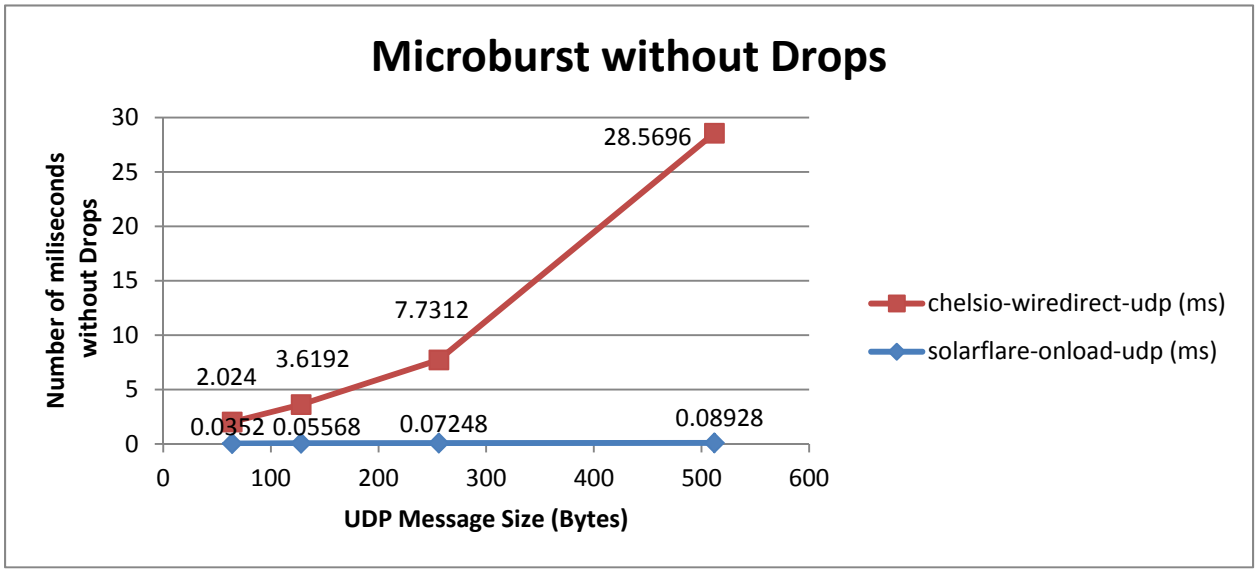
Chelsio T520-LL-CR vs. Solarflare SFN7122F

Microburst, Latency and Message Rate Competitive Benchmark Results

Executive Summary

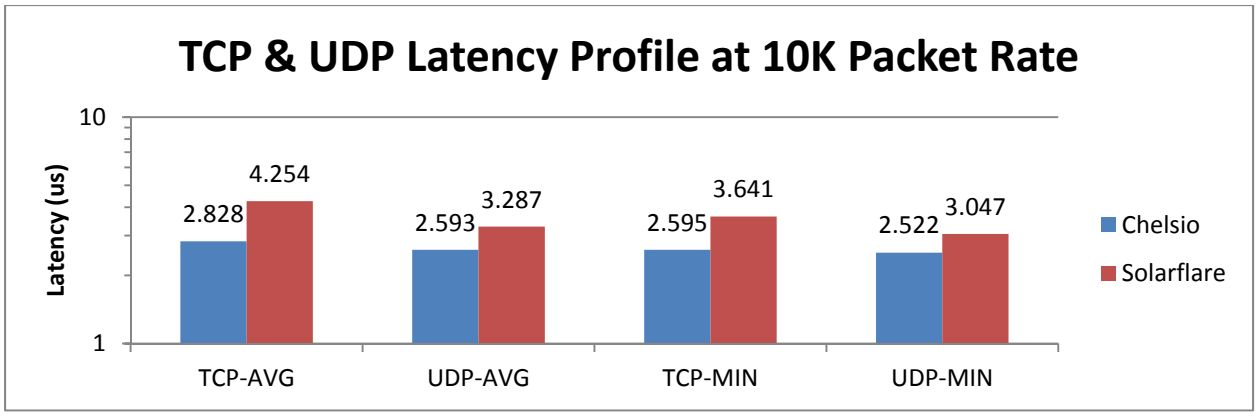
This paper provides an overview of the technology behind Chelsio's low latency solution, and contrasts it to the latest offering from Solarflare. Chelsio's solution is unique in combining hardware and software techniques to provide a robust high performance solution. This paper presents benchmark results which demonstrate a superior performance profile for Chelsio that is both more consistent, and more resilient to actual network conditions. The following two graphs show two representative metrics that highlight the advantages of Chelsio's solution.

