COMPUTING / HARDWARE

COVER

Low-Power Chips to Model a Billion Neurons

A miniature, massively parallel computer, powered by a million ARM processors, could produce the best brain simulations yet

By STEVE FURBER / AUGUST 2012

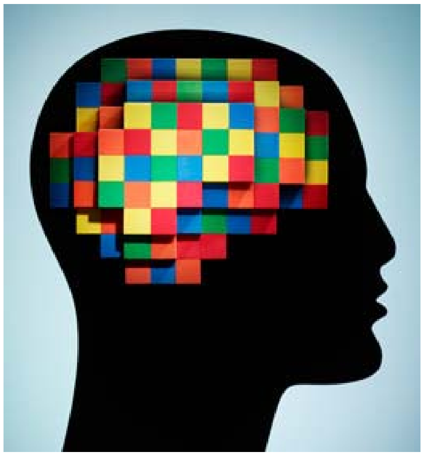


Photo: Dan Saelinger; Prop Stylist: Ariana Salvato

For all their progress, computers are still pretty unimpressive. Sure, they can pilot aircraft and simulate nuclear reactors. But even our best machines struggle with tasks that we humans find easy, like controlling limbs and parsing the meaning of this paragraph.

It's a little sobering, actually. The average human brain packs a hundred billion or so neurons—connected by a quadrillion (10^{15}) constantly changing synapses—into a space the size of a cantaloupe. It consumes a paltry 20 watts, much less than a typical incandescent lightbulb. But simulating this mess of wetware with traditional digital circuits would require a supercomputer that's a good 1000 times as powerful as the best ones we have available today. And we'd need the output of an entire nuclear power plant to run it.

Closing this computational gap is important for a couple of reasons. First, it can help us understand how the brain works and how it breaks down. There is only so much to learn on the coarse level, from imagers that show how the brain lights up when we remember a joke or tell a lie, and on the fine level, from laboratory studies of the basic biology of neurons and their wirelike dendrites and axons. All the real action happens at the intermediate level, where millions of networked neurons work in concert to produce behaviors you couldn't possibly predict by watching a handful of neurons fire. To make progress in this area you need computational muscle.

And second, it's quite likely that finding ways to mimic the brain could pave the way to a host of ultraspeedy, energy-efficient chips. By solving this grandest of all computational challenges, we may well learn how to handle many other difficult tasks, such as pattern recognition and robot autonomy.

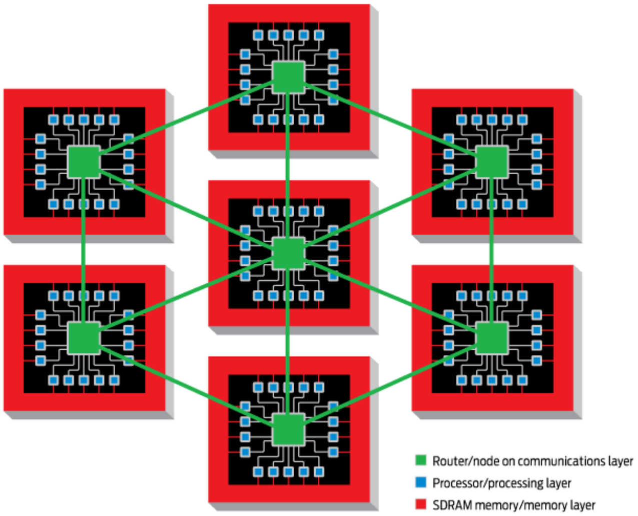
Fortunately, we don't have to rely on traditional, power-hungry computers to get us there. Scattered around the world are at least half a dozen projects dedicated to building brain models using specialized analog circuits. Unlike the digital circuits in traditional computers, which could take weeks or even months to model a single second of brain operation, these analog circuits can model brain activity as fast as or even faster than it really occurs, and they consume a fraction of the power. But analog chips do have one serious drawback—they aren't very programmable. The equations used to model the brain in an analog circuit are physically hardwired in a way that affects every detail of the design, right down to the placement of every analog adder and multiplier. This makes it hard to overhaul the model, something we'd have to do again and again because we still don't know what level of biological detail we'll need in order to mimic the way brains behave.

