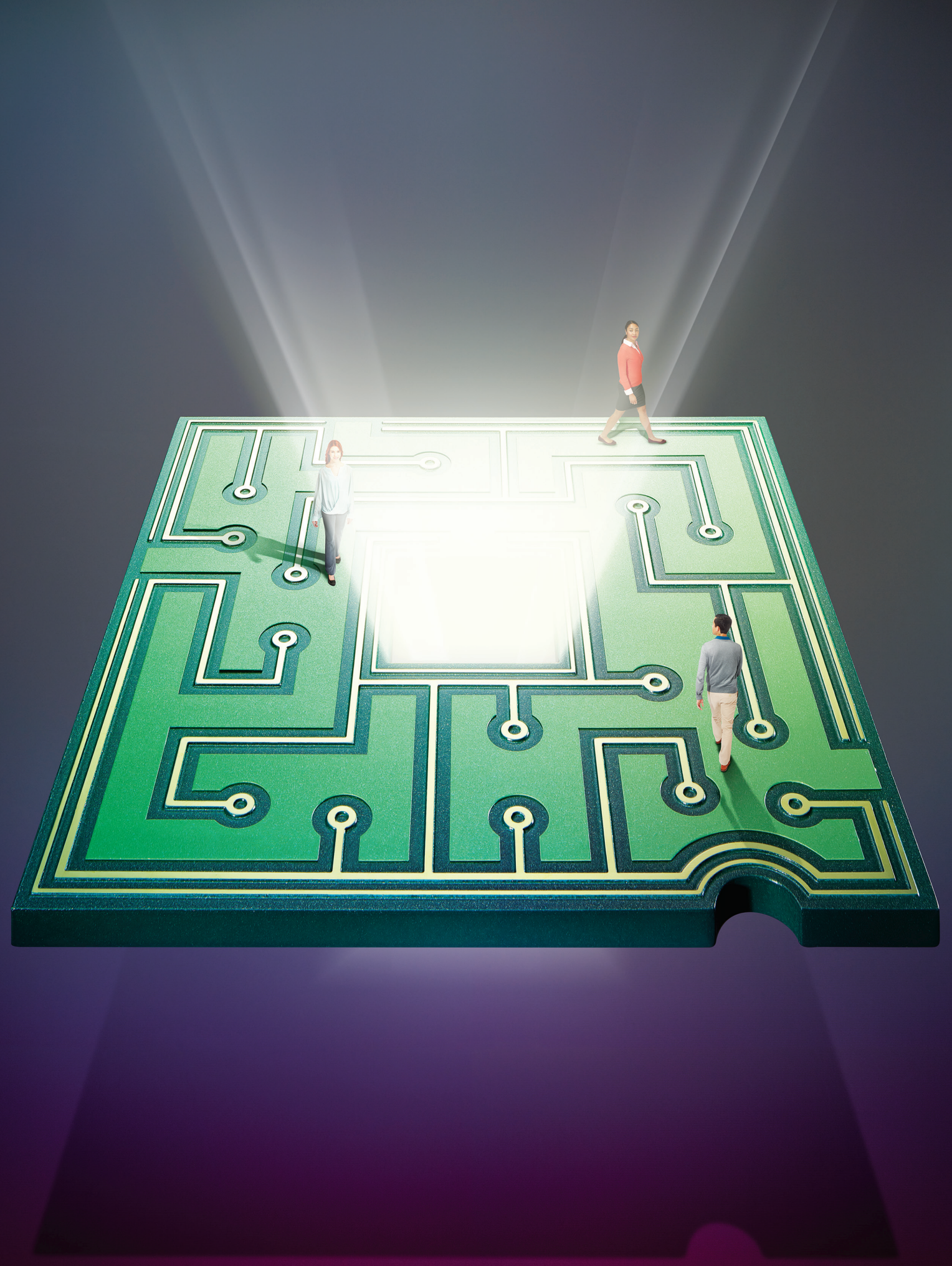


THE SEARCH FOR A NEW MACHINE

With the end of Moore's law in sight, chip manufacturers are spending billions to develop novel computing technologies

By John Pavlus

In a tiny, windowless conference room at the R&D headquarters of Intel, the world's dominant microprocessor and semiconductor manufacturer, Mark Bohr, the company's director of process architecture and integration, is coolly explaining how Moore's law, as it is commonly understood, is dead—and has been for some time. This might seem surprising, given that Bohr is literally in the Moore's law business: his job is to figure out how to make Intel's current 14-nanometer-wide transistors twice as small within the decade. But behind his round-rimmed glasses, Bohr does not even blink: "You have to understand that the era of traditional transistor scaling, where you take the same basic structure and materials and make it smaller—that ended about 10 years ago."

