure 3-1: Direct connections between the AMD reference systems used when creating this Tuning Guide

3.2 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. This is all much easier with AMD EPYC 8004 processors because they only support single-socket systems. Be sure to pay attention to the NUMA node and make sure your scripts are NUMA aware.

3.3 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.

3.3.1 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 six memory channels on a socket.

AMD standard BIOS defaults to NPS=1. Disabling **LLC as NUMA** 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 while still using 6-way memory channel interleaving. Setting NPS=1 and enabling **LLC as NUMA** is usually the best combination for NIC tuning. If you are concerned about latency, then set NPS=2 to interleave 3 memory channels.

Advanced > AMD CBS > DF Common Options > Memory Addressing > NUMA nodes per socket > NPS1

3.3.2 Last Level Cache (LLC) as NUMA Domain

AMD EPYC processors uses 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™ 8004 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

3.3.3 SMT

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

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

3.3.4 X2APIC

AMD EPYC 8004 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

3.3.5 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

3.3.6 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.

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

3.3.7 Memory Clock Speed

Setting the memory clock speed to **Auto** should select maximum memory speed using default BIOS settings. However, you are welcome to manually change this setting. Be sure to check the memory speed from the operating system before starting performance tests. In Windows, you can do this from the **Performance** tab of the **Task Manager**.

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

3.3.8 Slot Bifurcation

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

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

Slots 1-3 are represented in the field by position, as follows: 3x12

3.4 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, and AMD sees this paragraph as one of the most instrumental in this Tuning Guide.

3.4.1 Local NUMA Node Usage

Ensure maximum performance by using cores and memory that are in the same NUMA node as your network adapter. Opening **Device Manager**, selecting **Network Adapters**, and then selecting the **Details** tab of the NIC **Properties** window displays the **Numa Node** field in the device description menu. You can specify the NUMA node value to use as part of the command line later when you run `NTTTCP` to ensure that the NIC(s) is/are using local cores to execute the command. You can also check your work while the program runs via the **Performance** tab of the **Task Manager** with the **Numa Node** graphical view.

3.4.2 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.

3.4.3 Jumbo Packet Size

Maximum Transmission Units (MTUs) define the size of the data packet being transferred over the network fabric. The IEEE 802.3 standard sets a 1500-byte limit for Ethernet traffic. Jumbo frames (or jumbo packets) allow an MTU size of up to 9000 bytes. Your NIC vendor may support using larger than standard payloads. Increasing the packet size (MTU size) allows you to improve data throughput. However, be aware that the network switches and other parts of your network infrastructure may not support using payloads larger than 1500 bytes. You can adjust your NIC MTU by opening **Device Manager**, selecting **Network Adapters**, then selecting the **Advanced** tab of the **Properties** window for the desired NIC, and then specifying the desired **Jumbo Packet** size.

