10G Ethernet: The Foundation for Low-Latency, Real-Time Financial Services Applications and Other, Future Cloud Applications

Testing conducted by Solarflare® Communications and Arista Networks shows that microsecond, Infiniband-like latency can be achieved with commercially available 10G Ethernet Switch and Server Adapter Products

Bruce Tolley, PhD, Solarflare Communication

Abstract

Solarflare Communications, the leading vendor of standards-based 10 Gigabit Ethernet (10GbE) silicon and Arista Networks, the leading vendor of cloud networking solutions for large datacenter and computing environments, recently completed switch to server latency testing. The test found that the Solarstorm® server adapter in combination with the Arista DCS-7124Sswitch achieved 6.5 microsecond mean latency over 10GbE in a UDP multicast latency test. The Arista switch is the lowest latency switch on the market, demonstrating mean latency in this test of 560 nanoseconds. With back-to-back servers, the server adapter achieved impressive minimum latencies as low at 5.5 microseconds. The latency of the overall system was also very deterministic with 99% of the messages being delivered with latency less than 7.1 microseconds. Solarflare and Arista measured the performance of UDP multicast messaging using Solarflare developed benchmarks with commercially available products: the Solarflare Solarstorm SFN4112F 10 Gigabit server adapters and an Arista DCS-7124S 10 Gigabit Ethernet switch. The test platform used servers and processors typically found in use by financial firms today. Solarflare and Arista also expect low-latency UDP multicast to find application in public and private clouds that demand support for real-time interactive performance.

The Need for Low Latency in Automated, Real Time Trading

The rapid expansion of automated and algorithmic trading has increased the critical role of network and server technology in market trading, first in the requirement for low latency and second in the need for high throughput in order to process the high volume of transactions. Given the critical demand for information technology, private and public companies that are active in electronic markets continue to invest in their LAN and WAN networks and server infrastructure that carries market data and trading information.

In some trading markets, firms can profit from less than one millisecond of advantage over competitors, which drives them to search for sub-millisecond optimizations in their trading systems. The spread of automated trading across geographies and asset classes, and the resulting imperative to exploit arbitrage opportunities based on latency, has increased the focus on if not created an obsession with latency.

With this combination of forces, technologists, IT and data center managers in the financial services sector are constantly evaluating new technologies that can optimize performance. One layer of the technology stack that receives continuous scrutiny is messaging, i.e., the transmission of information from one process to another, over networks with specialized home-grown or commercial messaging middleware.

The ability to handle predictably the rapid growth of data traffic in the capital markets continues to be a major concern. As markets become more volatile, large volumes of traffic can overwhelm systems, increase latency unpredictably, and throw off application algorithms. Within limits, some algorithmic trading applications are more sensitive to the predictability of latency than they are to the mean latency. Therefore it is very important for the stack to perform not just with low latency but with bounded, predictable latency. Solarflare Communications and Arista Networks demonstrate in this paper that because of its low and predictable latency, a UDP multicast network built with 10 Gigabit Ethernet (10GigE) can become the foundation of messaging systems used in the financial markets.

Financial services applications and other applications that can take advantage of low-latency UDP multicast

Messaging middleware applications were named above as one key financial services application that produce and consume large amounts of multicast data that can take advantage of low-latency UDP multicast. Other applications in the financial services industry that can take advantage of low-latency UDP multicast data include:

- Market data feed handler software that takes as input multicast data feeds and uses multicasting as the distribution mechanism
- Caching/data distribution applications that use multicast for cache creation or to maintain data state
- Any application that makes use of multicast and requires high packets per second (pps) rates, low data distribution latency, low CPU utilization, and increased application scalability
-

Cloud Networking and the broader market implications of low latency to support real-time applications

