Researchers are Looking Beyond Digital Computing

They are using Biology and Light to design powerful, energy-efficient chips

The Economist, Sep 16th 2024


Brightly coloured 3D rendered illustration of melted motherboard. Illustration: Karan Singh

In 1945 John Von Neumann, a Hungarian polymath, proposed an “automatic digital computing system”. His design featured a central processing unit (CPU) for crunching numbers and a memory unit for storing instructions and data, linked by a communication pathway. Von Neumann dreamed of a computer where anything stored in memory would be instantly available to the CPU. In its absence, he came up with a clever fix: a memory hierarchy with small, fast memory close to the CPU and larger, slower memory farther away. Nearly 80 years later, his original blueprint still forms the basis of most modern processors.

But shuttling data between the processor and memory eats up time and energy, particularly for data-hungry artificial-intelligence (AI) models. An analysis by Amir Gholami and colleagues at the University of California, Berkeley found that in the past two decades, processor performance has tripled every two years, while memory access speed has increased only by about half (see chart). This means processors blaze through calculations faster than data can be fed from memory, creating a “von Neumann bottleneck”. This has led some to wonder if it is time for a new architecture. Every engineer, and every reader, carries with them the proof that such a thing is possible.

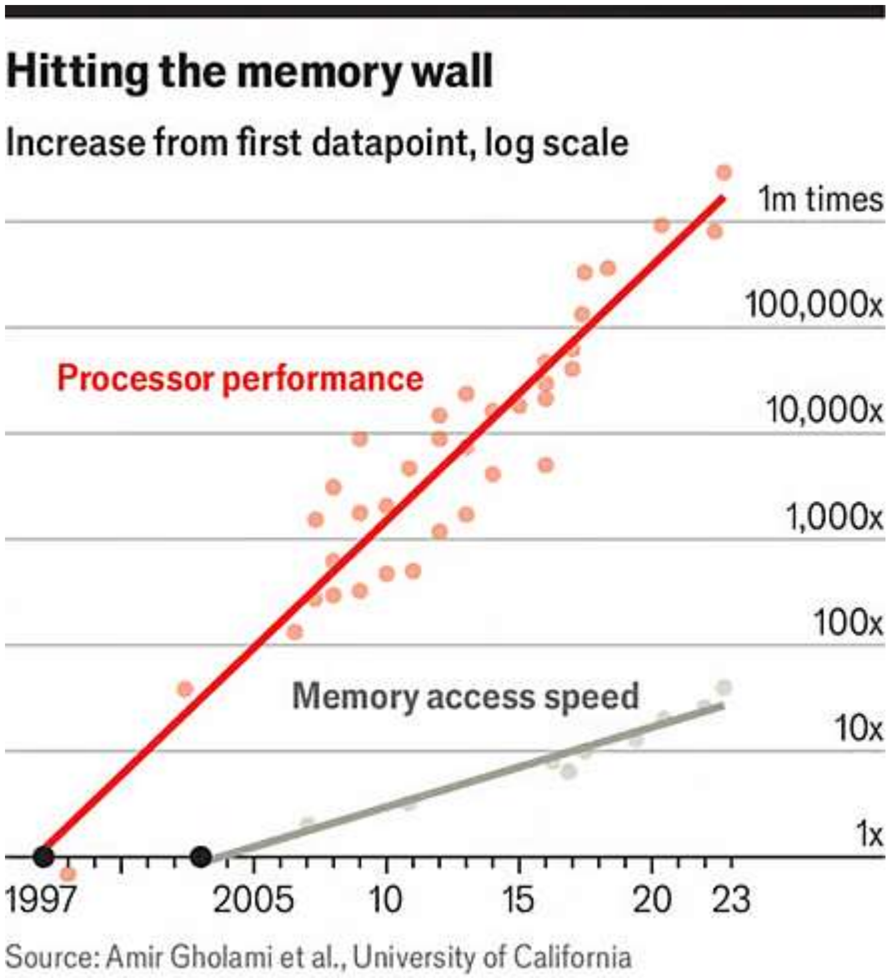

Chart: The Economist

Brains do not separate processing and storage. They run without the central co-ordinating clock von Neumann architectures use, and do more calculations in parallel than a computer, performing at exaflop speed—a billion billion calculations per second—using just 20 watts. To replicate this scale digitally, an artificial neural network would need **8 megawatts of power**. Machine-learning software already mimics the brain's parallel processing through neural networks. Is the next step to build hardware that mirrors the structure of the brain?

Sound and light show

"In-memory" computers are processors that use specialised memory devices capable of performing certain computations. The building block for this type of computer is a **memristor**, a type of a resistor whose conducting properties can be easily adjusted by applying a sufficiently high current or voltage. Crucially it retains its properties even after the current or voltage disappears, functioning as a memory. But unlike transistors which represent values as binary 1's and 0's, memristors record values on a continuum between the two, like values in the analogue world. When these devices are arranged in a grid of rows and columns, it is possible to perform a matrix multiplication in a single time step. In machine-learning applications, this allows weights to be stored within the computation unit, making processing more energy efficient.

Some think the future of computing lies not in silicon but in our skulls

Drawing on the brain's efficiency, processing units can be activated only when needed, to reduce energy consumption. "Neuromorphic" computing does away with a central clock—different neurons communicate when they are ready. These "spiking" neural networks are more efficient because they require less communication and computation. Joshua Yang of the University of Southern California believes this can be efficient and yield a "higher level of intelligence".

IBM and Intel have both designed chips that mimic this concept using current digital technology. IBM's NorthPole chip has no off-chip memory. The company claims that its brain-inspired chip is 25 times more energy efficient and 20 times faster than other specialist chips, called accelerators, for certain AI applications.

Another alternative is to use light, not electricity. Optical accelerators can process information much faster and using less power than their electrical cousins. But until recently these devices relied on components too bulky to be used with densely packed processors. Advances in **photonics** manufacturing have helped shrink these devices to nanoscale levels.

Mach won

At the heart of an optical computer is an old idea: the Mach–Zehnder interferometer (MZI), invented in the 1890s. This device takes a beam of light and splits it into two paths. Depending on the length of each path, the phase (i.e, the timing of the wave's crests and troughs) of the beam changes. The two beams are then recombined so that the amplitude, or strength, of the output beam is the amplitude of the input beam multiplied by a value that depends on the phase difference between the split beams. An optical accelerator has an array of MZIs laid out in a grid. Computation within these arrays occurs at the speed of light and the flow of light through the chip does not use energy.

Nick Harris, boss of Lightmatter, a California-based photonic-chip startup, points out that optical computers are not good for logical operations. But he says that, though they will "never run Windows", they are a great alternative for running neural networks because the energy benefit "scales exponentially".

Promising as these approaches are, analogue computers still need to talk to the digital world. Converting non-digital signals into binary 1s and 0s burns energy. But in inference applications, where a trained AI model answers user queries, speed trumps precision. This trade-off might be enough to bring analogue computers into the mainstream. ■

Related Articles

[Technology Quarterly by The Economist. Sept 21st, 2014]