FEATURE 2 August 2017

Brain box: Multitasking chips that can match the human mind

An electrical component once thought impossible is delivering machine minds that can think on their own

Raymond Biesinger

By **Liesbeth Venema**

SOMETHING is going on outside your window. Rough oblongs zoom past, one of which sidles along and stops, disgorging a series of small irregular shapes in many colours. High-pitched sounds drift up as they messily assemble into a row behind a taller shape.

Microseconds pass and all becomes clear: a group of schoolchildren has pulled up in a bus and lined up behind their teacher. Your brain has taken a chaos of sensory inputs and produced a lucid experience – as it does day in, day out, throughout your life.

If only computers could do the same. We may talk about artificial intelligence learning human smarts like driving cars or playing poker, but when it comes to quickly making sense of a huge, disordered set of information, we can't build an AI that even comes close to our brains. That's partly down to mysteries about the workings of the human mind that we're hard-pressed to explain. But it's also down to a bottleneck baked into

the architecture of nearly every computer for more than half a century.

Now we may be on the cusp of eliminating it thanks to a radical new computing paradigm, one that uses hardware that simultaneously stores and processes information – not unlike networks of neurons in the brain. Fulfil its promise and we could create machine minds that can parse rich streams of data in real time, spot patterns that elude us, and maybe learn without any help from humans.

Laptops, smartphones, tablets, you name it, they all adhere to an architecture that dates back to John von Neumann, one of the fathers of computing. Some 70 years ago, he proposed that computers should have separate processor and memory units. It may not sound like such a grand proposal, but it meant you no longer had to rewire a computer every time you wanted to run a fresh program. This division of labour has worked pretty well since then, allowing us to make ever-faster computers by souping up processors and memory in tandem.

But there is a catch. The von Neumann architecture means that whenever the processor needs information, it must retrieve it from the memory. That requires electrons to shuttle back and forth between the two, so the processor is often idle, waiting for data. This is one reason why your laptop probably has multiple "cores"; installing more processor units – each with their own connection to the memory – means they can request data simultaneously, speeding up the machine overall.

"A bottleneck has been baked into computer architecture for 70 years"

These days the bottleneck is really beginning to constrict us. There is more data to shuttle than ever, especially with the "big data" revolution looming. We are already glimpsing its promise: for example, there are algorithms that are better at predicting who will have a heart attack than standard medical risk assessments. Designed by researchers at the University of Nottingham, UK, the algorithms achieved this prowess by digesting the electronic medical records of nearly 400,000 people, a massive data-crunching task. And with the so-called internet of things encompassing ever more everyday objects – from traffic lights to fridges – machines will have even more scope to provide insights into our lives.

Manage that correctly and it could be wonderful. But computers are already overheating under the volume of data. According to a US Department of Energy report, between 5 and 15 per cent of the world's energy is spent on computing, much of it wasted on data trafficking. This is why the von Neumann bottleneck must be widened out or better still removed altogether.

Many attempts have been made to do that, for example, by developing programming languages that encode data more efficiently, reducing the number of electrons that need to shuttle back and forth. And in the 1980s, scientists started to consider using photons instead of electrons to encode information. Photons in optical cables travel faster than electrons in wires, so the data transfer time would be reduced. Others wanted to stick with electrons but cram more information in by encoding it into a

quantum mechanical property called spin. But so far neither strategy has quite come off, mainly because they are so complex to implement that the work involved outweighs any speed gains.

Raymond Biesinger

In short, we have been racking our brains for an answer for decades – which is ironic because our brains are themselves a supercomputer capable of amazing feats that need no more power than a 20-watt light bulb. They have nothing like the von Neumann bottleneck, of course, because the same network of neurons both stores information and processes it.

So how to copy them? There lies the rub. By no means do we have a complete picture of how the brain works, but there are probably at least three key features needed to mimic what it does.

First, it consists of a vast network of neurons with lots of connections called synapses. Second, those connections have synaptic plasticity; that is, they can be made stronger or weaker. We know that learning manifests itself as a strengthening of the connections between sets of neurons.

The third feature is called spike-time-dependent plasticity. This idea, less well understood than other features of the brain, says that a synapse only strengthens if the two neurons fire at similar times; if they fire out of sync, then it weakens. Over the long term this process builds strong connections between neurons that are working together to pass messages around, and weakens connections that don't seem to be important. It is thought that this is essential to the way our brains manage to learn independently. Imagine you see a green traffic light; you immediately know it means "go" because the sequence of neurons involved in that thought have developed strong connections over the years.

