



NVIDIA Mellanox NIC's Performance Report with DPDK 21.02

Rev 1.1

Notice

This document is provided for information purposes only and shall not be regarded as a warranty of a certain functionality, condition, or quality of a product. NVIDIA Corporation ("NVIDIA") makes no representations or warranties, expressed or implied, as to the accuracy or completeness of the information contained in this document and assumes no responsibility for any errors contained herein. NVIDIA shall have no liability for the consequences or use of such information or for any infringement of patents or other rights of third parties that may result from its use. This document is not a commitment to develop, release, or deliver any Material (defined below), code, or functionality.

NVIDIA reserves the right to make corrections, modifications, enhancements, improvements, and any other changes to this document, at any time without notice.

Customer should obtain the latest relevant information before placing orders and should verify that such information is current and complete.

NVIDIA products are sold subject to the NVIDIA standard terms and conditions of sale supplied at the time of order acknowledgement, unless otherwise agreed in an individual sales agreement signed by authorized representatives of NVIDIA and customer ("Terms of Sale"). NVIDIA hereby expressly objects to applying any customer general terms and conditions with regards to the purchase of the NVIDIA product referenced in this document. No contractual obligations are formed either directly or indirectly by this document.

NVIDIA products are not designed, authorized, or warranted to be suitable for use in medical, military, aircraft, space, or life support equipment, nor in applications where failure or malfunction of the NVIDIA product can reasonably be expected to result in personal injury, death, or property or environmental damage. NVIDIA accepts no liability for inclusion and/or use of NVIDIA products in such equipment or applications and therefore such inclusion and/or use is at customer's own risk.

NVIDIA makes no representation or warranty that products based on this document will be suitable for any specified use. Testing of all parameters of each product is not necessarily performed by NVIDIA. It is customer's sole responsibility to evaluate and determine the applicability of any information contained in this document, ensure the product is suitable and fit for the application planned by customer, and perform the necessary testing for the application in order to avoid a default of the application or the product. Weaknesses in customer's product designs may affect the quality and reliability of the NVIDIA product and may result in additional or different conditions and/or requirements beyond those contained in this document. NVIDIA accepts no liability related to any default, damage, costs, or problem which may be based on or attributable to: (i) the use of the NVIDIA product in any manner that is contrary to this document or (ii) customer product designs.

No license, either expressed or implied, is granted under any NVIDIA patent right, copyright, or other NVIDIA intellectual property right under this document. Information published by NVIDIA regarding third-party products or services does not constitute a license from NVIDIA to use such products or services or a warranty or endorsement thereof. Use of such information may require a license from a third party under the patents or other intellectual property rights of the third party, or a license from NVIDIA under the patents or other intellectual property rights of NVIDIA.

Reproduction of information in this document is permissible only if approved in advance by NVIDIA in writing, reproduced without alteration and in full compliance with all applicable export laws and regulations, and accompanied by all associated conditions, limitations, and notices.

Trademarks

NVIDIA, the NVIDIA logo, and Mellanox are trademarks and/or registered trademarks of NVIDIA Corporation in the U.S. and other countries. Other company and product names may be trademarks of the respective companies with which they are associated.

For the complete and most updated list of Mellanox trademarks, visit <http://www.mellanox.com/page/trademarks>.

Copyright

© 2021 NVIDIA Corporation. All rights reserved.



Table of Contents

1	About this Report.....	8
1.1	Target Audience	8
1.2	Terms and Conventions.....	8
2	Test Description	9
2.1	Hardware Components	9
2.2	Zero Packet Loss Test	9
2.3	Zero Packet Loss over SR-IOV Test	9
2.4	Single Core Performance Test	9
3	Test#1 Mellanox ConnectX-4 Lx 25GbE Throughput at Zero Packet Loss (2x 25GbE)	10
3.1	Test Settings.....	11
3.2	Test Results	12
4	Test#2 Mellanox ConnectX-5 25GbE Throughput at Zero Packet Loss (2x 25GbE).....	13
4.1	Test Settings.....	14
4.2	Test Results	15
5	Test#3 Mellanox ConnectX-5 Ex 100GbE Throughput at Zero Packet Loss (1x 100GbE)	16
5.1	Test Settings.....	17
5.2	Test Results	18
6	Test#4 Mellanox ConnectX-5 Ex 100GbE Single Core Performance (2x 100GbE)	19
6.1	Test Settings.....	20
6.2	Test Results	21
7	Test#5 Mellanox ConnectX-5 25GbE Single Core Performance (2x 25GbE)	22
7.1	Test Settings.....	23
7.2	Test Results	24
8	Test#6 Mellanox ConnectX-5 25GbE Throughput at Zero Packet Loss (2x 25GbE) using SR-IOV over VMware ESXi 7.0U1	25
8.1	Test Settings.....	27
8.2	Test Results	28
9	Test#7 Mellanox ConnectX-5 Ex 100GbE Throughput at Zero Packet Loss (1x 100GbE) using SR-IOV over KVM Hypervisor.....	29
9.1	Test Settings.....	30
9.2	Test Results	33
10	Test#8 Mellanox ConnectX-6Dx 25GbE Throughput at Zero Packet Loss (2x 25GbE)	34
10.1	Test Settings.....	35
10.2	Test Results	36
11	Test#9 Mellanox ConnectX-6 Dx 100GbE Throughput at Zero Packet Loss (1x 100GbE).....	37
11.1	Test Settings.....	38
11.2	Test Results	39
12	Test#10 Mellanox ConnectX-6Dx 100GbE Single Core Performance (2x 100GbE)	40
12.1	Test Settings.....	41
12.2	Test Results	42
13	Test#11 Mellanox ConnectX-6 Dx 100GbE PCIe Gen4 Throughput at Zero Packet Loss (2x 100GbE)	43
13.1	Test Settings.....	44
13.2	Test Results	45
14	Test#12 Mellanox ConnectX-6 Dx 100GbE Throughput at Zero Packet Loss (1x 100GbE) using SR-IOV over KVM Hypervisor.....	46

14.1	Test Settings	48
14.2	Test Results	50
15	Test#13 Mellanox ConnectX-6 Dx 200GbE PCIe Gen4 Throughput at Zero Packet Loss (1x 200GbE)	51
15.1	Test Settings	52
15.2	Test Results	53
16	Test#14 BlueField-2 25GbE Throughput at Zero Packet Loss (2x 25GbE)	54
16.1	Test Settings	55
16.2	Test Results	56
17	Test#15 Mellanox ConnectX-6 Lx 25GbE Throughput at Zero Packet Loss (2x 25GbE)	57
17.1	Test Settings	58
17.2	Test Results	59

List of Figures

Figure 1: Test #1 Setup – Mellanox ConnectX-4 Lx 25GbE Dual-Port connected to IXIA	10
Figure 2: Test #1 Results – Mellanox ConnectX-4 Lx 25GbE Dual-Port Throughput at Zero Packet Loss	12
Figure 3: Test #2 Setup – Mellanox ConnectX-5 25GbE Dual-Port connected to IXIA.....	13
Figure 4: Test #2 Results – Mellanox ConnectX-5 25GbE Dual-Port Throughput at Zero Packet Loss	15
Figure 5: Test #3 Setup – Mellanox ConnectX-5 Ex 100GbE connected to IXIA	16
Figure 6: Test #3 Results – Mellanox ConnectX-5 Ex 100GbE Throughput at Zero Packet Loss	18
Figure 7: Test #4 Setup – Two Mellanox ConnectX-5 Ex 100GbE connected to IXIA.....	19
Figure 8: Test #4 Results – Mellanox ConnectX-5 Ex 100GbE Single Core Performance	21
Figure 9: Test #5 Setup – Two Mellanox ConnectX-5 25GbE connected to IXIA	22
Figure 10: Test #5 Results – Mellanox ConnectX-5 25GbE Single Core Performance	24
Figure 11: Test #6 Setup – Mellanox ConnectX-5 25GbE connected to IXIA using ESXi SR-IOV	26
Figure 12: Test#6 Results – Mellanox ConnectX-5 25GbE Throughput at Zero Packet Loss using ESXi SR-IOV.....	28
Figure 13: Test #7 Setup – Mellanox ConnectX-5 Ex 100GbE connected to IXIA using KVM SR-IOV	30
Figure 14: Test #7 Results – Mellanox ConnectX-5 Ex 100GbE Throughput at Zero Packet Loss using KVM SR-IOV	33
Figure 15: Test #8 Setup – Mellanox ConnectX-6 Dx 25GbE Dual-Port connected to IXIA.....	34
Figure 16: Test #8 Results – Mellanox ConnectX-6Dx 25GbE Dual-Port Throughput at Zero Packet Loss	36
Figure 17: Test #9 Setup – Mellanox ConnectX-6 Dx 100GbE connected to IXIA.....	37
Figure 18: Test #9 Results – Mellanox ConnectX-5 Ex 100GbE Throughput at Zero Packet Loss	39
Figure 19: Test #10 Setup – Two Mellanox ConnectX-6 Dx 100GbE connected to IXIA	40
Figure 20: Test #10 Results – Mellanox ConnectX-6Dx 100GbE Single Core Performance	42
Figure 21: Test #11 Setup – Mellanox ConnectX-6 Dx 100GbE connected to IXIA.....	43
Figure 22: Test #11 Results – Mellanox ConnectX-6 Dx 100GbE dual port PCIe Gen4 Throughput at Zero Packet Loss	45
Figure 23 - Test #12 Setup – Mellanox ConnectX-6 Dx 100GbE connected to IXIA using KVM SR-IOV	47
Figure 24 - Test #12 Results – Mellanox ConnectX-6 Dx 100GbE Throughput at Zero Packet Loss using KVM SR-IOV.....	50
Figure 25 - Test #13 Setup – Mellanox ConnectX-6 Dx 200GbE connected to IXIA.....	51
Figure 26 - Test #13 Results – Mellanox ConnectX-6 Dx 200GbE dual port PCIe Gen4 Throughput at Zero Packet Loss	53
Figure 27 -Test #14 Setup – BlueField-2 25GbE Dual-Port connected to IXIA.....	54
Figure 28 - Test #14 Results – BlueField-2 25GbE Dual-Port Throughput at Zero Packet Loss.....	56
Figure 29 - Test #15 Setup – Mellanox ConnectX-6 Lx 25GbE Dual-Port connected to IXIA	57
Figure 30 - Test #15 Results – Mellanox ConnectX-6 Lx 25GbE Dual-Port Throughput at Zero Packet Loss	59

List of Tables

Table 1 - Document History	7
Table 2 - Terms, Abbreviations and Acronyms.....	8
Table 3: Test #1 Setup	10
Table 4: Test #1 Settings	11
Table 5: Test #1 Results – Mellanox ConnectX-4 Lx 25GbE Dual-Port Throughput at Zero Packet Loss	12
Table 6: Test #2 Setup	13
Table 7: Test #2 Settings	14
Table 8: Test #2 Results – Mellanox ConnectX-5 25GbE Dual-Port Throughput at Zero Packet Loss.....	15
Table 9: Test #3 Setup	16
Table 10: Test #3 Settings	17
Table 11: Test #3 Results – Mellanox ConnectX-5 Ex 100GbE Throughput at Zero Packet Loss	18
Table 12: Test #4 Setup	19
Table 13: Test #4 Settings	20
Table 14: Test #4 Results – Mellanox ConnectX-5 Ex 100GbE Single Core Performance	21
Table 15: Test #5 Setup	22
Table 16: Test #5 Settings	23
Table 17: Test #5 Results – Mellanox ConnectX-5 25GbE Single Core Performance.....	24
Table 18: Test #6 Setup	25
Table 19: Test#6 Settings	27
Table 20: Test#6 Results – Mellanox ConnectX-5 25GbE Throughput at Zero Packet Loss using ESXi SR-IOV.....	28
Table 21: Test #7 Setup	29
Table 22: Test #7 Settings	30
Table 23: Test #7 Results – Mellanox ConnectX-5 Ex 100GbE Throughput at Zero Packet Loss using KVM SR-IOV	33
Table 24: Test #8 Setup	34
Table 25: Test #8 Settings	35
Table 26: Test #8 Results – Mellanox ConnectX-6Dx 25GbE Dual-Port Throughput at Zero Packet Loss	36
Table 27: Test #9 Setup	37
Table 28: Test #9 Settings	38
Table 29: Test #9 Results – Mellanox ConnectX-6 Dx 100GbE Throughput at Zero Packet Loss.....	39
Table 30: Test #10 Setup	40
Table 31: Test #10 Settings	41
Table 32: Test #10 Results – Mellanox ConnectX-6 Dx 100GbE Single Core Performance.....	42
Table 33: Test #11 Setup	43
Table 34: Test #11 Settings	44
Table 35: Test #11 Results – Mellanox ConnectX-6 Dx 100GbE dual port PCIe Gen4 Zero Packet Loss Throughput.....	45
Table 36 - Test #12 Setup.....	46
Table 37 - Test #12 Settings	48
Table 38 - Test #12 Results – Mellanox ConnectX-6 Dx 100GbE Throughput at Zero Packet Loss using KVM SR-IOV.....	50
Table 39 - Test #13 Setup.....	51
Table 40 - Test #13 Settings	52
Table 41 - Test #13 Results – Mellanox ConnectX-6 Dx 200GbE single port PCIe Gen4 Throughput at Zero Packet Loss	53
Table 42 - Test #14 Setup.....	54
Table 43 - Test #14 Settings	55
Table 44 - Test #14 Results – BlueField-2 25GbE Dual-Port Throughput at Zero Packet Loss	56
Table 45 - Test #15 Setup.....	57
Table 46 - Test #15 Settings	58
Table 47 - Test #15 Results – Mellanox ConnectX-6 Lx 25GbE Dual-Port Throughput at Zero Packet Loss	59

Document History

Table 1 - Document History

Version	Date	Description of Change
1.0	07-Apr-2021	Initial report release
1.1	21-Feb-2023	Fix a typo in Test#13 results

1 About this Report

The purpose of this report is to provide packet rate performance data for NVIDIA® Mellanox® Network Interface Cards (NICs) and Data Processing Unit (DPU) (ConnectX®-4 Lx, ConnectX®-5, ConnectX®-5 Ex, ConnectX®-6 Lx, ConnectX®-6 Dx and BlueField-2) that has been achieved with the specified Data Plane Development Kit (DPDK) release. The report provides the measured packet rate performance as well as the hardware layout, procedures, and configurations for replicating these tests.

The document does not cover all network speeds available with the ConnectX® or BlueField® family of NICs / DPUs and is intended as a general reference of achievable performance for the specified DPDK release.

1.1 Target Audience

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

1.2 Terms and Conventions

The following terms, abbreviations, and acronyms are used in this document.

Table 2 - Terms, Abbreviations and Acronyms

Term	Description
DPU	Data Processing Unit
DUT	Device Under Test
MPPS	Million Packets Per Seconds
PPS	Packets Per Second
OFED	OpenFabrics Enterprise Distribution; An open-source software for RDMA & kernel bypass. Read more on Mellanox OFED here .
SR-IOV	Single Root IO Virtualization
ZPL	Zero Packet Loss

2 Test Description

2.1 Hardware Components

The following hardware components are used in the test setup:

- ▶ One of the following servers:
 - HPE® ProLiant DL380 Gen10 Server
 - Dell PowerEdge R6525
- ▶ One of the followings NICs, SmartNICs or DPUs:
 - Mellanox ConnectX-4 Lx, ConnectX-5, ConnectX-5 Ex, ConnectX-6 Lx, ConnectX-6 Dx Network Interface Cards (NICs) and BlueField-2 Data Processing Unit (DPU)
- ▶ IXIA® XM12 packet generator

2.2 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 testing 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 Device Under Test (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 9.00EA) is used with the IXIA packet generator.

2.3 Zero Packet Loss over SR-IOV Test

The test is conducted similarly to the bare-metal zero packet loss test with the distinction of having the DPDK application running in a Guest OS inside a VM utilizing SR-IOV virtual function.

2.4 Single Core Performance Test

Single Core performance tests utilize **testpmd** (http://www.dpdk.org/doc/guides/testpmd_app_ug), for testing the max throughput while using a single CPU core. The duration of the test is 60 seconds and the average throughput that is recorded during that time is used as the result of the test.

- ▶ Duration for each test is 60 seconds.
- ▶ Traffic of 8192 UDP flows is generated per port.
- ▶ IxNetwork (Version 9.00EA) is used with the IXIA packet generator.

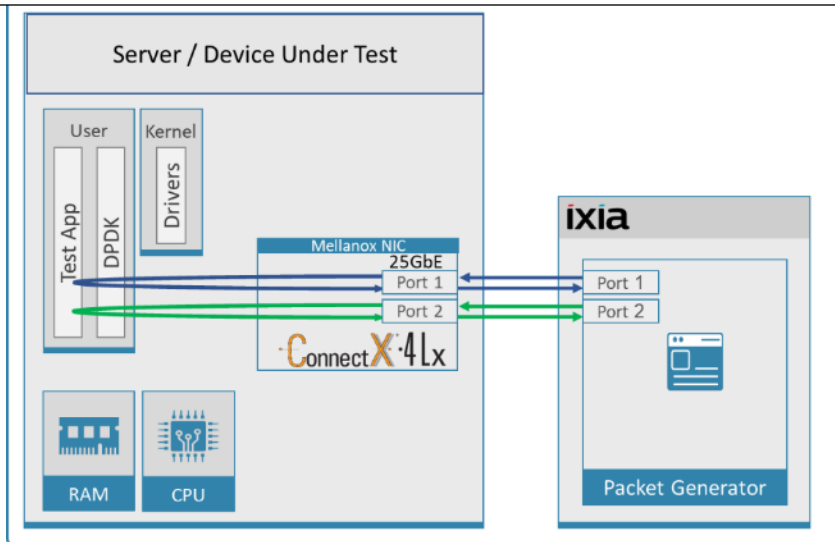
3 Test#1 Mellanox ConnectX-4 Lx 25GbE Throughput at Zero Packet Loss (2x 25GbE)

Table 3: Test #1 Setup

Item	Description
Test #1	Mellanox ConnectX-4 Lx 25GbE Dual-Port Throughput at zero packet loss
Server	HPE ProLiant DL380 Gen10
CPU	Intel(R) Xeon(R) Platinum 8168 CPU @ 2.70GHz 24 CPU cores * 2 NUMA nodes
RAM	384GB: 6 * 32GB DIMMs * 2 NUMA nodes @ 2666MHz
BIOS	U30 rev. 1.36 (02/15/2018)
NIC	One MCX4121A-ACAT - ConnectX-4 Lx network interface card 25GbE dual-port SFP28; PCIe3.0 x8; ROHS R6
Operating System	Ubuntu 20.04.2 LTS (Focal Fossa)
Kernel Version	5.4.0-65-generic.x86_64
GCC version	gcc (Ubuntu 9.3.0-17ubuntu1~20.04) 9.3.0
Mellanox NIC firmware version	14.29.2002
Mellanox OFED driver version	MLNX_OFED_LINUX-5.2-2.2.0.0
DPDK version	21.02
Test Configuration	1 NIC, 2 ports used on the NIC. Each port receives a stream of 8192 IP flows from the IXIA Each port has 4 queues assigned for a total of 8 queues 1 queue assigned per logical core with a total of 8 logical cores

The Device Under Test (DUT) is made up of the HPE server and the Mellanox ConnectX-4 Lx Dual-Port NIC. The DUT is connected to the IXIA packet generator which generates traffic towards the ConnectX-4 Lx NIC. The ConnectX-4 Lx data traffic is passed through DPDK to the test application **I3fwd** and is redirected to the opposite direction on the opposing port. IXIA measures throughput and packet loss.

