10Gb Ethernet: The Foundation for Low-Latency, Real-Time Financial Services Applications and Other, Latency-Sensitive Applications

Testing conducted by Solarflare and Arista Networks reveals single-digit microsecond TCP and UDP applications latency can be achieved with commercially available 10Gb Ethernet Switch and Server Adapter Products

Bruce Tolley, PhD, Solarflare

Abstract

Solarflare, the leader in application-intelligent 10 Gigabit Ethernet (10GbE) networking hardware and software, and Arista Networks, the leading vendor of ultra low-latency 10GbE and cloud networking solutions recently completed low-latency switch and adapter testing. The two companies collaborated to address the growing need for low-latency, high-performance 10GbE switch-to-server solutions for high-frequency trading and other demanding applications, such as public clouds, virtualization, and big data.

Conducted by Solarflare and Arista Networks in Solarflare's Cambridge, UK labs, the test results showed that single digit microsecond UDP and TCP application latency can be achieved with commercially available TCP/IP Ethernet products. Application performance of TCP/UDP messaging was measured using the Solarflare® SFN5122F 10GbE server adapter with OpenOnload® middleware and the Arista DCS-7124SX 10GbE switch. The testing used servers and processors widely available for clusters and data centers today. Running various scenarios, Solarflare and Arista Networks found:

- In half round trip testing, the Arista DCS-7124SX switch demonstrated mean application latency as low as 520 nanoseconds with 64-byte packets typical of messaging applications.
- With back-to-back servers, the Solarflare SFN5122F adapter achieved mean TCP latencies as low as 3.1 microseconds.
- In TCP testing, server-to-switch-to-server mean latency was as low as 3.6 microseconds.
- In UDP testing, with back-to-back servers, the Solarflare SFN5122F adapter achieved mean UDP latencies as low as 2.9 microseconds.
- The server-to-switch-to-server mean UDP latency was as low as 3.4 microseconds.

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 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, network engineer, 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 homegrown or commercial messaging middleware.

The ability to predictably handle 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 critical 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 (10GbE) 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 receive 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 provided by Arista switches and Solarflare 10GbE server adapters can deliver 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 financial trading, signals intelligence and transparent data acquisition, and seismic image processing in exploratory geophysics. 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 consumer cases, the requirement for low latency and real time application response will only become stronger in the next year or two.

Using standard Ethernet, the solution combines state-of-the-art Ethernet switching and server technologies to dramatically accelerate applications. Solarflare and Arista measured the latency performance of messaging using Solarflare-developed benchmarks with commercially available products: the Solarflare SFN5122F SFP+ 10 Gigabit server adapters and an Arista DCS-7124SX 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 in both switch-to-server adapter and server adapter-to-server adapter configurations. The adapters were run in kernel mode and in OpenOnload mode.

Solarflare's OpenOnload Defined

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. 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.

Application Transparency and Protocol Conformance

Solarflare has built application transparency and protocol conformance into its OpenOnload application acceleration middleware. To avoid the requirement of managing OpenOnload enabled sockets as a separate network interface, an abstraction is maintained of a single physical network interface for which particular network flows are accelerated onto a virtual interface mapped into a user-address space. Therefore, management of the physical network interface is performed using standard Linux networking tools and operations that maintains full access to and compatibility with the control plane of the Linux kernel. This architecture enables OpenOnload to maintain full support for standard enterprise networking features and topologies VLAN and teaming status, the ARP cache, ICMP notifications, and the IP routing tables.

Solarflare developed OpenOnload's TCP/UDP implementation from the ground up since 2002 specifically for hybrid user-space/kernel operation; Open Onload supports the latest Linux performance features and ensures robust and full RFC compliance.

Fundamental Findings

Exhibit 1 summarizes the results of TCP latency testing. The Arista DCS 7124SX is a very low-latency switch contributing a mean latency as low as 520 nanoseconds to the system latency. In the testing for the 64-byte message sizes typical of market data messaging systems, very low latency was observed. The Solarflare 10GbE server adapter in combination with the Arista DCS 7124SX switch achieved mean latency of 3.6 microseconds. The Solaflare adapters back to back achieved an amazingly low mean latency of 3.1 microseconds. This latency was also very deterministic with 99% of the messages being delivered with a mean less than 3.9 microseconds in the switch to server adapter configuration. Furthermore the standard deviation of latency (jitter) is reduced over 83%, decreasing from 479 nanoseconds to 79.

