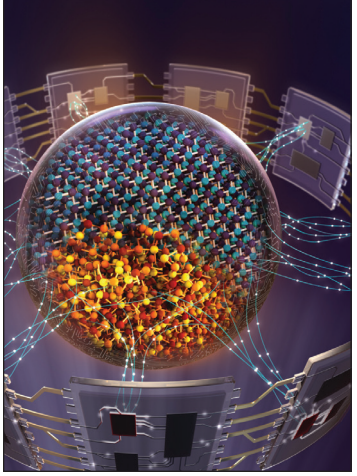


Computing power demand is expected to grow exponentially as millions in developing countries go online, along with a burgeoning number of Internet-connected devices.

Materials opportunities and challenges for low-energy computing: Devices

By **Prachi Patel**
Feature Editor: **Subhash L. Shinde**

Computing has progressed at a mind-bending pace in the past decades. As the world gets more digitally connected, and the amount of data processed, shared, and stored explodes, so do the energy use and environmental impact of computing.

Already, computing uses 5% of the electrical power the world consumes. Much of that goes to factory-sized server farms that store digital data from billions of smartphones and tablets as people stream movies, share photos and videos, and send emails. "These things don't come for free in terms of energy use," said Asif Khan, a professor of electrical and computer engineering at the Georgia Institute of Technology.

And the energy cost of computing is in for a steep increase. "Probably a third of the world's population has access to high-speed Internet now," he said. "Fast forward 10 years, and that number will more than double." Computing power demand is expected to grow exponentially as millions in developing countries go online, along with a burgeoning number of Internet-connected devices: phones, cars, robotic vacuum cleaners, and smart TVs. A recent study projects that the information and communications technologies industry could produce up to 5.5% of the world's carbon emissions by 2025 and 14% by 2040. "Unless we do something radical with our devices to curb energy use," said Khan.

During the past five decades, the semiconductor industry has gained performance and efficiency by shrinking silicon-based CMOS (complementary metal oxide semiconductor) devices, as governed by Moore's Law, which predicts the number of transistors on a chip doubling every two years while the costs halve. But, said Vijay Narayanan, IBM Fellow and senior manager, PCM & AI Materials, at the IBM T.J. Watson Research Center, "the amount of data being used has gone up exponentially, and at the same time, Moore's Law scaling has slowed down." For more powerful computers that use less energy per bit of calculation, researchers are exploring new materials technologies that can take computing beyond CMOS.

To understand why computing is energy-intensive, zoom in on one of the billions of transistors that flip digital zeroes and ones to perform the logic operations that happen behind the scenes as people map routes or upload photos. Traditional transistors are planar three-terminal devices with a silicon channel capped at either end with the source and drain, and a dielectric gate on top bridging the two. The voltage at the gate controls the flow of current between the source and drain on the order of milliamps. Each transistor requires 1 V to operate, using a million times more energy for that one-bit calculation as it theoretically should, based on a thermodynamic calculation carried out by Rolf Landauer, IBM, in 1961. Much of the electricity is converted into heat. "So at the device level, we're still stuck with a fundamental energy limit," said Khan.

This is a problem as chips get exponentially denser. The iPhone 10's microprocessor has approximately 7 billion transistors, twice as many as the iPhone 8. While the transistors have shrunk, their power efficiency has not improved at the same pace. The clock frequency of laptops increased exponentially from the 1990s until 2005, reaching 3 GHz, but has been stuck there since, because running processors faster generates too much heat. Not only would more energy-efficient computing save power and money, it is crucial to keep advancing the information revolution.

Materials advances have already helped. Swapping the traditional silicon dioxide dielectric at the gate with hafnium dioxide helped the industry continue scaling down transistors without compromising performance. As the gate became thinner, hafnium dioxide, with its higher dielectric constant, held back electrons that would leak through silicon dioxide, creating a power loss.

Chip makers are now dramatically changing device architecture to minimize power dissipation as material layers become ultra-thin. The big push, as was evident from the latest IEEE International Electronic Devices Meeting (IEDM) in December 2019, has been for three-dimensional devices.

In the early 2010s, semiconductor companies switched to FinFETs (field-effect transistors), which have a fin-like silicon channel sticking up from the surface with the gate wrapped around its three sides. Today's 7-nm transistors use this design, which reduced operating voltage from 1 V to 0.75 V. Major industry players are working on the next device architecture that will reduce device size to 3 nm.

The horizontally stacked nanosheet transistor design consists of nanometers-thick silicon sheets laid on their side fully wrapped by the gate dielectric, reducing leaks and offering better control. "The benefit is that it offers more than 25% performance improvement at the same power or more than 50% power saving for the same performance," said Narayanan.

But making nanosheets isn't easy. It involves removing material between layers of other material and then filling the gaps with metal and dielectric. IBM, Samsung, and others use an epitaxial reactor to make a superlattice—a periodic structure of alternating layers of two materials, in this case silicon-germanium (SiGe) and silicon.

