WHITE PAPER

July 2012

HP and Mellanox Benchmarking Report for Ultra Low Latency 10 and 40Gb/s Ethernet Interconnect

Benchmark Report

Executive Summary

Half roundtrip latencies on the Netperf benchmark of 1.43usec (UDP) and 2.08usec (TCP) for 64-byte messages, without rewriting low latency applications, is swinging the attention of the High Frequency Trading (HFT) community back to the integrated HP/Mellanox Technologies® solution of ProLiant Gen8 servers with ConnectX® FlexibleLOMs.

This breakthrough is achieved due to the confluence of architectural enhancements by HP, Intel, and Mellanox described in this report. The new TCP kernel bypass capability in VMA v6.1 with ConnectX-3 means that the existing DL360/DL380 G7 servers can be upgraded with this new adapter for an immediate latency reduction. But for the ultimate performance machine, the HP/Mellanox Gen8/ ConnectX-3 platform is shown in this report to have the best performance needed to make the trade.

The Four New 2012 Technologies Evaluated In This Benchmark

The global securities market continues its relentless drive to lower the latencies associated with high frequency trading strategies in order to gain the competitive advantage. Trading infrastructure solutions have the difficult task of delivering the lowest possible latency with minimum jitter while sustaining high message rate performance even under unpredictable traffic spikes.

The complete solution stack continues to be optimized by leading firms, including minimizing physical distances, optimizing trading software and adopting the newest hardware technology. As competition intensifies, trading applications are increasingly implemented with careful consideration of hardware design, ensuring that data is available in a processor's cache and server I/O is implemented using kernel bypass and RDMA techniques.

In 2012, four new advancements from HP, Intel and Mellanox have combined to further enable solutions to address this challenge of improving latency and maximizing throughput, while remaining within the cost envelope of industry-standard systems.

First, Intel's new E5-2600 family of processors (aka Sandy Bridge) brings important architectural changes which, when leveraged by the application, deliver faster computational and I/O performance. Computational examples include 33% more cores and memory channels, 66% larger L3 cache with 5x the bandwidth, 20% faster memory, and improved Turbo Boost operation. I/O examples include 2x bandwidth

expansion with PCIe Gen3, lower latency with on-die PCIe controllers, and Data Direct I/O Technology (DDIO) to bring data from a network adapter directly into cache. Taken together, these architectural improvements in Sandy Bridge more than compensate for the -16% reduction in the top bin processor clock speed for most HFT applications.

Second, HP's ProLiant Gen8 servers deliver breakthrough technologies relevant to ultra low latency systems.

- Highest frequency Intel Xeon E5-2600 processors in the ProLiant DL380p and DL360p models.
- Top speed 1600 MT/s Memory available in dual rank RDIMMs (and in even lower latency UDIMMs when only 1 DIMM/channel need be populated for up to 64GB of memory).
- Mellanox ConnectX-3 technology is available in three form factors: PCIe card, FlexibleLOM Network adapter card for rack mount servers, and C-Class BladeSystem mezzanine card.
- BIOS tuning options to disable System Management Interrupts and minimize system jitter.
- SmartArray P420i now 2X cache size; 6X performance with SSDs; 2X # of drives supported vs. G7
- Gen8 quality, reliability and manageability innovations, including NIC temperature sensors and agentless out-of-band data collection linked to the iLO4 management processor.

Third, Mellanox's ultra low-latency ConnectX-3 Virtual Protocol Interconnect (VPI) I/O network adapters for 10/40 Gigabit Ethernet and/or InfiniBand are factory integrated and supported by HP in these servers for global deployment. When linked to external 10/40 Gigabit Ethernet and InfiniBand switches from Mellanox with HP/Mellanox reliable cables, a complete HP/Mellanox end-to-end fabric is available for best-in-class low latency. The hardware-based stateless offload and flow steering engines in ConnectX-3 adapters reduce the CPU overhead of IP packet transport, freeing more processor cycles to work on the application.

