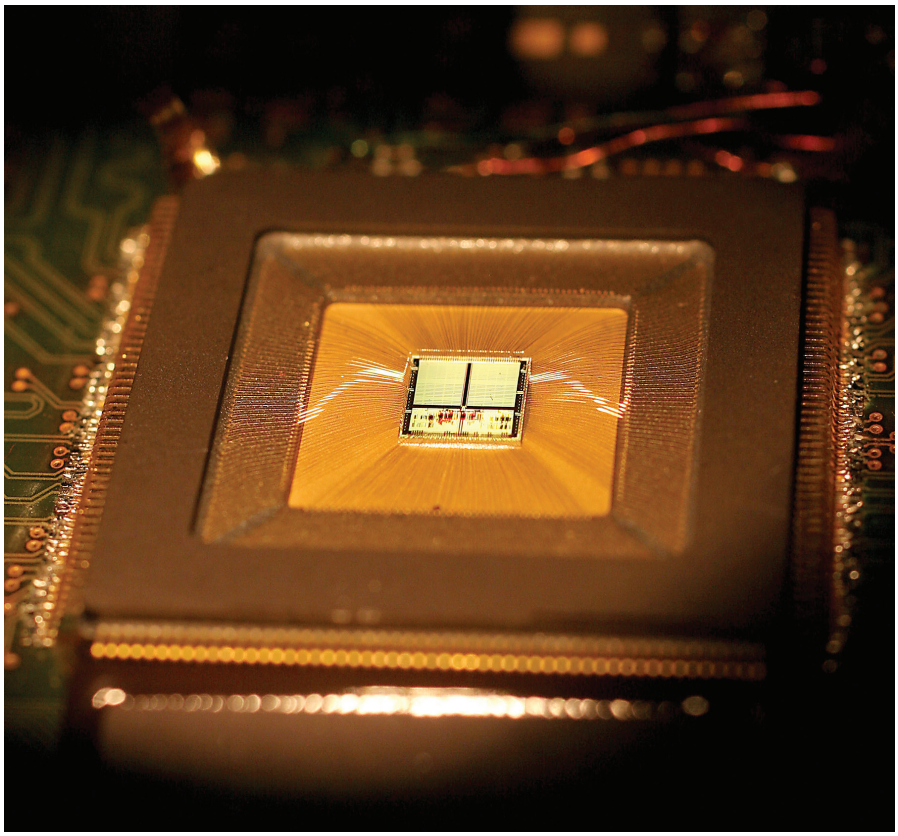
Science | DOI:10.1145/2601069 Don Monroe

Neuromorphic Computing Gets Ready for the (Really) Big Time

A technology inspired by biological principles but 'steamrolled for decades' prepares to take off as Moore's Law approaches its long-anticipated end.

As the long-predicted end of Moore's Law seems ever more imminent, researchers around the globe are seriously evaluating a profoundly different approach to large-scale computing inspired by biological principles. In the traditional von Neumann architecture, a powerful logic core (or several in parallel) operates sequentially on data fetched from memory. In contrast, "neuromorphic" computing distributes both computation and memory among an enormous number of relatively primitive "neurons," each communicating with hundreds or thousands of other neurons through "synapses." Ongoing projects are exploring this architecture at a vastly larger scale than ever before, rivaling mammalian nervous systems, and developing programming environments that take advantage of them. Still, the detailed implementation, such as the use of analog circuits, differs between the projects, and it may be several years before their relative merits can be assessed.

Researchers have long recognized the extraordinary energy stinginess of



"Spikey" is the first neuromorphic chip developed by the Electronic Vision(s) group at the University of Heidelberg, as part of the EU's Human Brain Project.

IMAGE COURTESY OF FACETS PROJECT. WWW.FACETS-PROJECT.ORG

biological computing, most clearly in a visionary 1990 paper by the California Institute of Technology (Caltech)'s Carver Mead that established the term "neuromorphic." Yet industry's steady success in scaling traditional technology kept the pressure off.

"Neuromorphic computing people got steamrolled for decades because Moore's Law just kept getting better and better ... so they could just never catch up," says Todd Hylton of Brain Corporation, San Diego, CA, who in 2008 established the SyNAPSE (Systems of Neuromorphic Adaptive Plastic Scalable Electronics) program while he was working for the U.S. Defense Advanced Research Projects Agency (DARPA). "That's not the case anymore."

"It's a design trade-off: power versus complexity," says Gill Pratt, who now runs the SyNAPSE program. By employing a huge number of computational elements, each can be dedicated to a particular subtask. The inputs change only rarely, avoiding the energy overhead for moving charge and transporting data. "We believe this is the key reason" that biological circuits are so much more efficient, Pratt says. "You can only get away with it if size is free," which is effectively true with modern integrated circuits.

