



29West Messaging Performance on 10-Gigabit Ethernet

September 22, 2008

Version 1.0

Executive Summary

This report highlights the performance of 29West Messaging on 10-gigabit Ethernet. Commodity Dell workstations running Linux were used along with the new Nexus 5020 switch from Cisco and NE020 10GbE adapters from NetEffect.

At this writing, 10-gigabit Ethernet is emerging as a compelling technology for server access to the network. Over 1 million 10-gigabit Ethernet ports shipped in 2007 with most being used for interconnecting network equipment. Many 29West customers are planning deployments of 10-gigabit Ethernet for server access so we have run a series of benchmarks to help our customers estimate the performance they may be able to achieve. Our key findings from these benchmarks include:

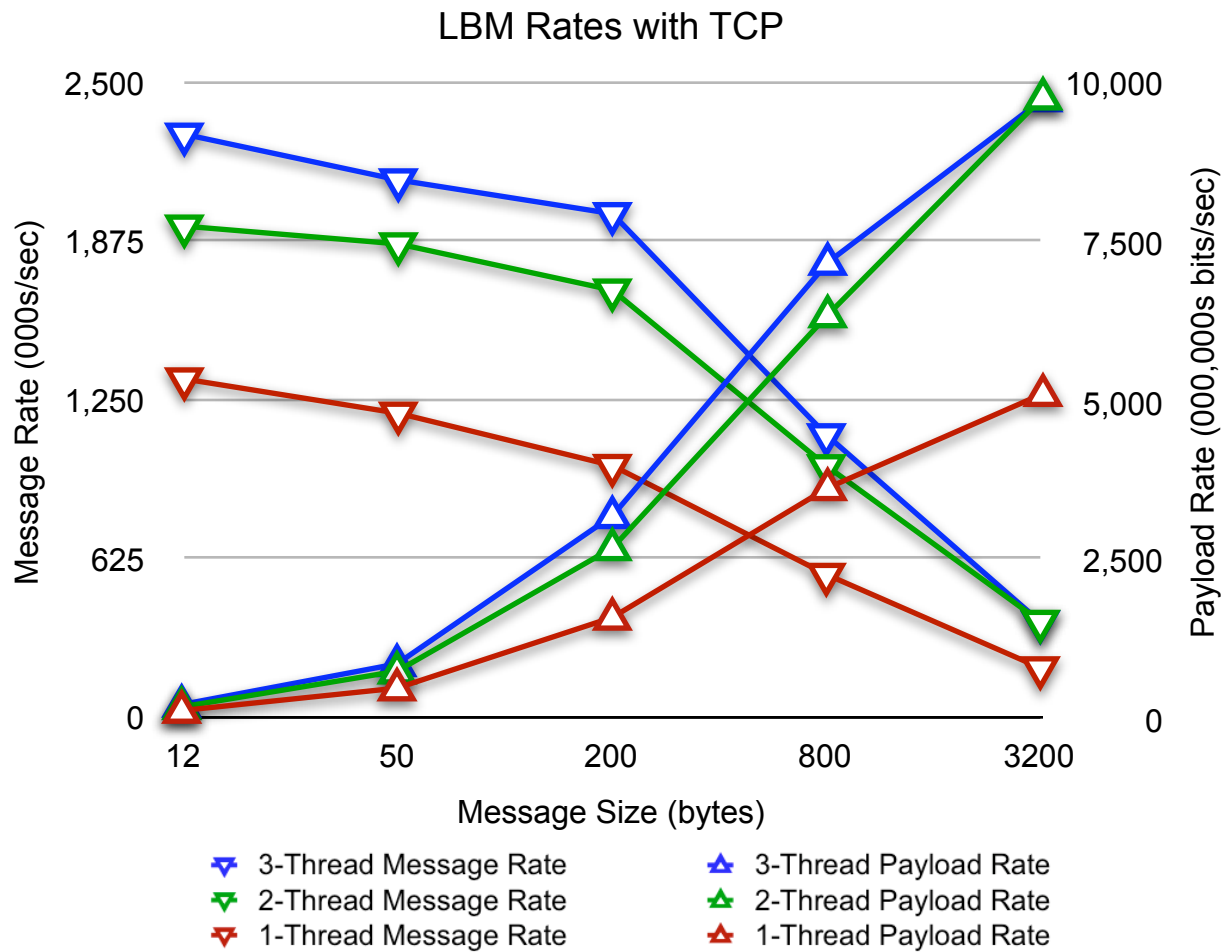
- **Payload delivery rates** as high as **9.8 gbps** for large messages
- **Message delivery rates** more than **2.3 million messages per second** for small messages
- **Average ping-pong latency** of **36 microseconds**
- **Scalable performance** as receivers are added when using UDP multicast
- **Message delivery latency** under **50 microseconds** for rates up to 110,000 messages per second

