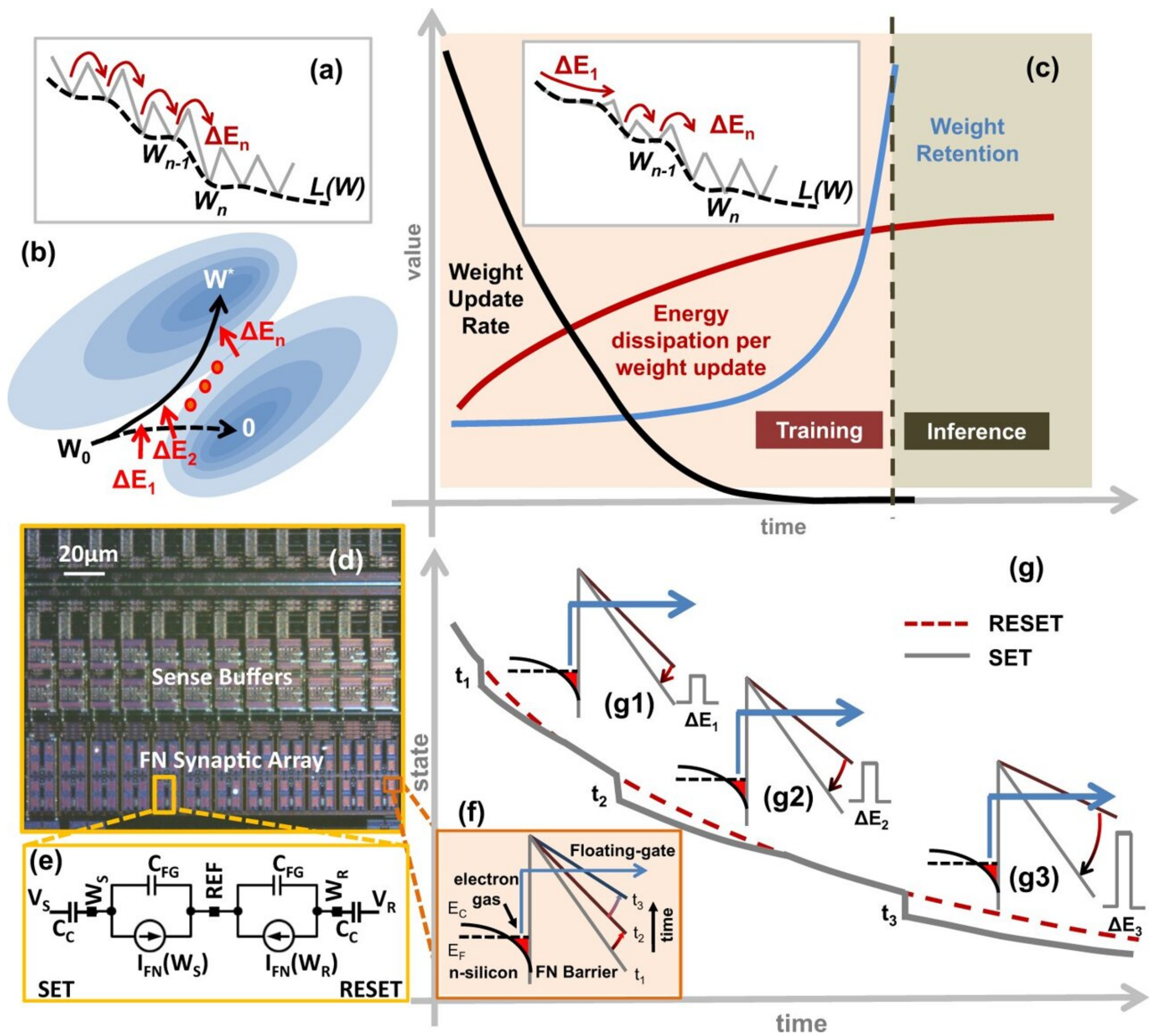
A nature-driven solution for more efficient AI

