Mellanox Messaging Accelerator (VMA)

Release Notes

Version 6.6.4

www.mellanox.com

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 PRODUCTO(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

Mellanox Technologies, Ltd.
Beit Mellanox
PO Box 586 Yokneam 20692
Israel
www.mellanox.com
Tel: +972 (0)74 723 7200
Fax: +972 (0)4 959 3245

Contents

List of Tables

1 Introduction

These release notes pertain to the Mellanox Messaging Accelerator (VMA) library for Linux, software version 6.6.4.

The VMA library accelerates TCP and UDP socket applications, by offloading traffic from the user-space directly to the network interface card (NIC) or Host Channel Adapter (HCA), without going through the kernel and the standard IP stack (kernel-bypass). VMA increases overall traffic packet rate, reduces latency, and improves CPU utilization.

1.1 System Requirements for VMA 6.6.4

The following table presents the currently certified combinations of stacks and platforms, and supported CPU architectures for VMA 6.6.4.

Table 1: System Requirements

Specification	Value
Network Adapter Cards	ConnectX®-3
Firmware	ConnectX-3 v2.31.5050 or newer
Driver Stack	MLNX-OFED v2.2-1.0.1
Tested Operating Systems and Kernels	RHEL 6 update 3 (2.6.32-279) RHEL 6 update 4 (2.6.32-358) Fedora 19 (3.9.5-301) SLES11 SP1 (2.6.32.12-0.7-default) SLES11 SP2 (3.0.13-0.27-default) SLES11 SP3 (3.0.76-0.11-default) Ubuntu 12.04 (3.2.0-39)
CPU Architecture	x86_64 (Intel Xeon)
Minimum memory requirements	1 GB of free memory for installation 800 MB per process running with VMA
Minimum disk space requirements	1 GB
Transport	Ethernet / InfiniBand / VPI

1.2 VMA Release Contents

Table 2: Release Contents

Item	Description
Binary RPM and DEB packages for 64-bit architecture for Linux distribution	libvma-6.6.4-0-x86_64.rpm libvma-6.6.4-0-x86_64.deb
Documentation	VMA Release Notes VMA Installation Guide VMA User Manual

2 New and Changed Features in 6.6.4

The following describe the main changes and new features in VMA 6.6.4.

- Added support to all Linux Operating Systems supported in MLNX_OFED 2.2-1.0.0
- Improved interrupt driven mode performance
- Added interrupt moderation and adaptive interrupt moderation support
- Added UDP software timestamps support
- Added support for accept4 system call
- Added support for SO_BINDTODEVICE socket option
- Added support for SOCK_NONBLOCK and SOCK_CLOEXEC socket() flags
- Added ring statistics to vma_stats

3 Certified Applications

The VMA library version 6.6.4 was successfully tested, and is certified to work with the applications listed in the following table.

Table 3: VMA Certified Applications

Application	Company / Source	Application Type	Notes
Memcached	Open Source	High-performance, distributed memory object caching system	Version 1.4.17 http://memcached.org/
Redis	Open Source	Advanced key-value store	http://redis.io/
sockperf	Mellanox (Open Source)	Bandwidth and Latency Benchmarking	Version 2.5.231 Included in VMA package. http://code.google.com/p/sockperf/
iperf	NLANR	Bandwidth Benchmarking	Version 2.0.5
netperf	Open Source	Bandwidth and Latency Benchmarking	Version 2.6.0
sfnt	Solarflaire	Bandwidth and Latency Benchmarking	Version 1.4.0
NetPIPE	Open Source	Network Protocol Independent Performance Evaluator	Version 3.7.2
UMS (formerly LBM)	Informatica	Message Middleware Infrastructures	Version 6.5
Opra FeedHandler	NYSE Technologies (WombatFS)	Market Data Infrastructures	Running with WDF/LBM/UMS/RV middleware

4 Resolved Issues

The following describe the issues that have been resolved in VMA 6.6.4:

- Fixed issues that caused multithread deadlocks and races in the system
- Fixed wrong usage of route gateway information
- Fixed buffer management issues and leaks
- Fixed multicast loopback filtering on RX flow

5 Known Issues

The following table describes known issues in VMA 6.6.4, and existing workarounds.

