

A computing hardware approach aspires to emulate the brain

David Kramer

Citation: [Physics Today](#) **76**, 1, 23 (2023); doi: 10.1063/PT.3.5155

View online: <https://doi.org/10.1063/PT.3.5155>

View Table of Contents: <https://physicstoday.scitation.org/toc/pto/76/1>

Published by the [American Institute of Physics](#)

ARTICLES YOU MAY BE INTERESTED IN

[The topology of data](#)

[Physics Today](#) **76**, 36 (2023); <https://doi.org/10.1063/PT.3.5157>

[The early universe in a quantum gas](#)

[Physics Today](#) **76**, 14 (2023); <https://doi.org/10.1063/PT.3.5152>

[Nazis, émigrés, and abstract mathematics](#)

[Physics Today](#) **76**, 44 (2023); <https://doi.org/10.1063/PT.3.5158>

[Ethics in physics: The need for culture change](#)

[Physics Today](#) **76**, 28 (2023); <https://doi.org/10.1063/PT.3.5156>

[Enceladus erupts](#)

[Physics Today](#) **76**, 62 (2023); <https://doi.org/10.1063/PT.3.5165>

[ITER's net loss](#)

[Physics Today](#) **76**, 12 (2023); <https://doi.org/10.1063/PT.3.5150>



focusing on renovating buildings to be more energy efficient and procuring carbon-free power. Transportation is also in their sights: Ideas being floated to encourage employees to cut emissions include providing electric bikes for campus transportation, offering parking only to carpoolers, and switching the lab's vehicles to electric cars, says Shannon Blair, who is on the lab's sustainability team. "Our government fleet is 1500 vehicles. It's 2% of our total emissions. That's tiny, but it's visible."

At Johnson's workplace in Utrecht, the ultralow-temperature freezers consume the equivalent energy of 60 average Dutch households. "We are facing a huge energy crisis in Europe, so the sustainability community is using that to get attention and to get institutions on board," she says. Increasing a freezer's temperature to -70°C from the typical -80°C uses about 30% less energy. "Through an ongoing in-house challenge, some groups have combined the contents of freezers and turned some off altogether."

Efficiency in cost and carbon

A few years ago, the Gemini Observatory telescopes got solar panels and energy-efficient equipment, including transformers, cooling systems, LED lights, and motion sensors. Solar panels provide 20% of the energy needed at Gemini South on Cerro Pachón in Chile, 12% at Gemini North on Mauna Kea in Hawaii, and 20% at the observatory's Hawaii of-

fices. The upgrades were intended to lower operating costs, but they also reduced the facilities' carbon footprint, says Inger Jorgensen, associate director of operations at NOIRLab, NSF's National Optical-Infrared Astronomy Research Laboratory, which comprises several telescopes and other facilities. "By next year they will have paid for themselves," she adds.

In its 2021 request to NSF for a five-year renewal grant, NOIRLab proposed to reduce staff travel by half compared with pre-COVID-19 levels and to use the consequently freed-up \$4.7 million on additional energy-efficient equipment. NSF agreed, and the changes, Jorgensen says, will reduce NOIRLab CO_2e emissions by 30%, from the estimated 8700 tons of CO_2e in 2019 to a target of 6200 tons by late 2027. That reduction "is equivalent to what 500 average US houses emit in a year," she says. "Every little piece makes a difference. And it shows it can be done."

Funding agencies have agency

Eichhorn and others want funding agencies to step in and use their leverage to nudge researchers and institutions to reduce their greenhouse gas emissions. A first step would be for the agencies to require applicants to estimate the carbon footprint of their proposed work. Eichhorn notes that while a growing number of universities and research institutions globally are doing so, the lack of standardization makes it diffi-

cult to compare them. "The day funding bodies say you have to estimate your carbon footprint, everyone will do it," says Lannelongue. "I haven't seen compulsory estimates yet, but things are moving in that direction."

Funding agencies could also reward proposals that include ways to reduce emissions. One incentive, suggests Johnson, could be to recognize institutions that behave sustainably—along the lines of the UK's Athena SWAN (Scientific Women's Academic Network) program, which recognizes good practices in advancing gender equality in higher education.