April 13 2022, by Brandie Jefferson



Motivation and principle of operation for the proposed synaptic memory device. (a) Conventional non-volatile analog memory where transition between analog static states (Wn−1, Wn) dissipates energy (ΔEn); (b) dynamic analog memory where an external energy (ΔE1, ΔE2… ΔEn) is used to modulate the trajectory of the memory states (W0) towards the optimal solution (W“); (c) desired analog synapse characteristic where the memory retention rate is traded-off with the write energy; reducing the energy dissipation per weight update in training phase by matching the dynamics of the dynamic analog memory to the weight decay as shown in (inset) where the height of the energy barrier (ΔEn) increases as the training progresses (d) micrograph of a fabricated DAM array along with (e) its equivalent circuit where the leakage current IFN is implemented by (f) the electron transport across a Fowler–Nordheim (FN) tunneling barrier; (g) implementation of the FN tunneling based DAM where dynamic states g1–g3 determines the energy dissipated (ΔE1, ΔE2, ΔE3) per memory update at time instance t1–t3 and memory retention rate. Credit: *Nature Communications* (2022). DOI: 10.1038/s41467-022-29320-6

Over its lifetime, the average car is responsible for emitting about 126,000 pounds of the greenhouse gas carbon dioxide (CO_2).

Compare those emissions with the carbon footprint left behind by artificial intelligence (AI) technology. In 2019, training top-of-the-line artificial intelligence was responsible for more than 625,000 pounds of CO_2 emissions. AI energy requirements have only gotten bigger since.

To reduce AI's energy footprint, Shantanu Chakrabartty, the Clifford W. Murphy Professor at the McKelvey School of Engineering at Washington University in St. Louis, has reported a prototype of a new kind of computer memory. The findings were published March 29 in the journal *Nature Communications*.

The co-first authors on this article are Darshit Mehta and Mustafizur Rahman, both members of Chakrabartty's research group.

A disproportionate amount of energy is consumed to train an AI, when the computer searches different configurations as it learns the best solution to a problem, such as correctly recognizing a face or translating a language.

Because of this energy use, most companies can't afford to train a new AI from scratch. Instead, they train it "enough" and then maybe tweak some parameters for different applications. Or, if a company is big enough, Chakrabartty said, they'll move their data centers to a more convenient, waterfront property.

All that energy use creates a lot of heat and needs a lot of water to keep cool. "They are boiling a lake, practically, to build a neural network," he said.

Instead of boiling, Chakrabartty's research group turned to quantum tunneling.

When a computer searches for an answer, the system is using electricity to flip billions of tiny switches "on" and "off" as it seeks the shortest path to the solution. Once a switch is flipped, the energy is dumped out and additional energy is used to hold the switch in position, or hold it as memory. It's this use of electricity that creates such a large carbon footprint.

Instead of feeding a constant stream of energy into a memory array, Chakrabartty lets the physical memory do what it does in the wild.

"Electrons naturally want to move to the lowest stable state," said Chakrabartty, who is also vice dean for research and graduate education in the Preston M. Green Department of Electrical & System Engineering.

And electrons do so using the least amount of energy possible. He uses that law of physics to his advantage. By setting the solution (say, the recognition of the word "water" as "agua,") as a stable state, electrons tunnel toward the right answer mostly on their own with just a little guidance, i.e. direction from the training algorithm.

In this way, the laws of nature dictate that the electrons will find the fastest, most energy-efficient route to the answer on their own. At the ground state, they are surrounded by a barrier large enough that the electrons almost certainly will not tunnel through.

Whereas modern memory uses a more brute-force approach by recording its route into memory and using energy at each flipped switch, Chakrabartty's learning-in-memory design simply lets electrons do what they do without much interference and hardly any additional energy.

Once they've reached the final barrier, the AI is said to have learned something.

"It's like trying to remember a song," he said. "In the beginning, you're searching your memory, looking everywhere for the song. But once you've found it, the memory is now fixed, and then you can't get it out of your head."

More information: Darshit Mehta et al, An adaptive synaptic array using Fowler–Nordheim dynamic analog memory, *Nature Communications* (2022). DOI: 10.1038/s41467-022-29320-6

Provided by Washington University in St. Louis