Table 4: VMA Known Issues

Subject	Description	Workaround
High Availability	VMA 6.6.4 supports only bonding Active/Passive mode, and only with `fail_over_mac=1`	N/A
VLAN and High Availability	VLAN on the bond interface does not function properly when bonding is configured with `fail_over_mac=1` due to a kernel bug	**[RedHat ONLY]** Configure bonding over VLAN interfaces instead. This solution is not applicable for SLES OSes
Issues with UDP fragmented traffic reassembly	RX UDP UC and MC traffic in Ethernet and RX UDP UC in InfiniBand with fragmented packages (message size is larger than MTU) is not offloaded by VMA and will pass through the Kernel network stack. There might be performance degradation.	N/A
VMA_TRACELEVEL =4 causes performance degradation	VMA_TRACELEVEL=4 debug mode prints more info, which causes higher latency.	For best performance, run VMA with a lower than 4 VMA_TRACELEVEL value .
Huge-page reserved resources	The system runs out of memory due to huge-page reserved resources.	Use Contiguous Pages instead of Huge Pages to gain performance improvements. The following parameter should be set as follow: `VMA_MEM_ALLOC_TYPE = 1` (this is the default mode).
VMA_PANIC while opening large number of sockets	The following VMA_PANIC will be displayed when there are not enough open files defined on the server: `VMA PANIC : si[fd=1023]:51:sockinfo() failed to create internal epoll (ret=-1 Too many open files)`	Verify that the number of max open FDs (File Descriptors) in the system (ulimit -n) is twice as number of needed sockets. VMA internal logic requires one additional FD per offloaded socket.
There is limited support for fork().	Using fork() in a program is limited to the following conditions: A Parent process can continue running without any limitations on memory access. A Child process can continue running if the child does not access any sockets created by the Parent process.	VMA supports fork() if VMA_FORK=1 (is enabled) and the Mellanox-supported OFED stack is used. In this case the child must not use any sockets created by the parent process. General fork support from kernel 2.6.16 and later is available, provided that

Subject	Description	Workaround
		applications do not use threads. The fork() function is supported, provided that the parent process does not run before the child exits or calls exec(). • You can ensure that the parent process does not run before the child exits, by calling wait(childpid). • You can ensure that the parent process does not run before the child calls exec(), using application-specific means. The Posix system() call is supported.
The VMA application does not exit when you press CTRL-C.	When a VMA-enabled application is running, there are several cases when it does not exit as expected with CTRL-C.	Enable SIGINT handling in VMA, by using: #export VMA_HANDLE_SIGINTR=1
Sockperf over VMA, server re-assign ip- no traffic	VMA does not support network interface or route changes during runtime	N/A
Packets loss occurs when running sockperf with max pps rate	The send rate is higher than the receive rate. Therefore when running one sockperf server with one sockperf client there will be packets loss.	Limit the sender max PPS per receiver capacity. Example below with the following configuration: : • O.S: Red Hat Enterprise • Linux Server release 6.2 (Santiago) Kernel \r on an \m • Kernel: 2.6.32-220.el6.x86_64 • link layer: InfiniBand 56G Ethernet 10G • GEN type: GEN3 • Architecture: x86_64 • CPU: 16 • Core(s) per socket: 8 • CPU socket(s): 2 • NUMA node(s): 2 • Vendor ID: GenuineIntel • CPU family: 6 • Model: 45 • Stepping: 7 • CPU MHz: 2599.926

