# Mellanox NIC's Performance Report with DPDK 19.08 for AMD 2nd Gen EPYC Processor

**Rev 1.0**

# Table of Contents

# List of Figures

# List of Tables

# Document Revision History

***Table 1: Document Revision History***

| Revision | Date | Description |
|---|---|---|
| 1.0 | 2-Dec-2019 | Initial report release for AMD Processor |

# About this Report

The purpose of this report is to provide packet rate performance data for Mellanox ConnectX-5 and ConnectX-6 Network Interface Cards (NICs) achieved with the specified Data Plane Development Kit (DPDK) release. The report provides both the measured packet rate performance and the procedures and configurations to replicate the results. This document does not cover all network speeds available with the ConnectX family of NICs and is intended as a general reference of achievable performance for the specified DPDK release.

## Target Audience

This document is intended for engineers implementing applications with DPDK to guide and help achieving optimal performance.

# 1 Test Description

## 1.1 General

Setup is made up of the following components:

1. Server - AMD “Daytona X” Rome Server reference platform with 2nd Gen EPYC processor
2. Mellanox ConnectX® NIC
3. IXIA® XM12 packet generator

## 1.2 Zero Packet Loss Test

Zero Packet Loss tests utilize **l3fwd** (http://www.dpdk.org/doc/guides/sample_app_ug/l3_forward.html) as the test application for maximum throughput with zero packet loss at various frame sizes based on RFC2544 https://tools.ietf.org/html/rfc2544.

The packet generator transmits a specified frame rate towards the DUT and counts the received frame rate sent back from the DUT. Throughput is determined with the maximum achievable transmit frame rate and is equal to the received frame rate i.e. zero packet loss.

- Duration for each test is 60 seconds.
- Traffic of 8192 IP flows is generated per port.
- IxNetwork (Version 8.51EA) is used with the IXIA packet generator.

## 1.3 Single Core Performance Test

Single Core performance tests utilize **testpmd** (http://www.dpdk.org/doc/guides/testpmd_app_ug), with this test the max throughput is tested with a single CPU core. Average throughput within test duration (60 seconds) are the results recorded in this test.

- Duration for each test is 60 seconds.
- Traffic of 8192 UDP flows is generated per port.
- IxNetwork (Version 8.51EA) is used with the IXIA packet generator.

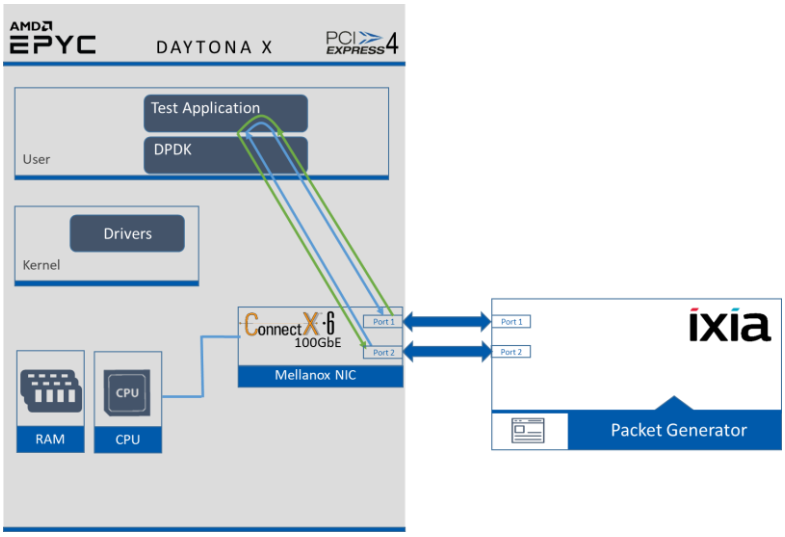
# 2 Test #1
# Mellanox ConnectX-6 2x100GbE Throughput at Zero Packet Loss (2x 100GbE)

