MEMORY UPGRADE

The demands of modern computing call for a seismic shift in data storage and retrieval

By Andrew Grant

Google operates data centers at 13 sites globally, including this server farm in Hamina, Finland. Storing and processing so much data requires loads of energy and a dedicated cooling system. At Hamina, seawater from the Gulf of Finland cools the computers. Creating new kinds of computer memory could cut the demand for energy and make searching data faster.

CONNIE ZHOU

Ramamoorthy Ramesh listens to Indian classical music on his smartphone, which is jammed with videos of his kids' soccer games. He streams Netflix movies on his tablet and, on his laptop, he uses Google to search the Internet several times a day. Like many of us, he's an active consumer of data in a data-centric world.

But Ramesh is also a materials scientist who has a thorough understanding of what's going on under the hood of his electronic devices, and he has a lingering concern: "The computer is very advanced, but it's not close to where it should be."

The problem, he says, is that today's users rely on computers that are much better at computing than at storing and recalling information. At the heart of every computer is a processor that carries out programmed instructions at blazing speeds so users can pay bills online, find a nearby Italian restaurant and post selfies on Instagram. But the processor also needs a place to store the results of its efforts for use milliseconds or years in the future. No existing memory technology can do both: keep up with the processor and store information for long periods of time. Modern memory devices, including random access memory (RAM), hard drives and flash, are some combination of too slow, too expensive and too energy-hogging.

This performance gap between processor and memory has existed since the first electronic computers were introduced more than a half-century ago. But those machines weren't asked to find obscure facts on the Internet, sort through patients' medical histories and mine personal profiles on social networks. The amount of data globally is expected to grow eight times larger within five years, according to IBM, and 90 percent of today's data is less than 2 years old. The era of Big Data has arrived.

For computers to successfully navigate through the barrage of superfluous data, Ramesh and a host of engineers and physicists believe they need to develop a next-generation memory device. They call it storage-class memory: a single gadget that combines speed, energy efficiency and high-density storage.

Access to storage-class memory would lead to smarter, faster mobile devices with better battery life. In the long run, storage-class memory could revolutionize the way computers work, Ramesh says, enabling processor-memory hybrids that compute and remember at the same time, kind of like the human brain. It would be the first makeover of computers' fundamental architecture since the 1940s, when the word transistor first entered the lexicon.

With so much at stake, technology industry giants such as IBM, Samsung and Hewlett-Packard, along with innovative smaller outfits like Crossbar and Micron, are spending billions of dollars to probe the bit-storing potential of tiny magnets, amorphous solids and miniature grids of wire. It's a competitive game full of hype and tricky science, yet a steady stream of advances suggests that storage-class memory may soon catch up and meet the lofty performance standards of the processor.

"Everybody is on this like gangbusters because their business is at risk," says Ramesh, who researches storage-class candidates at the University of California, Berkeley. "There's definitely going to be a pathway to storage-class memory. The question is: Which technology will take us there?"

Same old architecture

The modern computer as we know it emerged in 1945, when mathematician John von Neumann penned his "First Draft of a Report on the EDVAC." The electronic computer he envisioned centered on a processor that could perform hundreds of calculations per second. But if the processor were the master chef, it needed recipes (instructions to tell it what to do) as well as a place to store ingredients (data to be calculated) and keep finished meals (the results of the calculations) warm. Von Neumann assigned those responsibilities to memory, which at the time came in the form of magnetic tape and tubes of mercury.

Von Neumann's machine, which went live in 1951, took up more floor space than a thousand iPads set side by side and outweighed an African elephant. It had only a few kilobytes of memory, but that was plenty because the processor worked so slowly.

