WHITE PAPER

June 2011

Achieving and Measuring the Lowest Application to Wire Latency for High Frequency Trading

Benchmark Report

Executive Summary

This benchmark paper describes in detail the comparison between two leading technologies for minimizing application to wire latency for high frequency trading. Detailed latency measurements performed on identical server/OS platforms prove that Mellanox ConnectX family of Ethernet NICs together with VMA messaging acceleration software provide the fastest path from application to wire (and wire to application), with latencies as low as 2us end-to-end (application-to-application). The report goes on to show how these latencies are sustained at high message rates guaranteeing determinism at scale.

Introduction

As the high frequency trading market continues to grow and become more and more competitive, so does the race to zero latency. Aside from the ongoing need to be faster than the competition, additional trends have recently come into play increasing the drive for lower application latency:

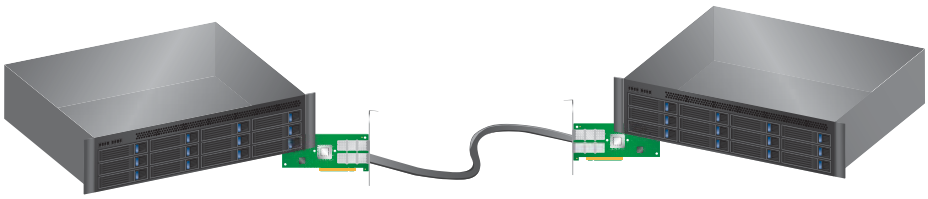
1. The ability and availability of latency measurement tools from mupltiple vendors able to provide sub-microsecond accuracy, makes latency reductions in the microsecond range interesting and applicable
2. Emerging regulations requiring pre-trade risk management increase the need to maintain each hop of latency (on a server or on a switch) as low as possible

In light of these market requirements Mellanox has performed a full, detailed benchmark comparison of the leading low-latency "application to wire" solutions. Such a solution must comprise of two basic components:

1. Hardware – low-latency adapter (NIC)
2. Software – low-latency communication path from application to NIC (typically via RDMA or OS bypass). The benchmark described here leverages OS bypass technologies which provide the greatest latency benefit without requiring modifications to application code

Setup

The setup was designed to measure application to wire latencies and isolate them from other network latencies. As such, the servers were connected directly, with no intermediate switching/routing layer. In order to achieve a true "apples to apples" comparison, the same servers, running the same operating systems were used for both setups. Below is a description of the two systems under test (SUTs).



Mellanox vs. Solarflare Systems Under Test		
	SUT 1 - Solarflare	SUT 2 - Mellanox
Latency Measurement Tool	Sockperf	
Acceleration Software	OpenOnload (OO) 20101111-u1 Built: Apr 10 201 1 17:58:34 (release)	VMA v5.0.4.0
Low Level Driver	SFC 3.0.8.2221	OFED 1.5.3
Adapter Firmware	3.0.7.2206	2.8.0000
Adapter (NIC)	SFN5122F	ConnectX-2 EN
Operating System	RHEL 6.0	
Memory	24GB	
Processor	Intel X5670 2.93Ghz (12 cores Hyper-threading)	
Server	2 x HP ProLiantSL390s G7	

Sockperf Latency Measurement Utility

Sockperf is an open source benchmarking utility targeted at high performance networking systems. It provides discrete packet latency measurement at sub-nanosecond resolution under loads of millions of packets per second. Sockperf covers most of the available socket API calls, and provides comprehensive logging and analysis capabilities.

