Windows® Network Tuning Guide for AMD EPYC™ 7002 Series Processor Based Servers

Application Note

Publication #	**56746**	Revision:	**0.90**
Issue Date:	**October 2019**		

Advanced Micro Devices

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, and combinations thereof are trademarks of Advanced Micro Devices, Inc. Other product names used in this publication are for identification purposes only and may be trademarks of their respective companies.

Linux is a registered trademark of Linus Torvalds.

Contents

List of Tables

Revision History

Date	Revision	Description
October 2019	0.90	Initial release.

Chapter 1 Introduction

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, one may have to make different adjustments to network tuning to optimize for a specific workload. With Ethernet® speeds going higher, up to 200 Gb, 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 recommends steps to check and modify when (or if) they prove to be beneficial for a given scenario. In this guide, the steps are focused on TCP/IP network performance. Appendix A provides tables of recommended tuning parameters as well as results measured in AMD labs.

One general rule of thumb for all performance testing is to ensure your memory subsystem is properly configured. All I/O uses data transfers into or out of memory, so the I/O bandwidth can never exceed the capabilities of the memory subsystem. For the maximum memory bandwidth on modern CPUs, you must populate at least one DIMM in every DDR channel. For AMD EPYC 7002 Series Processor-based servers, there are eight DDR4 memory channels on each CPU socket. So, for a single-socket platform, you must populate all eight memory channels. Likewise, on a dual-socket platform, you must populate 16 memory channels.

In addition to this document, AMD recommends consulting any tuning guide available from your NIC vendor. Vendors will sometimes enable specific tuning options for their devices with parameters that can be modified to further improve performance. One example could be the ability to enable or disable interrupt moderation, described in more detail in 2.1 Interrupt Moderation.

Chapter 2 Adapter Device Driver Tuning

2.1 Interrupt Moderation

Interrupt Moderation is a method used by some NICs to send a generate interrupt for multiple packets being transferred. This can be used for both transmit and receive side. The benefit will be lower CPU overhead and higher throughput, because the device driver will generally be more efficient when operating on a larger number of packets vs a single packet. The downside is longer latency. Therefore, for low latency environments, disable interrupt moderation for both transmit and receive side (some vendors combine the settings while others expose them separately).

2.2 Jumbo Packet Size

Maximum Transmission Units, or MTUs, define the size of the data packet being transferred over the fabric of a network. The limit for Ethernet set by the IEEE 802.3 standard is 1500. Jumbo frames, or jumbo packets, allow for an MTU size of up to 9000. In the properties of your NIC, if the NIC vendor allows, you can modify the jumbo packet size (MTU size) to improve your throughput. However, if your network infrastructure does not support above 1500 MTU, then the packet will be limited in size.

2.3 Device Driver

Ensure you have the latest device driver and firmware from your NIC vendor. AMD and the ecosystem are working closely together to optimize devices for the AMD EPYC 7002 Series processor, and sometimes that can result in updates to devices’ drivers and firmware.

Chapter 3 BIOS Options

3.1 x2APIC

With the introduction of the EPYC 7002 Series of processors, AMD has implemented an x2APIC controller. This has two benefits:

- Allows operating systems to work with the 256 CPU threads now available on AMD platforms
- Provides improved performance over the legacy APIC

AMD recommends, but not requires, that you enable the x2APIC mode in BIOS even for lower core count parts. (The AMD BIOS will enable x2APIC automatically when two 64-core processors are installed.) Microsoft Windows 2019 requires installation media released in early October 2019, or a later version. Earlier versions of Microsoft Windows do not support the AMD x2APIC implementation, and therefore require fewer than 256 total threads by disabling SMT if using dual 64-core processors.

3.2 Infinity Fabric P-States

Like processor cores, the EPYC 7002 Series processor Infinity Fabric has the ability to go into lower power states (P-states) when being lightly used. This saves power consumption of the overall socket, or allows power to be diverted to other portions of the processor. By default, to enable the best performance per watt, P-states are enabled in the processor. To disable the switching of P-states and force P0 all the time, you must go into the system BIOS and set APBDIS to 1. Higher bandwidth adapters may require the forcing of the P0 state to maintain the highest bandwidth.

3.3 Preferred I/O and PCIe® Relaxed Ordering

From PCIe 1.0, there have been both strict and relaxed methods of ordering PCIe packets within a system. The ordering of the packets helps maintain data coherency and consistency as data flows between endpoints and main memory. Relaxed ordering is enabled to allow packets to be retired out of order when possible. This maintains data consistency and improves performance in high-bandwidth cases. AMD introduced Preferred I/O as a new feature in the EPYC 7002 Series processors also to help with ordering of PCIe packets. Enabling this for high-bandwidth adapters can result in improved performance in some cases.

3.4 Last Level Cache (LLC) as NUMA

The EPYC line of processors has multiple Last Level Caches (LLCs), or L3 caches. While operating systems can handle the multiple LLCs and schedule jobs accordingly, AMD has created a BIOS option to enable the description of 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.

Appendix A Networking Tuning Recommendations and Results

Table 1 Network Tuning Recommendations provides the recommended values for each of the options described in the document. Not all adapters require modification from default BIOS or adapter property options.

Table 1 Network Tuning Recommendations

	Single Port **100 Gb Ethernet**	**Dual Port** **100 Gb Ethernet**	**Dual Port** **100 Gb InfiniBand®**
BIOS Options			
Local APIC Mode	x2APIC	x2APIC	x2APIC
APBDIS	1	1	1
Preferred I/O	Enabled	Enabled	Enabled
LLC as NUMA	Enabled	Enabled	Enabled
Adapter Properties Options			
Interrupt Moderation	Aggressive	Aggressive	Aggressive
Jumbo Packet	1514	9216	Default
Receive Completion Method	Polling	Polling	Default

AMD has tested several adapters at multiple speeds using the recommendations from Table 1 Network Tuning Recommendations, and those results are below in Table 2 Network Testing Results.

Table 2 Network Testing Results

Tested Adapter	Fabric Type	Port Speed	Total Ports	Bi-directional Bandwidth
Mellanox ConnectX-5	Ethernet	100 Gb	1	186.8 Gbps
Broadcom Stratus	Ethernet	100 Gb	1	181.8 Gbps
Mellanox ConnectX-5 (1)	Ethernet	100 Gb	2	295.0 Gbps
Mellanox ConnectX-5	InfiniBand	100 Gb	2	376.4 Gbps

1. Run the following command:
   - ⇒ Mlxconfig.exe -d mt4121_pciconf0 set PCI_WR_ORDERING=1