VMware vSphere Tuning Guide for AMD EPYC™ 7002 Series Processors

Publication #	**56779**	Revision:	**1.0**
Issue Date:		**November 2019**	
Author:		**Rajesh (Raj) Bhat**	

Advanced Micro Devices

Contents

Revision History

Date	Revision	Description
November, 2019	1.0	Initial public release.

Chapter 1 Introduction

This tuning guide provides detailed descriptions of parameters that can optimize performance on servers with AMD EPYC™ 7002 Series processors in them. The default configurations on hardware and BIOS from different OEM vendors may not provide the best possible performance on all OS platforms and for all workloads. To enable optimization on a per platform and workload level, this guide calls out

- BIOS settings that can impact performance
- Hardware configuration best practices
- Supported versions of operating systems and optimization hooks on them
- Workload specific settings in BIOS and operating systems for a variety of workloads

1.1 Prerequisites

This document is intended for a technical audience with a background of configuring servers.

- Administrative access to the Server's Management Interface (BMC) as well as the operating system is required.
- Familiarity with OEMs Server's Management Interface (BMC) is strongly recommended.
- Familiarity with the OS specific tools for configuration, monitoring and troubleshooting is strongly recommended.

1.2 History

The AMD EPYC™ 7002 Series Processors are built with leading-edge 7nm technology, Zen 2 core and microarchitecture. The AMD EPYC™ SoC offers a consistent set of features across 8 to 64 cores, including 128 lanes of PCIe® Gen 4, 8 memory channels and access to up to 4 TB of high-speed memory. AMD EPYC™ 7002 Series processors are built with the following specifications:

AMD EPYC™ 7002 Series	
Process technology	7nm
Max Processor speed	3.4GHz
Max number of cores	64
Max memory speed	3200MHz
Max memory capacity	4TB
Peripheral Component Interconnect	128 lanes (max) PCIeGen4

Chapter 2 Microarchitecture Overview

2.1 Microarchitecture

Processor cores, memory controllers, I/O controllers, and security are incorporated into a Multi-Chip Module (MCM) of the AMD EPYC™ 7002 Series Processors.

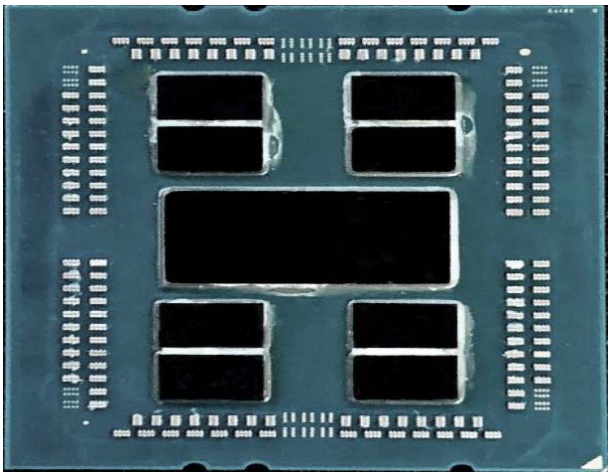

Figure 1 EPYC 7002 Configuration with 8 Core Complex Dies (CCDs) and central I/O Die (IOD)

2.2 Zen 2 core

The EPYC 7002 Series processor is based the new Zen2 processor core, that includes an L1 write-back cache. Each core can support Simultaneous Multi-threading (SMT), allowing 2 execution threads to execute simultaneously per core. Each core includes a private 512KB L2 cache.

2.3 Core Complex Die (CCD) and Core-Complex (CCX)

Up to four Zen2 cores share a 16MB (last level) L3 cache. While the two L3 Caches are on the same chiplet, they are separate. The 4 cores and their associated caches are referred to as a Core-Complex (CCX). Each Core Complex Die (CCD) contains 2 CCXs

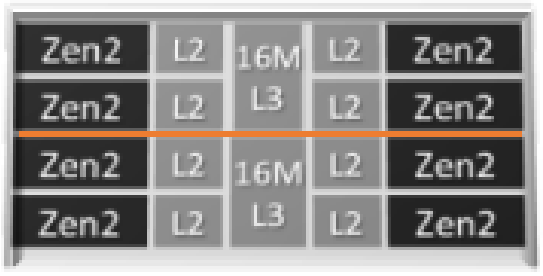

Figure 2 Two Core Complexes (CCXs) on a Core Complex Die (CCD)

Two CCDs may be abstracted as a quadrant. The CCDs connect to memory, I/O, and each other through the I/O Die (IOD). This die supports dual DDR4 memory channels.

