Engineering

Analog chip paves the way for sustainable AI

Hechen Wang

As the resources required by artificial intelligence increase unsustainably, an analog design provides an energy-efficient alternative to digital computer chips — and one that is ideally suited to neural-network computations. **See p.768**

From adjusting the tone of e-mails to generating stunning imagery, artificial intelligence (AI) has already cemented its place in everyday life. However, training a large language model, such as ChatGPT, and handling the daily queries it receives, requires data centres that generate thousands of tonnes of carbon emissions and cost millions of dollars (see go.nature.com/3qepnpv). Given these requirements, the development of generative AI is expected to stall soon if it continues to rely on standard computing hardware. Analog AI is a promising and sustainable solution that can minimize the time and energy spent on computations. On page 768, Ambrogio *et al.*[1] present an analog AI implementation for natural-language processing.

In a conventional digital computer, data are transferred from the machine's memory to its central processing unit (CPU) with every computation (Fig. 1). Semiconductor technology has now advanced to a point at which computation is dominated by the energy dissipated during this data shuttling. This energy can be anywhere between 3 times and 10,000 times that required for the actual computation, depending on where the memory is located relative to the processing unit[2].

Placing the processing units close to or inside the memory is thus the most intuitive way to improve the efficiency of AI computations. But this is difficult to accomplish with standard digital circuits, because 'multiply–accumulate' operations (which form the majority of neural-network computations) typically require chips comprising hundreds or thousands of transistors. The graphics processing units (GPUs) that were originally designed for computer-game graphics improve on this situation, at least in part, by enabling computations to be performed in parallel across multiple processing units — but each one still requires many components.

By contrast, analog computing schemes need only a few resistors or capacitors, so they can fit easily into memory, eliminating the need to move data. Ambrogio *et al.* designed an analog AI chip that is based on a technology known as phase-change memory, which relies on a material that switches between amorphous and crystalline phases when it is hit with electrical pulses. The two phases are analogous to the 1s and 0s in digital computers, but the device can also encode a state that sits somewhere between the two. This value is known as its synaptic weight, and it allows multiply–accumulate operations to be encoded in simple combinations of current and voltage — without having to move a single bit of data.

Ambrogio and colleagues' chip comprises 35 million phase-change memories that can store a total of 45 million synaptic weights — each of which is preprogrammed before the computations begin. The authors tailored their network-training techniques with the benefits and limitations of the hardware in mind, and configured the synaptic weights to optimize its output. Using these techniques, they were able to achieve 12.4 trillion operations per second for each watt of power, an energy efficiency that is tens or even hundreds of times higher than for the most powerful CPUs and GPUs. And they benchmarked this performance on natural-language-processing tasks, including spotting keywords in a set of 65,000 clips, compiled by Google, of thousands of people speaking 30 short words (see go.nature.com/3aef4ii).

Although analog chips show enormous promise for combating the sustainability problems associated with AI, the technology is still in its infancy. The path towards commercially viable analog AI products involves six steps. The first three are to design the

