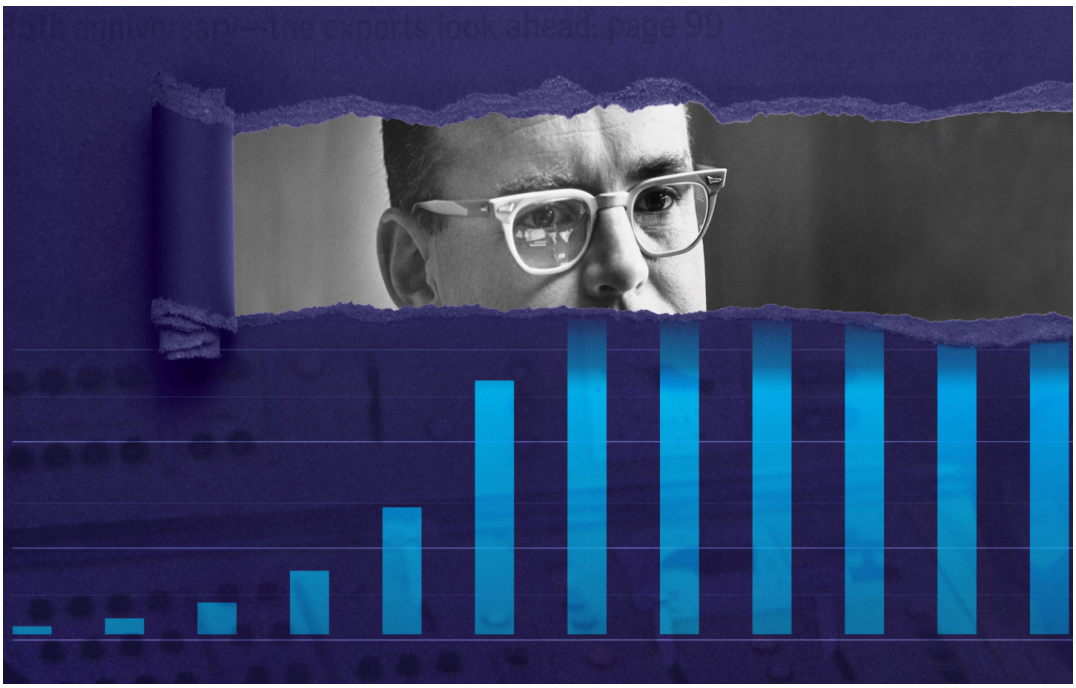
We’re not prepared for the end of Moore’s Law

It has fueled prosperity of the last 50 years. But the end is now in sight.

by

February 24, 2020



Moore's Law illustrationMS Tech

Gordon Moore’s 1965 forecast that the number of components on an integrated circuit would double every year until it reached an astonishing 65,000 by 1975 is the greatest technological prediction of