In 1965 Gordon Moore, then director of R&D at Fairchild Semiconductor, published the bluntly entitled document "Cramming More Components onto Integrated Circuits." Moore predicted that the number of transistors that could be built into a chip at optimal cost would double every year. A decade later he revised his prediction into what became known as Moore's law: every two years the number of transistors on a computer chip will double.

Integrated circuits make computers work. But Moore's law makes computers evolve. Because transistors are the "atoms" of electronic computation—the tiny switches that encode every 1 and 0 of a computer's memory and logic as a difference in voltage—if you double the number of them that can fit into the same amount of physical space, you can double the amount of computing you can do for the same cost. Intel's first general-purpose microprocessor, the 8080, helped to launch the PC revolution when it was released in 1974. The two-inch-long,

Photoillustration by Dan Saelinger

candy bar–shaped wafer contained 4,500 transistors. As of this writing, Intel's high-performance server central processing units (CPUs)—the highest-density chips commercially available—contain 4.5 billion transistors each. Inside Intel's Hillsboro, Ore., fabrication facilities, or "fabs," the company's latest manufacturing process can etch features as small as 14 nanometers into a wafer of silicon. That is thinner than a bacterium's flagellum. Such exponential growth in transistor density turned the room-sized, vacuum tube–powered calculating engines of the mid-20th century into the miniaturized silicon marvels of the early 21st.

But even Moore's law buckles under the laws of physics—and within a decade it will no longer be possible to maintain this unprecedented pace of miniaturization. That is why chip manufacturers such as Intel, IBM and Hewlett-Packard (HP) are dumping billions of R&D dollars into figuring out a post–Moore's law world. It will require blowing up basic assumptions about what our technology requires to function. Does a computer chip need to be a two-dimensional array of wires etched into silicon? IBM thinks not: it is seriously investigating carbon nanotubes and graphene as a computational substrate. What about electrons—are those necessary? IBM and HP are also placing bets on photonics, which uses pulses of light instead of voltage.

HP is going even further; it wants to extend the fundamental theory of electronics itself. The company has built a prototype computer, code-named "the Machine," that leverages the power of a long-sought missing link of electronics: the memristor. This component—mathematically predicted decades ago but only recently developed—allows the storage and random-access memory (RAM) functions of computers to be combined. The common metaphor of the CPU as a computer's "brain" would become more accurate with memristors instead of transistors because the former actually work more like neurons: they transmit and encode information as well as store it. Combining volatile memory and nonvolatile storage in this way could dramatically increase efficiency and diminish the so-called von Neumann bottleneck, which has constrained computing for half a century.

None of these technologies are ready to replace, or even augment, the chips in our laptops or phones in the next few years. But by the end of the decade at least one of them must be able to deliver computational performance gains that have a chance of taking over where traditional silicon circuit engineering inevitably trails off. The question is: Which one—and when?

BEYOND SILICON

THE IDEA BEHIND Moore's law is simple—halving the size of a transistor means you can get double the computing performance for the same cost. But there has always been more to it than that. Gordon Moore's 1965 paper may have predicted *what* would happen to transistor density every other year, but he never described *how* performance doubling would emerge from that increased density. It took another nine years for a scientist at IBM named Robert Dennard to publish an explanation now known as Dennard scaling. It describes how the power density of MOSFETs (which stands for "metal-oxide-semiconductor field-effect transistors," the dominant technology in 1974) stays constant as their physical size scales down. In other words, as transistors shrink, the amount of electric voltage and current required to switch them on and off shrinks, too.

For 30 years Dennard scaling was the secret driver of Moore's law, guaranteeing the steady PC performance gains that helped people start businesses, design products, cure diseases, guide spacecraft and democratize the Internet. And then it stopped working. Once fabs began etching features into silicon smaller than 65 nanometers (about half the length of an HIV virus), chip designers found that their transistors began to "leak" electrons because of quantum-mechanical effects. The devices were simply becoming too tiny to reliably switch between "on" and "off"—and any digital computer that cannot tell the difference between 1 and 0 has a serious problem. Not only that, researchers at IBM and Intel were discovering a so-called frequency wall that put a limit on how fast silicon-based CPUs could execute logical operations—about four billion times per second—without melting down from excess heat.