Subject	Description	Workaround

		`MC 1 socket max pps 3M`
		`MC 10 sockets (select) max pps 1.5M`
		`MC 20 sockets (select) max pps 1.5M`
		`MC 50 sockets (select) max pps 1M`
		`UC 1 socket max pps 2.8M`
		`UC 10 sockets max pps 1.5M`
		`UC 20 sockets max pps 1.5M`
		`UC 50 sockets max pps`

VMA behavior of epoll EPOLLET (Edge Triggered) and EPOLLOUT flags with TCP sockets	VMA behavior of epoll EPOLLET (Edge Triggered) and EPOLLOUT flags with TCP sockets differs between OS and VMA. • VMA - triggers EPOLLOUT event every received ACK (only data, not syn/fin) • OS - triggers EPOLLOUT event only after buffer was full.	N/A
VMA behavior of epoll EPOLLET (Edge Triggered) and EPOLLOUT flags with UDP sockets	VMA will trigger 2 ready events instead of 1 in case of epoll with EPOLLET and EPOLLOUT flags with UDP sockets.	N/A
VMA does not close connections upon process termination	VMA does not close connections (sends FIN) when its own process is terminated (e.g. “CTRL-C”)	N/A
MC traffic with VMA process and non VMA process on the same machine	When non offloaded process joins the same MC address as another VMA process on the same machine, non-offloaded process will not get traffic.	Run both processes with VMA
Epoll with EPOLLONESHOT	Occasionally, Epoll with EPOLLONESHOT does not function properly.	N/A

SFNT-STREAM UDP with poll muxer flag ends with an error on client side	Occasionally, when running UDP SFNT-STREAM client with poll muxer flag, the client side ends with an expected error:
	`ERROR: Sync messages at end of test lost`
	`ERROR: Test failed.`
	This only occurs with poll flag
		Set a higher acknowledgment waiting time value in the `sfnt-stream`.

SFNT-STREAM UDP	Occasionally, SFNT-STREAM UDP	Set a higher acknowledgment

Subject	Description	Workaround
client hanging issue	client hangs when running multiple times.	waiting time value in the `sfnt-stream`.

When using `VMA_TX_MAX_INLINE=0`, `post_send` fails.	When running VMA with `VMA_TX_MAX_INLINE=0`
	The following error will be received:
	`VMA ERROR : qpm[0x16f7660]:448:send() failed post_send (errno=11 Resource temporarily unavailable)`
	`VMA ERROR : qpm[0x16f7660]:450:send() bad_wr info: wr_id=0x10efa358, send_flags=0, addr=0x10c699d0, length=60, lkey=0x46d00, max_inline_data=0`
	In this scenario, the send operation will fail.
		Set the `VMA_TX_MAX_INLINE` value to a smaller message size that the used in the application.

MC loopback in InfiniBand	MC loopback in InfiniBand functions only between 2 different processes. It will not work between threads in the same process.	N/A
Ethernet loopback is not functional between the VMA and the OS	Ethernet loopback functions only if both sides are either off-loaded or not-offloaded.	N/A
Error when running netperf 2.4.4 with VMA	The following error may occur when running netperf TCP tests with VMA: `remote error 107 'Transport endpoint is not connected'`	Use netperf 2.6.0
A packet is not sent if the socket is closed immediately after send()	Occasionally, a packet is not sent if the socket is closed immediately after send() (also for blocking socket)	Wait several seconds after send() before closing the socket
Iomux call with empty sockets	It can take for VMA more time than the OS to return from an iomux call if all sockets in this iomux are empty sockets	N/A
TCP throughput with maximum rate	TCP throughput with maximum rate may suffer from traffic "hiccups".	Set the `mps = 1000000`
Netcat on SLES11 SP1	Netcat with VMA on SLES 11 SP1 does not function.	N/A
Issues with performance with some multi-threaded applications	Sharing of HW resources between the different working threads might cause lock contentions which can affect performance.	For best performance with VMA, use multi-processing, each with a single thread. This will best allocate and separate the HW resources between the working threads, and minimize contention.

