November 2024

LightGBM Inference

Benchmark Report

Blackcore ICON 3100-RL+ Server
AMD Alveo U50 Accelerator Card

Contents

References

[1] Xelera Technologies, "Xelera Silva," [Online]. Available: https://www.xelera.io/products/silva. [Accessed 10 4 2024].

[2] Blackcore Technologies, "ICON 3100-RL+," [Online]. Available: https://blackcoretech.com/firefly/file/get?id=NJHYtiGBecqcAtCCKj5wkw. [Accessed 10 04 2024].

1 Xelera Silva

Gradient Boosting frameworks such as XGBoost and LightGBM are widely used in financial trading systems, ransomware and DDOS detection systems, and recommender systems. Xelera Silva provides best-in-class latency and throughput for XGBoost and LightGBM and Random Forest inference by leveraging commercial off-the-shelf data-center grade FPGA accelerators.

The Xelera Silva software (Xelera Technologies) loads machine learning models from XGBoost, LightGBM, ONNX ML Tools. The models are executed for inference on AMD Alveo platforms. The user application interacts with the accelerator software via a C/C++, C# or Python API. This document describes the latency benchmark tests on a Blackcore Technologies ICON 3100-RL+ server (Blackcore Technologies). The tests are performed with the Xelera Silva 7.4.0 release:

- Test 1: single model inference
- Test 2: simultaneous and asynchronous inference with 4 models

2 Test 1: Single Model Inference

This benchmark validates the LightGBM model inference latency on the system specified in Table 1 below.

Table 1: System-under-test

Server	Blackcore ICON 3100-RL+ (Blackcore Technologies) CPU: Intel Core i9 processor 14900KS CPU Frequency: Up to 8 P-Cores @ 5.9GHz (SSE); up to 8 E-Cores @ 4.3GHz (SSE) CPU Cache: 36MB @ 5.0GHz Chipset: Z790 Memory: 128GB DDR5 Overclocked UDIMM
OS	Linux Rocky 9.4
PCIe interface	Gen3 x16
AMD Alveo Card	U50 with Xelera PCIe ULL shell
Driver	Xelera PCIe ULL 2.13.0
ML Inference Software	Xelera Silva 7.4.0

The Xelera Silva ML Inference software was compared against several other software frameworks for the acceleration of gradient boosting machine learning models. The compared software frameworks are listed in Table 2 below.

Table 2: Compared software frameworks

ML Inference Software	Version	Description
Intel oneDAL	2024.5.0	Intel CPU-optimized ML inference software
TL2cgen vanilla	TL2cgen 1.0.0 Treelite 4.3.0	TL2cgen framework with Treelite serialization without optimizations
TL2cgen ann	TL2cgen 1.0.0 Treelite 4.3.0	TL2cgen framework with Treelite serialization with a pre-trained branch predictor (branch annotation)
TL2cgen quant	TL2cgen 1.0.0 Treelite 4.3.0	TL2cgen framework with Treelite serialization with quantized (integer) threshold comparisons
TL2cgen quant/ann	TL2cgen 1.0.0 Treelite 4.3.0	TL2cgen framework with Treelite serialization with combined quantized (integer) threshold comparisons and branch annotation
Xelera Silva	7.4.0	FPGA-accelerated ML inference software

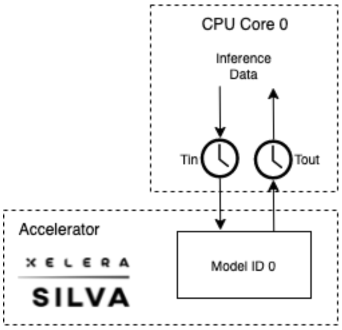
Xelera Silva is the only FPGA-accelerated ML inference software in this comparison. The other frameworks use only CPU optimizations to accelerate the inference of gradient boosting models, such as the use of vector extension instructions, branch prediction and integer comparisons.

2.1 Test Description

The roundtrip latency at the API interface (Tout – Tin) is measured when running the inference for the model configuration in Table 3.

Table 3: Model configuration

Model Type	LightGBM regression
Dataset	Synthetic Random
Number of Features	128
Number of Trees	1000
Number of Levels	8
Batch Size	1
Numerical Features	Yes
Categorical Features	No



For each software framework configuration, the test involves running inference on one model. Each process is pinned to a CPU core 0. The test is conducted 1,000,000 times.