Limiting scientists to listing only one invited in-person talk on grant applications would be an incentive to travel less, says Eichhorn. And conducting all grant or job interviews virtually would likewise reduce travel. Institutions and funding agencies could ease the requirement of taking the cheapest form of transportation to meetings and instead include the carbon budget in such decisions.

Remote conferences reduce emissions by up to 98%, Eichhorn notes. Even selecting a conference location based on where attendees will travel from can reduce a conference's carbon footprint by 20%, she notes. (See *PHYSICS TODAY*, September 2019, page 29.)

Says Eichhorn, "Reducing emissions in the science community will require creativity and culture change."

Toni Feder

A computing hardware approach aspires to emulate the brain

Neuromorphic computing promises energy savings, a deeper understanding of the human brain, and smarter sensors.

Imagine getting the performance of today's supercomputers but drawing just a few hundred watts instead of megawatts. Or computer hardware that can run models of neurons, synapses, and high-level functions of the human brain. Or a flexible patch that could be worn on the skin that could detect serious health disorders before symptoms develop. Those are a few applications that could be enabled by neuromorphic computing.

Today's high-performance computers have a von Neumann architecture, in

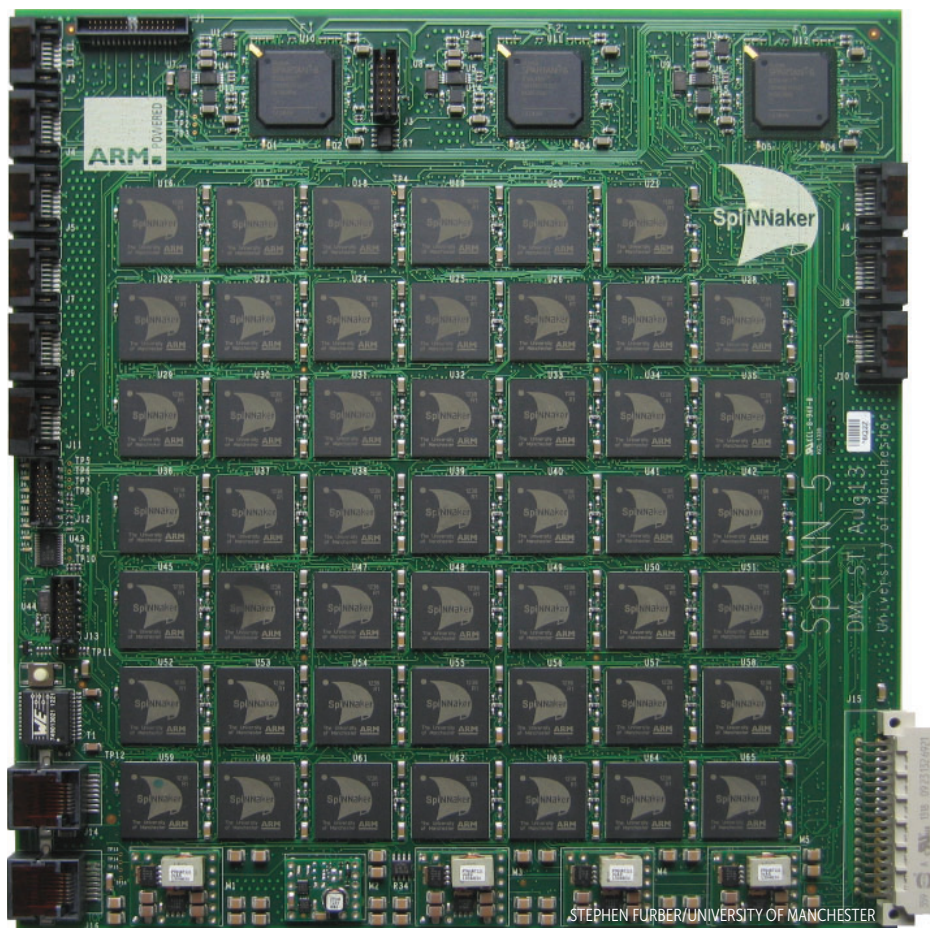
which the central processing or graphics processing units (CPUs and GPUs) are separate from memory units, with the data and instructions kept in memory. That separation creates a bottleneck that slows throughput. Accessing data from main memory also consumes a considerable amount of energy.