Technically, Moore's law could carry on (and did): Intel continued to cram tinier transistors into its wafers every two years. Yet it did not neatly translate into cheaper, faster computers.

Since 2000, chip engineers faced with these obstacles have been developing clever work-arounds. They have dodged the frequency wall by introducing multicore CPUs (a 10-gigahertz processor will burn itself up, but four, eight or 16 3-GHz processors working together will not). They have shored up leaky transistors with "tri-gates" that control the flow of current from three sides instead of one. And they have built systems that let CPUs outsource particularly strenuous tasks to special-purpose sidekicks (an iPhone 6's screen, for instance, is driven by its own four-core graphics processor). But these stopgaps will not change the fact that silicon scaling has less than a decade left to live.

That is why some chipmakers are looking for ways to ditch silicon. Last year IBM announced that it was allocating $3 billion to aggressively research various forms of postsilicon computing. The primary material under investigation is graphene: sheets of car-

John Pavlus is a writer and filmmaker focusing on science, technology and design topics. His work has appeared in *Wired, Nature, MIT Technology Review* and other outlets.

IN BRIEF

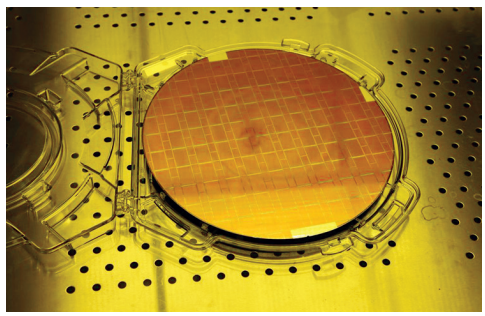
Progress in computing depends on Moore's law, the idea that every two years the number of transistors on a computer chip will double. But transistors can become only so small before engineers run up against the laws of physics, and that time is drawing near.

As a result, chip manufacturers are spending billions to develop fundamentally new computing architectures and processor designs, some based on new materials. Ideas long studied in laboratories are now being pursued with commercial fervor.

It is too early to tell which technologies will emerge as winners. The most likely outcome is that specialized technologies will come to perform specific tasks once assigned to a single central processor: Moore's law, singular, will be replaced by Moore's laws, plural.

bon just one atom thick. Like silicon, graphene has electronically useful properties that remain stable under a wide range of temperatures. Even better, electrons zoom through it at relativistic speeds. And most crucially, it scales—at least in the laboratory. Graphene transistors have been built that can operate hundreds or even thousands of times faster than the top-performing silicon devices, at reasonable power density, even below the five-nanometer threshold in which silicon goes quantum.

Yet unlike silicon, graphene lacks a "bandgap": the energy difference between orbitals in which electrons are bound to the atom and those in which the electrons are free to move around and participate in conduction. Metals, for example, have no bandgap: they are pure conductors. Without a bandgap, it is very difficult to stem the flow of current that turns a transistor from on to off—which means that a graphene device cannot reliably encode digital logic. "We have been the leaders in this area, but the results we have seen with graphene have not been encouraging," admits Supratik Guha, director of physical sciences at the IBM Thomas J. Watson Research Center. "Graphene has to be very cheap *and* provide some unique advantage for it to dislodge existing materials. It has very interesting properties, but it doesn't have a killer application that we've been able to identify."



COURTESY OF IBM (*TrueNorth microchip wafer*)

Dharmendra Modha, founder of IBM's cognitive computing group, wants to build computer chips that are at least as "smart" as a housefly.

Carbon nanotubes may hold more promise. When sheets of graphene are rolled into hollow cylinders, they can acquire a small bandgap that gives them semiconducting properties akin to what silicon has, reopening the possibility of using them for digital transistors. "We're cautiously optimistic," Guha says. "Carbon nanotubes as individual devices, when they're scaled down to 10 nanometers or so, outperform anything else available. If we look at our simulations of computing systems using carbon nanotubes, we expect that there may be a fivefold improvement [over silicon] in performance or energy efficiency."

But carbon nanotubes are delicate structures. If a nanotube's diameter or chirality—the angle at which its carbon atoms are "rolled"—varies by even a tiny amount, its bandgap may vanish, rendering it useless as a digital circuit element. Engineers must also be able to place nanotubes by the billions into neat rows just a few nanometers apart, using the same technology that silicon fabs rely on now. "For carbon nanotubes to become a worthy successor to silicon, we need to figure all this out within the next two or three years," Guha says.