Things got tricky, however, once processors sped up to undertake thousands, millions and then billions of calculations per second. No memory device could both exchange data with the processor billions of times a second and retain masses of information indefinitely. So engineers devised a solution: The fastest memory, which was also the most expensive, interacted directly with the processor and stored small amounts of the most urgent data. Information for the more distant future was relegated to cheaper, higher-capacity memory devices. By creating this memory hierarchy, engineers managed to keep von Neumann's basic architecture — memory that stores data plus instructions for the very busy processor — intact. "Today's computers would still be recognizable to von Neumann," says Neil Gershenfeld, a physicist and computer scientist at MIT.

In modern computers the processor's main helper is dynamic RAM, or DRAM, a chip that provides short-term, easily accessible information storage. Each DRAM cell consists of a capacitor that stores electrical energy and a transistor that serves as a swinging gate, controlling the flow of electricity to rapidly switch the capacitor between a charged state, which represents a 1, and uncharged, a 0.

DRAM, however, has an Achilles' heel: Capacitors can't hold electricity for very long. As a result, the DRAM chip requires an influx of energy 15 or so times a second to refill the capacitors. That continual need for a refresh means that the computer has got to be on for DRAM to function. It is no good for long-term storage.

Most systems use a hard disk drive for long-term memory. These drives use mechanical arms that write and read data onto cells on 3.5-inch-wide circular platters; the direction of magnetic orientation in each cell determines whether it is a 1 or a 0.

Hard drives are cheap and can store enormous amounts of data, but they are slow. It takes about 5 milliseconds for a bit (a 1 or a 0) from the processor to get stored on the disk — 5 million times as long as it takes the processor to do a calculation. In human terms, that's like a restaurant patron (the processor) deciding what to order and a waiter (the hard drive) needing more than a month to jot the order down. Forget about getting dessert.

On a more practical level, this lag explains why many computers take a couple of minutes to boot up when powered on: The operating system needs time to migrate from the hard drive to DRAM where the processor can access it.

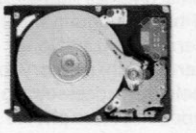
Engineers have spent decades trying to bridge the speed gap between processor and memory, and in 1988 computer chip giant Intel took the first step when it unveiled flash memory. Flash retains information when unpowered and can store data in cells 20 nanometers wide, enough to stockpile thousands of photos on a digital

Memory timeline

Electronics use memory technology invented last century. Engineers have squeezed lots more performance out of these devices, but the push is on to develop game-changing storage.

1945
John von Neumann publishes memory-intensive concept for EDVAC

1956
IBM introduces magnetic hard disk drives



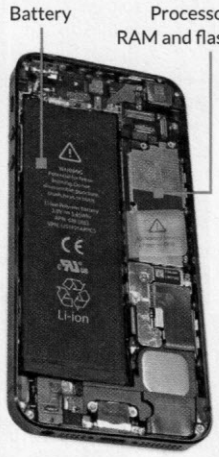
1967
IBM invents DRAM

1988
Intel introduces flash

2008
HP announces invention of memristor

2012
Micron Technology starts selling phase-change memory

FROM TOP: LANL/WIKIMEDIA COMMONS; VIKTORUS/ISTOCKPHOTO; APPALOOSA/WIKIMEDIA COMMONS; DNY59/ISTOCKPHOTO



A look inside Smartphones like Apple's iPhone 5 contain short-term (RAM) and long-term (flash) memory to help the processor run apps and store photos and music for later use. Storage-class memory would increase speed and capacity as well as extend battery life.

camera and hundreds of apps on a smartphone. It is also relatively fast (at least compared with a hard drive), so smartphones, laptops and tablets with flash memory boot up much faster than computers with a mechanical hard drive.

Intel claims that it can continue making flash cheaper and faster by shrinking and stacking memory cells. But computer scientists such as Darrell Long of the University of California, Santa Cruz say flash is nearing its performance limit. The time for a faster, cheaper and more energy-efficient replacement is now.