To help things along, my colleagues and I are building something a bit different: the first low-power, large-scale digital model of the brain. Dubbed SpiNNaker, for Spiking Neural Network Architecture, our machine looks a lot like a conventional parallel computer, but it boasts some significant changes to the way chips communicate. We expect it will let us model brain activity with speeds matching those of biological systems but with all the flexibility of a supercomputer.

Over the next year and half, we will create SpiNNaker by connecting more than a million ARM processors, the same kind of basic, energy-efficient chips that ship in most of today's mobile phones. When it's finished, SpiNNaker will be able to simulate the behavior of 1 billion neurons. That's just 1 percent as many as are in a human brain but more than 10 times as many as are in the brain of one of neuroscience's most popular test subjects, the mouse. With any luck, the machine will help show how our brains do all the incredible things that they do, providing insights into brain diseases and ideas for how to treat them. It should also accelerate progress toward a promising new way of computing.

Traditional CMOS chips were not invented with parallelism in mind, so it shouldn't come as a big surprise that they have trouble mimicking mammalian brains, the best parallel machines on Earth. A few comparisons show why brain modeling is such a thorny problem. The logic gate in an integrated circuit is typically connected to just a few neighboring devices, but the neurons in the brain receive signals from thousands—sometimes even hundreds of thousands—of other neurons, some clear on the other side of the brain. Also, neurons are always at the ready, responding as soon as they receive a signal. Silicon chips, by contrast, rely on global clocks to advance computation in discrete time steps, an approach that consumes a lot of power. To top it all off, while the connections between CMOS-based processors are fixed, the synapses that link neurons are always in flux. Connections are constantly being forged or reinforced or phased out.

Given all these differences, it's a wonder we can even begin to tackle the problem of simulating brain activity. But there have actually been some pretty impressive supercomputer models that have managed to reproduce neuron operation with great fidelity. The ongoing Blue Brain Project, led by Henry Markram at the École Polytechnique Fédérale de Lausanne, in Switzerland, is a prime example. The simulation, which began in 2005, now uses a 16 384-processor IBM BlueGene/P supercomputer and data collected from very detailed studies of brain tissue to simulate 10 000-neuron sections of the rat brain, each section no larger than the head of a pin.


🟩 Router/node on communications layer
🟦 Processor/processing layer
🟥 SDRAM memory/memory layer

STACKED DECK: SpiNNaker's machine architecture is divided into three fundamental layers. Each chip contains 18 cores that act like neurons, sending and receiving signals. All information on the connections' delays and strengths is stored in a layer of synchronous dynamic RAM (SDRAM) on each chip, and all signals pass through a separate router layer. *Click on the image to enlarge.*

Another team, led by Dharmendra Modha at IBM Almaden Research Center, in San Jose, Calif., works on supercomputer models of the cortex, the outer, information-processing layer of the brain, using simpler neuron models. In 2009, team members at IBM and Lawrence Livermore National Laboratory showed they could simulate the activity of 900 million neurons connected by 9 trillion synapses, more than are in a cat's cortex. But as has been the case for all such models, its simulations were quite slow. The computer needed many minutes to model a second's worth of brain activity.

One way to speed things up is by using custom-made analog circuits that directly mimic the operation of the brain. Traditional analog circuits—like the chips being developed by the BrainScaleS project at the Kirchhoff Institute for Physics, in Heidelberg, Germany—can run 10 000 times as fast as the corresponding parts of the brain. They're also fabulously energy efficient. A digital logic circuit may need thousands of transistors to perform a multiplication, but analog circuits need only a few. When you break it down to the level of modeling the transmission of a single neural signal, these circuits consume about 0.001 percent as much energy as a supercomputer would need to perform the same task. Considering you'd need to perform that operation 10 quadrillion times a second, that translates into some significant energy savings. While a whole brain model built using today's digital technology could easily consume more than US $10 billion a year in electricity, the power bill for a similar-scale analog system would likely come to less than $1 million.

Speed could actually be a disadvantage in some cases. If, for example, you want to develop a robot brain that can handle visual or audio inputs, it helps to have a neural model that works at about the same speed as a brain does so that it behaves at natural speeds. There are some ways around that problem. The Neurogrid project at Stanford, for example, builds brain models by operating analog circuits below their transistors' threshold voltage, which slows operations down to a biologically realistic rate. This approach comes with a caveat, however: It relies on fairly large circuits, which are hard to scale up in an economical way.

