Ultra-Low Latency (ULL) FPGA Framework

Benchmark Report

This document describes ULL FPGA Framework performance tests and results from Orthogone Technologies Inc.

Product Status: General Availability

Document version: 1.0.0
February 5, 2025

Intellectual property notice

This document is the property of Orthogone Technologies Inc. The data contained herein, in whole or in part, may not be duplicated, used, or disclosed outside the recipient for any purpose other than to conduct business and technical evaluation. This restriction does not limit the recipient’s right to use information contained in the data if it is obtained from another source without restriction.

Revision history

This document has been updated as follows:

Document Revision	Date	Comments
1.0	2025-02-05	Initial release

Table of contents

List of figures

List of tables

1. Introduction

The ULL FPGA Framework is a high-performance FPGA and software solution specifically designed for ultra-low latency (ULL) networking applications and primarily for high-performance financial applications.

The ULL FPGA Framework is an ultra-low-latency network interface and PCIe DMA (Circular Buffer Direct Memory Access - CBDMA) controller. It transfers data between the FPGA and a host through PCIe. Queues are used to channelize traffic. The networking data path protocol (TCP/UDP) is optimized in FPGA logic. Pure (FPGA) and Hybrid (FPGA+SW) acceleration are supported.

- Pure FPGA solution: Ultra-low latency, network, and On-chip application acceleration.
- Hybrid solution: Ultra-low latency, network, and On-chip application acceleration and host application acceleration.

Pure FPGA applications are used when the FPGA is the only source of packet generation and processing. Hybrid solutions applications are used when both the FPGA and the Host application can be the source of traffic generation and processing. In hybrid solutions, the software application handles non-latency-critical tasks, while the FPGA manages latency-critical operations to maximize ultra-low latency benefits.

This document presents the latency performance tests for Hybrid applications executed on two overclocked Blackcore Technologies servers (ICON 3100- RL+ and G3 SPR-M) and a standard server (Dell/EMC PowerEdge R750). As illustrated in the figure below, three tests have been performed:

- Test 1: 10G TCP/IP Full loopback (RTT latency)
- Test 2: 10G UDP/IP Full loopback (RTT latency)
- Test 3: PCIe DMA latency (RTT latency)

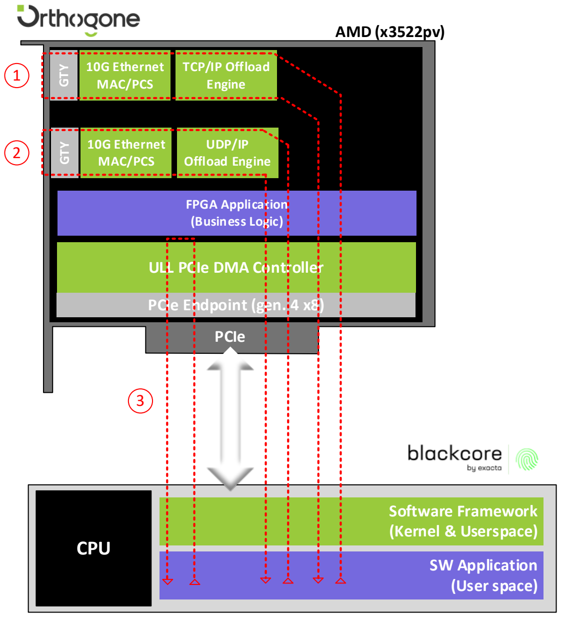


Figure 1 Test setup with local port loopback

The test application *oti-perf* is designed to allow for easier testing and better control of the computed and reported latency. The *oti-perf* tools can test for latency using this single server configuration. Special care must be taken on CPU core selection when running the *oti-perf* application with this test setup configuration. The client and server must not execute on the same CPU core because of performance options on the thread execution. For Test 1 and Test 2, loopback plugs are inserted in the DSFP ports to create physical loopbacks with minimal latency (< 1ns). A zero-latency loopback path is activated inside the FPGA to perform Test 3.

