УДК 004.31
Kirill Shanin[1], Nataliia Zhukova[2]
[1]student of group CST-110, National University «Zaporizhzhia Polytechnic»
[2]PhD (Philology), assistant prof. National University «Zaporizhzhia Polytechnic»

GOOGLE TENSOR PROCESSING UNIT

Machine learning has produced business and research breakthroughs ranging from network security to medical diagnoses. Tensor Processing Unit (TPU) created in order to make it possible for anyone to achieve similar breakthroughs. Cloud TPU is the custom-designed machine learning ASIC that powers Google products like Translate, Photos, Search, Assistant, and Gmail.

Cloud TPU is designed to run cutting-edge machine learning models with AI services on Google Cloud. And its custom high-speed network offers over 100 petaflops (quantity used to measure the performance of computers 1015) of performance in a single pod — enough computational power to transform your business or create the next research breakthrough. Training machine learning models is like compiling code: you need to update often, and you want to do so as efficiently as possible. ML models need to be trained over and over as apps are built, deployed, and refined.

Cloud TPU's robust performance and low cost make it ideal for machine learning teams looking to iterate quickly and frequently on their solutions. The minimum cloud-based TPU configuration consists of four 2-core chips and 64GB of HBM2 memory (high bandwidth memory provides higher bandwidth with less power consumption and significantly smaller footprint than DDR4 or GDDR5). The price is $6.50 per TPU per hour. Cloud TPU offering: Cloud TPU v2(180 teraflops, 64 GB High Bandwidth Memory (HBM) $4.50 / TPU hour), Cloud TPU v2 Pod (11.5

petaflops, 4 TB HBM, $8.00 / TPU hour), Cloud TPU v3 (420 teraflops, 128 GB HBM), Cloud TPU v3 Pod (100+ petaflops,32 TB HBM).

The potential of using Cloud TPU pods to accelerate deep learning research while keeping operational costs and complexity low is a big draw. It takes now a little over 24 hours to train models on our local GPU cluster. It will take depending on the size of the TPU pod, anywhere from 7 hours to 15 minutes. By using Cloud TPUs you can save money for fault-tolerant machine learning workloads, such as long training runs with checkpointing or batch prediction on large datasets. Preemptible Cloud TPUs are 70% cheaper than on-demand instances, making everything from your first experiments to large-scale hyperparameter searches more affordable than ever.