BREAKING DOWN THE MEMORY WALL

"WHAT'S THE MOST expensive real estate on the planet?" Andrew Wheeler asks. "This, right here." He points to a box drawn in black marker on a whiteboard, representing the die of a microchip. Tall, wiry and square-jawed in straight-leg jeans and a checked cotton shirt, Wheeler looks more like an ex-cowboy than the deputy director of HP Labs, Hewlett-Packard's research arm. He is explaining what most of the transistors occupying that premium real estate are actually used for. It is not for computation, he says. It is called "cache memory" or static RAM (SRAM), and all it does is store frequently accessed instructions. It is the silicon equivalent of the dock on your Mac—the place where you keep things you want to avoid digging for. Wheeler wants it to disappear. But he is getting ahead of himself. In the near term, he will settle for getting rid of your computer's hard drive and main memory.

According to HP, these three items—collectively known as the memory hierarchy, with SRAM at the top and hard drives at the bottom—are responsible for most of the problems faced by engineers grappling with Moore's law. Without high-speed, high-capacity memory to store bits and ship them as quickly as possible, faster CPUs do little good.

To break down this "memory wall," Wheeler's team in Palo Alto, Calif., has been designing a new kind of computer—the Machine—that avoids the memory hierarchy altogether by collapsing it into one unified tier. Each tier in the memory hierarchy is good at some things and bad at others. SRAM is extremely fast (so it can keep up with the CPU) but power-hungry and low-capacity. Main memory, or dynamic RAM (DRAM), is pretty fast, dense and durable—which is good, because this is the workbench that your computer uses to run active applications. Of course, cutting the power makes everything in DRAM disappear, which is why "nonvolatile" storage media such as flash and hard disks are necessary for saving data in long-term storage. They offer high capacity and low power consumption, but they are glacially slow (and flash memory wears out relatively quickly). Because each memory medium has overlapping trade-offs, modern computers link them together so that CPUs can shuttle data up and down the tiers as efficiently as possible. "It's an absolute marvel of engineering," Wheeler says. "But it's also a huge waste."

A universal memory that could combine the speed of SRAM, the durability of DRAM, and the capacity and power efficiency of flash storage has been a holy grail for engineers, designers and programmers for decades, Wheeler says. The Machine exploits an electronic component, the memristor, to cover the latter two items on the universal-memory wish list. Mathematically predicted in 1971, the memristor—which is a blend of the words "*mem*ory *resistor*" because the device's ability to conduct electricity depends on the amount of current that previously flowed through it—was long believed to be impossible to build. In 2008 HP announced that it had constructed a working memristor; the research program was internally fast-tracked and became the precursor of the Machine.

Pulsing a memristor memory cell with voltage can change its conductive state, creating the clear on/off distinction necessary for storing digital data. As with flash memory, that state persists when the current is removed. And like DRAM, the cells can be read and written at high speed while densely packed together.

To achieve SRAM-like performance, though, memristor cells would need to be placed adjacent to the CPU on the same die of silicon—a physically impractical arrangement with current technology. Instead HP plans to employ photonics—shipping bits as pulses of laser light instead of electric current—to connect its high-performance memristor memory to the standard SRAM caches on logic processors. It is not quite the holy grail of universal memory—the Machine collapses the memory hierarchy from three tiers to two—but it is close.

By combining RAM with nonvolatile storage, a memristor-based architecture like the Machine could enable massive increases in computer performance without relying on Moore's law–style miniaturization. The version of IBM's Watson supercomputer that beat human contestants on *Jeopardy* in 2011 needed 16 terabytes of DRAM—housed in 10 power-guzzling Linux server racks—to perform its feats in real time. The same amount of nonvolatile flash memory could fit into a shoe box while consuming the same amount of power as an average laptop. A memory architecture that combined both functions at once would allow enormous data sets to be held in active memory for real-time processing rather than diced into smaller, sequential chunks—and at much lower energy costs.

As more and more connected devices join the "Internet of Things," the problem of streaming countless petabytes of information to and from data centers for storage and processing will make Moore's law moot, Wheeler says. Yet if universal memory enables supercomputerlike capabilities in smaller, less energy-hungry packages, these data streams could be stored and preprocessed locally by the connected devices themselves. Silicon CPU elements might never get smaller than seven nanometers or faster than four gigahertz—but with the memory wall torn down, it may no longer matter.