In so-called neuromorphic systems, units known as neurons and synapses operate as both processors and memory. Just like neurons in the brain, artificial neurons only perform work when there is an input, or spike, to process. Neuro-

morphic systems are most often associated with machine learning and neural networks, but they can perform a variety of other computing applications.

Just a handful of large-scale neuromorphic computers are in operation today. The Spiking Neural Network Architecture (SpiNNaker) system located at the University of Manchester in the UK has been operating since 2011 and now has 450 registered users, says Stephen Furber, the Manchester computer engineer who led the computer's construction. The UK-government-funded 1-million-core platform was optimized to simulate neural networks.

A next-generation machine, dubbed



CIRCUIT BOARDS containing 48 SpiNNaker neuromorphic chips are at the heart of a 1-million-core computer built at the University of Manchester for the Human Brain Project. The Technical University of Dresden, in collaboration with Manchester, is developing a 10-million-core machine built around the more powerful SpiNNaker2 chip.

SpiNNaker2, is under construction in Dresden, Germany. It's being supported by the state of Saxony and by the Human Brain Project, the European Union's decade-long flagship program, whose goal is advancing neuroscience, computing, and medicine. That program was initiated in 2013 and will end in March. (See *PHYSICS TODAY*, December 2013, page 20.) Based on a more powerful SpiNNaker2 chip, the eponymous computer will consist of 10 million processors, each of which has 10 times the processing and storage capacity of the SpiNNaker chip, Furber says.

Targeted applications for SpiNNaker2 include remote learning, robotics interaction, autonomous driving, and real-time predictive maintenance for industry, says Christian Mayr, an electrical engineering professor at the Technical University of Dresden. Mayr coleads SpiNNaker2 with Furber. SpiNNcloud

Systems, a spin-off company Mayr cofounded to commercialize neuromorphic technology, is in discussions to supply a neuromorphic system to a "large smart city" customer that he declined to identify.

Germany is host to another large-scale neuromorphic platform, BrainScaleS (brain-inspired multiscale computation in neuromorphic hybrid systems) at Heidelberg University. That project also began as a component of the Human Brain Project.

Working with Intel, Sandia National Laboratories plans to complete assembly this spring of a neuromorphic computer consisting of 1 billion neurons. The human brain is estimated to contain 80 billion neurons. "There's a lot of reason to expect that we'll be able to achieve more biological-like capabilities as we get to that scale," says James Bradley Aimone, a Sandia computational neurological scientist.

Sandia has built a 128-million-neuron neuromorphic system, based on Intel's Loihi chip. Each Loihi chip houses 131 000 neurons. The billion-neuron machine will be based on Intel's Loihi 2 chips, which contain 1 million neurons each. (Loihi is named for an active underwater volcano in Hawaii.) The new machine is expected to draw less energy than a high-end workstation typically used for applications such as three-dimensional graphics, engineering design, and data science visualization, says Craig Vineyard, a Sandia researcher.

For Sandia, a nuclear weapons lab that hosts some of the world's largest high-performance computer (HPC) assets, energy savings and Moore's-law limitations are the main attractions of the neuromorphic approach. Conventional supercomputers are power hungry, and the potential to further scale their computational capacity is expected to be held in check by that growing appetite and the inability to further increase processor density, says Vineyard. The world's first exascale HPC, for example, is expected to draw 40 MW, enough power to supply 30 000 homes and businesses, when it begins full operation at Oak Ridge National Laboratory this year. Exascale is at least 10^{18} floating-point operations per second (FLOPS).

"Things like neuromorphic offer a viable path forward, because we can't just keep building larger and larger systems," Vineyard says.