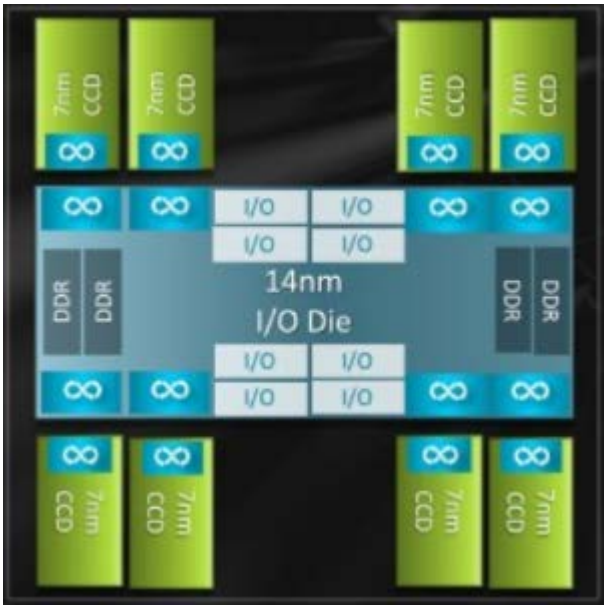

Figure 3 Single socket EPYC 7002 Processor internal connection between CCDs and Memory through memory IOD

2.4 Memory and I/O Layout

Each EPYC 7002 Series processor supports 8 memory channels. Each memory channel supports up to 2 DIMMs. Based upon BIOS settings these channels can be interleaved across a quadrant (2-way), all the way through 16-channel interleave, that is, across all memory channels of a 2-socket system. The system can have access to a maximum of 4TB of DDR4 memory at 3200MHz per processor.

The PCI subsystem provides up to 128 lanes of high speed I/O.

While all memory and I/O connect to the single I/O Die, they can be abstracted into separate quadrants each with 2 DIMM channels and 32 I/O lanes.

Two EPYC 7002 SoCs are interconnected via Socket to Socket Global Memory Interconnect (xGMI) links, part of the Infinity Fabric which connects all the components of the SoC together.

2.5 NUMA

The EPYC 7002 Series processors use a Non-Uniform Memory Access (NUMA) Micro-architecture. The four logical quadrants in an AMD EPYC 7002 Series processor (as described in *Core Complex Die (CCD) and Core-Complex (CCX)*) allow the processor to be partitioned into different NUMA domains. These domains are designated as NUMA per socket (NPS).

2.5.1 NPS1

The processor is a single NUMA domain, i.e. all the cores on the processor, all memory connected to it and all PCIe devices connected to the processor are in one NUMA domain. Memory is interleaved across the eight memory channels.

2.5.2 NPS2

The processor is partitioned into two NUMA domains. Half the cores and half the memory channels connected to the processor are grouped together into one NUMA domain. Memory is interleaved across the four memory channels in each NUMA domain.

2.5.3 NPS4

The processor is partitioned into four NUMA domains. Each logical quadrant of the processor is a NUMA domain. Memory is interleaved across the two memory channels in each quadrant. PCIe devices will be local to one of four NUMA domains on the processor depending on the quadrant of the IO die that has the PCIe root for that device.

2.5.4 L3 Cache as NUMA Domain

Each L3 Cache (as explained in *Core Complex Die (CCD) and Core-Complex (CCX)*) is exposed as a NUMA node. On a dual processor system, with up to 16 L3 Caches per processor, this setting will expose 32 NUMA domains.

Using BIOS settings, each server can be configured as NPS1, NPS2 or NPS4, with an additional option to configure L3 cache as NUMA nodes.

AMD EPYC 7002 Series processors are available in different core counts per processor and not all of them can support all NPS settings. See *https://developer.amd.com/wp-content/resources/56338_1.00_pub.pdf* for details on NUMA architecture and settings.