***Table 2: Test #1 Setup***

| Item | Description |
|---|---|
| Test | Test #1 – Mellanox ConnectX-6 2x100GbE Throughput at zero packet loss |
| Server | AMD "Daytona X" Rome Server Reference Platform |
| CPU | 2*AMD EPYC 7742 @ 2.25GHz (running @ 3.38GHz)<br>64 CPU cores ; PCIe Gen 4.0 |
| RAM | 256GB: 16 * 16GB DIMMs @ 2666MHz |
| BIOS | American Megatrends Inc. RDY1003A 08/23/2019 |
| NIC | One ConnectX-6 VPI adapter card; 200GbE; dual-port QSFP56; PCIe4.0 x16; |
| Operating System | Red Hat Enterprise Linux Server 7.6 |
| Kernel Version | Linux 3.10.0-957.el7.x86_64 |
| GCC version | 4.8.5 20150623 (Red Hat 4.8.5-36) (GCC) |
| Mellanox NIC firmware version | 20.26.1040 |
| Mellanox OFED driver version | MLNX_OFED_LINUX-4.7-1.0.0.1 |
| DPDK version | 19.08 |
| Test Configuration | 1 NIC, 2 ports used on NIC; Port has 8 queues assigned to it, 1 queue per logical core for a total of 16 logical cores for both ports.<br>Each port receives a stream of 8192 IP flows from the IXIA |

Device Under Test (DUT) is made up of the AMD "Daytona X" Rome Server reference platform and the Mellanox ConnectX-6 NIC with dual-ports. The DUT is connected to the IXIA packet generator which generates traffic towards the ConnectX-6 NIC.
The ConnectX-6 data traffic is passed via PCIe Gen 4 bus through DPDK to the test application **l3fwd** and is redirected to the opposite port. IXIA measures throughput and packet loss.