Figure 1: Test #1 Setup – Mellanox ConnectX-4 Lx 25GbE Dual-Port connected to IXIA



3.1 Test Settings

Table 4: Test #1 Settings

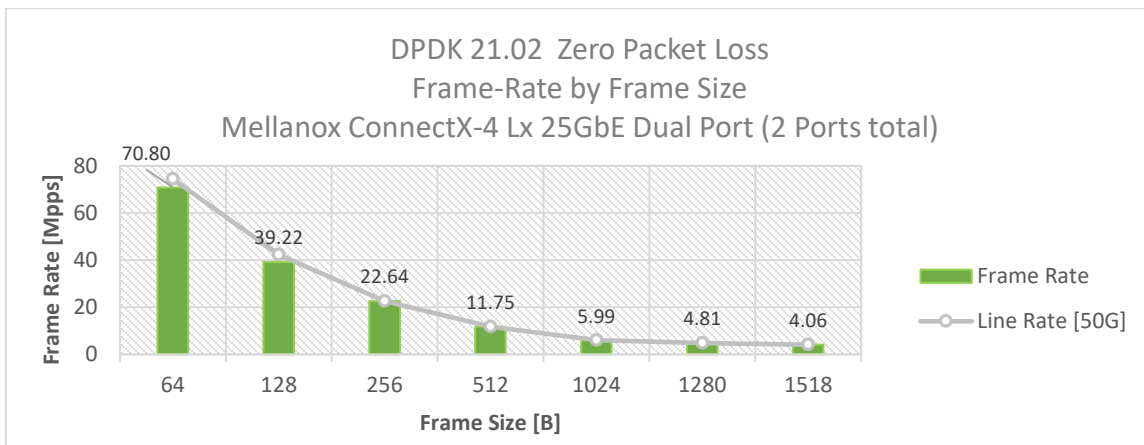
Item	Description
BIOS	<p>1) Workload Profile = "Low Latency";</p> <p>2) Jitter Control = Manual, 3400. (Setting turbo boost frequency to 3.4 GHz)</p> <p>See "Configuring and tuning HPE ProLiant Servers for low-latency applications": hpe.com > Search "DL380 gen10 low latency"</p>
BOOT Settings	<pre>isolcpus=24-47 intel_idle.max_cstate=0 processor.max_cstate=0 intel_pstate=disable nohz_full=24-47 rcu_nocbs=24-47 rcu_nocb_poll default_hugepagesz=1G hugepagesz=1G hugepages=64 audit=0 nosoftlockup</pre>
DPDK Settings	<p>Compile DPDK using: <pre>meson <build> -Dexamples=l3fwd ; ninja -C <build></pre></p> <p>During testing, l3fwd was given real-time scheduling priority.</p>
L3fwd settings	<p>Updated values /l3fwd/l3fwd.h:</p> <pre>#define RTE_TEST_RX_DESC_DEFAULT 4096 #define RTE_TEST_TX_DESC_DEFAULT 4096 #define MAX_PKT_BURST 64</pre>
Command Line	<pre>./build/examples/dpdk-l3fwd -c 0xff0000000000 -n 4 -a d8:00.0,txq_inline=200,txq_mpw_en=1 -a d8:00.1,txq_inline=200,txq_mpw_en=1 --socket-mem=0,8192 -- -p 0x3 -P -- config='(0,0,47),(0,1,46),(0,2,45),(0,3,44),(1,0,43),(1,1,42),(1,2,41),(1,3,40)' --eth- dest=0,00:52:11:22:33:10 --eth-dest=1,00:52:11:22:33:20</pre>
Other optimizations	<p>a) Flow Control OFF: "ethtool -A \$netdev rx off tx off"</p> <p>b) Memory optimizations: "sysctl -w vm.zone_reclaim_mode=0"; "sysctl -w vm.swappiness=0"</p> <p>c) Move all IRQs to far NUMA node: "IRQBALANCE_BANNED_CPUS=\$LOCAL_NUMA_CPUMAP irqbalance --oneshot"</p> <p>d) Disable irqbalance: "systemctl stop irqbalance"</p> <p>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"</p> <p>f) Set CQE COMPRESSION to "AGGRESSIVE": mlxconfig -d \$PORT_PCI_ADDRESS set CQE_COMPRESSION=1</p> <p>G) Disable Linux realtime throttling: echo -1 > /proc/sys/kernel/sched_rt_runtime_us</p>

3.2 Test Results

Table 5: Test #1 Results – Mellanox ConnectX-4 Lx 25GbE Dual-Port Throughput at Zero Packet Loss

Frame Size (Bytes)	Frame Rate (Mpps)	Line Rate [50G] (Mpps)	% Line Rate
64	70.80	74.4	95.17
128	39.22	42.23	92.88
256	22.64	22.64	100
512	11.75	11.75	100
1024	5.99	5.99	100
1280	4.81	4.81	100
1518	4.06	4.06	100

Figure 2: Test #1 Results – Mellanox ConnectX-4 Lx 25GbE Dual-Port Throughput at Zero Packet Loss



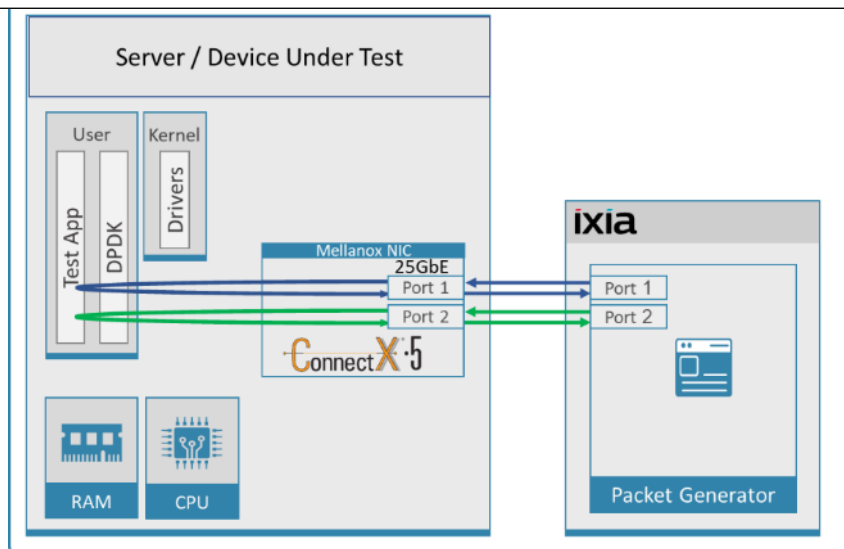
4 Test#2 Mellanox ConnectX-5 25GbE Throughput at Zero Packet Loss (2x 25GbE)

Table 6: Test #2 Setup

Item	Description
Test #2	Mellanox ConnectX-5 25GbE Dual-Port Throughput at zero packet loss
Server	HPE ProLiant DL380 Gen10
CPU	Intel(R) Xeon(R) Platinum 8168 CPU @ 2.70GHz 24 CPU cores * 2 NUMA nodes
RAM	384GB: 6 * 32GB DIMMs * 2 NUMA nodes @ 2666MHz
BIOS	U30 rev. 1.36 (02/15/2018)
NIC	One MCX512A-ACAT ConnectX-5 EN network interface card; 10/25GbE dual-port SFP28; PCIe3.0 x8; tall bracket; ROHS R6
Operating System	Ubuntu 20.04.2 LTS (Focal Fossa)
Kernel Version	5.4.0-65-generic.x86_64
GCC version	gcc (Ubuntu 9.3.0-17ubuntu1~20.04) 9.3.0
Mellanox NIC firmware version	16.29.2002
Mellanox OFED driver version	MLNX_OFED_LINUX-5.2-2.2.0.0
DPDK version	21.02
Test Configuration	1 NIC, 2 ports; Each port receives a stream of 8192 IP flows from the IXIA Each port has 4 queues assigned for a total of 8 queues 1 queue assigned per logical core with a total of 8 logical cores

The Device Under Test (DUT) is made up of the HPE server and the Mellanox ConnectX-5 Dual-Port NIC. The DUT is connected to the IXIA packet generator which generates traffic towards the ConnectX-5 NIC. The ConnectX-5 data traffic is passed through DPDK to the test application **l3fwd** and is redirected to the opposite direction on the same port. IXIA measures throughput and packet loss.

Figure 3: Test #2 Setup – Mellanox ConnectX-5 25GbE Dual-Port connected to IXIA



4.1 Test Settings

Table 7: Test #2 Settings

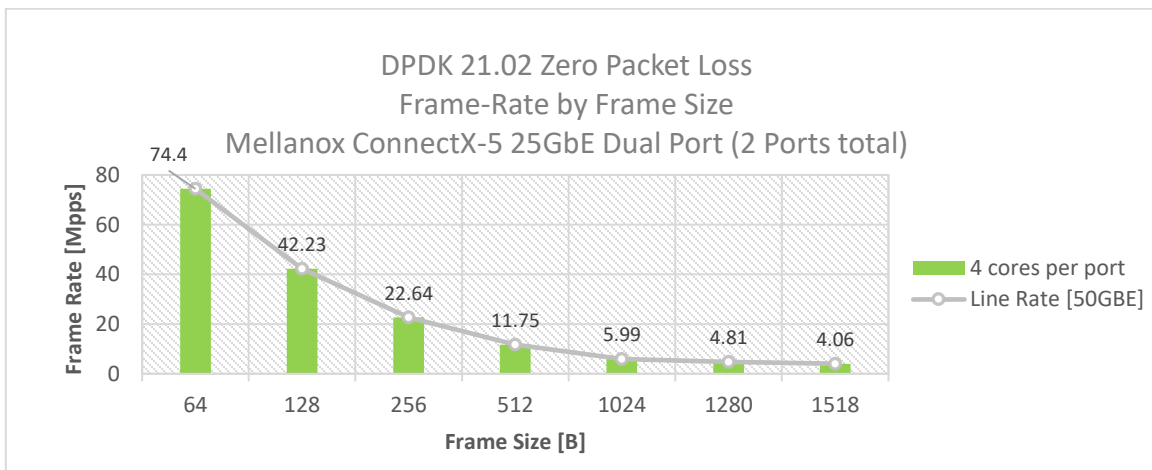
Item	Description
BIOS	<p>1) Workload Profile = "Low Latency";</p> <p>2) Jitter Control = Manual, 3400. (Setting turbo boost frequency to 3.4 GHz)</p> <p>See "Configuring and tuning HPE ProLiant Servers for low-latency applications": hpe.com > Search "DL380 gen10 low latency"</p>
BOOT Settings	<pre>isolcpus=24-47 intel_idle.max_cstate=0 processor.max_cstate=0 intel_pstate=disable nohz_full=24-47 rcu_nocbs=24-47 rcu_nocb_poll default_hugepagesz=1G hugepagesz=1G hugepages=64 audit=0 nosoftlockup</pre>
DPDK Settings	<p>Compile DPDK using: <pre>meson <build> -Dexamples=l3fwd ; ninja -C <build></pre></p> <p>During testing, l3fwd was given real-time scheduling priority.</p>
L3fwd settings	<p>Updated values /l3fwd/l3fwd.h:</p> <pre>#define RTE_TEST_RX_DESC_DEFAULT 4096 #define RTE_TEST_TX_DESC_DEFAULT 4096 #define MAX_PKT_BURST 64</pre>
Command Line	<pre>./build/examples/dpdk-l3fwd -c 0xff0000000000 -n 4 -a d8:00:0,mprq_en=1,rxqs_min_mprq=1 -a d8:00:1,mprq_en=1,rxqs_min_mprq=1 --socket-mem=0,8192 -- -p 0x3 -P --config='(0,0,47),(0,1,46),(0,2,45),(0,3,44),(1,0,43),(1,1,42),(1,2,41),(1,3,40)' --eth-dest=0,00:52:11:22:33:10 --eth-dest=1,00:52:11:22:33:20</pre>
Other optimizations	<p>a) Flow Control OFF: "ethtool -A \$netdev rx off tx off"</p> <p>b) Memory optimizations: "sysctl -w vm.zone_reclaim_mode=0"; "sysctl -w vm.swappiness=0"</p> <p>c) Move all IRQs to far NUMA node: "IRQBALANCE_BANNED_CPUS=\$LOCAL_NUMA_CPUMAP irqbalance --oneshot"</p> <p>d) Disable irqbalance: "systemctl stop irqbalance"</p> <p>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=3936"</p> <p>f) Set CQE COMPRESSION to "AGGRESSIVE": mlxconfig -d \$PORT_PCI_ADDRESS set CQE_COMPRESSION=1</p> <p>g) Disable Linux realtime throttling: echo -1 > /proc/sys/kernel/sched_rt_runtime_us</p>

4.2 Test Results

Table 8: Test #2 Results – Mellanox ConnectX-5 25GbE Dual-Port Throughput at Zero Packet Loss

Frame Size (Bytes)	Frame Rate (Mpps)	Line Rate [50G] (Mpps)	% Line Rate
64	74.40	74.40	100.00
128	42.23	42.23	100.00
256	22.64	22.64	100.00
512	11.75	11.75	100.00
1024	5.99	5.99	100.00
1280	4.81	4.81	100.00
1518	4.06	4.06	100.00

Figure 4: Test #2 Results – Mellanox ConnectX-5 25GbE Dual-Port Throughput at Zero Packet Loss



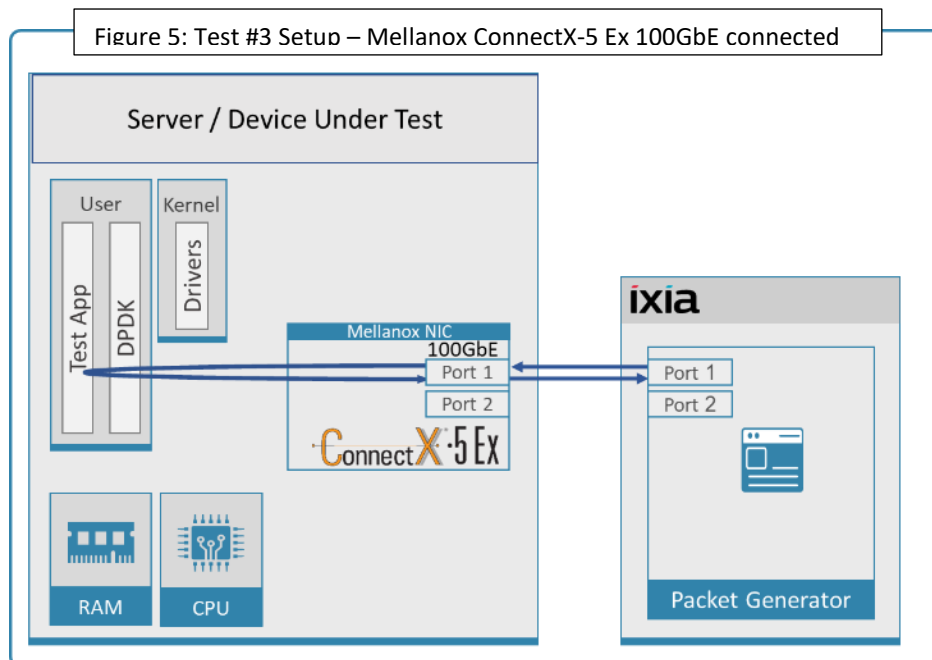
5 Test#3 Mellanox ConnectX-5 Ex 100GbE Throughput at Zero Packet Loss (1x 100GbE)

Table 9: Test #3 Setup

Item	Description
Test #3	Mellanox ConnectX-5 Ex 100GbE Throughput at zero packet loss
Server	HPE ProLiant DL380 Gen10
CPU	Intel(R) Xeon(R) Platinum 8168 CPU @ 2.70GHz 24 CPU cores * 2 NUMA nodes
RAM	384GB: 6 * 32GB DIMMs * 2 NUMA nodes @ 2666MHz
BIOS	U30 rev. 1.36 (02/15/2018)
NIC	One MCX516A-CDAT ConnectX-5 Ex network interface card 100GbE dual-port QSFP28; PCIe3.0/PCIe4 x16; ROHS R6
Operating System	Ubuntu 20.04.2 LTS (Focal Fossa)
Kernel Version	5.4.0-65-generic.x86_64
GCC version	gcc (Ubuntu 9.3.0-17ubuntu1~20.04) 9.3.0
Mellanox NIC firmware version	16.29.2002
Mellanox OFED driver version	MLNX_OFED_LINUX-5.2-2.2.0.0
DPDK version	21.02
Test Configuration	1 NIC, 1 port used on NIC; Port has 12 queues assigned to it, 1 queue per logical core for a total of 12 logical cores. Each port receives a stream of 8192 IP flows from the IXIA

The Device Under Test (DUT) is made up of the HPE server and the Mellanox ConnectX-5 Ex Dual-Port NIC (only the first port is used in this test). The DUT is connected to the IXIA packet generator which generates traffic towards the ConnectX-5 Ex NIC.

The ConnectX-5 Ex data traffic is passed through DPDK to the test application **l3fwd** and is redirected to the opposite direction on the same port. IXIA measures throughput and packet loss.



5.1 Test Settings

Table 10: Test #3 Settings

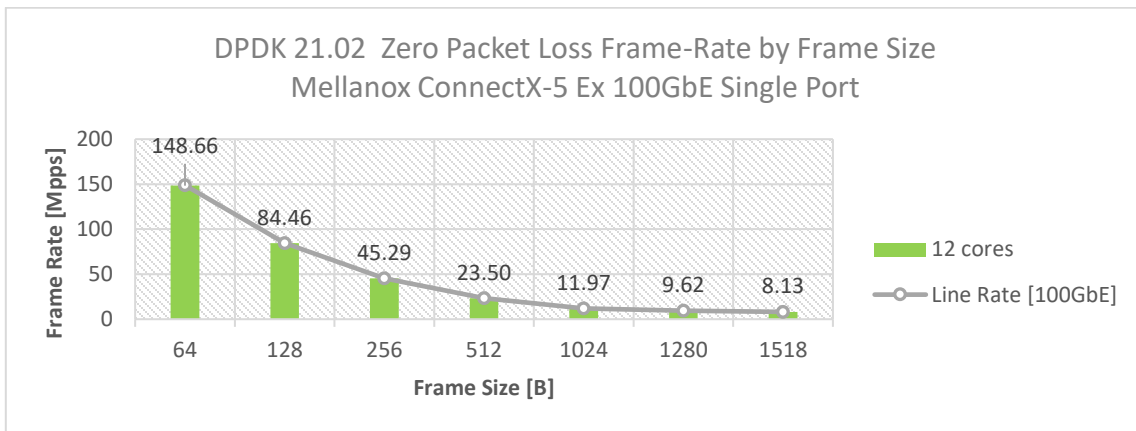
Item	Description
BIOS	<p>1) Workload Profile = "Low Latency";</p> <p>2) Jitter Control = Manual, 3400. (Setting turbo boost frequency to 3.4 GHz)</p> <p>See "Configuring and tuning HPE ProLiant Servers for low-latency applications": hpe.com > Search "DL380 gen10 low latency"</p>
BOOT Settings	<pre>isolcpus=24-47 intel_idle.max_cstate=0 processor.max_cstate=0 intel_pstate=disable nohz_full=24-47 rcu_nocbs=24-47 rcu_nocb_poll default_hugepagesz=1G hugepagesz=1G hugepages=64 audit=0 nosoflockup</pre>
DPDK Settings	<p>Compile DPDK using:</p> <pre>meson <build> -Dexamples=l3fwd ; ninja -C <build></pre> <p>During testing, l3fwd was given real-time scheduling priority.</p>
L3fwd settings	<p>Updated values /l3fwd/l3fwd.h:</p> <pre>#define RTE_TEST_RX_DESC_DEFAULT 4096 #define RTE_TEST_TX_DESC_DEFAULT 4096 #define MAX_PKT_BURST 64</pre>
Command Line	<pre>./build/examples/dpdk-l3fwd -c 0xffff00000000 -n 4 -a 0000:af:00.0,mprq_en=1,rxqs_min_mprq=1 --socket-mem=0,8192 -- -p 0x1 -P -- config='(0,0,47),(0,1,46),(0,2,45),(0,3,44),(0,4,43),(0,5,42),(0,6,41),(0,7,40),(0,8,39),(0,9,38),(0,10,37),(0,11,36)' --eth-dest=0,00:52:11:22:33:10</pre>
Other optimizations	<p>a) Flow Control OFF: "ethtool -A \$netdev rx off tx off"</p> <p>b) Memory optimizations: "sysctl -w vm.zone_reclaim_mode=0"; "sysctl -w vm.swappiness=0"</p> <p>c) Move all IRQs to far NUMA node: "IRQBALANCE_BANNED_CPUS=\$LOCAL_NUMA_CPUMAP irqbalance --oneshot"</p> <p>d) Disable irqbalance: "systemctl stop irqbalance"</p> <p>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"</p> <p>f) Set CQE COMPRESSION to "AGGRESSIVE": mlxconfig -d \$PORT_PCI_ADDRESS set CQE_COMPRESSION=1</p> <p>g) Disable Linux realtime throttling: echo -1 > /proc/sys/kernel/sched_rt_runtime_us</p>

