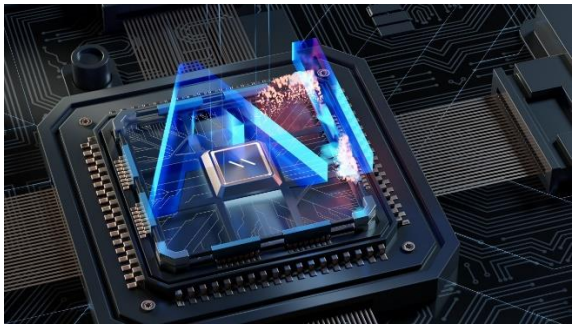

How GPU Cards Are Supporting Large Language Model Development

As artificial intelligence gathers pace, becoming ever more capable and powerful, large language models are increasingly valuable. It is these LLMs that deliver the understanding and language generation abilities AI systems need to function. They also provide the basis for generative AI structures – a vital part of modern AI.

But where are these large language models coming from? Well, they need to be developed, and this means we need high performance computing systems capable of handling this development.

And at the heart of the high performance system? An LLM GPU – the computing component that underpins the entire process. In this article, we'll look at these components in more detail and discover more about the role they play in development.

What Is a Large Language Model?

A large language model, or LLM, is essentially the foundation on which artificial intelligence – particularly generative artificial intelligence – is built.

These LLM structures are able to replicate human language patterns, both understanding and generating language in the same way that a person would. They are trained on enormous volumes of data, giving them a profound artificial understanding of how language works and how it is used.

This results in artificial language use that is remarkably close to genuine human speech, interaction, and cognition.

GPU vs. CPU – What's the Difference?

Pitting graphics processing units (GPUs) versus central processing units (CPUs) might not be the best way to go about this. While GPUs are more commonly associated with creating, training, and managing LLMs, both these components have a role to play.

The CPU is essentially the nerve center of the server. Without a CPU, the server cannot operate. As such, CPUs are vital to LLM development. The LLM needs a data server, and this server needs a CPU in order to function.

A GPU works in a similar way to a CPU. Both complete the calculations and tasks required to deliver data to the software that needs it. However, there are key differences. A GPU is able to perform more complex calculations by breaking tasks down into individual milestones, which

are then completed concurrently. So while a CPU will need to complete one milestone, then another, and then another, a GPU can complete all these separate calculations at the same time.

With this in mind, the CPU is still necessary to keep the server operating as it should. The GPU, on the other hand, achieves the power needed to deliver training data to the LLM.

Why GPU Cards Are Critical to LLM Development

Now we know what role GPU cards play in developing the high performance computing system required for large language model training, we can examine this role in more detail.

Here's why LLM GPU cards are so crucial to the future of artificial intelligence:

Enhanced Processing Power

First and foremost, a graphics processing unit offers significantly higher levels of processing power than a CPU does. For many tasks, the capabilities of a CPU are just fine, and users will have no problems when they use these components. However, developing a large language model is very different.

The process is so data-intensive, and requires so much computing power, that traditional CPUs simply won't cut it. Without the additional power provided by a GPU, large language model development is simply not possible.

Better Flexibility

LLM GPU cards don't just provide more power. They deliver more flexibility, too, when compared to processing units like CPUs.

This is because a GPU is able to carry out actions in parallel. With a GPU card, developers can complete a multitude of different tasks at the same time. Milestones can be reached simultaneously, which reduces the time taken to progress to the next stage of development. Projects are completed more swiftly, and developers have the flexibility they need to make changes and alterations when required.

A traditional CPU will complete tasks in series. When one task is completed, the processor can move on to the next one. This significantly reduces the speed and flexibility of development, and can lead to a major backlog in the process.

Applications of LLM GPU Cards

LLM GPU cards can be used in a number of different applications, each relating to a different stage of the process.

- Gathering and processing data required for large language model creation.
- Monitoring and testing the large language model once it is launched.
- Gathering and processing data for large language model fine-tuning and further development.
- Managing the functionality of generative AI apps and software, based on the large language model.

Explore the High Performance Computing System Components You Need

At Avalue, we recognize the power of digital technology. We also understand the rapid pace of development and the exciting potential of artificial intelligence, machine learning, and generative AI.

This is why we aim to put this power and potential into the hands of all businesses. To achieve this, we offer our on-premises AI chatbot training solution, giving organizations of all shapes and sizes the chance to develop smart applications of their own.

In addition to this, we provide high performance computing components so businesses can build, manage, and evolve their own high performance systems.