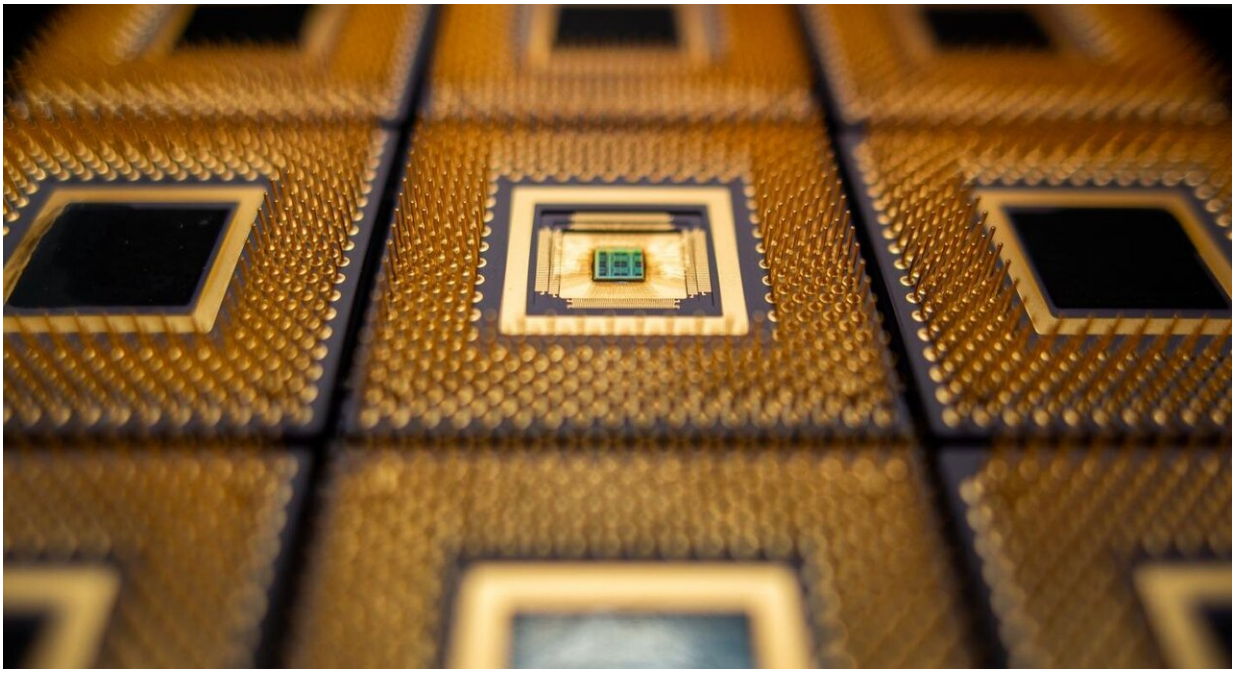
# Rethinking microchips' design pushes computing to the edge

February 24 2021, by John Sullivan



Princeton researchers have created a new chip that speeds artificial intelligence systems called neural nets while slashing power use. The chips could help bring advanced applications to remote devices such as cars and smartphones. Credit: Hongyang Jia/Princeton University

Responding to artificial intelligence's exploding demands on computer networks, Princeton University researchers in recent years have radically increased the speed and slashed the energy use of specialized AI

systems. Now, the researchers have moved their innovation closer to widespread use by creating co-designed hardware and software that will allow designers to blend these new types of systems into their applications.

"Software is a critical part of enabling new hardware," said Naveen Verma, a professor of electrical and computer engineering at Princeton and a leader of the research team. "The hope is that designers can keep using the same software system—and just have it work ten times faster or more efficiently."

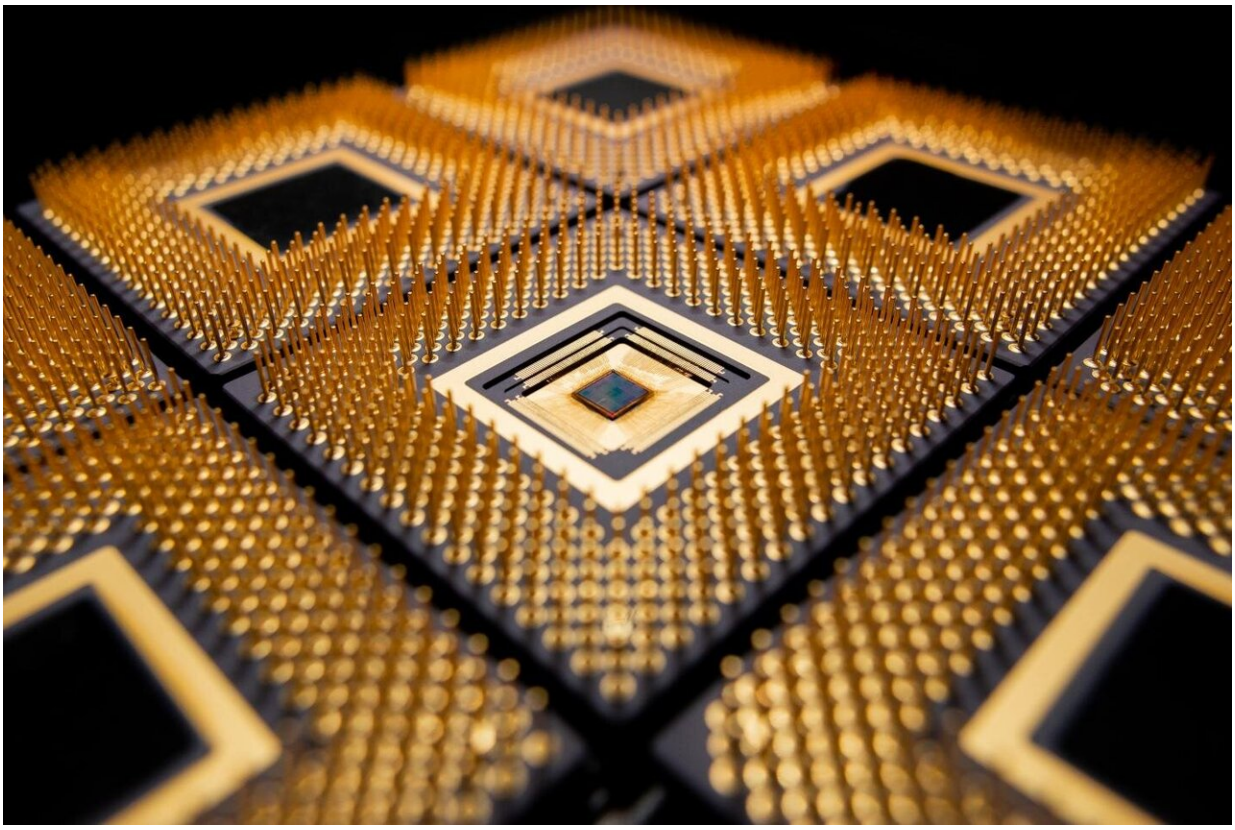
By cutting both power demand and the need to exchange data from remote servers, systems made with the Princeton technology will be able to bring artificial intelligence applications, such as piloting software for drones or advanced language translators, to the very edge of computing infrastructure.