As stated above, the low-latency UDP multicast solution provide by Arista switches and Solarflare Solarstorm server adapters can provide compelling benefit to any application that depends on multicast traffic where additional requirements exist for high throughput, low-latency data distribution, low CPU utilization, and increased application scalability. Typical applications that benefit from lower latency include medical imaging, radar and other data acquisition systems, and seismic image processing in oil and gas exploration. Yet moving forward, cloud networking is a market segment where requirements for throughput, low latency and real time application performance will also develop. The increasing deployment and build out of both public and private clouds will drive the increased adoption of social networking and Web 2.0 applications. These cloud applications will incorporate real-time media and video distribution and will need lower latency applications for both business to consumer (B2C) and business to business (B2B) needs. In both the business and the customer cases, the requirement for low latency and real time application response will only become stronger in the next year or two.

Solarflare's OpenOnload™ Defined

Solarflare and Arista measured the latency performance of messaging using Solarflare-developed benchmarks with commercially available products: the Solarflare Solarstorm SFN4112F SFP+ 10 Gigabit server adapters and an Arista DCS-7124S 10 Gigabit switch. A list of the hardware configurations and the benchmarks used is attached as an Appendix. The test platform used servers and processors typically found in use by financial firms today. The tests described below were run both switch to server adapter and server adapter to server adapter. The adapters were run in kernel mode and in OpenOnload mode.

OpenOnload is an open-source high-performance network stack for Linux created by Solarflare Communications. By improving the CPU efficiency of the servers, OpenOnload enables applications to leverage more server resources, resulting in dramatically accelerated application performance without changing the existing IT infrastructure. Using standard Ethernet, the solution combines state-of-the-art Ethernet switching and server technologies that dramatically accelerate applications. OpenOnload performs network processing at user-level and is binary-compatible with existing applications that use TCP/UDP with BSD sockets. It comprises a user-level shared library that implements the protocol stack, and a supporting kernel module.

Fundamental Findings

Network stack	Link	Min	Median	Mean	95th percentile	99th percentile
Kernel	Back to back	9801	10074	10185	10683	12075
Kernel	Switch	10233	10563	10745	11374	12704
Onload	Back to back	5517	5697	5884	6399	6498
Onload	Switch	6122	6327	6514	7030	7123

Exhibit 1: Half-Round Trip Latency in Nano Seconds

Exhibit 1 summarizes results of TCP latency testing. The Arista DCS7124 is a very low-latency switch contributing a mean latency of 560 nanoseconds to the system latency. In the testing for the 70 byte message sizes typical of market data messaging systems, very low latency was observed. The Solarstorm server adapter in combination with the Arista DCS 7124S switch achieved mean latency of 6.5 microseconds. The Solaflare adapters back to back achieved an amazingly low mean latency of 5.9 microseconds. This latency was also very deterministic with 95% of the messages being delivered with a mean less than 7.0 microsecond and 99% of the messages with mean latency less than 7.1 microseconds in the switch to server adapter configuration.