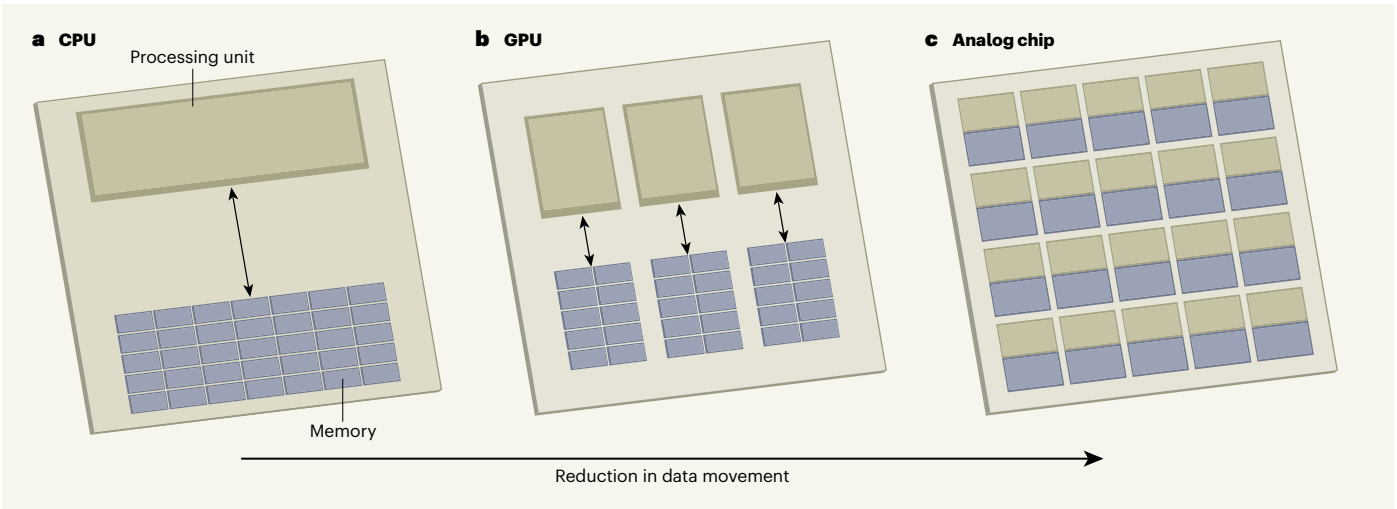

Figure 1 | Improving computational efficiency by reducing the need to move data. Conventional digital computers transfer data between a memory bank and a processing unit. **a**, Data are moved from central processing units (CPUs) to the memory every time a computation is made. **b**, Graphics processing units (GPUs) perform several computations simultaneously, but the data still need to be moved, which is time and energy inefficient, making large-scale neural-network computations for artificial intelligence (AI) unsustainable. **c**, Analog AI chips, such as the one designed by Ambrogio and colleagues[1], circumvent this problem by relocating the computing units inside the memory. The authors' implementation has an energy efficiency that is tens to hundreds of times higher than can be achieved by the most powerful digital processing units.

memory technology itself, the circuits that connect various chip elements and the architecture — the way in which the entire chip is laid out. Most research on this topic still focuses only on these steps. But the technology also requires the development of a compiler that can translate code into machine-level instructions, algorithms that are distinct from those developed for digital chips, and applications that are optimized for analog chips.

Even at the level of memory technology, the best way forwards is unclear. Phase-change memories are not the only options for analog chips. They are attractive in that they are non-volatile, which means that information can be retained even after the power is switched off. However, this feature can also be a burden. If the weights need to be reconfigured — for example, when the model size exceeds the on-chip memory or when AI training necessitates weights to be refreshed frequently — then non-volatile devices consume more power than do volatile memories, because the task needs higher voltages and complicated programming procedures with non-volatile devices. Volatile memories have therefore also been considered for use in analog chips, and have demonstrated higher efficiency than have non-volatile memories in some cases[3].

Circuit innovation is the next step to consider. A flaw in most analog-AI implementations so far is that they focus only on the multiply–accumulate operation and leave all other computing tasks in the digital domain. This means that data need to be converted from analog to digital, and vice versa, which slows computing down and limits performance. To overcome this, researchers either need to invent new techniques for converting data[4] or bring more digital operations into the analog domain.

After the circuit, the challenge is getting the architecture right. In the early 2010s, it became clear that GPUs were more efficient than CPUs for some applications. Analog AI chips represent the next step in this evolution: their throughput and efficiency are considerably better than those of CPUs and GPUs, but this comes at the expense of flexibility. A hybrid analog–digital architecture is the remedy, because digital components' flexibility enables them to bridge gaps that analog devices cannot, such as assigning computational resources[5] and correcting computation or storage errors.

These three steps would provide the hardware foundation for an analog AI chip. To further improve Ambrogio and colleagues' chip and unleash its full potential, the remaining three steps must also be tackled. The astonishing efficiency of the authors' chip merely reflects a theoretical maximum — in practice, the percentage of analog AI hardware that is actually used during computations can be very limited[6]. A customized compiler is therefore essential because it can segment tasks, then map each task efficiently to the available hardware to maximize performance.

