Google Skills

What do CPU, GPU, and TPU stand for?

These acronyms are the three main types of processors we'll be discussing:

- **CPU:** Central processing unit
- **GPU:** Graphics processing unit
- **TPU:** Tensor processing unit

What is a CPU and where can I find one?

A CPU is the general-purpose brain of a computer. It's a master of all trades, designed with a few powerful cores to quickly handle a vast range of tasks one by one. CPUs run your operating systems, web browsers, and just about every other application on your device.

The greatest benefit of CPUs is their flexibility—you can load any kind of software on a CPU for many different types of applications. For example, you can use them for word processing on a PC, controlling rocket engines, executing bank transactions, or classifying images with a neural network.

A CPU loads values from memory, performs a calculation on the values, and stores the result back in memory for every calculation. Memory access is slow when compared to the calculation speed and can limit the total throughput of CPUs—often referred to as the von Neumann bottleneck.

Based on the von Neumann architecture, a CPU works with software and memory like this:

View animation

Note: This animation is for conceptual purposes and does not reflect the actual behavior of processors.

What is a GPU and where can I find one?

A GPU is a specialist, originally created to handle the intense work of rendering graphics for video games. Unlike a CPU, which has a few powerful cores, a GPU has thousands of smaller cores, or 'Arithmetic Logic Units (ALUs)' A modern GPU usually contains between 2,500–5,000 cores.

The large number of processors means you can execute thousands of multiplications and additions at the same time—or 'in parallel.'

GPUs are found in **PCs (especially for gaming), workstations, and the cloud**. It's here where GPUs turned out to be a great fit for the math-heavy work of training AI models—where massive parallelism accelerates matrix operations in a neural network, providing an order of magnitude with higher throughput than a CPU.

View animation
Note: This animation is for conceptual purposes and does not reflect the actual behavior of processors.

But because GPUs are still general purpose processors that support different applications and softwares, they have the same problem as CPUs—for every calculation in the thousands of ALUs, a GPU must access registers or shared memory to read operands and store the intermediate calculation results.

What is a TPU and why did Google create it?

A **TPU** is a custom-built processor (known as a custom ASIC) that Google designed specifically for one job: AI.

As Google's products became more intelligent, their computational demand grew. We started thinking about TPUs about 10 years ago when we realized that if every Google user used voice search for just three minutes a day, we'd have needed to double its number of data centers. So we created TPUs to help perform the specific calculations needed for AI with maximum performance and power efficiency. They are laser-focused on running **tensor operations**.

TPUs can't run word processors, control rocket engines, or execute bank transactions. But they can handle massive matrix operations used in neural networks at fast speeds. In addition to powering Google services like Gemini, Search and YouTube, they are available to the public through **Google Cloud**. This means anyone from individual researchers to world-leading companies can access the power of Cloud TPUs for their own AI and machine learning (ML) projects.

How do TPUs work?

The primary task for TPUs is matrix processing—a combination of multiply and accumulate operations. TPUs contain thousands of multiply-accumulators that are directly connected to each other to form a large physical matrix. We call this a systolic array architecture.

Here's how it works:

1. The TPU host streams data into an infeed queue
2. The TPU loads data from the infeed queue and stores it in HBM memory
3. When the computation is complete, the TPU loads the results into the outfeed queue
4. The TPU host reads the results from the outfeed queue and stores them in the host's memory
5. To perform the matrix operations, the TPU loads the parameters from HBM memory into the Matrix Multiplication Unit (MXU). **View animation**
6. The TPU loads data from HBM memory

As each multiplication is executed, the result is passed to the next multiply-accumulator. The output is the summation of all multiplication results between the data and parameters. No memory access is required during the matrix multiplication process.

View animation

As a result, TPUs can achieve a high-computational throughput on neural network calculations. We'll explore these concepts in more depth shortly.

What is a "tensor"?

So glad you asked! A **tensor** is the fundamental data structure used in machine learning and AI. While the mathematical definition can get complex, you can simply think of it as a **multi-dimensional array of numbers**:

- A single number (a scalar) is a 0-dimensional tensor
- A list of numbers (a vector) is a 1-dimensional tensor
- A grid of numbers (a matrix) is a 2-dimensional tensor

Neural networks process information using these multi-dimensional arrays, and TPUs are hardware specifically engineered to perform these tensor calculations at blinding speed.

So, which one is best for AI: CPU, GPU, or TPU?

The answer is: **it depends**. With Google Cloud, you can choose from GPUs, TPUs, or CPUs to drive reliable, low-cost inference. It's not about which is 'better' but rather which is the **right tool for the job you're completing**. Broadly speaking:

- Use a **CPU** for general-purpose computing, logic, and sequential tasks. For smaller or more latency-tolerant AI workloads, CPUs can offer a cost-effective solution thanks to advancements in model quantization, software, and hardware frameworks. This makes them incredibly effective for preliminary model experimentation and preprocessing tasks. They're also worth considering for classical ML tasks like recommendation systems, image classification, and batch inference small language models.
- **Use TPUs** and **GPUs** for all other AI tasks—we recommend testing them both to see what's right for you.
- The future isn't about choosing one over the other, it's about using them together. Inference engines like **vLLM** are helping developers leverage the unique strengths of each processor in a complementary way—building more powerful and efficient AI systems for everyone. More on when to use TPUs and accelerator interoperability later in this course.