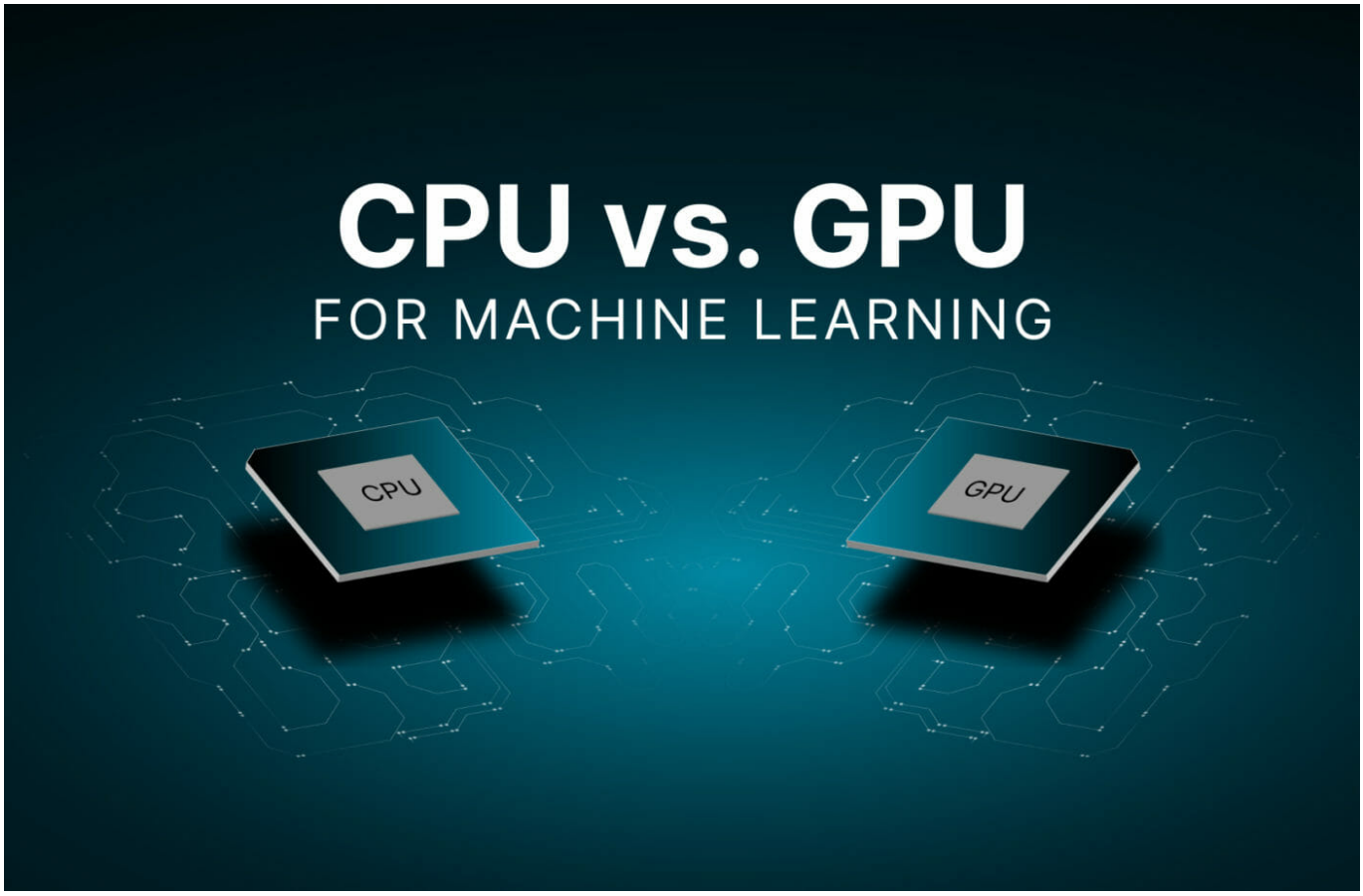
CPU vs. GPU for Machine Learning



While CPUs can process many general tasks in a fast, sequential manner, GPUs use parallel computing to break down massively complex problems into multiple smaller simultaneous calculations. This makes them ideal for handling the massively distributed computational processes required for machine learning.

In this article, we'll compare the differences between a CPU and a GPU, as well as the applications for each with machine learning, neural networks, and deep learning.

What Is a CPU?

A central processing unit, or CPU, is a processor that processes the basic instructions of a computer, such as arithmetic, logical functions, and I/O operations. It's typically a small but powerful chip integrated into the computer's motherboard.

A CPU is considered the computer's brain because it interprets and executes most of the computer's hardware and software instructions.

Standard components of a CPU include one or more cores, cache, memory management unit (MMU), and the CPU clock and control unit. These all work together to enable the computer to run multiple applications

at the same time.

The core is the central architecture of the CPU where all the computation and logic occur.

Traditionally, CPUs were single core, but today's CPUs are multicore, having two or more processors for enhanced performance. A CPU processes tasks sequentially with tasks divided among its multiple cores to achieve multitasking.

What Is a GPU?