Exhibit 1: Half-Round Trip TCP Latency in Nano Seconds (64 byte)

Network stack	Link	Min	Median	Mean	99th percentile	STD
Kernel	Back to back	7518	7987	8114	10332	490
Kernel	Switch to server	8209	8536	8663	10840	479
Onload	Back to back	2997	3103	3122	3376	91
Onload	Switch to server	3546	3679	3695	3925	79

Exhibit 2: Half-Round Trip UDP Latency in Nano Seconds (64 byte)

Network stack	Link	Min	Median	Mean	99th percentile	STD
Onload	Back to back	2866	2967	2976	3244	82
Onload	Switch to server	3370	3483	3496	3714	78

Exhibit 2 shows the results of the UDP latency testing. With very low UDP application latency, the Arista and Solarflare system can support the IP multicast traffic critical to high frequency trading market data as well as other point to multipoint traffic patters requiring high message rates and low latency. In half round trip testing, the Arista DCS-7124SX switch demonstrated mean application latency as low as 520 nanoseconds with 64-byte packets. In this UDP testing, with back-to-back servers, the Solarflare SFN5122F adapter achieved mean UDP latencies as low as 2.9 microseconds. For the server-to-switch link, the mean UDP latency was as low as 3.4 microseconds.

Exhibit 3: TCP Half Round Trip Latency

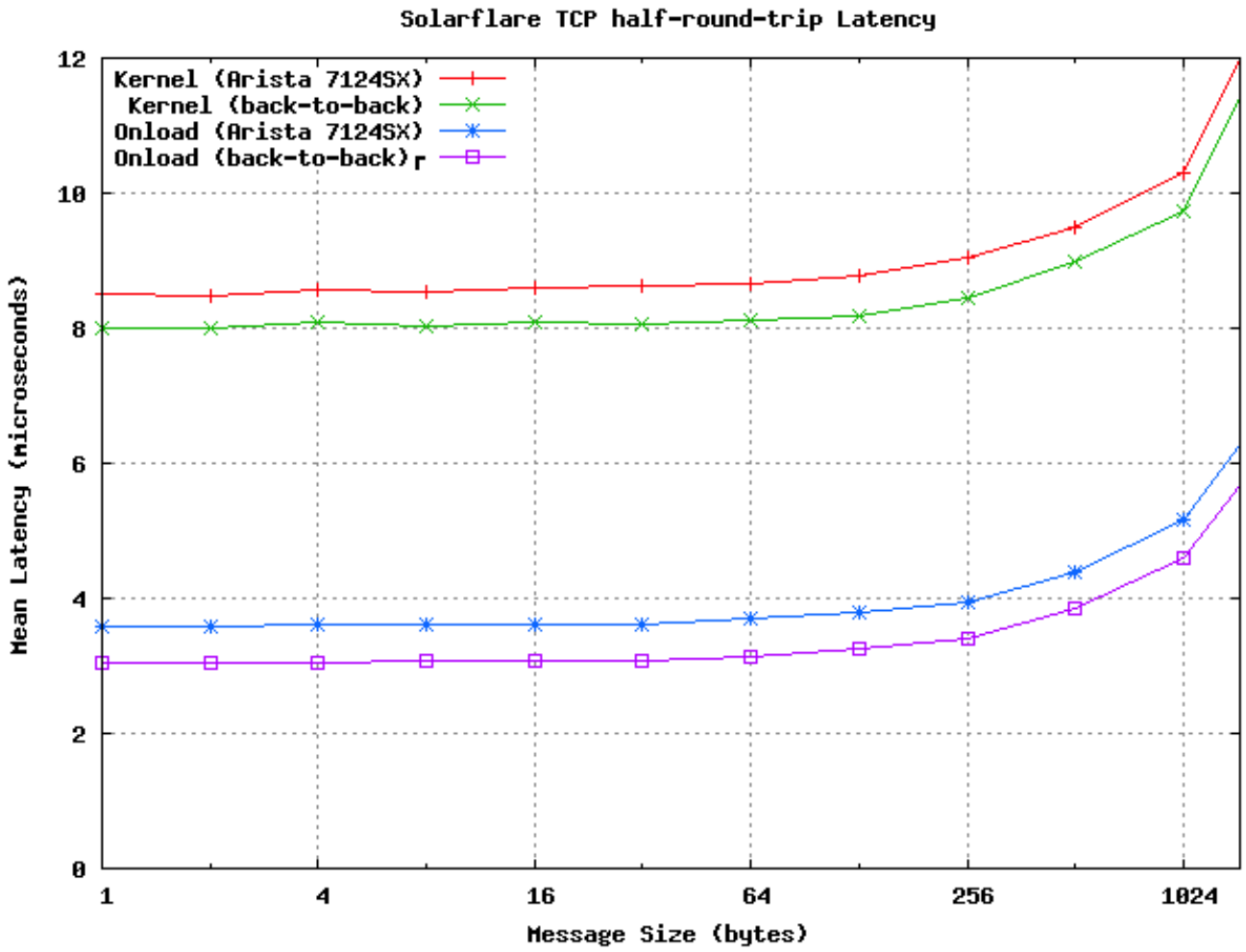


Exhibit 3 above plots TCP half round trip latency where the x-axis represents message size in bytes and the y-axis represents latency in microseconds. The data shows that the system demonstrated very low and deterministic latency from small up to very large message sizes of 1472 bytes. The latency is relatively constant even as the load on the system is increased from 1000s to several million packets per second. 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 3.5 microseconds, and with the server adapters back to back as low as 2.9 microseconds.