Finally, to maximize the benefits of low latency networking hardware for the end user application, the Mellanox Messaging Accelerator (VMA) Linux library has been enhanced in 2012. VMA has long been capable of performing socket acceleration via OS/kernel bypass for UDP unicast and multicast messages (typically associated with market data feeds from Exchanges), without rewriting the applications. In version 6 this dynamically-linked user-space library can now also accelerate any application using TCP messages (typically associated with orders/acknowledgements to and from the Exchanges), passing TCP traffic directly from the user-space application to the network adapter. Bypassing the kernel and IP stack interrupts delivers extremely low-latency because context switches and buffer copies are minimized. The result is high packet-per-second rates, low data distribution latency, low CPU utilization and increased application scalability.

This benchmark report quantifies the benefits of these four technologies with user-space benchmark applications and demonstrates that trading firms and exchanges can now choose one optimized network adapter and method of accelerating their applications across Ethernet and/or InfiniBand for all messaging protocols to achieve a competitive advantage.

Benchmark Objective

The objective of this benchmark is to provide trading infrastructure designers with accurate, specific data around the performance that can be expected with the above mentioned new technologies, using industry standard benchmarks (Netperf, Sockperf) running on G7 and Gen8 HP ProLiant Servers with Mellanox ConnectX-2 and ConnectX-3 adapters using 10 and 40GbE connectivity. Performance is measured with and without the VMA acceleration software to show how this application-transparent kernel bypass library can dramatically increase the value of fast networking hardware. Our intent is to provide the Financial Services community with credible, measured results on the basic metrics of performance on relevant server and I/O configurations, which enable them to evaluate the benefit of upgrading to these new technologies.

Testing Methodology

The tests were conducted on two HP ProLiant servers with Mellanox ConnectX adapter cards connected back to back. Roundtrip latencies were measured so that the same server clock did all timestamps and results were divided in half and reported as "half roundtrip" to represent the latency of a user space benchmark application to pass a message out through a ConnectX adapter.

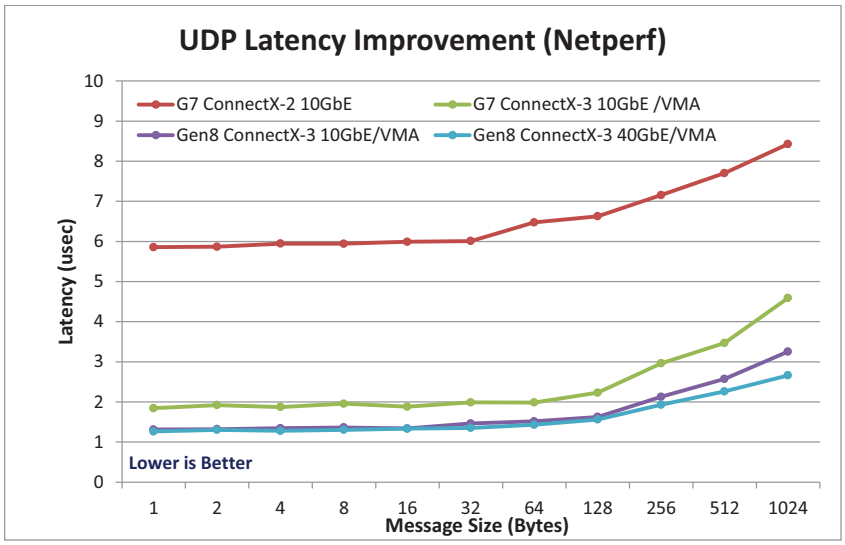
In all the measurements below Mellanox ConnectX-2 or ConnectX-3 adapter cards are used. The focus of the benchmarking is to measure latency, bandwidth and message rate. Both TCP and UDP transport results are measured, as well as the effect of port speeds at 10GbE and 40GbE. As a baseline, G7 servers with ConnectX-2 cards running 10GbE speeds are representative of the previous shipping configurations. The applications used are Netperf and Sockperf.