Chapter 3 Hardware Configuration Best Practices

3.1 Memory Configurations

For optimal performance, populate 8 DIMMs for 1 DPC (DIMMs per Channel) configuration, or 16 DIMMs for 2 DPC (DIMMs per Channel) configuration, per processor. Other configurations, such as 12 DIMMs per processor, does not provide optimal performance.

1 DPC configuration runs the memory DIMMs at a higher speed when compared to 2 DPC.

OEM servers supporting AMD EPYC 7002 Series processors are built to either support previous generations of AMD EPYC (EPYC 7001 series) or are specifically designed for AMD EPYC 7002. Contact your OEM to determine the characteristics of your servers.

3.1.1 Platforms that support previous generations of AMD EPYC

- Platforms are compatible with AMD EPYC 7001 Processors
- Contact your OEM to determine the Maximum Memory Bus Frequency supported on their platforms.
- For throughput sensitive applications, to obtain higher IO throughput, Maximum Memory Bus Frequency can be set to the maximum allowed, provided your Memory DIMM hardware supports it. However, in some cases, the Infinity Fabric Clock on these platforms may not synchronize with the maximum Memory Bus Frequency supported by the OEM. This unsynchronized behavior can lead to higher latency.
- For latency sensitive applications, better performance is obtained by setting the Maximum Memory Bus Frequency to 2667 MT/s or 2400 MT/s, since these frequencies synchronize with the Infinity Fabric Clock.

3.1.2 Platforms specifically designed for AMD EPYC 7002

- Platforms are specifically designed for AMD EPYC 7002 and are not compatible with AMD EPYC 7001 Processors
- The Maximum Memory Bus Frequency supported on these platforms is 3200 MT/s.
- For throughput sensitive applications to obtain higher IO throughput, Maximum Memory Bus Frequency can be set to the maximum allowed (3200 MT/s) provided your Memory DIMM hardware supports it. However, the Infinity Fabric Clock on these platforms does not optimally synchronize with Memory Bus Frequency of 3200 MT/s, which can lead to higher latency.
- For latency sensitive applications, better performance is obtained by setting the Maximum Memory Bus Frequency down to 2933 MT/s, 2667 MT/s or 2400 MT/s, since these frequencies synchronize with the Infinity Fabric Clock.

3.2 PCI Subsystem

For IO intensive workload, performance can be improved by placing the workload on the same socket that connects to the I/O Device used. For example, for a networking intensive operation, place the workload on the socket that the NIC connects to. Tools, such as lstopo on Linux, help determine the connectivity between PCI Devices and sockets.

If you plan to use SRIOV, then see vSphere Network Tuning Guide at *https://developer.amd.com/wp-content/resources/56763_0.90.pdf* for additional details.

Chapter 4 Performance Enhancements Using BIOS Settings

While application profile influences platform configuration, there are several areas on the platform BIOS that can be tuned to obtain better overall performance across the board. Evaluate all options explained below to examine its impact on your workload.

4.1 Prerequisites

- Administrative access to the Server's Management Interface (BMC) as well as the operating system is required.
- Familiarity with OEMs Server's Management Interface (BMC) is strongly recommended.
- Upgrade to the latest available vendor BIOS to get the most current fixes, features and performance benefits.
    - Refer to the Server Vendor's browser compatibility guidelines and need for maintenance windows if the BIOS is being upgraded using web browser.
    - Refer to the Server Vendor's guidelines on state of the machine and need for maintenance windows if the BIOS is being upgraded using the OS.
- Ensure that all the desired settings remain in place after every BIOS upgrade.

4.2 BIOS Settings for Performance

The following BIOS options provide optimal performance on a variety of virtual applications on VMware vSphere installations. They may not achieve additional objectives, such as optimal power consumption.

4.2.1 Performance and Power profiles

- ESXi can influence power profiles at an operating system level only if the BIOS power profile is set to “OS control mode” or equivalent. Using alternative power profiles in the BIOS, such as ”Max performance”, prevent ESXi from influencing power profiles within the operating system.
- OEM BIOS software sometimes have profiles that initialize several of the BIOS settings to match application requirements. To improve performance, pick the profile that favors performance on virtualized environments.
- Platforms often present a trade-off with Power savings and Performance. Turning off power saving schemes in the BIOS may result in improved performance of certain workloads at the expense of additional power consumed.
- Platforms may also provide ability to control Determinism. Power Deterministic allows you to extract the maximum performance from your processor (it disables locking your processor to a standardized performance level). Hence pick "Power Determinism" over "Performance

Determinism". Refer to *https://www.amd.com/system/files/2017-06/Power-Performance-Determinism.pdf* for more details.

