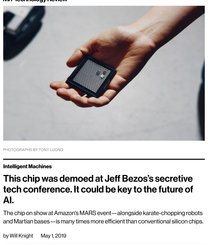


PHOTOGRAPHS BY TONY LUONG

Intelligent Machines

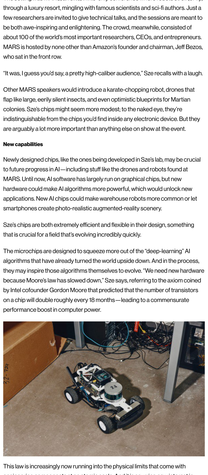
This chip was demoed at Jeff Bezos's secretive tech conference. It could be key to the future of AI.

The chip on show at Amazon's MARS event—alongside karate-chopping robots and Martian bases—is many times more efficient than conventional silicon chips.

by Will Knight May 1, 2019

Recently, on a dazzling morning in Palm Springs, California, Vivienne Sze took to a small stage to deliver perhaps the most nerve-racking presentation of her career.

She knew the subject matter inside-out. She was to tell the audience about the chips being developed in her lab at MIT, that promise to bring powerful artificial intelligence to a multitude of devices where power is limited, beyond the reach of the vast data centers where most AI computations take place. However, the event—and the audience—gave Sze pause.



The setting was MARS, an elite, invite-only conference where robots stroll (or fly) through a luxury resort, mingling with famous scientists and sci-fi authors. Just a few researchers are invited to give technical talks, and the sessions are meant to be both awe-inspiring and enlightening. The crowd, meanwhile, consisted of about 100 of the world's most important researchers, CEOs, and entrepreneurs. MARS is hosted by none other than Amazon's founder and chairman, Jeff Bezos, who sat in the front row.

"It was, I guess you'd say, a pretty high-caliber audience," Sze recalls with a laugh.

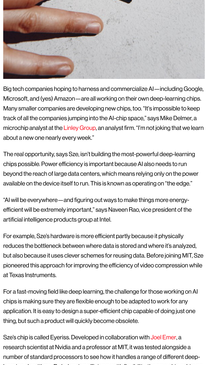
Other MARS speakers would introduce a karate-chopping robot, drones that flap like large, eerily silent insects, and even optimistic blueprints for Martian colonies. Sze's chips might seem more modest, to the naked eye, they're indistinguishable from the chips you'd find inside any electronic device. But they are arguably a lot more important than anything else on show at the event.

New capabilities

Newly designed chips, like the ones being developed in Sze's lab, may be crucial to future progress in AI—including stuff like the drones and robots found at MARS. Until now, AI software has largely run on graphical chips, but new hardware could make AI algorithms more powerful, which would unlock new applications. New AI chips could make warehouse robots more common or let smartphones create photo-realistic augmented-reality scenery.

Sze's chips are both extremely efficient and flexible in their design, something that is crucial for a field that's evolving incredibly quickly.

The microchips are designed to squeeze more out of the "deep-learning" AI algorithms that have already turned the world upside down. And in the process, they may inspire those algorithms themselves to evolve. "We need new hardware because Moore's law has slowed down," Sze says, referring to the axiom coined by Intel cofounder Gordon Moore that predicted that the number of transistors on a chip will double roughly every 18 months—leading to a commensurate performance boost in computer power.



This law is increasingly now running into the physical limits that come with engineering components at an atomic scale. And it is spurring new interest in alternative architectures and approaches to computing.

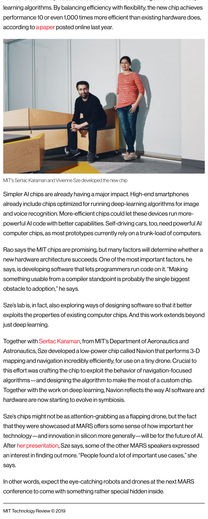
The high stakes attached to investing in next-generation AI chips—and maintaining America's dominance in chipmaking overall—aren't lost on the US government. Sze's microchips are being developed with funding from a Defense Advanced Research Projects Agency (DARPA) program meant to help develop new AI chip designs (see "The out-there AI ideas designed to keep the US ahead of China").

