Artist's rendition of Google's new data center in Hamina, Finland, which it is converting from a paper mill. Google is using machine learning to cut energy use in its data centers and other facilities. Google's Data Center Team

Chips Off the Old Block: Computers Are Taking Design Cues From Human Brains

New technologies are testing the limits of computer miniaturization. To cope with that, researchers have been looking for ideas from nature.

By Cade Metz

Nov. 18, 2021

SAN FRANCISCO — We expect a lot from our computers these days. They should talk to us, recognize everything from faces to spoken words and, in the near future, drive us from place to place. Computer engineers are enormous amounts of processing power, stretching the limits of even the most modern machines.

Now they're turning to a new way of thinking that takes a cue from biology as they respond to these growing demands. They are combining the very ideas of computers and are building machines that look more like the human brain, where a central brain store oversees the nervous system and offloads particular tasks — like hearing and seeing — to the surrounding cavity.

Such ideas are prompting the computer industry to engage, and this behind-the-scenes migration is a new kind of machine will have ramifications for everyone who uses a phone, drives a car, buys intelligent systems or searches, or the dream of machines that can perform the physical world by themselves can one day come true.

This migration could also diminish the power of Intel, the longtime giant of Silicon Valley, and a semiconductor industry that sits at the heart of all things tech. Intel has data centers that are humming with your iPhone to the virtual reality headsets and flying drones of tomorrow.

"This is an enormous change," said John Hennessy, the former president of Stanford University who wrote an influential book on computer design in the mid-1990s and is now a member of the Alphabet board of directors overseeing Google. "Everybody's looking at how they can put intelligence everywhere in the system."

Thanks for reading The Times.
Subscribe to The Times

The existing approach has had a pretty good run. For more than half a century, computer makers have built systems around a single, fast, fit-all chip — the central processing unit — from a company like Intel, one of the world's biggest semiconductor makers. That is what you'll find in the middle of your own laptop computer or smartphone.

Now, computer engineers are fashioning more complex systems, broken than functioning all tasks through one high-tech chip by Intel, newer machines are dividing work into tiny pieces and spreading it across a much vast farms of simpler, specialized chips that perform specific tasks.

Changes inside Google's giant data centers are a harbinger of what is to come for the rest of the industry. In the vast rooms of Google's servers, there is still a central processor. But enormous banks of custom-built chips work alongside them, running the computer algorithms that drive speech recognition and other forms of artificial intelligence.

Google reached this point out of necessity. For years, the company had operated the world's largest computer systems, spread across data centers and cables that stretched from California to Finland to Singapore. But for one Google researcher, it was much too small.

Xiaoyong Zhang, left, and Doug Berger of Microsoft are among the employees leading the company's efforts to design its own computer chips. Dan DeLong for The New York Times

In 2015, Jeff Dean, one of the company's most celebrated engineers, led a research team that explored the idea of neural networks, an essentially computer algorithm that can learn tasks on their own. They could spend hour of a number of days, for example, learning words spoken into smartphones or the faces in a photograph.

As a matter of months, the team was building a new service that could recognize spoken words for more accurately than Google's existing system. But there was a catch. If the network were than one billion phones that operated on Google's Android software used it every day for three minutes a day, the company would need to double the number of computers in its data centers, Mr. Dean said.

"We need another Google," Mr. Dean told Urs Hölzle, the Swiss-born computer scientist who oversaw the company's data center empire, according to someone who attended the meeting. Mr. Hölzle proposed an alternative: Google could build its own computer chip just for running this kind of artificial intelligence.

He and his team began this venture in secrecy, and they soon offered up what began as a dozen chips, or specialized computer processors, that would speed up the tasks of the tech landscape. Over the next five years, companies like Google, Apple and Samsung will build processors for a dozen or more A.I. chips. Microsoft is designing such a chip specifically for an artificial intelligence software. And everyone from Amazon to Intel is building autonomous cars that will need similar chips.

This trend toward specialized chips for a single computer architecture could lead to a "Cambrian explosion" of artificial intelligence, said Pat Gelsinger, who oversaw the creation of Intel's first processor for the Defense Department of Defense, and who now works on driverless cars at Pivotal. As he sees it, machines that spread computations across many specialized chips, he says, these chips can operate more like the human brain, which efficiently uses the energy at its disposal.