5.2 Test Results

Table 11: Test #3 Results – Mellanox ConnectX-5 Ex 100GbE Throughput at Zero Packet Loss

Frame Size (Bytes)	Frame Rate (Mpps)	Line Rate [100G] (Mpps)	% Line Rate
64	148.66	148.81	99.91
128	84.46	84.46	100.00
256	45.29	45.29	100.00
512	23.50	23.50	100.00
1024	11.97	11.97	100.00
1280	9.62	9.62	100.00
1518	8.13	8.13	100.00

Figure 6: Test #3 Results – Mellanox ConnectX-5 Ex 100GbE Throughput at Zero Packet Loss



6 Test#4 Mellanox ConnectX-5 Ex 100GbE Single Core Performance (2x 100GbE)

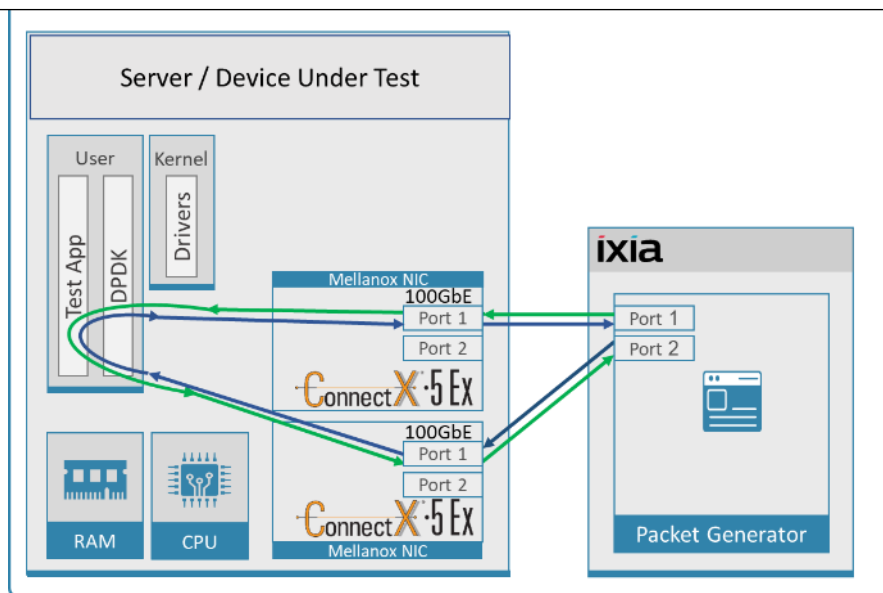
Table 12: Test #4 Setup

Item	Description
Test #4	Mellanox ConnectX-5 Ex 100GbE Single Core Performance
Server	HPE ProLiant DL380 Gen10
CPU	Intel(R) Xeon(R) Platinum 8168 CPU @ 2.70GHz 24 CPU cores * 2 NUMA nodes
RAM	384GB: 6 * 32GB DIMMs * 2 NUMA nodes @ 2666MHz
BIOS	U30 rev. 1.36 (02/15/2018)
NIC	Two MCX516A-CDAT- ConnectX-5 Ex network interface cards 100GbE dual-port QSFP28; PCIe3.0/PCIe4 x16; ROHS R6
Operating System	Ubuntu 20.04.2 LTS (Focal Fossa)
Kernel Version	5.4.0-65-generic.x86_64
GCC version	gcc (Ubuntu 9.3.0-17ubuntu1~20.04) 9.3.0
Mellanox NIC firmware version	16.29.2002
Mellanox OFED driver version	MLNX_OFED_LINUX-5.2-2.2.0.0
DPDK version	21.02
Test Configuration	2 NICs, each using 1 port Each port receives a stream of 8192 UDP flows from the IXIA 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.

The Device Under Test (DUT) is made up of the HPE server and two Mellanox ConnectX-5 Ex NICs utilizing one port each. The DUT is connected to the IXIA packet generator which generates traffic towards the first port of both ConnectX-5 Ex NICs.

The ConnectX-5 Ex data traffic is passed through DPDK to the test application **testpmd** and is redirected to the opposite direction on the opposing NIC's port. IXIA measures throughput and packet loss.

Figure 7: Test #4 Setup – Two Mellanox ConnectX-5 Ex 100GbE connected to IXIA



6.1 Test Settings

Table 13: Test #4 Settings

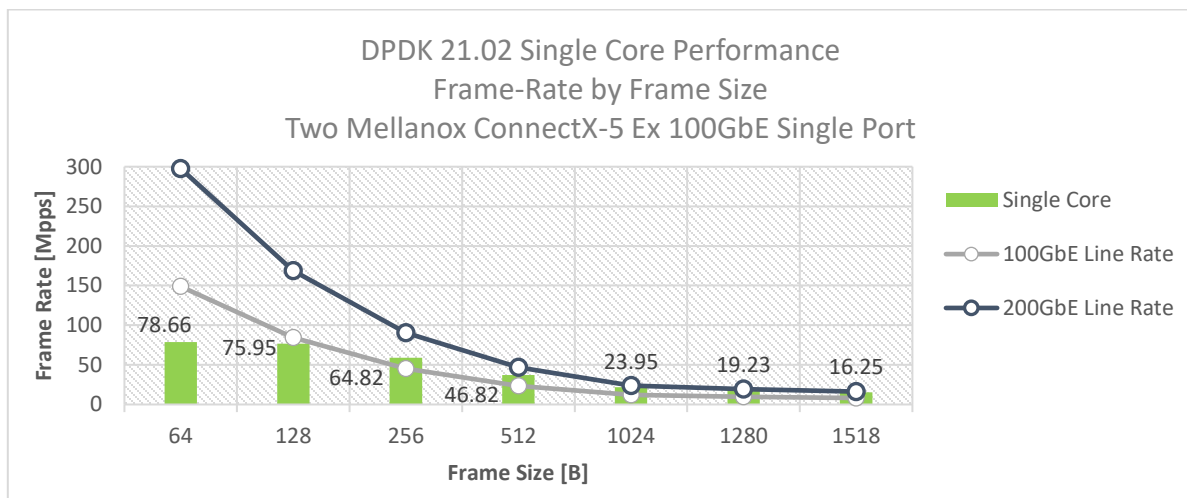
Item	Description
BIOS	<p>1) Workload Profile = "Low Latency";</p> <p>2) Jitter Control = Manual, 3400. (Setting turbo boost frequency to 3.4 GHz)</p> <p>See "Configuring and tuning HPE ProLiant Servers for low-latency applications": hpe.com > Search "DL380 gen10 low latency"</p>
BOOT Settings	<pre>isolcpus=24-47 intel_idle.max_cstate=0 processor.max_cstate=0 intel_pstate=disable nohz_full=24-47 rcu_nocbs=24-47 rcu_nocb_poll default_hugepagesz=1G hugepagesz=1G hugepages=64 audit=0</pre>
DPDK Settings	<p>Compile DPDK using: <pre>meson <build> ; ninja -C <build></pre></p> <p>During testing, testpmd was given real-time scheduling priority.</p>
Command Line	<pre>./build/app/dpdk-testpmd -c 0x110000000000 -n 4 -a 86:00.0 -a af:00.0 --socket-mem=0,8192 --port-numa-config=0,1,1,1 --socket-num=1 --burst=64 --txd=1024 --rxd=1024 --mbcache=512 --rxq=1 --txq=1 --nb-cores=1 -i -a --rss-udp --disable-crc-strip --record-core-cycles --record-burst-stats</pre>
Other optimizations	<p>a) Flow Control OFF: "ethtool -A \$netdev rx off tx off"</p> <p>b) Memory optimizations: "sysctl -w vm.zone_reclaim_mode=0"; "sysctl -w vm.swappiness=0"</p> <p>c) Move all IRQs to far NUMA node: "IRQBALANCE_BANNED_CPUS=\$LOCAL_NUMA_CPUMAP irqbalance --oneshot"</p> <p>d) Disable irqbalance: "systemctl stop irqbalance"</p> <p>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"</p> <p>f) Set CQE COMPRESSION to "AGGRESSIVE": mlxconfig -d \$PORT_PCI_ADDRESS set CQE_COMPRESSION=1</p> <p>g) Disable Linux realtime throttling: echo -1 > /proc/sys/kernel/sched_rt_runtime_us</p>

6.2 Test Results

Table 14: Test #4 Results – Mellanox ConnectX-5 Ex 100GbE Single Core Performance

Frame Size (Bytes)	Frame Rate (Mpps)	Line Rate [200G] (Mpps)	Line Rate [100G] (Mpps)	Throughput (Gbps)	CPU Cycles per packet <small>NOTE: Lower is Better</small>
64	78.67	297.62	148.81	40.277	31
128	75.95	168.92	84.46	77.768	32
256	64.82	90.58	45.29	132.755	31
512	46.84	46.99	23.50	191.845	32
1024	23.95	23.95	11.97	196.164	29
1280	19.23	19.23	9.62	196.880	32
1518	16.25	16.25	8.13	197.373	30

Figure 8: Test #4 Results – Mellanox ConnectX-5 Ex 100GbE Single Core Performance



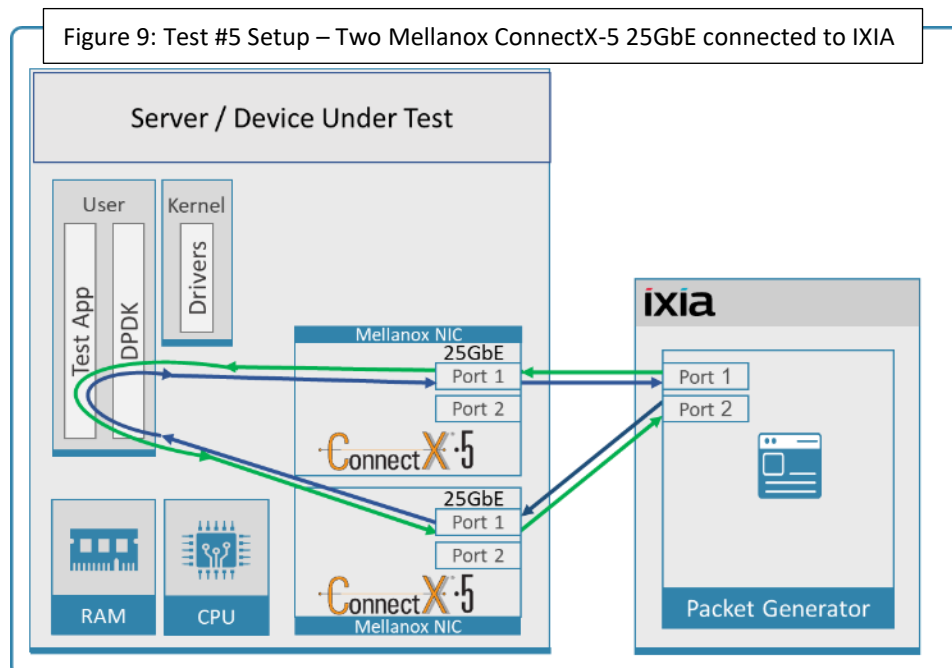
7 Test#5 Mellanox ConnectX-5 25GbE Single Core Performance (2x 25GbE)

Table 15: Test #5 Setup

Item	Description
Test #5	Mellanox ConnectX-5 25GbE Single Core Performance
Server	HPE ProLiant DL380 Gen10
CPU	Intel(R) Xeon(R) Platinum 8168 CPU @ 2.70GHz 24 CPU cores * 2 NUMA nodes
RAM	384GB: 6 * 32GB DIMMs * 2 NUMA nodes @ 2666MHz
BIOS	U30 rev. 1.36 (02/15/2018)
NIC	Two MCX512A-ACA ConnectX-5 EN network interface cards; 10/25GbE dual-port SFP28; PCIe3.0 x8; tall bracket; ROHS R6
Operating System	Ubuntu 20.04.2 LTS (Focal Fossa)
Kernel Version	5.4.0-65-generic.x86_64
GCC version	gcc (Ubuntu 9.3.0-17ubuntu1~20.04) 9.3.0
Mellanox NIC firmware version	16.29.2002
Mellanox OFED driver version	MLNX_OFED_LINUX-5.2-2.2.0.0
DPDK version	21.02
Test Configuration	2 NICs; 1 port used on each. Each port receives a stream of 8192 UDP flows from the IXIA 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.

The Device Under Test (DUT) is made up of the HPE server and two Mellanox ConnectX-5 25GbE NICs utilizing one port each. The DUT is connected to the IXIA packet generator which generates traffic towards the first port of both ConnectX-5 25GbE NICs.

The ConnectX-5 25GbE data traffic is passed through DPDK to the test application **testpmd** and is redirected to the opposite direction on the opposing NIC's port. IXIA measures throughput and packet loss.



7.1 Test Settings

Table 16: Test #5 Settings

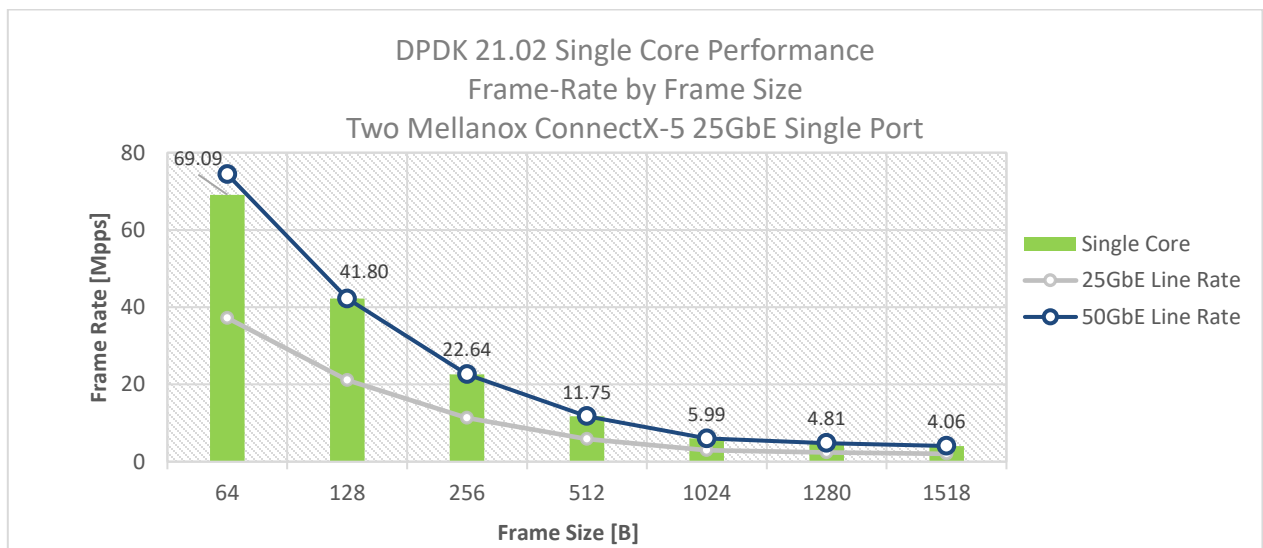
Item	Description
BIOS	<p>1) Workload Profile = "Low Latency"</p> <p>2) Jitter Control = Manual, 3400. (Setting turbo boost frequency to 3.4 GHz)</p> <p>See "Configuring and tuning HPE ProLiant Servers for low-latency applications": hpe.com > Search "DL380 gen10 low latency"</p>
BOOT Settings	<pre>isolcpus=24-47 intel_idle.max_cstate=0 processor.max_cstate=0 intel_pstate=disable nohz_full=24-47 rcu_nocbs=24-47 rcu_nocb_poll default_hugepagesz=1G hugepagesz=1G hugepages=64 audit=0 nosoftlockup</pre>
DPDK Settings	<p>Compile DPDK using: <pre>meson <build> ; ninja -C <build></pre></p> <p>During testing, testpmd was given real-time scheduling priority.</p>
Command Line	<pre>./build/app/dpdk-testpmd -c 0x300000000000 -n 4 -a d8:00.0 -a d9:00.0 --socket-mem=0,8192 -- port-numa-config=0,1,1,1 --socket-num=1 --burst=64 --txd=1024 --rxd=1024 --mbcache=512 -- rxq=1 --txq=1 --nb-cores=1 -i -a --rss-udp --disable-crc-strip --record-core-cycles --record-burst- stats</pre>
Other optimizations	<p>a) Flow Control OFF: "ethtool -A \$netdev rx off tx off"</p> <p>b) Memory optimizations: "sysctl -w vm.zone_reclaim_mode=0"; "sysctl -w vm.swappiness=0"</p> <p>c) Move all IRQs to far NUMA node: "IRQBALANCE_BANNED_CPUS=\$LOCAL_NUMA_CPUMAP irqlbalance --oneshot"</p> <p>d) Disable irqbalance: "systemctl stop irqbalance"</p> <p>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"</p> <p>f) Set CQE COMPRESSION to "AGGRESSIVE": mlxconfig -d \$PORT_PCI_ADDRESS set CQE_COMPRESSION=1</p> <p>g) Disable Linux realtime throttling: echo -1 > /proc/sys/kernel/sched_rt_runtime_us</p>

7.2 Test Results

Table 17: Test #5 Results – Mellanox ConnectX-5 25GbE Single Core Performance

Frame Size (Bytes)	Frame Rate (Mpps)	Line Rate [25G] (Mpps)	Line Rate [50G] (Mpps)	Throughput (Gbps)	CPU Cycles per packet <small>NOTE: Lower is Better</small>
64	69.24	37.2	74.4	35.45	26
128	41.82	21.11	42.23	42.82	26
256	22.64	11.32	22.64	46.368	26
512	11.75	5.87	11.75	48.12	27
1024	5.99	2.99	5.99	49.037	33
1280	4.81	2.4	4.81	49.226	28
1518	4.06	2.03	4.06	49.344	31

Figure 10: Test #5 Results – Mellanox ConnectX-5 25GbE Single Core Performance



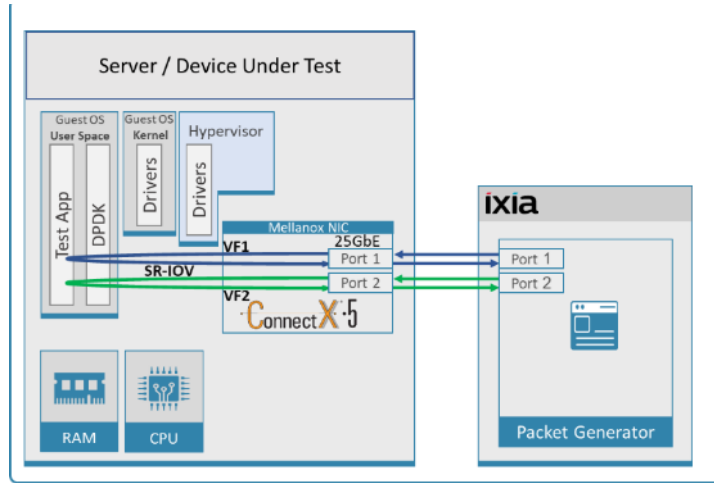
8 Test#6 Mellanox ConnectX-5 25GbE Throughput at Zero Packet Loss (2x 25GbE) using SR-IOV over VMware ESXi 7.0U1

Table 18: Test #6 Setup

Item	Description
Test #6	Mellanox ConnectX-5 25GbE Dual-Port Throughput at zero packet loss SRIOV over VMware ESXi 6.7U3
Server	HPE ProLiant DL380 Gen10
CPU	Intel(R) Xeon(R) Platinum 8168 CPU @ 2.70GHz 24 CPU cores * 2 NUMA nodes
RAM	384GB: 6 * 32GB DIMMs * 2 NUMA nodes @ 2666MHz
BIOS	U30 rev. 1.36 (02/15/2018)
NIC	One MCX512A-ACAT ConnectX-5 EN network interface card; 10/25GbE dual-port SFP28; PCIe3.0 x8; tall bracket; ROHS R6
Hypervisor	VMware ESXi 7.0U1
Hypervisor Build	VMware-VMvisor-Installer-7.0U1-16850804.x86_64.iso
Hypervisor Mellanox Driver	4.19.71.2-10EM.700.1.0.15843807
Guest Operating System	Red Hat Enterprise Linux Server release 7.7 (Maipo)
Guest Kernel Version	3.10.0-1062.el7.x86_64
Guest GCC version	4.8.5 20150623 (Red Hat 4.8.5-28) (GCC)
Guest Mellanox OFED driver version	MLNX_OFED_LINUX-5.2-2.2.0.0
Mellanox NIC firmware version	16.29.2002
DPDK version	21.02
Test Configuration	1 NIC, 2 ports with 1 VF per port (SR-IOV); Each port receives a stream of 8192 IP flows from the IXIA Each VF (SR-IOV) has 4 queues assigned for a total of 8 queues 1 queue assigned per logical core with a total of 8 logical cores.