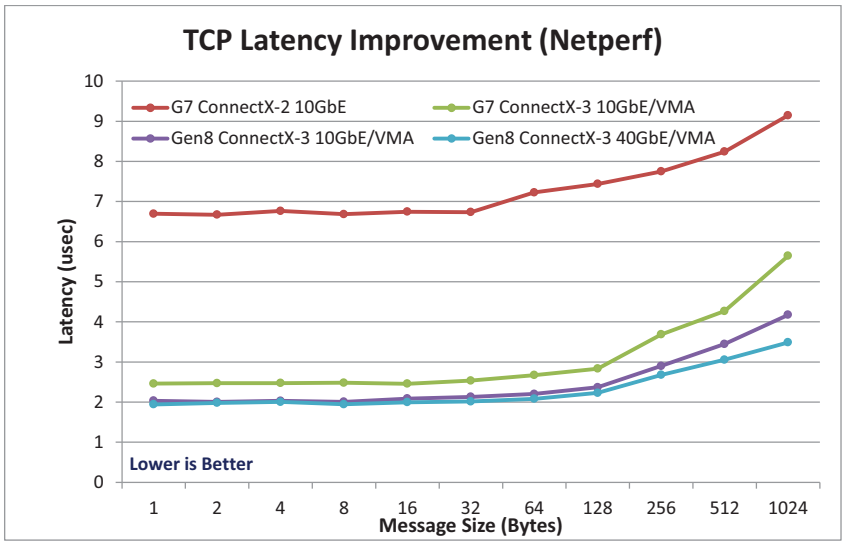
Netperf is a commonly used benchmark for measuring throughput and latency for different types of networking. These tests used v2.5.0 primarily; the main usage in this report is for measuring sockets API networking over 10GbE and 40GbE. See *www.netperf.org*.

Sockperf is another networking benchmark utility over socket API for measuring throughput and latency. These tests used v2.5.156, and it has the benefit of measuring the latency of every packet, allowing histograms of the various percentiles of the packets' latencies. See: *http://code.google.com/p/sockperf/*

The specific command lines used can be found in Appendix 2.

Netperf UDP/TCP Latency

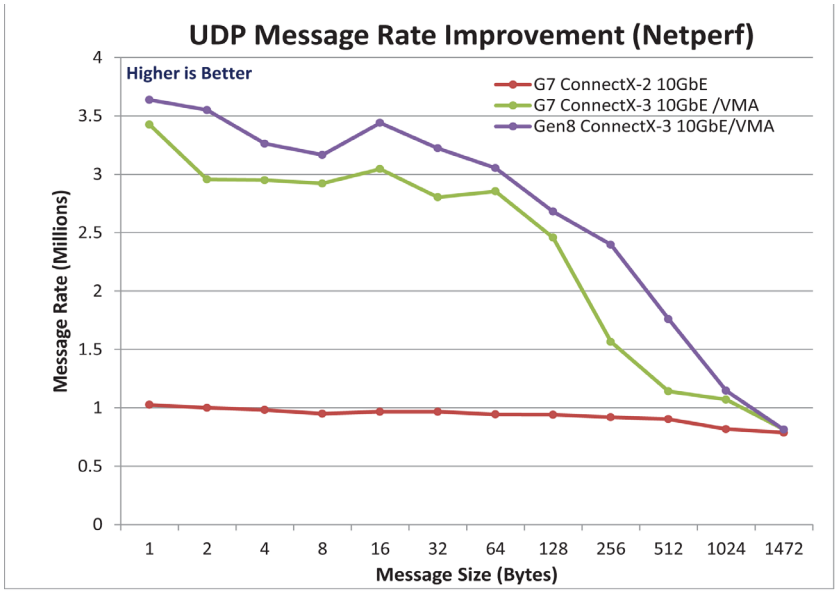




The key points to be learned from the Netperf latency charts are as follows:

- 64-Byte latencies of 1.43usec (UDP) and 2.08usec (TCP) were achieved with Gen8 servers, ConnectX-3 and VMA.
- Trends are similar for UDP and TCP, though UDP latency is lower by more than half a microsecond.
- The move from ConnectX-2 to ConnectX-3/VMA reduces TCP latency by about 4usec. This is mostly a result of VMA OS bypass.
- The move from G7/ConnectX-3/VMA to Gen8/ConnectX-3/VMA reduces latency by 0.5 - 1usec. This is mainly attributed to the move to the embedded PCIe I/O controller with PCIe Gen3 and Data Direct I/O capabilities of the new Intel Romley platform.
- 10GbE and 40GbE experience similar latencies on small message sizes, while on larger messages (200B and above) the latency benefit from using 40GbE increases as message size increases.