- Providing additional cooling to CPU helps drive higher workloads more effectively. Set BIOS options that enhance cooling. A side effect may be increased power consumed by the server.
- Always provide adequate power to the server by plugging in all the redundant power supply units on the server. This also helps keep the server running if one of the power supply units were to fail.

4.2.2 Processor

- Ensure that x2APIC support is enabled. This allows for more than 255 CPUs to be enabled on the platform (if applicable).
- Ensure that AMD SMT is enabled. This enables multi-threading, which effectively doubles the number of logical processors available to your workloads on vSphere.

4.2.3 NUMA configuration

In general, optimal NUMA settings can only be inferred by looking at the workload characteristics. Distributed workloads requiring clustering may benefit from BIOS settings that are different from settings that benefit small IO intensive workloads. Refer to section on <> for recommended settings on different workloads. See *NUMA* for details on NUMA, different NUMA settings and their benefits.

If your workload uses few VMs with few vCPUs per VM (such that total vCPUs in workload is less than quarter of the number of cores per socket), then the following settings tend to provide improved performance

- NPS (NUMA per socket) = 4
- "L3 cache as NUMA" turned on

If your workload uses several VMs, or are I/O intensive, or if your VM has a large number of vCPUs then the following settings tend to provide improved performance

- NPS (NUMA per socket) = 1
- "L3 cache as NUMA" turned off

4.2.4 Other settings

- Ensure that AMD IOMMU is turned on. IOMMU is essential for compatibility with IO devices, system performance and for technology such as SRIOV.
- Ensure that AMD Virtualization Technology is turned on. This setting is essential for Virtual Machines running on vSphere.

Chapter 5 Software Settings and Particulars

5.1 Supported Software Versions

5.1.1 vSphere/ESXi/VSAN

- vSphere 6.5 EP15 and above
- vSphere 6.7 u3 and above

5.1.2 NSX-V

- NSX-V 6.4.2 through 6.4.6 with vSphere 6.7 U3
- NSX-V 6.4.4 through 6.4.6 with vSphere 6.5 EP15

5.2 ESXi (OS) Advanced Configuration Parameters

Use the following advanced configuration parameters for performance:

- Numa.LocalityWeightActionAffinity=0
    - For more information, see *https://kb.vmware.com/s/article/2097369*.
- Numa.PreferHT=1
    - For memory latency sensitive workloads with low processor utilization or high inter-thread communication, use hyper-threads with fewer NUMA nodes instead of full physical cores spread over multiple NUMA nodes.
    - *https://kb.vmware.com/s/article/2003582*
- Workloads with VMs less than 4 vCPUs (or 8 with Numa.PreferHT enabled) provides significant performance by utilizing a localized L3 cache.
- In addition to high performance profile in the BIOS, ESXi power profile can be influenced by setting Power.CpuPolicy to “HighPerformance”. ESXi can influence power profiles at an operating system level only if the BIOS power profile is set to “OS control mode” or equivalent. Using alternative power profiles in the BIOS, such as ”Max performance”, prevent ESXi from influencing power profiles within the operating system.

5.3 Known Issues on vSphere for AMD EPYC 7002 Series Processors

The table below lists issues, impacted vSphere release versions, and links to related information for workarounds.

vSphere Release Version	Issue Description	Link to Related Articles or KB
6.5 EP15, 6.7 U3	ESXi spontaneously reboots after enabling SATA passthrough function in VMs	*KB Article Section: Problems with Device Assignment Dependencies*
6.5 EP15, 6.7 U3	Manually triggering a non-maskable interrupt (NMI) might not work on a vCenter Server system with an AMD EPYC 7002 series processor	*https://docs.vmware.com/en/VMware-vSphere/6.7/rn/vsphere-esxi-67u3-release-notes.html*
6.5 EP15, 6.7 U3	A virtual machine that has a PCI passthrough device assigned to it might fail to power on in a vCenter Server system with an AMD EPYC 7002 series processor.	*https://docs.vmware.com/en/VMware-vSphere/6.7/rn/vsphere-esxi-67u3-release-notes.html*
6.5 EP15, 6.7 U3	LINT1/NMI when using VPMC or vmkstats features on an AMD EPYC 7002 processor.	*https://kb.vmware.com/s/article/71314*
6.5 EP15, 6.7 U3	"Failed - Module 'NumVCPUs' power on failed" error powering on a virtual machine (74899)	*https://kb.vmware.com/s/article/74899*

