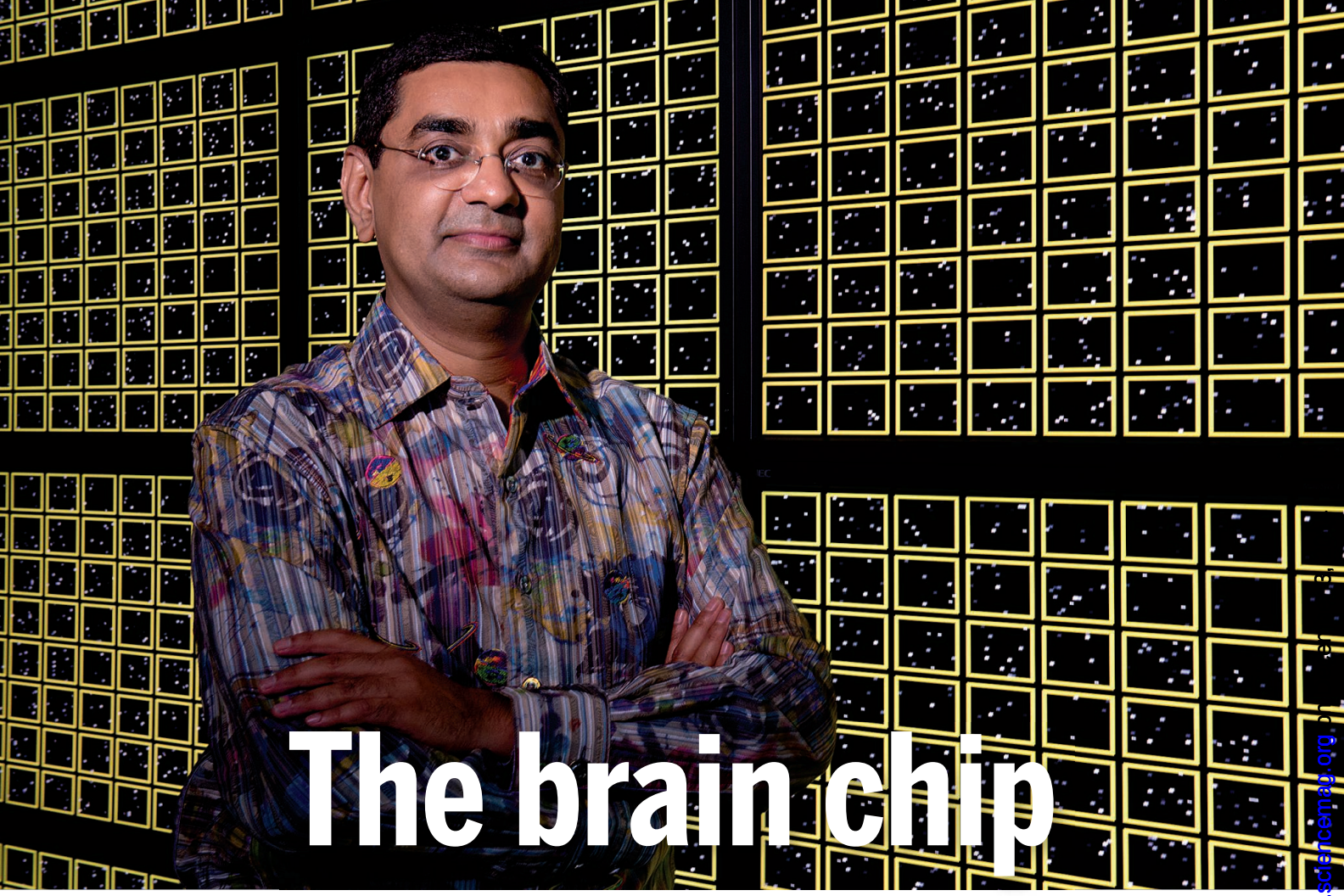


The brain chip

Microprocessors modeled on networks of nerve cells promise blazing speed at incredibly low power—if they live up to hopes

