MORE THAN MOORE

BY M. MITCHELL WALDROP

THE SEMICONDUCTOR INDUSTRY WILL SOON ABANDON ITS PURSUIT OF MOORE'S LAW. NOW THINGS COULD GET A LOT MORE INTERESTING.

Next month, the worldwide semiconductor industry will formally acknowledge what has become increasingly obvious to everyone involved: Moore's law, the principle that has powered the information-technology revolution since the 1960s, is nearing its end.

A rule of thumb that has come to dominate computing, Moore's law states that the number of transistors on a microprocessor chip will double every two years or so — which has generally meant that the chip's performance will, too. The exponential improvement that the law describes transformed the first crude home computers of the 1970s into the sophisticated machines of the 1980s and 1990s, and from there gave rise to high-speed Internet, smartphones and the wired-up cars, refrigerators and thermostats that are becoming prevalent today.

None of this was inevitable: chipmakers deliberately chose to stay on the Moore's law track. At every stage, software developers came up with applications that strained the capabilities of existing chips; consumers asked more of their devices; and manufacturers rushed to meet that demand with next-generation chips. Since the 1990s, in fact, the semiconductor industry has released a research road map every two years to coordinate what its hundreds of manufacturers and suppliers are doing to stay in step with the law — a strategy sometimes called More Moore. It has been largely thanks to this road map that computers have followed the law's exponential demands.

Not for much longer. The doubling has already started to falter, thanks to the heat that is unavoidably generated when more and more silicon circuitry is jammed into the same small area. And some even more fundamental limits loom less than a decade away. Top-of-the-line microprocessors currently have circuit features that are around 14 nanometres across, smaller than most viruses. But by the early 2020s, says Paolo Gargini, chair of the road-mapping organization, "even with super-aggressive efforts, we'll get to the 2–3-nanometre limit, where features are just 10 atoms across. Is that a device at all?" Probably not — if only because at that scale, electron behaviour will be governed by quantum uncertainties that will make transistors hopelessly unreliable. And despite vigorous research efforts, there is no obvious successor to today's silicon technology.

The industry road map released next month will for the first time lay out a research and development plan that is not centred on Moore's law. Instead, it will follow what might be called the More than Moore strategy: rather than making the chips better and letting the applications follow, it will start with applications — from smartphones and supercomputers to data centres in the cloud — and work downwards to see what chips are needed to support them. Among those chips will be new generations of sensors, power-management circuits and other silicon devices required by a world in which computing is increasingly mobile.

The changing landscape, in turn, could splinter the industry's long tradition of unity in pursuit of Moore's law. "Everybody is struggling with what the road map actually means," says Daniel Reed, a computer scientist and vice-president for research at the University of Iowa in Iowa City. The Semiconductor Industry Association (SIA) in Washington DC, which represents all the major US firms, has already said that it will cease its participation in the road-mapping effort once the report is out, and will instead pursue its own research and development agenda.

Everyone agrees that the twilight of Moore's law will not mean the end of progress. "Think about what happened to airplanes," says Reed. "A Boeing 787 doesn't go any faster than a 707 did in the 1950s — but they are very different airplanes", with innovations ranging from fully electronic controls to a carbon-fibre fuselage. That's what will happen with computers, he says: "Innovation will absolutely continue — but it will be more nuanced and complicated."

LAYING DOWN THE LAW

The 1965 essay[1] that would make Gordon Moore famous started with a meditation on what could be done with the still-new technology of integrated circuits. Moore, who was then research director of Fairchild Semiconductor in San Jose, California, predicted wonders such as home computers, digital wristwatches, automatic cars and "personal portable communications equipment" — mobile phones. But the heart of the essay was Moore's attempt to provide a timeline for this future. As a measure of a microprocessor's computational power, he looked at transistors, the on–off switches that make computing digital. On the basis of achievements by his company and others in the previous few years, he estimated that the number of transistors and other electronic components per chip was doubling every year.