***Figure 1: Test #1 Setup – Mellanox ConnectX-6 2x100GbE connected to IXIA***



## 2.1 Test Settings

***Table 3: Test #1 Settings***

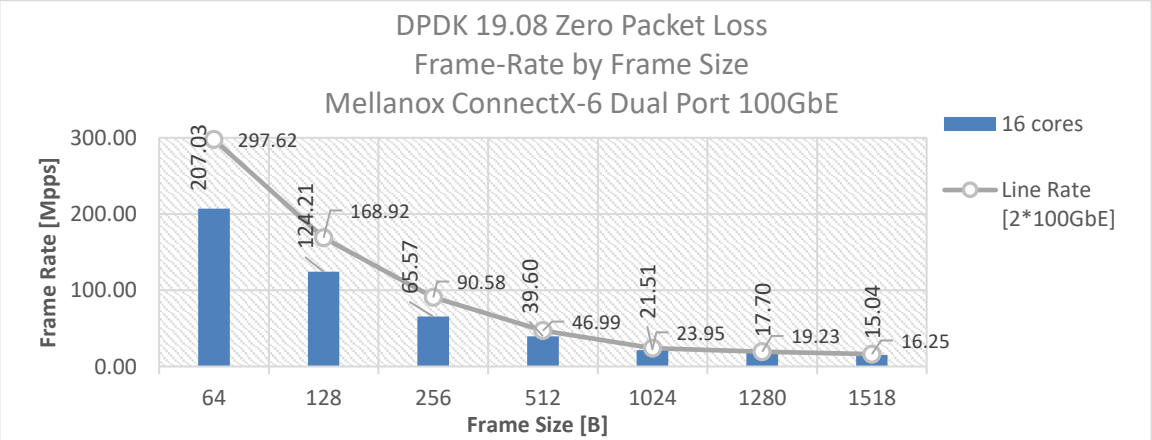
| Item | Description |
|---|---|
| BIOS | # Enable X2APIC - Advanced -> AMD CBS -> CPU Common -> Local APIC Mode -> X2APIC<br># Enable NPS2 - Advanced -> AMD CBS -> DF Common -> Memory Addressing -> NUMA Nodes Per Socket -> NPS2<br># Enable L3asNUMA - Advanced -> AMD CBS -> DF Common -> ACPI -> ACPI SRAT L3 Cache As NUMA Domain -> Enable<br># Increase DDR Frequency - Advanced -> AMD CBS -> UMC Common -> DDR4 Common -> DRAM Timing Config -> Accept -> Overclock -> Enabled -> Memory Clock Speed -> 1467MHz<br># Disable APB - Advanced -> AMD CBS -> NBIO Common -> SMU Common -> APBDIS=1<br>Advanced -> AMD CBS -> NBIO Common -> SMU Common -> Fixed SOC Pstate=P0<br># Set performance mode -<br>Advanced -> AMD CBS -> NBIO Common -> SMU Common -> Determinism Control -> Manual<br>Advanced -> AMD CBS -> NBIO Common -> SMU Common -> Determinism Slider -> Performance<br># Enable Preferred IO - Advanced -> AMD CBS -> NBIO Common -> Preferred IO -> Manual -> PCI bus number |
| BOOT Settings | crashkernel=auto selinux=0 rhgb quiet LANG=en_US.UTF-8 iommu=pt default_hugepagesz=1G hugepagesz=1G hugepages=128 isolcpus=64-95,192-223 nohz_full=60-250 numa_balancing=disable processor.max_cstate=0 nosoftlockup rcu_nocbs=60-250 |
| DPDK Settings | Enable mlx5 PMD before compiling DPDK:<br>In .config file generated by "make config",<br>set: "CONFIG_RTE_LIBRTE_MLX5_PMD=y"<br>During testing, l3fwd was given real-time scheduling priority. |
| L3fwd settings | Added /l3fwd/main.c:85:<br>#define RTE_TEST_RX_DESC_DEFAULT 4096<br>#define RTE_TEST_TX_DESC_DEFAULT 4096<br>Added /l3fwd/l3fwd.h:47:<br>#define MAX_PKT_BURST 64 |
| Command Line | ./examples/l3fwd/build/l3fwd -l 64-83 -n 4 --socket-mem=4096 -w 0000:c1:00.0,mprq_en=1,rxqs_min_mprq=1,mprq_log_stride_num=9,txq_inline_mpw=128 -w 0000:c1:00.1,mprq_en=1,rxqs_min_mprq=1,mprq_log_stride_num=9,txq_inline_mpw=128 -- -p 0x3 -P --config="(0,0,64),(0,1,65),(0,2,66),(0,3,67),(0,4,68),(0,5,69),(0,6,70),(0,7,71),(1,0,72),(1,1,73),(1,2,74),(1,3,75),(1,4,76),(1,5,77),(1,6,78),(1,7,79)" --eth-dest=0,00:11:22:33:44:50 --eth-dest=1,00:11:22:33:44:60 |
| Other optimizations | a) Flow Control OFF: "ethtool -A $netdev rx off tx off" (for both ports)<br>b) Memory optimizations: "sysctl -w vm.zone_reclaim_mode=0"; "sysctl -w vm.swappiness=0"<br>c) Move all IRQs to far NUMA node: "IRQBALANCE_BANNED_CPUS=$LOCAL_NUMA_CPUMAP irqbalance --oneshot"<br>d) Disable irqbalance: "systemctl stop irqbalance"<br>e) Change PCI MaxReadReq to 1024B for each port of each NIC:<br>Run "setpci -s $PORT_PCI_ADDRESS 68.w", it will return 4 digits ABCD --><br>Run "setpci -s $PORT_PCI_ADDRESS 68.w=3BCD"<br>f) Set CQE COMPRESSION to "AGGRESSIVE": mlxconfig -d $PORT_PCI_ADDRESS set CQE_COMPRESSION=1<br>g) Set PCI write ordering: mlxconfig -d $PORT_PCI_ADDRESS set PCI_WR_ORDERING=1<br>h) Disable Linux realtime throttling: echo -1 > /proc/sys/kernel/sched_rt_runtime_us<br>i) Disable auto neg for both ports: ethtool -s $PORT_PCI_ADDRESS autoneg off speed 100000 |

## 2.2 Test Results

***Table 4: Test #1 Results – Mellanox ConnectX-6 2x100GbE Throughput at Zero Packet Loss***

| Frame Size (Bytes) | Frame Rate (Mpps) | Line Rate [200G] (Mpps) | % Line Rate |
|---|---|---|---|
| 64 | 207.03 | 297.62 | 69.6 |
| 128 | 124.21 | 168.92 | 73.5 |
| 256 | 65.57 | 90.58 | 72.4 |
| 512 | 39.60 | 46.99 | 84.3 |
| 1024 | 21.51 | 23.95 | 89.7 |
| 1280 | 17.70 | 19.23 | 92.0 |
| 1518 | 15.04 | 16.25 | 92.6 |

