Effective low-cost High-Frequency Trading solutions from IBM and Mellanox
May 2011

A Low-Latency Solution for High-Frequency Trading from IBM and Mellanox

Vinit Jain
IBM Systems and Technology Group
Falke Bruinsma
IBM Software Group

Executive Overview

In the world of High-Frequency Trading (HFT), opportunities exist only fleetingly and therefore trading solutions must run at the lowest latency to be competitive. Low-latency 10 Gigabit Ethernet has become the interconnect of choice for HFT solutions. IBM® and Mellanox® have demonstrated a solution that performs at high throughput rates and low latency to facilitate High-Frequency Trading solutions.

This solution uses the IBM BNT® 8264 10/40Gb Ethernet Switch, coupled with IBM's WebSphere® MQ Low Latency Messaging (WMQLLM) software using the Mellanox ConnectX®-2 10 Gigabit Ethernet adapter with RoCE (RDMA over Converged Ethernet) drivers. This solution delivers a powerful combination of networking hardware and messaging software that meets the latency and throughput requirements of high-frequency trading environments.

This has been demonstrated through independently audited benchmarks published using the STAC-M2 Benchmark™ test. In addition to the STAC-M2 Benchmark test, further testing was performed by IBM to explore the latency and throughput performance that this solution delivers. The results show that the average latency of this solution does not exceed 5 µsec, with standard deviation remaining below 2 µsec. These results are obtained using a typical message size of 128 bytes and at message rates of 1 million messages / second.

High-Frequency Trading

HFT has gained a strong foothold in financial markets, driven by several factors, including advances in information technology that have been conducive to its growth. Unlike traditional traders who hold their positions long term, high-frequency traders hold their positions for shorter durations, which can be as little as a few seconds. They typically end their day with few to no positions carried over to the next business day.

HFT uses strategies such as statistical arbitrage and liquidity detection. Statistical arbitrage relies on the principle that stocks in a pair or pool of correlated stocks that diverge from their statistically expected behavior will converge and profit is achieved by taking positions based on this expectation. Liquidity detection is the strategy by which small trades are sent out to detect large orders that are not visible and then taking positions based on the expectation that the large orders will move the market. These strategies require programs to analyze massive amounts of market data using complex algorithms to exploit opportunities that exist for as little as a fraction of a second. The execution of these strategies requires information technology that can compute complex algorithms, and exchange messages at extremely low latencies—even at very high rates—to handle volume spikes without impacting system performance. In fact, it is at exactly those times when the volume spikes that HFT systems may take advantage of delays in other system.

Therefore IT organizations in the financial services industry face tremendous pressures to optimize the transaction lifecycle. There is a critical need for the underlying messaging infrastructures to deliver extremely low latency and very high message throughputs.

Solution

IT organizations are expected to deliver solutions offering low latency with high throughput without using specialized technology, to avoid high cost for capital and skills. The trend has been to favor solutions that use commodity hardware and software components. As a result, low-latency 10 Gigabit Ethernet has become the interconnect of choice.

IBM and Mellanox have demonstrated an ultralow-latency messaging solution that performs at high throughput rates with reliability. This solution stack, shown in *Figure 1*, addresses the requirements of the financial industry and delivers the solution using commodity hardware and software components.