The Device Under Test (DUT) is made up of the HPE server and the Mellanox ConnectX-5 NIC with dual-port. The DUT is connected to the IXIA packet generator which generates traffic towards the ConnectX-5 NIC. The ConnectX-5 data traffic is passed to VF1 (SR-IOV assigned to Port1) and VF2 (SR-IOV assigned to Port2) to VM running over ESXi 6.5 hypervisor. VM runs **l3fwd** over DPDK and is redirects traffic to the opposite direction on the same VF/port. IXIA measures throughput and packet loss.

Figure 11: Test #6 Setup – Mellanox ConnectX-5 25GbE connected to IXIA using ESXi SR-IOV



8.1 Test Settings

Table 19: Test#6 Settings

Item	Description
BIOS	<p>1) Workload Profile = "Low Latency";</p> <p>2) Jitter Control = Manual, 3400. (Setting turbo boost frequency to 3.4 GHz)</p> <p>3) Change "Workload Profile" to "Custom"</p> <p>4) Change VT-x, VT-d and SR-IOV from "Disabled" to "Enabled".</p> <p>See "Configuring and tuning HPE ProLiant Servers for low-latency applications": hpe.com > Search "DL380 gen10 low latency"</p>
BOOT Settings Guest OS	<pre>isolcpus=0-22 intel_idle.max_cstate=0 processor.max_cstate=0 intel_pstate=disable idle=poll nohz_full=0-22 rcu_nocbs=0-22 rcu_nocb_poll default_hugepagesz=1G hugepagesz=1G hugepages=16 nosoftlockup</pre>
Hypervisor settings	<p>1) Enable SRIOV via NIC configuration tool: (requires installation of mft-tools)</p> <pre>/opt/mellanox/bin/mlxconfig -d <PCI ID> set NUM_OF_VFS=2 SRIOV_EN=1 CQE_COMPRESSION=1</pre> <p>reboot</p> <p>2) Install Driver</p> <pre>esxcli software vib install -MLNX-NATIVE-ESX-ConnectX-4-5_4.17.70.1- 1OEM.670.0.0.8169922.zip</pre> <p>reboot</p> <pre>esxcfg-module -s 'max_vfs=1,1,1,1,1,1,1,1 supported_num_ports=8' nmlx5_core</pre> <p>reboot</p> <p>3) Virtual Hardware Configuration:</p> <p>"CPU": 23</p> <p>"Cores per Socket" : 1 (Sockets = 23) or 23 (Socket = 1)</p> <p>"Hardware virtualization": enabled</p> <p>"Scheduling Affinity": 25-47</p> <p>"CPU/MMU Virtualization": "Hardware CPU and MMU"</p> <p>"RAM": 32768 MB</p> <p>"Reservation": 32768 MB</p> <p>"Reserve all guest memory (All locked)": enabled</p> <p>VM options > Advanced > "Configuration Parameters" > "Edit Configuration" : Add parameter: numa.nodeAffinity = 1</p> <p>4) Create virtual switch:</p> <pre>Networking>Virtual Switches>Add standard virtual switch>Switch_SRIOV_1>Uplink : select vmnicXXXX matching the card under test</pre> <p>5) Add port group to Switch_SRIOV_XX (VLAN=0):</p> <pre>Networking>Port groups>Add port group>SRIOV_PG1>Switch_SRIOV_XX</pre> <p>6) Add 2xSRIOV network adapters to VM (same settings for both ports):</p> <p>Select correct port group created previously (SRIOV_PG1)</p> <p>Adapter Type: SR-IOV passthrough</p> <p>Physical function: select pci for the portX of the card under the test</p>
DPDK Settings on Guest OS	<p>Compile DPDK using:</p> <pre>meson <build> -Dexamples=l3fwd ; ninja -C <build></pre> <p>During testing, l3fwd was given real-time scheduling priority.</p>
L3fwd settings on Guest OS	<p>Updated values /l3fwd/l3fwd.h:</p>

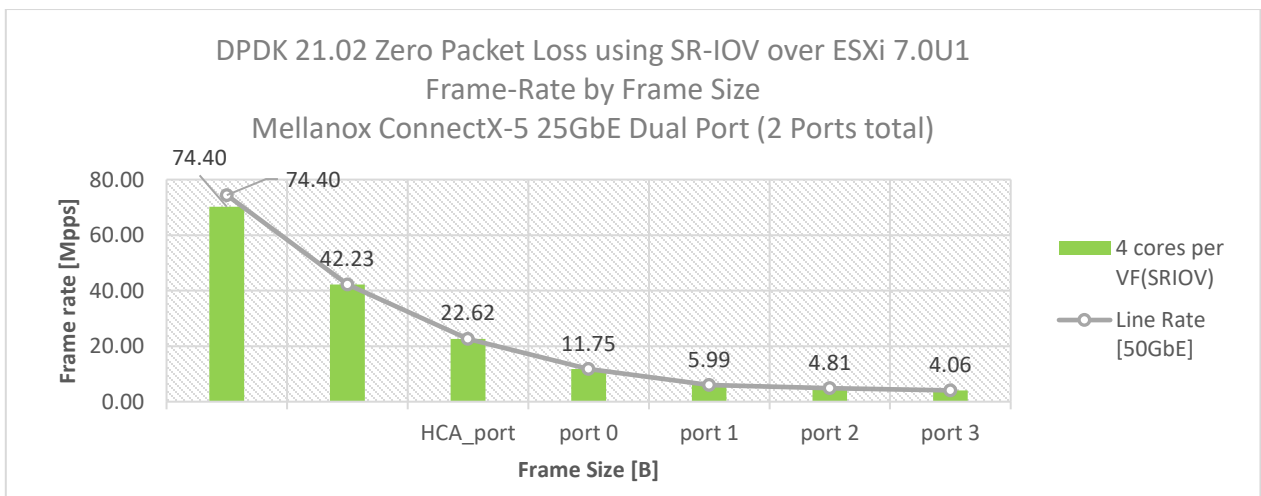
Item	Description
	<pre>#define RTE_TEST_RX_DESC_DEFAULT 2048 #define RTE_TEST_TX_DESC_DEFAULT 2048 #define MAX_PKT_BURST 64</pre>
Command Line on Guest OS	<pre>./build/examples/dpdk-l3fwd -c 0x7f8000 -n 4 -a 13:00.0,mprq_en=1,rxqs_min_mprq=1 -a 1b:00.0,mprq_en=1,rxqs_min_mprq=1 --socket-mem=8192 -- -p 0x3 -P --config='(0,0,22),(0,1,21),(0,2,20),(0,3,19),(1,0,18),(1,1,17),(1,2,16),(1,3,15)' --eth-dest=0,00:52:11:22:33:10 --eth-dest=1,00:52:11:22:33:20</pre>
Other optimizations on Guest OS	<p>a) Flow Control OFF: "ethtool -A \$netdev rx off tx off"</p> <p>b) Memory optimizations: "sysctl -w vm.zone_reclaim_mode=0"; "sysctl -w vm.swappiness=0"</p> <p>c) Move all IRQs to far NUMA node: "IRQBALANCE_BANNED_CPUS=\$LOCAL_NUMA_CPUMAP irqbalance --oneshot"</p> <p>d) Disable irqbalance: "systemctl stop irqbalance"</p> <p>e) Disable Linux realtime throttling: echo -1 > /proc/sys/kernel/sched_rt_runtime_us</p>

8.2 Test Results

Table 20: Test#6 Results – Mellanox ConnectX-5 25GbE Throughput at Zero Packet Loss using ESXi SR-IOV

Frame Size (Bytes)	Frame Rate (Mpps)	Line Rate [50G] (Mpps)	% Line Rate
64	74.4	74.4	100.00
128	42.23	42.23	100.00
256	22.62	22.64	100.00
512	11.75	11.75	100.00
1024	5.99	5.99	100.00
1280	4.81	4.81	100.00
1518	4.06	4.06	100.00

Figure 12: Test#6 Results – Mellanox ConnectX-5 25GbE Throughput at Zero Packet Loss using ESXi SR-IOV



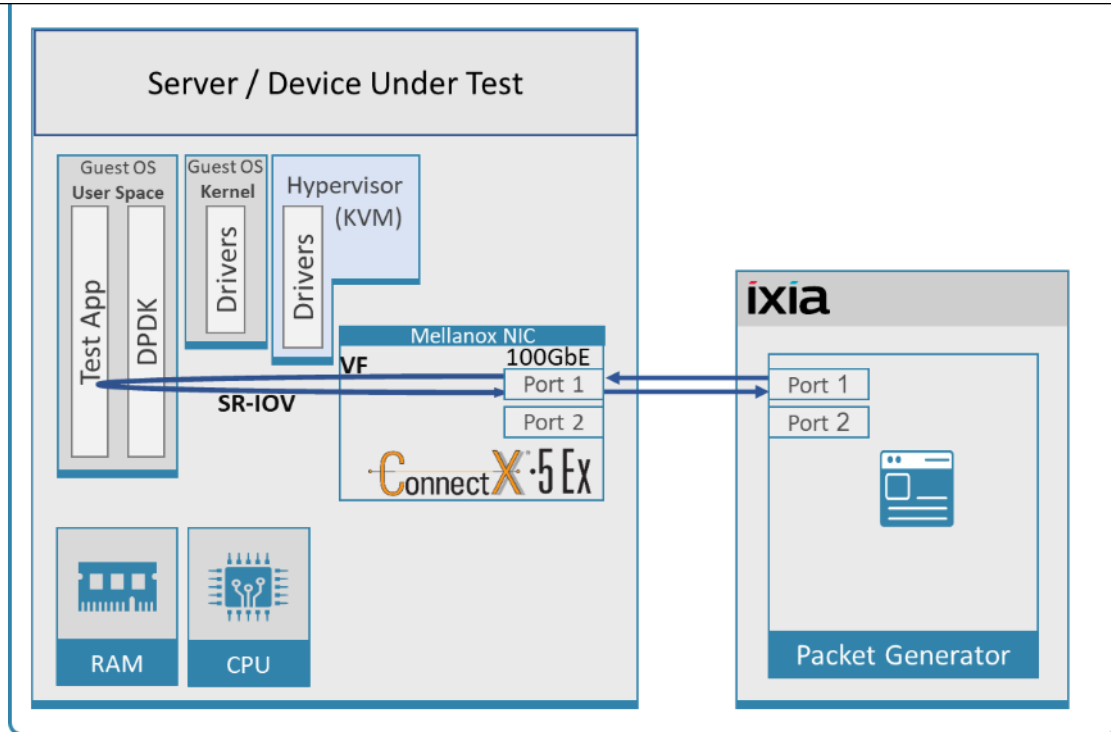
9 Test#7 Mellanox ConnectX-5 Ex 100GbE Throughput at Zero Packet Loss (1x 100GbE) using SR-IOV over KVM Hypervisor

Table 21: Test #7 Setup

Item	Description
Test #7	Mellanox ConnectX-5 Ex 100GbE Throughput at zero packet loss using SR-IOV over KVM Hypervisor
Server	HPE ProLiant DL380 Gen10
CPU	Intel(R) Xeon(R) Platinum 8168 CPU @ 2.70GHz 24 CPU cores * 2 NUMA nodes
RAM	384GB: 6 * 32GB DIMMs * 2 NUMA nodes @ 2666MHz
BIOS	U30 rev. 1.36 (02/15/2018)
NIC	One MCX516A-CDAT ConnectX-5 Ex network interface card 100GbE dual-port QSFP28; PCIe3.0/PCIe4 x16; ROHS R6
Hypervisor	Ubuntu 20.04.2 LTS (Focal Fossa) QEMU emulator version 4.2.1 (Debian 1:4.2-3ubuntu6.11)
Hypervisor Kernel Version	5.4.0-65-generic.x86_64
Hypervisor Mellanox Driver	MLNX_OFED_LINUX-5.2-2.2.0.0
Guest Operating System	Red Hat Enterprise Linux Server release 7.7 (Maipo)
Guest Kernel Version	3.10.0-1062.el7.x86_64
Guest GCC version	4.8.5 20150623 (Red Hat 4.8.5-28) (GCC)
Guest Mellanox OFED driver version	MLNX_OFED_LINUX-5.2-2.2.0.0
Mellanox NIC firmware version	16.29.2002
DPDK version	20.11
Test Configuration	1 NIC, 1 port over 1 VF (SR-IOV); VF has 12 queues assigned to it, 1 queue per logical core for a total of 12 logical cores. Each physical port receives a stream of 8192 IP flows from the IXIA directed to VF assigned to Guest OS.

The Device Under Test (DUT) is made up of the HPE server and the Mellanox ConnectX-5 Ex NIC with a dual-port (only first port used in this test) running Red Hat Enterprise Linux Server with qemu-KVM managed via libvirt, Guest OS running DPDK is based on Red Hat Enterprise Linux Server as well. The DUT is connected to the IXIA packet generator which generates traffic towards the ConnectX-5 Ex NIC. The ConnectX-5 Ex data traffic is passed through a virtual function (VF/SR-IOV) to DPDK running on the Guest OS, to the test application **l3fwd** and is redirected to the opposite direction on the same port. IXIA measures throughput and packet loss.

Figure 13: Test #7 Setup – Mellanox ConnectX-5 Ex 100GbE connected to IXIA using KVM SR-IOV



9.1 Test Settings

Table 22: Test #7 Settings

Item	Description
BIOS	<ol style="list-style-type: none"> 1) Workload Profile = "Low Latency"; 2) Jitter Control = Manual, 3400. (Setting turbo boost frequency to 3.4 GHz) 3) Change "Workload Profile" to "Custom" 4) Change VT-x, VT-d and SR-IOV from "Disabled" to "Enabled". <p>See "Configuring and tuning HPE ProLiant Servers for low-latency applications": hpe.com > Search "DL380 gen10 low latency"</p>
Hypervisor BOOT Settings	<pre>isolcpus=24-47 intel_idle.max_cstate=0 processor.max_cstate=0 nohz_full=24-47 rcu_nocbs=24-47 intel_pstate=disable default_hugepagesz=1G hugepagesz=1G hugepages=70 audit=0 nosoftlockup intel_iommu=on iommu=pt rcu_nocb_poll</pre>
Hypervisor settings	<ol style="list-style-type: none"> 1) Enable SRIOV via NIC configuration tool: (requires installation of mft-tools) <pre>mlxconfig -d /dev/mst/mt4121_pciconf1 set NUM_OF_VFS=1 SRIOV_EN=1 CQE_COMPRESSION=1 echo 1 > /sys/class/net/ens6f0/device/sriov_numvfs</pre> 2) Assign VF <pre>HCA_netintf=ens6f0 #assign a VF to the DUT device VF_PCI_address="0000:af:00.2" #VF PCI address echo \$VF_PCI_address > /sys/bus/pci/drivers/mlx5_core/unbind modprobe vfio-pci echo "\$(cat /sys/bus/pci/devices/\$VF_PCI_address/vendor) \$(cat /sys/bus/pci/devices/\$VF_PCI_address/device)" > /sys/bus/pci/drivers/vfio-pci/new_id</pre>

Item	Description
	<p># Now the VF may be assigned to Guest (passthrough) with libvirt virt-manager.</p> <p>3) Setting VF MAC - use the command below (find out the vf-index from "ip link show"), ip link set <<PF NIC interface>> <vf index> mac <MAC Address> : (mac is random) ip link set \$HCA_netintf vf 0 mac 00:52:11:22:33:42</p> <p>4) VM tuning: vcpupin and memory backing from hugepages: To persistently configure vcpu pinning and memory backing, add the below config to the VM's XML config before starting the VM. Add the following two elements to the XML: <cputune> and <memoryBacking> and also increase the number of cpus and memory: virsh edit <vmID> (to get vmID use - virsh list --all)</p> <p>Example xml configuration: (change "nodeset" and "cpuset" attributes to suit the local NUMA node in your setup)</p> <pre> <domain type='kvm' id='1'> <name>perf-dpdk-01-005-RH-7.4</name> <uuid>06f283fc-fd76-4411-8b6a-72fe94f50376</uuid> <memory unit='KiB'>33554432</memory> <currentMemory unit='KiB'>33554432</currentMemory> <memoryBacking> <hugepages> <page size='1048576' unit='KiB' nodeset='0'/> </hugepages> <nosharepages/> <locked/> </memoryBacking> <vcpu placement='static'>23</vcpu> <cputune> <vcpupin vcpu='0' cpuset='24'/> <vcpupin vcpu='1' cpuset='25'/> <vcpupin vcpu='2' cpuset='26'/> <vcpupin vcpu='3' cpuset='27'/> <vcpupin vcpu='4' cpuset='28'/> <vcpupin vcpu='5' cpuset='29'/> <vcpupin vcpu='6' cpuset='30'/> <vcpupin vcpu='7' cpuset='31'/> <vcpupin vcpu='8' cpuset='32'/> <vcpupin vcpu='9' cpuset='33'/> <vcpupin vcpu='10' cpuset='34'/> <vcpupin vcpu='11' cpuset='35'/> <vcpupin vcpu='12' cpuset='36'/> <vcpupin vcpu='13' cpuset='37'/> <vcpupin vcpu='14' cpuset='38'/> <vcpupin vcpu='15' cpuset='39'/> <vcpupin vcpu='16' cpuset='40'/> <vcpupin vcpu='17' cpuset='41'/> <vcpupin vcpu='18' cpuset='42'/> </pre>