Moore, who would later co-found Intel in Santa Clara, California, underestimated the doubling time; in 1975, he revised it to a more realistic two years[2]. But his vision was spot on. The future that he predicted started to arrive in the 1970s and 1980s, with the advent of microprocessor-equipped consumer products such as the Hewlett Packard hand calculators, the Apple II computer and the IBM PC. Demand for such products was soon exploding, and manufacturers were engaging in a brisk competition to offer more and more capable chips in smaller and smaller packages (see 'Moore's lore').

This was expensive. Improving a microprocessor's performance meant scaling down the elements of its circuit so that more of them could be packed together on the chip, and electrons could move between them more quickly. Scaling, in turn, required major refinements in photolithography, the basic technology for etching those microscopic elements onto a silicon surface. But the boom times were such that this hardly mattered: a self-reinforcing cycle set in. Chips were so versatile that manufacturers could make only a few types — processors and memory, mostly — and sell them in huge quantities. That gave them enough cash to cover the cost of upgrading their fabrication facilities, or 'fabs', and still drop the prices, thereby fuelling demand even further.

Soon, however, it became clear that this market-driven cycle could not sustain the relentless cadence of Moore's law by itself. The chip-making process was getting too complex, often involving hundreds of stages, which meant that taking the next step down in scale required a network of materials-suppliers and apparatus-makers to deliver the right upgrades at the right time. "If you need 40 kinds of equipment and only 39 are ready, then everything stops," says Kenneth Flamm, an economist who studies the computer industry at the University of Texas at Austin.

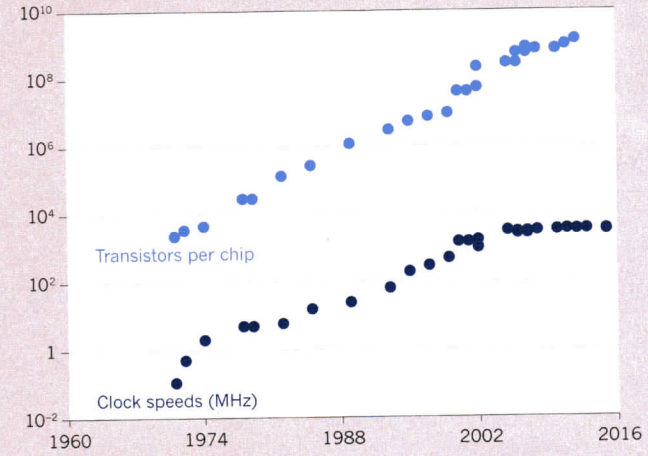
To provide that coordination, the industry devised its first road map. The idea, says Gargini, was "that everyone would have a rough estimate of where they were going, and they could raise an alarm if they saw roadblocks ahead". The US semiconductor industry launched the mapping effort in 1991, with hundreds of engineers from various companies working on the first report and its subsequent iterations, and Gargini, then the director of technology strategy at Intel, as its chair. In 1998, the effort became the International Technology Roadmap for Semiconductors, with participation from industry associations in Europe, Japan, Taiwan and South Korea. (This year's report, in keeping with its new approach, will be called the International Roadmap for Devices and Systems.)

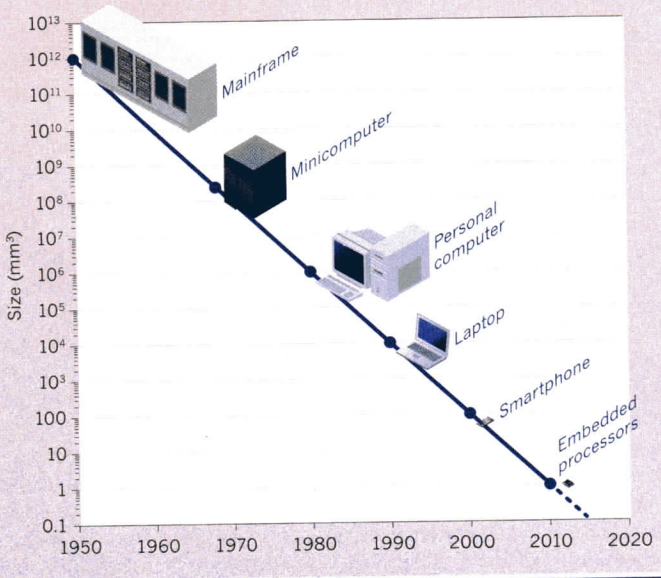
NATURE.COM
To hear more about what will come after Moore's law, visit: go.nature.com/nppjyx

