Executing low-power linear computations using nonlinear ferroelectric memristors
28 May 2020, by Ingrid Fadelli

Researchers at Toshiba Corporate R&D Center and Kioxia Corporation in Japan have recently carried out a study exploring the feasibility of using nonlinear ferroelectric tunnel junction (FTJ) memristors to perform low-power linear computations. Their paper, published in *Nature Electronics*, could inform the development of hardware that can efficiently run artificial intelligence (AI) applications, such as artificial neural networks.

"We all know that AI is slowly becoming an important part of many business operations and consumers’ lives," Radu Berdan, one of the researchers who carried out the study, told TechXplore. "Our team’s long-term objective is to develop more efficient hardware in order to run these very data-intensive AI applications, especially neural networks. Using our expertise in novel memory development, we are targeting (among others) memristor-based in-memory computing, which can alleviate some of the efficiency constraints of traditional computing systems."

Memristors are non-volatile electrical components used to enhance the memory of computer systems. These programmable resistors can be packed neatly into small but computationally powerful crossbar arrays that can be used to compute the core operations of artificial neural networks, acting as a memory and reducing their access to external data, thus ultimately enhancing their energy efficiency.

While researchers have been studying and developing memristor-based in-memory computing approaches for some time now, most systems proposed so far are difficult to scale up. The main reason for this lack of scalability is that to retain a high computational accuracy, these systems typically require a large device current and high power; thus, their original efficiency advantages are lost.
The device developed by Berdan and his colleagues operates at a far lower current than these previously proposed solutions. However, the researchers initially found that its nonlinear electronic characteristics made it incapable of performing accurate computations, at least in a more traditional sense.

"Taking inspiration from our previous work, where we developed a learning algorithm that exploits an undesirable aspect of practical devices (i.e., switching variability), we wanted to still utilize the seemingly unfit FTJ for computation," Berdan explained. "We then figured out that one of the device's flaws (i.e., its nonlinearity) can be corrected through the use of simple biasing circuits (logarithmic amplifiers), achieving both the benefits of low current and accurate computation through this circuit-device interaction."

The device developed by Berdan was manufactured and optimized via a standard fabrication process known as complementary metal-oxide semiconductor (CMOS). Its initial characterization was carried out inside a lab using high-accuracy parameter analyzers in a prober setup. The researchers also modeled their device's electrical characteristics in Python using scientific packages, such as scipy.

"In order to experimentally demonstrate our main result, the linear vector-matrix multiplication in an FTJ crossbar at low currents, we had to interface on wafer with a multi-input, multi-output crossbar," Berdan said. "This was a nontrivial task which required us to build our own PCB-based measurement platform and write the associated software and user interface. Once this was done, we were able to quickly demonstrate our hypothesis and perform more complex experiments with relative ease."

Berdan and his colleagues have introduced a method to perform linear computations in constant time using an ultra-low current non-linear FTJ crossbar that does not entail pulse-width modulation. The researchers also showed that the crossbar can be scaled up to perform large vector-matrix multiplication (VMM) operations, which are necessary for several practical applications. Their technique could bring memristor-based in-memory computing applications one step closer to the goal of mapping commercial software based on artificial neural networks straight onto hardware, including models composed of large, fully connected classification layers.

"Our purpose is to develop more efficient AI hardware for deployment on the cloud or at the edge," Berdan said. "Memristor-based in-memory computing is one development path toward this goal, and we are now focusing on system level architecture designs and further device optimization."


© 2020 Science X Network

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.