But as speedy and efficient as analog circuits are, they're not very flexible; their basic behavior is pretty much baked right into them. And that's unfortunate, because neuroscientists still don't know for sure which biological details are crucial to the brain's ability to process information and which can safely be abstracted away. That's certainly the case for dendritic trees—the branching network of neuron inputs that collect the signals arriving from all incoming connections. There is biological evidence that suggests that two such signals coming from the same branch affect a neuron differently than they would if they came from different branches, but it's still not clear whether this has a big impact on brain operation and needs to be included in models.

In 2005, my colleagues and I set out to find a good compromise between the shortcomings of the traditional digital and analog approaches to brain modeling. We wanted to come up with a system that would be capable of modeling brain activity in real time, as analog circuits do, yet be as programmable as a general-purpose digital computer.

We ended up with SpiNNaker, which received £5 million ($8 million) from the United Kingdom's Engineering and Physical Sciences Research Council in 2006. Four U.K. universities—Cambridge, Manchester, Sheffield, and Southampton—are involved in the project, along with three industry partners, ARM, Silistix, and Thales, which contributed the processor and interconnect technologies.

The basic idea behind SpiNNaker is pretty simple. The machine will consist of 57 600 custom-designed chips, each of which contains 18 low-power ARM9 processor cores. Such chips are, of course, eminently programmable. At the center of each chip, we place a specially designed router that receives and directs all the packets coming from the cores and forms links with neighboring chips. We stack 128 megabytes of synchronous dynamic RAM, or SDRAM, on top of each chip to hold the connectivity information for up to 16 million synaptic connections.

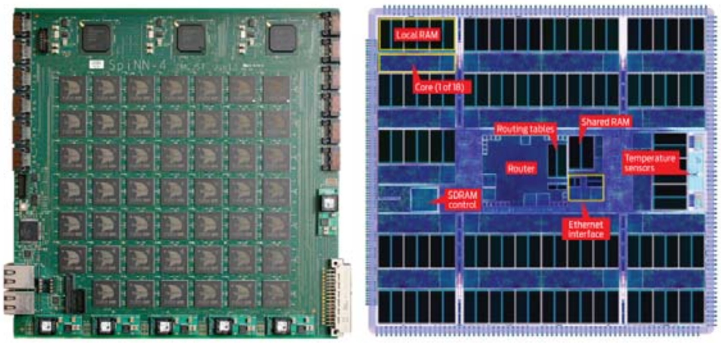


Image: Left: Norcott Technologies Limited; Right: University of Manchester

CHIPS AHOY: To keep SpiNNaker as compact as possible, the machine's chips are packed together in sets of 48 onto 23-centimeter-square boards [left]. A SpiNNaker chip contains 18 ARM9 cores [above], each with local RAM. Cores communicate with one another and with more-distant cores via a router at the center of each chip. All the information on the connectivity of the system is uploaded to these routers. *Click on the image to enlarge.*

As with most other brain models, SpiNNaker's operation is centered on the "spike"—an idealization of the electrical impulse sent out by firing neurons. The information needed to model a spike is tiny: You can condense it down to a single packet containing just 40 bits. But things get complicated when you set out to pass around as many of those packets as the brain does. To model even 1 percent of the human brain could involve wrangling 10 billion packets a second, each of which might need to be sent along to dozens of other chips containing hundreds of processors.

Such traffic is tough for even the best parallel computers to handle. Their architectures are optimized for quickly passing around big chunks of data from one point to another, but they perform very badly when juggling a great number of very small packets. The problem lies in the way the communications system is organized. Because they're designed to be very flexible and decentralized, conventional supercomputers off-load much of the routing information to the data packets they ferry. Each packet carries all its routing information in a header, and this header can't really be scaled down to fit the size of a packet. This approach works fine if you're transporting data in large chunks, but it becomes a burden when it comes to small packets. A tiny 40-bit spike may need to carry 10 times that amount of data in order to be routed properly. Setting up such headers on many tiny packets wastes a lot of energy and drastically reduces speed by clogging bandwidth.

We've eliminated this problem by taking those routing responsibilities away from the processors. In SpiNNaker, a processor modeling a spiking neuron sends a small packet that uniquely identifies the neuron to the router at the center of the chip. When a router receives a packet, it looks up the packet's unique identifier in a precomputed table that lists all the connections between neurons. Then the router passes copies of the packet out to other processors on the same chip or to routers on six adjacent chips. All the processors do is receive spikes and, if the total spike input is strong enough, generate new spikes.