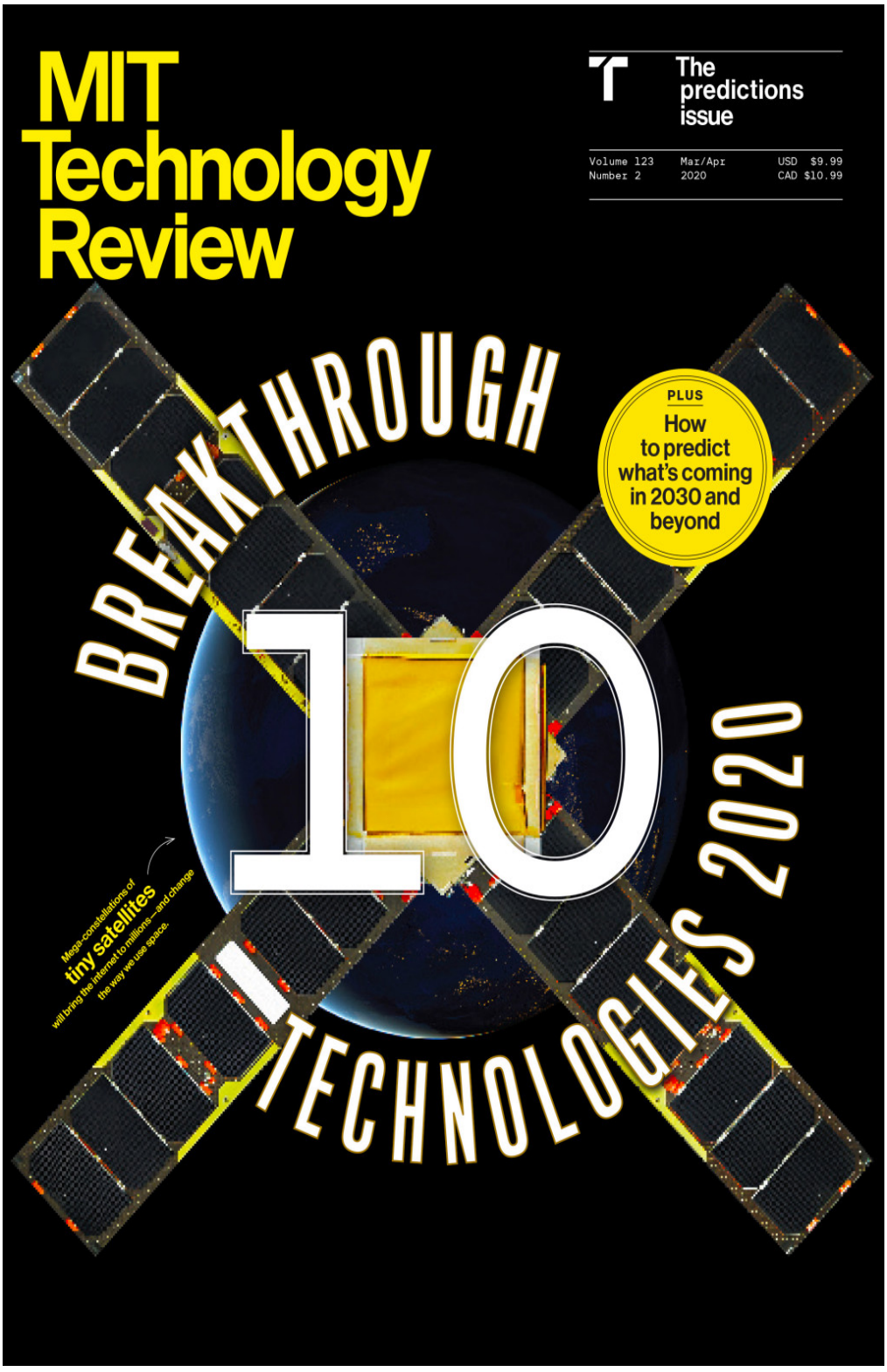
the last half-century. When it proved correct in 1975, he revised what has become known as Moore's Law to a doubling of transistors on a chip every two years.

Since then, his prediction has defined the trajectory of technology and, in many ways, of progress itself.

This story was part of our March 2020 issue
Moore's argument was an economic one. Integrated circuits, with multiple transistors and other electronic devices interconnected with aluminum metal lines on a tiny square of silicon wafer, had been invented a few years earlier by Robert Noyce at Fairchild Semiconductor. Moore, the company's R&D director, realized, as he wrote in 1965, that with these new integrated circuits, "the cost per component is nearly inversely proportional to the number of components." It was a beautiful bargain—in theory, the more transistors you added, the cheaper each one got. Moore also saw that there was plenty of room for engineering advances to increase the number of transistors you could affordably and reliably put on a chip.