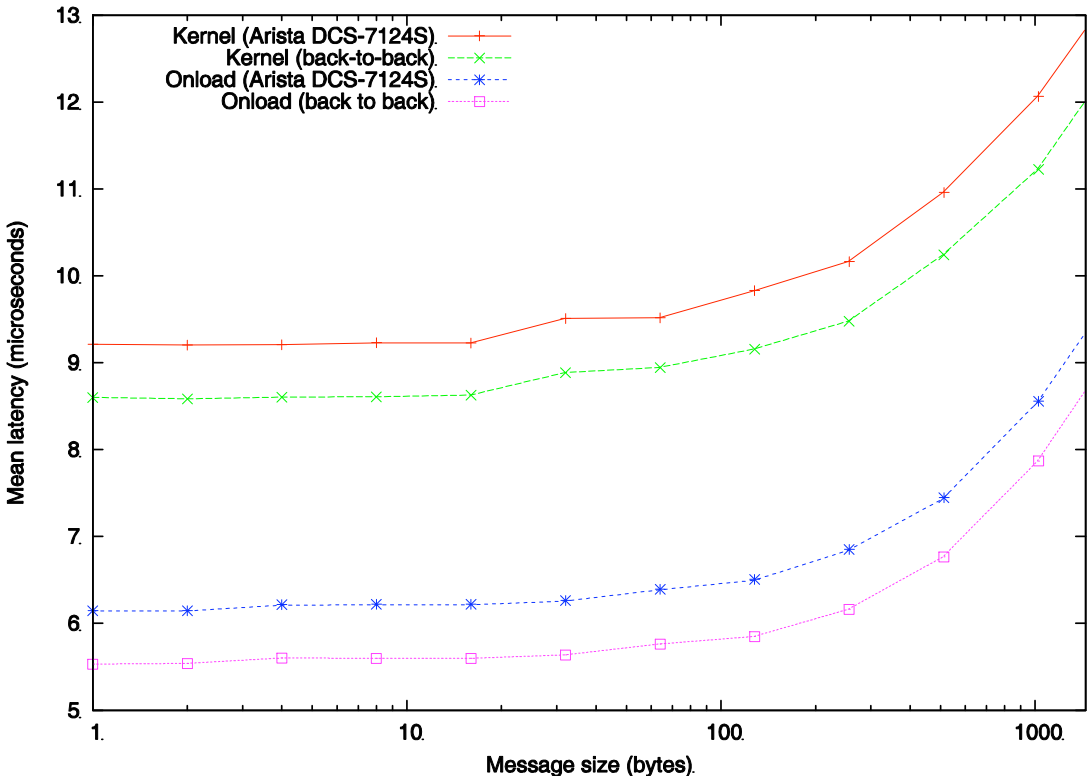

Exhibit 2: UDP Half Round Trip Latency

Exhibit 2 above plots UDP half round trip latency where the x axis represents message size in bytes and the y axis represents latency in microseconds. The data shows that the system demonstrated very low and deterministic latency from small up to very large message sizes of 1472 bytes. The data plot also shows very low latency in both kernel and OpenOnload mode. In OpenOnload mode with the switch and server adapter, minimum latencies go as low as 6 microseconds, and with the server adapters back to back as low as 5.5 microseconds.