In microburst tolerance testing at 256B packet size, Chelsio's adapter sustains more than 100 times larger bursts than Solarflare.



In latency testing at 10K packet/sec load, Chelsio shows more than 1.4us lower TCP average latency than the Solarflare, and about 0.7us lower average UDP latency.

** PRELIMINARY **



Overview

Chelsio is the leading provider of Ethernet network protocol offload technologies. Chelsio's Terminator *TCP Offload Engine (TOE) with WireDirect and RDMA* is the first and currently only network engine capable of full TCP/IP, UDP/IP, Multicast and RDMA offload at 1/10/40Gbps.

TCP Offload

The unique ability of a *TOE* to perform the full transport layer functionality in hardware is essential to obtaining tangible benefits. The vital aspect of offloading the transport layer is in it being the process-to-process layer, i.e. the data passed to the *TOE* comes straight from the application process, and the data delivered by the *TOE* goes straight to the application process. In contrast, lower layers only provide unreliable delivery services, limiting the effectiveness of offload at these levels.

User Space I/O

Application (user) space I/O has significant performance benefits. Chelsio's WireDirect implementation provides zero-copy, kernel bypass and application polling capabilities from user-space for the fastest application response to network data. WireDirect UDP uses Chelsio's RDMA Queue Pairs semantics to expose a UDP socket layer API to the application, while WireDirect TOE TCP uses Chelsio's TOE interface to expose a TCP socket layer API to the application. This allows user applications to achieve lower latency and higher message rates for TCP, UDP and Multicast. Thanks to WireDirect using hardware acceleration, it is uniquely capable of handling microbursts and providing deterministic performance under load.

RDMA

Chelsio's iWARP RDMA over Ethernet capability enables a user process on one system to transfer data directly between its virtual memory and the virtual memory of a process on another system without operating system intervention on either side of the communication. RDMA accomplishes this by offloading onto the "channel adapter" interface card the tasks traditionally performed by the operating system during network transfers. The result is high throughput, high message rate, low latency, and low CPU utilization message transfer.

Offload vs. Onload

Offload refers to moving compute intensive workloads off a system's CPU and onto specialized hardware, such as a graphics card, disk controller or network adapter. Although attempts have been made to market schemes that run such specialized workloads on the CPU, this has generally failed to compete with well integrated offload hardware. In contrast, the term "Onload" has recently been coined to refer to operating a system that lacks offload hardware, typically in the network interface.

In the context of networking, there are clear performance benefits to providing network access to user space. RDMA, in particular, is a protocol that provides a user-space I/O interface, which thanks to polling, enables low latency communications, and zero copy data transfer. In the RDMA paradigm, the network interface handles all protocol processing, and the CPU is practically bypassed therefore fully offloaded. This translates to substantial reduction in CPU utilization, allowing more power efficient and cost effective processors to be used to provide the same performance. Chelsio's adapters implement the RDMA over Ethernet standard (iWARP) in hardware.

The main motivation for the Onload push in networking is providing a user-space I/O capability, without hardware support from the network interface. However, allowing application direct access to hardware brings in complex protection issues and process management problems. By using the purpose built hardware RDMA infrastructure to support user-space I/O, Chelsio provides a cleaner and lighter weight solution. Thus, in the context of low latency applications, both Onload and Offload can provide the processing hooks necessary for achieving desired access times, with very different complexity and security characteristics.

It is noteworthy that most applications of high performance user space I/O in effect trade CPU utilization off for low latency. Applications typically run in a polling loop to detect and process incoming packets as quickly as possible. Therefore, Onload brings in no CPU utilization benefits, rather the opposite. Furthermore, while zero copy is usually touted as an Onload benefit, it only applies to small (unfragmented) UDP datagrams that can be processed individually and independently. TCP zero copy requires pre-processing payload and assembling it such that the byte stream can be placed in the correct order at the correct application buffer location. Therefore, marketing Onload as a generalized approach to move load back onto the CPU as an alternative to offload is promoting a step against the users' interests of system efficiency and associated power consumption.

