The New York Times | https://nyti.ms/2y52kvb

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

New technologies are testing the limits of computer semiconductors. To deal with that, researchers have gone looking for ideas from nature.

By CADE METZ SEPT. 16, 2017

SAN FRANCISCO — We expect a lot from our computers these days. They should talk to us, recognize everything from faces to flowers, and maybe soon do the driving. All this artificial intelligence requires an enormous amount of computing power, stretching the limits of even the most modern machines.

Now, some of the world's largest tech companies are taking a cue from biology as they respond to these growing demands. They are rethinking the very nature of computers and are building machines that look more like the human brain, where a central brain stem oversees the nervous system and offloads particular tasks — like hearing and seeing — to the surrounding cortex.

After years of stagnation, the computer is evolving again, and this behind-the-scenes migration to a new kind of machine will have broad and lasting implications. It will allow work on artificially intelligent systems to accelerate, so the dream of machines that can navigate the physical world by themselves can one day come true.

This migration could also diminish the power of Intel, the longtime giant of chip design and manufacturing, and fundamentally remake the $335 billion a year semiconductor industry that sits at the heart of all things tech, from the data centers that drive the internet to your iPhone to the virtual reality headsets and flying drones of tomorrow.

“This is an enormous change,” said John Hennessy, the former Stanford University president who wrote an authoritative book on computer design in the mid-1990s and is now a member of the board at Alphabet, Google’s parent company. “The existing approach is out of steam, and people are trying to re-architect the system.”

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

Now, computer engineers are fashioning more complex systems. Rather than funneling all tasks through one beefy chip made by Intel, newer machines are dividing work into tiny pieces and spreading them among vast farms of simpler, specialized chips that consume less power.

Changes inside Google’s giant data centers are a harbinger of what is to come for the rest of the industry. Inside most 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 network — an empire of data centers and cables that stretched from California to Finland to Singapore. But for one Google researcher, it was much too small.

In 2011, Jeff Dean, one of the company’s most celebrated engineers, led a research team that explored the idea of neural networks — essentially computer

algorithms that can learn tasks on their own. They could be useful for a number of things, like recognizing the words spoken into smartphones or the faces in a photograph.

In a matter of months, Mr. Dean and his team built a service that could recognize spoken words far more accurately than Google's existing service. But there was a catch: If the world's more than one billion phones that operated on Google's Android software used the new service just three minutes a day, Mr. Dean realized, Google would have to double its data center capacity in order to support it.

"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. So Mr. Dean proposed an alternative: Google could build its own computer chip just for running this kind of artificial intelligence.

But what began inside data centers is starting to shift other parts of the tech landscape. Over the next few years, companies like Google, Apple and Samsung will build phones with specialized A.I. chips. Microsoft is designing such a chip specifically for an augmented-reality headset. And everyone from Google to Toyota is building autonomous cars that will need similar chips.

This trend toward specialty chips and a new computer architecture could lead to a "Cambrian explosion" of artificial intelligence, said Gill Pratt, who was a program manager at Darpa, a research arm of the United States Department of Defense, and now works on driverless cars at Toyota. As he sees it, machines that spread computations across vast numbers of tiny, low-power 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 at Toyota's new 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, from the supercomputers used to create models for global warming to personal computers, internet servers and smartphones.

For years, the central processing units, or C.P.U.s, that ran 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 another, lesser-known law of chip design called Dennard scaling, named for the longtime 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.

Researchers in industry and academia were working to extend Moore's Law, exploring entirely new chip materials and design techniques. But Doug Burger, a researcher at Microsoft, had another idea: Rather than rely on the steady evolution of the central processor, as the industry had been doing since the 1960s, why not move some of the load onto specialized chips?

During his Christmas vacation in 2010, Mr. Burger, working with a few other chip researchers inside Microsoft, began exploring new hardware that could accelerate the performance of Bing, the company's internet search engine.

At the time, Microsoft was just beginning to improve Bing using machine-learning algorithms (neural networks are a type of machine learning) that could

improve search results by analyzing the way people used the service. Though these algorithms were less demanding than the neural networks that would later remake the internet, existing chips had trouble keeping up.

Mr. Burger and his team explored several options but eventually settled on something called Field Programmable Gate Arrays, or F.P.G.A.s.: chips that could be reprogrammed for new jobs on the fly. Microsoft builds software, like Windows, that runs on an Intel C.P.U. But such software cannot reprogram the chip, since it is hard-wired to perform only certain tasks.

With an F.P.G.A., Microsoft could change the way the chip works. It could program the chip to be really good at executing particular machine learning algorithms. Then, it could reprogram the chip to be really good at running logic that sends the millions and millions of data packets across its computer network. It was the same chip but it behaved in a different way.

