Mellanox NIC’s Performance Report with DPDK 17.02

Rev 1.0

NOTE:
THIS HARDWARE, SOFTWARE OR TEST SUITE PRODUCT ("PRODUCT(S)") AND ITS RELATED DOCUMENTATION ARE PROVIDED BY MELLANOX TECHNOLOGIES "AS-IS" WITH ALL FAULTS OF ANY KIND AND SOLELY FOR THE PURPOSE OF AIDING THE CUSTOMER IN TESTING APPLICATIONS THAT USE THE PRODUCTS IN DESIGNATED SOLUTIONS. THE CUSTOMER'S MANUFACTURING TEST ENVIRONMENT HAS NOT MET THE STANDARDS SET BY MELLANOX TECHNOLOGIES TO FULLY QUALIFY THE PRODUCT(S) AND/OR THE SYSTEM USING IT. THEREFORE, MELLANOX TECHNOLOGIES CANNOT AND DOES NOT GUARANTEE OR WARRANT THAT THE PRODUCTS WILL OPERATE WITH THE HIGHEST QUALITY. ANY EXPRESS OR IMPLIED WARRANTIES, INCLUDING, BUT NOT LIMITED TO, THE IMPLIED WARRANTIES OF MERCHANTABILITY, FITNESS FOR A PARTICULAR PURPOSE AND NONINFRINGEMENT ARE DISCLAIMED. IN NO EVENT SHALL MELLANOX BE LIABLE TO CUSTOMER OR ANY THIRD PARTIES FOR ANY DIRECT, INDIRECT, SPECIAL, EXEMPLARY, OR CONSEQUENTIAL DAMAGES OF ANY KIND (INCLUDING, BUT NOT LIMITED TO, PAYMENT FOR PROCUREMENT OF SUBSTITUTE GOODS OR SERVICES; LOSS OF USE, DATA, OR PROFITS; OR BUSINESS INTERRUPTION) HOWEVER CAUSED AND ON ANY THEORY OF LIABILITY, WHETHER IN CONTRACT, STRICT LIABILITY, OR TORT (INCLUDING NEGLIGENCE OR OTHERWISE) ARISING IN ANY WAY FROM THE USE OF THE PRODUCT(S) AND RELATED DOCUMENTATION EVEN IF ADVISED OF THE POSSIBILITY OF SUCH DAMAGE.

Mellanox Technologies
350 Oakmead Parkway Suite 100
Sunnyvale, CA 94085
U.S.A.
www.mellanox.com
Tel: (408) 970-3400
Fax: (408) 970-3403

Table of Contents

List of Figures

List of Tables

Document Revision History

Table 1: Document Revision History

Revision	Date	Description
1.0	11-May-2017	Initial report release

About this Report

The purpose of this report is to provide packet rate performance data for Mellanox ConnectX-4 and ConnectX-4 Lx 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. HPE® ProLiant DL380 Gen9 Server
2. Mellanox ConnectX® NIC
3. IXIA® XM12 packet generator

Tests utilize testpmd (http://dpdk.org/doc/guides/testpmd_app_ug/index.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.

1.2 Test Procedure

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 UDP flows is generated per port.
- IxNetwork (Version 8.00EA) is used with the IXIA packet generator.

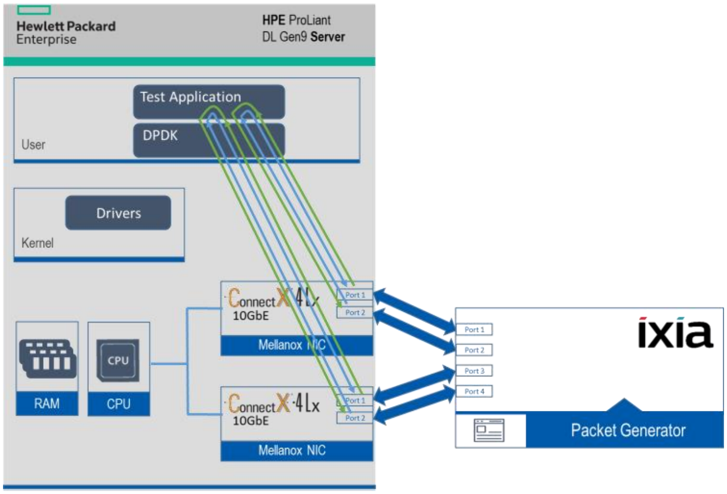
2 Test #1 Mellanox ConnectX-4 Lx 10GbE Throughput at Zero Packet Loss

