IBM WebSphere MQ Low Latency Messaging Software Tested With Arista 10 Gigabit Ethernet Switch and Mellanox® ConnectX®-2 EN with RoCE Adapter Delivers Reliable Multicast Messaging With Ultra Low Latency And High Throughput

Introduction

Microseconds matter in the ultra-competitive world of High Frequency Trading (HFT). Trading architectures that deliver the lowest latency solution are the difference between profit and loss for these financial firms. IBM, Mellanox and Arista have combined to demonstrate a low latency messaging transport solution that facilitates the high-speed delivery of market, trade, reference and event data, in or between front-, middle and back-office operations.

10 Gigabit Ethernet prevails as the interconnect technology of choice for these applications. With its full non-blocking throughput, record density, low latency, robust function, ease of operation and leading TCO, the Arista 7100 Series 10 Gigabit Ethernet switch is ideal for HFT applications. Arista's high-throughput, low latency 10Gb Ethernet switch, coupled with IBM's WebSphere MQ Low Latency Messaging (WMQLLM) software using the Mellanox ConnectX-2 10 Gigabit Ethernet adapter with RDMA over Converged Ethernet (RoCE) and OFED (OpenFabrics Enterprise Distribution) drivers, is a powerful combination of networking and messaging software that delivers the lowest latency solution for environments like High Frequency Trading.

Key Results Summary

- WebSphere MQ Low Latency Messaging (WMQLLM) 2.3.0 Reliable Multicast Messaging (RMM) using 10GbE maintained a latency of 4μs at rates up to 100K messages/second and 6μs at one million messages/second.
- Application throughput for the tested solution was on the order of 8.4 Gbps.
- Latency remained consistently low as the number of multicast groups increased from one to one thousand.
- Across all message sizes and message rates tested the average latency of LLM using 10GbE was 5X to 12X faster than LLM over 1GbE.

Test Goals

The main objective of this testing is to measure the single hop latency for WebSphere MQ Low Latency Messaging (WMQLLM) 2.3.0 using the 10GbE RoCE industry standard protocol with Mellanox's ConnectX-2 adapter (MT25448), as well as to determine the maximal end-to-end throughput rate. Maximal throughput and single hop latency were measured on a 10GbE network fabric.

System under test

WMQLLM

Software Version:	2.3.0

Machines

Vendor Model:	IBM x3650 M2 (quad core)
Processors:	2 x Intel Xeon Quad Core X5570 2.93 GHz
Cache:	8MB Level 2 cache
Front Side Bus Speed:	1333 MHz
Memory:	16GB
10GbE Adapter:	Mellanox ConnectX-2 MT25448 [ConnectX-2 EN with RoCE, PCIe 2.0 2.5 GT/s] (rev a0)
Driver:	OFED-1.5.1

Operating System

Version:	Linux SuSE 11 (2.6.18-164.el5) x86_64

Network Elements

Switch:	Arista 7124S

Network Accessories

Connectors:	SFP+
Cables:	LC to LC Fiber Optic cables

Single hop latency test

Test description

The test setup consists of two machines A and B connected through an Arista 7124S switch. The test is a simple reflector tests. On machine A an LLM transmitter sends messages to an LLM receiver on machine B over reliable multicast. The messages are immediately sent back to machine A using an LLM transmitter and are received by an LLM receiver. A time stamp is written to a subset of the messages before each message is submitted to the LLM transmitter on machine A and the time stamp is extracted by the LLM receiver on machine A after completing the round trip. The single hop latency is calculated as half 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. The test was repeated with variable message rates. Results are shown for message of 45 bytes. A diagram of the test configuration is shown below in Figure 1.

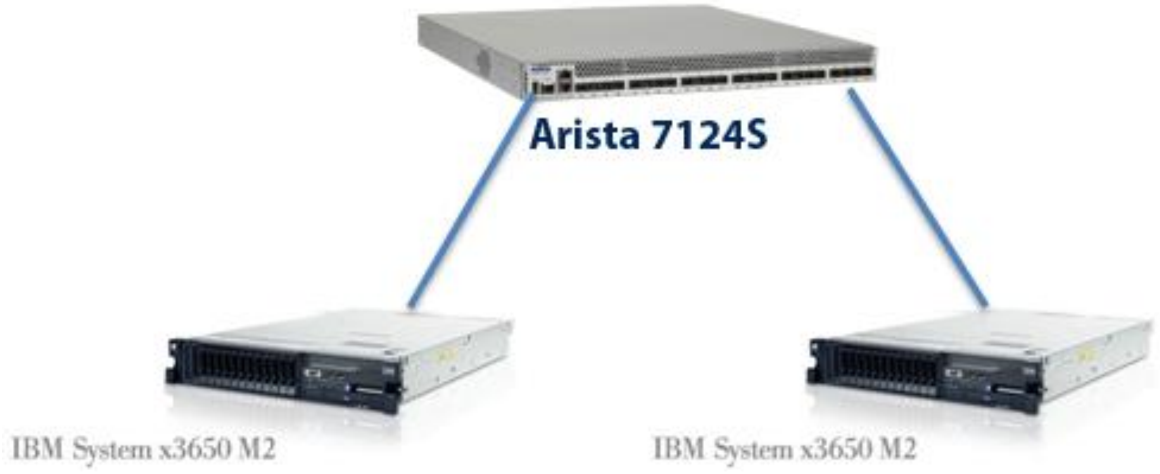

Figure 1: Test configuration

The results of the latency tests are presented in Table 1.