In addition to massive parallelism, the big neuromorphic projects explicitly track "spikes," the short pulses that carry information between biological neurons. When massively parallel artificial neural networks were studied starting in the 1980s, spikes were generally represented only by their average rate. Since the late 1990s, neuroscientists have found the detailed timing of the spikes in the brain conveys important information. None of the new chips reproduce the detailed spikes, only the time at which they occur, unsynchronized with any global system clock as in traditional chips. The guiding philosophy is not to reiterate or simulate the brain in complete detail, but to search for organizing principles that can be applied in practical devices.

Spikes from hundreds of neurons are transmitted, via synapses, as inputs for another neuron, which combines the spike information to compute its own probability to fire off a spike to neurons to which it connects. The parameters that define that input and the computation for each neuron specify the chip architecture. Some teams perform this computation with analog circuitry, which reduces the power consumption at the cost of increased sensitivity to noise and device variability. Other researchers consider the power savings from the massively parallel architecture much more important, and have implemented the circuits digitally.

The guiding philosophy is not to reiterate or simulate the brain in complete detail, but to search for organizing principles that can be applied in practical devices.

A key parameter is the synaptic strength, or "weight," which quantifies the contribution from each input neuron. The modification of synaptic weights, in response to local activity across a network, is a critical mechanism for learning. Narayan Srinivasa of HRL (formerly Hughes Research Laboratory) in Malibu, CA, who heads one of the two major collaborations funded under DARPA's SyNAPSE program, says replicating that learning, or plasticity, has been a central goal of his team. They designed chips that use the timing of the spikes to adjust the synaptic weights. The team found that a simulated chip discovers new features to look for in images without external coaching. "If the information changes, it will adapt," says Srinivasa. The latest challenge from DARPA, he says, is to train a version of their chip, with just 576 neurons and 73,000 synapses, to navigate a flying, hummingbird-scale, autonomous vehicle.

The other SyNAPSE collaboration, headed by Dharmendra Modha of IBM Research in Almaden, CA, has not yet focused on on-chip learning. Instead, the team is assembling a system of unprecedented scale, two billion cores with 100 trillion synapses—still smaller than a human brain, but comparable to that of a mouse. IBM implements neurons digitally, in part because the background leakage and variability of their state-of-the-art transistors is not well suited for analog circuits. In addition, doing everything digitally allows a perfect mapping between the circuit and simulations, Modha says. "Digital neurons are the key to success."

One requirement for exploiting these new chips is a new paradigm for programming them, Modha says. His team has developed a framework in which blocks of neurons are conceptually bound together to create "corelets" that perform a particular function, but whose inner workings are encapsulated in a fashion similar to object-oriented programming. This "end-to-end ecosystem" lets developers separately design corelets and combine them hierarchically into larger functional elements.

Modha says that even with digital neurons, "power is excellent" because of the massively parallel architecture. Yet in the Neurogrid project at Stanford University, Kwabena Boahen employs analog neurons, operating transistors in the ultra-low-power subthreshold mode championed by Carver Mead. These circuits often respond to electrical noise, he says, but researchers are looking for organizing principles so "reliability arises at the system level, not from the individual components." Like Modha, Boahen sees software support as critical to using the new systems, and is working with Chris Eliasmith at the University of Waterloo, Canada, to develop a "neural compiler" that lets developers move their proven algorithms to the new architecture.

Many neuromorphic projects mix two distinct goals: exploring biology and improving technology. However, "understanding the brain itself is different from building a better handset," says Tony Lewis of Qualcomm, in San Diego, CA. "I think you have to pick one or the other." His company has picked applications, announcing plans for a "neural core processor" called zeroth. The chip would work alongside traditional digital chips "to bring the sort of intelligence that people usually associate with the cloud down to the handset;" for example, image analysis and voice rec-

ognition. Lewis says the company does not intend to build these applications, just to make them possible. "We're enablers," he says, "not the guys who are going to write the final applications." Qualcomm announced its chip plans much earlier than usual in order to get feedback on the software tool chain that will let programmers take advantage of the new architecture when it is available.

Qualcomm also provides venture funding to collocated but independent Brain Corporation, founded by neuroscientist Eugene Izhikevich. Hylton, who moved to the company from DARPA in 2012, says it is eager both to understand brains and to develop autonomous robots, "but we're looking to solve it with a variety of approaches," not exclusively neuromorphic computing. In the end, some algorithms might be implemented using traditional hardware, as happened with neural networks. Hylton says that, although adaptation to novel environments is important, it will be important to share the learning to give each device a baseline capability. "All the robots don't have to go to school."