Energy-use comparisons between neuromorphic and classic supercomputing will vary depending on the application. But in some cases, a billion-neuron Loihi system should perform a petascale-equivalent calculation in the same amount of time for as little as 200 W. (Petascale is at least 10^{15} FLOPS.) That's a job for which the most power-efficient supercomputers require 20 kW, says Sandia's Aimone.

Some of today's very large deep-learning neural networks require hours just to train. Deep learning is a subfield of artificial intelligence (AI) that uses brain-inspired algorithms to help computers develop intelligence without explicit programming. The Generative Pre-trained Transformer 3 (GPT-3) language

model, for example, can generate text that is difficult to distinguish from that of a human. Training it is estimated to require more than 1 GWh. Equal or greater amounts of energy are consumed when deep-learning models are put to use. The human brain, with its vastly greater computing capacity, operates on 20–30 W, says Mayr.

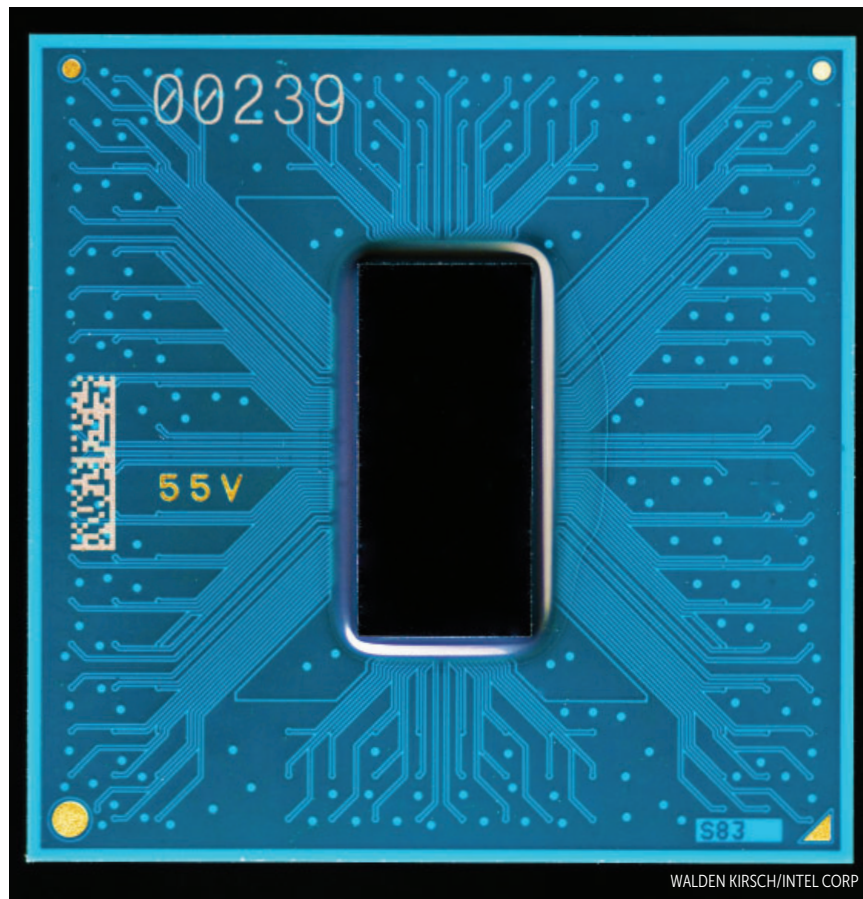
Unparalleled parallelism

Apart from energy savings, proponents of neuromorphic computing say it can offer equal or faster performance over classical HPCs for some applications. The number of processors in even the most powerful HPC machines pales in comparison with hundreds of millions of simulated neurons, though each neuron is far less computationally powerful than a GPU or CPU. Neuromorphic's unrivaled level of parallelism is well suited for calculating certain kinds of algorithms, such as Monte Carlo random-walk simulations, says Aimone. Those algorithms are used in modeling molecular dynamics in drug discovery, stock-market predictions, weather forecasts, and a host of other applications.

"Is it possible to spread out this exploration of where a stock price may go over the large population of neurons? It turns out that you can," Aimone says. Sandia demonstrated that a neuromorphic simulation of how radiation diffuses through materials performed on a Loihi system was nearly as fast as one accomplished on a CPU or GPU platform and at far lower energy cost.