Chapter 6 Workload Optimization on vSphere

6.1 VMware vSAN

VMware vSAN uses HCIBench as the benchmark to measure the IOPS and latency of different IO workloads. The settings suggested below have provided optimal results on AMD EPYC Series processors using HCIBench. Please note that HCIBench is a synthetic workload. Hence settings that result in improved results in the benchmarks may not necessarily translate to similar gains on real life workloads. For more details on the performance using the settings called out below, refer to the VMware vSAN Tuning Guide for AMD EPYC™ 7002 Series Processors.

6.1.1 Optimize latency

VMware VSAN is a latency sensitive application. You can reduce latency on your VSAN cluster with the following settings

- Use DIMM speeds as recommended in *Memory Configurations* to synchronize with Infinity Fabric Clock speeds
- Choose a performance profile that optimizes performance. See *Performance and Power profiles* for more details.
- xGMI is the communication link between the two sockets in a dual socket server. Set xGMI Link Width to maximum (16). See section 2.1.2 of https://developer.amd.com/wp-content/resources/56745_0.80.pdf for more details.

6.1.2 NUMA configuration

VSAN workloads perform better with fewer NUMA nodes on the system.

- Configure one NUMA domain per socket : NPS1
- Disable "L3 Cache as NUMA" in the BIOS

See *NUMA configuration* for more details on NUMA

6.1.3 Hardware configuration

- Using NVMe drives for disk group cache improves the throughput and reduces the latency significantly
- Strive to keep all VSAN disk groups and VSAN network ports on socket 0 in a dual socket server. See *PCI Subsystem* for more details.
- Use at least 25Gb NIC to avoid any networking I/O bottlenecks
- Follow best practices quoted in https://docs.vmware.com/en/VMware-vSphere/6.7/vsan-671-planning-deployment-guide.pdf

Chapter 7 References

7.1 vSphere performance best practices

https://www.vmware.com/content/dam/digitalmarketing/vmware/en/pdf/techpaper/performance/vsphere-esxi-vcenter-server-67-performance-best-practices.pdf

https://www.vmware.com/content/dam/digitalmarketing/vmware/en/pdf/techpaper/performance/Perf_Best_Practices_vSphere65.pdf

7.2 ESXi (vSphere's OS) commonly used tools

- esxtop
  - Performance monitoring tool
  - *http://www.yellow-bricks.com/esxtop/*
- vm-support
  - Support logs gathering utility
  - *https://kb.vmware.com/s/article/653*
- esxcli
  - Command line utility to query the OS
  - *https://www.virten.net/2018/04/vmware-esxi-6-7-esxcli-command-reference/*

7.3 Optimized workload runs on vSphere with AMD EPYC™ 7002 Series Processors

- TPCxV : *http://www.tpc.org/tpcx-v/results/tpcx-v_result_detail-5302.asp*
- VMmark SAN : *https://www.vmware.com/content/dam/digitalmarketing/vmware/en/pdf/vmmark/2019-08-07-HPE-ProLiant-DL385Gen10.pdf*
- VMmark VSAN : *https://www.vmware.com/content/dam/digitalmarketing/vmware/en/pdf/vmmark/2019-08-07-HPE-ProLiant-DL325Gen10.pdf*

7.4 Other resources

HCIBench : *https://flings.vmware.com/hcibench*

7.5 Glossary

BMC - Baseboard Management Controller

CCD - Core Complex Die

CCX - Core-Complex

DIMM - Dual In-line Memory Module

DPC - DIMMs per Channel

IOD - I/O Die

MCM - Multi-Chip Module

NMI - non-maskable interrupt

NPS - NUMA Per Socket

NUMA - Non-Uniform Memory Access

PCIe - Peripheral Component Interconnect Express

SMT - Simultaneous Multi-threading

VM – Virtual Machine

xGMI - Socket to socket Global Memory Interconnect