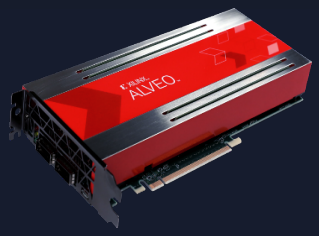
Decision Tree Ensembles

Random Forest, XGBoost and LightGBM Inference Acceleration

SOLUTION BRIEF

XILINX ALVEO™



- High-throughput Random Forest, XGBoost and LightGBM Inference
- One order of magnitude throughput over CPUs and GPUs
- One order of magnitude cost savings over CPUs and GPUs
- Usable with zero code change

INTRODUCTION

Gradient boosting frameworks such as XGBoost, LightGBM and CatBoost, as well as Random Forest algorithms are widely used techniques in recommender systems, search engines and payment platforms.

XGBoost, LightGBM, CatBoost, and Random Forest algorithms are based on learned decision tree ensembles. Such decision trees are fed with training data in order to teach them to ask the right questions about a data set: For example, if the decision tree shall predict whether a user will like a certain movie recommended to him on a website, the tree learns which features of the movie (*i.e.* the data set) are relevant to the user. After the training phase, new data are applied to the decision trees in order to make predictions autonomously without human interaction (**inference phase**).

In applications in which the request rate can amount to thousands of simultaneous predictions, there are two performance metrics in addition to the prediction accuracy:

- the sustained throughput (simultaneous queries)
- the response time of a single query

Xelera provides an accelerator software which offloads the inference phase to data center-grade Field-Programmable Gate Arrays (FPGAs) in order to increase the throughput at a guaranteed query response time.

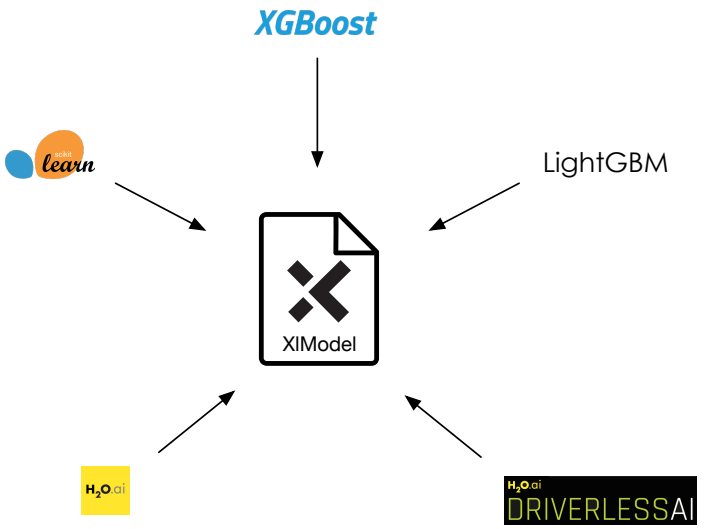
KEY BENEFITS

- **Acceleration**: One order of magnitude throughput over CPUs and GPUs
- **Cost savings:** One order of magnitude cost savings over CPUs and GPUs
- **Integration:** Integration with standard machine learning frameworks, usable with zero code change

SOLUTION OVERVIEW

The Decision Tree Accelerator speeds up the inference phase of gradient boosting trees and random forest algorithms. It works with models created with **XGBoost**, **LightGBM**, **Scikit Learn**, **H2O.ai** and **H2O Driverless AI**. The software allows data scientists and engineers to build fast, scalable and cost-efficient machine learning infrastructure, and it does **not require changes in the use of the machine learning frameworks**.

The Decision Tree Accelerator uses the fine-grained parallelism of FPGAs to execute the machine learning models significantly faster than on CPUs or GPUs. The acceleration of the machine learning inference enables more throughput per server node compared to running the frameworks without acceleration.



H2O® is a trademark of H2O.ai.