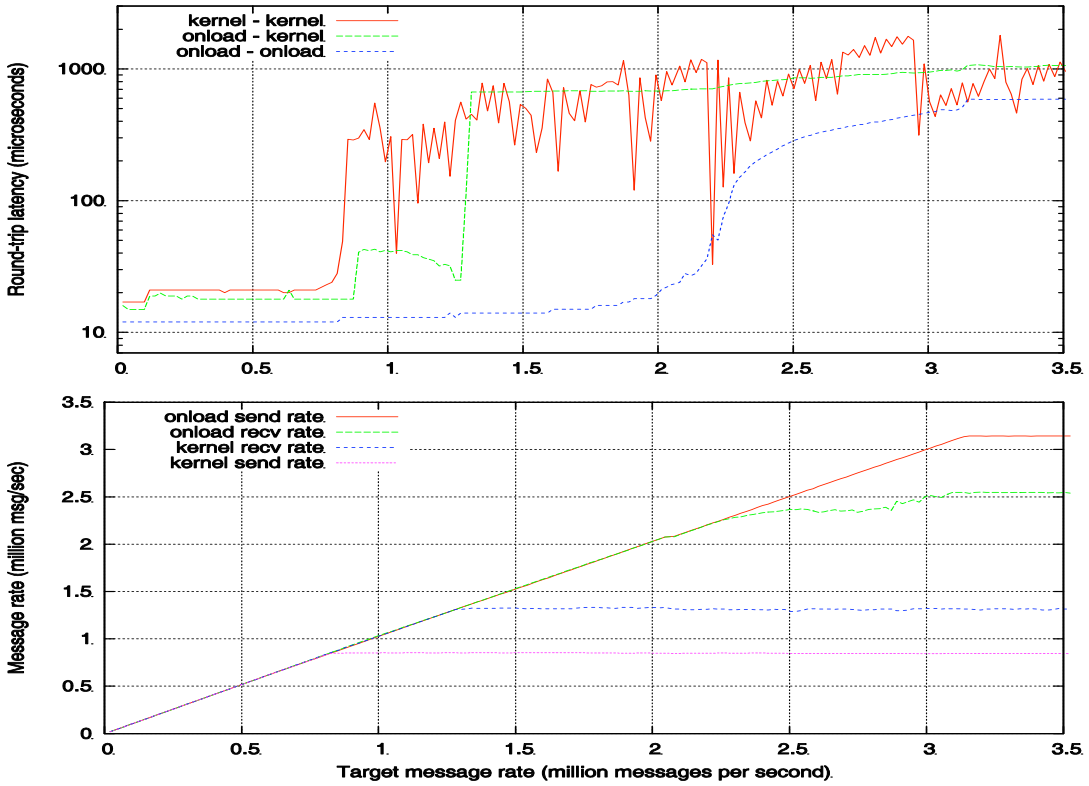

Exhibit 3: Message Rates Achieved with Upstream UDP

Exhibit 3 above shows two plots of performance versus desired data rate of UDP multicast performance with and without OpenOnload. This test simulates a traffic pattern that is common in financial services applications. In the test, the system streams small messages from a sender to a receiver. The receiver reflects a small proportion of the messages back to the sender, which the sender uses to calculate the round-trip latency. The x-axis shows the target message rate that the sender is trying to achieve. The y-axis shows one-way latency (including a switch) and the achieved message rate. The kernel results are measured with Solarflare server adapters without OpenOnload. The plot combines results from three runs: kernel to kernel, OpenOnload to kernel, and OpenOnload to OpenOnload. The OpenOnload to kernel test is needed in order to fully stress the kernel receive performance.

The top plot labeled Round Trip Latency shows the improved, deterministic low latency achieved with the Solarflare adapter, OpenOnload, and the Arista switch. The y-axis shows the round trip latency while the x axis shows the desired message rate in millions of messages per second at the receiver. With OpenOnload, not only is the system performing at much lower latency, but also the latency is predictable and deterministic over the range of expected message rates. This is precisely the attribute desired in trading systems or any other application demanding real time performance.

The second plot in Exhibit 3, Message Rate Achieved shows the Solarflare OpenOnload system's ability to scale and perform as the message rate is increased. This is in contrast to the kernel stack where the greater CPU processing overheads of the stack limit performance as higher levels of load are put on the system.

Arista DCS 7124: Industry Leading Performance, Scalability, and High Availability

The Arista DCS-7124S is recognized in the industry as a best-in-class multicast switch with the lowest packet latencies on the market. The Arista 7100 Series of Datacenter Ethernet switches feature the industry's highest density, lowest latency 10 Gigabit Ethernet switching solution and the first with an extensible modular network operating system. With breakthrough price-performance, the Arista 7100 Series enables 10 Gigabit Ethernet to be deployed everywhere in the data center, which can significantly improve server utilization and consequently data center power efficiency.

Arista switches run the EOS™ (Extensible Operating System), designed from the ground up to provide a foundation for the business needs of next-generation datacenters and cloud networks. EOS is a highly modular software design based on a unique multi-process state sharing architecture that completely separates networking state from the processing itself. This enables fault recovery and incremental software updates on a fine-grain process basis without affecting the state of the system.

Arista EOS provides extremely robust and reliable data center communication services while delivering security, stability, openness, modularity and extensibility. This unique combination offers the opportunity to significantly improve the functionality and evolution of next generation data centers.

The Solarflare Solution

The Solarstorm 10GbE PCI-Express server adapter was designed to support virtualization and iSCSI protocol acceleration with line-rate performance and low power consumption of just 7 Watts. In both kernel and OpenOnload modes, the adapter supports financial services and HPC applications that demand very low latency and very high throughput. Tests were performed using the standard kernel IP stack as well as Solarflare OpenOnload technology.