Microburst Test Results

This section looks at the latest Chelsio and Solarflare low latency adapters and tests their tolerance to traffic burstiness. The adapter microburst test shows the duration of time that an adapter can sustain back-to-back packets at wire rate without any drops. In the context of High Frequency Trading, avoiding packet drops is critical to successful operation.

An adapter capable of sustaining an unlimited back-to-back stream with no packet drop is impervious to microburst issues. The test used here sends packets with a particular message size at line rate (no gaps) to determine the number of packets that are received without packet drops. This data is converted to microburst time with the following formula:

```
Burst Time = (46 (Ethernet/IP/UDP Header Size) + UDP Message Size) * number of packets
without drop * bit time * 8
```

For example:

```
Chelsio 64 Byte UDP Message:   (46+64)*23000*.1ns*8= 2024000ns or 2.0240ms
Solarflare 64 Byte UDP Message: (46+64)*400*.1ns*8=     35200ns or 0.0352ms
```

Tests were run with IXIA traffic generator and a UDP application loopback to show the microburst performance of Chelsio's T520-LL-CR Dual Port 10G adapter and the Solarflare SFN7122F Dual Port 10G adapter.

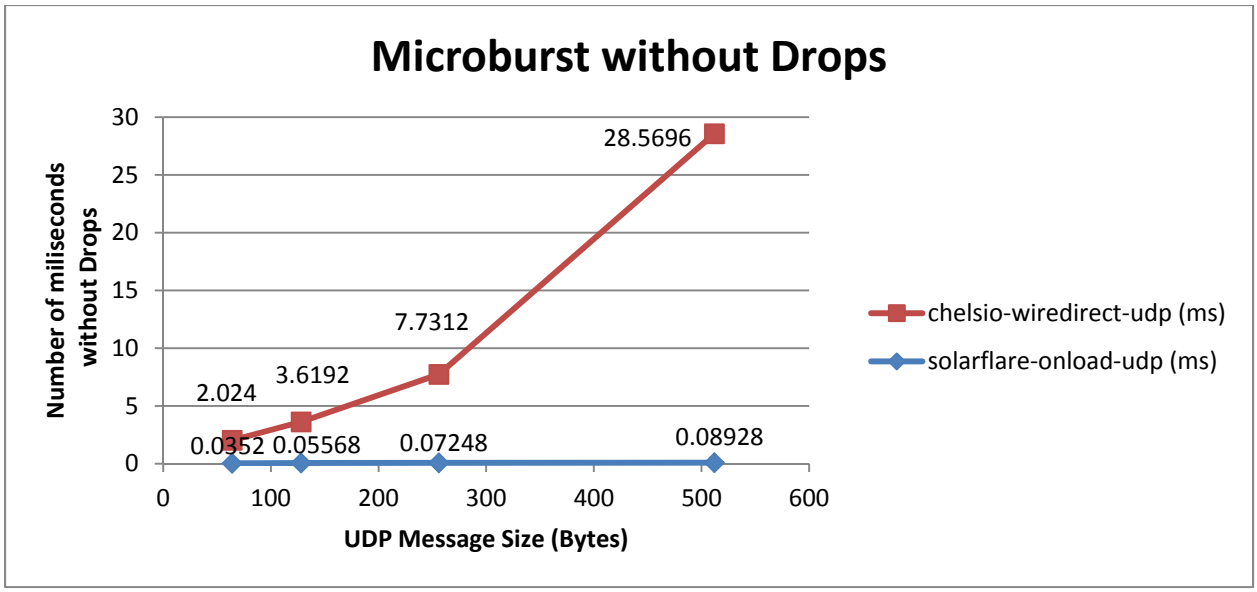
Test commands:

Chelsio
UDP Offload Stack Test Command: CXGB4_SOCK_SPIN_COUNT=2000000 CXGB4_SOCK_POLL_SPIN_COUNT=2000000 CXGB4_SOCK_RQ_DEPTH=19000 CXGB4_SOCK_SQ_DEPTH=32 taskset –c 1 wdload ./udpserver 63

Solarflare
UDP Onload Stack Test Command: taskset –c 1 onload –p latency ./udpserver 63

Setup:

Supermicro X9DR3-F, 2 socket x 8 core (16 core) Intel(R) Xeon(R) CPU E5-2687W 0 @ 3.10GHz connected back to back using RHEL 6.4 64 bit, T5 edc_only config file and modprobe -a t4_tom iw_cxgb4 rdma_ucm.