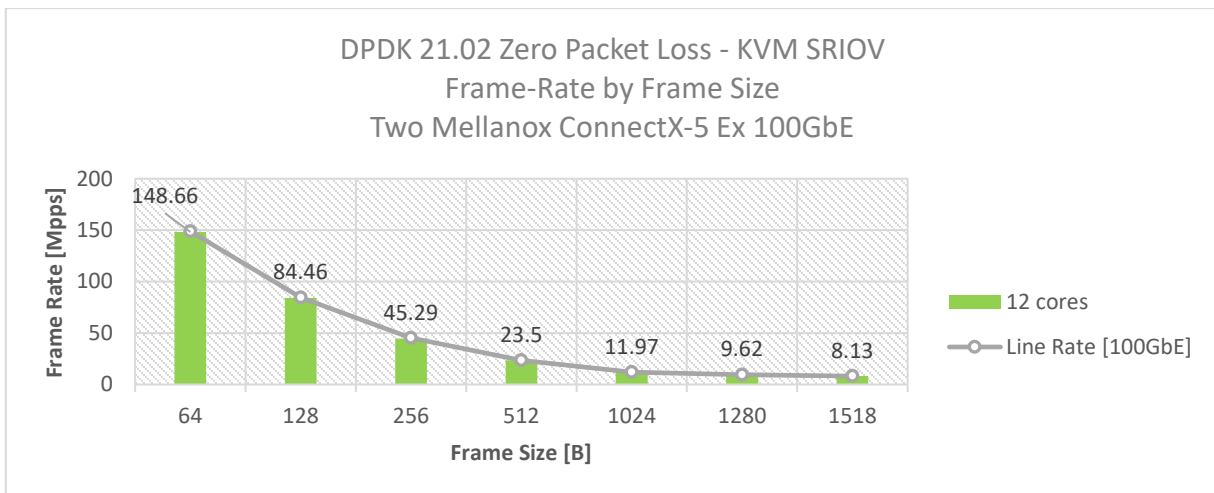
Item	Description
	<pre><vcupin vcpu='19' cpuset='43'/> <vcupin vcpu='20' cpuset='44'/> <vcupin vcpu='21' cpuset='45'/> <vcupin vcpu='22' cpuset='46'/> </cputune></pre>
Other optimizations on Hypervisor	<p>a) Flow Control OFF: "ethtool -A \$netdev rx off tx off"</p> <p>b) Memory optimizations: "sysctl -w vm.zone_reclaim_mode=0"; "sysctl -w vm.swappiness=0"</p> <p>c) Move all IRQs to far NUMA node: "IRQBALANCE_BANNED_CPUS=\$LOCAL_NUMA_CPUMAP irqbalance --oneshot"</p> <p>d) Disable irqbalance: "systemctl stop irqbalance"</p> <p>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"</p> <p>f) Disable Linux realtime throttling: echo -1 > /proc/sys/kernel/sched_rt_runtime_us</p>
Guest BOOT Settings	<pre>isolcpus=0-22 intel_idle.max_cstate=0 processor.max_cstate=0 intel_pstate=disable idle=poll nohz_full=0-22 rcu_nocbs=0-22 rcu_nocb_poll default_hugepagesz=1G hugepagesz=1G hugepages=16 nosoftlockup</pre>
Other optimizations on Guest OS	<p>a) Flow Control OFF: "ethtool -A \$netdev rx off tx off"</p> <p>b) Memory optimizations: "sysctl -w vm.zone_reclaim_mode=0"; "sysctl -w vm.swappiness=0"</p> <p>c) Move all IRQs to far NUMA node: "IRQBALANCE_BANNED_CPUS=\$LOCAL_NUMA_CPUMAP irqbalance --oneshot"</p> <p>d) Disable irqbalance: "systemctl stop irqbalance"</p> <p>e) Disable Linux realtime throttling: echo -1 > /proc/sys/kernel/sched_rt_runtime_us</p>
DPDK Settings on Guest OS	<p>Compile DPDK using: meson <build> -Dexamples=l3fwd ; ninja -C <build></p> <p>During testing, l3fwd was given real-time scheduling priority.</p>
L3fwd settings on Guest OS	<p>Updated values /l3fwd/l3fwd.h:</p> <pre>#define RTE_TEST_RX_DESC_DEFAULT 2048 #define RTE_TEST_TX_DESC_DEFAULT 2048 #define MAX_PKT_BURST 64</pre>
Command Line on Guest OS	<pre>./build/examples/dpdk-l3fwd -c 0x3ffc00 -n 4 -a 00:07:0,mprq_en=1,rxqs_min_mprq=1,mprq_log_stride_num=8 --socket-mem=8192 -- -p 0x1 -P -- config='(0,0,21),(0,1,20),(0,2,19),(0,3,18),(0,4,17),(0,5,16),(0,6,15),(0,7,14),(0,8,13),(0,9,12),(0,10,11),(0,11,10)' --eth-dest=0,00:52:11:22:33:10</pre>

9.2 Test Results

Table 23: Test #7 Results – Mellanox ConnectX-5 Ex 100GbE Throughput at Zero Packet Loss using KVM SR-IOV

Frame Size (Bytes)	Frame Rate (Mpps)	Line Rate [100G] (Mpps)	% Line Rate
64	148.66	148.81	99.91
128	84.46	84.46	100
256	45.29	45.29	100
512	23.50	23.50	100
1024	11.97	11.97	100
1280	9.62	9.62	100
1518	8.13	8.13	100

Figure 14: Test #7 Results – Mellanox ConnectX-5 Ex 100GbE Throughput at Zero Packet Loss using KVM SR-IOV



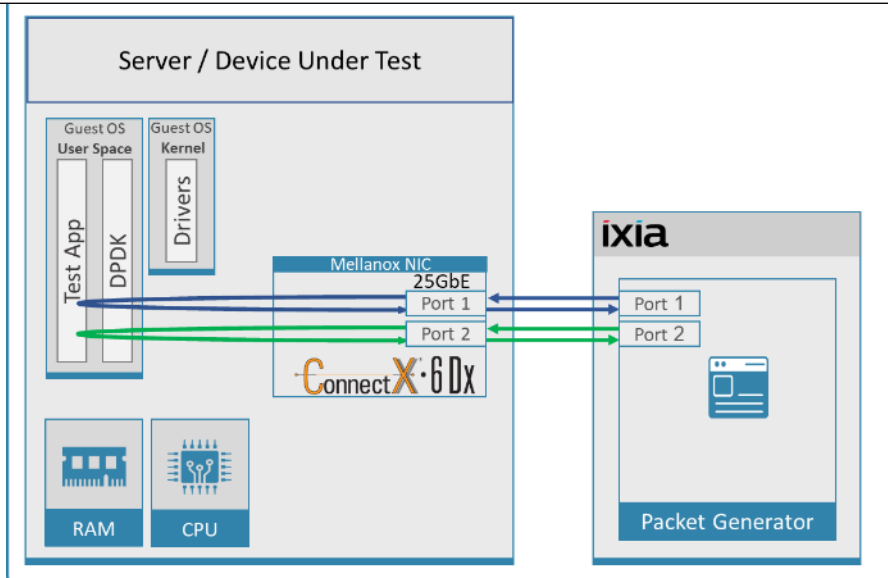
10 Test#8 Mellanox ConnectX-6Dx 25GbE Throughput at Zero Packet Loss (2x 25GbE)

Table 24: Test #8 Setup

Item	Description
Test #8	Mellanox ConnectX-6Dx 25GbE Dual-Port Throughput at zero packet loss
Server	HPE ProLiant DL380 Gen10
CPU	Intel(R) Xeon(R) Platinum 8168 CPU @ 2.70GHz 24 CPU cores * 2 NUMA nodes
RAM	384GB: 6 * 32GB DIMMs * 2 NUMA nodes @ 2666MHz
BIOS	U30 rev. 1.36 (02/15/2018)
NIC	One MCX623102AN-ADAT ConnectX-6 Dx EN adapter card; 25GbE; Dual-port SFP28; PCIe 4.0/3.0 x16
Operating System	Ubuntu 20.04.2 LTS (Focal Fossa)
Kernel Version	5.4.0-65-generic.x86_64
GCC version	gcc (Ubuntu 9.3.0-17ubuntu1~20.04) 9.3.0
Mellanox NIC firmware version	22.29.2002
Mellanox OFED driver version	MLNX_OFED_LINUX-5.2-2.2.0.0
DPDK version	21.02
Test Configuration	1 NIC, 2 ports; Each port receives a stream of 8192 IP flows from the IXIA Each port has 4 queues assigned for a total of 8 queues 1 queue assigned per logical core with a total of 8 logical cores

The Device Under Test (DUT) is made up of the HPE server and the Mellanox ConnectX-5 Dual-Port NIC. The DUT is connected to the IXIA packet generator which generates traffic towards the ConnectX-5 NIC. The ConnectX-5 data traffic is passed through DPDK to the test application **I3fwd** and is redirected to the opposite direction on the same port. IXIA measures throughput and packet loss.

Figure 15: Test #8 Setup – Mellanox ConnectX-6 Dx 25GbE Dual-Port connected to IXIA



10.1 Test Settings

Table 25: Test #8 Settings

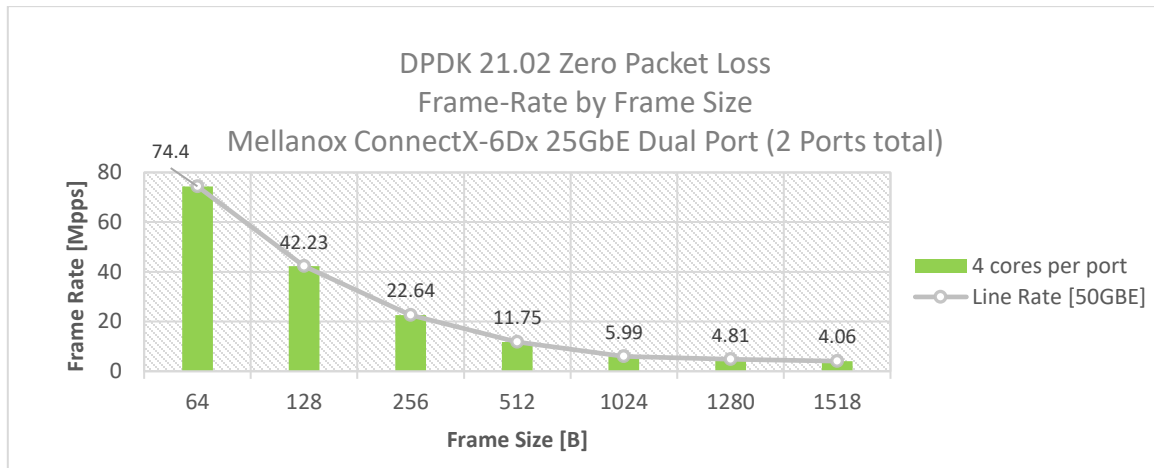
Item	Description
BIOS	<p>1) Workload Profile = "Low Latency";</p> <p>2) Jitter Control = Manual, 3400. (Setting turbo boost frequency to 3.4 GHz)</p> <p>See "Configuring and tuning HPE ProLiant Servers for low-latency applications": hpe.com > Search "DL380 gen10 low latency"</p>
BOOT Settings	<pre>isolcpus=0-23 intel_idle.max_cstate=0 processor.max_cstate=0 intel_pstate=disable nohz_full=0-23 rcu_nocbs=0-23 rcu_nocb_poll default_hugepagesz=1G hugepagesz=1G hugepages=64 audit=0 nosoftlockup idle=poll</pre>
DPDK Settings	<p>Compile DPDK using: <code>meson <build> -Dexamples=l3fwd ; ninja -C <build></code></p> <p>During testing, l3fwd was given real-time scheduling priority.</p>
L3fwd settings	<p>Updated values /l3fwd/l3fwd.h:</p> <pre>#define RTE_TEST_RX_DESC_DEFAULT 4096 #define RTE_TEST_TX_DESC_DEFAULT 4096 #define MAX_PKT_BURST 64</pre>
Command Line	<pre>./build/examples/dpdk-l3fwd -c 0xff0000 -n 4 -a 37:00.0,mprq_en=1,rxqs_min_mprq=1 -a 37:00.1,mprq_en=1,rxqs_min_mprq=1 --socket-mem=8192 -- -p 0x3 -P -- config='(0,0,23),(0,1,22),(0,2,21),(0,3,20),(1,0,19),(1,1,18),(1,2,17),(1,3,16)' --eth- dest=0,00:52:11:22:33:10 --eth-dest=1,00:52:11:22:33:20</pre>
Other optimizations	<p>a) Flow Control OFF: <code>"ethtool -A \$netdev rx off tx off"</code></p> <p>b) Memory optimizations: <code>"sysctl -w vm.zone_reclaim_mode=0"; "sysctl -w vm.swappiness=0"</code></p> <p>c) Move all IRQs to far NUMA node: <code>"IRQBALANCE_BANNED_CPUS=\$LOCAL_NUMA_CPUMAP irqbalance --oneshot"</code></p> <p>d) Disable irqbalance: <code>"systemctl stop irqbalance"</code></p> <p>e) Change PCI MaxReadReq to 1024B for each port of each NIC: Run <code>"setpci -s \$PORT_PCI_ADDRESS 68.w"</code>, it will return 4 digits ABCD --> Run <code>"setpci -s \$PORT_PCI_ADDRESS 68.w=3936"</code></p> <p>f) Set CQE COMPRESSION to "AGGRESSIVE": <code>mlxconfig -d \$PORT_PCI_ADDRESS set CQE_COMPRESSION=1</code></p> <p>g) Disable Linux realtime throttling: <code>echo -1 > /proc/sys/kernel/sched_rt_runtime_us</code></p>

10.2 Test Results

Table 26: Test #8 Results – Mellanox ConnectX-6Dx 25GbE Dual-Port Throughput at Zero Packet Loss

Frame Size (Bytes)	Frame Rate (Mpps)	Line Rate [50G] (Mpps)	% Line Rate
64	74.40	74.40	100.00
128	42.23	42.23	100.00
256	22.64	22.64	100.00
512	11.75	11.75	100.00
1024	5.99	5.99	100.00
1280	4.81	4.81	100.00
1518	4.06	4.06	100.00

Figure 16: Test #8 Results – Mellanox ConnectX-6Dx 25GbE Dual-Port Throughput at Zero Packet Loss



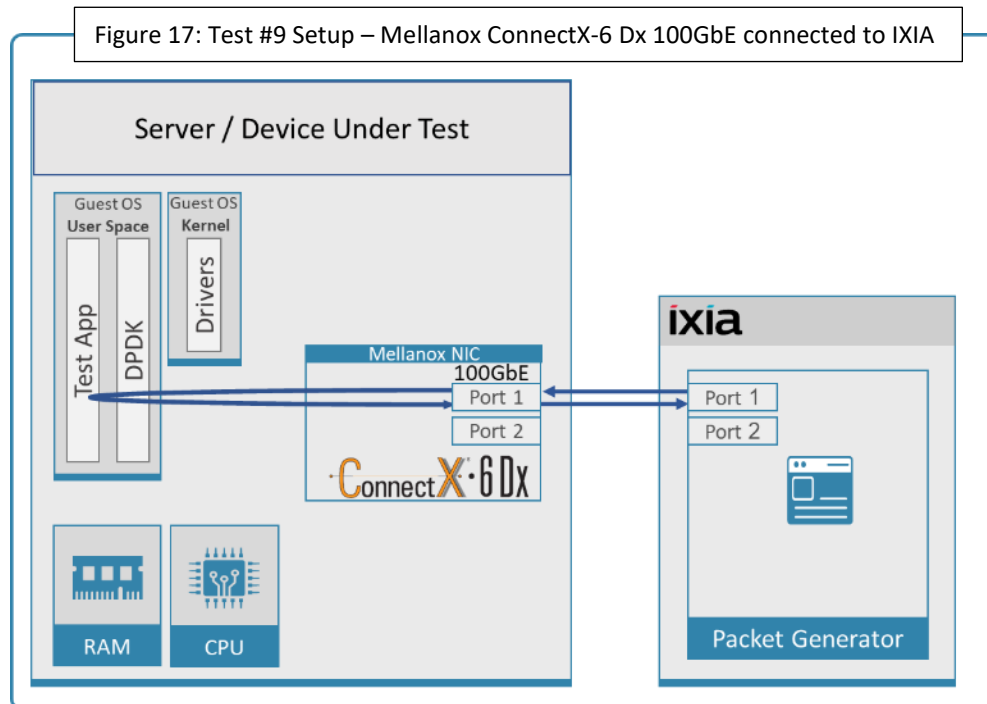
11 Test#9 Mellanox ConnectX-6 Dx 100GbE Throughput at Zero Packet Loss (1x 100GbE)

Table 27: Test #9 Setup

Item	Description
Test #9	Mellanox ConnectX-6 Dx 100GbE Throughput at zero packet loss
Server	HPE ProLiant DL380 Gen10
CPU	Intel(R) Xeon(R) Platinum 8168 CPU @ 2.70GHz 24 CPU cores * 2 NUMA nodes
RAM	384GB: 6 * 32GB DIMMs * 2 NUMA nodes @ 2666MHz
BIOS	U30 rev. 1.36 (02/15/2018)
NIC	One MCX623106AN-CDAT ConnectX-6 Dx EN adapter card; 100GbE; Dual-port QSFP56; PCIe 4.0/3.0 x16;
Operating System	Ubuntu 20.04.2 LTS (Focal Fossa)
Kernel Version	5.4.0-65-generic.x86_64
GCC version	gcc (Ubuntu 9.3.0-17ubuntu1~20.04) 9.3.0
Mellanox NIC firmware version	22.29.2002
Mellanox OFED driver version	MLNX_OFED_LINUX-5.2-2.2.0.0
DPDK version	21.02
Test Configuration	1 NIC, 1 port used on NIC; Port has 12 queues assigned to it, 1 queue per logical core for a total of 12 logical cores. Each port receives a stream of 8192 IP flows from the IXIA

The Device Under Test (DUT) is made up of the HPE server and the Mellanox ConnectX-6 Dx Dual-Port NIC (only the first port is used in this test). The DUT is connected to the IXIA packet generator which generates traffic towards the ConnectX-6Dx NIC.

The ConnectX-6Dx data traffic is passed through DPDK to the test application **l3fwd** and is redirected to the opposite direction on the same port. IXIA measures throughput and packet loss.