Subject	Description	Workaround
Segmentation fault on NetPIPE exit.	Known NetPIPE bug - Netpipe is trying to access read-only memory.	Upgrade to NetPIPE 3.7 or later.
When exiting, VMA logs errors when the VMA_HANDLE_SIGINTR is enabled.	If VMA runs when VMA_HANDLE_SIGINTR is enabled, an error message might be written upon exiting.	Ignore the error message, or run VMA with VMA_HANDLE_SIGINTR disabled.
VMA ping-pong latency degradation as PPS is lowered	VMA suffers from high latency in low message rates.	Use VMA_RX_POLL=-1
No support for direct broadcast	VMA does not support broadcast traffic.	Use libvma.conf to pass broadcast through OS
There is no non-valid pointer handling in VMA	Directing VMA to access non-valid memory area will cause a segmentation fault.	N/A
First connect/send operation might take more time than expected	VMA allocates resources on the first connect/send operation, which might take up to several tens of milliseconds.	N/A
Calling select() after shutdown (write) returns socket ready to write, while select() is expected to return timeout	Calling select upon shutdown of socket will return “ready to write” instead of timeout.	N/A
VMA does not raise sigpipe	VMA does not raise sigpipe in connection shutdown.	N/A
When there are no packets in the socket, it takes longer to return from the read call	VMA polls the CQ for packets; if no packets are available in the socket layer, it takes longer.	N/A
Select with more than 1024 sockets is not supported		Compile VMA with SELECT_BIG_SETSIZE defined.

6 Related Documentation

- Mellanox Messaging Accelerator (VMA) Library for Linux User Manual (DOC-00393)
- Mellanox VMA Installation Guide (DOC-10055)
- Performance Tuning Guidelines for Mellanox Network Adapters (DOC 3368) – available at www.mellanox.com

7 Performance Data

7.1 Ethernet Performance Data

The performance envelope of the VMA library is described in the following sections.

7.1.1 VMA Benchmark Configuration with ConnectX®-3

The following table describes the setup for the VMA library benchmarking.

Table 5: Benchmark Setup

Specifications	Details
CPU	Intel(R) Xeon(R) CPU E5-2687W 0 @ 3.10GHz
CPU Architecture	x86_64
I/O Expansion slots	PCI Express Gen 3
Network Adapter Type	ConnectX®-3
Line Rate	40 Gb/sec Ethernet
OS	Red Hat, 2.6.32-358.el6.x86_64
OFED Software Stack	MLNX_OFED 2.1-1.0.0
Connectivity	Back to back
Netperf	2.6.0

7.1.2 VMA UDP Unicast Throughput (Netperf UDP STREAM) Benchmark Using ConnectX®-3

The following table shows the UDP unicast throughput benchmark results from the test application, *netperf UDP STREAM*.

Table 6: VMA UDP Unicast Throughput Benchmark Results

VMA 6.6.4 with ConnectX-3 PCI Gen3	
Message Size [Bytes]	**Bandwidth [Gb/s]**
16	0.527
32	1.057
64	1.981
128	3.690
256	7.597
512	14.092
1024	24.813
1472	33.639

7.1.3 VMA TCP Unicast Throughput (Netperf TCP STREAM) Benchmark Using ConnectX®-3

The following table shows the TCP unicast throughput benchmark results from the test application, *netperf TCP STREAM*.

Table 7: VMA TCP Unicast Throughput Benchmark Results

VMA 6.6.4 with ConnectX-3 PCI Gen3	
Message Size [Bytes]	**Bandwidth [Gb/s]**
16	1.713

VMA 6.6.4 with ConnectX-3 PCI Gen3	
32	3.233
64	5.828
128	9.532
256	14.323
512	18.681
1024	21.913
1460	24.880

7.1.4 VMA UDP Unicast Latency (Netperf UDP RR) Benchmark with ConnectX®-3

The following table shows the UDP unicast latency benchmark results from the test application, *netperf UDP RR*. These results are the RTT/2 for a ping-pong test.