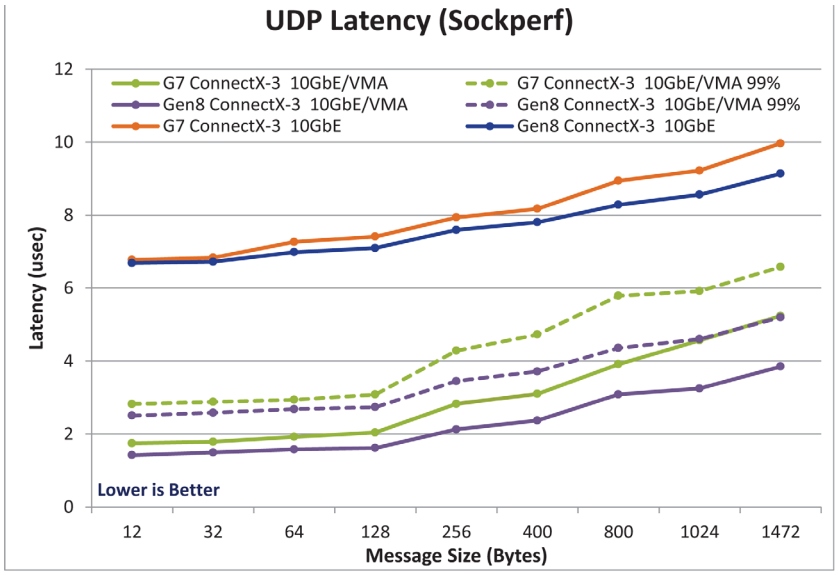
Netperf UDP Message Rate

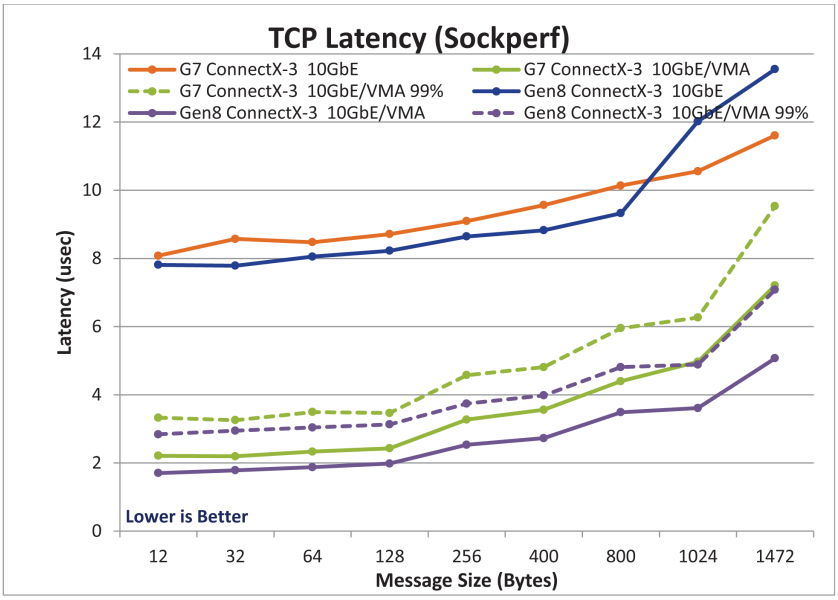


The key points to be learned from the Netperf message rate chart are as follows:

- Maximum message rates of 3.64 million messages per second were achieved with Gen8 servers, ConnectX-3 and VMA.
- The move from ConnectX-2 to ConnectX-3/VMA increases the message rate by 2 million messages per second. The message rate is mostly bound by CPU processing. Using VMA significantly reduces the amount of CPU involvement in the processing of each message.
- The move from G7/ConnectX-3/VMA to Gen8/ConnectX-3/VMA increases the message rate by an additional 0.5 - 1 million messages per second. This is mainly attributed to the move to the embedded PCIe I/O controller with PCIe Gen3 and Data Direct I/O capabilities of the new Intel Romley platform.