Table 2: Test #1 Setup

Item	Description
Test	Test #2 – Mellanox ConnectX-4 Lx 10GbE Throughput at zero packet loss
Server	HPE ProLiant DL380 Gen 9
CPU	Intel® Xeon® CPU E5-2697A v4 (Broadwell) @ 2.60GHz 16 cpus * 2 NUMA nodes
RAM	256GB: 4 * 32GB DIMMs * 2 NUMA nodes @ 2400MHz
BIOS	P89 v2.00 (12/27/2015)
NIC	Two of MCX4121A-XCA - ConnectX-4 Lx network interface card; 10GbE dual-port SFP28; PCIe3.0 x8; ROHS R6
Operating System	Red Hat Enterprise Linux Server 7.2 (Maipo)
Kernel Version	3.10.0-327.el7.x86_64
GCC version	4.8.5 20150623 (Red Hat 4.8.5-4) (GCC)
Mellanox NIC firmware version	14.18.2000
Mellanox OFED driver version	MLNX_OFED_LINUX-4.0-2.0.0.1
DPDK version	17.02.0
Test Configuration	2 NICs, 2 ports used on each NIC. Each port has 1 queue assigned for a total of 4 queues. 1 queue assigned per logical core for a total of 4 logical cores for 4 ports. Each port receives a stream of 8192 UDP flows from the IXIA

Device Under Test (DUT) is made up of the HPE server and two Mellanox ConnectX-4 Lx NICs with two 10GbE ports each (total 4 ports). The DUT is connected to the IXIA packet generator which generates traffic towards each of the ConnectX-4 Lx NIC ports. The ConnectX-4 Lx received data traffic is passed through DPDK to the test application testpmd and is redirected to the opposite port on the same NIC. IXIA measures throughput with zero packet loss.