Tailored algorithms are similarly crucial. Analog computing is inherently prone to generating errors because it is vulnerable to problems such as thermal noise, manufacturing imperfections and variations in the thermal and electrical environment of the device. This means that performance is compromised when analog-AI chips use algorithms that are designed for conventional digital computing — an issue for which there are two promising solutions. First, researchers can mitigate the impact of errors by using algorithm optimization techniques to relax the required computational precision[7]. Alternatively, they can embrace algorithms that leverage analog errors, such as are involved in Bayesian neural networks[8], which use statistical inference methods to improve the performance of ordinary neural networks.

Finally, developing dedicated applications for analog-AI chips is a key step towards making them commercially viable — and it is a challenging one. It took decades to shape the computational ecosystems in which CPUs and GPUs operate so successfully, and it will probably take years to establish the same sort of environment for analog AI. The good news is that Ambrogio and colleagues, together with other researchers in this area, are steering the ship, and have set sail towards realizing this goal.

Hechen Wang is at Intel Labs, Intel Corporation, Hillsboro, Oregon 97124, USA.
e-mail: hechen.wang@intel.com

1. Ambrogio, S. *et al*. *Nature* **620**, 768–775 (2023).
2. Horowitz, M. *IEEE Int. Solid-State Circuits Conf. Digest Tech. Pap.* **57**, 10–14 (2014).
3. Wang, H. *et al*. *IEEE J. Solid-State Circuits* **58**, 1037–1050 (2023).
4. Wu, P.-C. *et al*. *IEEE Int. Solid-State Circuits Conf. Digest Tech. Pap.* **66**, 126–128 (2023).
5. Chen, G. K., Knag, P. C., Tokunaga, C. & Krishnamurthy, R. K. *IEEE J. Solid-State Circuits* **58**, 1117–1128 (2023).
6. Sze, V., Chen, Y.-H., Yang, T.-J. & Emer, J. S. *IEEE Solid-State Circuits Mag.* **12**, 28–41 (2020).
7. Keller, B. *et al*. *IEEE J. Solid-State Circuits* **58**, 1129–1141 (2023).
8. Dorrance, R., Dasalukunte, D., Wang, H., Liu, R. & Carlton, B. *IEEE J. Solid-State Circuits* https://doi.org/10.1109/JSSC.2023.3283186 (2023).

The author declares no competing interests.

Genetics

Nuclear genome regulates mitochondrial DNA

Sonia Boscenco & Ed Reznik

A genetic analysis provides the most-detailed glimpse yet of how genetic variants in nuclear DNA regulate the copy number and variability of DNA housed in organelles called mitochondria. **See p.839**

Mitochondria are cellular organelles that house their own DNA. Each DNA molecule is built from 16,569 nucleotides — a stark contrast to the 3 billion nucleotides needed to build each copy of the nuclear genome (nuDNA)[1]. There are thousands of copies of mitochondrial DNA (mtDNA) in each cell, and each copy encodes 13 proteins that are essential for cellular energy production[1]. The genetic integrity of mtDNA is therefore crucial for healthy cell function. However, the proteins required for maintaining and replicating mtDNA are encoded by nuDNA[2]. On page 839, Gupta *et al*.[3] shed light on how nuDNA mediates maintenance and regulation of mtDNA, with consequences for human disease, physiology and evolution.

Gupta and colleagues initially focused on a distinguishing aspect of mtDNA: its high copy number per cell. Shifts in mtDNA copy number (mtCN) are associated with ageing-related disorders[4] and cancer[5], and low mtCN in blood cells has been linked with an elevated risk of a range of diseases, from type 2 diabetes[6] to heart attack[7]. The authors searched for factors associated with mtCN variation using the UK Biobank — a large-scale publicly available genetic database. The group discovered that nearly 25% of the variation in mtCN between people in this cohort could be ascribed to blood-cell composition, potentially reflecting different mtCN levels in different blood-cell types. After adjusting for blood-cell composition and other covariates, the associations between mtCN and disease that had previously been reported were no longer apparent. Instead, the authors found that mtCN simply decreased with age.