The graph above contrasts the microburst handling capabilities of the two adapters. It is unquestionably apparent from the data that the Chelsio adapter is far superior at sustaining high input rates, whereas the Solarflare adapter is burst intolerant. This in turn means that the Solarflare latency numbers are only seen when ingress feeds are trickling in, perhaps one packet at a time, which is not the case in practice. In fact, feed activity is notoriously bursty, and rather than getting low latency, users face the scenario of packet drop and the risk of losing critical data at the worst possible moment, with resulting financial shortfall.

Latency vs. Message Rate Test Results

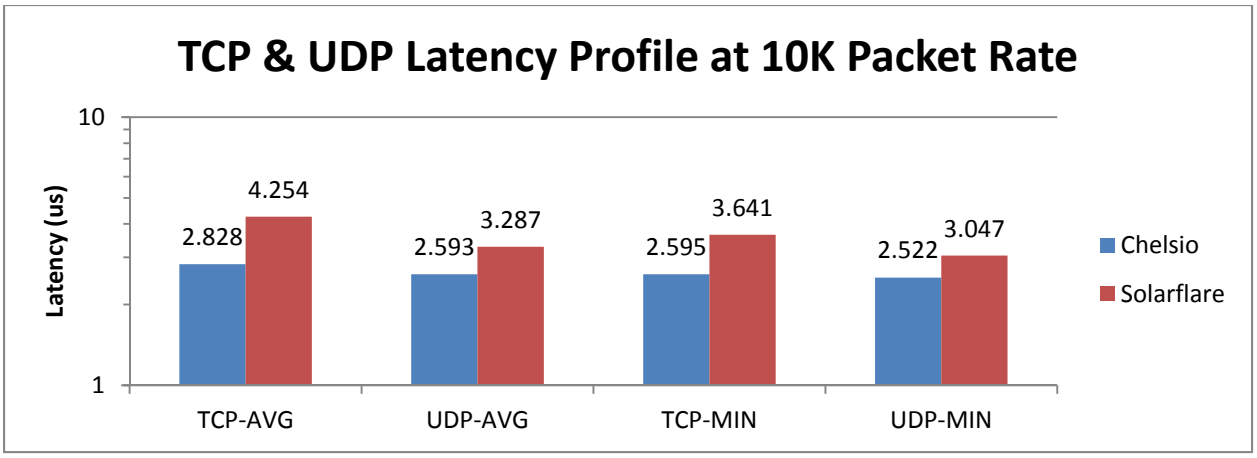
Tests were run with *sockperf* at the 256B message size common in HFT applications, to compare the latency and message rate performance of Chelsio's T520-LL-CR Dual Port 10G adapter and Solarflare's SFN7122F Dual Port 10G adapter.

The systems used were the following:

```
Supermicro X9DR3-F, 2 socket x 8 core (16 core) Intel(R) Xeon(R) CPU E5-2687W 0 @
3.10GHz connected back to back using RHEL 6.4 64 bit, T5 edc_only config file and
modprobe -a t4_tom iw_cxgb4 rdma_ucm
```

UDP & TCP Latency at 10K Packet Rate

The following graphs compares the UDP and TCP minimum and average latencies at 10K packet rate.