OpenOnload is an open-source high-performance network stack for Linux created by Solarflare Communications. As Exhibit 4 shows, the OpenOnload software provides an optimized TCP/UDP stack into the application domain which can communicate directly with the Solarflare server adapter. With Solarflare's OpenOnload, the adapter provides the application with protected, direct access to the network, bypassing the OS kernel, and hence reducing networking overheads and latency.

The typical TCP/UDP/IP stack resides as part of the kernel environment and suffers performance penalties due to context switching between the kernel and application layers, the copying of data between kernel and application buffers, and high levels of interrupt handling.

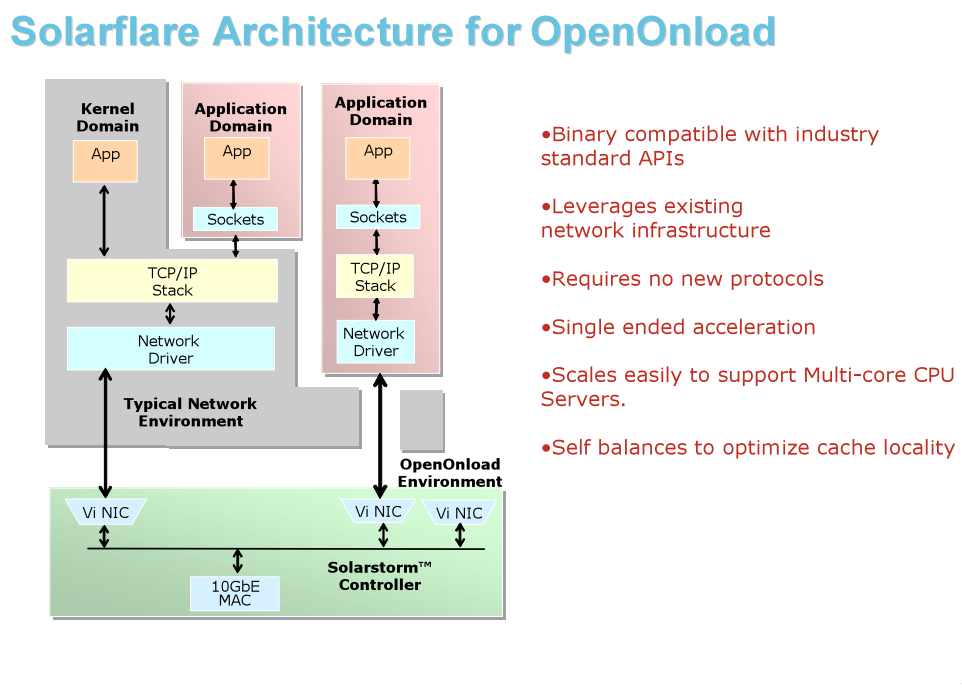

Exhibit 4: The Solarflare Architecture for OpenOnload

The kernel TCP/UDP/IP and OpenOnload stacks can co-exist in the same environment. This co-existence allows applications that require a kernel-based stack to run simultaneously with OpenOnload. This coexistence feature was leveraged as part of the testing where the benchmarks were run through both the kernel and OpenOnload stacks in back to back fashion using the same build and without having to reboot the systems.

Conclusions

The findings analyzed in this white paper represent the results of testing of transmit latency of a configuration with the Solarflare Solarstorm server adapter with OpenOnload and the Arista DCS-7124S at transmission rates up to 3 million messages/second (mps). For the 70-byte message sizes typical of market data messaging systems, very low latency was observed:

- Mean did not exceed 6.5 microseconds with switch
- Mean did not exceed 5.8 microseconds without switch
- 99th percentile did not exceed 7.1 microseconds with switch

The system demonstrated therefore very bounded jitter and very low UDP multicast latency which delivers very predictable messaging systems.

With Solarflare's server adapter and OpenOnload technology, off-the-shelf 10GbE hardware can be used as the foundation of messaging systems for electronic trading with no need to re-write applications or use proprietary, specialized hardware.

