Advanced chips will be key weapons in war of the AI world

INSIDE BUSINESS

TECHNOLOGY

Richard Waters

As an all-out war for AI dominance breaks out in the tech industry, Wall Street has placed an early bet on who the biggest winners will be: the companies that make the weapons that will be used by all combatants. That means, specifically, the advanced chips needed for "generative AI" systems such as the ChatGPT chatbot and image-generating systems such as Dall-E.

And investors are not betting on just any weapon maker. Shares in Nvidia, whose graphical processing units — or GPUs — dominate the market for training large AI models, have surged 55 per cent this year. They have also doubled since October, when Nvidia was under a cloud from a combination of the crypto bust (its chips were widely used by crypto miners), a collapse in PC sales and a badly managed product transition in data centre chips.

A "picks and shovels" investment strategy makes sense when it is still hard to tell how a new technology will play out. The Big Tech companies are gearing up to wield expensive new AI systems against each other with no clear sign yet of how to gain a lasting edge. The one sure thing is that a lot of advanced silicon will be deployed and energy consumed. But what type of silicon will it be — and who will be best placed to supply it?

It seems safe to say that GPUs will be in high demand, benefiting Nvidia and, to a lesser extent, AMD (whose shares are up 30 per cent this year). Besides the job of training large AI models, GPUs are also likely to be more widely used in inferencing — the job of comparing real-world data against a trained model to provide a useful answer.

Until now, AI inferencing has been a healthy market for companies such as Intel that make CPUs (processors that can handle a wider range of tasks but are less efficient to run). But the AI models used in generative systems are likely to be too large for CPUs, requiring more powerful GPUs to handle this task, according to Karl Freund at Cambrian AI Research.

Five years ago, it was far from certain that Nvidia would be in this position. With the computational demands from machine learning rising exponentially, a spate of start-ups emerged to make specialised AI "accelerators". These so-called ASICs — application-specific integrated circuits, designed to perform just one task but in the most efficient way — suggested a better way to handle an intensive data-crunching operation.

> If the AI market ends up in the hands of a small number of tech giants, Nvidia's long-term prospects will be crimped

Yet predictions that GPUs would fail to match this purpose-built hardware have proved wrong and Nvidia remains on top. That owes much to its Cuda software, which is used for running applications on the company's GPUs, tying developers to Nvidia chips and reducing the incentive to buy from AMD.

Nvidia also has a new product hitting the market at the right time, in the form of its new H100 chip. This has been specifically designed to handle transformers, the AI technique behind recent big advances in language and vision models. For designers of ASICs, changes in underlying architecture like this are hard to handle. Redesigning each new generation of chips is expensive and it can be hard to sell enough to amortise the development costs.

But the competition is about to get more fierce. Microsoft's success in harnessing OpenAI research to take an early lead in generative AI owes a lot to the specialised hardware it has built to run the OpenAI models. These are based on GPUs, but the chip industry has been rife with speculation that the software giant is now designing its own AI accelerators.

If it does, it certainly won't be alone. Google decided eight years ago to design its own chips, known as tensor processing units, or TPUs, to handle its most intensive AI work. Amazon and Meta have followed. The idea of transformers originated at Google, suggesting that it, at least, will have optimised its latest chips to work with the new AI models.

Another threat could come from OpenAI itself. The research company behind ChatGPT has developed its own software, called Triton, to help developers run their neural networks on GPUs. That could reduce the need for Nvidia's Cuda — one step towards turning its chips into a commodity and giving developers such as OpenAI the chance to deploy their models on any hardware.

If the AI market ends up in the hands of a small number of giant tech companies, each with ample economic incentive to design their own specialised chips, Nvidia's long-term prospects will be crimped. But it has defied doubters before and, for now, is well placed for the tech world's latest bout of AI mania.

richard.waters@ft.com