***Figure 2: Test #1 Results – Mellanox ConnectX-6 2x100GbE Throughput at Zero Packet Loss***



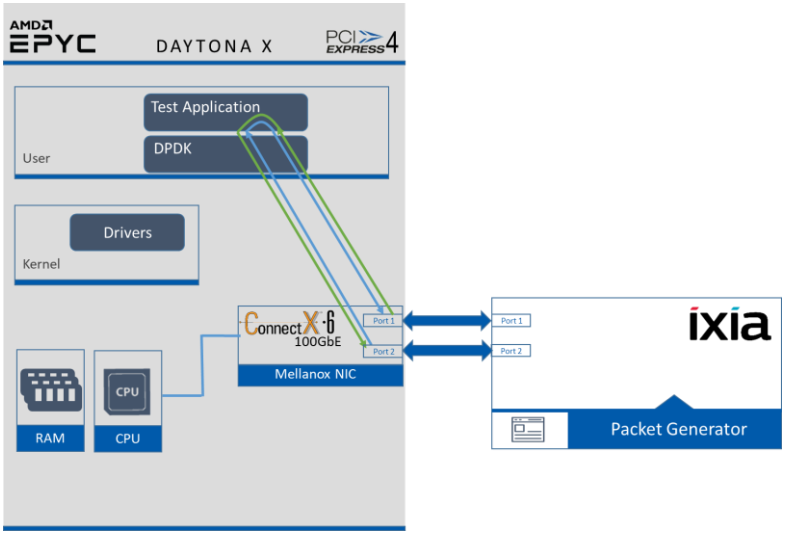
# 3 Test #2
# Mellanox ConnectX-6 Single Core Performance (2x 100GbE)

***Table 5: Test #2 Setup***

| Item | Description |
|---|---|
| Test | Test #2– Mellanox ConnectX-6 2x100GbE Single Core Performance |
| Server | AMD “Daytona X” Rome Server Reference Platform |
| CPU | 2*AMD EPYC 7742 @ 2.25GHz (running @ 3.38GHz)<br>64 CPU cores ; PCIe Gen 4.0 |
| RAM | 256GB: 16 * 16GB DIMMs @ 2666MHz |
| BIOS | American Megatrends Inc. RDY1003A 08/23/2019 |
| NIC | One ConnectX-6 VPI adapter card; 200GbE; dual-port QSFP56; PCIe4.0 x16; |
| Operating System | Red Hat Enterprise Linux Server 7.6 |
| Kernel Version | Linux 3.10.0-957.el7.x86_64 |
| GCC version | 4.8.5 20150623 (Red Hat 4.8.5-36) (GCC) |
| Mellanox NIC firmware version | 20.26.1040 |
| Mellanox OFED driver version | MLNX_OFED_LINUX-4.7-1.0.0.1 |
| DPDK version | 19.08 |
| Test Configuration | 1 NICs, using 2 port<br>Each port receives a stream of 8192 UDP flows from the IXIA<br>Each port has 1 queue assigned, a total of two queues for two ports and both queues are assigned to the same single logical core. |

Device Under Test (DUT) is made up of the AMD “Daytona X” Rome Server reference platform and a Mellanox ConnectX-6 NIC utilizing two ports. The DUT is connected to the IXIA packet generator which generates traffic towards the first port and second ports of the ConnectX-6 NIC.
The ConnectX-6 data traffic is passed via PCIe Gen 4 bus, through DPDK to the test application **testpmd** and is redirected to the opposite port. IXIA measures throughput and packet loss.