Figure 1: Test #1 Setup – Mellanox ConnectX-4 Lx 10GbE connected to IXIA



2.1 Test Settings

Table 3 : Test #1 Settings

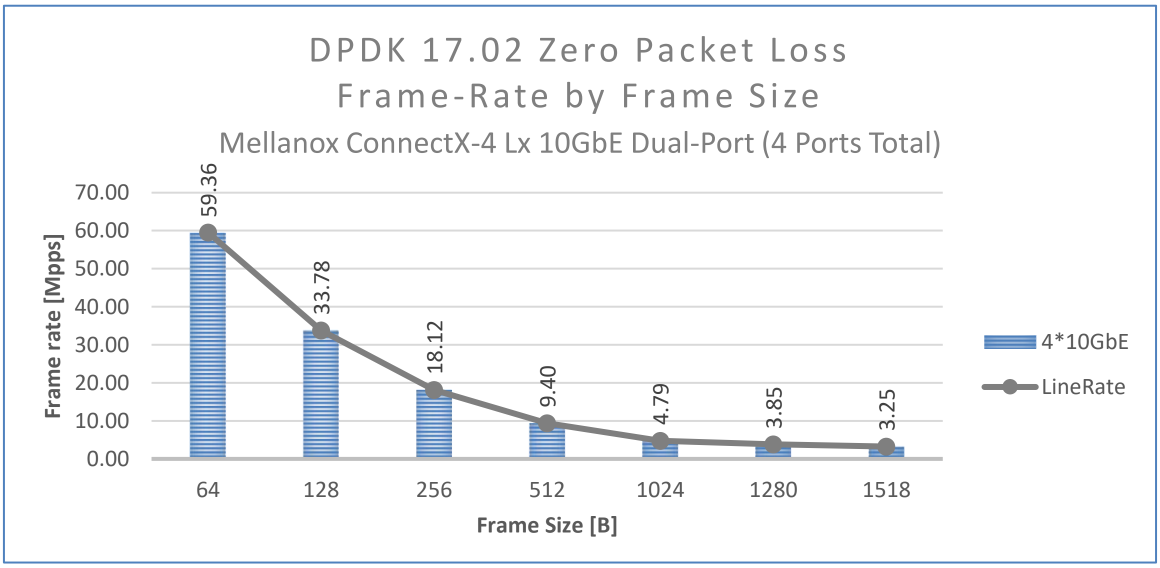
Item	Description
BIOS	Boot in “Legacy BIOS mode” Power Profile PERFORMANCE; C-states OFF; P-states OFF; TurboBoost ON; HyperThreading OFF; Virt OFF; VT-d OFF; SR-IOV OFF; SMI OFF
BOOT Settings	isolcpus=0-7,16-23 intel_idle.max_cstate=0 processor.max_cstate=0 intel_pstate=disable nohz_full=0-7,16-23 rcu_nocbs=0-7,16-23 rcu_novb_poll default_hugepagesz=1G hugepagesz=1G hugepages=64 audit=0 nosoftlockup
DPDK Settings	Enable mlx5 PMD before compiling DPDK: In .config file generated by "make config", set: "CONFIG_RTE_LIBRTE_MLX5_PMD=y" During testing, testpmd was given real-time scheduling priority.
Command Line	/root/dpdk/build/app/testpmd -c 0x80f00000 --master-lcore=31 -n 4 -w 05:00.0 -w 05:00.1 -w 0b:00.0 -w 0b:00.1 --socket-mem=8192,256 -- --port-numa-config=0,0,1,0,2,0,3,0 --socket-num=0 --burst=64 --txd=4096 --rxd=4096 --mbcache=512 --rxq=1 --txq=1 --nb-cores=4 -i -a --rss-udp
Other optimizations	a) Flow Control OFF: "ethtool -A $netdev rx off tx off" b) Memory optimizations: "sysctl -w vm.zone_reclaim_mode=0"; "sysctl -w vm.swappiness=0" c) Move all IRQs to far NUMA node: "IRQBALANCE_BANNED_CPUS=$LOCAL_NUMA_CPUMAP irqbalance --oneshot" d) Disable irqbalance: "systemctl stop irqbalance" e) Change PCI MaxReadReq to 1024B for each port of each NIC: Run "setpci -s $PORT_PCI_ADDRESS 68.w", it will return 4 digits ABCD --> Run "setpci -s $PORT_PCI_ADDRESS 68.w=3BCD"

2.2 Test Results

Table 4: Test #1 Results – Mellanox ConnectX-4 Lx 10GbE Throughput at Zero Packet Loss

Frame Size (Bytes)	Throughput (Mpps)	Line Rate Throughput (Mpps)	% Line Rate
64	59.36	59.52	99.72
128	33.78	33.78	100
256	18.12	18.12	100
512	9.40	9.40	100
1024	4.79	4.79	100
1280	3.85	3.85	100
1518	3.25	3.25	100

Figure 2: Test #2 Results – Mellanox ConnectX-4 Lx 10GbE Throughput at Zero Packet Loss