2. Test setup and tuning recommendations

The hardware configurations for the servers are summarized in the table below:

	Blackcore Technologies		**Dell/EMC**
	ICON 3100-RL+	G3 SPR-M	PowerEdge R750
CPU	Intel® Core™ i9-14900KS	Intel® Xeon® w7-2495X	Intel® Xeon® Gold 5315Y
CPU Frequency	Up to 8 cores @ 5.9GHz	Up to 24 cores @ 4.8GHz	Turbo Boost activated, 3.5 GHz
RAM	DDR5 64GB DIMM Synchronous Registered (Buffered)	DDR5 128GB DIMM Synchronous Registered (Buffered)	DDR4-2933 4x 8GB RDIMM, 3200MT/s, Single Rank
PCI Express (PCIe)	PCIe Gen 4 x8		
FPGA Card	AMD-Xilinx x3522pv		

Table 1 Server hardware configurations

The installed OS is:

- Kernel version: 5.14.0-522.el9.x86
- CentOS Stream 9 Linux release (Core)

ULL Component	**Version**
FPGA ULL-MAC/PCS core	1.7
FPGA ULL PCIe DMA Core	4.0
FPGA ULL-TOP core (TCP/UDP)	3.1
SW ULL driver (oti-ull)	1.1.0
SW PCIe DMA driver (oti_cbdma)	3.0.0

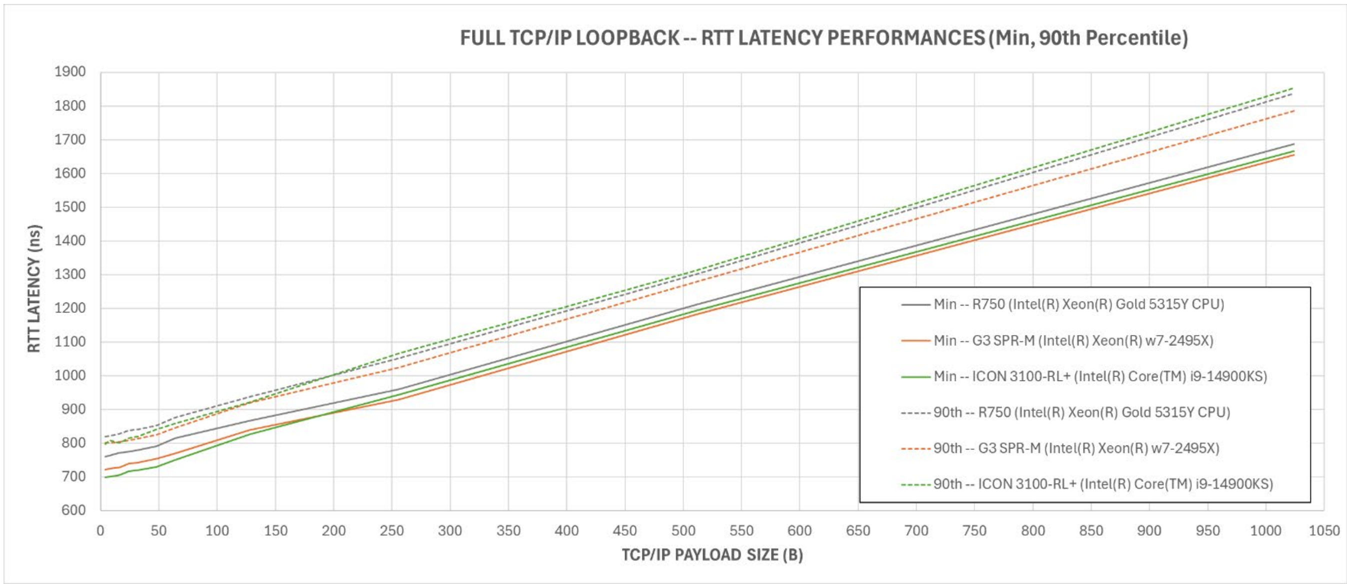
Table 2 FPGA/SW Version description

The performance tuning recommendations are provided in the document *OTI ULL Software Framework User Guide.*

