Wide & Deep

CTR Model for Recommender System

SOLUTION BRIEF

- 13 Bucketized key + 24 Indicator + 6 Embedding (60K Look Up Table)
- 14.7k QPS
- 1.8ms Latency

INTRODUCTION

Google published its implementation of wide & deep learning on TensorFlow in 2016 and widely used by internet companies to significantly increases CTR (click-through rate) of advertisement.

Wide & Deep Learning for Recommender System can be viewed as a search ranking system, where the input query is a set of user and contextual information, and the output is a ranked list of items. Given a query, the recommendation task is to find the relevant items in database and then rank the items based on certain objectives, such as clicks or purchases.

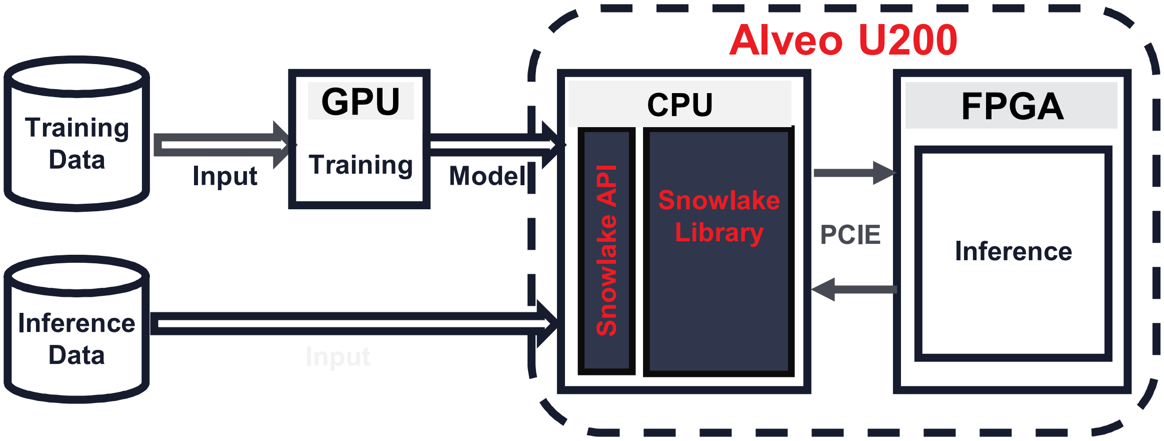
Compared with traditional Recommender System, memorization and generalization are added in Wide & Deep. As a result, the accuracy of prediction would be enhanced rapidly.

KEY BENEFITS

- 2.4x higher QPS
- 5.7x lower latency
- 82% lower TCO

SOLUTION OVERVIEW

Explanation and/or image to show what is important about the solution. If there is an image, put it here below. Imagery on page1 will help entice readers to learn more and turn the page.



Adaptable. Intelligent.

雪湖科技 snowlake-tech.com XILINX.

Wide & Deep

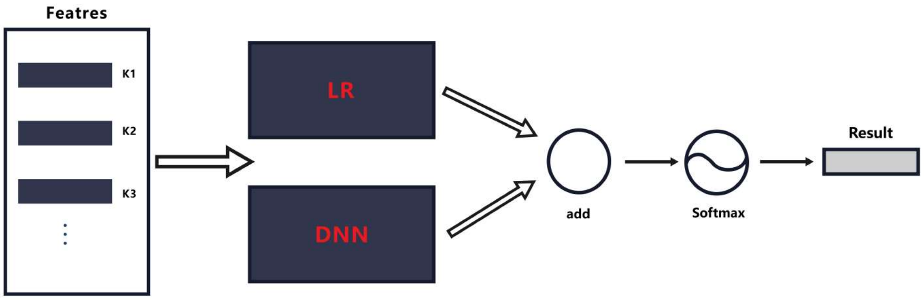
CTR Model for Recommender System

SOLUTION DETAILS

The product is designed to improve the performance of Wide & Deep model of Recommender System by using FPGA. It can achieve the lower latency and higher throughput than CPU to meet the requirement of rapidly increasing data to be dealt with.

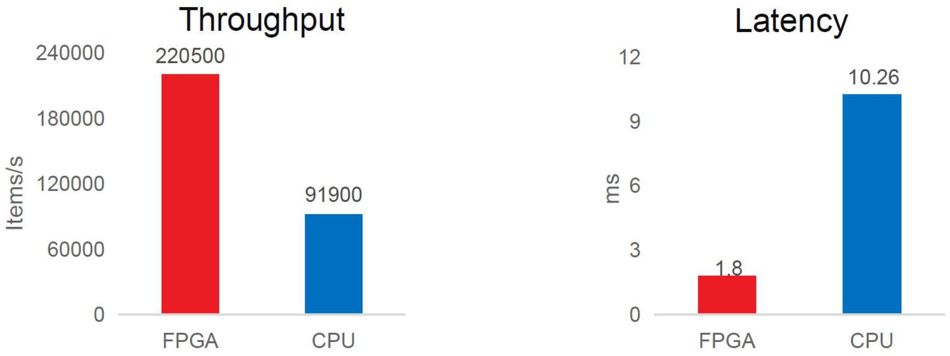
According to operators of model in TensorFlow, we utilize distributed pipeline calculation to implement this system on Alveo™ U200. With these optimizations, the inference performance can be dramatically improved.

In addition, a tool kit is developed for model deployment. It can be easily applied to renew model and deploy the system. At the same time, our system supports various models based on Google open source to guarantee system compatibility.



RESULTS

Compared with CPU, the implementation of system can achieve the higher throughput and lower latency. Comprehensively, it can achieve more than 5 times TCO (total cost of ownership).



TAKE THE NEXT STEP

Learn more about Xilinx [Alveo accelerator cards]
Learn more about Snowlake-Tech http://www.snowlake-tech.com/
Reach out to Snowlake-Tech sales – (yinqingyu@snowlake-tech.com)