In SpiNNaker we cannot implement anything like the hundreds of thousands of physical connections that are sometimes found among individual neurons. However, we can make up for that weakness by exploiting the computational power of the cores as well as the millionfold speed advantage that signals moving along metal wires have over biological ones. Because modeling a single spike requires only a fraction of the core's time, we can save on space and power by packing about a thousand simulated neurons onto every processor. The output signals generated by the interaction of those thousand neurons come in the form of spikes, which we send out using only the wires that connect each of the processors and routers. We can keep all these overlapping signals in order by using careful multiplexing.

By designing our machine in this way, we've thrown out some of the central axioms of parallel processing. Two of the key ones are the need for synchronization among the many programs running on the processors in the machine, and the expectation of deterministic operation—the idea that if you run the same program twice, you get exactly the same result both times. All the processors in SpiNNaker run in real time, and no attempt is made to impose global synchrony using a central clock. This approach mimics the asynchronous way the brain works. Communications between processors are initiated whenever a sender wants to send, and signals arrive at the receiver unheralded and must be handled, ready or not. This means that, just as in the brain, the precise ordering of signals is unknown, and the results can differ in minor ways from one run to the next.

The basic operation of SpiNNaker involves mapping a problem onto the machine—setting up the connectivity graphs in the machine's routing hardware—and then letting the model run with the spikes flying where and when they may.

Building a digital computer in this way comes with a lot of flexibility advantages. With SpiNNaker, there is effectively no difference between communicating with a nearby processor and one that's many chips away. We can upload any neural network we'd like, and the exact way that processors are connected should have no bearing on how fast that neural network can be modeled. In a sense, the SpiNNaker machine could be considered a rewirable computer—an enormous version of the field-programmable gate array chip, or FPGA, specialized for neurons. With appropriate tweaking, it should be able to model any part of the brain we choose.

Our full 57 600-chip machine won't be finished until the end of 2013, but we've already made some progress. Since we accepted delivery of the first SpiNNaker test chip in May 2011, we've built circuit boards containing four such chips, for a total of 72 processor cores. We've mounted this prototype system onto a simple wheeled robot and shown that the robot can perform real-time processing of basic visual information, like following the path of a white line of tape. It's certainly not a difficult task for a modern computer, but it shows that SpiNNaker chips can be connected to form a real-time neural network and can interact with the world through real-world sensors and actuators. We recently received the first 48-node boards, which will be used to build the upcoming system.

When complete, the full million-processor SpiNNaker machine will occupy 10 or so standard 19-inch racks and consume 50 to 100 kilowatts of power. That's still about a hundred times as much as a comparable analog model would need, but then again it's only about a hundredth the power you'd need for an equivalent supercomputer. We also have room to improve. To save money, our processors were built using a decade-old, 130-nanometer chip manufacturing process. If the project produces good results, we could move to a much smaller feature size for our integrated circuits and potentially drop power consumption by a factor of 10.

Although we've carefully designed and simulated the machine, there are still quite a few engineering questions we'll have to sort out once the machine is built. We'll need to figure out the best way to divide large networks into pieces that are small enough to be mapped onto a single processor, to cope with run-time faults, and to package everything into software that our neuroscientist and psychologist collaborators will find easy to use.

Even partial progress toward understanding how the brain works could yield dramatic benefits. We work with psychologists who use neural networks to model and test treatments for reading disorders caused by strokes or similar brain damage. These networks are trained to read—translate text to speech—and are then selectively damaged to reproduce the clinical pathology. SpiNNaker will allow these models to become more detailed and sophisticated, which could help psychologists select better therapies.

We also expect that our computer architecture could help fields outside of brain modeling that can also benefit from computers capable of dividing problems into a very large number of small processes. Some areas that could benefit include computer graphics, circuit modeling, and drug discovery.

SpiNNaker won't get us all the way to full-scale simulations of the human brain. But the machine's communications architecture could help pave the way for better-networked analog chips that could get us there. It will also help show us what information we need to make good models. Then we can really put our brains to use.

For more about the author, see the Back Story, "The Brain Maker."

This article originally appeared in print as "To Build a Brain."