"In the brain, energy efficiency is the key," he said during a recent interview in the company's research center in Silicon Valley.

Change on the Horizon

There are many kinds of silicon chips. There are chips that store information. There are chips that perform basic tasks in toys and televisions. And there are chips that run various processes for computers and smartphones, including processors that are based on a widely used processor company known as ARM and smartphones.

For years, the central processing units, or C.P.U.s, that run PCs and similar devices were where the money was. And there had not been much need for change.

In accordance with Moore's Law, the oft-quoted maxim from Intel co-founder Gordon Moore, the number of transistors on a computer chip had doubled every two years or so, and that provided steadily improved performance for decades. As performance improved, chips consumed about the same amount of power, according to research from Robert Dennard, a fellow at the company who is now named for the legendary IBM researcher Robert Dennard.

By 2010, however, doubling the number of transistors was taking much longer than Moore's Law predicted. Dennard's scaling maxim had also been upended as chip designers ran into the limits of the physical materials they used to build processors. The result: If a company wanted more computing power, it could not just upgrade its processors. It needed more computers, more space and more electricity.

Jeff Dean, one of Google's most celebrated engineers, said the company should build its own computer chip specifically for artificial intelligence. Erin Kunkel for The New York Times

Because systems that rely on neural networks can learn largely on their own, they can evolve more quickly than traditional services. They are not as reliant on engineers writing software lines of code that explain how they should behave.

But there is a wrinkle. Training neural networks this way requires extensive trial and error. To create one that is able to recognize cats, for example, you must show it many different photos of cats, identifying the cats and improving the system every time it gets one wrong. At any given time, this process unfolds over hundreds of algorithms. That requires extensive computing power, and companies like Microsoft see standard-issue chips to do it, the same chips they use to help because the Internet runs smoothly, to be hard and too much electrical power is consumed.

So, the leading Internet companies have been running their neural networks with help from another type of chip called a graphics processing unit, or G.P.U., a piece of hardware that, initially, was built to — very specifically designed to create images for games and other software, and they worked hand-in-hand with the chip — usually made by Intel — at the center of a computer. G.P.U.s can process the math required by neural networks far more efficiently than C.P.U.s.

Nvidia is thriving as a result, and it is now selling large numbers of G.P.U.s to the internet giants in the United States and the biggest online companies around the world, in China most notably. The Chinese internet giant Baidu purchased from data center which is 1400 million over the last year.

"This is a little like being right there at the beginning of the Internet," Jensen Huang said in a recent interview in other words, the tech landscape is changing rapidly, and his Nvidia is at the heart of that change.

Creating Specialized Chips

G.P.U.s are the primary vehicles that companies use to teach their neural networks a particular task, but this is not the end of the process. Once a neural network is trained for a task, it must perform it, and that requires a different kind of computing power.

After training a speech-recognition algorithm, for example, Microsoft offers it up as an online service, and it actually starts identifying commands that people speak into their smartphones. G.P.U.s are too quick and too power-hungry for that purpose. So, many companies are now building chips specifically to do what the other chips have learned.

Google built its own specialty chip, a Tensor Processing Unit, or T.P.U., to speed up its services based on the neural networks it reorganizational specialized chips from Intel, which was acquired by Intel, so that it can run our neural networks more easily.

Other companies are following suit. Qualcomm, which specializes in chips for smartphones, and a number of start-ups are also working on A.I. chips, hoping to grab their piece of the rapidly expanding market. The tech research firm IDC predicts that revenue from servers equipped with alternative chips will reach $6.8 billion by 2024, about 10 percent of the overall server market.

Across Microsoft's global network of machines, Mr. Burger pointed out, alternative chips are still a relatively modest part of the equation. And that leaves room for more experimentation. The tech giant has developed software that distributes Google's neural networks and other computer tasks across the chips designed for Google's servers.

But then, for the production of engineering who leads hardware and software efforts for Google's neural networks, Google's data centers, Google is exploring systems commonly referred accounting the internet, driverless cars and other services that will demand even more speed and accuracy that is not possible now.

In other words, a driverless car needs cameras and radar and lasers. But it also needs a brain.

Subscribe for $1.00 a week. Ends soon.