2.2 Results

For each software framework configuration, the test involves running inference on one model. Figure 1 shows the latency statistics of Xelera Silva in comparison to the third-party software frameworks. The graphs show the fraction of inference measurements (y-axis) below a specified latency (x-axis).

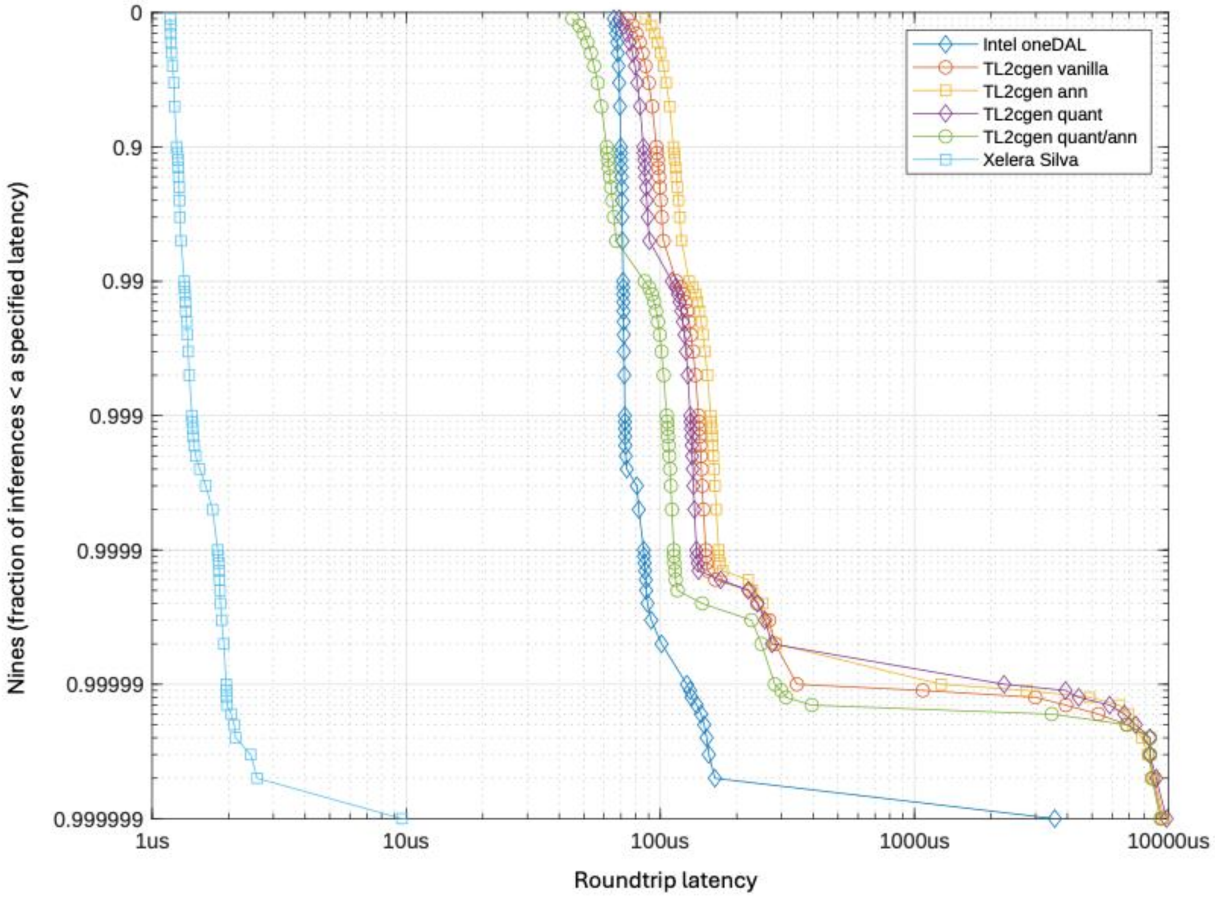


Figure 1: Latency comparison for single-model inference

Table 4 compares the median latency (50th percentile) and the 99th percentile latency of the graphs above.