Exhibit 4: Message Rates Achieved with Upstream UDP

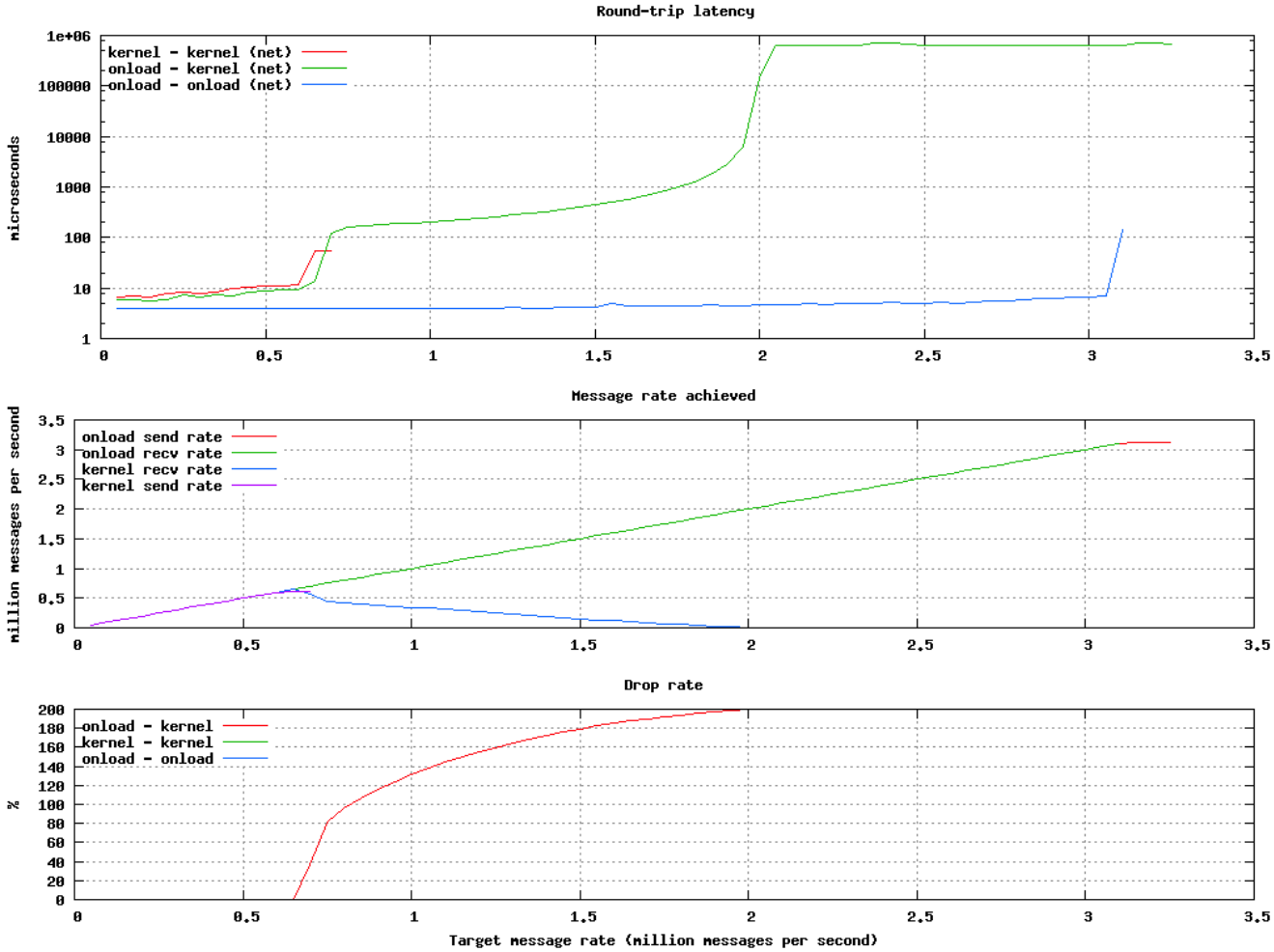


Exhibit 4 above shows two plots of performance versus desired data rate of UDP multicast performance with and without OpenOnload. This test simulates a traffic pattern that is common in financial services applications. In the test, the system streams small messages from a sender to a receiver. The receiver 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 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 4, 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-7124SX 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 Data Center 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 throughout the data center, which can significantly improve server utilization and consequently data center power efficiency.

