

New neuromorphic chip for AI on the edge, at a small fraction of the energy and size of today's computing platforms

August 17 2022



A team of international researchers designed, manufactured and tested the NeuRRAM chip. Credit: David Baillot/University of California San Diego

An international team of researchers has designed and built a chip that



runs computations directly in memory and can run a wide variety of AI applications–all at a fraction of the energy consumed by computing platforms for general-purpose AI computing.

The NeuRRAM neuromorphic chip brings AI a step closer to running on a broad range of edge devices, disconnected from the cloud, where they can perform sophisticated cognitive tasks anywhere and anytime without relying on a network connection to a centralized server. Applications abound in every corner of the world and every facet of our lives, and range from smart watches, to VR headsets, smart earbuds, smart sensors in factories and rovers for space exploration.

The NeuRRAM chip is not only twice as energy efficient as the state-ofthe-art "compute-in-memory" chips, an innovative class of hybrid chips that runs computations in memory, it also delivers results that are just as accurate as conventional digital chips. Conventional AI platforms are a lot bulkier and typically are constrained to using large data servers operating in the cloud.

In addition, the NeuRRAM chip is highly versatile and supports many different <u>neural network</u> models and architectures. As a result, the chip can be used for many different applications, including <u>image recognition</u> and reconstruction as well as <u>voice recognition</u>.

"The conventional wisdom is that the higher efficiency of compute-inmemory is at the cost of versatility, but our NeuRRAM chip obtains efficiency while not sacrificing versatility," said Weier Wan, the paper's first corresponding author and a recent Ph.D. graduate of Stanford University who worked on the chip while at UC San Diego, where he was co-advised by Gert Cauwenberghs in the Department of Bioengineering.

The research team, co-led by bioengineers at the University of



California San Diego, presents their results in the Aug. 17 issue of *Nature*.

Currently, AI computing is both power hungry and computationally expensive. Most AI applications on edge devices involve moving data from the devices to the cloud, where the AI processes and analyzes it. Then the results are moved back to the device. That's because most edge devices are battery-powered and as a result only have a limited amount of power that can be dedicated to computing.

By reducing power consumption needed for AI inference at the edge, this NeuRRAM chip could lead to more robust, smarter and accessible edge devices and smarter manufacturing. It could also lead to better data privacy as the transfer of data from devices to the cloud comes with increased security risks.

On AI chips, moving data from memory to computing units is one major bottleneck.

"It's the equivalent of doing an eight-hour commute for a two-hour work day," Wan said.

To solve this data transfer issue, researchers used what is known as resistive random-access memory, a type of non-volatile memory that allows for computation directly within memory rather than in separate computing units. RRAM and other emerging memory technologies used as synapse arrays for neuromorphic computing were pioneered in the lab of Philip Wong, Wan's advisor at Stanford and a main contributor to this work. Computation with RRAM chips is not necessarily new, but generally it leads to a decrease in the accuracy of the computations performed on the chip and a lack of flexibility in the chip's architecture.

"Compute-in-memory has been common practice in neuromorphic



engineering since it was introduced more than 30 years ago," Cauwenberghs said. "What is new with NeuRRAM is that the extreme efficiency now goes together with great flexibility for diverse AI applications with almost no loss in accuracy over standard digital generalpurpose compute platforms."

A carefully crafted methodology was key to the work with multiple levels of "co-optimization" across the abstraction layers of hardware and software, from the design of the chip to its configuration to run various AI tasks. In addition, the team made sure to account for various constraints that span from memory device physics to circuits and network architecture.

"This chip now provides us with a platform to address these problems across the stack from devices and circuits to algorithms," said Siddharth Joshi, an assistant professor of computer science and engineering at the University of Notre Dame, who started working on the project as a Ph.D. student and postdoctoral researcher in Cauwenberghs lab at UC San Diego.





A close up of the NeuRRAM chip. Credit: David Baillot/University of California San Diego

Chip performance

Researchers measured the chip's energy efficiency by a measure known as energy-delay product, or EDP. EDP combines both the amount of energy consumed for every operation and the amount of times it takes to complete the operation. By this measure, the NeuRRAM chip achieves 1.6 to 2.3 times lower EDP (lower is better) and 7 to 13 times higher computational density than state-of-the-art chips.