And not just because flash could be close to maxing out. The demands placed on computers today are vastly different than in the past. Computers with the von Neumann hierarchy excel at taking a set of data, modifying it in some way and then placing it back into memory; data processing takes precedence over the actual contents of the data. Now the bigger challenge is finding jewels and trends in vast amounts of largely non-essential data. "Instead of crunching numbers for a bank, we're trying to find an answer to a question," says Paul Franzon, an electrical engineer at North Carolina State University in Raleigh. Computers need to be able to swiftly store and analyze large datasets.

Motivated by these factors, researchers began about a decade ago to hunt for a type of memory that combines the swiftness of DRAM with the capacity and longevity of disk drives.

The right material

The search for a breakthrough memory device begins in the labs of materials scientists and condensed matter physicists. Any material used to build the next-generation memory device will have to effectively distinguish between 1s and 0s by having two distinct electrical or magnetic states.

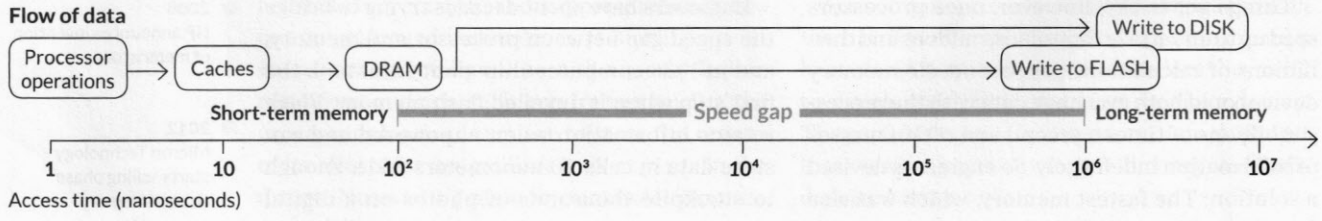
Ferroelectric RAM, or FRAM, functions like DRAM, with a kick. Each memory unit has a capacitor to store electricity and a transistor to switch between 1 and 0. FRAM's added benefit is that its capacitors are made from materials such as lead zirconate titanate and bismuth ferrite, which can hold charge without needing constant refreshes. "A decade ago, people thought it was very straightforward: FRAM would win the race for storage-class memory," Ramesh says.

But FRAM has some glaring weaknesses. While ferroelectric materials make great capacitors, they do not integrate easily with other components made of silicon. "You can't put FRAM on a silicon wafer directly," Ramesh says, making cheap manufacturing a challenge. Scientists are also concerned about the reliability of FRAM over the long run, though Ramesh and colleagues recently developed a technique that allowed FRAM chips to read and write data millions of times without any signs of degradation (*SN: 7/13/13, p. 11*). That's about 1,000 times better than flash and would allow most users to safely store data for decades.

Hewlett-Packard claims that its next-generation memory device is 100 times faster than flash and can hold at least twice as much data. Plus its device, made of titanium dioxide, gets along just fine with silicon. The HP memristor, short for memory resistor, changes its electrical properties depending on the direction of the current going through it and then remembers those charges when the power is off. HP made headlines in May 2008 when a team led by Stanley Williams introduced the memristor in *Nature* and demonstrated its potential for fast, high-capacity storage (*SN: 5/24/08, p. 13*). Still enthusiastic in 2010, Williams told an HP publication: "We believe that the memristor is a universal memory that over time could replace flash, DRAM and even hard drives."

HP won't say when it will bring the memristor to market, but it could happen as early as next year. Ramesh says HP has to address concerns about the device's long-term reliability. Meanwhile, in August a small company called Crossbar, in Santa Clara, Calif., announced that it has

Closing the gap Computer performance has skyrocketed over the decades, yet memory is still a long way from keeping up with the processor, which can run through billions of 1s and 0s per second. The processor relies on DRAM and expensive low-capacity devices called caches to store the most urgent data. Everything else gets passed on to long-term storage such as flash and hard disk drives, but these work at a fraction of the processor's speed. Researchers want to create storage-class memory that combines the cost and robustness of hard disks with the speed of DRAM. SOURCE: GEOFFREY W. BURR



FROM TOP: BRITTANY MCCRIGLER; M. ATAROD

developed a similar type of fast memory called resistive RAM. The company claims to have produced a commercially viable postage stamp–sized chip that can hold a terabyte of data (that's 10^{12} bytes), though it hasn't announced when its product will be available for sale.

