XELERA

Ultra-low Latency XGBoost / LightGBM Inference

Xelera Silva

XELERA SILVA

- Fastest XGBoost and LightGBM inference
- 100x better latency than CPUs and GPUs
- Ultra-low, guaranteed and controllable latency
- Usable through a C/C++ or Python API

INTRODUCTION

Gradient Boosting frameworks such as XGBoost, LightGBM and CatBoost, as well as Random Forest algorithms are widely used in ransomware and DDOS detection systems, recommender systems and trading systems. **Xelera Silva** provides best-in-class throughput and latency for XGBoost, LightGBM and Random Forest inference by leveraging COTS datacenter-grade FPGA accelerators. The accelerator comes in two flavors: a latency- and a throughput-optimized flavor. This product brief describes the latency-optimized version.

ALGORITHMIC TRADING

Algorithmic trading uses decision algorithms to automate trading instructions. The automated decisions are often made by XGBoost and LightGBM models. Low execution latency is an important goal in these systems. Silva provides an API for trading software frameworks that performs inference of XGBoost and LightGBM models with a latency of a few microseconds. This enables clients to execute trading instructions faster than their competitors.

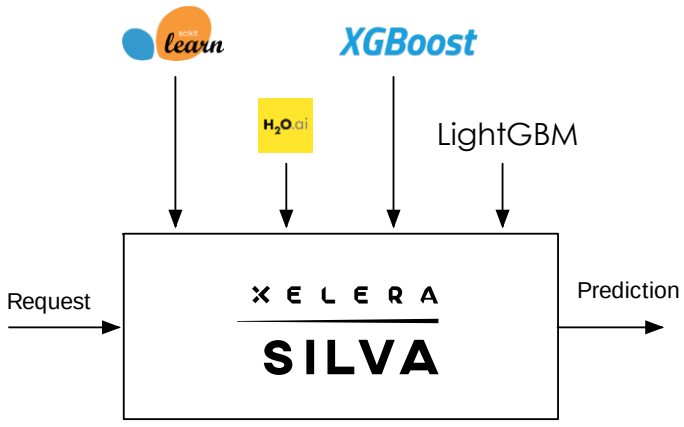
NETWORK THREAT DETECTION

The network is the first instance that must be protected to secure the enterprise IT system. Firewalls using Gradient Boosting Decision Trees perform inspection of the network flow data to detect suspicious patterns (e.g., a DDOS attack). Silva provides the necessary inference speed to ensure that the computationally intensive classification of network flows can be performed at the required bandwidth to avoid slowing down the network system.

OVERVIEW

Xelera Silva speeds up the inference of Gradient Boosting Decision Tree algorithms. It accelerates XGBoost and LightGBM Regression/Binomial Classification models:

- **Acceleration**: Single-digit microseconds latency; up to 9 fully concurrent models per accelerator
- **Guaranteed latency**: The FPGA accelerator kernels come with zero jitter.
- **Seamless adoption:** Usable with standard machine learning frameworks in a bring-your-own-model fashion; usable with a C/C++ or Python API
- **Deployment**: Built to run out-of-the-box on COTS server and datacenter FPGA hardware, and on public cloud instances