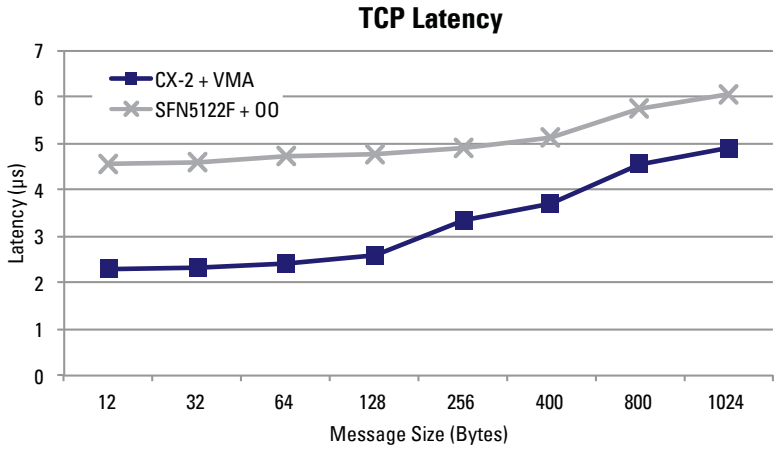
For access to Sockperf code and documentation go to: http://code.google.com/p/sockperf/

Results

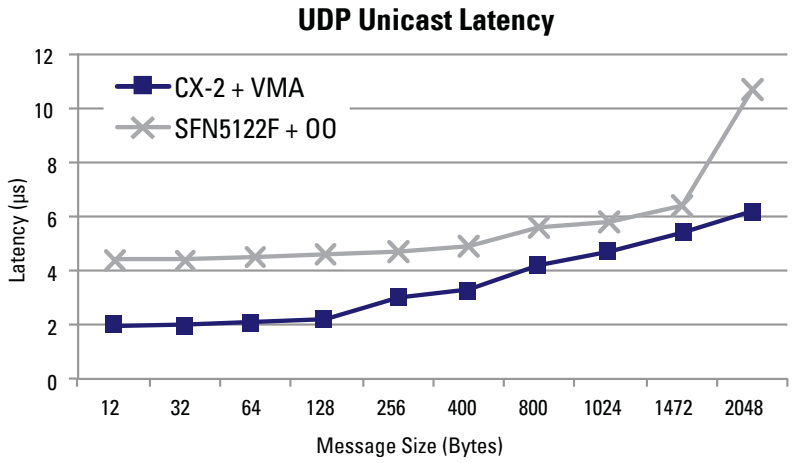
While measuring pure latency has some interest, benchmarks that attempt to simulate real world conditions, should measure how latency is sustained under various conditions. The benchmark tests described in this report measured how latency is affected by variations in Message size and Message rate.

Latency vs. Message Size

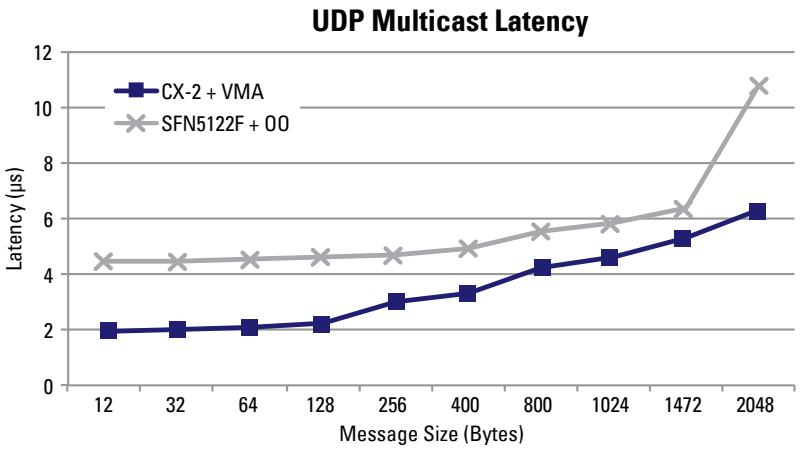
As larger messages require more processing time both from CPUs (allocating buffers) and from networks (serializing data onto the wire), it is expected for increasing message sizes to have some impact on latency, but not necessarily a large one. Though typical message sizes in high frequency trading do not exceed 300 Bytes, the results below include a full comparison of latencies up to 2048 Bytes, showing lower latencies for the Mellanox solution on all message sizes with all transport protocols.