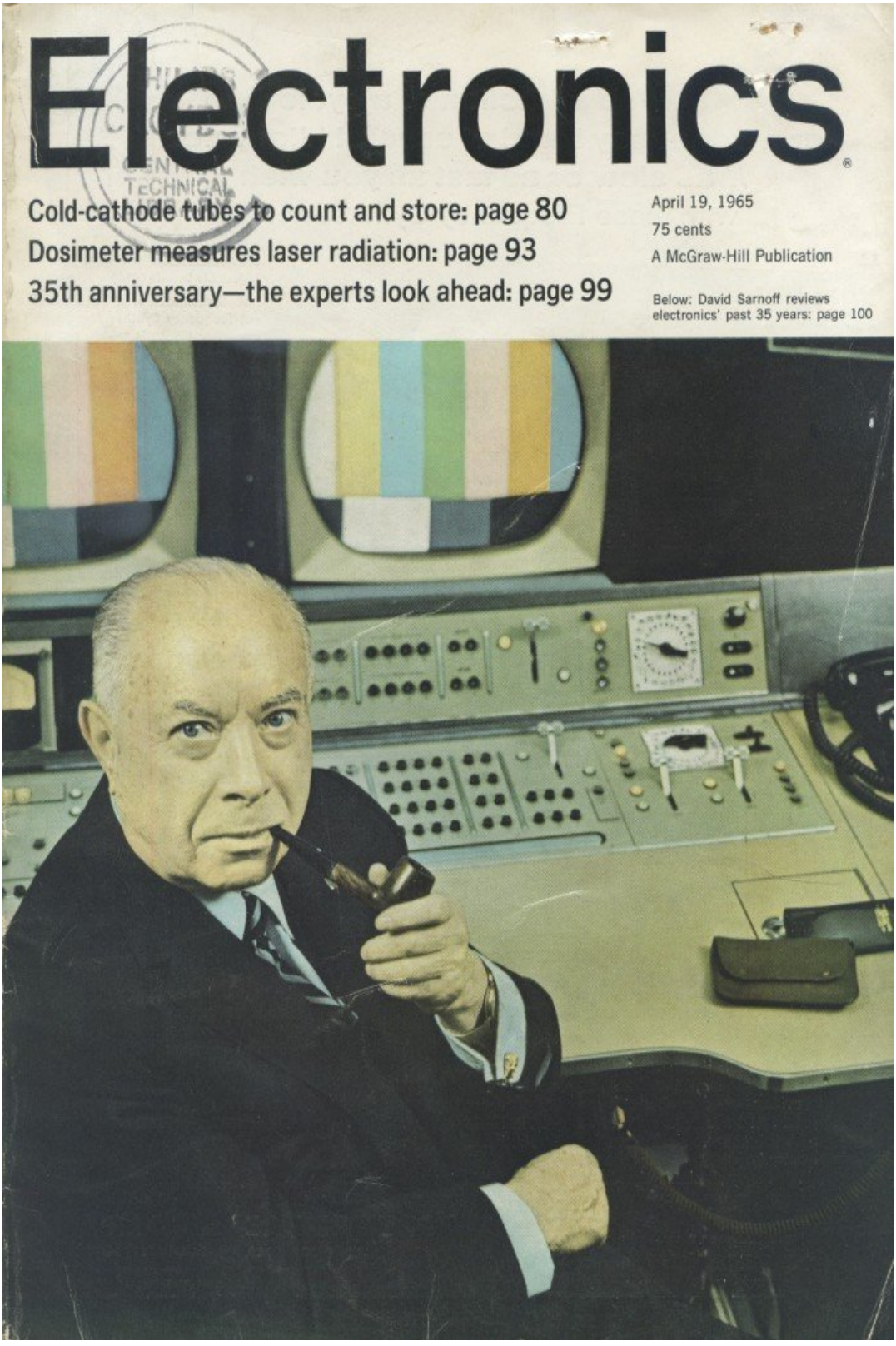
Soon these cheaper, more powerful chips would become what economists like to call a general purpose technology—one so fundamental that it spawns all sorts of other innovations and advances in multiple industries. A few years ago, leading economists credited the information technology made possible by integrated circuits with a third of US productivity growth since 1974. Almost every technology we care about, from smartphones to cheap laptops to GPS, is a direct reflection of Moore's prediction. It has also fueled today's breakthroughs in artificial intelligence and genetic medicine, by giving machine-learning techniques the ability to chew through massive amounts of data to find answers.

But how did a simple prediction, based on extrapolating from a graph of



the number of transistors by year—a graph that at the time had only a few data points—come to define a half-century of progress? In part, at

least, because the semiconductor industry decided it would.



