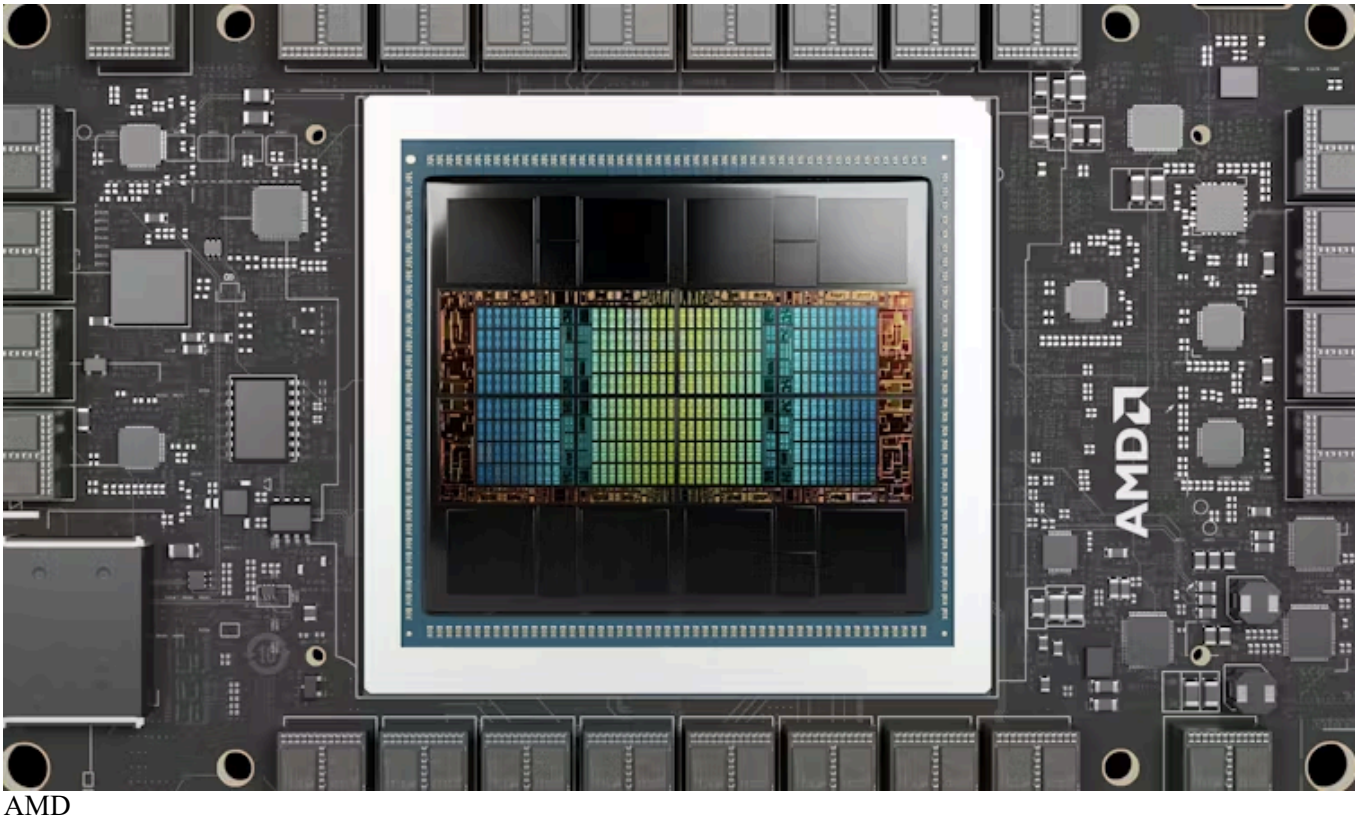
What is a GPU? An expert explains the chips powering the AI boom, and why they're worth trillions

What is a GPU? An expert explains the chips powering the AI boom, and why they're worth trillions



AMD

Conrad Sanderson, *CSIRO*

As the world rushes to make use of the latest wave of AI technologies, one piece of high-tech hardware has become a surprisingly hot commodity: the graphics processing unit, or GPU.

