SOLARFLARE®

Application Nanosecond TCP Send (ANTS): From request to response in less than 250ns

SolarflareWhitePaper

Introduction

Certain classes of network application fall into the category of high-performance request-response applications. These applications require either very low latency (<< 1μs) between receipt of request and transmission of response, the ability to handle a very large number of such transactions (>>100ktps), or both.

FPGAs can accelerate such applications by speeding up the algorithmic processing and eliminating the cost of moving messages between the network interface and host processor. However, if the response has to be provided as part of a TCP stream, handling the TCP stack on the FPGA becomes a significant overhead in FPGA resource and adds to test and verification workload.

Solarflare's patented ANTS feature combines the ApplicationOnload™ Engine (AOE) product with the OpenOnload® (or EnterpriseOnload®) Delegated Send feature. The greater part of the complexity of TCP processing may now be handled in a non-time critical way in software on the host CPU, while requiring the FPGA to implement little more than a simple filter and packet buffer in RTL. The complete system remains fully RFC compliant.

This white paper describes one possible system setup which leverages this feature and explains its implementation.

OpenOnload Delegated Send

OpenOnload-201502 introduced the Delegated Send API, which enables an application to use OpenOnload's standard sockets API and TCP/IP stack for socket creation, management, demultiplexing and all usual data send/receive operations, but also to benefit from sending directly and consequently being able to achieve the absolute best latency on the critical path. When using the delegated send API, sockets are created and managed through the POSIX API as normal. When an application knows it will need to make a critical path TCP send, it uses the delegated send API in OpenOnload. OpenOnload provides the application with the current TCP/IP headers and details of how much data is authorized to be sent. Leveraging AOE's FPGA, the application is then free to make its send(s) at the appropriate time using the FPGA. Having sent one or more TCP segments, the application gives OpenOnload a copy of the message(s) sent. This allows OpenOnload to update protocol state for the socket, and save the messages in case they need to be retransmitted. TCP packets that are received are delivered directly to OpenOnload and processed as normal.

Request-Response Application

The application described below is required to act on incoming data which forms part of a TCP or UDP stream, process it according to rules and/or accumulated state, and provide a response over a TCP connection to the requester. Such an application could be part of a control loop

sales@solarflare.com

US 1.949.581.6830 x2930

UK +44 (0)1223 477171

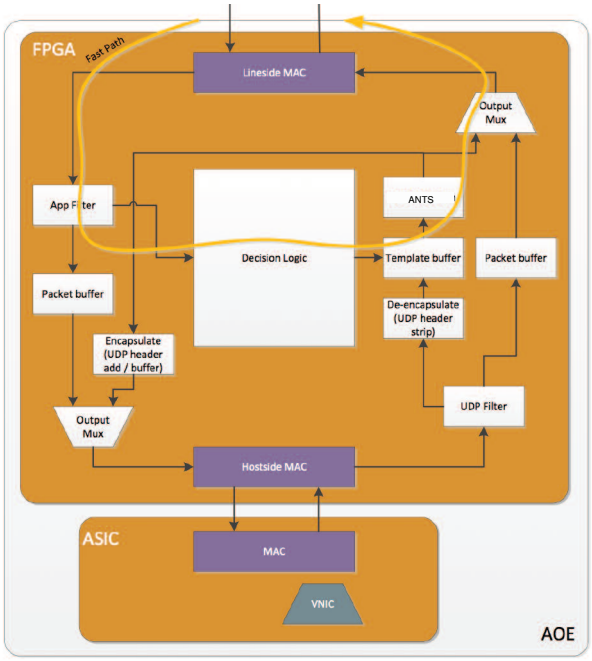
HK +852 2624-8868

www.solarflare.com

needing a very fast response – a financial trading application, queries to a hardware-searched memory database, optical actuators in a multi-mirror telescope or other applications. This paper will describe a "tick-to-trade" application in more detail.

FPGA Setup

In addition to the decision logic (i.e. the heart of the application) in RTL, a few simple RTL blocks are required to complete a tick-to-trade application. These include packet buffers (FIFOs), multiplexers, packet filters and the ANTS block. The rest of the support is contained in a software component running on a host PC or server.