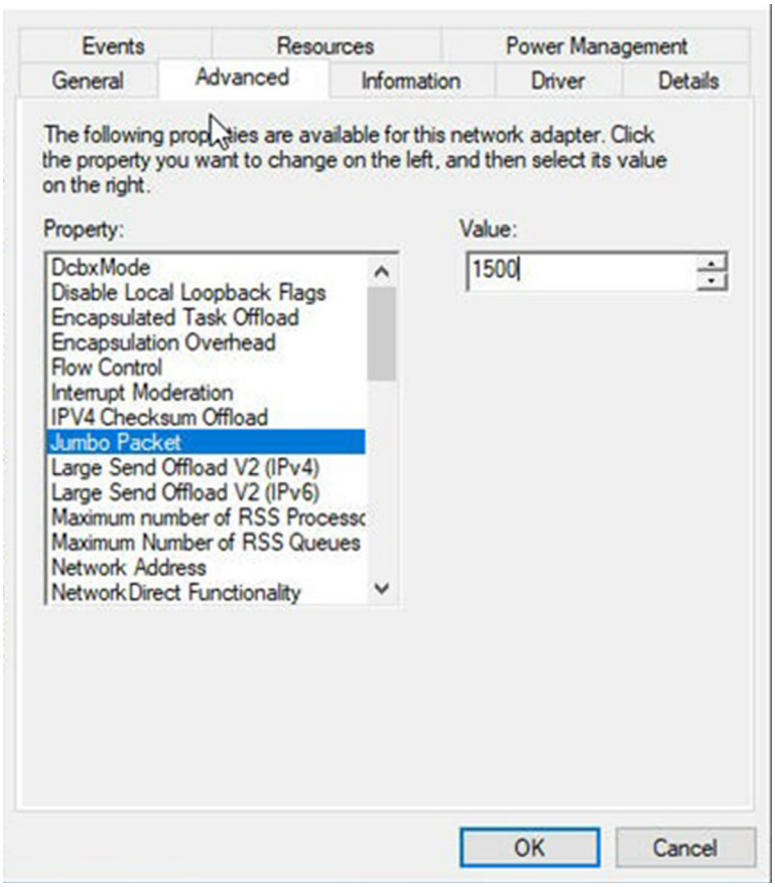

Figure 3-2: Setting the NIC Jumbo Packet value

3.4.4 Interrupt Moderation

Some NICs use Interrupt Moderation to generate an interrupt for multiple packets being transferred. This can be used for both transmitting and receiving. The benefit is lower CPU overhead and higher throughput because the device driver will generally be more efficient when operating on a larger number of packets versus a single packet. The downside is longer latency. Disable interrupt moderation for both the transmit and receive side (some vendors combine the settings while others expose them separately) for low-latency environments. You can adjust the RX and TX moderation settings via Device Manager.

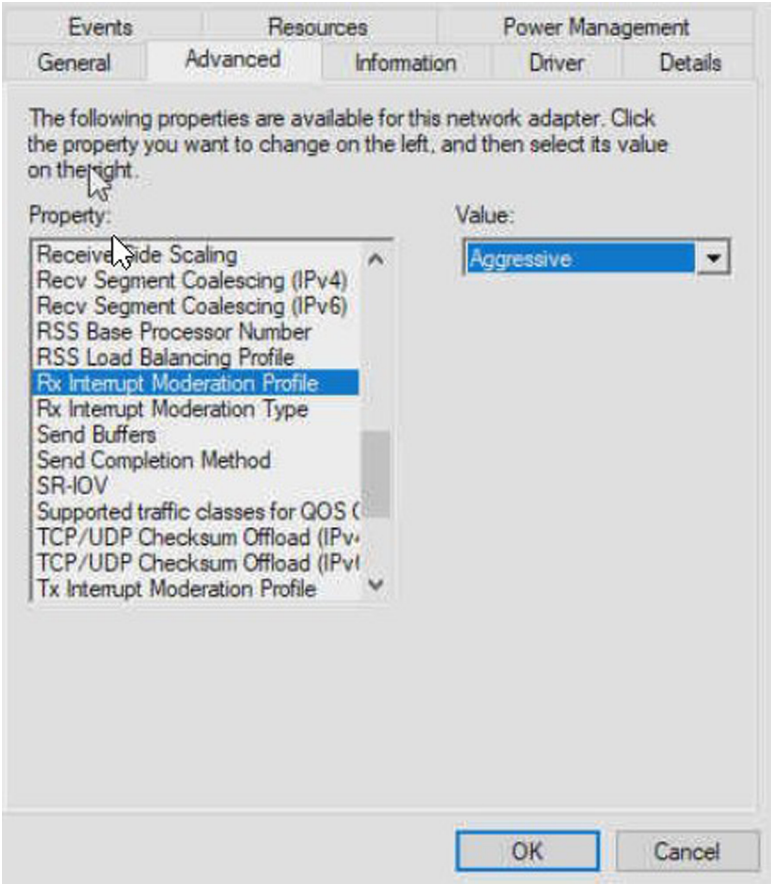

Figure 3-3: Specifying the NIC interrupt moderation

Chapter 4 Additional Information

4.1 Recommendations

Table 4-1 provides recommended values for each of the options described in this Tuning Guide. Not all adapters require modifying the default BIOS, OS, or adapter settings.