Table 8: VMA UDP Unicast Latency Benchmark Results

VMA 6.6.4 with ConnectX-3 PCI Gen3	
Message Size [Bytes]	**Latency [usec]**
16	1.125
32	1.175
64	1.250
128	1.310
256	1.640
512	1.840
1024	2.155

7.1.5 VMA TCP Latency (Netperf TCP RR) Benchmark with ConnectX®-3

The following table shows the TCP latency benchmark results from the test application, *netperf TCP RR*. These results are the RTT/2 for a ping-pong test.

Table 9: VMA TCP Latency Benchmark Results

VMA 6.6.4 with ConnectX-3 PCI Gen3	
Message Size [Bytes]	**Latency [usec]**
16	1.650
32	1.700
64	1.755
128	1.850
256	2.090
512	2.310
1024	2.665

7.2 InfiniBand Performance Data

7.2.1 VMA Benchmark Configuration with ConnectX®-3

The following table describes the setup for the VMA library benchmarking.

Table 10: Benchmark Setup

Specifications	Details
CPU	Intel(R) Xeon(R) CPU E5-2687W 0 @ 3.10GHz
CPU Architecture	x86_64
I/O Expansion slots	PCI Express Gen 3
Network Adapter Type	ConnectX®-3
Line Rate	56 Gb/sec InfiniBand
OS	Red Hat, 2.6.32-358.el6.x86_64
OFED Software Stack	MLNX_OFED 2.1-1.0.0
Connectivity	Back to back
netperf	2.6.0

7.2.2 VMA UDP Unicast Throughput (Netperf UDP STREAM) Benchmark Using ConnectX®-3

The following table shows the UDP unicast throughput benchmark results from the test application, *netperf UDP STREAM*.

Table 11: VMA UDP Unicast Throughput Benchmark Results

VMA 6.6.4 with ConnectX-3 PCI Gen3	
Message Size [Bytes]	**Bandwidth [Gb/s]**
16	0.539
32	1.045
64	2.054
128	3.670
256	7.259
512	13.353
1024	23.817
1472	31.469

7.2.3 VMA TCP Unicast Throughput (Netperf TCP STREAM) Benchmark Using ConnectX®-3

The following table shows the TCP unicast throughput benchmark results from the test application, *netperf TCP STREAM*.

Table 12: VMA TCP Unicast Throughput Benchmark Results

VMA 6.6.4 with ConnectX-3 PCI Gen3	
Message Size [Bytes]	**Bandwidth [Gb/s]**
16	1.715
32	3.251
64	5.779
128	8.255
256	14.052
512	18.123
1024	21.311
1460	23.907

7.2.4 VMA UDP Unicast Latency (Netperf UDP RR) Benchmark with ConnectX®-3

The following table shows the UDP unicast latency benchmark results from the test application, *netperf UDP RR*. These results are the RTT/2 for a ping-pong test.

Table 13: VMA UDP Unicast Latency Benchmark Results

VMA 6.6.4 with ConnectX-3 PCI Gen3	
Message Size [Bytes]	**Latency [usec]**
16	1.145
32	1.175
64	1.220
128	1.320
256	1.700
512	1.845
1024	2.125

7.2.5 VMA TCP Latency (Netperf TCP RR) Benchmark with ConnectX®-3

The following table shows the TCP latency benchmark results from the test application, *netperf TCP RR*. These results are the RTT/2 for a ping-pong test.

Table 14: VMA TCP Latency Benchmark Results

VMA 6.6.4 with ConnectX-3 PCI Gen3	
Message Size [Bytes]	**Latency [usec]**
16	1.650
32	1.675
64	1.755
128	1.860
256	2.140
512	2.270
1024	2.590

8 Submitting a Service Request

The Mellanox Support Center is at your service. You may access Warranty Service through our Web Request Form by using the following link: http://www.mellanox.com/content/pages.php?pg=support_index.