The April 1965 Electronics Magazine in which Moore's article appeared. Wikimedia

Moore wrote that “cramming more components onto integrated circuits,” the title of his 1965 article, would “lead to such wonders as home computers—or at least terminals connected to a central computer—automatic controls for automobiles, and personal portable communications equipment.” In other words, stick to his road map of squeezing ever more transistors onto chips and it would lead you to the promised land. And for the following decades, a booming industry, the government, and armies of academic and industrial researchers poured money and time into upholding Moore’s Law, creating a self-fulfilling prophecy that kept progress on track with uncanny accuracy. Though the pace of progress has slipped in recent years, the most advanced chips today have nearly 50 billion transistors.

Every year since 2001, MIT Technology Review has chosen the 10 most important breakthrough technologies of the year. It’s a list of technologies that, almost without exception, are possible only because of the computation advances described by Moore’s Law.

For some of the items on this year’s list the connection is obvious: consumer devices, including watches and phones, infused with AI; climate-change attribution made possible by improved computer modeling and data gathered from worldwide atmospheric monitoring systems; and cheap, pint-size satellites. Others on the list, including quantum supremacy, molecules discovered using AI, and even anti-aging treatments and hyper-personalized drugs, are due largely to the computational power available to researchers.

But what happens when Moore’s Law inevitably ends? Or what if, as some suspect, it has already died, and we are already running on the fumes of the greatest technology engine of our time?

RIP

“It’s over. This year that became really clear,” says Charles Leiserson, a computer scientist at MIT and a pioneer of parallel computing, in which multiple calculations are performed simultaneously. The newest Intel fabrication plant, meant to build chips with minimum feature sizes of 10 nanometers, was much delayed, delivering chips in 2019, five years after the previous generation of chips with 14-nanometer features. Moore’s Law, Leiserson says, was always about the rate of progress, and “we’re no longer on that rate.” Numerous other prominent computer scientists have also declared Moore’s Law dead in recent years. In early 2019, the CEO of the large chipmaker Nvidia agreed.

In truth, it’s been more a gradual decline than a sudden death. Over the decades, some, including Moore himself at times, fretted that they could see the end in sight, as it got harder to make smaller and smaller transistors. In 1999, an Intel researcher worried that the industry’s goal of making transistors smaller than 100 nanometers by 2005 faced fundamental physical problems with “no known solutions,” like the quantum effects of electrons wandering where they shouldn’t be.

For years the chip industry managed to evade these physical roadblocks. New transistor designs were introduced to better corral the electrons. New lithography methods using extreme ultraviolet radiation were invented when the wavelengths of visible light were too thick to precisely carve out silicon features of only a few tens of nanometers. But progress grew ever more expensive. Economists at Stanford and MIT have calculated that the research effort going into upholding Moore’s Law has risen by a factor of 18 since 1971.

Likewise, the fabs that make the most advanced chips are becoming prohibitively pricey. The cost of a fab is rising at around 13% a year, and is expected to reach $16 billion or more by 2022. Not coincidentally, the number of companies with plans to make the next generation of chips

has now shrunk to only three, down from eight in 2010 and 25 in 2002.

> *Finding successors to today's silicon chips will take years of research.If you're worried about what will replace moore's Law, it's time to panic.*

Nonetheless, Intel—one of those three chipmakers—isn't expecting a funeral for Moore's Law anytime soon. Jim Keller, who took over as Intel's head of silicon engineering in 2018, is the man with the job of keeping it alive. He leads a team of some 8,000 hardware engineers and chip designers at Intel. When he joined the company, he says, many were anticipating the end of Moore's Law. If they were right, he recalls thinking, "that's a drag" and maybe he had made "a really bad career move."

But Keller found ample technical opportunities for advances. He points out that there are probably more than a hundred variables involved in keeping Moore's Law going, each of which provides different benefits and faces its own limits. It means there are many ways to keep doubling the number of devices on a chip—innovations such as 3D architectures and new transistor designs.

These days Keller sounds optimistic. He says he has been hearing about the end of Moore's Law for his entire career. After a while, he "decided not to worry about it." He says Intel is on pace for the next 10 years, and he will happily do the math for you: 65 billion (number of transistors) times 32 (if chip density doubles every two years) is 2 trillion transistors. "That's a 30 times improvement in performance," he says, adding that if software developers are clever, we could get chips that are a hundred times faster in 10 years.