Table 4: Latency statistics

ML Inference Software	**Median latency (50th percentile)**	**99th percentile latency**
Intel oneDAL	68.02 us	71.36 us
TL2cgen vanilla	85.05 us	115.60 us
TL2cgen ann	100.02 us	129.80 us
TL2cgen quant	77.72 us	110.95 us
TL2cgen quant/ann	53.30 us	86.81 us
Xelera Silva	1.19 us	1.34 us

The following figure show the latency characteristic of Xelera Silva in detail.

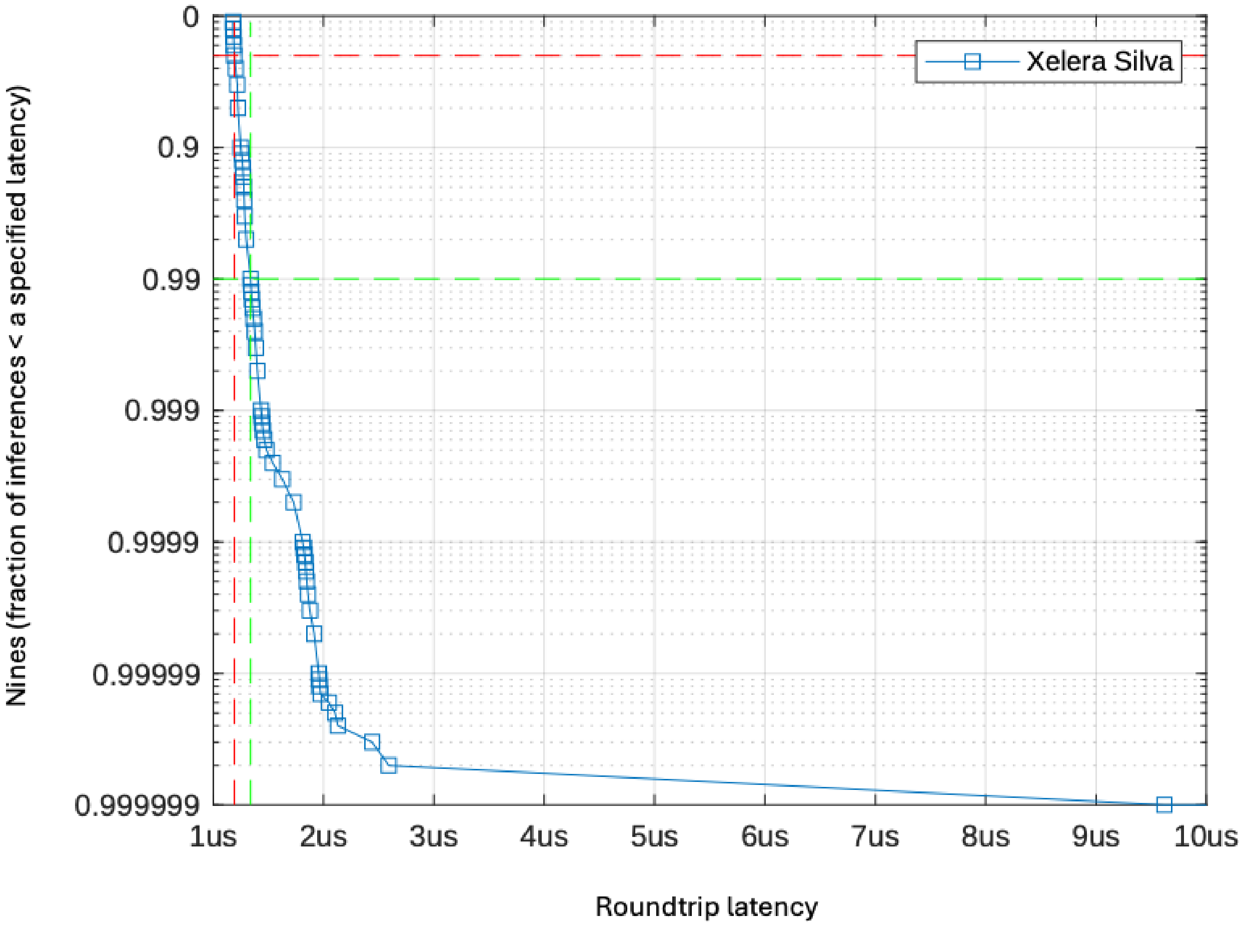


Figure 2: Latency statistic of Xelera Silva in single-model execution

2.3 Key Findings

The single-model benchmark demonstrated that Xelera Silva running on the AMD Alveo U50 FPGA achieved a **median latency of 1.19 microseconds** and a **99th percentile latency of 1.34 microseconds**. By comparison, CPU-based inference frameworks such as Intel oneDAL and Treelite/TL2cgen recorded median latencies in the range of **53 to 100 microseconds**, with 99th percentile latencies extending up to **130 microseconds**. These results show that Xelera Silva consistently delivers **over forty times lower latency** than leading CPU-optimized implementations, setting a new benchmark for low-latency inference in gradient boosting models.