***Figure 3: Test #2 Setup – Mellanox ConnectX-6 2x100GbE connected to IXIA***



## 3.1 Test Settings

***Table 6: Test #2 Settings***

| Item | Description |
|---|---|
| BOOT Settings | # Enable X2APIC - Advanced -> AMD CBS -> CPU Common -> Local APIC Mode -> X2APIC<br># Enable NPS2 - Advanced -> AMD CBS -> DF Common -> Memory Addressing -> NUMA Nodes Per Socket -> NPS2<br># Enable L3asNUMA - Advanced -> AMD CBS -> DF Common -> ACPI -> ACPI SRAT L3 Cache As NUMA Domain -> Enable<br># Increase DDR Frequency - Advanced -> AMD CBS -> UMC Common -> DDR4 Common -> DRAM Timing Config -> Accept -> Overclock -> Enabled -> Memory Clock Speed -> 1467MHz<br># Disable APB - Advanced -> AMD CBS -> NBIO Common -> SMU Common -> APBDIS=1<br>Advanced -> AMD CBS -> NBIO Common -> SMU Common -> Fixed SOC Pstate=P0<br># Set performance mode -<br>Advanced -> AMD CBS -> NBIO Common -> SMU Common -> Determinism Control -> Manual<br>Advanced -> AMD CBS -> NBIO Common -> SMU Common -> Determinism Slider -> Performance<br># Enable Preferred IO - Advanced -> AMD CBS -> NBIO Common -> Preferred IO -> Manual -> PCI bus number |
| DPDK Settings | Enable mlx5 PMD before compiling DPDK:<br>In .config file generated by "make config",<br>set: "CONFIG_RTE_LIBRTE_MLX5_PMD=y"<br>set: "CONFIG_RTE_TEST_PMD_RECORD_CORE_CYCLES=y"<br>During testing, testpmd was given real-time scheduling priority. |
| Command Line | ./testpmd -l 80-83 -n 4 --socket-mem=4096 -w 0000:c1:00.0 -w 0000:c1:00.1 -- --burst=64 --txd=512 --rxd=512 --mbcache=512 --rxq=1 --txq=1 --nb-cores=1 --rss-udp --disable-crc-strip --forward-mode=io -a -i |
| Other optimizations | a) Flow Control OFF: "ethtool -A $netdev rx off tx off" (for both ports)<br>b) Memory optimizations: "sysctl -w vm.zone_reclaim_mode=0"; "sysctl -w vm.swappiness=0"<br>c) Move all IRQs to far NUMA node: "IRQBALANCE_BANNED_CPUS=$LOCAL_NUMA_CPUMAP irqbalance --oneshot"<br>d) Disable irqbalance: "systemctl stop irqbalance"<br>e) Change PCI MaxReadReq to 1024B for each port of each NIC:<br>Run "setpci -s $PORT_PCI_ADDRESS 68.w", it will return 4 digits ABCD --><br>Run "setpci -s $PORT_PCI_ADDRESS 68.w=3BCD"<br>f) Set CQE COMPRESSION to "AGGRESSIVE": mlxconfig -d $PORT_PCI_ADDRESS set CQE_COMPRESSION=1<br>g) Set PCI write ordering: mlxconfig -d $PORT_PCI_ADDRESS set PCI_WR_ORDERING=1<br>h) Disable Linux realtime throttling: echo -1 > /proc/sys/kernel/sched_rt_runtime_us<br>i) Disable auto neg for both ports: ethtool -s $PORT_PCI_ADDRESS autoneg off speed 100000 |

## 3.2 Test Results

***Table 7: Test #2 Results – Mellanox ConnectX-6 2x100GbE Single Core Performance***

| Frame Size (Bytes) | Frame Rate (Mpps) | Line Rate [100G] (Mpps) | Throughput (Gbps) | CPU Frequency (GHz) | CPU Cycles per packet<br>NOTE: Lower is Better |
|---|---|---|---|---|---|
| 64 | 62.44 | 148.81 | 31.96 | 3.38 | 34 |
| 128 | 61.20 | 84.46 | 62.67 | 3.38 | 36 |
| 256 | 57.06 | 45.29 | 116.86 | 3.38 | 38 |
| 512 | 37.50 | 23.50 | 153.60 | 3.38 | 35 |
| 1024 | 20.90 | 11.97 | 171.21 | 3.38 | 35 |
| 1280 | 16.26 | 9.62 | 166.50 | 3.38 | 36 |
| 1518 | 13.26 | 8.13 | 161.03 | 3.38 | 37 |

***Figure 4: Test #2 Results – Mellanox ConnectX-6 2x100GbE Single Core Performance***