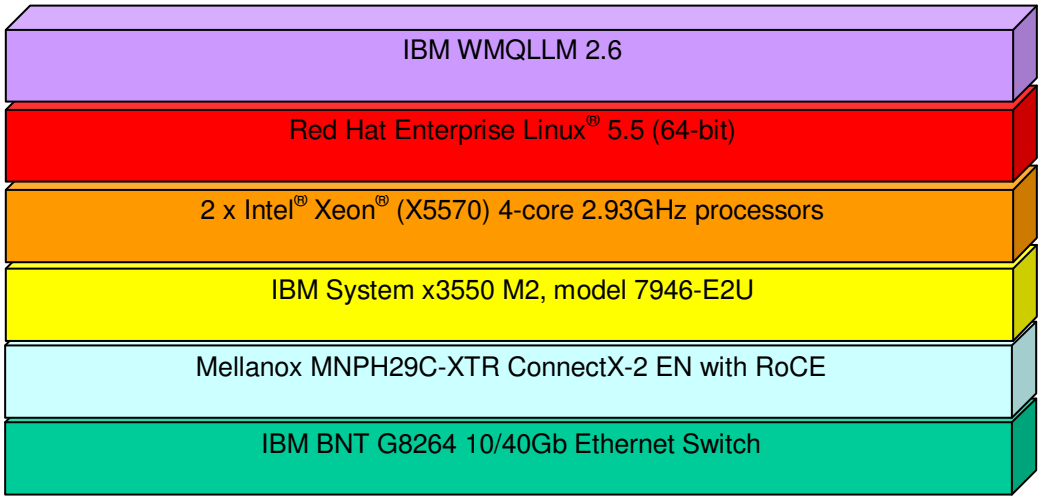

Figure 1. *Solution Stack Components*

This solution uses the IBM BNT 8264 10 Gigabit Ethernet switch, coupled with IBM's WMQLLM software using the Mellanox ConnectX-2 10 Gigabit Ethernet adapter with RoCE drivers. This solution delivers a powerful combination of networking hardware and messaging software that meets the latency and throughput requirements of high-frequency trading environments. This has been demonstrated through independently audited benchmarks published using the STAC-M2 Benchmark test. In this paper we present additional test results demonstrating the latency characteristics of this solution.

IBM BNT RackSwitch G8264 10/40 GbE Switch

The IBM BNT RackSwitch™ G8264 is high-performance switch designed to meet the demanding requirements of high-frequency trading systems. It provides line-rate, high-bandwidth switching, filtering, and traffic forwarding, without delaying data. This switch offers up to 64 10GbE ports and up to 4 40GbE ports, 1.2 Terabits per second of non-blocking bidirectional throughput in a 1U footprint. In addition to a rich set of Layer-2 and Layer-3 connectivity, the G8264 supports the newest protocols, including Data Center Bridging / Converged Enhanced Ethernet (DCB/CEE) for support of Fibre Channel over Ethernet (FCoE). Redundant power and fans, along with numerous high-availability features, enable the RackSwitch G8264 to always be available for business-critical traffic.

The single-chip design and the default cut-through mode are key to enabling extremely low deterministic latency and jitter. Large data-center-grade buffers enable congestion free operation. Furthermore, the G8264 delivers best-of-breed performance and function, including Layer-3 with a standard 40GbE interconnect into the core, rather than taking the approach of a proprietary core interconnection, as chosen by some Ethernet vendors.

IBM Websphere MQ Low Latency Messaging

WebSphere MQ Low Latency Messaging is a transport fabric product engineered for the rigorous latency and throughput requirements typical of today's financial trading environments. The product is daemonless and provides peer-to-peer transport for one-to-one, one-to-many and many-to-many data exchange. It also exploits the IP multicast infrastructure to enable scalable resource conservation and timely information distribution.

Designed to dramatically improve throughput and reduce latency while maximizing system reliability, WMQLLM can help high-frequency trading organizations enhance the responsiveness of their existing trade infrastructure while developing new solutions for emerging business opportunities. Several factors contribute to the high performance enabled by WMQLLM. For example, a unique method of message packetization enables delay-free, high-speed data delivery. Unique batching technology dynamically optimizes packetization for reliable delivery and lowest latency, based on throughput, message sizes, receiver, and system feedback. In addition, very compact packet headers leave more network bandwidth for application data.

WMQLLM supports high performance interconnects, such as 10 Gigabit Ethernet and InfiniBand® to enable higher throughput with lower latency, reduced latency variability, and low CPU utilization.