The European Union is also pursuing neuromorphic computing in a big way, as one piece of the billion-euro Human Brain Project (HBP). However, "the European sense is that this whole area is still held back because the basic principles of operation in the brain are still a mystery, and if only we could crack that, then we'd make giant leaps forward," says Steve Furber of the University of Manchester in the U.K. "On the U.S. side…there's a sense that 'we can build little cognitive units, let's go see what applications we can find.'" For a decade, Furber has been running the SpiNNaker (Spiking Neural Network Architecture) project, which combines ARM cores in a massively parallel network. HBP funding will move the project to a much larger scale, a million cores.

SpiNNaker's digital architecture is complementary to the other HBP neuromorphic project, headed by Karlheinz Meier at the University of Heidelberg, which uses mixed-signal neurons previously explored in the BrainScaleS (Brain-inspired multiScale computation in neuromorphic hybrid Systems) project. To reduce the need to drive off-chip interconnections, this project uses wafer-scale integration of 200,000 neurons and 50 million synapses with a relatively-cheap 180 nm process on an 8-inch-diameter wafer. "We decided to be faster than biology: not milliseconds, but tens of nanoseconds," Meier says, which lets researchers study slow processes like learning. "In our system, you need 10 seconds [to simulate] a day." With the support of HBP, the team is building a 20-wafer system that will have four million neurons and one billion synapses, to be completed in two years.

Even the most ambitious of the new neuromorphic projects are idealized abstractions of biological brains; they assume that only the timing of spikes is important, they replace the three-dimensional "wetware" geometry with two-dimensional networks of fast wires carrying multiplexed events, and the feedback paths that help fine-tune brain performance are largely missing. Still, "it's a very exciting time," says DARPA's Pratt. The scale of these projects "allows you to show things you don't see when they're small."

Will spike-based neuromorphic systems be more successful than the artificial neural networks of a quarter century ago? Actually, "neural networks were not a failure," says Meier. "The algorithms were incorporated into deep learning," like that employed by Google for facial recognition.

Yet neuromorphic systems hold the promise of going beyond algorithms to specialized hardware that cashes in the enormous complexity of today's integrated circuits to do large computations with low-power mobile devices. C

Further Reading

"A New Era of Computing Requires a New Way to Program Computers," Dharmendra S. Modha, http://bit.ly/1i2rxqh

"Low-Power Chips to Model a Billion Neurons," Steve Furber, *IEEE Spectrum*, August 2012, http://bit.ly/1fuotGv

"Cognitive Computing," Dharmendra S. Modha, Rajagopal Ananthanarayanan, Steven K. Esser, Anthony Ndirango, Anthony J. Sherbondy, Raghavendra Singh, *CACM*, August 2011, http://bit.ly/1i2rtqz

"A Computer that Works like the Human Brain," a TEDx talk with bioengineer Kwabena Boahen, http://bit.ly/1gOY0UM

Don Monroe is a science and technology writer based in Murray Hill, NJ.

© 2014 ACM 0001-0782/14/06 $15.00

ACM Member News

MAAREK LEADS YAHOO! SEARCH ENGINE TEAMS IN ISRAEL, INDIA

New ACM Fellow Yoelle Maarek is vice president of research and the head of Yahoo Labs in Haifa, Israel, where the search engine expert leads a team of 60 scientists and research engineers co-located in Israel and India in utilizing big data to solve complex search problems and develop new products.

Her team's dual mission: positively impact Yahoo systems and the academic research community. The French-raised Maarek is well suited to these tasks; she earned her Ph.D. in computer science from Technion – Israeli Institute of Technology, and did stints at IBM Research in the U.S. and Israel concentrating on search technologies. In 2006, she joined Google to open its first Haifa-based Engineering Center; she moved to Yahoo in 2009.

A typical day for Maarek is "a mixture of technical contributions, design reviews, writing papers, developing code, testing and debunking theories and technologies, and talking to partners and colleagues worldwide."

Maarek thrives on research challenges, "especially those involving huge amounts of data and hundreds of millions of users." That means conducting data-driven experiments to demonstrate working principles for new search engine functionality and products. She tries to maintain equilibrium amid the triangle of conflicting usage data factors: Rightsizing big data + Personalization + Data Privacy.

"Our goal is to invent something that no one has done yet," she said.

Living in the world's Fertile Crescent is advantageous and invigorating. "I have the best of all worlds. I love my job. I can communicate with Yahoo developers in California and India, take my team to the beach to work, and go hiking with my family," Maarek said.

—Laura DiDio