TCP Latency Measurement		
Message Size	**ConnectX-2 EN + VMA**	**SFN5122F + 00**
12	2.29	4.555
32	2.304	4.572
64	2.393	4.723
128	2.584	4.769
256	3.335	4.903
400	3.699	5.139
800	4.554	5.74
1024	4.915	6.033
1472	7.246	7.53
2048	7.431	7.815



UDP Unicast Latency Measurement		
Message Size	**ConnectX-2 EN + VMA**	**SFN5122F + 00**
12	1.951	4.418
32	1.975	4.435
64	2.08	4.532
128	2.246	4.62
256	3.012	4.719
400	3.323	4.931
800	4.207	5.54
1024	4.645	5.804
1472	5.347	6.401
2048	6.235	10.714



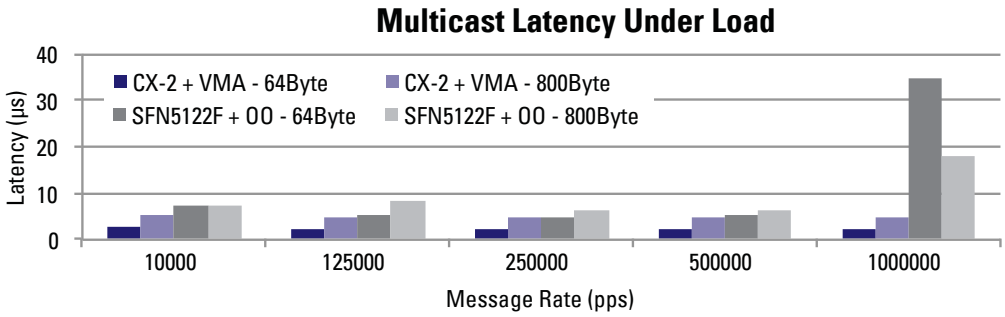
UDP Multicast Latency Measurement		
Message Size	**ConnectX-2 EN + VMA**	**SFN5122F + OO**
12	1.935	4.434
32	1.981	4.445
64	2.084	4.545
128	2.243	4.61
256	3.022	4.724
400	3.333	4.925
800	4.25	5.531
1024	4.643	5.819
1472	5.334	6.412
2048	6.334	10.751

Latency vs. Message Rate

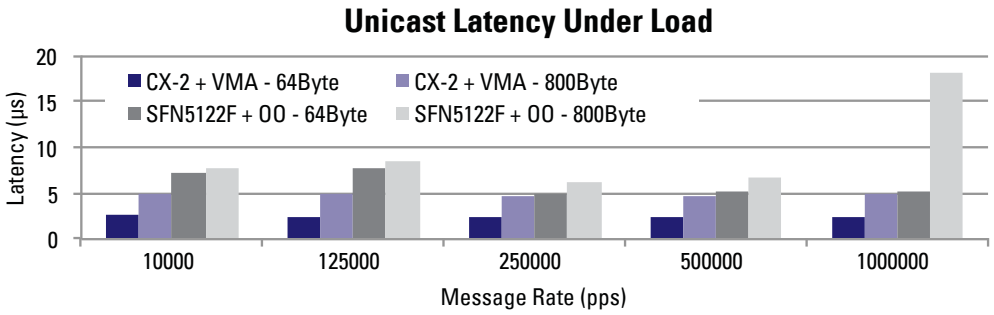
Message rates are expected to be independent of messaging latency to an extent. If the sending CPU is able to generate messages sufficiently fast, and the receiving CPU is fast enough to process every incoming message at the moment it is notified of its existence, then the measured latency consists, essentially, of only the time it takes to send, propagate and receive an individual message.

If, however, the desired message rate causes either CPU (usually the receiving CPU) to be overburdened, additional queuing latency will be added in order to make higher rates possible.

The measurements below show how latency varies as a function of message rate for two different message sizes – a large one (800 Bytes) and a small one (64 Bytes) – and for two different transport protocols unicast (UDP) and multicast. The results clearly show that the Mellanox solution maintains consistent latencies under all message rates, while the Solarflare solution experiences higher variations between the different message rates, and significant latency spikes when reaching a million packets per second.