Mellanox ConnectX-2 10GbE Server Adapter with RoCE

Mellanox ConnectX-2 EN Ethernet Network Interface Cards (NICs) deliver low latency, high throughput and low CPU utilization leveraging the RoCE standard. RoCE is based on the IBTA RoCE specifications, and utilizes the Open Fabrics Enterprise Distribution (OFED) verbs interface as the software interface between application layer and ConnectX-2 EN hardware. RoCE takes advantage of transport services support of various modes of communication, such as reliable connected services and datagram services. RoCE uses well-defined verb operations, including kernel bypass, send/receive semantics, RDMA read/write, user-level multicast, user-level I/O access, zero copy and atomic operations. The ConnectX-2 EN adapters with RoCE are widely used in financial services for removing I/O bottlenecks, lowering latency and jitter, and increasing message rates for high-frequency trading, market data distribution and real-time risk management.

Performance Testing

STAC-M2 Benchmark Test

The STAC-M2 Benchmark specifications test the ability of a solution to handle real-time market data in a variety of configurations found in typical trading environments. The specifications are defined by end-user IT executives within the financial industry with input from vendors of high-performance messaging solutions. The STAC-M2 Benchmarks provide key performance metrics such as latency, throughput, power efficiency, and CPU/memory consumption under several scenarios, including both undisturbed flow and exception conditions like slow consumers. STAC Report™ Highlights are made available at the STAC website (*http://www.stacresearch.com*) for systems that have been independently audited by STAC. The full STAC Report™ is available in the STAC Vault™ to STAC's premium end-user subscribers.

IBM and Mellanox completed an audited STAC-M2 Benchmark using the solution described in this paper. The latency results below are from three of the 10 test sequences that were carried out. The tests we describe the results for here are:

- **BASELINE** — Each of five consumers has a unique watchlist, with one consumer per motherboard. BASELINE emulates applications such as a smart order router that partitions symbols across servers.
- **OVERLAP** — This is similar to the BASELINE test except for some overlap in the consumer watchlists. OVERLAP emulates deployments such as multiple black-box applications that are independent of one another.
- **FLEXIBLE** — Emulates a general enterprise deployment, where costs matter more than latency. FLEXIBLE requires 15 consumer applications, with no restriction on the number of consumers per motherboard and with some overlap in the consumer watchlists.

STAC-M2 Latency IBM BNT G8264 10/40 GbE and Mellanox ConnectX-2 with RoCE (SUT ID LLM110421)				
Test Description (Spec ID)	Mean (µsec)	Med (µsec)	99P (µsec)	STDV (µsec)
SupplyToReceive Latency (Hybrid) at base rate in the 1:5 setup with no watchlist overlap **(STAC.M2.v1.0.BASELINE.LAT1)**	6	6	9	1
SupplyToReceive Latency (Hybrid) at base rate in the 1:5 setup with some watchlist overlap **(STAC.M2.v1.0.OVERLAP.LAT1)**	6	6	9	1
SupplyToReceive Latency (Hybrid) at base rate in the setup with flexible Consumer resources (**STAC.M2.v1.0.FLEXIBLE.LAT1)**	7	7	9	1

Table 1. *STAC-M2 Latency Results*

This solution exhibited the best mean latencies ever published for the foregoing benchmarks in a STAC-M2 Report.

Single-Hop Latency Test

In addition to the STAC-M2 Benchmark test, further testing was performed by IBM to explore the performance such a solution can deliver in terms of latency.

As shown in *Figure 2*, the test is a reflector test and the setup consists of two machines, A and B, connected through an IBM BNT G8264 RackSwitch.