We ran tests using TCP and UDP multicast transport protocols. The benefits of stateless TCP offload features in NICs and accompanying kernel support for these features can be seen in the TCP throughput results. Vendors are now developing kernel bypass libraries that will offer better performance for both UDP and TCP. 29West is planning a follow-up report when such libraries are available and reliable.

The rest of this report has sections covering the effect of message size and number of threads on message rates, the effect of message rates on latency, descriptions of the test setup, and conclusions.

Message Rates versus Message Size and Number of Threads

The two largest factors in achievable message rates are message payload size and the number of threads sending and receiving simultaneously. This section shows the impact of these factors on message delivery rates and on payload delivery rates. The following graphs show test results for different message sizes across the horizontal axis. Message sizes double at each point along the axis, giving the effect of plotting on a logarithmic scale. The left vertical axis shows message delivery rates (measured in thousands of messages delivered per second). The right vertical axis shows payload delivery rates (measured in millions of payload bits delivered per second).



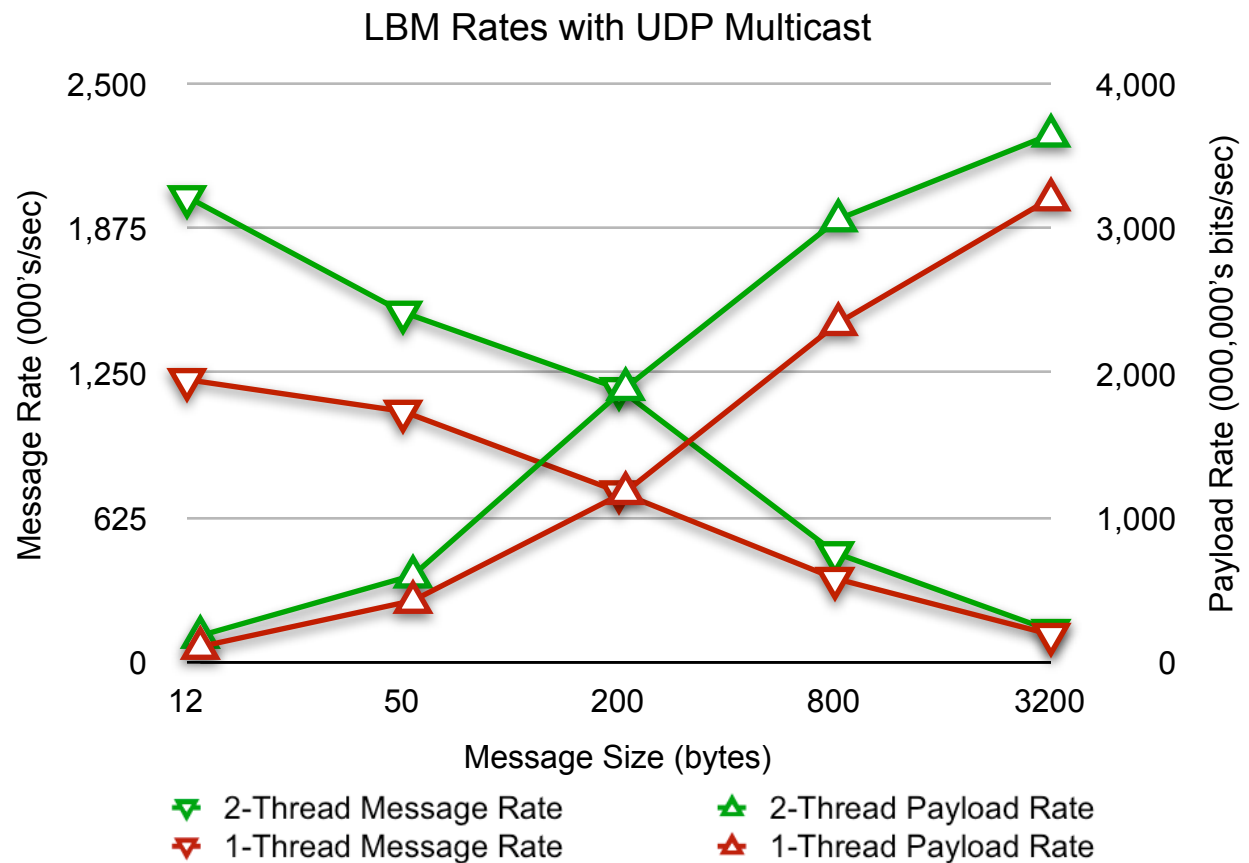
Interpretation of LBM Rates with TCP

Modern CPUs can generate over 2 million messages per second when using small message sizes. Such rates can be maintained through message sizes up to 200 bytes.

The work of sending small messages is mostly CPU work. This is shown by the 50% jump in message rate between the 1-thread (red) line and 2-thread (green) line at the left edge of the chart. We see less impressive gains past 2 threads due to contention for resources in the kernel. Payload rates are low for small message sizes because a 10-gigabit network is difficult to saturate with small messages, even for modern CPUs. However, at message sizes above 50 Bytes, the payload rate easily exceeds 1 gbps, thus showing the value of 10-gigabit server access to the network.

As message size increases, the constant work per message becomes a smaller percentage of the time taken to move the payload bits of the message. Payload rates go up as message rates go down.

The upper right section of the graph shows the theoretical maximum network speed of 9.8 gbps was achieved for large message sizes.



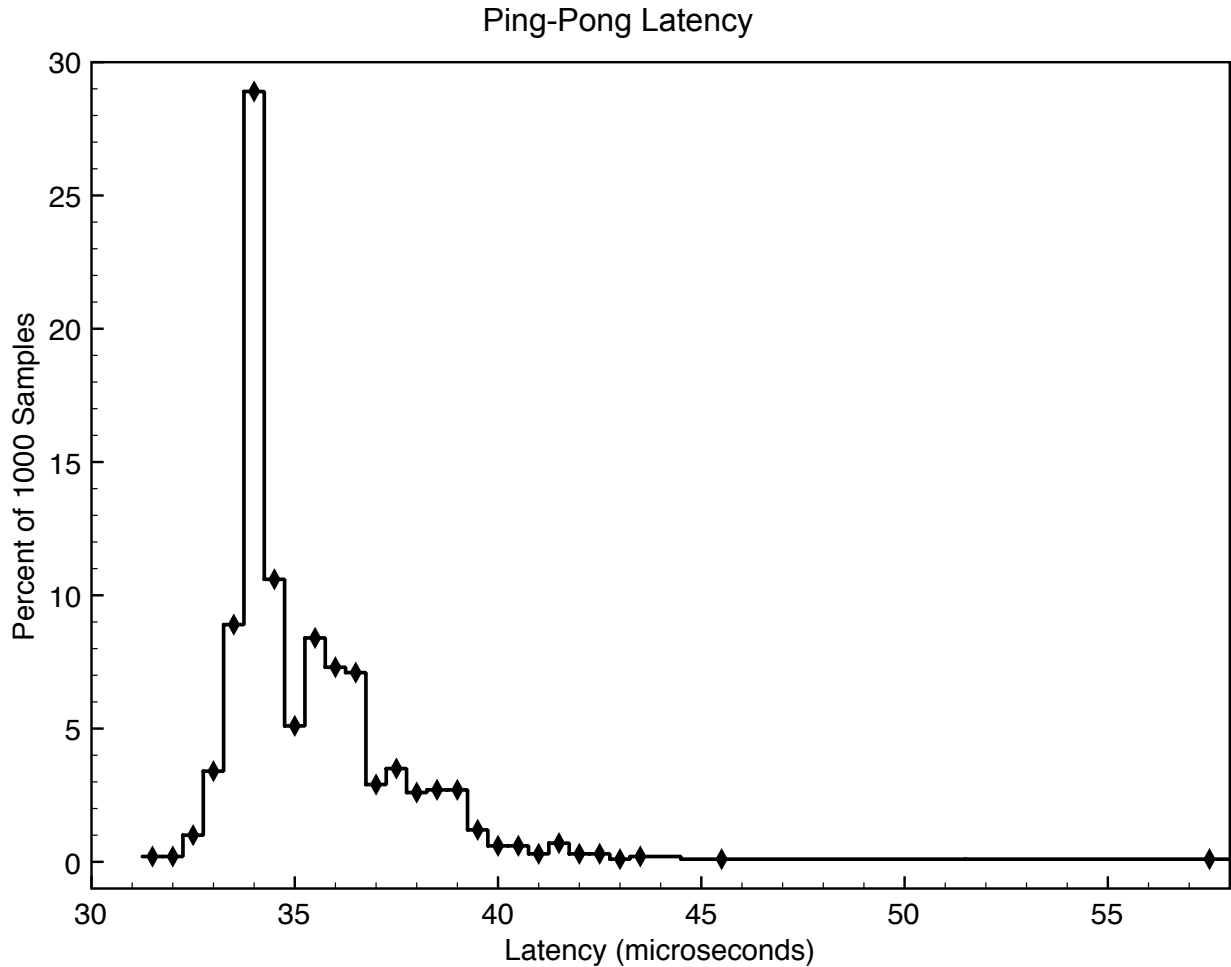
Interpretation of LBM Rates with UDP Multicast