3 Test #2 Mellanox ConnectX-4 Lx 40GbE Throughput at Zero Packet Loss

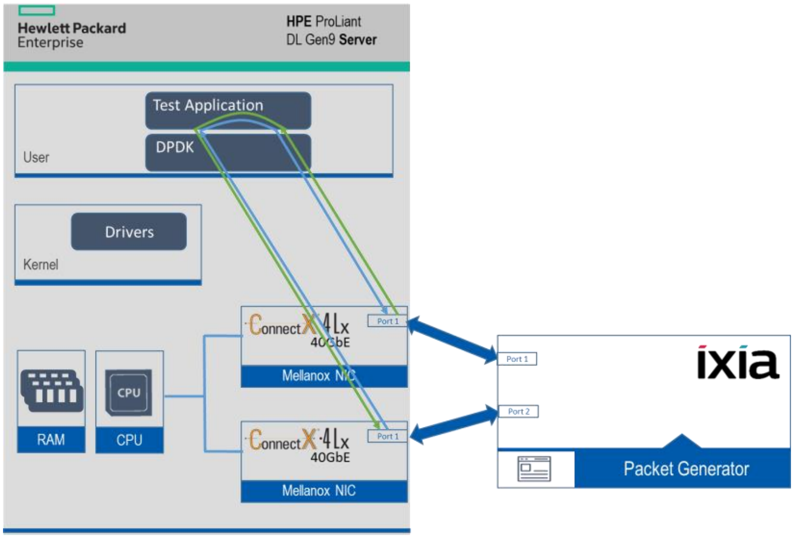
Table 5: Test #2 Setup

Item	Description
Test	Test #1 – Mellanox ConnectX-4 Lx 40GbE Throughput at zero packet loss
Server	HPE ProLiant DL380 Gen 9
CPU	Intel(R) Xeon(R) CPU E5-2697A v4 (Broadwell) @ 2.60GHz 16 cpus * 2 NUMA nodes
RAM	256GB: 4 * 32GB DIMMs * 2 NUMA nodes @ 2400MHz
BIOS	P89 v2.00 (12/27/2015)
NIC	Two of MCX4131A-BCA - ConnectX-4 Lx network interface card; 40GbE single-port QSFP28; PCIe3.0 x8; ROHS R6
Operating System	Red Hat Enterprise Linux Server 7.2 (Maipo)
Kernel Version	3.10.0-327.el7.x86_64
GCC version	4.8.5 20150623 (Red Hat 4.8.5-4) (GCC)
Mellanox NIC firmware version	14.18.2000
Mellanox OFED driver version	MLNX_OFED_LINUX-4.0-2.0.0.1
DPDK version	17.02.0
Test Configuration	2 NICs, 1 port used on each NIC. Each port has 2 queues assigned for a total of 4 queues 1 queue assigned per logical core with a total of 4 logical cores and 4 queues for 2 ports Each port receives a stream of 8192 UDP flows from the IXIA

Device Under Test (DUT) is made up of the HPE server and the two Mellanox ConnectX-4 Lx NICs with one 40GbE port each (total of 2 ports). The DUT is connected to the IXIA packet generator which generates traffic towards each of the ConnectX-4 Lx NICs.

The ConnectX-4 Lx data traffic is passed through DPDK to the test application testpmd and is redirected to the opposite card's port. IXIA measures throughput and packet loss.

