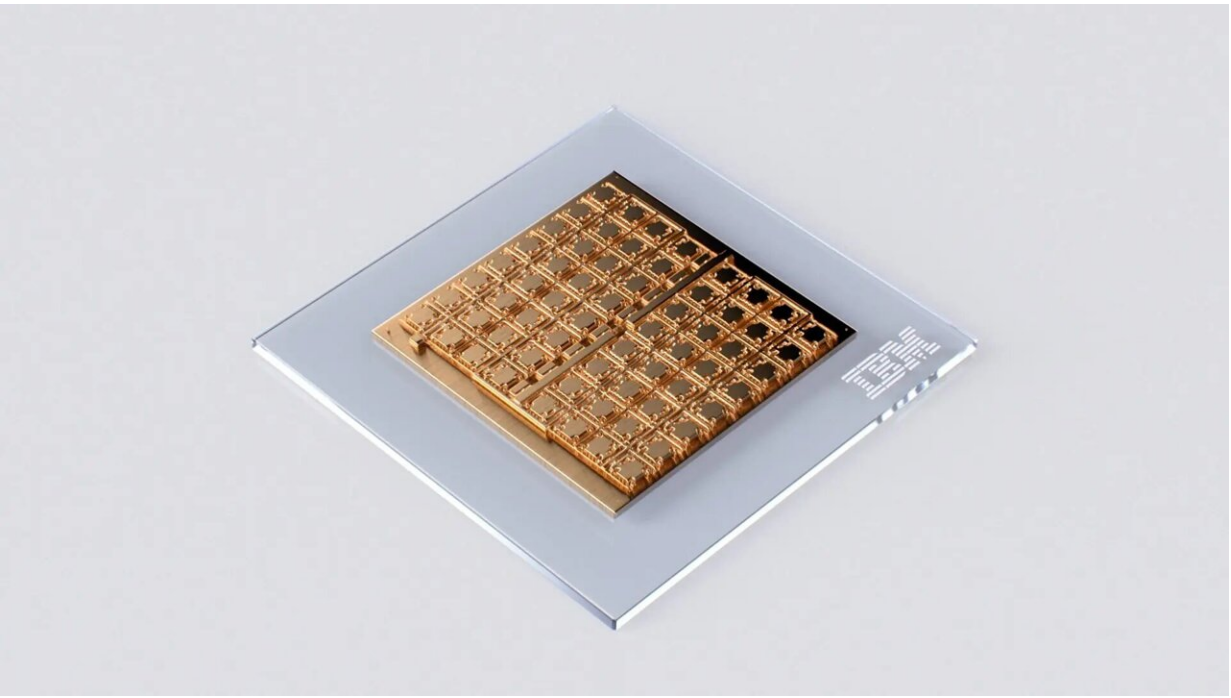
IBM develops a new 64-core mixed-signal in-memory computing chip

August 27 2023, by Ingrid Fadelli



A rendering of the IBM HERMES Project Chip. Credit: Le Gallo et al.

For decades, electronics engineers have been trying to develop increasingly advanced devices that can perform complex computations faster and consuming less energy. This has become even more salient after the advent of artificial intelligence (AI) and deep learning algorithms, which typically have substantial requirements both in terms of data storage and computational load.

A promising approach for running these algorithms is known as analog in-memory computing (AIMC). As suggested by its name, this approach consists of developing electronics that can perform computations and store data on a single chip. To realistically achieve both improvements in speed and energy consumption, this approach should ideally also support on-chip digital operations and communications.

Researchers at IBM Research Europe recently developed a new 64-core mixed-signal in-memory computing chip based on phase-change memory devices that could better support the computations of deep neural networks. Their 64-core chip, presented in a paper in *Nature Electronics*, has so far attained highly promising results, retaining the accuracy of deep learning algorithms, while reducing computation times and energy consumption.

"We have been investigating how to use phase-change memory (PCM) devices for computing for more than 7 years, starting from when we first showed how to implement neuronal functions with individual PCM devices," Manuel Le Gallo, one of the authors of the paper, told Tech Xplore.

"Since then we showed that a lot of applications could benefit from using PCM devices as compute elements, such as scientific computing and deep neural network inference, for which we demonstrated little to no accuracy loss in hardware/software implementations using prototype PCM chips. With this new chip, we wanted to go a step forward towards an end-to-end analog AI inference accelerator chip."

To create their new in-memory computing chip, Le Gallo and his colleagues combined PCM-based cores with digital computing processors, connecting all cores and digital processing units via an on-chip digital communication network. Their chip consists of 64 analog PCM-based cores, each of which contains a 256-by-256 crossbar array

of synaptic unit cells.

"We integrated compact, time-based analog-to-digital converters in each core to transition between the analog and digital worlds," Le Gallo explained. "Each core is also integrated with lightweight digital processing units that perform rectified linear unit (reLU) neuronal activation functions and scaling operations. A global digital processing unit is integrated in the middle of the chip that implements long-short term memory (LSTM) network operations."

A unique characteristic of the team's chip is that the memory cores contained inside it and its global processing unit are connected via a digital communication network. This allows it to perform all computations associated with individual layers of a neural network on-chip, significantly reducing computation times and power consumption.

To evaluate their chip, Le Gallo and his colleagues carried out a highly comprehensive study, running deep learning algorithms on their chip and testing its performance. The results of their evaluation were tremendously promising, as when running on the chip and tested on the CIFAR-10 image dataset, deep neural networks trained to complete image recognition tasks achieved a remarkable accuracy of 92.81%.

"We believe this to be the highest level of accuracy of any currently reported chips using similar technology," Le Gallo said. "In the paper, we also showed how we can seamlessly combine analog in-memory computing with several digital processing units and a digital communication fabric. The measured throughput per area for 8-bit input-output matrix multiplications of 400 GOPS/mm^2 of the chip is more than 15 times higher than previous multi-core, in-memory computing chips based on resistive memory, while achieving comparable energy efficiency."

The recent work by IBM Research Europe is a further step towards the development of AIMC chips that can support the needs and demands of deep learning algorithms. In the future, the design introduced by Le Gallo and his colleagues could be updated further to enable an even better performance.

"Using our learning from this chip and another 34-tile chip that was presented at VLSI in 2021, we have designed an end-to-end analog AI inference accelerator architecture which was published earlier this year in IEEE Transactions on VLSI systems," Le Gallo added. "Our vision combines many analog in-memory computing tiles with a mix of special-purpose, digital compute-cores connected with a massively-parallel 2D mesh. In conjunction with sophisticated hardware-aware training we have developed in recent years, we expect these accelerators to deliver software-equivalent neural network accuracies across a wide variety of models in the years ahead."

More information: Manuel Le Gallo et al, A 64-core mixed-signal in-memory compute chip based on phase-change memory for deep neural network inference, *Nature Electronics* (2023). DOI: 10.1038/s41928-023-01010-1.

© 2023 Science X Network

Citation: IBM develops a new 64-core mixed-signal in-memory computing chip (2023, August 27) retrieved 9 May 2024 from https://techxplore.com/news/2023-08-ibm-core-mixed-signal-in-memory-chip.html

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.