LLM over RDMA (RoCE)				
Mellanox ConnectX-2 (MTU=2K)				
Rate [msgs/sec]	median [usec]	average [usec]	max [usec]	std [usec]
10,000	4.00	4.50	15.50	0.90
50,000	4.00	4.50	26.00	1.10
100,000	4.00	4.50	41.00	2.10
250,000	4.50	5.00	88.00	7.30
500,000	4.50	6.00	180.00	23.30
1,000,000	6.00	7.50	140.00	18.10

Table 1: Single hop latency results for LLM over 10GbE using RoCE

Key point:

- As we increase the message rate to 1,000,000 msgs/sec, the median latency remains consistently low for the tested solution, ranging from 4 microseconds to 6 microseconds.

Throughput test

Test description

The test setup consists of two machines A and B connected through an Arista 7124S switch. On machine A an LLM transmitter sends messages to an LLM receiver on machine B over reliable multicast. The reported throughput is the maximal rate at which messages can be delivered from the sender to the receiver without any loss.

All throughput tests ran for 5 minutes. The throughput test has been repeated with different message sizes.

The results of the latency tests are presented in Table 2. It important to note that the reported throughput is that of the application and does not include the transport and network overhead.

LLM over UDP	
Mellanox ConnectX-2 (Ethernet MTU=1500 bytes)	
Msg size [bytes]	[Gbit/sec]
45	8.40
100	8.20
1,000	8.46

Table 2: Max throughput for LLM over 10GbE

Key point:

- The tested solution can deliver close to a full 10GbE of network throughput across different message sizes.

Multicast Scalability Test

The main objective of this test is to determine multicast scalability, in terms of the number of different multicast groups, of the Arista switch, when used in conjunction with IBM's WMQLLM software running on an IBM Blade HS-21 (quad core) system with the Mellanox ConnectX-2 Ethernet adapter.

Test Description

The test setup consists of four machines A, B, C and D connected through an Arista 7124S switch. A multiple multicast group reflector tests is running between machines C and D to load the switch with multicast data. While the load test is running the basic latency test (described in the 'single hop latency test' subsection) is running, using a single multicast group, between machines A and B and produces latency results. The test is repeated with the load test between C and D configured to use different number of groups.

The results of the multicast scalability tests are presented in Table 3.

LLM over UDP		
Rate [msgs/sec]	Multicast Groups	Median Latency [usec]
100,000	1	13.50
	16	12.50
	64	13.00
	256	13.00
	512	12.50
	1,000	12.50

Table 3: Single hop latency results of LLM over UDP in the multicast scalability test

Key point:

- As the number of multicast groups grows, the tested solution is able to scale to efficiently deliver multiple flows of data to multiple groups of subscribers while maintaining consistently low latency.

Summary

The results reported here clearly show the significant performance benefits of a 10GbE based solution combining Arista’s switching technology, Mellanox ConnectX-2 EN with RoCE adapters, and IBM's WebSphere MQ Low Latency Messaging.

This unique combination of networking technology and messaging software provides the lowest latency solution for today's demanding HFT environments with a solution that scales to very high message rates.

As well, for applications that require high bandwidth, for example, exchanges or trading applications that need to store information to a transaction logger, or for system architectures anticipating increased future traffic volumes, the tested solution delivers close to 100% utilization of available network bandwidth.

Finally, the ability to forward traffic with consistently low latency as the number of multicast groups increases is important in financial market environments where multiple flows of data need to be delivered effectively to multiple subscribers to maximize application performance. The tested solution clearly demonstrated this capability.

About Arista

Arista Networks was founded to deliver cloud networking solutions for large datacenter and computing environments. Arista Networks ignited the low-latency 10GbE Ethernet revolution with the Arista 7100 and reinvented the modular data center Ethernet switch with the Arista 7500. Arista leads the data center Ethernet switching industry with innovation in switching hardware, silicon based performance, and the EOS platform, a pioneering new software architecture with self-healing and live in-service software upgrade capabilities. Arista markets its products worldwide through distribution partners, systems integrators and resellers with a strong dedication to partner and customer success.

For more information, visit http://www.aristanetworks.com.

About IBM

For more information, visit http://www.ibm.com/financialmarkets/fasterdata.

About Mellanox

Mellanox Technologies is a leading supplier of end-to-end connectivity solutions for servers and storage that optimize data center performance. Mellanox products deliver market-leading bandwidth, performance, scalability, power conservation and cost-effectiveness while converging multiple legacy network technologies into one future-proof solution. For the best in performance and scalability, Mellanox connectivity solutions are a preferred choice for Fortune 500 data centers and the world’s most powerful supercomputers. Founded in 1999, Mellanox Technologies is headquartered in Sunnyvale, California and Yokneam, Israel.

For more information, visit Mellanox at www.mellanox.com.