But innovation in chipmaking has been spurred mostly by the emergence of deep learning, a very powerful way for machines to learn to perform useful tasks. Instead of giving a computer a set of rules to follow, a machine basically programs itself. Training data is fed into a large, simulated artificial neural network, which is then tweaked so that it produces the desired result. With enough training, a deep-learning system can find subtle and abstract patterns in data. The technique is applied to an ever-growing array of practical tasks, from face recognition on smartphones to predicting disease from medical images.

The new chip race

Deep learning is not so reliant on Moore's law. Neural nets run many mathematical computations in parallel, so they run far more effectively on the specialized video-game graphics chips that perform parallel computations for rendering 3-D imagery. But microchips designed specifically for the computations that underpin deep learning should be even more powerful.

The potential for new chip architectures to improve AI has stirred up a level of entrepreneurial activity that the chip industry hasn't seen in decades (see "The Race to Power AI's Silicon Brains" and "China has never had a real chip industry. Making AI chips could change that").



Big tech companies hoping to harness and commercialize AI—including Google, Microsoft, and (yes) Amazon—are all working on their own deep-learning chips. Many smaller companies are developing new chips, too. "It's impossible to keep track of all the companies jumping into the AI-chip space," says Mike Demler, a microchip analyst at the Linley Group, an analyst firm. "I'm not joking that we learn about a new one nearly every week."

The real opportunity, says Sze, isn't building the most-powerful deep-learning chips possible. Power efficiency is important because AI also needs to run beyond the reach of large data centers, which means relying only on the power available on the device itself to run. This is known as operating on "the edge."

"AI will be everywhere—and figuring out ways to make things more energy-efficient will be extremely important," says Naveen Rao, vice president of the artificial intelligence products group at Intel.

For example, Sze's hardware is more efficient partly because it physically reduces the bottlenecks between where data is stored and where it's analyzed, but also because it uses clever schemes for reusing data. Before joining MIT, Sze pioneered this approach for improving the efficiency of video compression while at Texas Instruments.

For a fast-moving field like deep learning, the challenge for those working on AI chips is making sure they are flexible enough to be adapted to work for any application. It is easy to design a super-efficient chip capable of doing just one thing, but such a product will quickly become obsolete.

Sze's chip is called Eyeriss. Developed in collaboration with Joel Emer, a research scientist at Nvidia and a professor at MIT, it was tested alongside a number of standard processors to see how it handles a range of different deep-learning algorithms. By balancing efficiency with flexibility, the new chip achieves performance 10 or even 1,000 times more efficient than existing hardware does, according to a paper posted online last year.

MIT's Sertac Karaman and Vivienne Sze developed the new chip.

Simpler AI chips are already having a major impact. High-end smartphones already include chips optimized for running deep-learning algorithms for image and voice recognition. More-efficient chips could let these devices run more-powerful AI code with better capabilities. Self-driving cars, too, need powerful AI computer chips, as most prototypes currently rely on a trunk-load of computers.

Rao says the MIT chips are promising, but many factors will determine whether a new hardware architecture succeeds. One of the most important factors, he says, is developing software that lets programmers run code on it. "Making something usable from a compiler standpoint is probably the single biggest obstacle to adoption," he says.

Sze's lab is, in fact, also exploring ways of designing software so that it better exploits the properties of existing computer chips. And this work extends beyond just deep learning.

Together with Sertac Karaman, from MIT's Department of Aeronautics and Astronautics, Sze developed a low-power chip called Navion that performs 3-D mapping and navigation incredibly efficiently, for use on a tiny drone. Crucial to this effort was crafting the chip to exploit the behavior of navigation-focused algorithms—and designing the algorithm to make the most of a custom chip. Together with the work on deep learning, Navion reflects the way AI software and hardware are now starting to evolve in symbiosis.

Sze's chips might not be as attention-grabbing as a flapping drone, but the fact that they were showcased at MARS offers some sense of how important her technology—and innovation in silicon more generally—will be for the future of AI. After her presentation, Sze says, some of the other MARS speakers expressed an interest in finding out more. "People found a lot of important use cases," she says.

In other words, expect the eye-catching robots and drones at the next MARS conference to come with something rather special hidden inside.