Sockperf UDP/TCP Latency (showing 99 percentile)

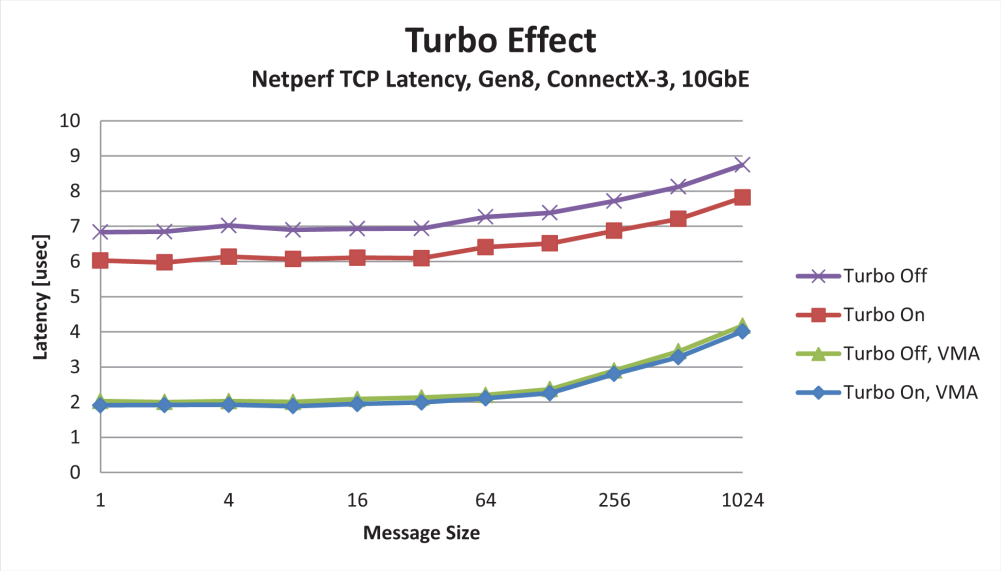




The key points to be learned from the Sockperf 99% latency chart are as follows:

- 64-Byte latencies of 1.58usec (UDP) and 1.875usec (TCP) were achieved with Gen8 servers, ConnectX-3 and VMA.
- Trends are similar for UDP and TCP, though UDP latency is lower by more than half a microsecond.
- The move to VMA not only significantly reduces average latency, but also 99% is significantly lower than average latency without VMA. Using VMA significantly reduces the amount of CPU involvement in the processing of each message, therefore reducing variance originated in the CPU/OS.
- The move from G7/ConnectX-3/VMA to Gen8/ConnectX-3/VMA not only reduces average latency, but also 99% latency, by more or less the same rate. This is mainly attributed to the move to the embedded PCIe I/O controller with PCIe Gen3 and Data Direct I/O capabilities of the new Intel Romley platform.

TurboBoost Latency Effect



The key points to be learned from the Turbo Effect chart are as follows:

- When using Turbo Boost to raise the processor clock from 2.9GHz to 3.3GHz, a latency gain is observed when the messages go through the OS/kernel stack.
- When VMA is used to bypass the OS/kernel and send the socket calls directly to the adapter, the Turbo effect is, logically, negligible.
- We recognize that that Turbo Boost may offer significant gains to the computational trading algorithms themselves, but that is not the subject of this benchmark report.

Conclusion

HP customers with existing DL360/DL380 G7 servers in production can gain an immediate benefit in TCP latency reduction by deploying HP ConnectX-3 PCIe cards and linking their applications to the VMA library. Demo copies of VMA can be used to prove this gain.

All firms seeking the fastest solution in the world should evaluate the performance and reliability benefits of the HP/Mellanox Gen8 solution to confirm the results in this report and determine that all of your requirements are met. We are ready to address any questions and/or issues for your success.

With Mellanox VMA v6.1 and upcoming v6.3 release now fulfilling both the UDP and TCP message acceleration requirements of financial services companies, HP is adding VMA to HP's price list in August, 2012 for easy procurement and support of this library.