A top-of-the-line GPU can sell for tens of thousands of dollars, and leading manufacturer NVIDIA has seen its market valuation soar past US$2 trillion as demand for its products surges.

GPUs aren't just high-end AI products, either. There are less powerful GPUs in phones, laptops and gaming consoles, too.

By now you're probably wondering: what is a GPU, really? And what makes them so special?

What is a GPU?

GPUs were originally designed primarily to quickly generate and display complex 3D scenes and objects, such as those involved in video games and computer-aided design software. Modern GPUs also handle tasks such as decompressing video streams.

The "brain" of most computers is a chip called a central processing unit (CPU). CPUs can be used to generate graphical scenes and decompress videos, but they are typically far slower and less efficient on these tasks compared to GPUs. CPUs are better suited for general computation tasks, such as word processing and browsing web pages.

How are GPUs different from CPUs?

A typical modern CPU is made up of between 8 and 16 "cores", each of which can process complex tasks in a sequential manner.

GPUs, on the other hand, have thousands of relatively small cores, which are designed to all work at the same time ("in parallel") to achieve fast overall processing. This makes them well suited for tasks that require a large number of simple operations which can be done at the same time, rather than one after another.

Traditional GPUs come in two main flavours.

First, there are standalone chips, which often come in add-on cards for large desktop computers. Second are GPUs combined with a CPU in the same chip package, which are often found in laptops and game consoles such as the PlayStation 5. In both cases, the CPU controls what the GPU does.

Why are GPUs so useful for AI?

It turns out GPUs can be repurposed to do more than generate graphical scenes.

Many of the machine learning techniques behind artificial intelligence (AI), such as deep neural networks, rely heavily on various forms of "matrix multiplication".

This is a mathematical operation where very large sets of numbers are multiplied and summed together. These operations are well suited to parallel processing, and hence can be performed very quickly by GPUs.

What's next for GPUs?

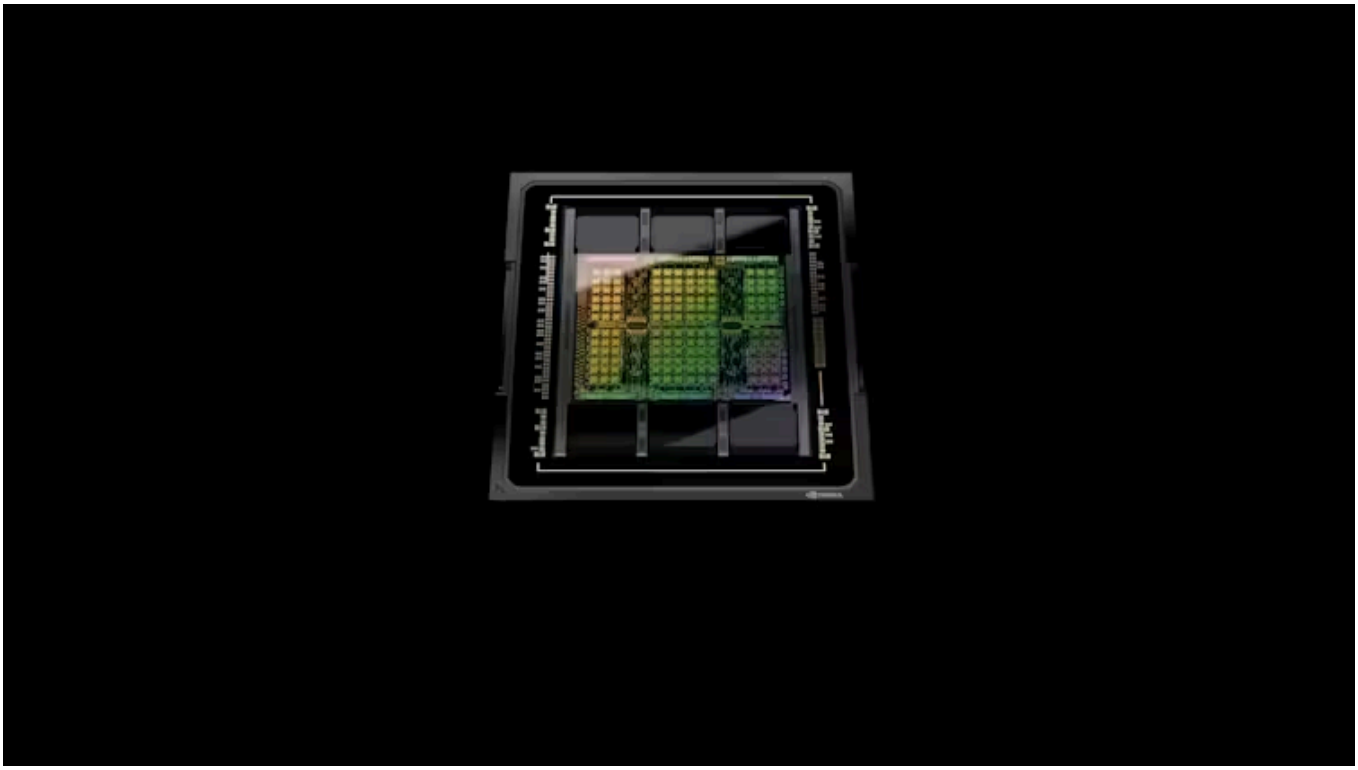
The number-crunching prowess of GPUs is steadily increasing, due to the rise in the number of cores and their operating speeds. These improvements are primarily driven by improvements in chip manufacturing by companies such as TSMC in Taiwan.