Figure 3: Test #2 Setup – Mellanox ConnectX-4 Lx 40GbE connected to IXIA



3.1 Test Settings

Table 6: Test #2 Settings

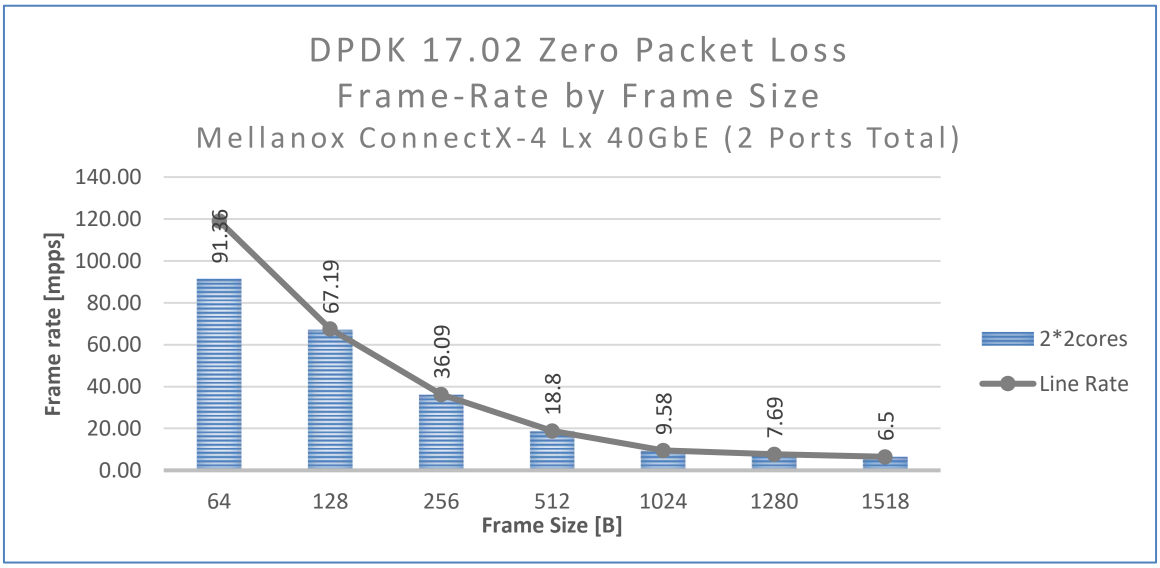
Item	Description
BIOS	Boot in “Legacy BIOS mode” Power Profile PERFORMANCE; C-states OFF; P-states OFF; TurboBoost ON; HyperThreading OFF; Virt OFF; VT-d OFF; SR-IOV OFF; SMI OFF
BOOT Settings	isolcpus=0-7,16-23 intel_idle.max_cstate=0 processor.max_cstate=0 intel_pstate=disable nohz_full=0-7,16-23 rcu_nocbs=0-7,16-23 rcu_novb_poll default_hugepagesz=1G hugepagesz=1G hugepages=64 audit=0 nosoftlockup
DPDK Settings	Enable mlx5 PMD before compiling DPDK: In .config file generated by "make config", set: "CONFIG_RTE_LIBRTE_MLX5_PMD=y" During testing, testpmd was given real-time scheduling priority.
Command Line	/root/dpdk/build/app/testpmd -c 0xff00ff02 --master-lcore=1 -n 4 -w 84:00.0 -w 81:00.0 --socket-mem=256,8192 -- --port-numa-config=0,1,1,1 --socket-num=1 --burst=64 --txd=4096 --rxd=4096 --mbcache=512 --rxq=2 --txq=2 --nb-cores=4 -i -a --rss-udp
Other optimizations	a) Flow Control OFF: "ethtool -A $netdev rx off tx off" b) Memory optimizations: "sysctl -w vm.zone_reclaim_mode=0"; "sysctl -w vm.swappiness=0" c) Move all IRQs to far NUMA node: "IRQBALANCE_BANNED_CPUS=$LOCAL_NUMA_CPUMAP irqbalance --oneshot" d) Disable irqbalance: "systemctl stop irqbalance" e) Change PCI MaxReadReq to 1024B for each port of each NIC: Run "setpci -s $PORT_PCI_ADDRESS 68.w", it will return 4 digits ABCD --> Run "setpci -s $PORT_PCI_ADDRESS 68.w=3BCD"

3.2 Test Results

Table 7: Test #2 Results – Mellanox ConnectX-4 Lx 40GbE Throughput at Zero Packet Loss

Frame Size (Bytes)	Throughput (Mpps)	Line Rate Throughput (Mpps)	% Line Rate
64	91.36	119.05	76.74
128	67.19	67.57	99.44
256	36.09	36.23	99.61
512	18.80	18.80	100
1024	9.58	9.58	100
1280	7.69	7.69	100
1518	6.50	6.50	100

Figure 4: Test #2 Results – Mellanox ConnectX-4 Lx 40GbE Throughput at Zero Packet Loss