Rate tests with UDP show very similar results with TCP for small message sizes since CPUs limit performance rather than the network. However for larger message sizes, the benefits of stateless TCP offload features can be seen in TCP rates that are about double what we measured with UDP. We tried using 3 threads to get higher rates, but saw little improvement due to thread contention for kernel resources.

In summary, we find that 10-gigabit Ethernet delivers almost 4x the UDP multicast performance of 1 gigabit Ethernet and over 10x the TCP performance of 1 gigabit Ethernet. These rates were achieved without relying on kernel bypass libraries that allow direct hardware access from user space to implement socket calls.

Ping-Pong Latency

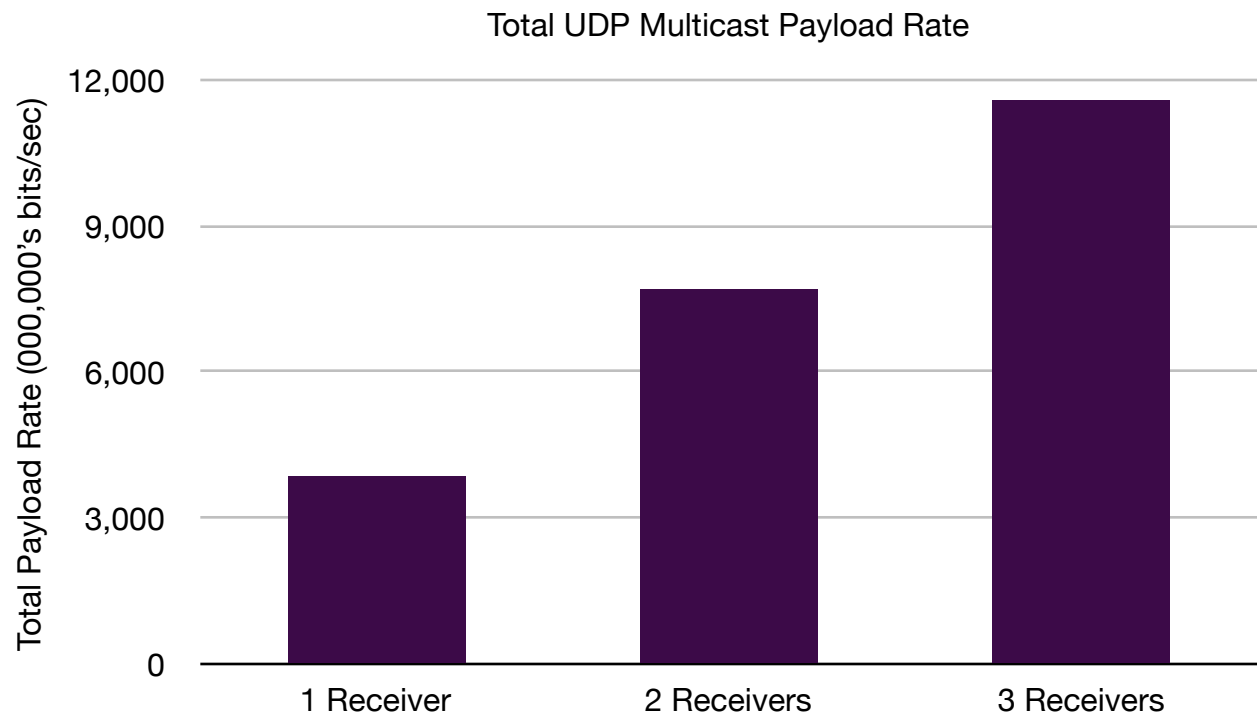
We measured the average time taken for a 16-byte message to move from a sending application to a receiving application at 36 microseconds. The outlier at 57 microseconds is the first sample taken. We assume the first sample takes longer due to cache misses. The most often measured value (nearly 1/3 of the samples) was 34 microseconds. We believe that most of this time is context switching to kernel mode and execution of the IP stack in the kernel.