ILLUSTRATION BY REBECCA MOCK

MOORE'S LORE

For the past five decades, the number of transistors per microprocessor chip — a rough measure of processing power — has doubled about every two years, in step with Moore's law (top). Chips also increased their 'clock speed', or rate of executing instructions, until 2004, when speeds were capped to limit heat. As computers increase in power and shrink in size, a new class of machines has emerged roughly every ten years (bottom).





"The road map was an incredibly interesting experiment," says Flamm. "So far as I know, there is no example of anything like this in any other industry, where every manufacturer and supplier gets together and figures out what they are going to do." In effect, it converted Moore's law from an empirical observation into a self-fulfilling prophecy: new chips followed the law because the industry made sure that they did.

And it all worked beautifully, says Flamm — right up until it didn't.

HEAT DEATH

The first stumbling block was not unexpected. Gargini and others had warned about it as far back as 1989. But it hit hard nonetheless: things got too small.

"It used to be that whenever we would scale to smaller feature size, good things happened automatically," says Bill Bottoms, president of Third Millennium Test Solutions, an equipment manufacturer in Santa Clara. "The chips would go faster and consume less power."

But in the early 2000s, when the features began to shrink below about 90 nanometres, that automatic benefit began to fail. As electrons had to move faster and faster through silicon circuits that were smaller and smaller, the chips began to get too hot.

That was a fundamental problem. Heat is hard to get rid of, and no one wants to buy a mobile phone that burns their hand. So manufacturers seized on the only solutions they had, says Gargini. First, they stopped trying to increase 'clock rates' — how fast microprocessors execute instructions. This effectively put a speed limit on the chip's electrons and limited their ability to generate heat. The maximum clock rate hasn't budged since 2004.

Second, to keep the chips moving along the Moore's law performance curve despite the speed limit, they redesigned the internal circuitry so that each chip contained not one processor, or 'core', but two, four or more. (Four and eight are common in today's desktop computers and smartphones.) In principle, says Gargini, "you can have the same output with four cores going at 250 megahertz as one going at 1 gigahertz". In practice, exploiting eight processors means that a problem has to be broken down into eight pieces — which for many algorithms is difficult to impossible. "The piece that can't be parallelized will limit your improvement," says Gargini.

Even so, when combined with creative redesigns to compensate for electron leakage and other effects, these two solutions have enabled chip manufacturers to continue shrinking their circuits and keeping their transistor counts on track with Moore's law. The question now is what will happen in the early 2020s, when continued scaling is no longer possible with silicon because quantum effects have come into play. What comes next? "We're still struggling," says An Chen, an electrical engineer who works for the international chipmaker GlobalFoundries in Santa Clara, California, and who chairs a committee of the new road map that is looking into the question.

That is not for a lack of ideas. One possibility is to embrace a completely new paradigm — something like quantum computing, which promises exponential speed-up for certain calculations, or neuromorphic computing, which aims to model processing elements on neurons in the brain. But none of these alternative paradigms has made it very far out of the laboratory. And many researchers think that quantum computing will offer advantages only for niche applications rather than for the everyday tasks at which digital computing excels. "What does it mean to quantum-balance a chequebook?" wonders John Shalf, head of computer-science research at the Lawrence Berkeley National Laboratory in Berkeley, California.