"From the algorithm side, we've recognized that neuromorphic systems provide computational advantages, but that only becomes apparent at large scale," says Aimone. That's partly due to the overhead associated with setting up the machine to solve a new problem, adds Vineyard. "It's not a magic solution, so research is identifying where those advantages are."

Neuromorphic computers won't pose a threat to HPCs, says Rick Stevens, associate director for computing, environmental, and life sciences at Argonne National Laboratory. "There are a handful of examples that have been demonstrated where you can do interesting problems. But it's nowhere near a general-purpose platform that can replace a conventional supercomputer." Neuromorphic hardware is particularly well suited to simulate



INTEL'S LOIHI second-generation neuromorphic chip was unveiled in September 2021. With up to 1 million neurons per chip, it supports new classes of neuro-inspired algorithms and has 15 times the storage density, faster processing speed, and improved energy efficiency compared with the predecessor Loihi chip. Sandia National Laboratories is building a 1-billion-neuron neuromorphic computer based on Loihi 2 architecture.

computational neuroscience problems, he says, "because that's the computational model it's directly implementing."

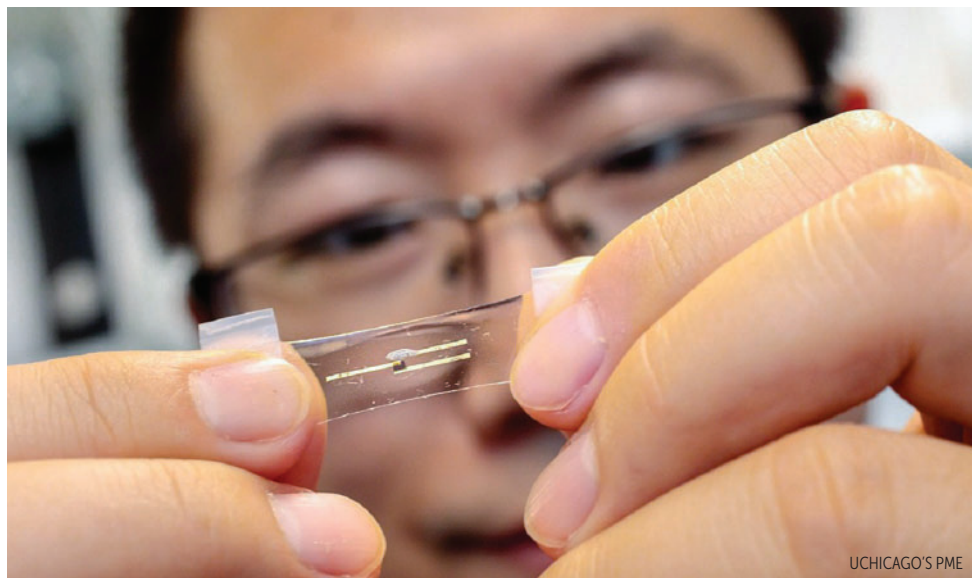
But for the hundreds of applications that general-purpose supercomputers can perform, the alternative hardware has yet to show it offers advantages or even works well, Stevens says. And there are companies building specialized accelerators for deep learning that can implement abstractions of neurons without making any claims about being neuromorphic.

Comparing neuromorphic computing power with that of HPCs is not straightforward. "They are different animals," says Furber. Each of SpiNNaker's processors, for instance, is capable of delivering 200 million instructions per second, so the million-core machine can deliver 200 trillion instructions per second. HPCs are measured in FLOPS, but the Manchester machine has no floating-

point hardware, he says.

"[Neuromorphic] is kind of similar to quantum in that it's a technology waiting to prove itself at scale," Stevens says. "Loihi is a great research project, but it's not at the point where commercial groups are going to deploy large-scale versions to replace existing computing."

