SOLARFLARE®

WHITE PAPER

TCPDirect

TCPDirect Delivers Lowest Possible Latency between the Application and the Network

TCPDirect is a new user-space, kernel bypass application library implementing TCP and UDP over IP that is now included as part of the Onload™ product. TCPDirect is focused on minimizing network stack overhead and hence offering applications requiring it, the absolutely lowest possible latency between the application and the network via the new Flareon® Ultra 8000-series adapters.

Flexibility vs. Latency

Fundamentally, the latency through any TCP/IP stack, even written to be low-latency, is a function of the number of processor and memory operations that must be performed between the application sending/receiving and the network adapter serving it. Given that the TCP/IP protocol is extremely feature-rich and relatively complex, implementation trade-offs must be made between scalability, feature support and latency. Independently of the stack implementation, going via the kernel imposes system calls, context switches and in most cases, interrupts which have a negative impact on latency.

The solution to this overhead is to take advantage of Solarflare's kernel bypass architecture where the TCP/IP stack resides in the address space of the user-mode application. Onload is an example of a low-latency kernel bypass TCP/IP implementation that is very feature-rich and can support millions of open connections. However, some key use-cases, particularly in the High Frequency Trading sector where minimizing latency is extremely important, do not require a full TCP/IP feature-set and only require a very limited numbers of connections.

Architected to Minimize Latency

TCPDirect is designed from the ground up to minimize the number of processor and memory operations between the user-space application and a Solarflare Flareon Ultra 8000-series adapter. It does this in a number of ways such as by employing much smaller data structures internally that can be virtually guaranteed to be cache-resident as it supports far fewer open sockets than Onload.

TCPDirect provides an optimized zero-copy proprietary API, removing the need to copy packet payloads as part of sending or receiving. Each stack is completely self-contained which removes all processor operations other than those required to get the payload between the application and the adapter. In contrast, Onload provides transparent communication between kernel and accelerated sockets.

TCPDirect's user-space C API, is designed to provide conceptually similar semantics to a subset of the BSD Sockets API in a more efficient fashion. Some key concepts are:

- TCPDirect supports multithreading
- Applications can parallelize and partition network access via multiple TCPDirect stack instances, typically one per thread, which are created explicitly
- TCP and UDP sockets have their own set of API calls
- Both passive (listening) and active (connecting) TCP sockets are supported
- Sockets may be addressed individually or via a multiplexer (similar to epoll)
- All data-path calls are non-blocking to give the application total control
- A transmit mode is provided allowing a number of different pre-populated possible segments/datagrams to be submitted to the adapter and one of them triggered by the application and pushed out of the adapter directly rather than pushed from the host.

As TCPDirect is a zero-copy API, the application has direct access to the memory containing the actual packet buffers in which packet payloads may be written to/read from.

Ease of Implementation

To obtain the latency benefits of TCPDirect, an application is modified or written natively to take advantage of the TCPDirect API defined in its header files. The TCPDirect library may then either be statically linked into the application or loaded dynamically when the application runs.

Solarflare has provided a number of example applications with source code and make files demonstrating the key features of opening/closing sockets and sending and receiving data via the TCPDirect API.

Troubleshooting is facilitated by a TCPDirect-specific version of Onload's stackdump tool which provides non-intrusive real-time access to a host of stack state and statistics. Architecturally TCPDirect keeps state information in shared memory which stackdump can read independently when required imposing negligible impact on the TCPDirect running code.

Summary

TCPDirect is delivered as part of Onload and provides a subset of Onload's extremely rich feature set targeted at obtaining the lowest possible latency between the application and network on the Flareon Ultra 8000-series adapters. The TCPDirect stack component of a Half-Round-Trip mean latency can be as low as 15ns (UDP) and 22ns (TCP).

TCPDirect provides a proprietary zero-copy API conceptually similar to BSD sockets as part of the rigorous focus on minimizing application to network latency.

TCPDirect is not a perfect fit for all low-latency use-cases as it necessarily requires certain compromises in its focus of latency. For cases where other features are required, Onload provides the ideal solution.

Latency Comparisons

The following table shows each stack's Half-Round-Trip (HRT) latency over and above the adapter hardware latency.

Kernel	1500 ns
Onload	300 ns
TCPDirect	20 ns

US +1 949 581 6830 x0
UK +44 (0) 1223 477171
HK +852 2624 8868
sales@solarflare.com
www.solarflare.com

SF-117079-AN Issue 3
TCPDirect WP 121016