MATERIAL DIFFERENCES

A different approach, which does stay in the digital realm, is the quest to find a 'millivolt switch': a material that could be used for devices at least as fast as their silicon counterparts, but that would generate much less heat. There are many candidates, ranging from 2D graphene-like compounds to spintronic materials that would compute by flipping electron spins rather than by moving electrons. "There is an enormous research space to be explored once you step outside the confines of the established technology," says Thomas Theis, a physicist who directs the nanoelectronics initiative at the Semiconductor Research Corporation (SRC), a research-funding consortium in Durham, North Carolina.

Unfortunately, no millivolt switch has made it out of the laboratory either. That leaves the architectural approach: stick with silicon, but configure it in entirely new ways. One popular option is to go 3D. Instead of etching flat circuits onto the surface of a silicon wafer, build skyscrapers: stack many thin layers of silicon with microcircuitry etched into each. In principle, this should make it possible to pack more computational power into the same space. In practice, however, this currently works only with memory chips, which do not have a heat problem: they use circuits that consume power only when a memory cell is accessed, which is not that often. One example is the Hybrid Memory Cube design, a stack of as many as eight memory layers that is being pursued by an industry consortium originally

launched by Samsung and memory-maker Micron Technology in Boise, Idaho.

Microprocessors are more challenging: stacking layer after layer of hot things simply makes them hotter. But one way to get around that problem is to do away with separate memory and microprocessing chips, as well as the prodigious amount of heat — at least 50% of the total — that is now generated in shuttling data back and forth between the two. Instead, integrate them in the same nanoscale high-rise.

This is tricky, not least because current-generation microprocessors and memory chips are so different that they cannot be made on the same fab line; stacking them requires a complete redesign of the chip's structure. But several research groups are hoping to pull it off. Electrical engineer Subhasish Mitra and his colleagues at Stanford University in California have developed a hybrid architecture that stacks memory units together with transistors made from carbon nanotubes, which also carry current from layer to layer[3]. The group thinks that its architecture could reduce energy use to less than one-thousandth that of standard chips.

GOING MOBILE

The second stumbling block for Moore's law was more of a surprise, but unfolded at roughly the same time as the first: computing went mobile.

Twenty-five years ago, computing was defined by the needs of desktop and laptop machines; supercomputers and data centres used essentially the same microprocessors, just packed together in much greater numbers. Not any more. Today, computing is increasingly defined by what high-end smartphones and tablets do — not to mention by smart watches and other wearables, as well as by the exploding number of smart devices in everything from bridges to the human body. And these mobile devices have priorities very different from those of their more sedentary cousins.

Keeping abreast of Moore's law is fairly far down on the list — if only because mobile applications and data have largely migrated to the worldwide network of server farms known as the cloud. Those server farms now dominate the market for powerful, cutting-edge microprocessors that do follow Moore's law. "What Google and Amazon decide to buy has a huge influence on what Intel decides to do," says Reed.

Much more crucial for mobiles is the ability to survive for long periods on battery power while interacting with their surroundings and users. The chips in a typical smartphone must send and receive signals for voice calls, Wi-Fi, Bluetooth and the Global Positioning System, while also sensing touch, proximity, acceleration, magnetic fields — even fingerprints. On top of that, the device must host special-purpose circuits for power management, to keep all those functions from draining the battery.

The problem for chipmakers is that this specialization is undermining the self-reinforcing economic cycle that once kept Moore's law humming. "The old market was that you would make a few different things, but sell a whole lot of them," says Reed. "The new market is that you have to make a lot of things, but sell a few hundred thousand apiece — so it had better be really cheap to design and fab them."

Both are ongoing challenges. Getting separately manufactured technologies to work together harmoniously in a single device is often a nightmare, says Bottoms, who heads the new road map's committee on the subject. "Different components, different materials, electronics, photonics and so on, all in the same package — these are issues that will have to be solved by new architectures, new simulations, new switches and more."