11.1 Test Settings

Table 28: Test #9 Settings

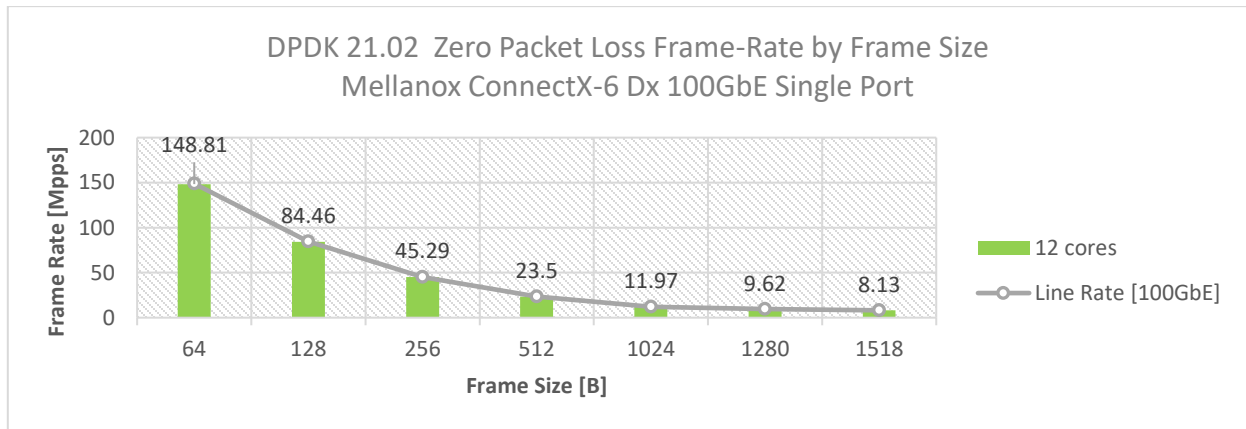
Item	Description
BIOS	<p>1) Workload Profile = "Low Latency";</p> <p>2) Jitter Control = Manual, 3400. (Setting turbo boost frequency to 3.4 GHz)</p> <p>See "Configuring and tuning HPE ProLiant Servers for low-latency applications": hpe.com > Search "DL380 gen10 low latency"</p>
BOOT Settings	<pre>isolcpus=24-47 intel_idle.max_cstate=0 processor.max_cstate=0 intel_pstate=disable nohz_full=24-47 rcu_nocbs=24-47 rcu_nocb_poll default_hugepagesz=1G hugepagesz=1G hugepages=64 audit=0 nosoflockup</pre>
DPDK Settings	<p>Compile DPDK using:</p> <pre>meson <build> -Dexamples=l3fwd ; ninja -C <build></pre> <p>During testing, l3fwd was given real-time scheduling priority.</p>
L3fwd settings	<p>Updated values /l3fwd/l3fwd.h:</p> <pre>#define RTE_TEST_RX_DESC_DEFAULT 4096 #define RTE_TEST_TX_DESC_DEFAULT 4096 #define MAX_PKT_BURST 64</pre>
Command Line	<pre>./build/examples/dpdk-l3fwd -c 0xffff00000000 -n 4 -a 0000:af:00.0,mprq_en=1,mprq_log_stride_num=8 --socket-mem=0,8192 -- -p 0x1 -P --config='(0,0,47),(0,1,46),(0,2,45),(0,3,44),(0,4,43),(0,5,42),(0,6,41),(0,7,40),(0,8,39),(0,9,38),(0,10,37),(0,11,36)' --eth-dest=0,00:52:11:22:33:10</pre>
Other optimizations	<p>a) Flow Control OFF: "ethtool -A \$netdev rx off tx off"</p> <p>b) Memory optimizations: "sysctl -w vm.zone_reclaim_mode=0"; "sysctl -w vm.swappiness=0"</p> <p>c) Move all IRQs to far NUMA node: "IRQBALANCE_BANNED_CPUS=\$LOCAL_NUMA_CPUMAP irqbalance --oneshot"</p> <p>d) Disable irqbalance: "systemctl stop irqbalance"</p> <p>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"</p> <p>f) Set CQE COMPRESSION to "AGGRESSIVE": mxconfig -d \$PORT_PCI_ADDRESS set CQE_COMPRESSION=1</p> <p>g) Disable Linux realtime throttling: echo -1 > /proc/sys/kernel/sched_rt_runtime_us</p>

11.2 Test Results

Table 29: Test #9 Results – Mellanox ConnectX-6 Dx 100GbE Throughput at Zero Packet Loss

Frame Size (Bytes)	Frame Rate (Mpps)	Line Rate [100G] (Mpps)	% Line Rate
64	148.81	148.81	100.00
128	84.46	84.46	100.00
256	45.29	45.29	100.00
512	23.50	23.50	100.00
1024	11.97	11.97	100.00
1280	9.62	9.62	100.00
1518	8.13	8.13	100.00

Figure 18: Test #9 Results – Mellanox ConnectX-5 Ex 100GbE Throughput at Zero Packet Loss



12 Test#10 Mellanox ConnectX-6Dx 100GbE Single Core Performance (2x 100GbE)

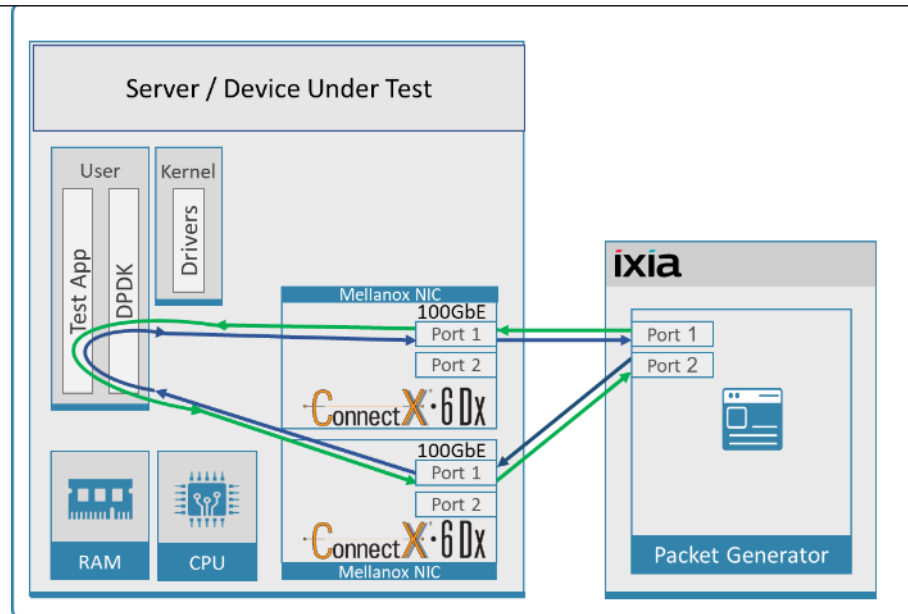
Table 30: Test #10 Setup

Item	Description
Test #10	Mellanox ConnectX-6Dx 100GbE Single Core Performance
Server	HPE ProLiant DL380 Gen10
CPU	Intel(R) Xeon(R) Platinum 8168 CPU @ 2.70GHz; 24 CPU cores * 2 NUMA nodes
RAM	384GB: 6 * 32GB DIMMs * 2 NUMA nodes @ 2666MHz
BIOS	U30 rev. 1.36 (02/15/2018)
NIC	Two MCX623106AN-CDAT ConnectX-6 Dx EN adapter cards; 100GbE; Dual-port QSFP56; PCIe 4.0/3.0 x16;
Operating System	Ubuntu 20.04.2 LTS (Focal Fossa)
Kernel Version	5.4.0-65-generic.x86_64
GCC version	gcc (Ubuntu 9.3.0-17ubuntu1~20.04) 9.3.0
Mellanox NIC firmware version	22.29.2002
Mellanox OFED driver version	MLNX_OFED_LINUX-5.2-2.2.0.0
DPDK version	21.02
Test Configuration	2 NICs; 1 port used on each. Each port receives a stream of 8192 UDP flows from the IXIA 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.

The Device Under Test (DUT) is made up of the HPE server and two Mellanox ConnectX-6 Dx 100GbE NICs utilizing one port each. The DUT is connected to the IXIA packet generator which generates traffic towards the first port of both ConnectX-6 Dx 100GbE NICs.

The ConnectX-6 Dx 100GbE data traffic is passed through DPDK to the test application **testpmd** and is redirected to the opposite direction on the opposing NIC's port. IXIA measures throughput and packet loss.

Figure 19: Test #10 Setup – Two Mellanox ConnectX-6 Dx 100GbE connected to IXIA



12.1 Test Settings

Table 31: Test #10 Settings

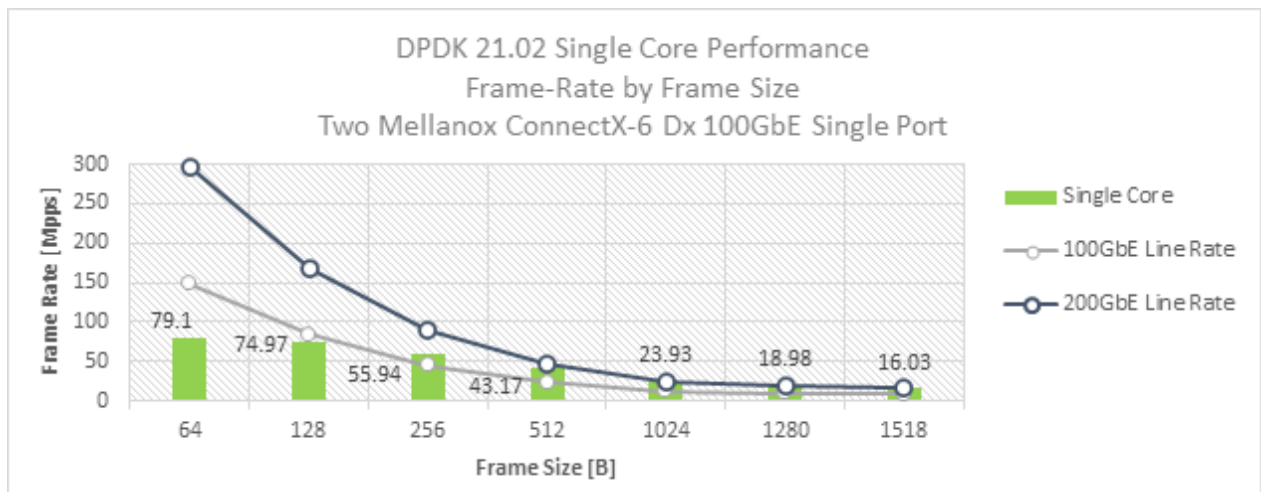
Item	Description
BIOS	<p>1) Workload Profile = "Low Latency"</p> <p>2) Jitter Control = Manual, 3400. (Setting turbo boost frequency to 3.4 GHz)</p> <p>See "Configuring and tuning HPE ProLiant Servers for low-latency applications": hpe.com > Search "DL380 gen10 low latency"</p>
BOOT Settings	<pre>isolcpus=24-47 intel_idle.max_cstate=0 processor.max_cstate=0 intel_pstate=disable nohz_full=24-47 rcu_nocbs=24-47 rcu_nocb_poll default_hugepagesz=1G hugepagesz=1G hugepages=64 audit=0 nosoftlockup</pre>
DPDK Settings	<p>Compile DPDK using: <code>meson <build> ; ninja -C <build></code></p> <p>During testing, testpmd was given real-time scheduling priority.</p>
Command Line	<pre>./build/app/dpdk-testpmd -c 0x110000000000 -n 4 -a 86:00.0,decap_en=0 -a af:00.0,decap_en=0 --socket-mem=0,8192 --port-numa-config=0,1,1,1 --socket-num=1 --burst=64 --txd=1024 --rx=1024 --mbcache=512 --rxq=1 --txq=1 --nb-cores=1 -i -a --rss-udp --disable-crc-strip --record-core-cycles --record-burst-stats</pre>
Other optimizations	<p>a) Flow Control OFF: <code>"ethtool -A \$netdev rx off tx off"</code></p> <p>b) Memory optimizations: <code>"sysctl -w vm.zone_reclaim_mode=0"; "sysctl -w vm.swappiness=0"</code></p> <p>c) Move all IRQs to far NUMA node: <code>"IRQBALANCE_BANNED_CPUS=\$LOCAL_NUMA_CPUMAP irqbalance --oneshot"</code></p> <p>d) Disable irqbalance: <code>"systemctl stop irqbalance"</code></p> <p>e) Change PCI MaxReadReq to 1024B for each port of each NIC: Run <code>"setpci -s \$PORT_PCI_ADDRESS 68.w"</code>, it will return 4 digits ABCD --> Run <code>"setpci -s \$PORT_PCI_ADDRESS 68.w=3BCD"</code></p> <p>f) Set CQE COMPRESSION to "AGGRESSIVE": <code>mlxconfig -d \$PORT_PCI_ADDRESS set CQE_COMPRESSION=1</code></p> <p>g) Disable Linux realtime throttling: <code>echo -1 > /proc/sys/kernel/sched_rt_runtime_us</code></p>

12.2 Test Results

Table 32: Test #10 Results – Mellanox ConnectX-6 Dx 100GbE Single Core Performance

Frame Size (Bytes)	Frame Rate (Mpps)	Line Rate [200G] (Mpps)	Line Rate [100G] (Mpps)	Throughput (Gbps)	CPU Cycles per packet NOTE: Lower is Better
64	79.1	297.62	148.81	40.488	30
128	74.97	168.92	84.46	76.762	29
256	55.94	90.58	45.29	114.565	28
512	43.17	46.99	23.50	176.835	29
1024	23.93	23.95	11.97	196.06	30
1280	18.98	19.23	9.62	194.356	31
1518	16.03	16.25	8.13	194.702	31

Figure 20: Test #10 Results – Mellanox ConnectX-6Dx 100GbE Single Core Performance



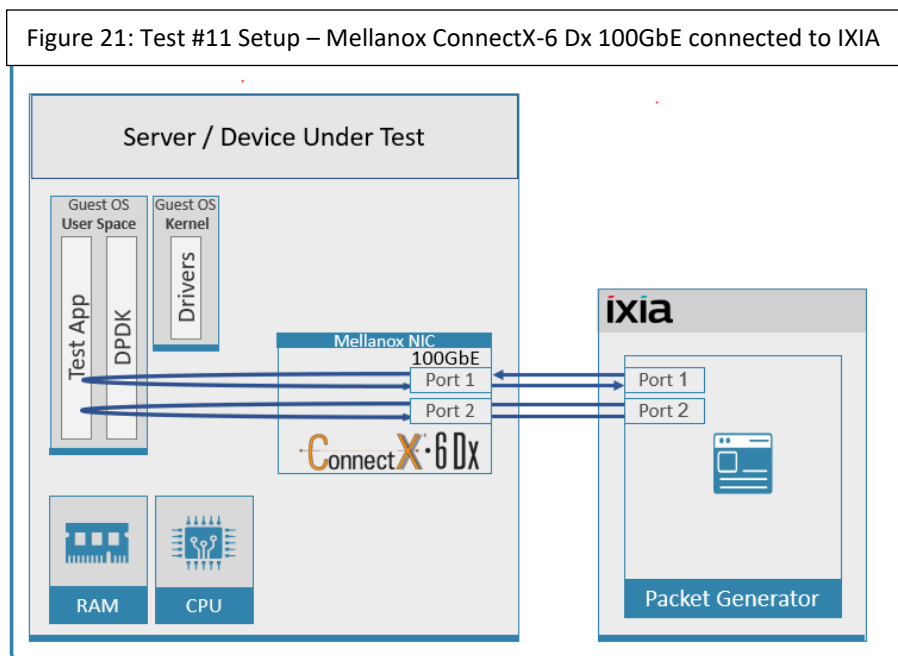
13 Test#11 Mellanox ConnectX-6 Dx 100GbE PCIe Gen4 Throughput at Zero Packet Loss (2x 100GbE)

Table 33: Test #11 Setup

Item	Description
Test #11	Mellanox ConnectX-6 Dx 100GbE Dual-Port PCIe Gen 4 Throughput at zero packet loss
Server	Dell PowerEdge R6525
CPU	2x AMD EPYC 7742 64-Core Processor
RAM	512GB: 16 * 32GB DIMMs @ 3200MHz
BIOS	Vendor: Dell Inc. version 1.5.5 (07/27/2020)
NIC	One MCX623106PN-CDAT ConnectX-6 Dx EN adapter card; 100GbE; Dual-port QSFP56; with PPS In/Out; PCIe 4.0 x16
Operating System	Ubuntu 20.04.2 LTS (Focal Fossa)
Kernel Version	5.4.0-65-generic.x86_64
GCC version	gcc (Ubuntu 9.3.0-17ubuntu1~20.04) 9.3.0
Mellanox NIC firmware version	22.29.2002
Mellanox OFED driver version	MLNX_OFED_LINUX-5.2-2.2.0.0
DPDK version	21.02
Test Configuration	1 NIC, 2 port used on NIC; each port has 8 queues assigned to it, 1 queue per logical core for a total of 16 logical cores. Each port receives a stream of 8192 IP flows from the IXIA

The Device Under Test (DUT) is made up of the Dell server and the Mellanox ConnectX-6 Dx Dual-Port NIC (both ports are used in this test). The DUT is connected to the IXIA packet generator which generates traffic towards the ConnectX-6 Dx NIC ports.

The ConnectX-6 Dx data traffic is passed via PCIe Gen 4 bus through DPDK to the test application **l3fwd** and is redirected to the opposite direction on the same port. IXIA measures throughput and packet loss.



13.1 Test Settings

Table 34: Test #11 Settings

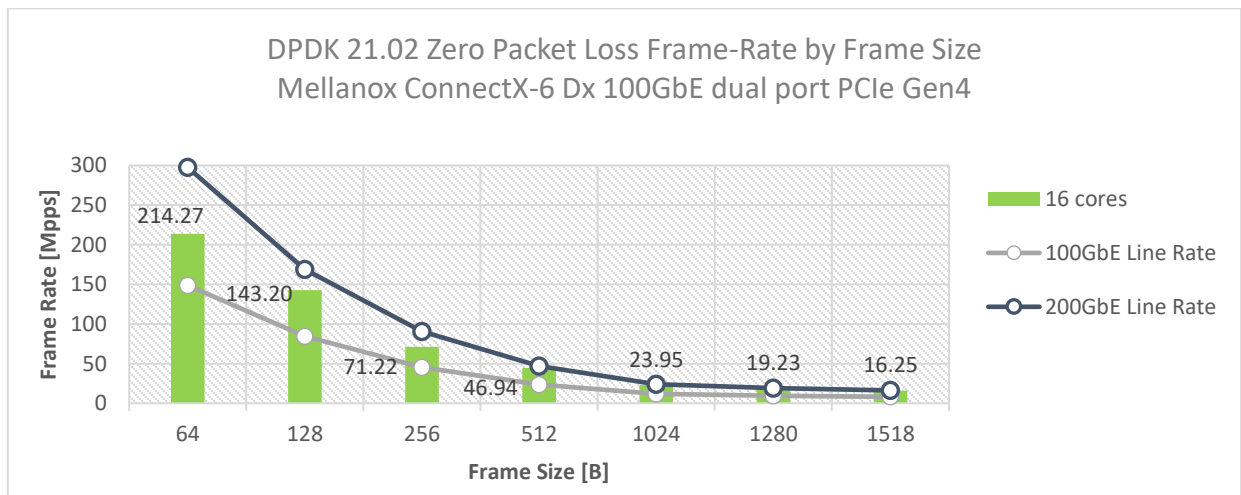
Item	Description
BIOS	System Profile Settings > System Profile> Custom System Profile Settings > CPU Power Management > Maximum Performance System Profile Settings > Memory Frequency > Maximum Performance System Profile Settings > Alg. Performance Boost Disable (ApbDis) > Enabled System Profile Settings ApbDis Fixed Socket P-State > P0 Processor Settings > NUMA Nodes Per Socket > 2 Processor Settings > L3 cache as NUMA Domain > Enabled Processor Settings > x2APIC Mode > Enabled Bios settings > Integrated devices> Pcie Preferred IO Device >Enabled Bios settings > Integrated devices> Pcie Preferred IO Device Bus > xx
BOOT Settings	iommu=pt default_hugepagesz=1G hugepagesz=1G hugepages=128 isolcpus=16-127 nohz_full=16-127 numa_balancing=disable processor.max_cstate=0 nosoftlockup rcu_nocbs=16-127
DPDK Settings	Compile DPDK using: meson <build> -Dmax_numa_nodes=32 -Dexamples=l3fwd ; ninja -C <build> During testing, l3fwd was given real-time scheduling priority.
L3fwd settings	Updated values /l3fwd/l3fwd.h: <pre>#define RTE_TEST_RX_DESC_DEFAULT 4096 #define RTE_TEST_TX_DESC_DEFAULT 4096 #define MAX_PKT_BURST 64 #define NB_SOCKETS 32</pre>
Command Line	<pre>./build/examples/dpdk-l3fwd -l 96-111 -n 4 -- socketmem=0,4096 -a 0000:a1:00:0,mprq_en=1,rxqs_min_mprq=1,mprq_log_stride_num=9,txq_inline_mpw=128,rxq_pkt_ pad_en=1 -a 0000:a1:00:1,mprq_en=1,rxqs_min_mprq=1,mprq_log_stride_num=9,txq_inline_mpw=128,rxq_pkt_ pad_en=1 -- -p 0x3 -P -- config='(0,0,111),(0,1,110),(0,2,109),(0,3,108),(0,4,107),(0,5,106),(0,6,105),(0,7,104),(1,0,103),(1,1,10 2),(1,2,101),(1,3,100),(1,4,99),(1,5,98),(1,6,97),(1,7,96)' --eth-dest=0,00:52:11:22:33:10 -- ethdest=1,00:52:11:22:33:20</pre>
Other optimizations	a) Flow Control OFF: "ethtool -A \$netdev rx off tx off" (for both ports) 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 4096B 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= 5BCD " f) Set CQE COMPRESSION to "AGGRESSIVE": mlxconfig -d \$PORT_PCI_ADDRESS set CQE_COMPRESSION=1 g) Set PCI write ordering: mlxconfig -d \$PORT_PCI_ADDRESS set PCI_WR_ORDERING=1 h) Disable Linux real-time throttling: echo -1 > /proc/sys/kernel/sched_rt_runtime_us i) Disable auto neg for both ports: ethtool -s \$PORT_PCI_ADDRESS autoneg off speed 100000