At the opposite extreme from supercomputers, neuromorphic computers may benefit so-called edge-computing applications where energy conservation is a must. They include satellites, remote sensing stations, weather buoys, and visual monitors for intrusion detection. Instead of sending data on a regular clock cycle, spiking neuromorphic smart sensors would transmit only when something is detected or when a threshold value is crossed. "It should be smart about what it collects, what it transmits, and wake up if something is going on," says Aimone. "That requires computation." Adds



A SKIN-LIKE SENSOR developed by Argonne National Laboratory and the University of Chicago features stretchable neuromorphic electronics. The technology could lead to precision medical sensors that would attach to the skin and perform health monitoring and diagnosis. Holding the device is the project's principal investigator, Sihong Wang.

Stevens, "You're trying to go from a sensor input to some digital compact classification or representation of what that sensor is doing."

In November, Argonne and the University of Chicago's Pritzker School of Molecular Engineering announced the development of a skin-like wearable patch featuring flexible and stretchable neuromorphic circuitry. If developed further, such wearable electronics hold promise for detecting possible emerging health problems, such as heart disease, cancer, or multiple sclerosis, according to a lab press release. Devices might also perform a personalized analysis of tracked health data while minimizing the need for their wireless transmission.

In one test, the research team built an AI device and trained it to distinguish healthy electrocardiogram signals from four different signals indicating health problems. After training, the device was more than 95% effective at correctly identifying the electrocardiogram signals.

A tall order

In Europe, a major motivation for neuromorphic R&D has been to improve understanding of how the brain works, and that's no small task. "First and foremost the goal is fundamental research to see if we can learn from biology a different way of computing, and if this alternative

way of computing can help in neuroscience as a research platform," says Johannes Schemmel, a Heidelberg University researcher who heads BrainScaleS.

"We have a very massive neural network in the brain, but it's on the scale of 99% idle," says John Paul Strachan, who leads the neuromorphic compute nodes subinstitute at the Peter Grünberg Institute at the Jülich Research Center in Germany.

SpiNNaker and Loihi systems are fully digital. But BrainScaleS is a hybrid: It has analog signals for emulating individual neurons and digital ones for communications among neurons. "We've developed electronic circuits from transistors that behave similar to neuron synapses in the biological brain," says Schemmel. "They are all continuous analog quantities."

At higher levels, however, "we use digital communication between the neurons, because in principle there is no real analog communication possible," Schemmel says.

The brain is highly sparse: When a neuron fires in response to a stimulus, its signal is transmitted only to the thousands of other neurons it connects to, not to the billions of others in the brain. Sparsity is critical to brain function. "If all our neurons were firing and communicating, we'd heat up and die," says Strachan. But working with sparse data

isn't an ideal fit for HPCs. "If the hardware has been designed to optimize for dense computations, it will be idle or doing a bunch of multiply-by-zero operations," he says. That means that simulating one second of just a tiny portion of the brain on an HPC today requires minutes of processing time.

The brain has various mechanisms to keep processing to the absolute minimum required, says Mayr, "but AI networks do a lot of irrelevant stuff. Take a video task, where every new frame of a video only contains maybe 2–3% new information, and even that can be compressed. All the rest is rubbish; you don't need it. But with a conventional AI HPC chip-based approach, you have to compute all of it."

"One of the biggest research tasks in computational neuroscience is to merge the function of the brain with what works in AI. This is the holy grail," says Schemmel. Ideally, AI training would use localized learning rules that work at the level of neurons and synapses. Such rules could also permit robots to learn without needing to be uplinked to computers.

Limited funding is a restraint on the mostly academic field. Hardware is expensive to build, and it's not surprising that Intel has the most advanced neuromorphic chips, says Schemmel. Algorithms are needed, and their development lags the state of hardware. "We now have large-scale platforms that can support spiking networks at bigger scales than most people can work out what's useful to do with them," says Furber.

Interchangeability is another issue. "We haven't come up yet with a neat software framework that the neuromorphic guys can all subscribe to," says Mayr. "We need to standardize a lot more."

Yet perhaps the greatest limitation on R&D is a paucity of trained scientists, particularly computational neuroscientists. "Students have to have knowledge on a lot of different levels. You need longer to train; you can't compartmentalize problems like you can do in software engineering nowadays," says Schemmel.

David Kramer 