Multicast Latency Under Load		
Message Size	**Rate**	**Latency (us)**
ConnectX-2 EN + VMA - 64Byte		
64	10000	2.587
64	125000	2.439
64	250000	2.37
64	500000	2.294
64	1000000	2.267
ConnectX-2 EN + VMA - 800Byte		
800	10000	5.153
800	125000	4.769
800	250000	4.607
800	500000	4.621
800	1000000	4.85
SFN5122F + OO - 64Byte		
64	10000	7.458
64	125000	5.313
64	250000	4.989
64	500000	5.096
64	1000000	34.486
SFN5122F + OO - 800Byte		
800	10000	7.558
800	125000	8.277
800	250000	6.079
800	500000	6.303
800	1000000	18.056



Unicast Latency Under Load		
Message Size	**Rate**	**Latency (us)**
ConnectX-2 EN + VMA - 64Byte		
64	10000	2.607
64	125000	2.448
64	250000	2.399
64	500000	2.375
64	1000000	2.327
64	2500000	3.076
ConnectX-2 EN + VMA - 800Byte		
800	10000	4.835
800	125000	4.843
800	250000	4.662
800	500000	4.714
800	1000000	4.994
SFN5122F + OO - 64Byte		
64	10000	7.337
64	125000	7.758
64	250000	4.9
64	500000	5.224
64	1000000	5.291
SFN5122F + OO - 800Byte		
800	10000	7.802
800	125000	8.382
800	250000	6.229
800	500000	6.605
800	1000000	18.104

Summary

Increased competition in the high frequency trading market along with stronger regulations is driving the demand for lower application latency. The detailed benchmark described in this document clearly shows the leadership of the Mellanox ConnectX family of adapters along with VMA acceleration software, outperforming competitive solutions by 2X, with no modifications required to customer applications. These advantages are achieved through superior micro-processing at the adapter layer along with years of experience fine-tuning software accelerators to the needs of the high frequency trading market.

Product Descriptions

About Mellanox

Mellanox Technologies is a leading supplier of end-to-end InfiniBand and Ethernet connectivity solutions and services for servers and storage. Mellanox products optimize data center performance and deliver industry-leading bandwidth, scalability, power conservation and cost-effectiveness while converging multiple legacy network technologies into one future-proof architecture. The company offers innovative solutions that address a wide range of markets including HPC, enterprise, mega warehouse data centers, cloud computing, Internet and Web 2.0.

Mellanox ConnectX® Ethernet Network Interface Cards (NIC)

Mellanox ConnectX EN family of 10 and 40 Gigabit Ethernet Network Interface Cards (NIC) deliver high-bandwidth and industry leading Ethernet connectivity for performance-driven server and storage applications in Enterprise Data Centers, Web 2.0, High-Performance Computing, and Embedded environments. Clustered databases, web infrastructure, and high frequency trading are just a few applications that will achieve significant throughput and latency improvements resulting in faster access, real-time response and more users per server. ConnectX improves network performance by increasing available bandwidth while decreasing the associated transport load on the CPU especially in virtualized server environments.

VMA

VMA is a dynamically-linked user-space Linux library for accelerating unicast or multicast messaging traffic. Applications that utilize standard BSD sockets use the library to offload network processing from a server's CPU. The traffic is passed directly to the ConnectX EN 10GigE Network Interface Cards (NIC) from the application user space, bypassing the kernel and IP stack and thus minimizing context switches, buffer copies and interrupts resulting in extremely low latency.

Looking Ahead

As servers continue to advance with faster clock rates, faster access to memory, more cores, and faster I/O busses, Mellanox end-to-end interconnect solutions will continue to be the first ones to provide the required networking speed, leading the way to new technologies, such as 40/100GbE, 56/100Gb/s InfiniBand, and beyond.

Mellanox TECHNOLOGIES

350 Oakmead Parkway, Suite 100, Sunnyvale, CA 94085
Tel: 408-970-3400 • Fax: 408-970-3403
www.mellanox.com