Appendix-1: Server Configuration and Tuning Details

Benchmark Test Details

- Back-to-back configuration (no Switch), run in 10GbE/40GbE/IB modes.
- ½ Round Trip
- Netperf v2.5.0; Sockperf v2.5.156

Server Configuration (Gen8)

- Servers and FlexibleNetwork Adapter Cards:
  - Two Servers: HP Gen8 DL380p
  - CPU: Intel(R) Xeon (R) CPU E5-2690 2.90GHz

  - Memory: 32GB (8 x4GB 1600MHz)
  - ConnectX-3 PCIe Gen3 FDR FlexibleLOM (HP Part# 649282-B21)
  - Benchmarks ran on the Processor where PCIe card is inserted (unless noted).
- SW Versions:
  - OS: RHEL6.1
  - ConnectX-3 FW: 2.10.2220
  - Driver: OFED-VMA 1.5.3-0008
  - VMA: 6.1.6
- Tuning:
  - iLO4 version 1.05 and BIOS firmware 2012.02.21
  - TurboBoost disabled (except for specific TurboBoost benchmark).
  - Both SMIs turned off; tuned per HP Low Latency Tuning White Paper #581608-004
  - For white paper, search pub# on HP.com or send e-mail to: *low.latency@hp.com*

Server Configuration (G7)

- Servers and FlexibleNetwork Adapter Cards:
  - Two Servers: HP Gen7 DL380
  - CPU: Intel(R) Xeon (R) X5687 3.60GHz
  - Memory: 48GB (8 x4GB 1600MHz)
  - ConnectX-3 PCIe Card in Gen2 slot (HP Part #649281-B21)
  - ConnectX-2 PCIe Card in Gen2 slot (HP Part #516937-B21)
- SW Versions:
  - OS: RHEL6.1
  - ConnectX-3 FW: 2.10.2220
  - ConnectX-3 FW: 2.9.1000
  - ConnectX-2 Driver: MLNX_OFED 1.5.3-3
  - ConnectX-3 Driver: OFED-VMA 1.5.3-0008
  - VMA: 6.1.6
- Tuning:
  - Standard HP G7 low latency tuning, including both SMIs turned off.

Appendix-2: Benchmark Tests Command Lines Used

Netperf

Four different Netperf tests are:

Measurement	Netperf Test Name
TCP Bandwidth	TCP_STREAM
TCP Latency	TCP_RR
UDP Bandwidth	UDP_STREAM
UDP Latency	UDP_RR

The following command lines are used on the client side:

- TCP_STREAM: netperf -n 16 -H <peer ip> -c -C -P 0 -t TCP_STREAM -l 10 -T 2,2 -- -m <message size>
- TCP_RR: netperf -n 16 -H <peer ip> -c -C -P 0 -t TCP_RR -l 10 -T 2,2 -- -r <message size>
- UDP_STREAM: netperf -n 16 -H <peer ip> -c -C -P 0 -t UDP_STREAM -l 10 -T 2,2 -- -m <message size> -s 128K -S 128K

Where:

- -n 16 : 16 cpu cores on the local host
- -c -C: Measure CPU utilization on local and remote hosts
- -l 10: Test duration of 10 seconds
- -T 2,2: Bind the client and server to cpu 2 (arbitrary)

The command line used on the server side for all tests was: "netserver –D –f".

Sockperf

Four different Sockperf tests are:

Measurement	Sockperf Test Name
TCP Bandwidth	TCP throughput (tp)
TCP Latency	TCP ping-pong (pp)
UDP Bandwidth	UDP throughput (tp)
UDP Latency	UDP ping-pong (pp)

The following command lines are used on the client side:

- TCP Bandwidth: sockperf tp --tcp -m <message size> -i <peer ip> -t 10
- TCP Latency: sockperf pp --tcp -m <message size> -i <peer ip> -t 10
- UDP Bandwidth: sockperf tp -m <message size> -i <peer ip> -t 10
- UDP Latency: sockperf pp -m <message size> -i <peer ip> -t 10

Where:

- -t 10: Test duration of 10 seconds.

The command line used on the server side for UDP tests was: "sockperf sr -i <interface ip to bind>".

The command line used on the server side for TCP tests was: "sockperf sr -i <interface ip to bind> --tcp".

Mellanox TECHNOLOGIES

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