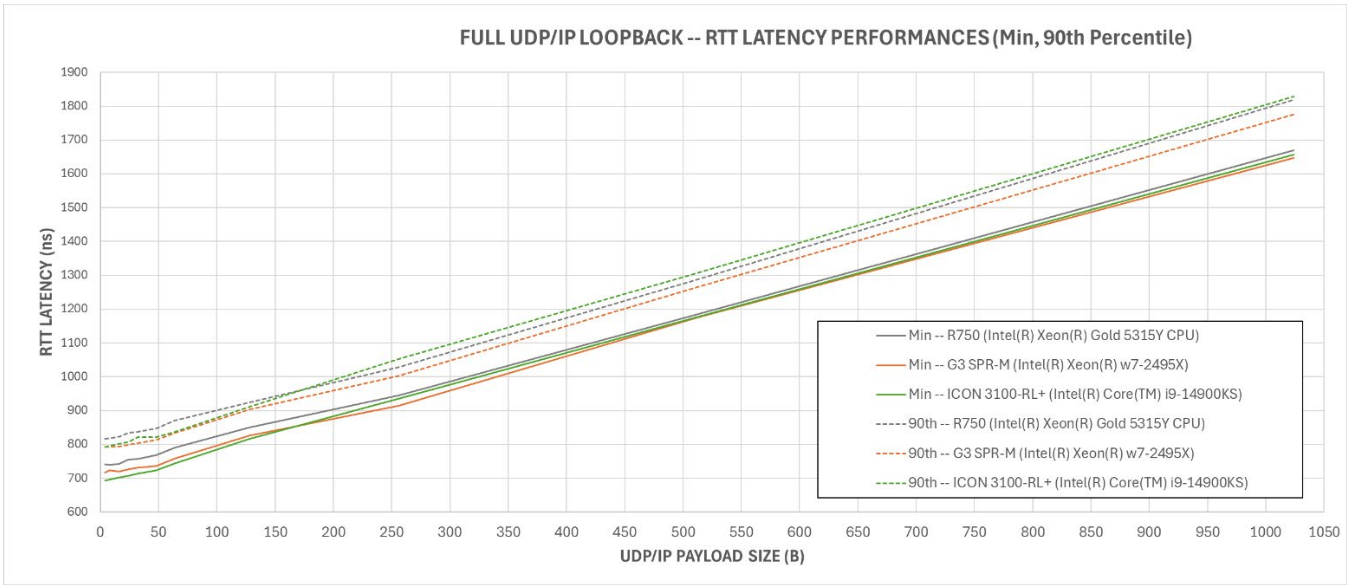
3. Test results

Test 1: 10G TCP/IP Full loopback (RTT latency)

The following results were obtained using Cut-Through (CT) data flow for TCP/IP and UDP/IP. RTT latency results are presented as a function of TCP and UDP payload sizes across different percentiles (minimum and 90th percentile). A minimum latency of under 700ns and a 90th percentile latency of under 800ns can be achieved with small payloads on the ICON 3100-RL+ server. The G3 SPR-M server delivers similar performance while offering additional processing capabilities due to its higher core count.

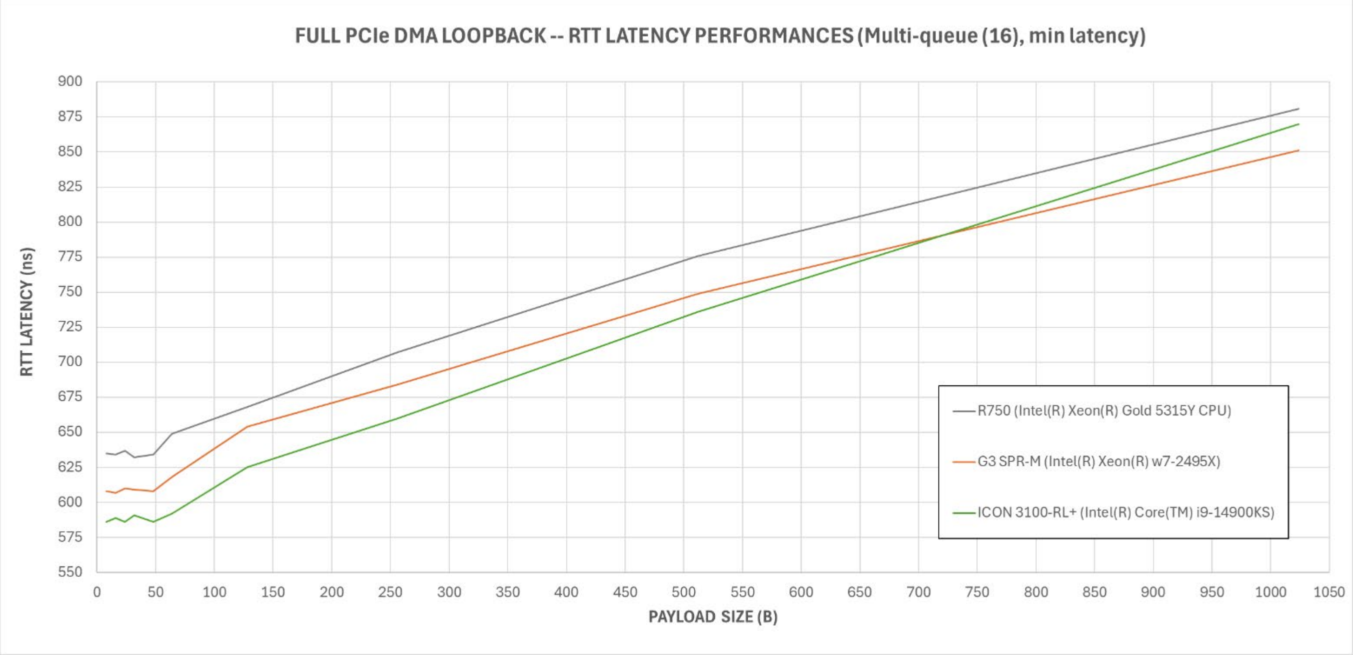


Test 2: 10G UDP/IP Full loopback (RTT latency)

