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

Application Note

Publication #	**56763** Revision: **0.90**
Issue Date:	**November 2019**

Advanced Micro Devices

Contents

List of Tables

Revision History

Date	Revision	Description
November 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 and hypervisors 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. Please consult *Memory Population Guidelines for AMD EPYC 7002 Series Processors* (PID# 56502) for more details.

In addition to this document, AMD recommends consulting any tuning guide available from your NIC vendor and operating system or hypervisor vendor. Vendors will sometimes enable specific tuning options for their devices with parameters that can be modified to further improve performance.

Chapter 2 Adapter Device Driver Tuning

2.1 Device Driver

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

2.2 Single Root I/O Virtualization (SR-IOV)

The PCI Sig® introduced Single Root I/O Virtualization (SR-IOV) in 2007 as an extension to the PCIe® Revision 2.0 base specification. SR-IOV brought the ability to have a single PCIe endpoint (device), or Physical Function (PF), to contain multiple Virtual Functions (VF) within it. One or more VFs can then be assigned to a virtual machine by the hypervisor, providing direct access to specific hardware portions of the endpoint obviating the need for the hypervisor to use full virtualization or even para-virtualization techniques for performing typical PCIe transactions, such as having a NIC send or receive data. This can result in a noticeable performance improvement by removing another layer of software from the overall transaction. While not all PCIe endpoints support SR-IOV, there are more vendors adding the support on a regular basis. You should enable SR-IOV in the device driver of the NIC and assign VFs to individual Virtual Machines.

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.)

3.2 Single Root I/O Virtualization (SR-IOV)

Please see 2.4 Single Root I/O Virtualization (SR-IOV) for details on SR-IOV. For a PCIe endpoint (adapter) to use SR-IOV, it must be enabled in BIOS in addition to the device driver. In EPYC platforms, there is a BIOS option for both enabling SR-IOV and another feature called PCIe Alternative Routing-ID Interpretation (ARI). ARI must be enabled in conjunction with SR-IOV, and some platform vendors have combined the two into a single BIOS option.

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

	Dual Port 25 Gb Ethernet
BIOS Options	
Local APIC mode	x2APIC
SR-IOV	Enabled
PCIe ARI Support	Enabled
Adapter Options within OS	
SR-IOV	Enabled

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	Bidirectional Bandwidth
Mellanox® ConnectX-4®	Ethernet	25 Gb	2	94 Gbps

Appendix B Example of Testing Dual Port 25 Gb NIC

B.1 Equipment Needed

To measure the network performance properly, AMD used two different EPYC 7002 based servers: one as the System Under Test (SUT) and one as the client system or target system. In both cases, we used the following configuration:

CPU	AMD EPYC 7742 Processor
Memory	512 GB of DDR4 DRAM 3200 MT/s
Guest OS	Ubuntu 19.04
Host OS	VMware ESXi 6.7.0 build-14212230
Network Cards	Mellanox ConnectX4 FW 14.23.1020 connected to Socket 0, nmlx5_core driver Version: 4.17.15.16-1OEM.670.0.0.8169922
Network Modes	Ethernet

AMD set up these two systems back-to-back. What this means is that we connected Port 1 of the Mellanox adapter in the SUT to Port 1 of the Mellanox adapter in the Client. Likewise, the two Port 2s were connected. No switch was used in between the SUT and Client.

B.2 Required Tools

The Mellanox Firmware Tools (MFT), specifically ***mst*** for ESXi is required to update firmware and set SR-IOV configuration. It can be found by visiting:
https://www.mellanox.com/page/management_tools

B.3 NIC Settings and Driver Configuration

1. Upgrade to the latest firmware available for your NIC by visiting:
   https://www.mellanox.com/page/firmware_table_ConnectX4EN