4 Test #3
Mellanox ConnectX-4 100GbE Throughput at Zero Packet Loss

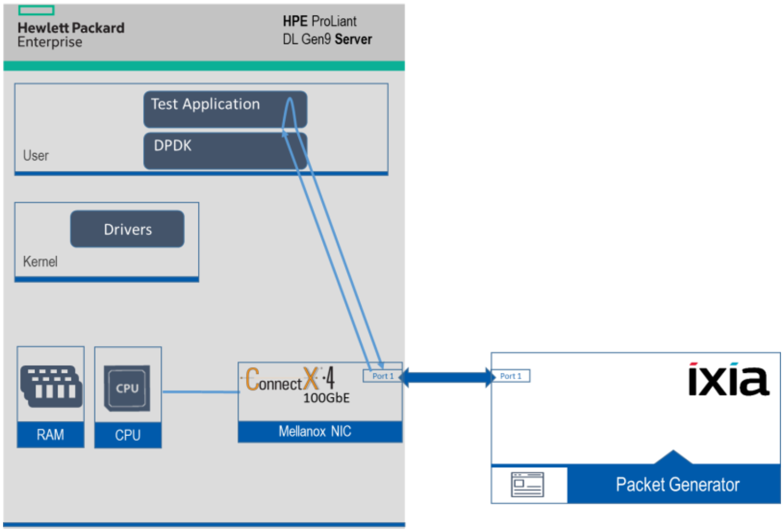
Table 8: Test #3 Setup

Item	Description
Test	Test #1 – Mellanox ConnectX-4 100GbE Throughput at zero packet loss
Server	HPE ProLiant DL380 Gen 9
CPU	Intel(R) Xeon(R) CPU E5-2697A v4 (Broadwell) @ 2.60GHz 16 cpus * 2 NUMA nodes
RAM	256GB: 4 * 32GB DIMMs * 2 NUMA nodes @ 2400MHz
BIOS	P89 v2.00 (12/27/2015)
NIC	One MCX415A-CCAT- ConnectX-4 network interface card 100GbE single-port QSFP28; PCIe3.0 x16; ROHS R6
Operating System	Red Hat Enterprise Linux Server 7.2 (Maipo)
Kernel Version	3.10.0-327.el7.x86_64
GCC version	4.8.5 20150623 (Red Hat 4.8.5-4) (GCC)
Mellanox NIC firmware version	12.18.2000
Mellanox OFED driver version	MLNX_OFED_LINUX-4.0-2.0.0.1
DPDK version	17.02.0
Test Configuration	1 NIC, 1 port used on NIC, The port has 8 queues assigned to it, 1 queue per logical core for a total of 8 logical cores. Each port receives a stream of 8192 UDP flows from the IXIA

Device Under Test (DUT) is made up of the HPE server and the Mellanox ConnectX-4 NIC with a single port. The DUT is connected to the IXIA packet generator which generates traffic towards the ConnectX-4 NIC.

The ConnectX-4 data traffic is passed through DPDK to the test application testpmd and is redirected to the opposite direction on the same port. IXIA measures throughput and packet loss.