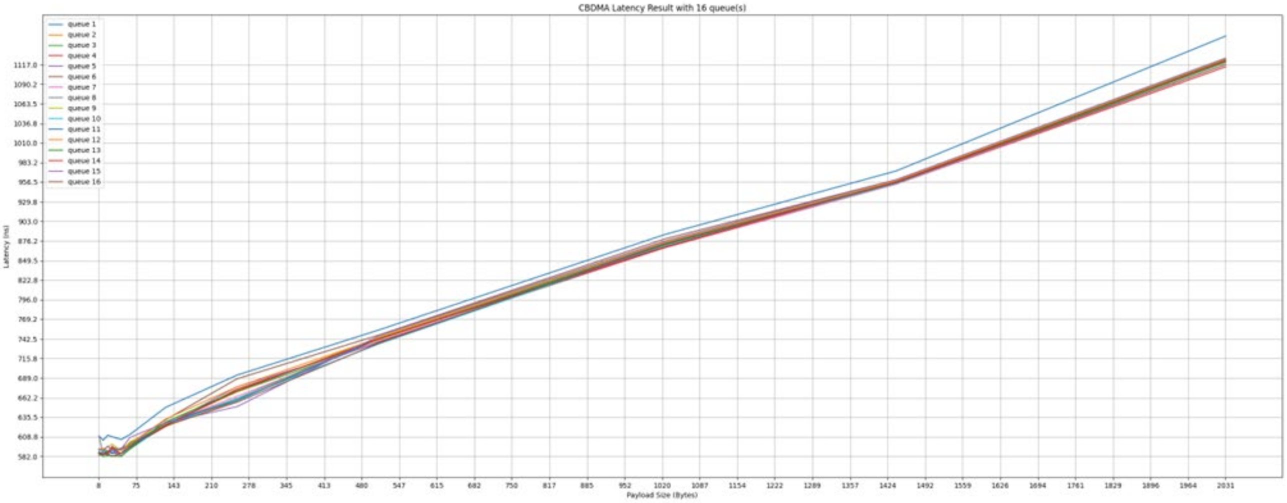


Test 3: PCIe DMA latency (RTT latency)

To better assess the latency benefits of overclocked servers, we focus on measuring only the latency of the PCIe DMA controller and its associated software framework. In this mode, test packets are generated by host software and looped back by the FPGA application at the DMA controller's AXI-4 interface. The software tool creates 16 queues to transfer data in parallel. Although all queues exhibit similar performance, we present the lowest latency results for various payload sizes across our three server configurations.



Notably, the Blackcore ICON 3100-RL+ server achieves an RTT latency as low as 585ns for small payloads, compared to 635ns for the standard Dell/EMC R750 server. A latency gain of approximately 50ns remains consistent for payload sizes up to around 512B. The following figure shows the minimum latency for all 16 queues tested on the Blackcore ICON 3100-RL+ server. As mentioned earlier, the latencies are similar across all queues. This is illustrated in the figure below.



4. Conclusion

In February 2025, Orthogone Technologies and Blackcore Technologies announced a strategic partnership aimed at setting new benchmarks in ultra-low latency network acceleration. By combining Orthogone's ULL FPGA Framework with Blackcore's overclocked high-performance servers, sub-700ns full TCP/IP loopback latency performance is now achievable. As detailed in this report, the performance of the joint solution is self-evident in terms of latency and consistency.

- 699ns Full loopback, minimum RTT latency for 64B frame (6B TCP/IP payload)
- 798ns Full loopback, 90% percentile RTT latency for 64B frame (6B TCP/IP payload)

Finally, latency improvements are still possible with readily available FPGA solutions such as the AMD-Xilinx UL3524 and UL3422, which can improve end-to-end latency by approximately 25ns for all payload sizes. Likewise, latency optimizations are still possible with the use of PCIe Gen 5 and beyond in future hardware platforms.

Additional product information can be found at the following URL links:

- Orthogone Technologies ULL FPGA Framework: https://orthogone.com/ull-fpga-framework/
- Blackcore Technologies overclocked servers: https://blackcoretech.com/