Microsoft sends a new kind of AI processor into the cloud

reader comments

38 with 26 posters participating

Share this story

Microsoft rose to dominance during the ’80s and ’90s thanks to the success of its Windows operating system running on Intel’s processors, a cosy relationship nicknamed “Wintel”.

Now Microsoft hopes that another another hardware–software combo will help it recapture that success—and catch rivals Amazon and Google in the race to provide cutting-edge artificial intelligence through the cloud.

Microsoft hopes to extend the popularity of its Azure cloud platform with a new kind of computer chip designed for the age of AI. Starting today, Microsoft is providing Azure customers with access to chips made by the British startup Graphcore.

Graphcore, founded in Bristol, UK, in 2016, has attracted considerable attention among AI researchers—and several hundred million dollars in investment—on the promise that its chips will accelerate the computations required to make AI work. Until now it has not made the chips publicly available or shown the results of trials involving early testers.

Microsoft, which put its own money into Graphcore last December as part of a $200 million funding round, is keen to find hardware that will make its cloud services more attractive to the growing number of customers for AI

applications.

Unlike most chips used for AI, Graphcore's processors were designed from scratch to support the calculations that help machines to recognize faces, understand speech, parse language, drive cars, and train robots. Graphcore expects it will appeal to companies running business-critical operations on AI, such as self-driving-car startups, trading firms, and operations that process large quantities of video and audio. Those working on next-generation AI algorithms may also be keen to explore the platform's advantages.

Microsoft and Graphcore today published benchmarks that suggest the chip matches or exceeds the performance of the top AI chips from Nvidia and Google using algorithms written for those rival platforms. Code written specifically for Graphcore's hardware may be even more efficient.

The companies claim that certain image-processing tasks work many times faster on Graphcore's chips, for example, than on its rivals using existing code. They also say they were able to train a popular AI model for language processing, called BERT, at rates matching those of any other existing hardware.

BERT has become hugely important for AI applications involving language. Google recently said that it is using BERT to power its core search business. Microsoft says it is now using Graphcore's chips for internal AI research projects involving natural language processing.

Karl Freund, who tracks the AI chip market at Moor Insights, says the results show the chip is cutting-edge but still flexible. A highly-specialized chip could outperform one from Nvidia or Google but would not be programmable enough for engineers to develop new applications. "They've done a good job making it programmable, he says. "Good performance in both training and inference is something they've always said they would do, but it is really, really hard."

Freund adds that the deal with Microsoft is crucial for Graphcore's business, because it provides an on-ramp for customers to try the new hardware. The chip may well be superior to existing hardware for some applications, but it takes a lot of effort to redevelop AI code for a new platform. With a couple of exceptions, Freund says, the chip's benchmarks are not eye-popping enough to lure companies and researchers away from the hardware and software they are already comfortable using.

Graphcore has created a software framework called Poplar, which allows existing AI programs to be ported to its hardware. Plenty of existing algorithms may still be better-suited to software that runs on top of rival hardware, though. Google's Tensorflow AI software framework has become the de facto standard for AI programs in recent years, and it was written specifically for Nvidia and Google chips. Nvidia is also expected to release a new AI chip next year, which is likely to have better performance.

Nigel Toon, cofounder and CEO of Graphcore, says the companies began working together a year after his company's launch, through Microsoft Research Cambridge in the UK. His company's chips are especially well-suited to tasks that involve very large AI models or temporal data, he says. One customer in finance supposedly saw a 26-fold performance boost in an algorithm used to analyze market data thanks to Graphcore's hardware.

A handful of other, smaller companies also announced today

that they are working with Graphcore chips through Azure. This includes Citadel, which will use the chips to analyze financial data, and Qwant, a European search engine that wants the hardware to run an image-recognition algorithm known as ResNext.

The AI boom has already shaken up the market for computer chips in recent years. The best algorithms perform parallel mathematical computations, which can be done more effectively on a graphics chips (or GPUs) that have hundreds of simple processing cores as opposed to conventional chips (CPUs) that have a few complex processing cores.

The GPU-maker Nvidia has ridden the AI wave to riches, and Google announced in 2017 that it would develop its own chip, the Tensor Processing Unit, which is architecturally similar to a GPU but optimized for Tensorflow.

Graphcore's chips, which it calls intelligence processing units (IPUs), have many more cores than GPUs or TPUs. They also feature memory on the chip itself, which removes a bottleneck that comes with moving data onto a chip for processing and off again.

Facebook is also working on its own AI chips. Microsoft has previously touted reconfigurable chips made by Intel and customized by its engineers for AI applications. A year ago, Amazon revealed it was also getting into chipmaking, but with a more general-purpose processor optimized for Amazon's cloud services.

More recently, the AI boom has sparked a flurry of startup hardware companies to develop more specialized chips. Some of these are optimized for specific applications such as autonomous driving or surveillance cameras. Graphcore and a few others offer much more flexible chips, which are crucial for developing AI applications but also much more challenging to produce. The company's last investment round gave the

company a valuation of $1.7 billion.

Graphcore’s chips might first find traction with top AI experts who are able to write the code needed to exploit their benefits. Several prominent AI researchers have invested in Graphcore, including Demis Hassabis, cofounder of DeepMind, Zoubin Ghahramani, a professor at the University of Cambridge and the head of Uber’s AI lab, and Peiter Abbeel, a professor at UC Berkeley who specializes in AI and robotics. In an interview with WIREDlast December, AI visionary Geoffrey Hinton discussed the potential for Graphcore chips to advance fundamental research.

Before long, companies may be tempted to try out the latest thing, too. As Graphcore’s CEO Toon says, “Everybody’s trying to innovate, trying to find an advantage.”

This story originally appeared on wired.com.

Listing image by Graphcore

https://arstechnica.com/?p=1602809