Phase-change memory is already finding its way into electronics. The device is made of a compound of germanium, antimony and tellurium with electric properties that change depending on its temperature: It can behave as a normal solid or an amorphous, flowing substance. The idea is to melt or solidify the compound depending on whether the memory is storing a 1 or a 0. Though some researchers worry that the device requires too much energy to repeatedly heat and melt the compound, Micron Technology, headquartered in Boise, Idaho, began selling rudimentary phase-change memory for basic cell phones last year.

There are other storage-class memory technologies in play too. Samsung is working on spin-transfer torque RAM, which uses electric currents to shift the magnetic orientation of a thin layer of material. And IBM is exploring racetrack memory, which relies on current darting through a tiny grid of wire to manipulate even smaller magnetic cells that can switch between 1 and 0.

All of these memory upgrades face technical challenges, but there are economic ones as well. Manufacturers have churned out hard drives, flash and DRAM for years, and they won't rush to adopt a risky technology. "Modern semiconductor development is extremely expensive," says Geoffrey Burr, who studies storage-class memory at IBM's Almaden Research Center in San Jose, Calif. Companies will invest, he says, only if it's almost definite that the technology would work as expected and sell in large numbers.

Smarter devices

The road to creating storage-class memory has been bumpier than most researchers expected, but their eyes are still on the prize. They know that when storage-class memory finally makes it to market, life will change for consumers and businesses.

"You could have a terabyte of memory in your mobile device," Franzon says, or 30 to 60 times as much storage as most current smartphones. "It would dramatically change the user experience." For example, he says, people could store thousands of movies on their phones rather than having to stream them online.

Better storage-class memory is about much more than upgrading smartphones, however. Tech giants like Google and Facebook operate vast data centers that use plodding hard drives and power-hogging DRAM chips to store and analyze petabytes — more than a million billion 1s and 0s — of search terms, likes and relationship statuses. The energy costs for these mammoth facilities are steep; they require their own power plants and cooling facilities to keep them humming. In 2010, Google's servers used 2.3 million megawatt-hours of energy, enough to supply 200,000 homes for a year. Replacing hard drives and DRAM with storage-class memory would speed up servers and slash their energy needs.

30–60
times more storage

What 1 terabyte of memory would mean for smartphone users

Plenty of other Big Data users would benefit as well. Doctors could quickly sift through medical records and studies to diagnose patients and prescribe treatments. Scientists could look for patterns in genetic sequences as well as astronomical images. (The Large Synoptic Survey Telescope in Chile, scheduled to begin scanning the skies within a decade, is expected to produce 30 terabytes of data — equivalent to about 4 million high-quality photographs — per night.) And government defense agencies would surely love a computer that rapidly ferrets out terrorist networks and identifies threatening messages.

Some computer scientists say that the real fun will begin once a storage-class memory device — whether souped-up flash, memristors, phase-change or an untapped mystery material — rises above its competitors. Ramesh contrasts the hierarchical approach of present-day computers, which masks the shortcomings of memory, with the complex multitasking that takes place in the human brain. If engineers can finally build memory that works in tandem with the processor, he says, then they can think about creating devices that compute and recall at the same time. Such a seismic shift would further optimize computers to do the jobs we ask of them and, perhaps finally, lead to a machine that even a visionary like von Neumann wouldn't recognize. ■

Explore more

- John L. Hennessy and David A. Patterson. *Computer Architecture*. Elsevier, 2012. bit.ly/1bpwe06