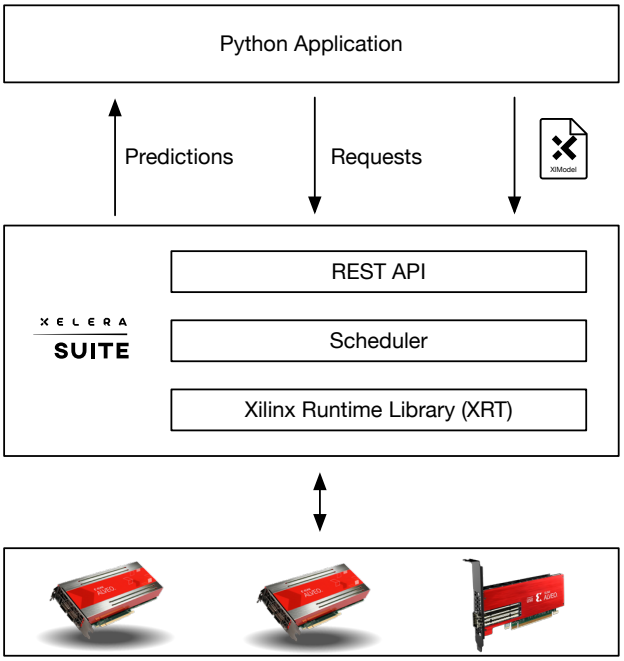
Decision Tree Ensembles

Random Forest, XGBoost and LightGBM Inference Acceleration

SOLUTION DETAILS

The Decision Tree Inference Accelerator consists of two parts:

1. The **Model Compiler** converts trained gradient boosting machine and trained random forest models from **XGBoost**, **LightGBM**, **Scikit Learn**, **H2O.ai** and **H2O Driverless AI** automatically into a unified model format (XIModel).
2. The **Acceleration Software**, based on Xelera Suite, loads the compiled XIModel and executes it on Xilinx Alveo U200, U250 and U50 platforms, and on AWS F1 cloud instances. The acceleration software provides a Python interface compatible to the aforementioned machine learning frameworks. The user application sends inference requests to a REST API and receives the corresponding predictions. Internal to the software layer, the Scheduler dispatches the requests to one or several FPGAs, and splits the workload across several server nodes if needed. The scheduling and FPGA access is transparent to the user. Knowledge of FPGAs is not required to use the accelerator.

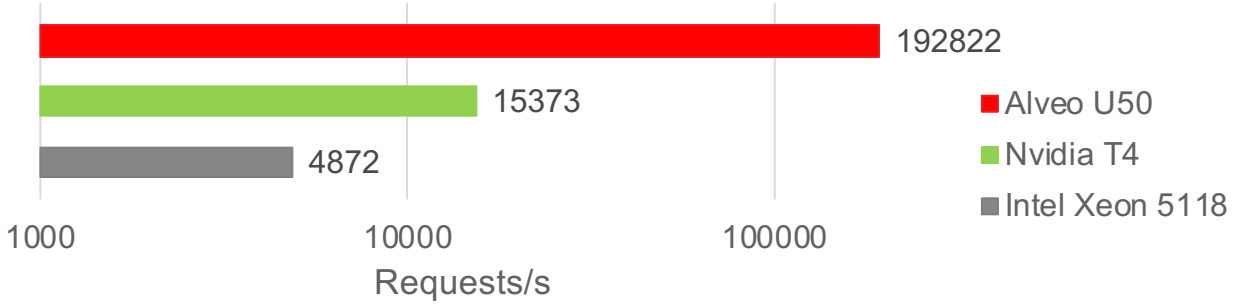


RESULTS

The graph below shows a throughput 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 throughput depends on the parameters of the input data set, the algorithm, and the machine learning model.

Data set: Flight (https://www.transtats.bts.gov/OT_Delay/OT_DelayCause1.asp) | Number of features: 10 | Maximum tree depth: 8 levels | Number of trees: 300

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



TAKE THE NEXT STEP

Request a free trial (Docker container or shared cloud image) by sending an email to sales@xelera.io.