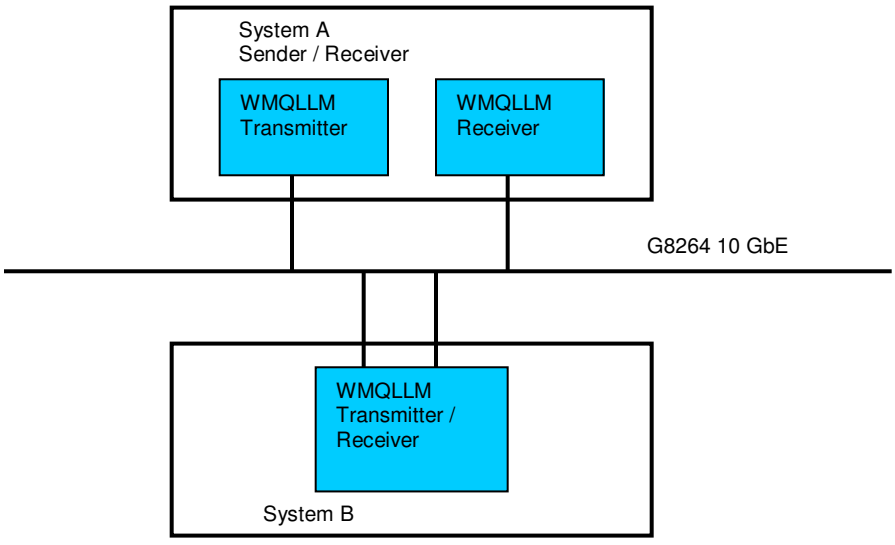

Figure 2. *Reflector Test Layout*

On System A, the “sender” sends packets at the rate being tested to the “reflector” on System B. The “reflector” receives every packet, and only forwards to the “receiver” on System A those packets that have a time stamp for latency measurement. The “receiver” on System A extracts the time stamps from the reflected packet and uses it to measure round-trip time. The single-hop

latency is calculated as half of the round trip time. The standard deviation is calculated using the round trip time.

All latency tests ran for 5 minutes. Approximately 300,000 latency samples were recorded for each 5-minute test. From these 300,000 samples latency statistics were calculated. Two test parameters were used to vary the workload for this testing: message size and message rate. *Table 2* shows the results for each message rate and size test combination.

LLM Latency using IBM BNT G8264 10/40 GbE and Mellanox ConnectX-2 with RoCE					
Msg Rate [msgs/sec]	Msg Size [bytes]	Single Hop			RTT
		Average [µsec]	Median [µsec]	99P [µsec]	Std Dev [µsec]
10,000	45	3.6	3.0	4.5	0.7
100,000		3.6	4.0	4.5	0.9
1,000,000		4.3	4.0	5.5	2.2
10,000	128	4.5	4.0	4.5	0.7
100,000		4.7	5.0	5.5	0.8
1,000,000		5.5	5.0	6.5	1.9
10,000	512	5.7	5.0	7.0	0.9
100,000		5.9	6.0	7.0	0.9
1,000,000		8.0	8.0	8.0	1.9

Table 2. *Single-hop latency*

Key Takeaway

The average latency of this solution remains in the extremely low range of 3.6 – 8.0 µsec with standard deviation remaining less than 2 µsec, even as message sizes grow large and at very high rates.

Conclusions

These results clearly show that a messaging solution stack created using IBM's WMQ Low Latency Messaging, IBM BNT G8264 10GbE switch, and Mellanox's ConnectX-2 EN with RoCE adapters, delivers the technology that High-Frequency Trading applications need. This result has been audited by STAC and conforms to the STAC M2 benchmark. Defined by trading firms, the STAC M2 benchmark represents typical requirements of a trading system infrastructure. It measures latency and throughput in both optimal and failure scenarios. The results of the STAC audit and subsequent tests by IBM confirm that the components described in this paper produce the lowest latency solution while scaling to very high message rates.

References

High Performance Business Computing in Financial Institutions, Sue Gouws Korn, CFA, Christopher G. Willard, Ph.D, Addison Snell, February 2011

High-frequency trading, Deutsche Bank Research, February 7, 2011

STAC Report Highlights: IBM WMQ LLM with IBM x3550 servers, Mellanox ConnectX-2 EN and IBM BNT G8264 10/40Gb Ethernet Switch, May 2011

Develop high-volume, low-latency finance solutions with IBM WebSphere MQ Low Latency Messaging, Financial industry solutions White paper, IBM, October 2009

For More Information

IBM System Networking	http://**ibm.com**/systems/networking
IBM System x Servers	http://**ibm.com**/systems/x
IBM Systems Director Service and Support Manager	http://**ibm.com**/support/electronic
IBM System x and BladeCenter Power Configurator	http://**ibm.com**/systems/bladecenter/resources/powerconfig.html
IBM Standalone Solutions Configuration Tool	http://**ibm.com**/systems/x/hardware/configtools.html
IBM Configuration and Options Guide	http://**ibm.com**/systems/x/hardware/configtools.html
IBM ServerProven Program	http://**ibm.com**/systems/info/x86servers/serverproven/compat/us
Technical Support	http://**ibm.com**/server/support
Other Technical Support Resources	http://**ibm.com**/systems/support