13.2 Test Results

Table 35: Test #11 Results – Mellanox ConnectX-6 Dx 100GbE Dual-Port PCIe Gen4 Zero Packet Loss Throughput

Frame Size (Bytes)	Frame Rate (Mpps)	Line Rate [200G] (Mpps)	Line Rate [100G] (Mpps)	% Line Rate
64	214.27	297.62	148.81	72
128	143.20	168.92	84.46	84.78
256	71.22	90.58	45.29	78.64
512	46.94	46.99	23.50	99.9
1024	23.95	23.95	11.97	100.00
1280	19.23	19.23	9.62	100.00
1518	16.25	16.25	8.13	100.00

Figure 22: Test #11 Results – Mellanox ConnectX-6 Dx 100GbE Dual-Port PCIe Gen4 Throughput at Zero Packet Loss



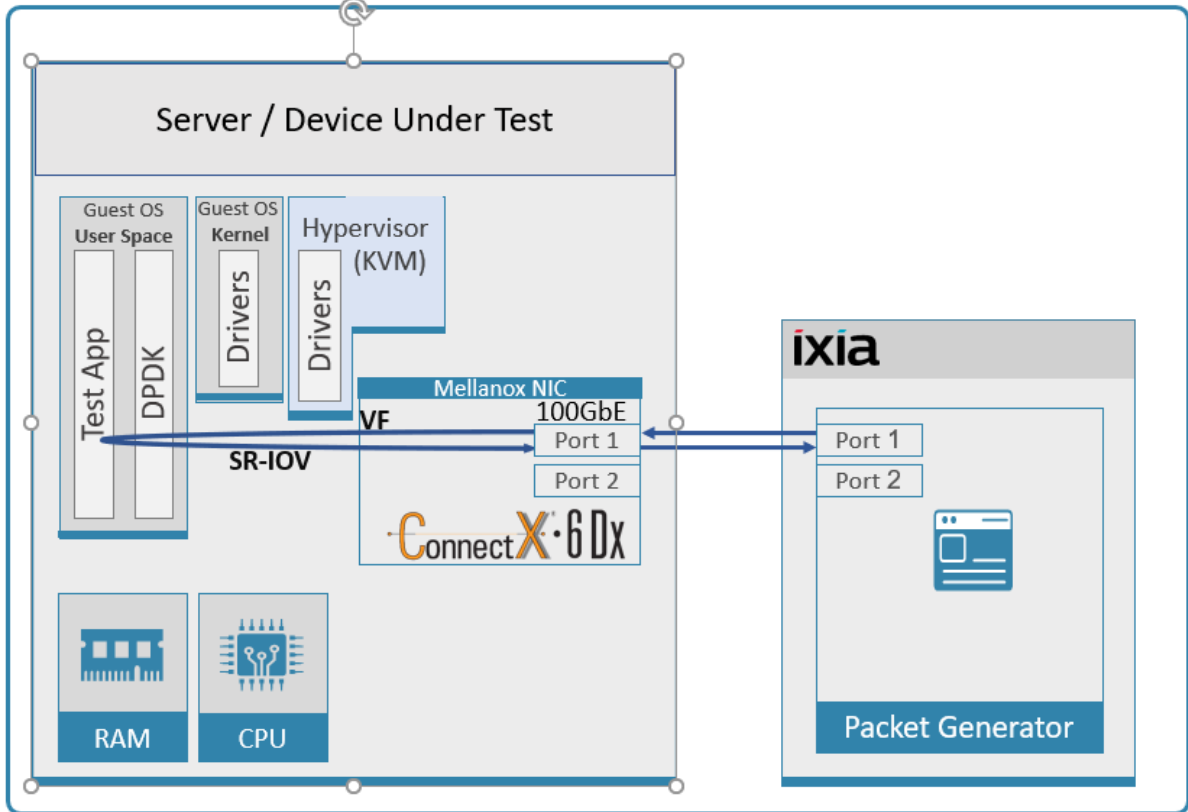
14 Test#12 Mellanox ConnectX-6 Dx 100GbE Throughput at Zero Packet Loss (1x 100GbE) using SR-IOV over KVM Hypervisor

Table 36 - Test #12 Setup

Item	Description
Test #12	Mellanox ConnectX-6 Dx 100GbE Throughput at zero packet loss using SR-IOV over KVM Hypervisor
Server	HPE ProLiant DL380 Gen10
CPU	Intel(R) Xeon(R) Platinum 8168 CPU @ 2.70GHz 24 CPU cores * 2 NUMA nodes
RAM	384GB: 6 * 32GB DIMMs * 2 NUMA nodes @ 2666MHz
BIOS	U30 rev. 1.36 (02/15/2018)
NIC	One MCX623106AN-CDAT ConnectX-6 Dx EN adapter card; 100GbE; Dual-port QSFP56; PCIe 4.0/3.0 x16;
Hypervisor	Ubuntu 20.04.2 LTS (Focal Fossa) QEMU emulator version 4.2.1 (Debian 1:4.2-3ubuntu6.11)
Hypervisor Kernel Version	5.4.0-65-generic.x86_64
Hypervisor Mellanox Driver	MLNX_OFED_LINUX-5.2-2.2.0.0
Guest Operating System	Red Hat Enterprise Linux Server release 7.7 (Maipo)
Guest Kernel Version	3.10.0-1062.el7.x86_64
Guest GCC version	4.8.5 20150623 (Red Hat 4.8.5-28) (GCC)
Guest Mellanox OFED driver version	MLNX_OFED_LINUX-5.2-2.2.0.0
Mellanox NIC firmware version	22.29.2002
DPDK version	21.02
Test Configuration	1 NIC, 1 port over 1 VF (SR-IOV); VF has 12 queues assigned to it, 1 queue per logical core for a total of 12 logical cores. Each physical port receives a stream of 8192 IP flows from the IXIA directed to VF assigned to Guest OS.

The Device Under Test (DUT) is made up of the HPE server and the Mellanox ConnectX-6 Dx NIC with a dual- port (only first port used in this test) running Red Hat Enterprise Linux Server with qemu-KVM managed via libvirt, Guest OS running DPDK is based on Red Hat Enterprise Linux Server as well. The DUT is connected to the IXIA packet generator which generates traffic towards the ConnectX-6 Dx NIC. The ConnectX-6 Dx data traffic is passed through a virtual function (VF/SR-IOV) to DPDK running on the Guest OS, to the test application **l3fwd** and is redirected to the opposite direction on the same port. IXIA measures throughput and packet loss.

Figure 23 - Test #12 Setup – Mellanox ConnectX-6 Dx 100GbE connected to IXIA using KVM SR-IOV



14.1 Test Settings

Table 37 - Test #12 Settings

Item	Description
BIOS	<p>1) Workload Profile = "Low Latency";</p> <p>2) Jitter Control = Manual, 3400. (Setting turbo boost frequency to 3.4 GHz)</p> <p>3) Change "Workload Profile" to "Custom"</p> <p>4) Change VT-x, VT-d and SR-IOV from "Disabled" to "Enabled".</p> <p>See "Configuring and tuning HPE ProLiant Servers for low-latency applications": hpe.com > Search "DL380 gen10 low latency"</p>
Hypervisor BOOT Settings	<pre>isolcpus=24-47 intel_idle.max_cstate=0 processor.max_cstate=0 nohz_full=24-47 rcu_nocbs=24-47 intel_pstate=disable default_hugepagesz=1G hugepagesz=1G hugepages=70 audit=0 nosoftlockup intel_iommu=on iommu=pt rcu_nocb_poll</pre>
Hypervisor settings	<p>1) Enable SRIOV via NIC configuration tool: (requires installation of mft-tools)</p> <pre>mlxconfig -d /dev/mst/mt4121_pciconf1 set NUM_OF_VFS=1 SRIOV_EN=1 CQE_COMPRESSION=1</pre> <pre>echo 1 > /sys/class/net/ens5f0/device/sriov_numvfs</pre> <p>2) Assign VF</p> <pre>HCA_netintf=ens5f0 #assign a VF to the DUT device</pre> <pre>VF_PCI_address="0000:af:00.2" #VF PCI address</pre> <pre>echo \$VF_PCI_address > /sys/bus/pci/drivers/mlx5_core/unbind</pre> <pre>modprobe vfio-pci</pre> <pre>echo "\$(cat /sys/bus/pci/devices/\$VF_PCI_address/vendor) \$(cat /sys/bus/pci/devices/\$VF_PCI_address/device)" > /sys/bus/pci/drivers/vfio-pci/new_id</pre> <p># Now the VF may be assigned to Guest (passthrough) with libvirt virt-manager.</p> <p>3) Setting VF MAC - use the command below (find out the vf-index from "ip link show"), ip link set <<PF NIC interface>> <vf index> mac <MAC Address> : (mac is random)</p> <pre>ip link set \$HCA_netintf vf 0 mac 00:52:11:22:33:42</pre> <p>4) VM tuning: vcpupin and memory backing from hugepages:</p> <p>To persistently configure vcpu pinning and memory backing, add the below config to the VM's XML config before starting the VM. Add the following two elements to the XML: <cputune> and <memoryBacking> and also increase the number of cpus and memory: virsh edit <vmID> (to get vmID use - virsh list --all)</p> <p>Example xml configuration: (change "nodeset" and "cpuset" attributes to suit the local NUMA node in your setup)</p> <pre><domain type='kvm' id='1'> <name>perf-dpdk-01-005-RH-7.4</name> <uuid>06f283fc-fd76-4411-8b6a-72fe94f50376</uuid> <memory unit='KiB'>33554432</memory> <currentMemory unit='KiB'>33554432</currentMemory> <memoryBacking> <hugepages> <page size='1048576' unit='KiB' nodeset='0'/> </hugepages> </memoryBacking> <nosharepages/> </domain></pre>

Item	Description
	<pre> <locked/> </memoryBacking> <vcpu placement='static'>23</vcpu> <cputune> <vcpupin vcpu='0' cpuset='24'/> <vcpupin vcpu='1' cpuset='25'/> <vcpupin vcpu='2' cpuset='26'/> <vcpupin vcpu='3' cpuset='27'/> <vcpupin vcpu='4' cpuset='28'/> <vcpupin vcpu='5' cpuset='29'/> <vcpupin vcpu='6' cpuset='30'/> <vcpupin vcpu='7' cpuset='31'/> <vcpupin vcpu='8' cpuset='32'/> <vcpupin vcpu='9' cpuset='33'/> <vcpupin vcpu='10' cpuset='34'/> <vcpupin vcpu='11' cpuset='35'/> <vcpupin vcpu='12' cpuset='36'/> <vcpupin vcpu='13' cpuset='37'/> <vcpupin vcpu='14' cpuset='38'/> <vcpupin vcpu='15' cpuset='39'/> <vcpupin vcpu='16' cpuset='40'/> <vcpupin vcpu='17' cpuset='41'/> <vcpupin vcpu='18' cpuset='42'/> <vcpupin vcpu='19' cpuset='43'/> <vcpupin vcpu='20' cpuset='44'/> <vcpupin vcpu='21' cpuset='45'/> <vcpupin vcpu='22' cpuset='46'/> </cputune> </pre>
Other optimizations on Hypervisor	<p>a) Flow Control OFF: "ethtool -A \$Netdev rx off tx off"</p> <p>b) Memory optimizations: "sysctl -w vm.zone_reclaim_mode=0"; "sysctl -w vm.swappiness=0"</p> <p>c) Move all IRQs to far NUMA node: "IRQBALANCE_BANNED_CPUS=\$LOCAL_NUMA_CPUMAP irqbalance --oneshot"</p> <p>d) Disable irqbalance: "systemctl stop irqbalance"</p> <p>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"</p> <p>f) Disable Linux realtime throttling: echo -1 > /proc/sys/kernel/sched_rt_runtime_us</p>
Guest BOOT Settings	<pre> isolcpus=0-22 intel_idle.max_cstate=0 processor.max_cstate=0 intel_pstate=disable idle=poll nohz_full=0-22 rcu_nocbs=0-22 rcu_nocb_poll default_hugepagesz=1G hugepagesz=1G hugepages=16 nosoftlockup </pre>
Other optimizations on Guest OS	<p>a) Flow Control OFF: "ethtool -A \$Netdev rx off tx off"</p> <p>b) Memory optimizations: "sysctl -w vm.zone_reclaim_mode=0"; "sysctl -w vm.swappiness=0"</p> <p>c) Move all IRQs to far NUMA node: "IRQBALANCE_BANNED_CPUS=\$LOCAL_NUMA_CPUMAP irqbalance --oneshot"</p> <p>d) Disable irqbalance: "systemctl stop irqbalance"</p> <p>e) Disable Linux realtime throttling: echo -1 > /proc/sys/kernel/sched_rt_runtime_us</p>

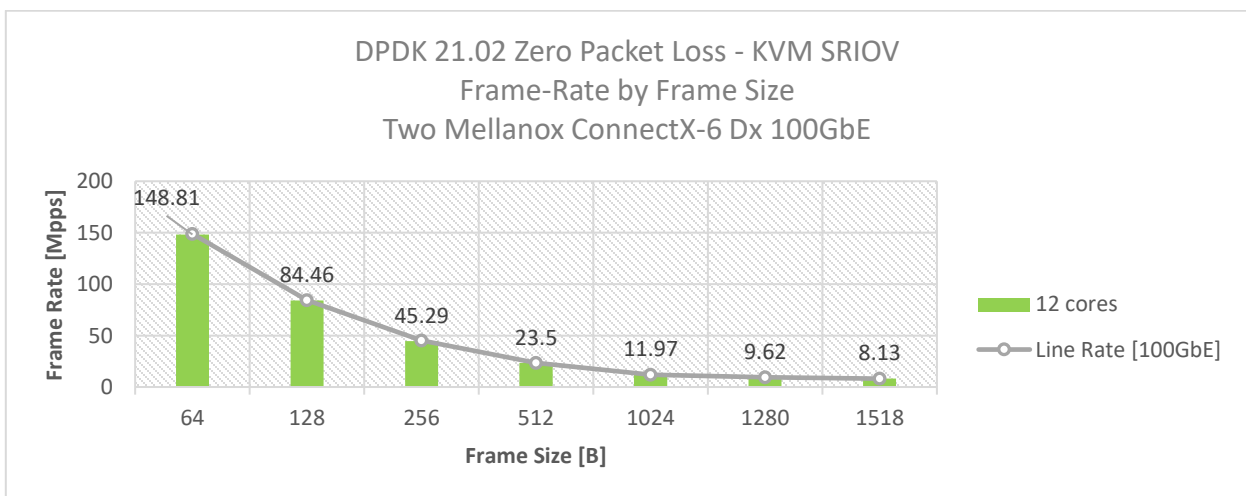
Item	Description
DPDK Settings on Guest OS	Compile DPDK using: meson <build> -Dexamples=l3fwd ; ninja -C <build> During testing, l3fwd was given real-time scheduling priority.
L3fwd settings on Guest OS	Updated values /l3fwd/l3fwd.h: #define RTE_TEST_RX_DESC_DEFAULT 2048 #define RTE_TEST_TX_DESC_DEFAULT 2048 #define MAX_PKT_BURST 64
Command Line on Guest OS	./build/examples/dpdk-l3fwd -c 0x3ffc00 -n 4 -w 00:07.0,mprq_en=1,rxqs_min_mprq=1,mprq_log_stride_num=8 --socket-mem=8192 -- -p 0x1 -P -- config='(0,0,21),(0,1,20),(0,2,19),(0,3,18),(0,4,17),(0,5,16),(0,6,15),(0,7,14),(0,8,13),(0,9,12),(0,10,11),(0,11,10)' --eth-dest=0,00:52:11:22:33:10

14.2 Test Results

Table 38 - Test #12 Results – Mellanox ConnectX-6 Dx 100GbE Throughput at Zero Packet Loss using KVM SR-IOV

Frame Size (Bytes)	Frame Rate (Mpps)	Line Rate [100G] (Mpps)	% Line Rate
64	148.81	148.81	100.00
128	84.46	84.46	100.00
256	45.29	45.29	100.00
512	23.50	23.50	100.00
1024	11.97	11.97	100.00
1280	9.62	9.62	100.00
1518	8.13	8.13	100.00

Figure 24 - Test #12 Results – Mellanox ConnectX-6 Dx 100GbE Throughput at Zero Packet Loss using KVM SR-IOV



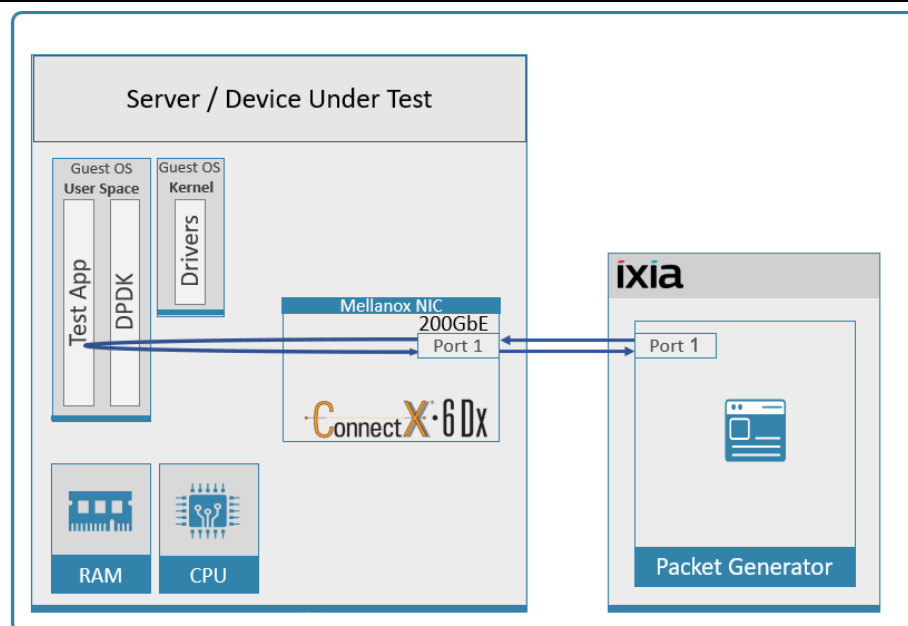
15 Test#13 Mellanox ConnectX-6 Dx 200GbE PCIe Gen4 Throughput at Zero Packet Loss (1x 200GbE)

Table 39 - Test #13 Setup

Item	Description
Test #13	Mellanox ConnectX-6 Dx 200GbE single-port PCIe Gen4 throughput at zero packet loss
Server	Dell PowerEdge R6525
CPU	2x AMD EPYC 7742 64-Core Processor
RAM	512GB: 16 * 32GB DIMMs @ 3200MHz
BIOS	Vendor: Dell Inc. version 1.5.5 (07/27/2020)
NIC	One MCX623105AN-VDAT ConnectX-6 Dx EN adapter card, 200GbE, Single-port QSFP56, PCIe 4.0 x16, No Crypto
Operating System	Ubuntu 20.04.2 LTS (Focal Fossa)
Kernel Version	5.4.0-65-generic.x86_64
GCC version	gcc (Ubuntu 9.3.0-17ubuntu1~20.04) 9.3.0
Mellanox NIC firmware version	22.29.2002
Mellanox OFED driver version	MLNX_OFED_LINUX-5.2-2.2.0.0
DPDK version	21.02
Test Configuration	1 NIC, 1 port used on NIC; Port has 16 queues assigned to it, 1 queue per logical core for a total of 16 logical cores. Each port receives a stream of 8192 IP flows from the IXIA

The Device Under Test (DUT) is made up of the Dell server and the Mellanox ConnectX-6 Dx Single-Port NIC . The DUT is connected to the IXIA packet generator which generates traffic towards the ConnectX-6 Dx NIC port. The ConnectX-6 Dx data traffic is passed via PCIe Gen 4 bus through DPDK to the test application **13fwd** and is redirected to the opposite direction on the same port. IXIA measures throughput and packet loss.