BEYOND VON NEUMANN

EVEN IF HP SUCCEEDS in its gambit to build universal memory, computers will still remain what they have always been since ENIAC, the first general-purpose computer, was built in 1946: extremely fast numerical calculators. Their essential design, formalized by mathematician John von Neumann in 1945, consists of a processing unit to execute instructions, a memory bank to store those instructions and the data they are to operate on, and a connection, or "bus," linking them. This von Neumann architecture is optimized for executing symbolic instructions in a linear sequence—also known as doing arithmetic. The first "computers" were human beings paid to sit in a room and work out calculations by hand, so it is no coincidence that electronic computers were designed to automate that tedious and error-prone process.

But today we increasingly need computers to do jobs that do not map well to linear mathematical instructions: tasks such as recognizing objects of interest in hours of video footage or guiding autonomous robots through unstable or dangerous territory. These tasks have more in common with the sensing, pattern-matching abilities of biological brains than mechanical calculators. Organisms must extract actionable information from a dynamic environment in real time; if a housefly were forced to pass discrete instructions back and forth, one by one, between separate memory and processor modules in its brain, it would never complete the computation in time to avoid getting swatted.

Dharmendra Modha, founder of IBM's cognitive computing group, wants to build computer chips that are at least as "smart" as that housefly—and as energy-efficient. The key, he explains, has been to scrap the calculatorlike von Neumann architecture. Instead his team has aimed to mimic cortical columns in the mammalian brain, which process, transmit and store information in the same structure, with no bus bottlenecking the connection. IBM's recently unveiled TrueNorth chip contains more than five billion transistors arranged into 4,096 neurosynaptic cores that model one million neurons and 256 million synaptic connections.

What that arrangement buys is real-time pattern-matching performance on the energy budget of a laser pointer. Modha points to a video monitor in the corner of the demo room at the IBM Almaden research campus in San Jose, Calif. The scene on it looks like surveillance footage from a camera that needs a hard reboot: cars, pedestrians, and a bicycle or two are frozen in place on a traffic roundabout; one of the pedestrians is highlighted by a red box superimposed on the image. After a minute, the cars, people and bikes lurch into a different frozen position, as if the footage suddenly skipped ahead.

It turns out that we do not want stand-alone, oraclelike "thinking machines" as much as late 20th-century science-fiction writers thought we would.

"You see, it's not a still image," Modha explains. "That's a video stream from Stanford's campus being analyzed by a laptop simulating a TrueNorth chip. It's just running about 1,000 times slower than real time." The actual TrueNorth chip that usually runs the video stream was being used for an internal training session in an auditorium next door, so I was not witnessing the chip's real performance. If I were, Modha says, the video feed would be playing at real-time speed, and the little red boxes would smoothly track the pedestrians as they entered and exited the frame.

Just like their von Neumann architecture counterparts, neurosynaptic devices such as TrueNorth have their own inherent weaknesses. "You wouldn't want to run iOS with this chip," Modha says. "I mean, you could, but it would be horribly inefficient—just like the laptop is inefficient at processing that video stream." IBM's goal is to harness the efficiencies of both architectures—one for precise and logical calculation, the other for responsive, associative pattern matching—into what it describes as a holistic computing system.

In this vision, the classic formulation of Moore's law still matters. Modha's team has already tiled 16 TrueNorth chips into a board, and by the end of this year the group plans to

SCIENTIFIC AMERICAN ONLINE See a slide show of new chip technologies at ScientificAmerican.com/may2015/moores-law

stack eight boards together into a 100-watt, toaster oven–sized device whose computational power "would require an entire data center to simulate."

In other words, silicon and transistor counts still matter—but what matters more is how they are arranged. "That was our insight: by rearranging the bricks, you get a very different functionality in the building," Modha says. "A lot of people believed, including us at first, that you really needed to change the technology to get the gains. In fact, it became clear that while new technology may bring gains, a new architecture gave orders-of-magnitude gains in performance that were easy pickings in comparison."

MOORE'S LAWS

BACK AT BUILDING RA3 in Hillsboro, Michael C. Mayberry, Intel's director of components research, is dispelling another myth about Moore's law: it was never really about transistors. "Cost per function is the game," he says. Whether it is measured in transistors per square centimeter of silicon, instructions of code executed per second, or performance per watt of power, all that matters is doing ever more work with ever fewer resources. It is no surprise that on its own Web site, Intel describes Moore's law not as a technological trend or force of nature but as a business model.

"When someone asks me, 'What keeps you up at night about Moore's law?' I say, 'I sleep fine,'" Mayberry says. "When Dennard scaling ended, that doesn't mean we stopped. We just changed. If you look forward 15 years, we can see several changes coming, but it doesn't mean we're going to stop." What Intel, IBM and HP all agree on is that the future of computational performance—that is, how the industry will collectively deliver increased function at decreased cost—will cease to look like a line or a curve and will instead look more like the multibranching tree of biological evolution itself.