The size of individual transistors – the basic components of any computer chip – is decreasing, allowing more transistors to be placed in the same amount of physical space.

However, that is not the entire story. While traditional GPUs are useful for AI-related computation tasks, they are not optimal.

Just as GPUs were originally designed to accelerate computers by providing specialised processing for graphics, there are accelerators that are designed to speed up machine learning tasks. These accelerators are often referred to as "data centre GPUs".

Some of the most popular accelerators, made by companies such as AMD and NVIDIA, started out as traditional GPUs. Over time, their designs evolved to better handle various machine learning tasks, for example by supporting the more efficient "brain float" number format.



NVIDIA's latest GPUs have specialised functions to speed up the 'transformer' software used in many modern AI applications. NVIDIA

Other accelerators, such as Google's Tensor Processing Units and Tenstorrent's Tensix Cores, were designed from the ground up for speeding up deep neural networks.

Data centre GPUs and other AI accelerators typically come with significantly more memory than traditional GPU add-on cards, which is crucial for training large AI models. The larger the AI model, the more capable and accurate it is.

To further speed up training and handle even larger AI models, such as ChatGPT, many data centre GPUs can be pooled together to form a supercomputer. This requires more complex software in order to properly harness the available number crunching power. Another approach is to create a single very large accelerator, such as the "wafer-scale processor" produced by Cerebras.

Are specialised chips the future?

CPUs have not been standing still either. Recent CPUs from AMD and Intel have built-in low-level instructions that speed up the number-crunching required by deep neural networks. This additional functionality mainly helps with "inference" tasks – that is, using AI models that have already been developed elsewhere.

To train the AI models in the first place, large GPU-like accelerators are still needed.

It is possible to create ever more specialised accelerators for specific machine learning algorithms. Recently, for example, a company called Groq has produced a "language processing unit" (LPU) specifically designed for running large language models along the lines of ChatGPT.

However, creating these specialised processors takes considerable engineering resources. History shows the usage and popularity of any given machine learning algorithm tends to peak and then wane – so expensive specialised hardware may become quickly outdated.

For the average consumer, however, that's unlikely to be a problem. The GPUs and other chips in the products you use are likely to keep quietly getting faster.

Conrad Sanderson, Research Scientist & Team Leader, *CSIRO*