Scalable Performance as Receivers Are Added

The unicast addressing used by TCP means that a source feeding a group of receivers can never feed them faster than the speed of the source's switch port. For example, a source might generate 10 gbps of traffic toward 3 receivers, but the sum of receive rates across all receivers could never exceed 10 gbps. If all 3 receivers want a copy of the same message, then 3 copies must be sent with TCP.

In contrast, UDP multicast allows a source to send 1 message that will be copied 3 times in a switch and delivered to each of the receivers. We configured a UDP multicast source to run at about 4 gbps and measured the total receive rate as receivers were added. We saw the expected linear growth in receive rate.



Our sources could only generate about 4 gbps so we could not test UDP multicast at wire speed. However the Nexus 5020 switch showed no signs of stress at these rates and we are confident that multicast copying would work at wire speed just as it does in Cisco's 1 gbps switches.

Our lab only has 4 servers with NetEffect 10 gbps NICs so we couldn't test scalability beyond 3 receivers. Again, we are confident that all ports of the Nexus 5020 could receive wire-speed copies of the same message.

Latency versus Message Rates

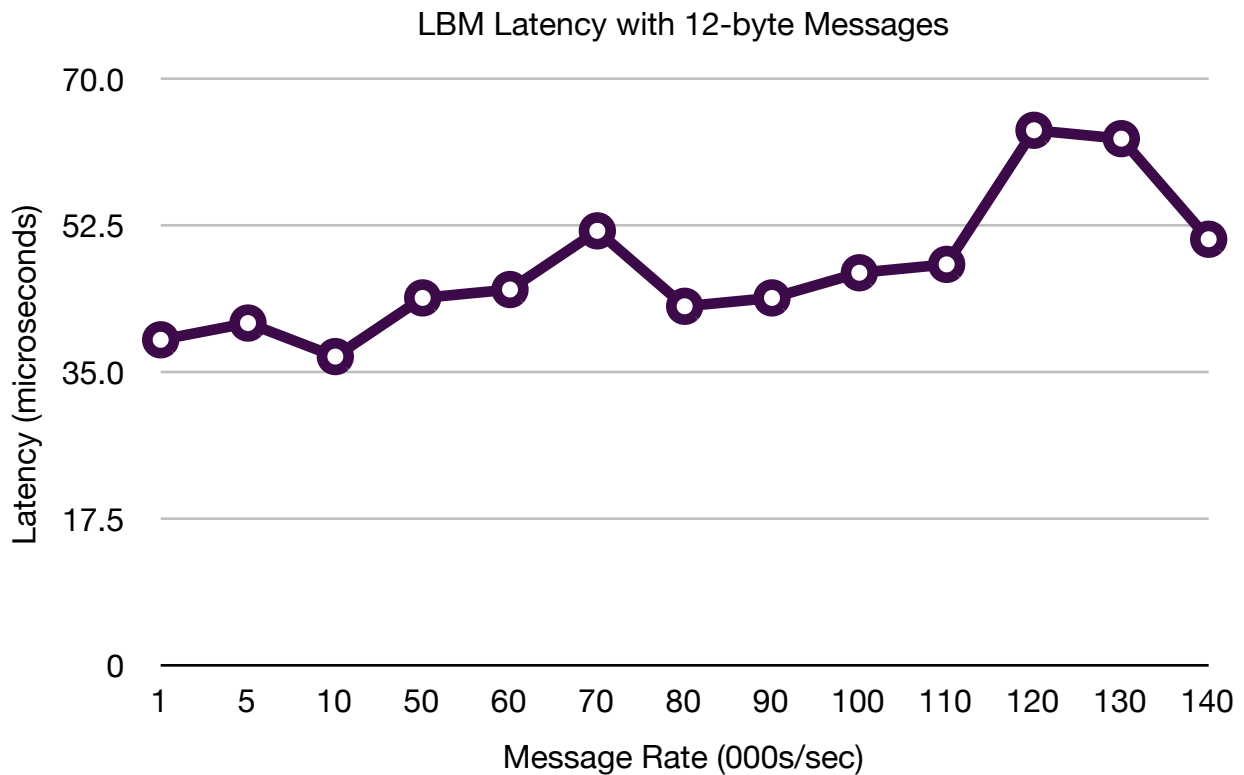
It is easy to measure throughput independent of latency as we have done earlier in this report. Such tests do not measure latency, but do check for lost messages.