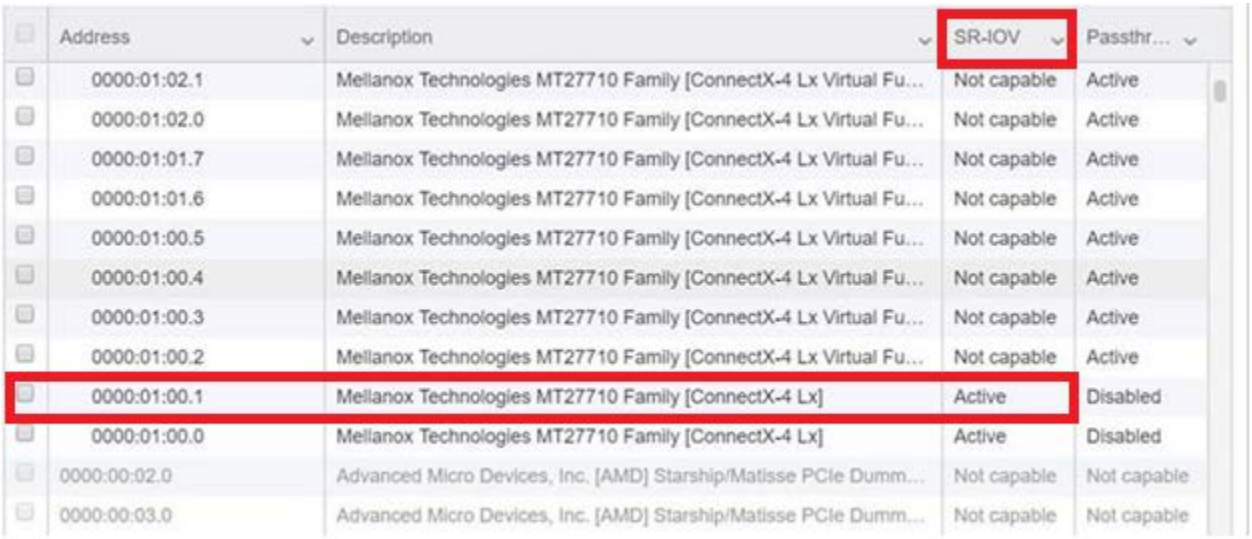
2. Upgrade to the latest NIC driver for ESXi by visiting:
   https://www.mellanox.com/page/products_dyn?product_family=29&ssn=h2e10vtnj8ckehai14o082ihg3

3. Enable SR-IOV in the adapter and set the number of VFs
   `/opt/mellanox/bin/mlxconfig -d <device_name> set SRIOV_EN=1 NUM_OF_VFS=12`
   `<device_name>` can be found by running `/opt/Mellanox/bin/mst status -v`

4. Set the number of VFs on the NIC driver

```
esxcli system module parameters set -m nmlx5_core -p max_vfs=4,4
```

5. Using the VMware vSphere Web Client, verify that SR-IOV is “Active” on the Physical Function (PF) and Passthrough is “Active” on the VFs that you created above for both the SUT and Client systems.



Address	Description	SR-IOV	Passthr...
0000:01:02.1	Mellanox Technologies MT27710 Family [ConnectX-4 Lx Virtual Fu...	Not capable	Active
0000:01:02.0	Mellanox Technologies MT27710 Family [ConnectX-4 Lx Virtual Fu...	Not capable	Active
0000:01:01.7	Mellanox Technologies MT27710 Family [ConnectX-4 Lx Virtual Fu...	Not capable	Active
0000:01:01.6	Mellanox Technologies MT27710 Family [ConnectX-4 Lx Virtual Fu...	Not capable	Active
0000:01:00.5	Mellanox Technologies MT27710 Family [ConnectX-4 Lx Virtual Fu...	Not capable	Active
0000:01:00.4	Mellanox Technologies MT27710 Family [ConnectX-4 Lx Virtual Fu...	Not capable	Active
0000:01:00.3	Mellanox Technologies MT27710 Family [ConnectX-4 Lx Virtual Fu...	Not capable	Active
0000:01:00.2	Mellanox Technologies MT27710 Family [ConnectX-4 Lx Virtual Fu...	Not capable	Active
0000:01:00.1	Mellanox Technologies MT27710 Family [ConnectX-4 Lx]	Active	Disabled
0000:01:00.0	Mellanox Technologies MT27710 Family [ConnectX-4 Lx]	Active	Disabled
0000:00:02.0	Advanced Micro Devices, Inc. [AMD] Starship/Matisse PCIe Dumm...	Not capable	Not capable
0000:00:03.0	Advanced Micro Devices, Inc. [AMD] Starship/Matisse PCIe Dumm...	Not capable	Not capable