Figure 5: Test #3 Setup – Mellanox ConnectX-4 100GbE connected to IXIA



4.1 Test Settings

Table 9: Test #3 Settings

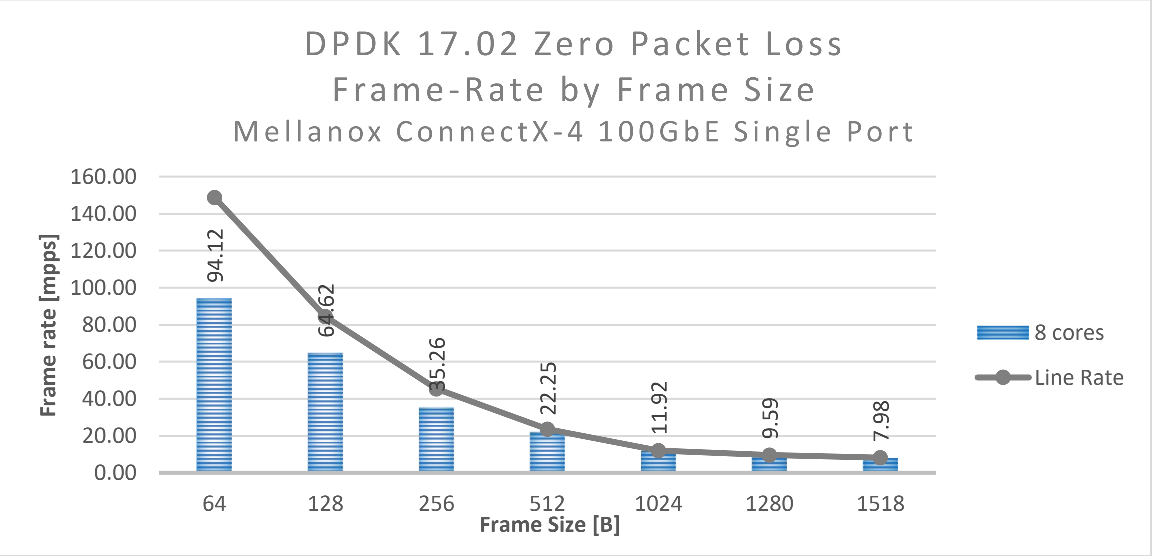
Item	Description
BIOS	Boot in “Legacy BIOS mode” Power Profile PERFORMANCE; C-states OFF; P-states OFF; TurboBoost ON; HyperThreading OFF; Virt OFF; VT-d OFF; SR-IOV OFF; SMI OFF
BOOT Settings	isolcpus=0-7,16-23 intel_idle.max_cstate=0 processor.max_cstate=0 intel_pstate=disable nohz_full=0-7,16-23 rcu_nocbs=0-7,16-23 rcu_novb_poll default_hugepagesz=1G hugepagesz=1G hugepages=64 audit=0 nosoftlockup
DPDK Settings	Enable mlx5 PMD before compiling DPDK: In .config file generated by "make config", set: "CONFIG_RTE_LIBRTE_MLX5_PMD=y" During testing, testpmd was given real-time scheduling priority.
Command Line	/root/dpdk/build/app/testpmd -c 0xff008000 -n 4 -w 88:00.0,txq_inline=128 --socket-mem=256,8192 -- --port-numa-config=0,1 --socket-num=1 --burst=64 --txd=4096 --rxd=4096 --mbcache=512 --rxq=8 --txq=8 --nb-cores=8 -i -a --rss-udp
Other optimizations	a) Flow Control OFF: "ethtool -A $netdev rx off tx off" b) Memory optimizations: "sysctl -w vm.zone_reclaim_mode=0"; "sysctl -w vm.swappiness=0" c) Move all IRQs to far NUMA node: "IRQBALANCE_BANNED_CPUS=$LOCAL_NUMA_CPUMAP irqbalance --oneshot" d) Disable irqbalance: "systemctl stop irqbalance" e) Change PCI MaxReadReq to 1024B for each port of each NIC: Run "setpci -s $PORT_PCI_ADDRESS 68.w", it will return 4 digits ABCD --> Run "setpci -s $PORT_PCI_ADDRESS 68.w=3BCD"

4.2 Test Results

Table 10: Test #3 Results – Mellanox ConnectX-4 100GbE Throughput at Zero Packet Loss

Frame Size (Bytes)	Throughput (Mpps)	Line Rate Throughput (Mpps)	% Line Rate
64	94.12	148.81	63.24
128	64.62	84.46	76.50
256	35.26	45.29	77.85
512	22.25	23.5	94.68
1024	11.92	11.97	99.58
1280	9.59	9.61	99.79
1518	7.98	8.13	98.15

Figure 6: Test #3 Results – Mellanox ConnectX-4 100GbE Throughput at Zero Packet Loss


DPDK 17.02 Zero Packet Loss
Frame-Rate by Frame Size
Mellanox ConnectX-4 100GbE Single Port
Frame rate [mpps]
Frame Size [B]
8 cores
Line Rate