It is also easy to measure ping-pong latency, as we have done earlier in this report. Such tests do not generate high message rates because most of the test runtime is spent waiting for 1 message to arrive before the next can be sent.

The ping-pong latency test is appealing because it does not require precisely synchronized clocks. However, most real-world applications do not respond to every message they receive. The work of responding to an incoming message may delay processing of a following message.

29West has developed benchmarking tools that allow us to look for a relationship between latency and message rate. Instead of responding to every incoming message, only every 113th incoming message triggers a response. This limits the impact of replying to less than 1% of messages.

The following chart shows latency on the left vertical axis. Note the horizontal axis has irregular spacing to show detail.



The chart shows message rates changing by two orders of magnitude from 1,000 to 100,000 messages per second while latency changes only 30%. These tests used UDP multicast and sent each message in its own datagram. We found it necessary to enable interrupt coalescing on the receive side to run at rates over 70,000 messages per second. We attribute the rise in average latency above 100,000 messages per second to ever larger numbers of arriving messages being coalesced into a single interrupt.

Conclusions

10-gigabit Ethernet, once reserved for network-to-network connections, is now a viable alternative for server access to high-speed networks. Latency is low, as is latency jitter. Throughput can reach wire speed with TCP and 3,200-byte messages while UDP rates seem to be limited to about 40% of that until reliable kernel bypass libraries are available. UDP multicast provides the advantage of wire-speed copying for all ports of modern switches, delivering linear growth in receive rates as receivers are added.

The benchmark numbers given here show what we measured using equipment in our testing labs. The numbers that really matter are those run on your production equipment. We invite you to try our messaging software on your production equipment through our free evaluation program.

Detailed Test Setup

Hardware Setup

The system consists of two Dell Precision T3400 and two Dell Precision 390 machines networked with 10-gigabit Ethernet.

These computers had the following configuration:

Hardware Configuration Table

Vendor Model	Dell Precision T3400
Processors	Intel® Core™ 2 Quad Q6600 @ 2.40 GHz
RAM	2 GB
Cache	8192 KB (2 x 4096 KB)
Operating System	Centos 5, 2.6.18-53.1.21.e15
Vendor Model	Dell Precision 390
Processors	Intel® Core™ 2 E6600 @ 2.40 GHz
RAM	2 GB
Cache	4096 KB
Operating System	RHEL 4, 2.6.9-67.0.7.ELsmp

The network had the following specifications:

Network Configuration Table

Ethernet Switch	Cisco Nexus 5020
Ethernet NIC	NetEffect NE020.LP.1.XSR

Software Configurations

The measurements were performed using version 3.3.6 of 29West, Inc. Latency Busters® Messaging.

