CPUs and GPUs have a lot in common. They are **both silicon-based microprocessors**. At the same time, they are substantially different, and they are deployed for different roles.

Any data scientist or machine learning enthusiast who has been trying to elicit performance of her learning models at scale will at some point hit a cap and start to experience various degrees of processing lag.

Tasks that take minutes with smaller training sets may now take more hours—in some cases weeks—when datasets get larger. You’ll need the best hardware, and while researching you will come across and may get confused with CPUs, GPUs, and ASICs.

**A Short History**

A central processing unit (CPU) is essentially the brain of any computing device, carrying out the instructions of a program by performing control, logical, and input/output (I/O) operations.

The first CPU, the 4004 unit, was developed by Intel just 50 years ago in the 1970s. Most CPUs then were designed with one “core,” meaning that only one operation could be performed at a time. Years later, owing to vast improvements in chip design, research, and manufacturing, the computing market advanced to dual and multi-core CPUs which were faster because they could now perform two or more operations at a time.

Today’s CPUs, however, have just a handful of cores, and its basic design and purpose—to process complex computations—has not really changed. Essentially, they’re mostly applicable for problems that require parsing through or interpreting complex logic in code.

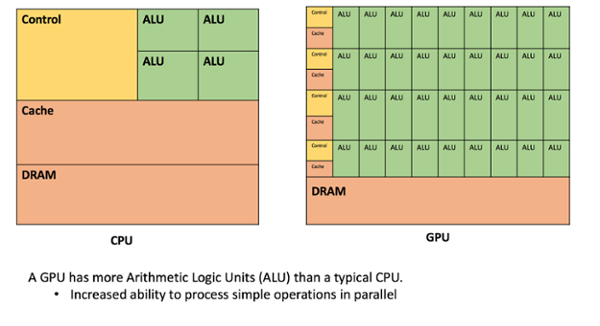
**GPU's Rise**

#### What are CPUs and GPUs?

A CPU (central processing unit) is often called the “brain” or the “heart” of a computer. It is required to run the majority of engineering and office software. However, there is a multitude of tasks that can overwhelm a computer’s central processor. That is when using GPU becomes essential for computing.

A GPU (graphics processing unit) is a specialized type of microprocessor, **primarily designed for quick image rendering.** GPUs appeared as a response to graphically intense applications that put a burden on the CPU and degraded computer performance. They became a way to offload those tasks from CPUs, but modern graphics processors are powerful enough to perform **rapid mathematical calculations** for many other purposes apart from rendering.

A graphical processing unit (GPU), on the other hand, has smaller-sized but many more logical cores (arithmetic logic units or ALUs, control units and memory cache) whose basic design is to process a set of simpler and more identical computations in parallel.

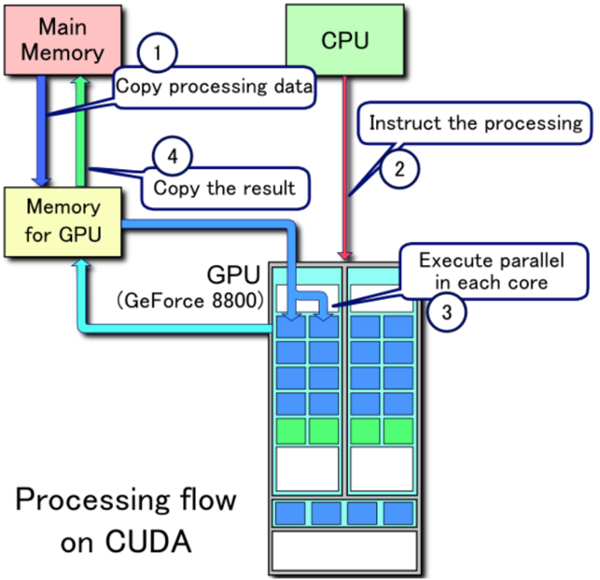


*Figure 1: CPU vs GPU*

While GPUs have certainly been around as long as gaming applications since the 1970s, it wasn’t until Nvidia released its GeForce processor line of “graphics processing units” or GPUs that they became more popular.

Initially designed mainly as dedicated graphical rendering workhorses of computer games, GPUs were later enhanced to accelerate other geometric calculations (for instance, transforming polygons or rotating verticals into different coordinate systems like 3D).

Nvidia created a parallel computing architecture and platform for its GPUs called [CUDA](https://developer.nvidia.com/cuda-zone), which gave developers access and the ability to express simple processing operations in parallel through code.



*Figure 2: GPU Parallel Architecture (Wikipedia)*

Moreover, most of these computations involved matrix and vector operations, the same type of mathematics that is used in data science today. It wasn’t too long before engineers and non-gaming scientists studied how GPUs might be also used for non-graphical calculations.

**Accessing CPUs and GPUs**

CPUs are accessible today to data science practitioners on the cloud, using serverless microservices, or “[backend-as-a-service” or BAAS](https://www.martinfowler.com/articles/serverless.html) architectures. Developers can add API-driven machine learning services to any application with diverse libraries on computer vision, speech or language, as well as integration with modern tools like data lakes or stream processing.

How would you choose among these? As with any data science project, it depends. There are tradeoffs to consider, between speed, reliability, and cost. As a general rule, GPUs are a safer bet for fast machine learning because, at its heart, data science model training is composed of simple matrix math calculations, the speed of which can be greatly enhanced if the computations can be carried out in parallel.

In other words, CPUs are best at handling single, more complex calculations sequentially, while GPUs are better at handling multiple but simpler calculations in parallel.

GPU compute instances will typically cost 2-3x that of CPU compute instances, so unless you’re seeing 2-3x performance gains in your GPU-based training models, I would suggest going with CPUs.