That is because our vision of computers themselves is evolving. It turns out that we do not want stand-alone, oraclelike "thinking machines" as much as late 20th-century science-fiction writers thought we would. What is really dying is not Moore's law but the era of efficient general-purpose computation that Moore's law described and enabled. "Cramming everything into the box that you can," as Mayberry puts it.

Instead the relentless pursuit of lower cost per function will be driven by so-called heterogeneous computing, as Moore's law splits into Moore's laws. Companies such as IBM, Intel, HP and others will integrate not just circuits but entire systems that can handle the multiplying demands of distinct computational workloads. Bernard S. Meyerson of IBM says that people buy functions, not computer chips; indeed, they are less and less interested in buying computers at all. We just want our tools to compute, or "think," in ways that make them helpful in the contexts in which we use them. So instead of HAL, the superintelligent computer from *2001: A Space Odyssey,* we have Google Now on a smartphone telling us when to leave for the airport to catch a flight.

Futurists such as Nick Bostrom (author of *Superintelligence: Paths, Dangers, Strategies*) presume that Moore's law will cause generalized artificial intelligence to take off and coalesce into a kind of all-knowing, omnipotent digital being. But heterogeneous computing suggests that computation is more likely to diffuse outward into formerly "dumb" objects, systems and niches—imbuing things such as cars, network routers, medical diagnostic equipment and retail supply chains with the semiautonomous flexibility and context-specific competence of domestic animals. In other words, in a post–Moore's law world, computers will not become gods—but they will act like very smart dogs.

And just as a Great Dane is not built to do the job of a terrier, a graphics processor is not built to do the work of a CPU. HP's Wheeler offers the example of multiple special-purpose processing cores "bolted onto" a petabyte-scale pool of universal memory—a hybrid of processing power and massive memory that works in much the same way that dedicated graphics accelerators and memory caches are marshaled around centralized CPU resources now. Meanwhile IBM's Modha envisions golf ball–size devices, consisting of cognitive chips fastened to cheap cameras, that could be dropped into natural disaster sites to detect highly specific patterns such as the presence of injured children. Computer scientist Leon Chua of the University of California, Berkeley, who first theorized the existence of memristors in 1971, says that HP's efforts to collapse the memory hierarchy and IBM's research on reimagining the CPU are complementary responses to what he calls "the Great Data Bottleneck." "It's incredible that the computers we've been using for everything for the past 40 years are all still based on the same idea" of the calculatorlike von Neumann architecture, he says. The two-front transition to heterogeneous computing is "inevitable," he asserts, and "will create an entirely new economy"—not least because post–Moore's law, post–von Neumann computing will require entirely new methods of programming and designing systems. So much of modern computer science, engineering and chip design is concerned with masking the inherent limitations that the memory hierarchy and von Neumann architecture impose on computation, Chua says, that once those limitations are removed, "every computer programmer will have to go back to school."

What Chua, Modha and Wheeler never mention in these near-future visions are transistors—or the predictable performance gains that the world has come to expect from them. According to IBM's Meyerson, what Moore's law has accurately described for half a century—a tidy relation between increased transistor density and decreased cost per function—may turn out to be a temporary coincidence. "If you look at the past 40 years of semiconductors, you can see a very constant heartbeat," Meyerson says. "It's not that progress won't continue. But this technology has now developed an arrhythmia." SA

MORE TO EXPLORE

Cramming More Components onto Integrated Circuits. Gordon E. Moore in *Electronics*, Vol. 38, No. 8, pages 114–117; April 19, 1965.

Memristor: The Missing Circuit Element. L. O. Chua in *IEEE Transactions on Circuit Theory*, Vol. 18, No. 5, pages 507–519; September 1971.

Design of Ion-Implanted MOSFET's with Very Small Physical Dimensions. R. H. Dennard et al. in *IEEE Journal of Solid-State Circuits*, Vol. 9, No. 5, pages 256–268; October 1974.

Carbon Nanotubes: The Route toward Applications. Ray H. Baughman et al. in *Science*, Vol. 297, pages 787–792; August 2, 2002.

FROM OUR ARCHIVES

The Next 20 Years of Microchips. The Editors; January 2010.

Just Add Memory. Massimiliano Di Ventra and Yuriy V. Pershin; February 2015.