Enabling financial trading customers to implement highly predictable systems, Solarflare's and Arista's10GbE solutions provide a competitive advantage and offer increased overall speeds, more accurate trading and higher profits. Now, financial firms can use off-the-shelf Ethernet, TCP/IP, UDP and multicast solutions to accelerate market data systems without requiring the implementation of

new wire protocols or changing applications. By leveraging the Solarstorm server adapter with OpenOnload, IT managers are able to build market data delivery systems designed to handle increasing message rates, while reducing message latency and jitter between servers.

Summary

Solarflare Communications and Arista Networks have demonstrated performance levels with 10 Gigabit Ethernet that enables Ethernet to serve as the foundation of messaging systems used in the financial markets. Now, financial firms can use off-the-shelf Ethernet, TCP/IP, UDP and multicast solutions to accelerate market data systems without requiring the implementation of new wire protocols or changing applications. With off the shelf 10GbE gear, Solarflare's server adapter and the Arista switch can be used as the foundation of messaging systems for electronic trading and the support of low-latency UDP multicast with no need to re-write applications or use proprietary, specialized hardware. IT and data center managers can deploy plain old Ethernet solutions today.

Moving forward, Solarflare Communications and Arista Networks also expect high performance 10G Ethernet solutions with low-latency UDP multicast to become an important technology component of public and private clouds that rely on real time media distribution for business to consumer and business-to-business applications.

About Solarflare Communications

Solarflare Communications is a semiconductor company delivering the next level of high-performance 10 Gigabit Ethernet (10GbE). As the leading provider of standards-compliant 10GbE silicon, Solarflare's robust and power-efficient solutions are cost effective and easy to deploy. Ready for primetime, Solarflare's 10 Gigabit Ethernet solutions make possible next-generation applications such cloud computing, server virtualization, network convergence and low-latency UDP multicast (with OpenOnload™) for market data applications. The privately held company is headquartered in Irvine, California, with a development center in Cambridge, UK, and has announced relationships with Accton, Citrix, CommScope, Delta Networks Inc., Panduit, SMC Networks and VMware. For more information, contact Solarflare at productinfo@solarflare.com or 949-581-6830.

About Arista Networks

Arista Networks delivers cloud networking solutions for large datacenter and computing environments. Arista offers the best-of-breed 10 Gigabit Ethernet switches that redefine scalability, robustness, and price-performance. At the core of Arista's platform is the Extensible Operating System (EOS™), pioneering new software architecture with self-healing and live in-service software upgrade capabilities. For more information, please visit www.aristanetworks.com or contact Arista at info@aristanetworks.com or 650-462-5000.

Appendix: Hardware Configuration and List of Benchmarks

Hardware		Type	Comments
Servers			
	Processor	2 x Intel Xeon x5482 (3.2GHz)	
	RAM	4 GBytes	
	Chipset	Harpertown 32K L1, 6 Mb L2	
Switch			
	Model	Arista DCS7124	
NICs			
	Model	Solarstorm SFN4112F	SFP+
	Vendor	Solarflare	
	PCI Bus	PCI-e x8	

Software		Type	Comments
Operating System			
	Linux	Red Hat Enterprise Linux Client release 5.1 (64 Bit)	
Middleware			
	Solarflare	OpenOnload	
Benchmarks			
	Latency	udprtt	Generates Traffic and measures Round Trip latency with respect to different UDP payloads
	Message Throughput	udpstream	Generates traffic and measures message rate throughput
	Bandwidth	Udpsend/udpswallow	Generates traffic and measures bandwidth

Acknowledgements

The author would like to acknowledge David Riddoch, PhD, Solarflare Communications, for running the benchmarks and his technical contributions to the paper as well as Mansour Karam, Arista Networks, Steve Pope, PhD, Solarflare Communications, and George Zimmerman, PhD, Solarflare Communications, for their technical reviews and contributions to the paper.