"Until the memristor came along, we'd never been able to

mimic a synapse”

In truth we have been trying to ape the way the brain computes for a long time. And this field, today known as neuromorphic computing, has seen some neat progress.

One of our earliest efforts was the gloriously named Mark 1 Perceptron, unveiled in 1958 by a researcher named Frank Rosenblatt. This wardrobe-sized array of electronics was organised in a network reminiscent of neurons. Rosenblatt showed cards bearing circles or triangles to the machine’s camera for it to name the shape, and he would then correct its mistakes. Within 50 tries, the perceptron had learned to output one signal for circles and another for triangles.

The perceptron was limited by the electrical engineering of the day, so its neural network wasn’t particularly extensive, nor were its abilities that exciting. But things have moved on considerably. Today, Google DeepMind’s neural networks can pull off impressive feats, such as when its AlphaGo program defeated the best human player of the game Go last year.

The Mark 1 Perceptron was an early neural network
Frederic Lewis/Archive Photos/Getty Images

The thing about DeepMind’s neural network, however, is that it is entirely simulated in software, and runs on standard silicon electronics. So although it learns in a similar way to a network of neurons, it does not get around the von Neumann bottleneck.

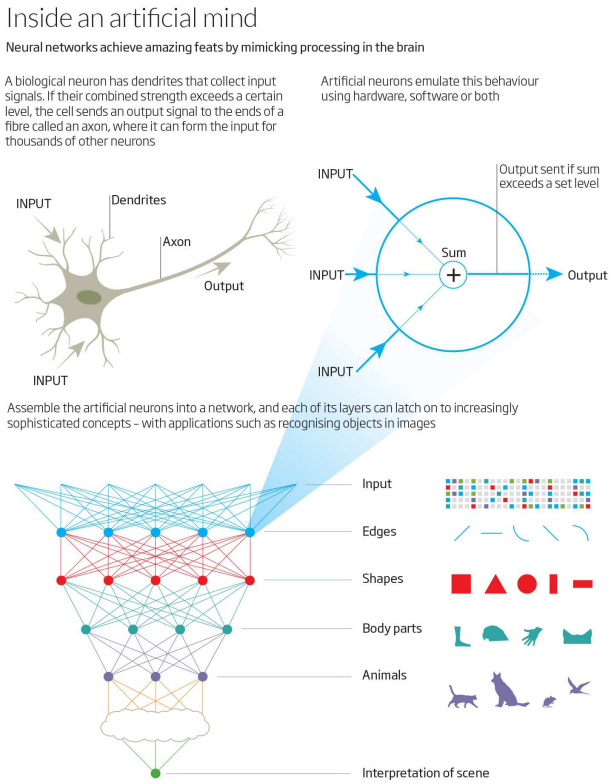
IBM’s TrueNorth chip, which appeared in 2014, goes further. It boasts 5.5 billion silicon transistors arranged into a brain-like architecture of 1 million interconnected “neurons”. It could recognise in real time objects like cars and bicycles in videos, using about as much power as a smartphone does in sleep mode. That sounds impressive, but if the chip were scaled up to match the 100 billion neurons of the human brain, it would consume 10,000 times more energy than the brain. “It is actually a wasteful approach,” says Giacomo Indiveri, a neuromorphic engineer at the University of Zurich,

Switzerland.

In short, while we've managed to mimic some of the features of the brain, we've never been able to combine all three in a physical system. The TrueNorth chip, for example, has a lot of highly connected "neurons" but it can't adjust the strength of the connections between them, except with software.

That failing is down to the fact that conventional electronics hasn't delivered a device that truly mimics a synapse. But we have a way out of that impasse, thanks to an idea that surfaced almost half a century ago.