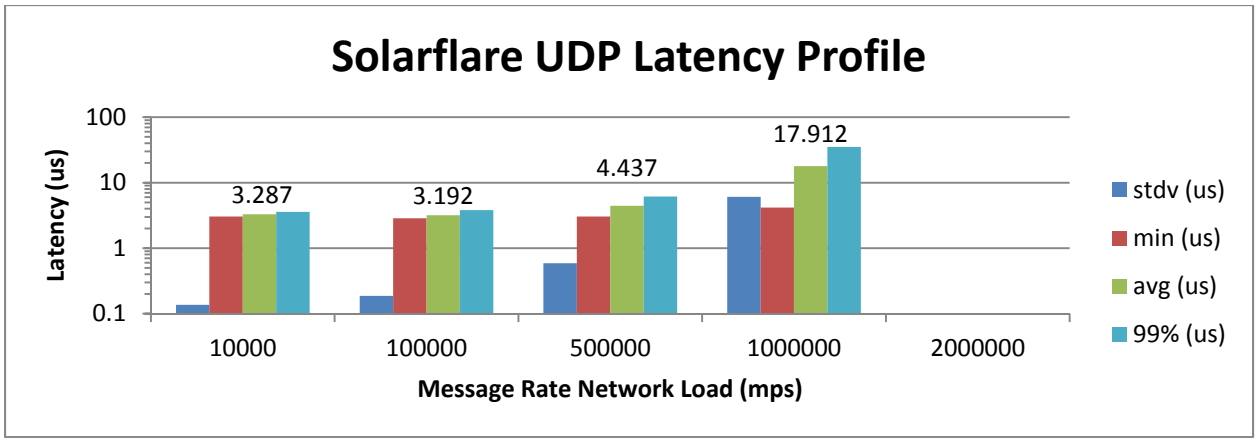
Chelsio's solution is as much as 1.4us lower latency than Solarflare at 10K packet rate. The latency under load test is a more realistic measure for latency in a typical HFT environment than a test with 1 packet in flight.

UDP Latency

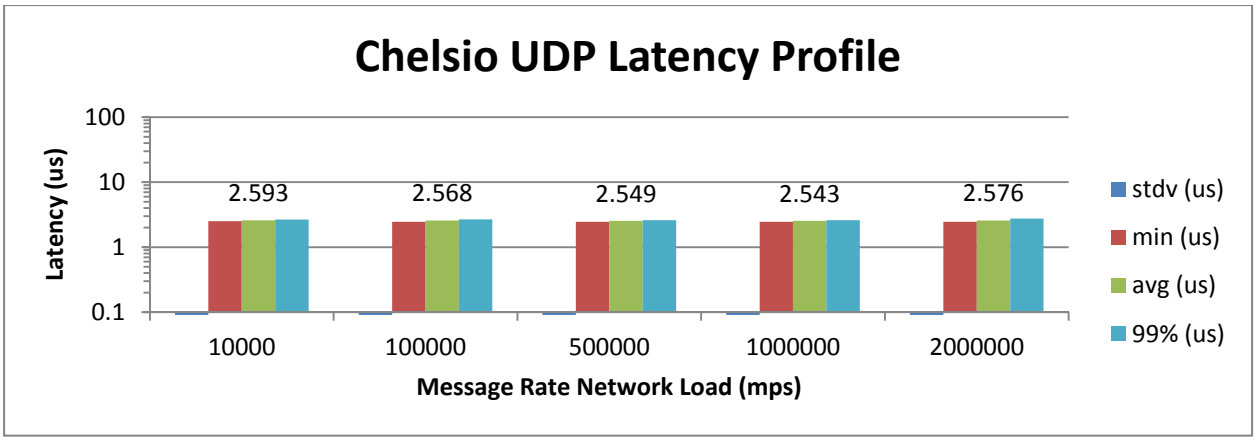
The UDP latency tests used the following command lines:

```
Solarflare
taskset –c 1 onload -p latency sockperf ul -i 4.4.4.1 --msg-size=256 --mps=<rate>

Chelsio
CXGB4_SOCK_SPIN_COUNT=2000000 CXGB4_SOCK_POLL_SPIN_COUNT=2000000 CXGB4_SOCK_RQ_DEPTH=64
CXGB4_SOCK_SQ_DEPTH=8 PROT=UDP tasket –c 1 wdload sockperf ul -i 2.2.2.1 --msg-size=256 --
mps=<rate>
```



The Solarflare adapter is not capable of sustaining network loads at more than 1MPPS without dropping packets, therefore the missing datapoint at 2MPPS. Furthermore, its latency profile deteriorates significantly as the load is increased, along with increased variance, symptomatic of decreased stability as the capacity limit is approached.