3 Test 2: Simultaneous and Asynchronous Inference with 4 Models

This benchmark validates the LightGBM model inference latency on the system specified in the Table 1 below. In this test, 4 models are executed simultaneously on the FPGA accelerator. Each model is accessed by the host software via an individual process.

Table 5: System-under-test

Server	Blackcore ICON 3100-RL+ (Blackcore Technologies) CPU: Intel Core i9 processor 14900KS CPU Frequency: Up to 8 P-Cores @ 5.9GHz (SSE); up to 8 E-Cores @ 4.3GHz (SSE) CPU Cache: 36MB @ 5.0GHz Chipset: Z790 Memory: 128GB DDR5 Overclocked UDIMM
OS	Linux Rocky 9.4
PCIe interface	Gen3 x16
AMD Alveo Card	U50 with Xelera PCIe ULL shell
Driver	Xelera PCIe ULL 2.13.0
ML Inference Software	Xelera Silva 7.4.0

3.1 Test Description

The roundtrip latency at the API interface ($Tout_x$ – Tin_x) is measured when running the inference for the model configuration in Table 6.

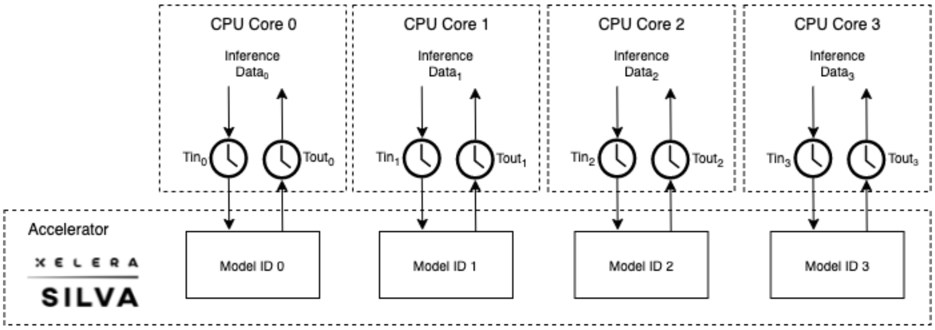


Table 6: Model configuration

Model Type	LightGBM regression
Dataset	Synthetic Random
Number of Features	128
Number of Trees	1000
Number of Levels	8
Batch Size	1
Numerical Features	Yes
Categorical Features	No

For each model configuration, the test involves running inference with four models (IDs from 0 to 3) simultaneously in the **asynchronous** mode (independent processes accessing the models). Each process is pinned to a CPU core (0 to 3). The test is conducted 1,000,000 times.

3.2 Results

The Figure below shows the latency statistics of Xelera Silva when running inference with 4 models at the same time. The graphs show the fraction of inference measurements (y-axis) below a specified latency (x-axis) for each of the 4 concurrent model inferences.