By **Robert F. Service**, *in San Jose, California*

On a computer monitor in IBM's brain lab here, a video taken from atop a tower on the Stanford University campus shows a steady stream of cars, bikes, buses, trucks, and pedestrians as they come and go. As each shape enters the scene, it's briefly surrounded by a splash of color: purple for cyclists, green for pedestrians, dark blue for cars, sky blue for trucks, and yellow for buses. The colors signal the judgments made by a postage stamp–sized computer chip, which surveys the ever-changing scene and identifies each on-screen target. "It gets almost everything right," says Dharmendra Modha, an electrical and computer engineer who leads the project at IBM's Almaden Research Center in the hills beyond Silicon Valley. A bicyclist entering from the right is quickly wrapped in purple. But as the cyclist stops, dismounts, and starts to walk his bike, the color shifts to pedestrian green. Modha smiles. "We got a little lucky with that one," he says.

Easy for a human, such pattern recognition is a tour de force for a computer. On page 668 of this issue, Modha and colleagues at five IBM research centers and Cornell University describe the chip responsible for it: the first-ever production-scale "neuromorphic" computer chip designed to work more like a mammalian brain than like the processors in a laptop or smart phone. The new chip, called TrueNorth, marks a radical departure in chip design and promises to make computers better able to handle complex tasks such as image and voice recognition—jobs at which conventional chips struggle.

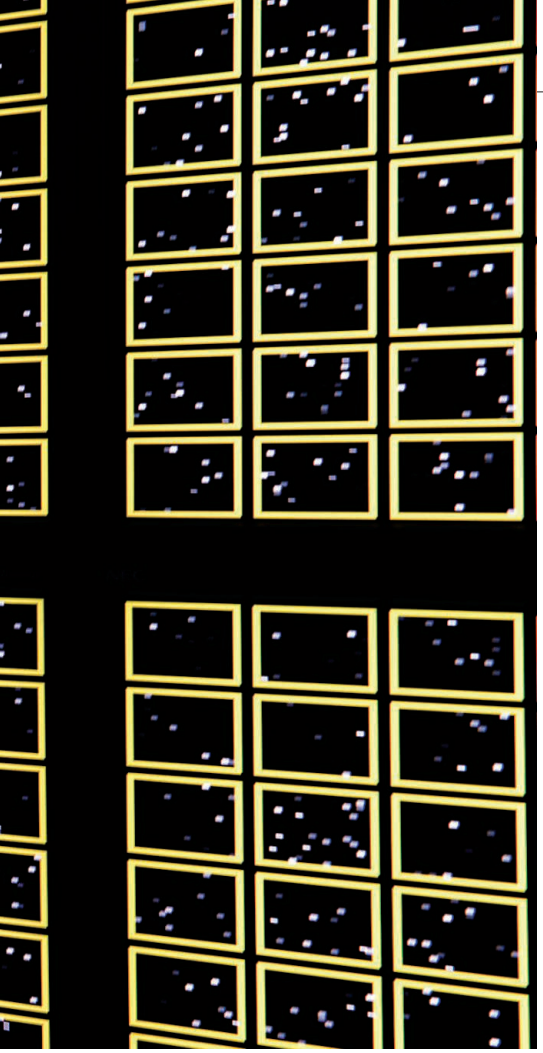
PODCAST

To hear a podcast with author Robert F. Service, see http://scim.ag/pod_6197.

Progress in brain-inspired computing has been building for several years. Several U.S. and European labs are working on different versions of the technology. But outsiders say TrueNorth has the potential to propel neuromorphic computing from an alluring research endeavor to a real-world technology.

TrueNorth contains 5.4 billion transistors wired together to form an array of 1 million digital "neurons" that talk to one another via 256 million "synapses." Like the brains of organisms, this neural network architecture accomplishes complex tasks such as pattern recognition far more efficiently than conventional chips can. "It's a tremendous achievement," says Wei Lu, an electrical engineer and computer scientist at the University of Michigan, Ann Arbor. Horst Simon, a computer scientist and deputy director of the Lawrence Berkeley National

PHOTO: DEANNE FITZMAURICE



Dharmendra Modha's team at IBM aims to give computers the perception skills of biological organisms.

Laboratory in California, agrees. "This is a qualitative breakthrough to go from one computing paradigm to the next," he says.