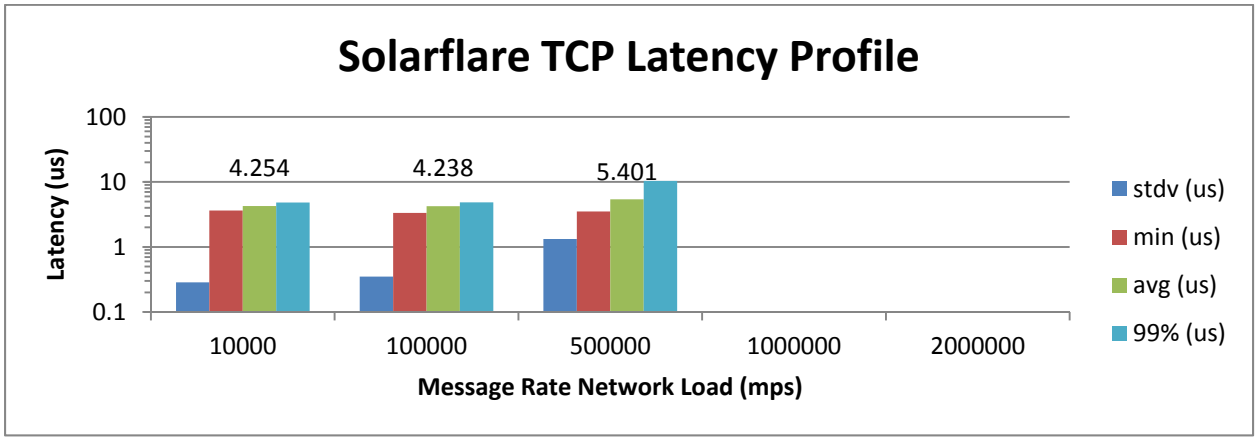
In contrast, Chelsio's solution is capable of sustaining loads in excess of 2MPPS, more than twice the limit of Solarflare. In addition, Chelsio's performance profile shows a remarkably consistent performance level, with minimum, average and 99% data points virtually indistinguishable, showing high tolerance for network load. This is a key difference between the two adapters, and can translate to dramatic difference in returns in actual HFT applications.

TCP Latency

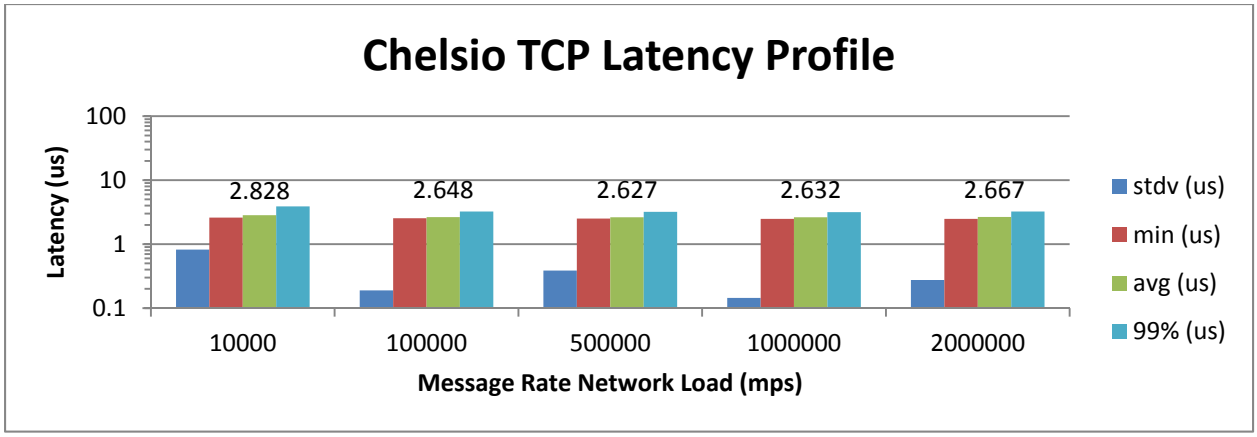
The same tests were run with TCP, using the following commands:

```
Solarflare
taskset -c 1 onload -p latency sockperf ul -i 4.4.4.1 --msg-size=256 --mps=<rate> --tcp

Chelsio
CXGB4_SOCK_SPIN_COUNT=2000000 CXGB4_SOCK_POLL_SPIN_COUNT=2000000 CXGB4_SOCK_RQ_DEPTH=64
CXGB4_SOCK_SQ_DEPTH=8 PROT=TCP taskset -c 1 wdload sockperf ul -i 2.2.2.1 --msg-size=256 --
mps=<rate> --tcp
```



Again, the Solarflare adapter is not capable of sustaining high network loads without drops, therefore the missing datapoints at 1M and 2MPPS. The results are worse than for UDP, with higher latency and worse variance at any particular load point.