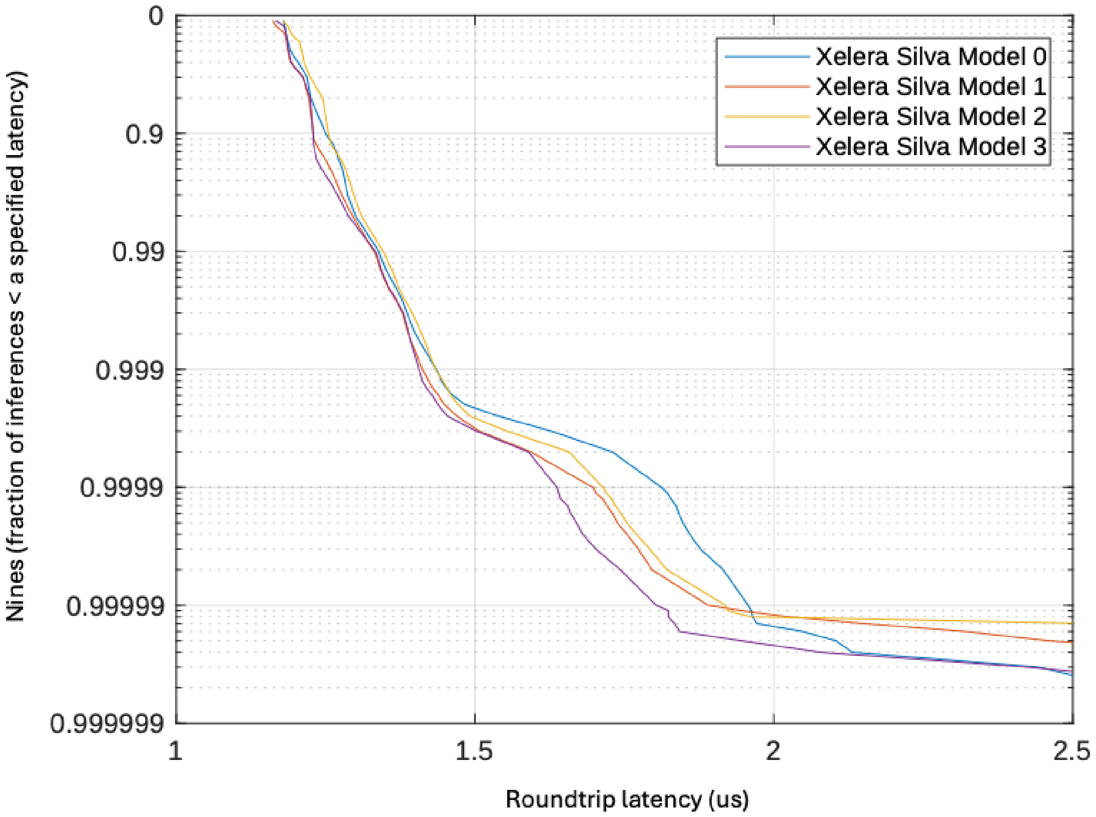


Figure 3: Latency statistic of Xelera Silva in multi-model execution

Table 7 compares the median latency (50th percentile) and the 99th percentile latency of the graphs above.

Table 7: Latency statistics

Model ID	Median latency (50th percentile)	99th percentile latency
0	1.191 us	1.339 us
1	1.187 us	1.331 us
2	1.210 us	1.346 us
3	1.188 us	1.333 us

3.3 Key Findings

The multi-model test, where four LightGBM models were executed concurrently in asynchronous mode, Xelera Silva maintained **stable and consistent performance** across all models. Each model exhibited median latencies around **1.19 microseconds** and 99th percentile latencies near **1.34 microseconds**, effectively identical to the single-model results. Even under one million iterations per model, the latency distribution remained tightly bounded with no long-tail deviations, confirming the **deterministic and scalable performance** of FPGA acceleration for concurrent inference workloads.

4 Summary and Conclusions

This benchmarking study evaluated the inference latency of Xelera Silva for LightGBM models on a Blackcore ICON 3100-RL+ server equipped with an AMD Alveo U50 FPGA.

Two test scenarios were considered: single-model inference and simultaneous multi-model inference with four models running asynchronously. In both scenarios, Xelera Silva consistently outperformed state-of-the-art CPU-based frameworks.

The single-model test demonstrated **sub-microsecond inference latency**, more than 40x faster than optimized CPU software libraries.

The multi-model test confirmed that this performance is sustained even when multiple models are executed concurrently, with no observable degradation in latency or stability.

A key differentiator is the **deterministic nature of FPGA acceleration**, as evidenced by the tight latency distribution across one million iterations. Unlike CPU-based inference, which typically exhibits wider and less predictable latency ranges, Xelera Silva in conjunction with the Blackcore ICON 3100-RL+ server delivers stable and repeatable results ideally suited for latency-sensitive environments.

Overall, the results validate Xelera Silva as a high-performance, low-latency, and scalable inference solution. Its ability to maintain **microsecond-level response times** under both single and concurrent workloads makes it particularly valuable for real-time applications such as financial trading, network security, and recommendation systems, where every microsecond matters.