How AI Can Keep Accelerating After Moore’s Law

New ideas in chip design look likely to keep software getting smarter.

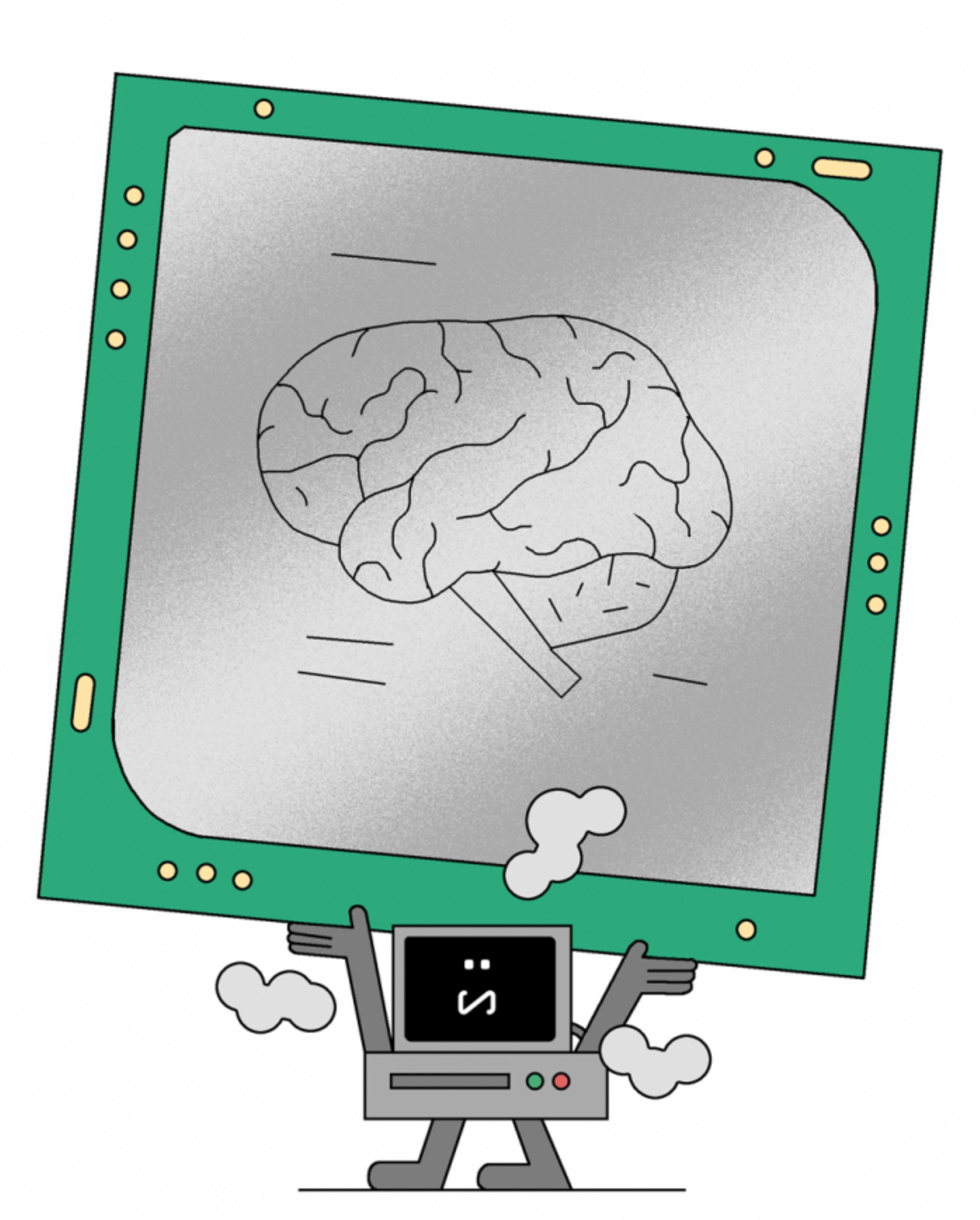
Tom Simonite

Google CEO Sundar Pichai was obviously excited when he spoke to developers about a blockbuster result from his machine-learning lab earlier this month. Researchers had figured out how to automate some of the work of crafting machine-learning software, something that could make it much easier to deploy the technology in new situations and industries.

But the project had already gained a reputation among AI researchers for another reason: the way it illustrated the vast computing resources needed to compete at the cutting edge of machine learning.

A paper from Google’s researchers says they simultaneously used as many as 800 of the powerful and expensive graphics processors that have been crucial to the recent uptick in the power of machine learning (see “10 Breakthrough Technologies 2013: Deep Learning”). They told *MIT Technology Review* that the project had tied up hundreds of the chips for two weeks solid—making the technique too resource-intensive to be more than a research project even at Google.

A coder without ready access to a giant collection of GPUs would need deep pockets to replicate the experiment. Renting 800 GPUs from Amazon’s cloud computing service for just a week would cost around $120,000 at the listed prices.



Andrea Chronopoulos

Feeding data into deep learning software to train it for a particular task is much more resource intensive than running the system afterwards, but that still takes significant oomph. "Computing power is a bottleneck right now for machine learning," says Reza Zadeh, an adjunct professor at Stanford University and founder and CEO of Matroid, a startup that helps companies use software to identify objects like cars and people in security footage and other video.

The sudden thirst for new power to drive AI comes at a time when the computing industry is adjusting to the loss of two things it has relied on for 50 years to keep chips getting more powerful. One is Moore's Law, which forecast that the number of transistors that could be fitted into a given area of a chip would double every two years. The other is a phenomenon called Dennard scaling, which describes how the amount of power that transistors use scales down as they shrink.

Neither holds true today. Intel has slowed the pace at which it introduces generations of new chips with smaller, denser transistors (see "Moore's Law Is Dead. Now What?"). And the usual efficiency gains that transistors showed as they got smaller came to a halt in the mid-2000s, making power consumption a major headache.

The good news for those betting on AI is that graphics chips have so far managed to defy gravity. At the recent conference of leading graphics chipmaker Nvidia, CEO Jensen Huang displayed a chart showing how his chips' performance has continued to accelerate exponentially while growth in the performance of general purpose processors, or CPUs, has slowed.

Doug Burger, a distinguished engineer at Microsoft's NExT division that works on commercializing new technology, says a similar gap is opening between conventional and machine-learning software. "You're starting to see a [performance] plateau for general software—it has stopped improving at historical rates—but this AI stuff is still increasing rapidly," he says.

Burger thinks that trend will continue. Engineers have kept GPUs getting more powerful because they can be more specialized to the particular math they need to perform for graphics or machine learning, he says.

The same idea is behind a project Burger led at Microsoft, which is putting more power behind AI software by using reconfigurable chips called FPGAs. It also motivates the startups—and giants such as Google—creating new chips customized to power machine learning (see "Google Reveals a Powerful New AI Chip and Supercomputer").

In the longer term, more radical changes in how computer chips work will be required to keep AI getting more powerful. Creating chips that don't add accurately is one option. Prototypes have shown that they can make computers more efficient without undermining the accuracy of results from machine-learning software (see "Why a Chip That's Bad at Math Could Help Computers Tackle Harder Problems").

Chip designs that directly copy from biology could also be crucial. IBM and others have built prototype chips that compute using spikes of current, similar to how our neurons fire (see "Thinking in Silicon"). Even simple animals, Burger points out, use little energy to do things beyond what today's robots and software can accomplish—evidence that computers have much further to go.

"Look at the computation a cockroach does," he says. "There are existence proofs that show many more orders of magnitude of performance and efficiency are available. We can have decades of scaling left in AI."