LBM Product Background

29West, Inc. Latency Busters® Messaging (LBM) product is a messaging system and API designed for high message rates and low latency. It contains a number of innovative latency-reducing features, chief among them the elimination of daemons and servers. 29West's messaging products can use a variety of transport protocols including TCP, reliable unicast and reliable multicast. This report focuses on the performance of 29West's reliable multicast protocol under varying message sizes and rates.

29West messaging utilizes a number of key architectural advancements to provide performance far in excess of traditional messaging products. In particular, it:

- Eliminates the need for messaging routers and intermediate messaging daemons, removing latencies incurred from such hops.
- Leverages advancements in operating system and hardware capabilities by having the messaging layer share the application's address space, rather than requiring special daemon processes.
- Utilizes the network infrastructure for message routing, as opposed to using software to duplicate routing functions within the messaging layer.

About 29West

As applications evolve to require new data delivery models, more performance, and lower latency, the messaging layer must be adaptable and flexible in its delivery mechanism and designed from the ground up for efficiency. 29West was founded in 2002 in response to the financial services industry's need for a new messaging design to handle the growing data rates and ever lower latency requirements. The team at 29West started with 20 years of financial market data delivery experience, a clean sheet of paper and a single goal: to build the highest throughput, lowest latency, most stable messaging product for the financial markets.

Headquartered in Warrenville, Illinois (a suburb of Chicago), with offices in New York, London and Tokyo. 29West, Inc. employs some of the most innovative messaging architects in the industry. Led by President and Founder Mark Mahowald, a former Talarian executive and founder of WhiteBarn, and principle software architect Todd Montgomery, co-founder of GlobalCast Communications and also formerly of WhiteBarn, Talarian and Tibco, the 29West team has been involved in financial market data distribution from the first digital trading floors of the mid 1980s.

Since June of 2004, when we first announced Latency Busters® Messaging (LBM) and our first customer, we have been setting the performance standard in the market with innovative design ideas and the absolute highest performance along with customer focused support. Deployed in mission-critical production applications at over 100 financial market firms world-wide, 29West is the performance leader in high-performance messaging.

With LBM and now Ultra Messaging® for the Enterprise (UME), the next-generation design for persistent messaging and applications that require delivery confirmation, "guaranteed delivery" and durable subscription messaging, 29West offers the application developer the most-flexible and lowest-latency messaging solutions in the industry. Whether your application requires a streaming messaging model or you have a need for persistence and delivery confirmation, our design breakthroughs remove the choke points found in legacy messaging and unleash the power of your network to efficiently route high-speed messaging traffic.

Many banks, hedge funds and exchanges have ported their applications to the 29West messaging framework, and are attesting to dramatic performance improvements as well as greater network stability when driving unusually high data loads. With server bottlenecks, daemons, extra data copies and other messaging software choke points removed, 29West Messaging is proving to be the next-generation network solution for the financial services industry.

Acknowledgements

29West would like to thank Cisco for providing the Nexus 5020 switch used in collecting the data for this report.

29West would like to thank NetEffect for providing the NE020 10GbE adapters used in this test.

All trademarks and copyrights are the property of their respective owners.

Contact Information

For more information on 29West regarding our messaging software or to request a Free Software Evaluation, please call or e-mail us at an office listed below. If you would like to be added to our mailing list or receive copies of networking and messaging white papers we create, please e-mail us at sales@29West.com.

Headquarters:

29W110 Butterfield Road
Suite 306
Warrenville, IL 60555
Phone: 630-836-2990 Fax: 630-836-7508
E-mail: info@29west.com

New York:

One Liberty Plaza, 23rd Floor
New York, NY 10006
Phone: 212-835-9485
E-mail: kclancey@29west.com

London:

2 Finch Lane
London, EC3V 3NA
Phone: 44 (0) 207 763 7001
E-mail: annalisa@29west.com

Tokyo:

12th floor Yurakucho ITOCiA
2-7-1 Yurakucho, Chiyoda-ku
Tokyo, 100-0006 Japan
Phone: +81(0)3 6860 4692
E-mail: swatanabe@29west.com

Hong Kong:

1201 Jubilee Centre
18 Fenwick Street
Wanchai, Hong Kong
Phone: +852 2376 3232
E-mail: kennethlau@serisys.com