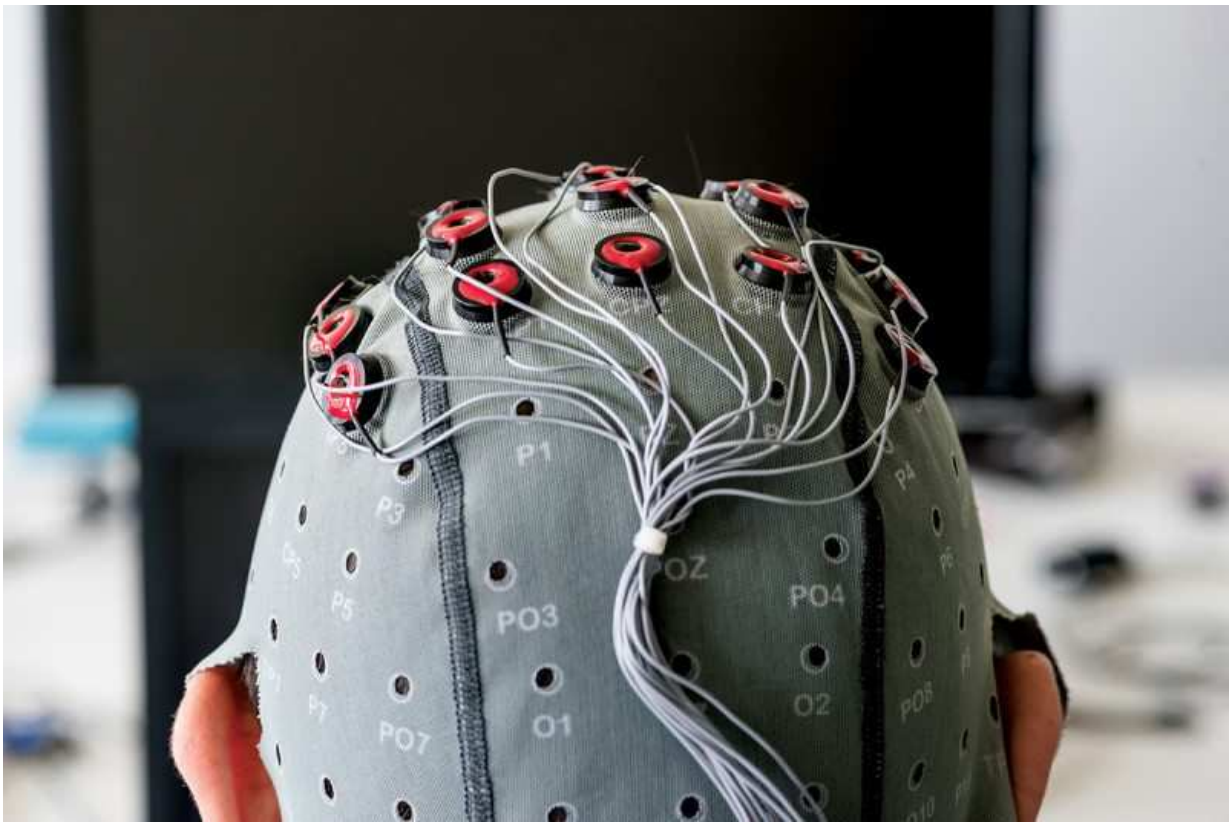
In 1971 Leon Chua, an electrical engineer at the University of California, Berkeley, was looking at the equations connecting the basic circuit components students learn about – the resistor, capacitor and inductor – when he noticed that there was another way the terms could be arranged. This produced an equation for a fourth component whose resistance would vary depending on the current. Chua called it a "memristor" because its resistance seemed to display a memory. But with no material or device known to behave in this way, his idea was largely forgotten.



Then, about a decade ago, a team at Hewlett Packard led by Stan Williams was working on a new type of memory that would, unlike a desktop computer, retain its data when the power was switched off. The researchers were investigating devices based on thin films of titanium when they noticed that their resistance was changing in strange ways depending on the current passing through them. Eventually, they realised it wasn't just electrons that were moving within the films, but also atoms, which subtly and reversibly changed the material's structure and so its resistance. In other words, the team had inadvertently created Chua's memristor (*Nature*, vol 453, p 80).

Williams's work helped explain why memristance had never been seen before; it only manifests itself on tiny scales. But with that unearthed, a range of other materials have now been found to act as memristors, including some polymers.

The advent of real life memristors has animated researchers for several reasons, including opening up the possibility of computing in more sophisticated, efficient languages than 0s and 1s (see "Beyond binary").



Memristors could be perfect for making brain interfaces
Erik Tham/Getty

But soon real action was happening on the neuromorphic computing scene. Shortly after Williams's discovery, Wei Lu, an engineer at the University of Michigan, took the crucial step and showed that memristors can act as plastic synapses. He used a device made of several thin layers of silicon, one of them with a smattering of silver ions, and showed this can mimic that second feature of the brain. Lu later showed that memristors can simulate the third ingredient too; the memristor synapse could be strengthened or weakened depending on the exact timing of applied electrical spikes.

This work shows that it is "really an exciting time for neuromorphic engineering", says Indiveri. Beatriz Noheda, a physicist at the University of Groningen in the Netherlands, agrees. "It's time to give up on silicon transistors," she says, and focus on developing full-blown memristor-based neural networks.