Figure 1. Confirming SR-IOV Is Enabled

B.4 Setting Up the Virtual Machines

Two identical Virtual Machines (VMs) need to be created on both the SUT and Client ESXi systems. Each of the VMs should be configured with four vCPUs and 64 GB vRAM. If assistance is required in setting up the VMs, consult the following VMware document: *https://docs.vmware.com/en/VMware-vSphere/6.7/com.vmware.vsphere.vm_admin.doc/GUID-AE8AFBF1-75D1-4172-988C-378C35C9FAF2.html*

1. To get accurate measurements, we want to ensure the VM’s vCPUs are all within the same NUMA node. For that, we will pin, or affinitize, the vCPUs to specific logical CPU threads. VM affinity can be changed in VM properties, and we set the following values:
   - VM 1 is affinitized to Logical threads 0, 2, 4, 6
   - VM 2 is affinitized to Logical threads 8, 10, 12, 14
2. Install Ubuntu 19.04 as the Guest OS and install iPerf 2.09. iPerf will be used to measure the network bandwidth.

3. Next, we will setup the networking for the VM. Figure 2. Network Topology of the System is a pictorial representation of how we will setup the network adapter.

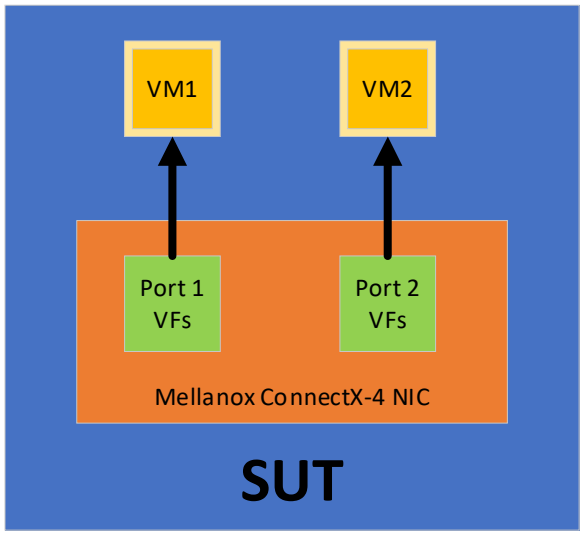


Figure 2. Network Topology of the System

   a. Using the VM Edit Configuration Wizard within the vSphere Web Client, attach a VF from the Port 1 VF pool of the NIC to VM 1 (Edit Setting > Add PCIe Device)

      - For assistance in adding a PCIe device to a VM, consult the following VMware doc: *https://docs.vmware.com/en/VMware-vSphere/6.7/com.vmware.vsphere.vm_admin.doc/GUID-5B3CAB26-5D06-4A99-92A0-3A04C69CE64B.html*

   b. Next, attach any VF from the Port 2 VF pool of the NIC to VM2 (Edit Setting > Add PCIe Device)

   c. Finally, configure IP settings to the newly added interfaces inside the VM. They will each present themselves as separate network adapters. The subnets of ports connected between the SUT and Client must be the same for them to be able to communicate. You can use a simple ping test between the VMs on both SUT and Client to confirm communication is established.

B.5 Running iPerf

iPerf has a server task and a client task. The server must be setup prior to starting the client, and then the client generates the load. Both tasks will present throughput results but when AMD looks at the results, we look at the throughput seen by the server.

1. Start an iPerf server instance on all VMs. So, that would be a total of four iPerf server instances combined between the SUT and Client

2. Prepare an iPerf client instance on each VM. For the iPerf client instance, specify two threads per instance. The target server will be the direct connected VM on the opposite machine. Make the time be a minimum of 60 seconds to get a good average throughput.

3. Launch each iPerf client as close to simultaneously as possible.

4. After this completes, add the throughput for all four server instances to find your total aggregate throughput measured.