**The Future**

Application Specific Integrated Circuits, or ASICs, are on the horizon next for processor design—single-purpose chips customized specifically for one type of function.

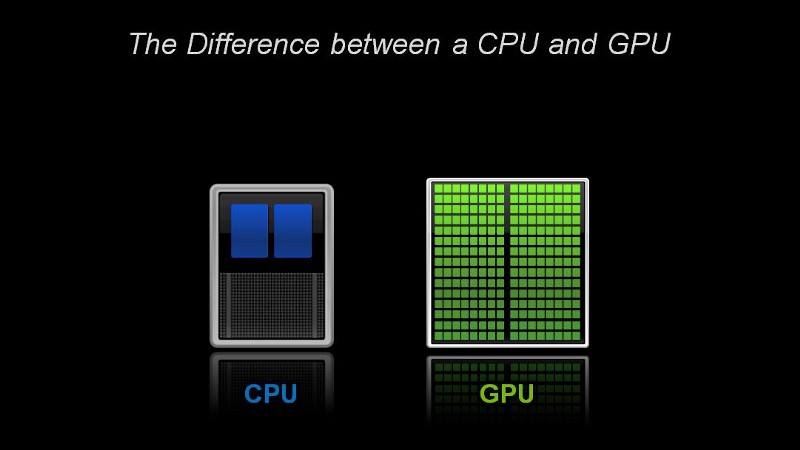
One great example of ASICs is a class of processors that are being developed for [Bitcoin mining](https://en.bitcoin.it/wiki/List_of_Bitcoin_mining_ASICs) and involve solving mathematically intensive[“hash” computations](https://coinsutra.com/hash-rate-or-hash-power/) to find blocks (in a blockchain). One of the more notable Bitcoin ASICs is [BitFury](https://bitfury.com/content/downloads/16-nm-asic-datasheet-2018-04-003.pdf)’s, whose 16nm multi-core ASIC can achieve over 80 GH/s, or over 80 billion hashes per second.

In a [recent paper](https://www2.deloitte.com/content/dam/Deloitte/global/Images/infographics/technologymediatelecommunications/gx-deloitte-tmt-2018-nextgen-machine-learning-report.pdf), Deloitte Global predicts that by the end of 2018, over 25% of all chips to accelerate machine learning will be ASICs (and FPGAs). When this happens, machine-learning enabled applications are likely to cause big changes in industry while expanding to new areas.

**What is the difference?**

CPUs and GPUs process tasks in different ways. Regarding interrelations, they are often compared with brain and brawn. A CPU (the brain) can work on a variety of different calculations, while a GPU (the brawn) is best at focusing all the **computing abilities on a specific task.** That is because a CPU consists of a few cores (up to 24) optimized for sequential serial processing. It is designed to maximize the performance of a single task within a job; however, the range of tasks is wide. On the other hand, a **GPU uses thousands of smaller and more efficient cores for a massively parallel architecture aimed at handling multiple functions at the same time.**

Modern GPUs provide superior processing power, memory bandwidth and efficiency over their CPU counterparts. They are 50–100 times faster in tasks that require **multiple parallel processes, such as machine learning and big data analysis.**



Source: blogs.nvidia.com

#### What problems are GPUs suited to address?

GPU computing [is defined](http://www.nvidia.com/object/what-is-gpu-computing.html) as the use of a **GPU together with a CPU** to accelerate scientific, analytics, engineering, consumer, and enterprise applications.

For many years, GPUs have powered the display of images and motion on computer displays, but they are technically capable of doing more. Graphics processors are brought into play when **massive calculations are needed on a single task**.

That task may include:

* **Games**

A graphics processing unit is essential for fast, graphic-intensive rendering of the gaming world. Rendering of special effects and sophisticated 3D graphics in real time requires some serious computing power. The tasks of modern games become too heavy for CPU graphics solution. Games even made a step further with virtual reality, which is so believable because GPUs can quickly render and maintain realistic images with proper lighting and shading.

* **3D Visualization**

GPUs drive viewport performance in 3D visualization applications such as **computer-aided design (CAD).** Software that lets you visualize objects in 3 dimensions relies on GPUs to draw those models in real time as you rotate or move them.

* **Image Processing**

GPUs can accurately process millions of images to find differences and similarities. This ability is extensively used in industries such as border control, security, and medical x-ray processing. For example, in 2010, the US military linked together more than 1,700 Sony PlayStation 3TM systems to process high-resolution satellite imagery more quickly.

* **Big Data**

With thousands of computational cores and 10–100x application throughput compared to CPUs alone, graphics units are the choice for processing big data for scientists and industry. GPUs are used to depict data as interactive visualization, and they integrate with other datasets in order to explore volume and velocity of data. For example, we are now able to power up **gene mapping by processing data** and analyzing co-variances to understand the relationship between different combinations of genes.

* **Deep Machine Learning**

Machine learning has been around for some time now, but powerful and efficient GPU computing has raised it to a new level. Deep learning is the use of sophisticated neural networks to create systems that can perform feature detection from massive amounts of unlabeled training data. GPUs can process tons of training data and train neural networks in areas like image and video analytics, speech recognition and natural language processing, self-driving cars, computer vision and much more.

**GPUs are not replacements for CPU architecture. Rather, they are powerful accelerators for existing infrastructure.** GPU-accelerated computing offloads compute-intensive portions of the application to the GPU, while the remainder of the code still runs on the CPU. From a user’s perspective, applications just run much faster. While general-purpose computing is still the CPU’s domain, GPUs are the hardware backbone of nearly all intensive computational applications.