Legal Information

IBM Systems and Technology Group
Dept. USA
3039 Cornwallis Road
Research Triangle Park, NC 27709

Produced in the USA
May 2011

For a copy of applicable product warranties, write to: Warranty Information, P.O. Box 12195, RTP, NC 27709, Attn: Dept. JDJA/B203. IBM makes no representation or warranty regarding third-party products or services including those designated as ServerProven® or ClusterProven®. Telephone support may be subject to additional charges. For onsite labor, IBM will attempt to diagnose and resolve the problem remotely before sending a technician.

IBM, the IBM logo, ibm.com, BNT, ClusterProven, RackSwitch, ServerProven, and WebSphere, are trademarks of IBM Corporation in the United States and/or other countries. If these and other IBM trademarked terms are marked on their first occurrence in this information with a trademark symbol (® or ™), these symbols indicate U.S. registered or common law trademarks owned by IBM at the time this information was published. Such trademarks may also be registered or common law trademarks in other countries. For a list of additional IBM trademarks, please see http://**ibm.com**/legal/copytrade.shtml.

InfiniBand is a trademark of InfiniBand Trade Association.

Intel, the Intel logo, and Xeon are trademarks or registered trademarks of Intel Corporation or its subsidiaries in the United States and other countries.

Linux is a registered trademark of Linus Torvalds.

Mellanox and ConnectX-2 are trademarks or registered trademarks of Mellanox.

STAC and all STAC names are trademarks or registered trademarks of the Securities Technology Analysis Center, LLC.

Other company, product and service names may be trademarks or service marks of others.

IBM reserves the right to change specifications or other product information without notice. References in this publication to IBM products or services do not imply that IBM intends to make them available in all countries in which IBM operates. IBM PROVIDES THIS PUBLICATION "AS IS" WITHOUT WARRANTY OF ANY KIND, EITHER EXPRESS OR IMPLIED, INCLUDING THE IMPLIED WARRANTIES OF MERCHANTABILITY AND FITNESS FOR A PARTICULAR PURPOSE. Some jurisdictions do not allow disclaimer of express or implied warranties in certain transactions; therefore, this statement may not apply to you.

This publication may contain links to third party sites that are not under the control of or maintained by IBM. Access to any such third party site is at the user's own risk and IBM is not responsible for the accuracy or reliability of any information, data, opinions, advice or statements made on these sites. IBM provides these links merely as a convenience and the inclusion of such links does not imply an endorsement.

Information in this presentation concerning non-IBM products was obtained from the suppliers of these products, published announcement material or other publicly available sources. IBM has not tested these products and cannot confirm the accuracy of performance, compatibility or any other claims related to non-IBM products. Questions on the capabilities of non-IBM products should be addressed to the suppliers of those products.

MB, GB and TB = 1,000,000, 1,000,000,000 and 1,000,000,000,000 bytes, respectively, when referring to storage capacity. Accessible capacity is less; up to 3GB is used in service partition. Actual storage capacity will vary based upon many factors and may be less than stated.

Performance is in Internal Throughput Rate (ITR) ratio based on measurements and projections using standard IBM benchmarks in a controlled environment. The actual throughput that any user will experience will depend on considerations such as the amount of multiprogramming in the user's job stream, the I/O configuration, the storage configuration and the workload processed. Therefore, no assurance can be given that an individual user will achieve throughput improvements equivalent to the performance ratios stated here.

Maximum internal hard disk and memory capacities may require the replacement of any standard hard drives and/or memory and the population of all hard disk bays and memory slots with the largest currently supported drives available. When referring to variable speed CD-ROMs, CD-Rs, CD-RWs and DVDs, actual playback speed will vary and is often less than the maximum possible.

QCW03013-USEN-00