Figure 25 - Test #13 Setup – Mellanox ConnectX-6 Dx 200GbE connected to IXIA

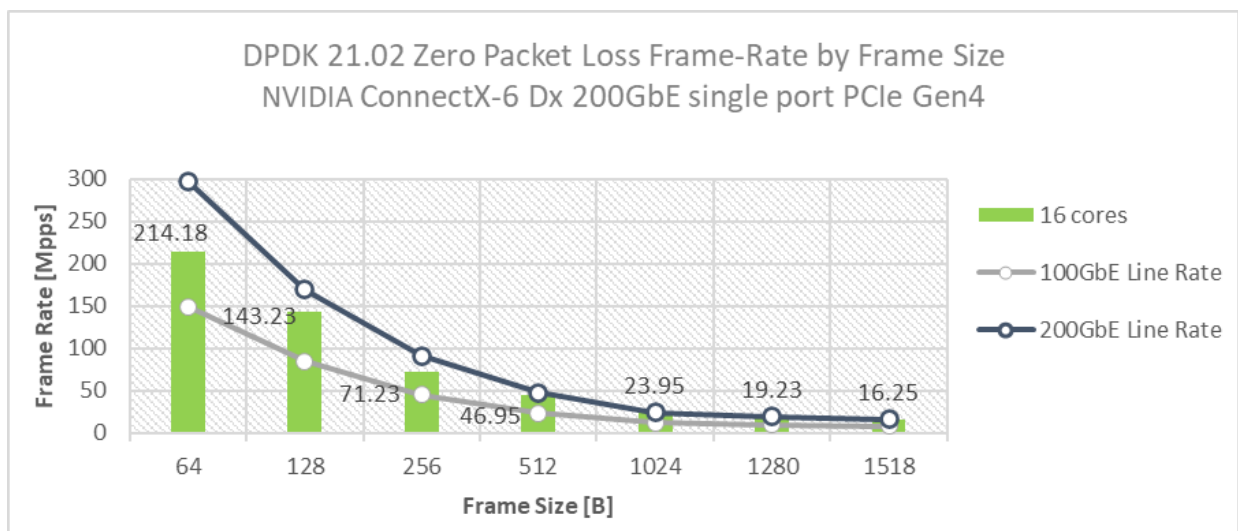


15.2 Test Results

Table 41 - Test #13 Results – Mellanox ConnectX-6 Dx 200GbE single port PCIe Gen4 Throughput at Zero Packet Loss

Frame Size (Bytes)	Frame Rate (Mpps)	Line Rate [200G] (Mpps)	Line Rate [100G] (Mpps)	% Line Rate
64	214.18	297.62	148.81	71.96
128	143.23	168.92	84.46	84.79
256	71.23	90.58	45.29	78.64
512	46.46	46.99	23.50	98.86
1024	23.93	23.95	11.97	99.91
1280	19.23	19.23	9.62	100
1518	16.24	16.25	8.13	99.91

Figure 26 - Test #13 Results – Mellanox ConnectX-6 Dx 200GbE dual port PCIe Gen4 Throughput at Zero Packet Loss



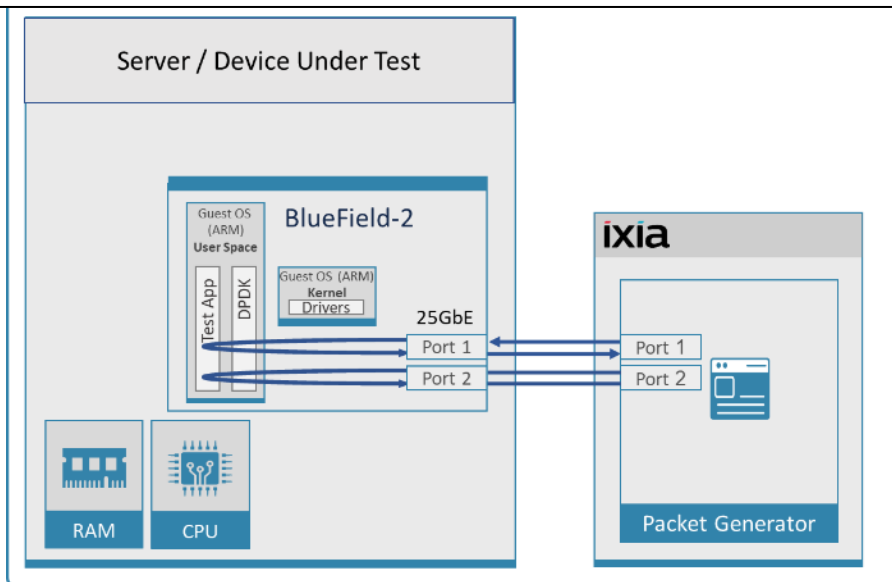
16 Test#14 BlueField-2 25GbE Throughput at Zero Packet Loss (2x 25GbE)

Table 42 - Test #14 Setup

Item	Description
Test #14	BlueField-2 25GbE Dual-Port Throughput at zero packet loss
Server	HPE ProLiant DL380 Gen10
Data Processing Unit (DPU)	One MBF2H332A-AEEOT_A1 BlueField-2 P-Series SmartNIC; 25GbE; Dual-port SFP56; PCIe Gen3/4 x8
DPU hosted CPUs	BlueField-2 A1 A72 @2.5GHz , 8 Cores-Processor
DPU RAM	DDR On-board Memory 16GB
DPU BIOS	U30 rev. 1.36 (02/15/2018)
Operating System	BlueField-2, Ubuntu20.04-MLNX_OFED-5.2-2.2.0.0-3.5.1.11601-1-aarch64
DPU Kernel Version	5.4.0-1007-bluefield.aarch64
DPU GCC version	gcc (Ubuntu 9.3.0-17ubuntu1~20.04) 9.3.0
Mellanox NIC/DPU firmware version	24.29.2002
Mellanox OFED driver version	MLNX_OFED_LINUX-5.2-2.2.0.0
DPDK version	21.02
Test Configuration	1 NIC/DPU, 2 ports; Each port receives a stream of 7500 UDP flows from the IXIA 1 queue assigned per logical core with a total of 2,4 and 8 logical cores

The Device Under Test (DUT) is made up of the HPE server and one BlueField-2 25GbE DPU utilizing two ports. It is connected to the IXIA packet generator which generates traffic towards both ports of the BlueField-2 25GbE DPU. BlueField-2 25GbE data traffic is passed through DPDK to the test application **testpmd** that is running on the ARM cores (**embedded in the DPU**) and is redirected to the opposite direction using the second port. IXIA measures throughput and packet loss. The test measured the results while using 1,2,4,6 or 7 ARM cores.

Figure 27 -Test #14 Setup – NVIDIA BlueField-2 25GbE Dual-Port connected to IXIA



16.1 Test Settings

Table 43 - Test #14 Settings

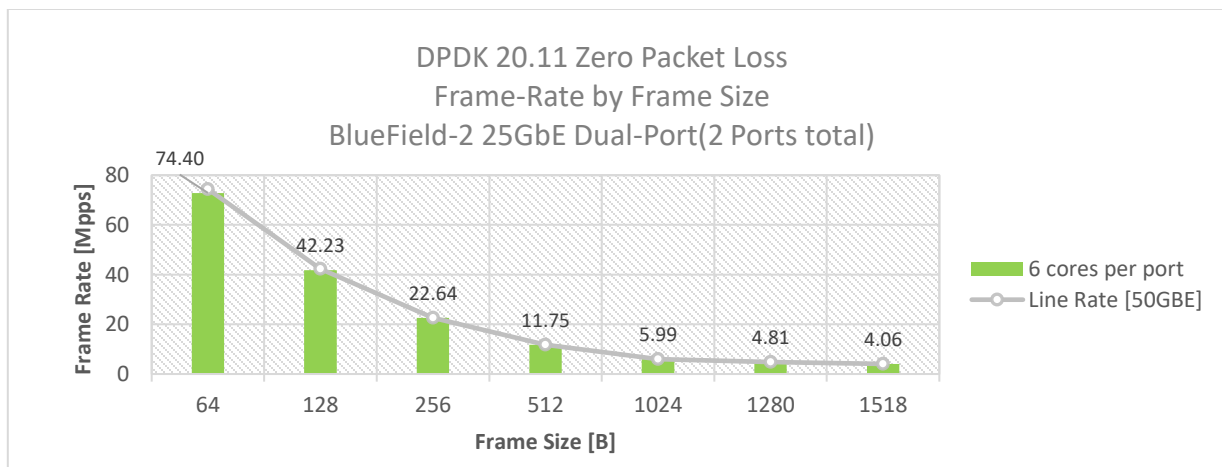
Item	Description
BIOS	<p>1) Workload Profile = "Low Latency";</p> <p>2) Jitter Control = Manual, 3400. (Setting turbo boost frequency to 3.4 GHz)</p> <p>See "Configuring and tuning HPE ProLiant Servers for low-latency applications": hpe.com > Search "DL380 gen10 low latency"</p>
DPU BOOT Settings	<pre>ro crashkernel=auto console=ttyAMA1 console=hvc0 console=ttyAMA0 earlycon=pl011,0x01000000 earlycon=pl011,0x01800000 modprobe.blacklist=mlx5_core,mlx5_ib isolcpus=1-7 nohz_full=1-7 rcu_nocbs=1-7</pre>
DPDK Settings	<p>Compile DPDK using:</p> <pre>meson <build> ; ninja -C <build></pre>
Command Lines	<p>1 Core:</p> <pre>/build/app/dpdk-testpmd -c 0x5 --master-lcore=0 -n 4 -w 03:00.0 -w 03:00.1 --socket-mem=1024 --burst=64 --txq=1 --rxq=1 --rxd=1024 --txd=1024 --mbcache=512 --nb-cores=1 -i -a --rss-udp --port-topology=loop</pre> <p>2 Cores:</p> <pre>/build/app/dpdk-testpmd -c 0x15 --master-lcore=0 -n 4 -w 03:00.0 -w 03:00.1 --socket-mem=1024 --burst=64 --txq=2 --rxq=2 --rxd=1024 --txd=1024 --mbcache=512 --nb-cores=2 -i -a --rss-udp --port-topology=loop</pre> <p>4 Cores:</p> <pre>/build/app/dpdk-testpmd -c 0xab --master-lcore=0 -n 4 -w 03:00.0 -w 03:00.1 --socket-mem=1024 --burst=64 --txq=4 --rxq=4 --rxd=1024 --txd=1024 --mbcache=512 --nb-cores=4 -i -a --rss-udp --port-topology=loop</pre> <p>6 Cores:</p> <pre>/build/app/dpdk-testpmd -c 0x7f --master-lcore=0 -n 4 -w 03:00.0 -w 03:00.1 --socket-mem=1024 --burst=64 --txq=6 --rxq=6 --rxd=1024 --txd=1024 --mbcache=512 --nb-cores=6 -i -a --rss-udp --port-topology=loop</pre>
Other optimizations	<p>a) Flow Control OFF: "ethtool -A \$netdev rx off tx off"</p> <p>b) Memory optimizations: "sysctl -w vm.zone_reclaim_mode=0"; "sysctl -w vm.swappiness=0"</p> <p>c) Move all IRQs to far NUMA node: "IRQBALANCE_BANNED_CPUS=\$LOCAL_NUMA_CPUMAP irqbalance --oneshot"</p> <p>d) Disable irqbalance: "systemctl stop irqbalance"</p> <p>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=3900"</p> <p>f) Set CQE COMPRESSION to "AGGRESSIVE": mlxconfig -d \$PORT_PCI_ADDRESS set CQE_COMPRESSION=1</p> <p>g) Disable Linux realtime throttling: echo -1 > /proc/sys/kernel/sched_rt_runtime_us</p>

16.2 Test Results

Table 44 - Test #14 Results – BlueField-2 25GbE Dual-Port Throughput at Zero Packet Loss

Frame Size (Bytes)	Line Rate [50G] (Mpps)	Frame Rate (Mpps)				Line rate % (6 Cores)
		1 Core	2 Cores	4 Cores	6 Cores	
64	74.40	24	46.24	73.96	74.40	100.00
128	42.23	23.81	42.06	42.18	42.23	100.00
256	22.64	22.60	22.62	22.64	22.64	100.00
512	11.75	11.74	11.75	11.75	11.75	100.00
1024	5.99	5.99	5.99	5.99	5.99	100.00
1280	4.81	4.81	4.81	4.81	4.81	100.00
1518	4.06	4.06	4.06	4.06	4.06	100.00

Figure 28 - Test #14 Results – BlueField-2 25GbE Dual-Port Throughput at Zero Packet Loss



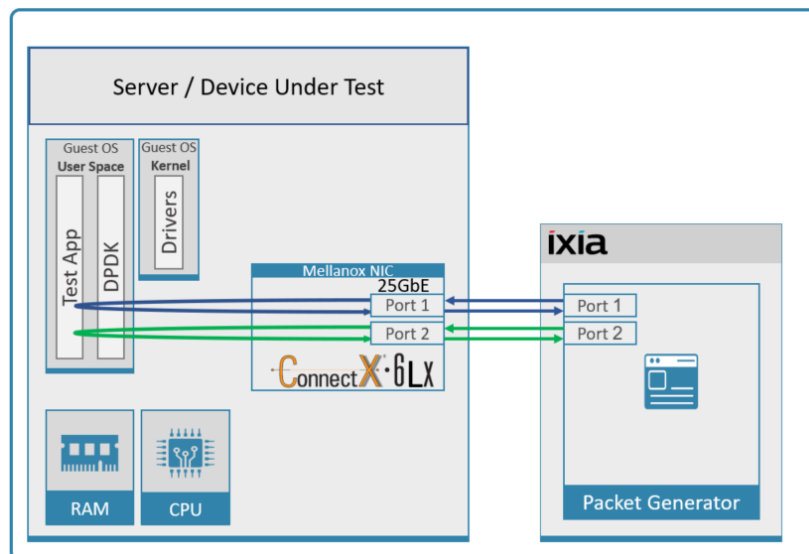
17 Test#15 Mellanox ConnectX-6 Lx 25GbE Throughput at Zero Packet Loss (2x 25GbE)

Table 45 - Test #15 Setup

Item	Description
Test #8	Mellanox ConnectX-6 Lx 25GbE Dual-Port Throughput at zero packet loss
Server	HPE ProLiant DL380 Gen10
CPU	Intel(R) Xeon(R) Platinum 8168 CPU @ 2.70GHz 24 CPU cores * 2 NUMA nodes
RAM	384GB: 6 * 32GB DIMMs * 2 NUMA nodes @ 2666MHz
BIOS	U30 rev. 1.36 (02/15/2018)
NIC	One MCX631102AN-ADAT, ConnectX-6 Lx EN adapter card, 25GbE, Dual-port SFP28, PCIe 4.0 x8, No Crypto
Operating System	Ubuntu 20.04.2 LTS (Focal Fossa)
Kernel Version	5.4.0-65-generic.x86_64
GCC version	gcc (Ubuntu 9.3.0-17ubuntu1~20.04) 9.3.0
Mellanox NIC firmware version	22.29.2002
Mellanox OFED driver version	MLNX_OFED_LINUX-5.2-2.2.0.0
DPDK version	21.02
Test Configuration	1 NIC, 2 ports; Each port receives a stream of 8192 IP flows from the IXIA Each port has 4 queues assigned for a total of 8 queues 1 queue assigned per logical core with a total of 8 logical cores

The Device Under Test (DUT) is made up of the HPE server and the Mellanox ConnectX-6 Lx Dual-Port NIC. The DUT is connected to the IXIA packet generator which generates traffic towards the ConnectX-6 Lx NIC. The ConnectX-6 Lx data traffic is passed through DPDK to the test application **l3fwd** and is redirected to the opposite direction on the same port. IXIA measures throughput and packet loss.

Figure 29 - Test #15 Setup – Mellanox ConnectX-6 Lx 25GbE Dual-Port connected to IXIA



17.1 Test Settings

Table 46 - Test #15 Settings

Item	Description
BIOS	<p>1) Workload Profile = "Low Latency";</p> <p>2) Jitter Control = Manual, 3400. (Setting turbo boost frequency to 3.4 GHz)</p> <p>See "Configuring and tuning HPE ProLiant Servers for low-latency applications": hpe.com > Search "DL380 gen10 low latency"</p>
BOOT Settings	<pre>isolcpus=0-23 intel_idle.max_cstate=0 processor.max_cstate=0 intel_pstate=disable nohz_full=0-23 rcu_nocbs=0-23 rcu_nocb_poll default_hugepagesz=1G hugepagesz=1G hugepages=64 audit=0 nosoftlockup idle=poll</pre>
DPDK Settings	<p>Compile DPDK using: <code>meson <build> -Dexamples=l3fwd ; ninja -C <build></code></p> <p>During testing, l3fwd was given real-time scheduling priority.</p>
L3fwd settings	<p>Updated values /l3fwd/l3fwd.h:</p> <pre>#define RTE_TEST_RX_DESC_DEFAULT 4096 #define RTE_TEST_TX_DESC_DEFAULT 4096 #define MAX_PKT_BURST 64</pre>
Command Line	<pre>./build/examples/dpdk-l3fwd -c 0xff0000 -n 4 -a 12:00.0,mprq_en=1,rxqs_min_mprq=1 -a 12:00.1,mprq_en=1,rxqs_min_mprq=1 --socket-mem=8192 -- -p 0x3 -P -- config='(0,0,23),(0,1,22),(0,2,21),(0,3,20),(1,0,19),(1,1,18),(1,2,17),(1,3,16)' --eth- dest=0,00:52:11:22:33:10 --eth-dest=1,00:52:11:22:33:20</pre>
Other optimizations	<p>a) Flow Control OFF: <code>"ethtool -A \$netdev rx off tx off"</code></p> <p>b) Memory optimizations: <code>"sysctl -w vm.zone_reclaim_mode=0"; "sysctl -w vm.swappiness=0"</code></p> <p>c) Move all IRQs to far NUMA node: <code>"IRQBALANCE_BANNED_CPUS=\$LOCAL_NUMA_CPUMAP irqbalance --oneshot"</code></p> <p>d) Disable irqbalance: <code>"systemctl stop irqbalance"</code></p> <p>e) Change PCI MaxReadReq to 1024B for each port of each NIC: Run <code>"setpci -s \$PORT_PCI_ADDRESS 68.w"</code>, it will return 4 digits ABCD --> Run <code>"setpci -s \$PORT_PCI_ADDRESS 68.w=3936"</code></p> <p>f) Set CQE COMPRESSION to "AGGRESSIVE": <code>mlxconfig -d \$PORT_PCI_ADDRESS set CQE_COMPRESSION=1</code></p> <p>g) Disable Linux realtime throttling: <code>echo -1 > /proc/sys/kernel/sched_rt_runtime_us</code></p>

17.2 Test Results

Table 47 - Test #15 Results – Mellanox ConnectX-6 Lx 25GbE Dual-Port Throughput at Zero Packet Loss

Frame Size (Bytes)	Frame Rate (Mpps)	Line Rate [50G] (Mpps)	% Line Rate
64	74.40	74.40	100.00
128	42.23	42.23	100.00
256	22.64	22.64	100.00
512	11.75	11.75	100.00
1024	5.99	5.99	100.00
1280	4.81	4.81	100.00
1518	4.06	4.06	100.00

Figure 30 - Test #15 Results – Mellanox ConnectX-6 Lx 25GbE Dual-Port Throughput at Zero Packet Loss