The figure above comparing WD-TCP and TCP Onload shows similar conclusions to the ones for UDP. Chelsio's solution offers a latency advantage that is maintained as the load is increased, and demonstrates much higher capacity at high load.

In summary, the results show that Chelsio offers an equivalent or better raw latency, and support the microburst test conclusion that Chelsio's adapter enjoys a significantly more robust performance profile than Solarflare.

Latency Test Results

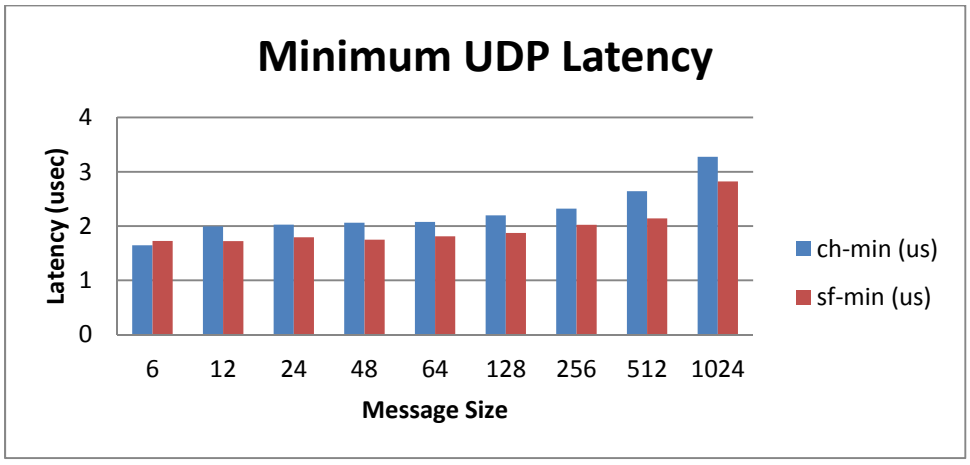
Tests were run with sockperf to compare the TCP and UDP latency performance of Chelsio's T520-LL-CR Dual Port 10G adapter and Solarflare's SFN7122F Dual Port 10G adapter.

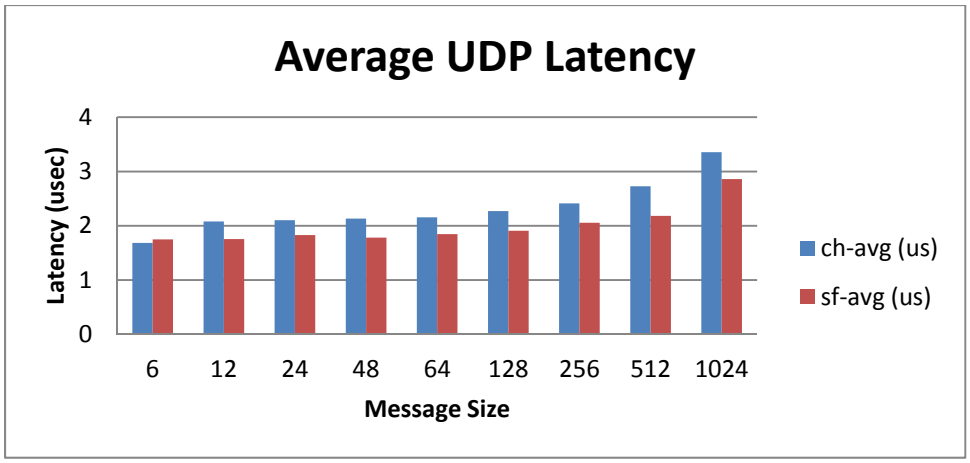
During testing, it became apparent that the Solarflare part exhibits measurable variability and performance impairments under load, possibly due to heat issues. In order to reduce the variability, the results below are averages of 5 runs of every test with each adapter. The system used were the following

Supermicro X9DR3-F, 2 socket x 8 core (16 core) Intel(R) Xeon(R) CPU E5-2687W 0 @ 3.10GHz connected back to back using RHEL 6.4 64 bit, T5 edc_only config file and modprobe -a t4_tom iw_cxgb4 rdma_ucm

UDP Latency

The minimum and average UDP latency for various message sizes are compared below.