Subhash L. Shinde, University of Notre Dame, USA
Prachi Patel, prachi@lekh.org

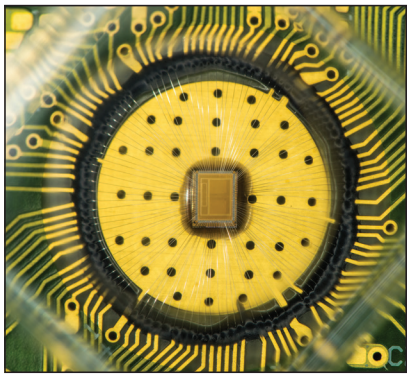
https://doi.org/10.1557/mrs.2020.70 Published online by Cambridge University Press

The SiGe is etched out, and then atomic layer deposition is used to make an atoms-thick layer of dielectric and metal around the sheets.

Scaling down devices has its limits, though, said Khan. "We need to fundamentally alter the way transistors operate." He and others are trying to reinvent the transistor by introducing new materials with new physical properties, specifically using ferroelectric materials that show negative capacitance. Capacitors store charge in response to a voltage. The stored charge creates an electric field that, in conventional capacitors, opposes the voltage. But in negative capacitance ferroelectrics, the electric field aligns with the voltage, boosting it.

Replacing a transistor's gate insulator with a ferroelectric could amplify the gate voltage, enabling low-power operation. Purdue University researchers proposed the idea in 2008. It was slow to catch on, but all major companies are now pursuing ferroelectrics. So far, the materials have needed thick layers to work and have slowed down devices. The challenge is finding the right material and practically integrating it.

The good news is that hafnium dioxide, which is already used as the gate dielectric in today's transistors, fits the bill. But while it is used now in amorphous form, it needs to be crystalline for ferroelectricity. Researchers are working on thinning it down, controlling its grain size, or mixing it with other materials.



Phase-change memory for next-generation computing. Courtesy: IBM Research.

Zooming out again from devices and looking ahead, significantly denting computing's energy use will require radically changing the way computers work, especially as artificial intelligence (AI) finds its way into smartphones and self-driving cars. Traditional von Neumann architecture separates logic and memory, expending power to move data between the two. With data-intensive AI, that power consumption is daunting.

The industry is pursuing new computer architectures that bring together computations and memory and that could enable brain-like, or neuromorphic, computing. "There is a lot of material innovation and process ingenuity needed to adapt to these new architectures," said Narayanan.

Take, for example, the task of finding a dog in a picture. Neural networks accomplish this by first being trained on a labeled data set, forming connections, much like synapses in biological neurons. Each connection carries a weight, so a neural network can have billions of weights. "In normal computing, you'd move both the weighted matrix and new data into the logic part, cross the two with a matrix operation, and then move back to memory," said Matthew Marinella, who conducts brain-inspired computing research at Sandia National Laboratories. Moving the data from memory to processor takes up to 100 times more energy than running the computation itself.

Researchers and companies are trying to map entire neural networks into memory arrays. Encoding the weights of the neural networks in the memory elements would allow in-memory computing that is faster and uses lower power.

The main requirement for this is a nonvolatile memory made of materials that can hold multiple states. There is a broad range of options to choose from, and researchers seem to come up with new ones every few months. "We're evaluating a whole suite of materials," said Narayanan. "Foremost is phase-change memory." That typically consists of germanium, antimony, and tellurium-based materials that transition from amorphous to crystalline under electric pulses, during which the material can have different conductance values.

They and others are also looking at resistive random-access memory (RRAM) devices, simple metal–insulator–metal structures that store data as different levels of resistance. People have explored oxides of hafnium, nickel, titanium, and vanadium for these devices, which are also called memristors. In another type of RRAM, microscopic conductive paths called filaments can be created or broken in the medium by an applied voltage. RRAM maker Crossbar, Inc. is commercializing this type of memory, which can be built right on top of silicon-integrated circuits. The devices are made of layers of tungsten, amorphous silicon, and silver; applied voltage creates a silver filament that spans the amorphous silicon to reach the tungsten and make the cell conductive.

Ferroelectrics have recently attracted attention for nonvolatile memory. Ferroelectrics become polarized in an electric field and retain the polarization when the field is removed. Ferroelectric perovskites such as $PbZrO_x$, $SrBi_2Ta_2O_9$, $BaTiO_3$, and $BiFeO_3$ have all been investigated, but they face issues, including thickness, CMOS compatibility, and reliability. Some researchers are looking at ferroelectric perovskites that can hold three or four stable polarization states.

IBM researchers are now working on a novel memory element called electrochemical RAM (ECRAM) that can achieve dozens of states. Unveiled at the 2018 IEDM, the ECRAM looks like a typical transistor, except it uses tungsten trioxide instead of silicon, and the dielectric is lithium phosphorous oxynitride, a solid electrolyte used in experimental thin-film lithium-ion batteries. A current at the gate drives lithium ions into the tungsten. "Think of it like a battery," said Narayanan. "We send in some kind of ion into a substrate, and that will change resistance, and depending on how the ions are intercalated, you can get a spectrum of conductances."

Meanwhile, at the 2019 IEDM, Purdue University's Peide Ye and his colleagues reported the first device that can both process and store information. They turned to a recently discovered ferroelectric, alpha indium selenide, which is also a semiconductor instead of an insulator, such as traditional ferroelectrics. Using it, they made transistors that can store data as well as form memristor-like devices.

As Moore's Law comes to an end, and the energy demand of AI-driven computing escalates, it's clear that the industry will have to adopt new paths forward with the help of new materials and methods. "It is a bit of chaos right now," said Marinella. "It'll be interesting to see which way the industry heads." □

https://doi.org/10.1557/mrs.2020.70 Published online by Cambridge University Press