Microsoft started to install the chips en masse in 2015. Now, just about every new server loaded into a Microsoft data center includes one of these programmable chips. They help choose the results when you search Bing, and they help Azure, Microsoft's cloud-computing service, shuttle information across its network of underlying machines.

Teaching Computers to Listen

In fall 2016, another team of Microsoft researchers — mirroring the work done by Jeff Dean at Google — built a neural network that could, by one measure at least, recognize spoken words more accurately than the average human could.

Xuedong Huang, a speech-recognition specialist who was born in China, led the effort, and shortly after the team published a paper describing its work, he had dinner in the hills above Palo Alto, Calif., with his old friend Jen-Hsun Huang, (no relation), the chief executive of the chipmaker Nvidia. The men had reason to celebrate, and they toasted with a bottle of champagne.

Xuedong Huang and his fellow Microsoft researchers had trained their speech-recognition service using large numbers of specialty chips supplied by Nvidia, rather than relying heavily on ordinary Intel chips. Their breakthrough would not have been possible had they not made that change.

"We closed the gap with humans in about a year," Microsoft's Mr. Huang said. "If we didn't have the weapon — the infrastructure — it would have taken at least five years."

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 endless 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 words as well as a human can, researchers must train it repeatedly, tweaking the algorithms and improving the training data over and over. At any given time, this process unfolds over hundreds of algorithms. That requires enormous computing power, and if companies like Microsoft use standard-issue chips to do it, the process takes far too long because the chips cannot handle the load and too much electrical power is consumed.

So, the leading internet companies are now training their neural networks with help from another type of chip called a graphics processing unit, or G.P.U. These low-power chips — usually made by Nvidia — were originally designed to render 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 of the United States and the biggest online companies around the world, in China most notably. The company's quarterly revenue from data center sales tripled to $409 million over the past year.

"This is a little like being right there at the beginning of the internet," Jen-

Hsun Huang said in a recent interview. In other words, the tech landscape is changing rapidly, and 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 that is only part 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 not quite as efficient during this stage of the process. 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. Nvidia is building a similar chip. And Microsoft has reprogrammed specialized chips from Altera, which was acquired by Intel, so that it too can run 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 2021, 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 operation. And Bart Sano, the vice president of engineering who leads hardware and software development for Google's network, said much the same about the chips deployed at its data centers.

Mike Mayberry, who leads Intel Labs, played down the shift toward alternative

processors, perhaps because Intel controls more than 90 percent of the data-center market, making it by far the largest seller of traditional chips. He said that if central processors were modified the right way, they could handle new tasks without added help.

But this new breed of silicon is spreading rapidly, and Intel is increasingly a company in conflict with itself. It is in some ways denying that the market is changing, but nonetheless shifting its business to keep up with the change.

Two years ago, Intel spent $16.7 billion to acquire Altera, which builds the programmable chips that Microsoft uses. It was Intel's largest acquisition ever. Last year, the company paid a reported $408 million buying Nervana, a company that was exploring a chip just for executing neural networks. Now, led by the Nervana team, Intel is developing a dedicated chip for training and executing neural networks.

"They have the traditional big-company problem," said Bill Coughran, a partner at the Silicon Valley venture capital firm Sequoia Capital who spent nearly a decade helping to oversee Google's online infrastructure, referring to Intel. "They need to figure out how to move into the new and growing areas without damaging their traditional business."

Intel's internal conflict is most apparent when company officials discuss the decline of Moore's Law. During a recent interview with The New York Times, Naveen Rao, the Nervana founder and now an Intel executive, said Intel could squeeze "a few more years" out of Moore's Law. Officially, the company's position is that improvements in traditional chips will continue well into the next decade.

Mr. Mayberry of Intel also argued that the use of additional chips was not new. In the past, he said, computer makers used separate chips for tasks like processing audio.

But now the scope of the trend is significantly larger. And it is changing the market in new ways. Intel is competing not only with chipmakers like Nvidia and Qualcomm, but also with companies like Google and Microsoft.

Google is designing the second generation of its T.P.U. chips. Later this year, the company said, any business or developer that is a customer of its cloud-computing service will be able to use the new chips to run its software.

While this shift is happening mostly inside the massive data centers that underpin the internet, it is probably a matter of time before it permeates the broader industry.

The hope is that this new breed of mobile chip can help devices handle more, and more complex, tasks on their own, without calling back to distant data centers: phones recognizing spoken commands without accessing the internet; driverless cars recognizing the world around them with a 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.

Follow Cade Metz on Twitter: @CadeMetz

A version of this article appears in print on September 17, 2017, on Page BU1 of the New York edition with the headline: Chip Off the Old Block.