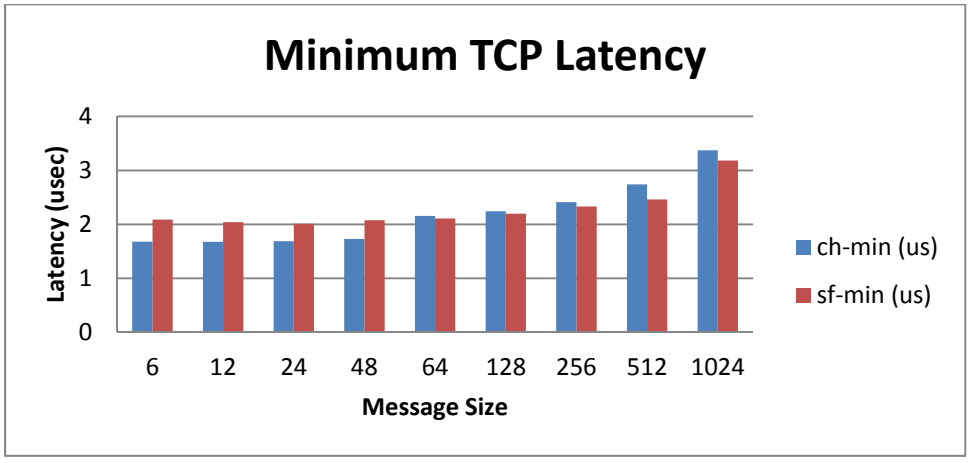


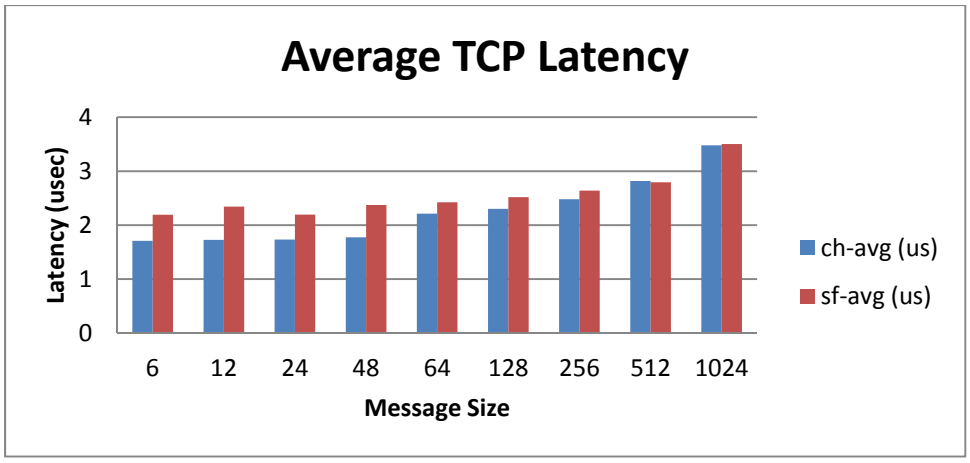


The test results show that Chelsio and Solarflare offer comparable UDP latency results in ping pong tests (single packet in flight). However, as shown in the previous section, when ingress rates are increased, Chelsio's performance is clearly superior, and extends to much higher loads.

TCP Latency

The minimum and average TCP latency for various message sizes are compared below.





Chelsio provides better results than Solarflare in the TCP test. Furthermore, as shown in the previous section, Chelsio's performance is delivered over a much wider band of network conditions.

Conclusions

This paper provided an overview of the technology behind Chelsio's low latency solution, and contrasted it to the latest Solarflare offering. Chelsio's solution is unique in combining hardware and software techniques to provide a robust high performance solution. Representative benchmark results clearly show a superior performance profile for Chelsio that is both more consistent, and more resilient to actual network conditions.

In High Frequency Trading environments, packet drop and latency are critical performance metrics that directly translate into financial results. In selecting an adapter for this market, it is essential to consider the adapter's performance profile beyond limited benchmarking scenarios that may not exhibit the same adversarial conditions as real life traffic scenarios. In a very competitive market, the advantage goes to the participant that can absorb traffic conditions and react in deterministic time. This paper demonstrates that Chelsio's solution is currently the best fit for these requirements. Chelsio also offers high performance 40GbE solutions that provide increased capacity and lower latency at large message sizes.