Arista switches run EOS™ (Extensible Operating System), the world's most advanced network operating system which is designed from the ground up to provide a foundation for the business needs of next-generation data centers 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.

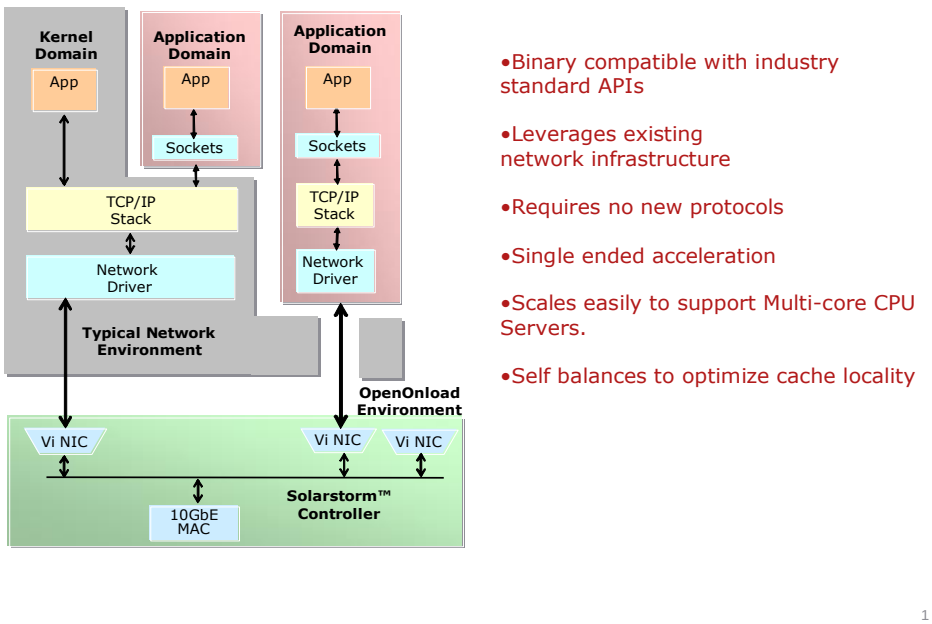
The Solarflare Solution

The SFN5122F 10GbE server adapter was designed to support low latency and high performance and low power consumption of just 7 Watts. In both kernel and OpenOnload modes, the adapter supports financial services and HPC applications which demand very low latency and very high throughput. Tests were performed using the standard kernel IP stack as well as Solarflare OpenOnload middleware.

OpenOnload is an open-source high-performance network stack for Linux created by Solarflare Communications. As Exhibit 5 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 5: 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 half round trip latency of a configuration with the Solarflare server adapter with OpenOnload and the Arista DCS-7124SX at transmission rates up to 3 million messages/second (mps). The findings analyzed in this white paper represent the results of testing of latency of a configuration with the Solarflare server adapter with OpenOnload and the Arista DCS-7124SX at transmission rates up to 3 million messages/second (mps). For the 64-byte message sizes typical of market data messaging systems, very low TCP and UDP application latency was observed:

- Mean did not exceed 3.6 microseconds with switch
- Mean did not exceed 3.1 microseconds without switch
- 99th percentile did not exceed 3.9 microseconds with switch
- For UDP performance, the numbers were 3.4, 2.9, and 3.7 respectively

The system demonstrated 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 Solarflare 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 and Arista Networks have demonstrated performance levels with 10 Gigabit Ethernet that enable 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. Therefore, IT and data center managers can deploy standard 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

Solarflare is the leading provider of application-intelligent networking I/O products that bridge the gap between applications and the network, delivering improved performance, increased scalability, and higher return on investment. The company's solutions are widely used in scale-out server environments such as high frequency trading, high performance computing, cloud, virtualization and big data. Solarflare's products are available from leading distributors and value-added resellers, as well as from Dell, IBM, and HP. Solarflare is headquartered in Irvine, California and operates an R&D facility in Cambridge, UK.

About Arista Networks

Arista Networks delivers cloud networking solutions for large datacenter and computing environments. Arista offers 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, a 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: List of Benchmarks

Software		Type	Comments
Operating System			
	Linux	Red Hat Enterprise Linux Client release 6.1 (64 Bit)	
Middleware			
	Solarflare	OpenOnload	
Benchmarks			
	Latency	sfnt-pingpong	Generates Traffic and measures Round Trip latency with respect to different UDP payloads
	Message Throughput	sfnt-stream	Generates traffic and measures message rate throughput
	Bandwidth	Udpsend/udpswallow	Generates traffic and measures bandwidth

Acknowledgement

The author would like to acknowledge Mounir Maaref of Solarflare, for running the benchmarks and his technical contributions to the paper.