CPU Vs GPU Vs TPU

Specialization→ → →

10xAI

- **The CPU** is a general purpose processor
- It contains one ALU*[Arithmetic Logic Unit]* per core
- It is not designed for any specific type of task

Since it is too general purpose, it has no idea what the next instruction can be. So it read the instruction everytime from the program

-This is the point that is Traded-Off in GPU and even further with TPU

- **GPU** contains 2000-5000 ALU to do the calculations in parallel

<u>But its communication with CPU is expensive</u>
=> Bigger the calculation we sent to GPU better the Throughput/Latency ratio

- It gives a huge improvement over CPU for Neural Network by computing the Tensor multiplications/ addition in parallel

But is is not specialized for Deep Learning Or Neural Network

For every single calculation(*e.g. let's say all the multiplication of 1st layers weights and input data)* in the thousands of ALUs, its needs to communicate with CPU

- **TPU**[*Tensor Processing Unit*] is a matrix processor specialized for Neural Network work loads.

=>TPUs can't run any other software but they can handle the massive multiplications and additions for Neural Network

- Basically its hardware is designed for sequential Matrix multiplication and summation
- During the whole process of massive calculations and data passing, no memory access is required at all.

Read more -
[Google cloud]
[Blog]