Still, even if Intel and the other remaining chipmakers can squeeze out a few more generations of even more advanced microchips, the days when you could reliably count on faster, cheaper chips every couple of years are clearly over. That doesn't, however, mean the end of

computational progress.

Time to panic

Neil Thompson is an economist, but his office is at CSAIL, MIT's sprawling AI and computer center, surrounded by roboticists and computer scientists, including his collaborator Leiserson. In a new paper, the two document ample room for improving computational performance through better software, algorithms, and specialized chip architecture.

One opportunity is in slimming down so-called software bloat to wring the most out of existing chips. When chips could always be counted on to get faster and more powerful, programmers didn't need to worry much about writing more efficient code. And they often failed to take full advantage of changes in hardware architecture, such as the multiple cores, or processors, seen in chips used today.

Thompson and his colleagues showed that they could get a computationally intensive calculation to run some 47 times faster just by switching from Python, a popular general-purpose programming

language, to the more efficient C. That's because C, while it requires more work from the programmer, greatly reduces the required number of operations, making a program run much faster. Further tailoring the code to take full advantage of a chip with 18 processing cores sped things up even more. In just 0.41 seconds, the researchers got a result that took seven hours with Python code.

That sounds like good news for continuing progress, but Thompson worries it also signals the decline of computers as a general purpose technology. Rather than "lifting all boats," as Moore's Law has, by offering ever faster and cheaper chips that were universally available, advances in software and specialized architecture will now start to selectively target specific problems and business opportunities, favoring those with sufficient money and resources.

Indeed, the move to chips designed for specific applications, particularly in AI, is well under way. Deep learning and other AI applications increasingly rely on graphics processing units (GPUs) adapted from gaming, which can handle parallel operations, while companies like Google, Microsoft, and Baidu are designing AI chips for their own particular needs. AI, particularly deep learning, has a huge appetite for computer power, and specialized chips can greatly speed up its performance, says Thompson.

But the trade-off is that specialized chips are less versatile than traditional CPUs. Thompson is concerned that chips for more general computing are becoming a backwater, slowing "the overall pace of computer improvement," as he writes in an upcoming paper, "The Decline of Computers as a General Purpose Technology."

At some point, says Erica Fuchs, a professor of engineering and public policy at Carnegie Mellon, those developing AI and other applications will miss the decreases in cost and increases in performance delivered by Moore's Law. "Maybe in 10 years or 30 years—no one really knows when—you're going to need a device with that additional computation

power," she says.

The problem, says Fuchs, is that the successors to today's general purpose chips are unknown and will take years of basic research and development to create. If you're worried about what will replace Moore's Law, she suggests, "the moment to panic is now." There are, she says, "really smart people in AI who aren't aware of the hardware constraints facing long-term advances in computing." What's more, she says, because application--specific chips are proving hugely profitable, there are few incentives to invest in new logic devices and ways of doing computing.

Wanted: A Marshall Plan for chips

In 2018, Fuchs and her CMU colleagues Hassan Khan and David Hounshell wrote a paper tracing the history of Moore's Law and identifying the changes behind today's lack of the industry and government collaboration that fostered so much progress in earlier decades. They argued that "the splintering of the technology trajectories and the short-term private profitability of many of these new splinters" means we need to greatly boost public investment in finding the next great computer technologies.

If economists are right, and much of the growth in the 1990s and early 2000s was a result of microchips—and if, as some suggest, the sluggish productivity growth that began in the mid-2000s reflects the slowdown in computational progress—then, says Thompson, "it follows you should invest enormous amounts of money to find the successor technology. We're not doing it. And it's a public policy failure."

There's no guarantee that such investments will pay off. Quantum computing, carbon nanotube transistors, even spintronics, are enticing possibilities—but none are obvious replacements for the promise that Gordon Moore first saw in a simple integrated circuit. We need the

research investments now to find out, though. Because one prediction is pretty much certain to come true: we're always going to want more computing power.