For many of the special-purpose circuits, design is still something of a cottage industry — which means slow and costly. At the University of California, Berkeley, electrical engineer Alberto Sangiovanni-Vincentelli and his colleagues are trying to change that: instead of starting from scratch each time, they think that people should create new devices by combining large chunks of existing circuitry that have known functionality[4]. "It's like using Lego blocks," says Sangiovanni-Vincentelli. It's a challenge to make sure that the blocks work together, but "if you were to use older methods of design, costs would be prohibitive".

Costs, not surprisingly, are very much on the chipmakers' minds these days. "The end of Moore's law is not a technical issue, it is an economic issue," says Bottoms. Some companies, notably Intel, are still trying to shrink components before they hit the wall imposed by quantum effects, he says. But "the more we shrink, the more it costs".

Every time the scale is halved, manufacturers need a whole new generation of ever more precise photolithography machines. Building a new fab line today requires an investment typically measured in many billions of dollars — something only a handful of companies can afford. And the fragmentation of the market triggered by mobile devices is making it harder to recoup that money. "As soon as the cost per transistor at the next node exceeds the existing cost," says Bottoms, "the scaling stops."

Many observers think that the industry is perilously close to that point already. "My bet is that we run out of money before we run out of physics," says Reed.

Certainly it is true that rising costs over the past decade have forced a massive consolidation in the chip-making industry. Most of the world's production lines now belong to a comparative handful of multinationals such as Intel, Samsung and the Taiwan Semiconductor Manufacturing Company in Hsinchu. These manufacturing giants have tight relationships with the companies that supply them with materials and fabrication equipment; they are already coordinating, and no longer find the road-map process all that useful. "The chip manufacturer's buy-in is definitely less than before," says Chen.

> "MY BET IS THAT WE RUN OUT OF MONEY BEFORE WE RUN OUT OF PHYSICS."

Take the SRC, which functions as the US industry's research agency: it was a long-time supporter of the road map, says SRC vice-president Steven Hillenius. "But about three years ago, the SRC contributions went away because the member companies didn't see the value in it." The SRC, along with the SIA, wants to push a more long-term, basic research agenda and secure federal funding for it — possibly through the White House's National Strategic Computing Initiative, launched in July last year.

That agenda, laid out in a report[5] last September, sketches out the research challenges ahead. Energy efficiency is an urgent priority — especially for the embedded smart sensors that comprise the 'Internet of things', which will need new technology to survive without batteries, using energy scavenged from ambient heat and vibration. Connectivity is equally key: billions of free-roaming devices trying to communicate with one another and the cloud will need huge amounts of bandwidth, which they can get if researchers can tap the once-unreachable terahertz band lying deep in the infrared spectrum. And security is crucial — the report calls for research into new ways to build in safeguards against cyberattack and data theft.

These priorities and others will give researchers plenty to work on in coming years. At least some industry insiders, including Shekhar Borkar, head of Intel's advanced microprocessor research, are optimists. Yes, he says, Moore's law is coming to an end in a literal sense, because the exponential growth in transistor count cannot continue. But from the consumer perspective, "Moore's law simply states that user value doubles every two years". And in that form, the law will continue as long as the industry can keep stuffing its devices with new functionality.

The ideas are out there, says Borkar. "Our job is to engineer them." ■

M. Mitchell Waldrop *is a features editor for* Nature.

1. Moore, G. E. *Electronics* **38,** 114–117 (1965).
2. Moore, G. E. *IEDM Tech. Digest* 11–13 (1975).
3. Sabry Aly, M. M. *et al. Computer* **48**(12), 24–33 (2015).
4. Nikolic, B. *41th Eur. Solid-State Circuits Conf.* (2015); available at http://go.nature.com/wwljk7
5. *Rebooting the IT Revolution: A Call to Action* (SIA/SRC, 2015); available at http://go.nature.com/urvkhw