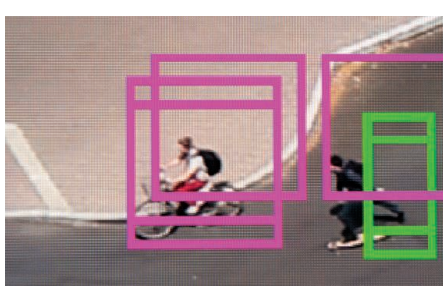
TO US, PERCEPTION seems so simple. We peer out the window of an office building and easily distinguish between a person on a bike and one on a skateboard. When we read a sentence that mentions a calico and a tabby, we understand that both are cats without having to be told. And with little effort we pick out the voice of the person we are talking to at a noisy cocktail party. By comparison, modern computers, for all their powers of calculation, remain infants at such tasks. State-of-the-art algorithms can crack these challenges, but they require massive computing power. Google, for example, recently demonstrated a setup capable of recognizing cats and human faces in video clips. But the task required 16,000 processing chips and about 100 kilowatts of power. Our brains, by contrast, use just tens of watts.

Today's chips are based on the same architecture that was developed 7 decades ago by the Hungarian-born polymath John von Neumann. In 1945, von Neumann laid out modern computing's basic design, with separate processing, memory, and control units. The architecture excels at performing sequences of logical operations and is ideal for crunching numbers and running spreadsheets and word processors. But it struggles when trying to integrate and process large amounts of data, as vision and language processing demand.

The difficulty arises in the way conventional chips carry out a task. To do anything useful, they must pull data in from stored memory, manipulate it, and send the result back to storage before tackling the next operation. Moving all that data back and forth demands power and creates traffic bottlenecks. For decades, engineers have compensated by shrinking the transistors, communication lines, and other devices on chips. That shortened the distance data needed to travel, reduced the power demands of individual devices, and sped them up.

But this strategy is all but tapped out. Individual device features in the latest chips are as small as 14 nanometers across—the width of fewer than 100 atoms, and close to the limits set by physics. To keep computer power on its upward trajectory, makers have resorted to tiling multiple processor chips side by side. That approach compounds the need to shuttle data—and the challenge of matching the skills of biology.

In 2012, for example, Modha and his colleagues used an IBM supercomputer called Sequoia at Lawrence Livermore National Laboratory in California to simulate the network communication in a human-scale brain. The simulation used conventional circuitry programmed to emulate the communication among 500 billion neurons and 100 trillion synapses. All of Sequoia's 1.5 million processor chips and 1.5 petabytes (1.5 quadrillion bytes) of memory were devoted to the task—and even so, the simulation ran at only 1/1500 the speed of a real brain. If the simulation had been scaled up to keep pace with actual "wetware," it would have required 12 gigawatts of power, Modha says—the power consumption of Los Angeles and New York City combined.



One neuromorphic chip can spot patterns, such as clues that distinguish cyclists from pedestrians, as handily as an array of power-hungry processors.

Biology takes a different approach. Individual neurons in our brains communicate with thousands of other neurons through chemical signals at connection points called synapses. When the combined chemical signals from a neuron's partners exceed a certain threshold, it fires, creating an electrical spike that triggers it to pass chemical signals to other partners. And the communication ripples through the brain's network, which in humans includes an estimated 100 billion neurons and 100 trillion synapses.