"Memristor networks pull off all three of the brain's most crucial functions"

It might seem that would be simply a case of scaling up Lu's work. Although his efforts involved only a single synapse with an input and output neuron, it showed that memristors could pull off all three crucial functions of the brain. The way forward would be to build networks with more and more layers of networked memristor neurons; with each added layer, the network can "think" in more sophisticated concepts (see diagram).

Not so fast, says Geoffrey Burr at IBM's Almaden research lab in California. He says the spike-time-dependent plasticity Lu has demonstrated is all very well on a small scale,

but neuroscientists aren't sure how this feature plays a role in learning on a large scale in the brain. "It must happen somehow," he says. "But we aren't even close to figuring it out." That means that implementing it in a large artificial neural network is no guarantee of getting closer to brain-like computing.

Burr prefers to stick with networks that don't have spike-time-dependent plasticity baked in. The one he uses is like those that power Google DeepMind's neural networks, which have plastic synapses controlled by software. But by running them on memristors rather than transistors he can potentially use thousands of times less energy.

In 2014, Burr constructed just such a network with almost 165,000 synapses. Training it on a database of handwritten letters, he then showed that it could accurately recognise them. Burr's memristors were made from a chalcogenide glass, a material that can switch between phases where its atoms are more or less ordered, altering its conductivity. Such phase-change memristors are becoming so reliable that chip manufacturer Intel this year began selling memory devices based on them.

Others think memristors could lead to machines that can learn entirely on their own. That includes Themis Prodromakis, a nanoelectronics researcher at the University of Southampton, UK. Starting small, last year he built a network of four input and two output neurons connected by memristor synapses. He could feed it electrical signals such as "1001" or "0110" – akin to showing circles or triangles to the 1950s perceptron. But unlike that machine, which required a human to tell it whether it had guessed the right shape, Prodromakis's network had spike-time-dependent plasticity, and learned all on its own to fire off one output neuron when it saw 1001 and the other for 0110. This worked even with noisy input signals, an important win given that real-life data is messy.

Finally, we seem to be recreating with memristors the essence of what your brain does when you look out of a window. This is independent learning with no bottlenecks.

Suitably scaled up, such self-teaching systems could screen data in real time, for example, monitoring the behaviour of self-driving cars or the integrity of bridges or nuclear plants. This could reduce the need for more sprawling data storage centres, like the ones that store data for social networks. These are sometimes built near the Arctic because they require so much cooling. But if our data is parsed in real time by networks of memristors then maybe we don't need to keep it.

Computers made from memristors have one more potential benefit: because they work akin to our brains, they may be easier to interface with them. There are already silicon-based devices out there that pick up brain activity and relay it to things in the physical world, enabling paralysed people to control exoskeletons, for instance, or let someone control a computer while dreaming.

But many challenges remain. The behaviour of neurons in the brain is complex in the extreme, and existing neural interfaces find all that information hard to handle. "The electronics to process this very rich, high-bandwidth data, becomes unbearable," says Prodromakis. Memristors could be the ideal solution because they only record signals that spike significantly, ignoring the noisy background. This excites Prodromakis, who has recently started developing memristor-based neural interfaces with Galvani Bioelectronics, a UK company formed last year in a £540 million partnership between

GlaxoSmithKline and a Google subsidiary.

One of the biggest questions hanging over memristor networks is whether we could ever manufacture them efficiently. Silicon circuits are pumped out by well-oiled factories; would anything like that be possible for memristors? The first step to answering that question is to properly scope out the best materials to make them from, and Noheda is now setting up a research centre in Groningen to do just that. If she and the other memristor champions are successful then the computers of the future could be made from a component that, for 40 years, we thought didn't even exist.

Beyond binary

Computers speak a simple language known as binary. The lexicon is built from digital 0s and 1s, so the "C" letter at the start of this box would be represented as an elaborate code: "01000011".

The dominance of binary is partly due to computers being built from transistors, electrical switches that either allow current to flow or not, and nothing in between. These two well-defined states stand in neatly for 0 and 1.

But there's a newer electrical component on the scene called a memristor (see main story). These devices are becoming more and more useful in computers built to mimic the brain, and they are plenty more versatile than the transistor. Rather than being simply on or off, they can adopt several different states of resistance. Last year, researchers led by Vikas Rana at the Peter Grünberg Institute in Jülich, Germany, got a set of memristors successfully perfoming calculations in a ternary language, which uses the digital eqivalent of 0s, 1s and 2s. This means memristors could allow computers to compute much more efficiently. And it doesn't have to stop at base 3; memristors can reliably adopt at least seven, and possibly more, resistive states.

Liesbeth Venema is a senior editor at the journal *Nature*