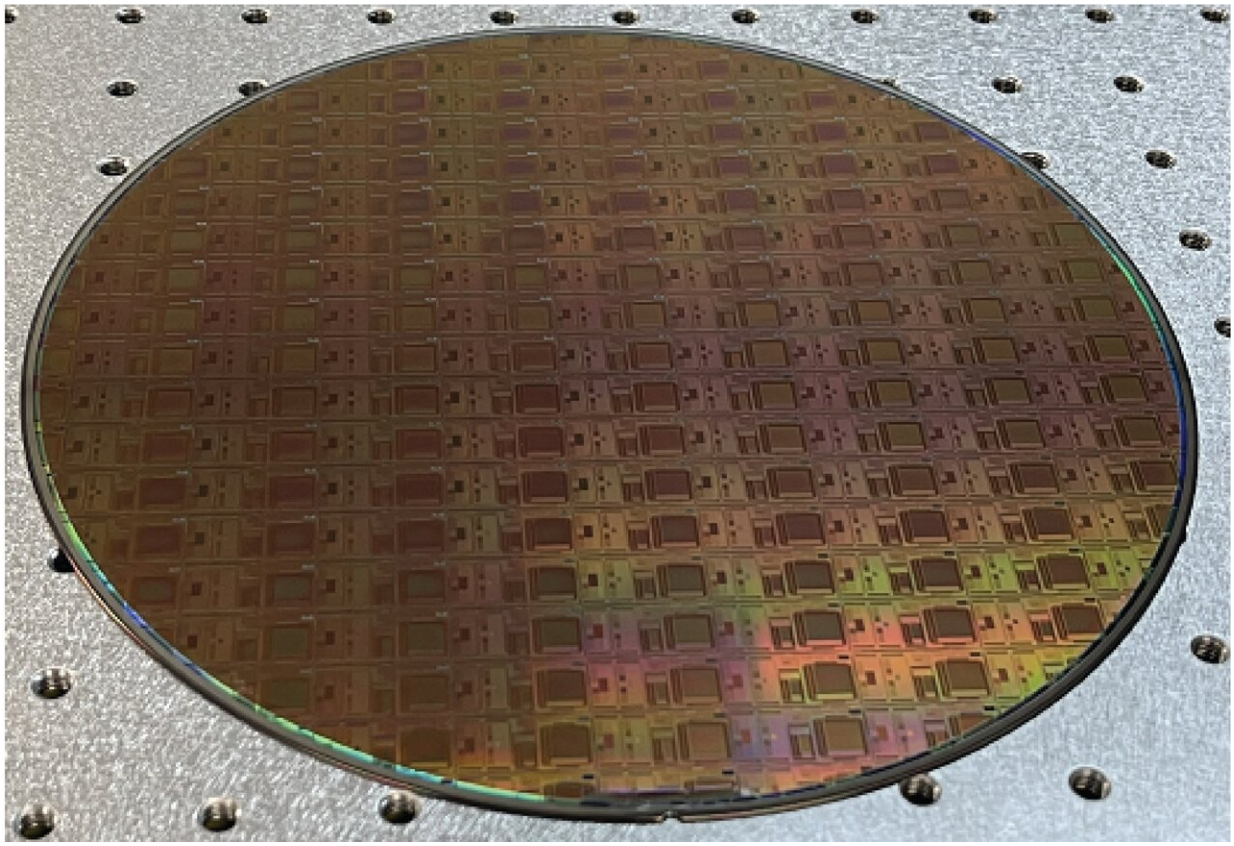
New chip design to provide greatest precision in memory to date

March 29 2023



New chip design to provide greatest precision in memory to date, will enable powerful AI in your portable devices. Credit: Joshua Yang of USC and TetraMem

Everyone is talking about the newest AI and the power of neural

networks, forgetting that software is limited by the hardware on which it runs. But it is hardware, says USC Professor of Electrical and Computer Engineering Joshua Yang, that has become "the bottleneck." Now, Yang's new research with collaborators might change that. They believe that they have developed a new type of chip with the best memory of any chip thus far for edge AI (AI in portable devices).

For approximately the past 30 years, while the size of the neural networks needed for AI and data science applications doubled every 3.5 months, the hardware capability needed to process them doubled only every 3.5 years. According to Yang, hardware presents a more and more severe problem for which few have patience.

Governments, industry, and academia are trying to address this hardware challenge worldwide. Some continue to work on hardware solutions with silicon chips, while others are experimenting with new types of materials and devices. Yang's work falls into the middle—focusing on exploiting and combining the advantages of the new materials and traditional silicon technology that could support heavy AI and data science computation.

The researchers' new paper in *Nature* focuses on the understanding of fundamental physics that leads to a drastic increase in memory capacity needed for AI hardware. The team led by Yang, with researchers from USC (including Han Wang's group), MIT, and the University of Massachusetts, developed a protocol for devices to reduce "noise" and demonstrated the practicality of using this protocol in integrated chips. This demonstration was made at TetraMem, a startup company co-founded by Yang and his co-authors (Miao Hu, Qiangfei Xia, and Glenn Ge), to commercialize AI acceleration technology.

According to Yang, this new memory chip has the highest information density per device (11 bits) among all types of known memory

technologies thus far. Such small but powerful devices could play a critical role in bringing incredible power to the devices in our pockets. The chips are not just for memory but also for the processor. Millions of them in a small chip, working in parallel to rapidly run your AI tasks, could only require a small battery to power it.

The chips that Yang and his colleagues are creating combine silicon with metal oxide memristors in order to create powerful but low-energy intensive chips. The technique focuses on using the positions of atoms to represent information rather than the number of electrons (which is the current technique involved in computations on chips). The positions of the atoms offer a compact and stable way to store more information in an analog, instead of digital fashion. Moreover, the information can be processed where it is stored instead of being sent to one of the few dedicated "processors," eliminating the so-called 'von Neumann bottleneck' existing in current computing systems. In this way, says Yang, computing for AI is "more energy-efficient with a higher throughput."

How it works

Yang explains that electrons that are manipulated in traditional chips are "light." This lightness makes them prone to moving around and being more volatile. Instead of storing memory through electrons, Yang and collaborators are storing memory in full atoms. Here is why this memory matters. Normally, says Yang, when one turns off a computer, the information memory is gone—but if you need that memory to run a new computation and your computer needs the information all over again, you have lost both time and energy.

This new method, focusing on activating atoms rather than electrons, does not require battery power to maintain stored information. Similar scenarios happen in AI computations, where a stable memory capable of

high information density is crucial. Yang imagines this new tech that may enable powerful AI capability in edge devices, such as Google Glasses, which he says previously suffered from a frequent recharging issue.

Further, by converting chips to rely on atoms as opposed to electrons, chips become smaller. Yang adds that with this new method, there is more computing capacity at a smaller scale. Moreover, this method, he says, could offer "many more levels of memory to help increase information density."

To put it in context, right now, ChatGPT is running on a cloud. The new innovation, followed by some further development, could put the power of a mini version of ChatGPT in everyone's personal device. It could make such high-powered tech more affordable and accessible for all sorts of applications.

More information: Mingyi Rao et al, Thousands of conductance levels in memristors integrated on CMOS, *Nature* (2023). DOI: 10.1038/s41586-023-05759-5

Provided by University of Southern California