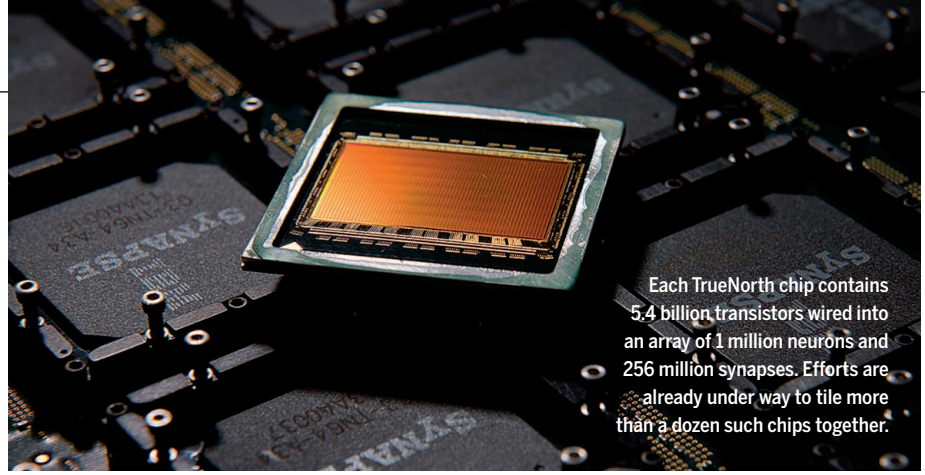
The setup needs no program: Connections to other neurons that fire regularly are reinforced, whereas those that don't are pared away. In that way, the architecture itself learns. The strategy is also efficient, because once a nerve fires, it "forgets" all about the inputs that pushed it over the threshold and passes on only the fact that it fired. And it enables the brain to distribute information processing tasks. Separate groups of neurons in the visual cortex, for example, respond to horizontal and vertical edges in a scene and pass those perceptions to other neurons that integrate the information. This midlevel processing minimizes the need to shuttle data from one place to another.

The upshot, says neuromorphic chip expert Gill Pratt, is that whereas conventional computer processors base their performance mainly on speed, the wetware in a brain relies on the complexity of its network—which lowers energy consumption. That's a survival advantage, says Pratt, who heads the SyNAPSE (Systems of Neuromorphic Adaptive Plastic Scalable Electronics) program at the Defense Advanced Research Projects Agency in Arlington, Virginia, which funds the IBM effort and a companion program at HRL Laboratories LLC in Malibu, California. "In nature, the most precious resource for most animals is food," he explains.

Nature's emphasis on parallel processing and complexity is at the heart of most attempts at neuromorphic computing. The term was coined in the 1980s by California Institute of Technology electrical engineer Carver Mead to describe efforts to mimic neurological architectures using analog computer circuits—devices that, like neurons, send signals only after inputs from their neighbors reach a predetermined threshold. But today, researchers apply the term more broadly to a range of analog, digital, hybrid, and even software systems. Like wetware, all neuromorphic systems distribute processing and memory tasks broadly across the chip to minimize the need to ferry data back and forth to central logic and memory hubs.

One effort, run by Narayan Srinivasa at HRL Laboratories, follows some of Mead's

CREDIT: DEANNE FITZMAURICE



Each TrueNorth chip contains 5.4 billion transistors wired into an array of 1 million neurons and 256 million synapses. Efforts are already under way to tile more than a dozen such chips together.

inspiration with analog circuitry. Srinivasa's team has developed a silicon chip with analog circuitry containing 576 neurons and 73,000 synapses. In June, Srinivasa's team showed that the chip could interpret visual signals well enough to pilot a palm-sized helicopter through a building. The chip picked up signals from an imaging sensor and used that information to determine whether it was in a new room or one that it had flown through previously. Because the setup uses analog circuitry, Srinivasa says, the chip can strengthen connections over time and thus learn and improve its performance.

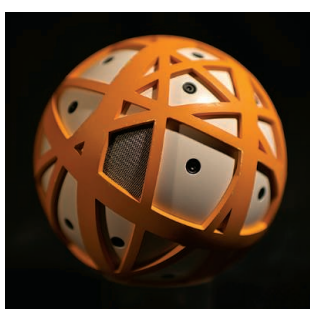
Another project is being run by Stephen Furber of the University of Manchester in the United Kingdom. Known as SpiNNaker (Spiking Neural Network Architecture), Furber's system is a supercomputer constructed from conventional digital-based low-power computer chips, the sort commonly found in smart phones. SpiNNaker now consists of 20,000 chips, each of which represents 1000 neurons. This fall, Furber says, he expects that number will rise to 100,000 chips representing 100 million neurons, and eventually a 1-million-chip system representing 1 billion neurons—about 1% of the neurons in the human brain. Although similar in concept to the Sequoia simulation, SpiNNaker's design is expected to model brain activity at speeds matching biology.