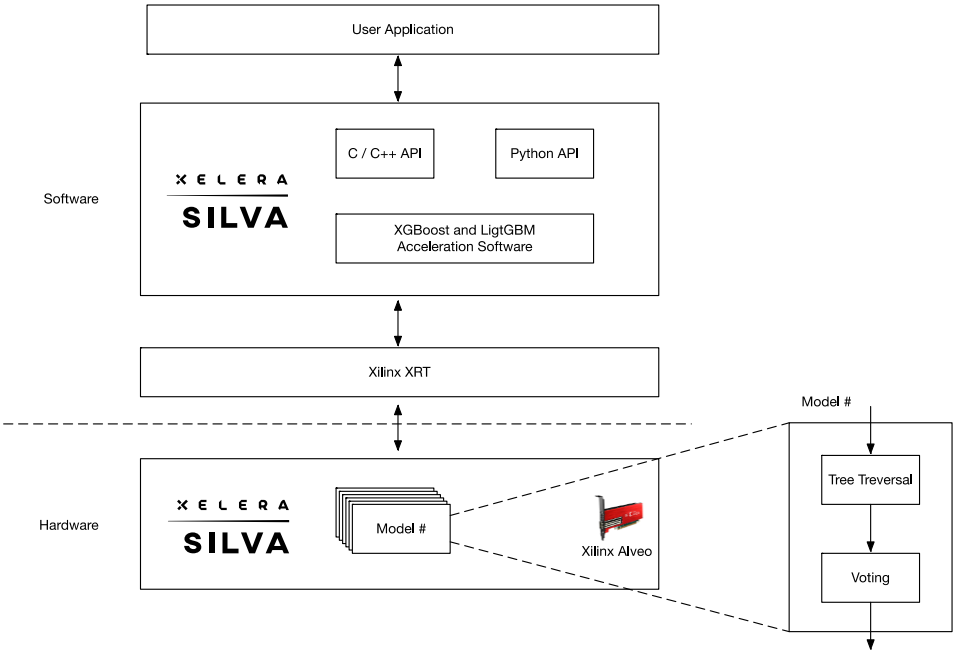
H2O and Sklearn are only supported by the throughput-optimized flavor.
H2O® is a trademark of H2O.ai.

Ultra-low Latency XGBoost / LightGBM Inference

Xelera Silva

USAGE DETAILS

Xelera Silva executes XGBoost and LightGBM models on Xilinx Alveo Data Center acceleration cards. The user application interacts with the acceleration software via a C/C++ or Python interface. The host software registers the models into the hardware accelerator and submits data for low-latency inference to it. The acceleration kernel on the FPGA contains up to 9 accelerators which can execute XGBoost and/or LightGBM models simultaneously. Each accelerator consists of low latency (sub micro-second) parallel tree traversal and voting units.

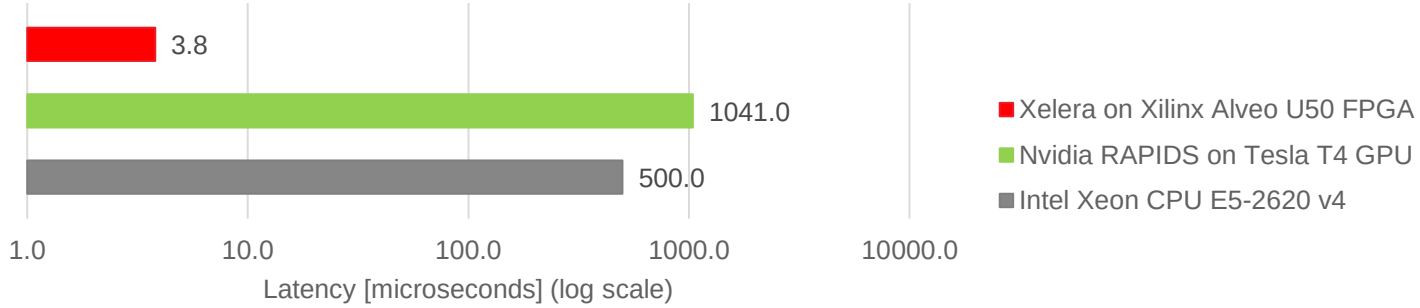


BENCHMARK

The graph below shows a latency comparison for an **XGBoost** regression between FPGA, GPU and CPU platforms, measured using a publicly available data set. The benchmark is exemplary in that the latency depends on the parameters of the input data set, the algorithm, and the machine learning model.

Data set: Malware Detection (Kaggle) (https://www.kaggle.com/c/malware-detection/overview) | Number of features: 23 | Maximum tree depth: 6 levels | Number of trees per model: 1000 | Number of concurrent models: 9 | Batch size: 1

- FPGA acceleration software: Xelera Silva
- GPU acceleration software: Nvidia RAPIDS CuML
- CPU: XGBoost (no hardware acceleration)



TAKE THE NEXT STEP

Request a free trial (on-premises or cloud):
sales@xelera.io

25 October 2021