A GPU, or graphics processing unit, is a computer processor that uses accelerated calculations to render intensive high-resolution images and graphics. While originally designed for rendering 2D and 3D images, videos, and animations on a computer, today's GPUs are used in applications far beyond graphics processing, including big analytics and machine learning. This kind of computing is often called "GPGPU," or "General Purpose GPU."

GPUs function similarly to CPUs and contain similar components (e.g., cores, memory, etc). They can be integrated into the CPU or they can be discrete (i.e., separate from the CPU with its own RAM).

GPUs use parallel processing, dividing tasks into smaller subtasks that are distributed among a vast number of processor cores in the GPU. This results in faster processing of specialized computing tasks.

GPUs vs. FPGAs: What's the Difference?

CPU vs. GPU: What's the Difference?

The fundamental difference between GPUs and CPUs is that CPUs are ideal for performing sequential tasks quickly, while GPUs use parallel processing to compute tasks simultaneously with greater speed and efficiency.

CPUs are general-purpose processors that can handle almost any type of calculation. They can allocate a lot of power to multitask between several sets of linear instructions to execute those instructions faster.

While CPUs can perform sequential tasks on complex computations quickly and efficiently, they are less efficient at parallel processing across a wide range of tasks.

GPUs are excellent at handling specialized computations and can have thousands of cores that can run operations in parallel on multiple data points. By batching instructions and pushing vast amounts of data at high volumes, they can speed up workloads beyond the capabilities of a CPU.

In this way, GPUs provide massive acceleration for specialized tasks such as machine learning, data analytics, and other artificial intelligence (AI) applications.

How Does a GPU Work?

While CPUs typically have fewer cores that run at high speeds, GPUs have many processing cores that operate at low speeds. When given a task, a GPU will divide it into thousands of smaller subtasks and process them concurrently, instead of serially.

In graphics rendering, GPUs handle complex mathematical and geometric calculations to create realistic visual effects and imagery. Instructions must be carried out simultaneously to draw and redraw images

hundreds of times per second to create a smooth visual experience.

GPUs also perform pixel processing, a complex process that requires phenomenal amounts of processing power to render multiple layers and create the intricate textures necessary for realistic graphics.

It is this high level of processing power that makes GPUs suitable for machine learning, AI, and other tasks that require hundreds or thousands of complex computations. Teams can increase compute capacity with high-performance computing clusters by adding multiple GPUs per node that can divide tasks into thousands of smaller subtasks and process them all at the same time.

How to Accelerate Apache Spark with RAPIDS on GPU

CPU vs. GPU for Machine Learning

Machine learning is a form of artificial intelligence that uses algorithms and historical data to identify patterns and predict outcomes with little to no human intervention. Machine learning requires the input of large continuous data sets to improve the accuracy of the algorithm.

While CPUs aren't considered as efficient for data-intensive machine learning processes, they are still a cost-effective option when using a GPU isn't ideal.

Such use cases include machine learning algorithms, such as time series data, that don't require parallel computing, as well as recommendation systems for training that need lots of memory for embedding layers. Some algorithms are also optimized to use CPUs over GPUs.

The more data, the better and faster a machine learning algorithm can learn. The technology in GPUs has advanced beyond processing high-performance graphics to use cases that require high-speed data processing and massively parallel computations. As a result, GPUs provide the parallel processing necessary to support the complex multistep processes involved in machine learning.

CPU vs. GPU for Neural Networks

Neural networks learn from massive amounts of data in an attempt to simulate the behavior of the human brain. During the training phase, a neural network scans data for input and compares it against standard data so that it can form predictions and forecasts.

Because neural networks work primarily with massive data sets, training time can increase as the data set grows. While it's possible to train smaller-scale neural networks using CPUs, CPUs become less efficient at processing these large volumes of data, causing training time to increase as more layers and parameters are added.

Neural networks form the basis of deep learning (a neural network with three or more layers) and are designed to run in parallel, with each task running independently of the other. This makes GPUs more suitable for processing the enormous data sets and complex mathematical data used to train neural networks.

CPU vs. GPU for Deep Learning

A deep learning model is a neural network with three or more layers. Deep learning models have highly flexible architectures that allow them to learn directly from raw data. Training deep learning networks with

large data sets can increase their predictive accuracy.

CPUs are less efficient than GPUs for deep learning because they process tasks in order one at a time. As more data points are used for input and forecasting, it becomes more difficult for a CPU to manage all of the associated tasks.

Deep learning requires a great deal of speed and high performance and models learn more quickly when all operations are processed at once. Because they have thousands of cores, GPUs are optimized for training deep learning models and can process multiple parallel tasks up to three times faster than a CPU.

Power Machine Learning with Next-gen AI Infrastructure

GPUs play an important role in the development of today's machine learning applications. When choosing a GPU for your machine learning applications, there are several manufacturers to choose from, but NVIDIA, a pioneer and leader in GPU hardware and software (CUDA), leads the way.

AIRI//S™ is modern AI infrastructure architected by Pure Storage® and NVIDIA and powered by the latest NVIDIA DGX systems and Pure Storage FlashBlade//S™.

AIRI//S is an out-of-the-box AI solution that simplifies your AI deployment to deliver simple, fast, next-generation, future-proof infrastructure to meet your AI demands at any scale.

Simplify AI at scale with AIRI//S.