Operation

At the start of day, the Controlling Software Application (CSA) on the host initializes itself, and establishes a control plane with the FPGA. A TCP connection is established with the exchange using standard socket calls (the FPGA just passes these packets through). The CSA then calls o_d_s_prepare()[1]. The OpenOnload stack generates network headers and passes them back to the CSA, which forwards them on to the ANTS block on the FPGA.

The CSA may also prepare messages (or message templates) that will form the TCP responses, and store these in a buffer on the FPGA. The CSA also pre-computes the Internet checksum for these messages. Alternatively the FPGA application may generate response messages itself.

Fast Path

Trading information is received by the FPGA directly after a relevant packet emerges from the FPGA's line-side MAC. It uses an appropriate filter (TCP/UDP port, multicast group etc.) to identify packets of interest, parses the headers and payload and then creates a trading decision. Other traffic, including TCP ACKs, passes through to the host.

Once the algorithm has decided to trade, a pre-formed response template is selected and final trade data is inserted. The pre-computed checksum is updated as these modifications are made, allowing cut-through operation which saves further latency. The response message and checksum are passed to the ANTS block, which prepends the network headers and launches the newly formed TCP segment to the line-side MAC. The message is simultaneously encapsulated in UDP and passed back to the CSA.

[1] o_d_s_ is an abbreviation for onload_delegated_send_

SOLARFLARE®

sales@solarflare.com

US 1.949.581.6830 x2930

UK +44 (0)1223 477171

HK +852 2624-8868

www.solarflare.com

SolarflareWhitePaper

Cleanup

The CSA removes the encapsulation and passes the message payload to OpenOnload using o_d_s_complete(). OpenOnload stores these messages in the socket's retransmit queue so that it can retransmit them if any are lost in the network. Otherwise they are freed when acknowledged by the receiver.

The FPGA application can send further TCP messages up to the limit of the send and congestion window, so the CSA calls o_d_s_prepare() again as needed to get up-to-date values and passes these to the FPGA. The same mechanism is also used to update the network headers when they change.

Latency

Request-response latency is largely determined by the wire-to-wire time for the RX/TX MAC, PCS and PMA, and by the application itself. The latency of the ANTS critical path itself is only three clock cycles (<= 15ns). Realistic state-of-the-art wire-to-wire latency for MACs in Stratix V FPGAs are now less than 100ns, depending on vendor. The application latency is application dependent, but some applications might realistically produce a response in under 150ns.

FPGA Resources

TCP cores for FPGA typically consume large amounts of logic and memory resources, and often require off-chip RAM to buffer packets. The ANTS logic consumes only 1.3% of the AOE's FPGA, does not use off-chip RAM and requires less than 100 bytes of state per TCP connection. This maximizes the FPGA resources that the application can leverage and allows ANTS to scale to thousands of TCP connections.

Summary

- ANTS provides the ability to execute a TCP send from the AOE FPGA with virtually no latency overhead.
- Leveraging ANTS to perform all of a request/response application in an FPGA has the potential to outperform any other method of implementing that process.
- RTL development is time-consuming and costly. If TCP is involved, the time and FPGA development costs can increase significantly. Using Delegated Send helps keep development focused solely on the application, reducing time-to-market, hardware and development costs.
- Robust, scalable, standards-compliant solution.

About Solarflare

Solarflare is the leading provider of application-intelligent networking I/O software and hardware that accelerate, monitor and secure network data. With over 1,400 customers worldwide, the company's solutions are widely used in scale-out server environments such as electronic trading, high performance computing, content delivery, cloud, virtualization and big data. Solarflare's products are available from leading distributors and value-added resellers, as well as from Dell, HP, IBM and Lenovo. Solarflare is headquartered in Irvine, California, and operates R&D facilities in Cambridge, UK, San Diego, USA and New Delhi, India.

SF-114903-CD Issue 3
ANTS_WP_121415

SOLARFLARE®

sales@solarflare.com

US 1.949.581.6830 x2930

UK +44 (0)1223 477171

HK +852 2624-8868

www.solarflare.com