	Single Port 100/200 Gb Ethernet	Dual Port 100/200 Gb Ethernet	100/200/400 Gb InfiniBand®
BIOS Options			
Local APIC Mode	x2APIC	x2APIC	x2APIC
LLC as NUMA	Enabled	Enabled	Enabled
Determinism mode	Performance	Performance	Performance
10-bit tag	Enabled	Enabled	Enabled
Adapter Options			
Interrupt Moderation	Aggressive	Aggressive	Aggressive
Jumbo Packet	1514	9216	Default
Relaxed Ordering	Enabled	Enabled	Enabled
Receive Completion Method	Polling	Polling	Default
OS Options			
Ring Buffers	maximum	maximum	maximum
Large Receive Offload (lro)	enabled	enabled	enabled
Transmit Queue Length (txqueuelen)	20,000	20,000	20,000
Interrupts	combined 16	combined 16	combined 16

Table 4-1: NIC configuration recommendations

Table 4-2 lists several adapters that AMD has tested using Microsoft Windows Server 2022. 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 "Sunstone" reference design with BIOS versions RSS1001D and RSS1009C using NTTTCP.

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 200 Gbps Network Interface Card
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	ConnectX-7 VPI Single Port InfiniBand & Ethernet Adapter*
Testing completed with Windows Server 2022. *Ethernet performance is still being evaluated in cooperation with NVIDIA.		

Table 4-2: Tested network adapters

Chapter 5

Processor Identification

Figure 5-1 shows the processor naming convention for AMD EPYC 8004 Series Processors and how to use this convention to identify particular processors models:

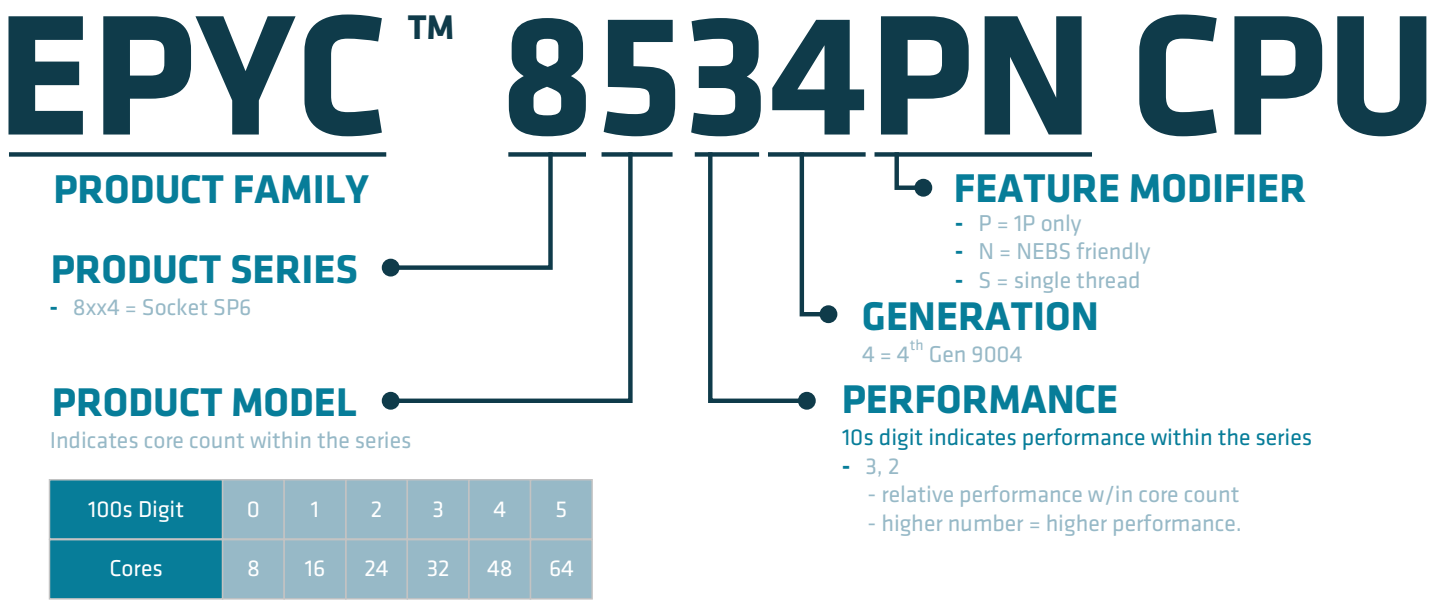

Figure 5-1: AMD EPYC SoC naming convention

5.1 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 Model 10h corresponds to an "A" part "Zen 4" CPU.
  - **8xx4:** Models A0h–AFh
- **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.

5.2 New Software-Visible Features

AMD EPYC 8004 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-bit 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*.

5.2.1 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 8004 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 8004 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.