"To make AI accessible to the real-time and often personal process all around us, we need to address latency and privacy by moving the computation itself to the edge," said Verma, who is the director of the University's Keller Center for Innovation in Engineering Education. "And that requires both energy efficiency and performance."

Two years ago, the Princeton research team fabricated a new chip designed to improve the performance of neural networks, which are the essence behind today's artificial intelligence. The chip, which performed tens to hundreds of times better than other advanced microchips, marked a revolutionary approach in several measures. In fact, the chip was so different than anything being used for neural nets that it posed a challenge for the developers.

"The chip's major drawback is that it uses a very unusual and disruptive architecture," Verma said in a 2018 interview. "That needs to be

reconciled with the massive amount of infrastructure and design methodology that we have and use today."



The new chip is based on analog computing, which uses circuits to mimic equations being solved rather than generate 1s and 0s like a digital computer. Credit: Hongyang Jia/Princeton University

Over the next two years, the researchers worked to refine the chip and to create a software system that would allow artificial intelligence systems to take advantage of the new chip's speed and efficiency. In a presentation to the International Solid-State Circuits Virtual Conference on Feb. 22, lead author Hongyang Jia, a graduate student in Verma's research lab, described how the new software would allow the new chips

to work with different types of networks and allow the systems to be scalable both in hardware and execution of software.

"It is programmable across all these networks," Verma said. "The networks can be very big, and they can be very small."

Verma's team developed the new chip in response to growing demand for artificial intelligence and to the burden AI places on computer networks. Artificial intelligence, which allows machines to mimic cognitive functions such as learning and judgment, plays a critical role in new technologies such as image recognition, translation, and self-driving vehicles. Ideally, the computation for technology such as drone navigation would be based on the drone itself, rather than in a remote network computer. But digital microchips' power demand and need for memory storage can make designing such a system difficult. Typically, the solution places much of the computation and memory on a remote server, which communicates wirelessly with the drone. But this adds to the demands on the communications system, and it introduces security problems and delays in sending instructions to the drone.

To approach the problem, the Princeton researchers rethought computing in several ways. First, they designed a chip that conducts computation and stores data in the same place. This technique, called in-memory computing, slashes the energy and time used to exchange information with dedicated memory. The technique boosts efficiency, but it introduces new problems: because it crams the two functions into a small area, in-memory computing relies on analog operation, which is sensitive to corruption by sources such as voltage fluctuation and temperature spikes. To solve this problem, the Princeton team designed their chips using capacitors rather than transistors. The capacitors, devices that store an electrical charge, can be manufactured with greater precision and are not highly affected by shifts in voltage. Capacitors can also be very small and placed on top of memory cells, increasing

processing density and cutting energy needs.

But even after making analog operation robust, many challenges remained. The analog core needed to be efficiently integrated in a mostly digital architecture, so that it could be combined with the other functions and software needed to actually make practical systems work. A digital system uses off-and-on switches to represent ones and zeros that computer engineers use to write the algorithms that make up computer programming. An analog computer takes a completely different approach. In an article in the IEEE Spectrum, Columbia University Professor Yannis Tsividis described an analog computer as a physical system designed to be governed by equations identical to those the programmer wants to solve. An abacus, for example, is a very simple analog computer. Tsividis says that a bucket and a hose can serve as an analog computer for certain calculus problems: to solve an integration function, you could do the math, or you could just measure the water in the bucket.

Analog computing was the dominant technology through the Second World War. It was used to perform functions from predicting tides to directing naval guns. But analog systems were cumbersome to build and usually required highly trained operators. After the emergency of the transistor, digital systems proved more efficient and adaptable. But new technologies and new circuit designs have allowed engineers to eliminate many shortcomings of the analog systems. For applications such as neural networks, the analog systems offer real advantages. Now, the question is how to combine the best of both worlds.Verma points out that the two types of systems are complimentary. Digital systems play a central role while neural networks using analog chips can run specialized operations extremely fast and efficiently. That is why developing a software system that can integrate the two technologies seamlessly and efficiently is such a critical step.

"The idea is not to put the entire network into in-memory computing," he said. "You need to integrate the capability to do all the other stuff and to do it in a programmable way."

Provided by Princeton University