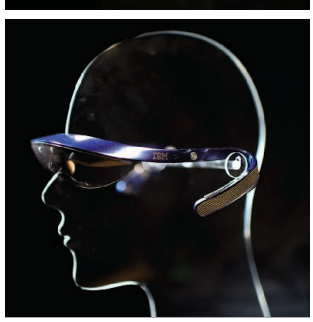
Furber says he and his colleagues aren't building SpiNNaker for particular applications. For now, he says, the machine will serve as a testbed for computer scientists and neuroscientists to model brain function in a way that connects their small-scale knowledge of neurons with the global insights of brain imaging and psychology. With SpiNNaker, "you can ask very detailed questions that are very difficult to ask in biology," Furber says.

IBM'S TRUENORTH ALSO uses conventional digital devices. But in this case, Modha and his colleagues wired them in an entirely new hardware architecture. So far, that's enabled them to represent 16 times as many neurons on a single chip as previous efforts could, thus speeding up its parallel processing at low power.

Modha and his colleagues unveiled their first attempt in 2011: a pencil tip–sized chip containing 256 neurons and 262,000 synapses, a unit they refer to as a core. Their current chip contains an array of 4096 cores. "This is the largest chip IBM has ever made by a factor of 3," Modha says. And because they used more advanced chipmaking techniques to build it, each of TrueNorth's cores is just 1/15 the size of the previous generation and consumes 1/100 as much power. That makes it more than 1000 times as efficient as chips made with the conventional architecture. Whereas a typical chip sucks down 50 to 100 watts per square centimeter of chip space, TrueNorth sips just 20 milliwatts for the same area of circuitry.

Modha and his colleagues are already working to build on their success. They are testing arrays of up to 16 TrueNorth chips and are getting ready to push beyond that. Because each chip already represents an array of cores wired together, it's a straightforward task to tile more and more chips together to increase the computational power. On the drawing board are collections of 64, 256, 1024, and 4096 chips. "It's only limited by money, not imagination," Modha says.





Neuromorphic chips could lead to rolling robots (*top*) that beam back data from hazardous environments or to navigation aids for the blind.

That's just the hardware. To control it, Modha and his colleagues have also created a novel programming language—a type of machine language for synaptic chips. Modha says engineers have already used it to write hundreds of simple software routines called corelets that tell TrueNorth's neural network how to carry out common computational tasks. Vision corelets that spot features such as vertical and horizontal edges helped TrueNorth distinguish the cars, cyclists, and pedestrians in the brain lab video, Modha says. Equipped with an array of these basic functions, programmers should be able to rework conventional, inefficient neural network programs to run on the IBM chips. Thousands of such programs already exist for carrying out tasks such as visual and auditory recognition. The IBM chip should be able to run them far more quickly while drawing less power.

IBM hopes to commercialize the chips "sooner rather than later" and is looking into partnerships with other companies, Modha says. Meanwhile, he says, IBM plans to give computer scientists access to the chips to explore the cornucopia of new applications they should make possible. Berkeley Lab's Simon thinks that in time the chips could lead to a new generation of power-efficient supercomputers. "There's a huge potential here for addressing different types of calculations that are currently done on [conventional chips] but less efficiently," Simon says.

Modha's team has dreamed up some applications of its own. One of them, displayed in the team's brain lab across the room from the video monitor tracking Stanford traffic, imagines an advanced version of Google Glass that processes visual information and communicates it to visually impaired wearers. A blind person, for example, might receive auditory cues to avoid objects in her path. In a second example, called Tumbleweed, a robot outfitted with multiple sensors rolls its way around a dangerous environment, such as the inside of a damaged nuclear reactor, and beams back visual, temperature, chemical, and radiation data. In the nearer term, Furber suggests, neuromorphic chips could be integrated into smart phones to improve their visual and voice recognition.

To encourage such applications, IBM has set up a virtual school, called Synapse University, where computer scientists and researchers can learn how to program the new chips to do whatever they want. "If IBM can do that, I'm sure lots of people will have fun with it and do real science," Furber says. And eventually, perhaps, computers will begin to emerge from their infancy in carrying out everyday tasks that we take for granted. ■

PHOTOS: DEANNE FITZMAURICE