Researchers ran various AI tasks on the chip. It achieved 99% accuracy on a handwritten digit recognition task; 85.7% on an image classification



task; and 84.7% on a Google speech command recognition task. In addition, the chip also achieved a 70% reduction in image-reconstruction error on an image-recovery task. These results are comparable to existing digital chips that perform computation under the same bit-precision, but with drastic savings in energy.

Researchers point out that one key contribution of the paper is that all the results featured are obtained directly on the hardware. In many previous works of compute-in-memory chips, AI benchmark results were often obtained partially by software simulation.

Next steps include improving architectures and circuits and scaling the design to more advanced technology nodes. Researchers also plan to tackle other applications, such as spiking neural networks.

"We can do better at the device level, improve circuit design to implement additional features and address diverse applications with our dynamic NeuRRAM platform," said Rajkumar Kubendran, an assistant professor for the University of Pittsburgh, who started work on the project while a Ph.D. student in Cauwenberghs' research group at UC San Diego.

In addition, Wan is a founding member of a startup that works on productizing the compute-in-memory technology. "As a researcher and an engineer, my ambition is to bring research innovations from labs into practical use," Wan said.

New architecture

The key to NeuRRAM's energy efficiency is an innovative method to sense output in memory. Conventional approaches use voltage as input and measure current as the result. But this leads to the need for more complex and more power hungry circuits. In NeuRRAM, the team



engineered a neuron circuit that senses voltage and performs analog-todigital conversion in an energy efficient manner. This voltage-mode sensing can activate all the rows and all the columns of an RRAM array in a single computing cycle, allowing higher parallelism.

In the NeuRRAM architecture, CMOS neuron circuits are physically interleaved with RRAM weights. It differs from conventional designs where CMOS circuits are typically on the peripheral of RRAM weights. The neuron's connections with the RRAM array can be configured to serve as either input or output of the neuron. This allows neural network inference in various data flow directions without incurring overheads in area or power consumption. This in turn makes the architecture easier to reconfigure.

To make sure that accuracy of the AI computations can be preserved across various neural network architectures, researchers developed a set of hardware algorithm co-optimization techniques. The techniques were verified on various neural networks including convolutional neural networks, long short-term memory, and restricted Boltzmann machines.

As a neuromorphic AI chip, NeuroRRAM performs parallel distributed processing across 48 neurosynaptic cores. To simultaneously achieve high versatility and high efficiency, NeuRRAM supports dataparallelism by mapping a layer in the neural network model onto multiple cores for parallel inference on multiple data. Also, NeuRRAM offers model-parallelism by mapping different layers of a model onto different cores and performing inference in a pipelined fashion.





The NeuRRAM chip uses an innovative architecture that has been co-optimized across the stack. Credit: David Baillot/University of California San Diego

An international research team

The work is the result of an international team of researchers.

The UC San Diego team designed the CMOS circuits that implement the neural functions interfacing with the RRAM arrays to support the synaptic functions in the chip's architecture, for high efficiency and versatility. Wan, working closely with the entire team, implemented the design; characterized the chip; trained the AI models; and executed the experiments. Wan also developed a software toolchain that maps AI applications onto the chip.



The RRAM synapse array and its operating conditions were extensively characterized and optimized at Stanford University.

The RRAM array was fabricated and integrated onto CMOS at Tsinghua University.

The Team at Notre Dame contributed to both the design and architecture of the chip and the subsequent machine learning model design and training.

More information: Weier Wan, A compute-in-memory chip based on resistive random-access memory, *Nature* (2022). DOI: 10.1038/s41586-022-04992-8. www.nature.com/articles/s41586-022-04992-8

Provided by University of California - San Diego

Citation: New neuromorphic chip for AI on the edge, at a small fraction of the energy and size of today's computing platforms (2022, August 17) retrieved 5 May 2024 from <u>https://techxplore.com/news/2022-08-neuromorphic-chip-ai-edge-small.html</u>

This document is subject to copyright. Apart from any fair dealing for the purpose of private study or research, no part may be reproduced without the written permission. The content is provided for information purposes only.