

The great shrink

How did computers go from giant, room-filling slabs of machinery to tiny gadgets you can fit in your pocket? Tim Cross considers the miracle of “Moore’s Law”, and asks what will happen when computer chips stop shrinking

In 1971, Intel, then an obscure firm in what would only later come to be known as Silicon Valley, released a chip called the 4004. It was the world’s first commercially available microprocessor, which meant it sported all the electronic circuits necessary for advanced number-crunching in a single, tiny package. It was a marvel of its time, built from 2,300 tiny transistors, each around 10,000 nanometres (or billionths of a metre) across – about the size of a red blood cell. A transistor is an electronic switch that, by flipping between “on” and “off”, provides a physical representation of the 1s and 0s that are the fundamental particles of information. In 2015, Intel, by then the world’s leading chipmaker, with revenues of more than \$55bn that year, released its Skylake chips. The firm no longer publishes exact numbers, but the best guess is that they have 1.5 billion–2 billion transistors apiece. Spaced 14 nanometres apart, each is so tiny as to be literally invisible, for they are more than an order of magnitude smaller than the wavelengths of light that humans use to see.

Everyone knows that modern computers are better than old ones.

But it is hard to convey just how much better, for no other consumer technology has improved at anything approaching a similar pace. The standard analogy is with cars: if the car from 1971 had improved

at the same rate as computer chips, by 2015 new models would have had top speeds of about 420 million miles per hour – fast enough to drive round the world in less than a fifth of a second. This blistering progress is a consequence of an observation first made in 1965 by one of Intel’s founders, Gordon Moore. Moore noted that the number of components that could be crammed onto an integrated circuit was doubling every year. Later amended to every two years, “Moore’s Law” has become a self-fulfilling prophecy that sets the pace for the entire computing industry. But it’s also a force that is nearly spent. Shrinking a chip’s components gets harder each time you do it, and with modern transistors having features measured in mere dozens of atoms, engineers are simply running out of room. For the law to hold until 2050, engineers would have to figure out how to build computers from components smaller than an atom of hydrogen, the smallest element there is. That, as far as anyone knows, is impossible.

Moreover, the benefits of Moore’s Law are dwindling. Shrinking chips no longer makes them faster or more efficient in the way that it used to. And the rising cost of the ultra-sophisticated equipment needed to make the chips is eroding the financial gains. Moore’s second law, more light-hearted than his first, states that



Great beginnings: the IBM 7090 mainframe computer could fill a large room

“The transistors in this chip are literally invisible, an order of magnitude smaller than the wavelengths of light humans use to see”

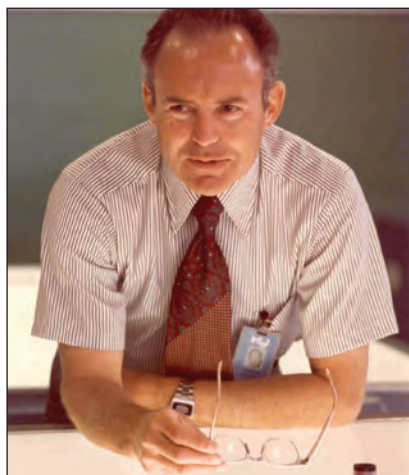
the cost of a “foundry”, as such factories are called, doubles every four years. A modern one leaves little change from \$10bn. Even for Intel, that is a lot of money. The result is a consensus among Silicon Valley’s experts that Moore’s Law is near its end. Bob Colwell, a former chip designer at Intel, thinks the industry may be able to get down to chips whose components are just five nanometres apart by the early 2020s – “but you’ll struggle to persuade me that they’ll get much further than that”. One of the most powerful technological forces of the past 50 years, in other words, will soon have run its course.

There are other ways of making computers better besides shrinking their components. The end of Moore’s Law does not mean that the computer revolution will stall. But it does mean that the coming decades will look very different from the preceding ones, for none of the alternatives is as reliable, or as repeatable, as the great shrinkage of the past half-century. Moore’s Law has made computers smaller, transforming them from room-filling behemoths to svelte, pocket-filling slabs. It has also made them more frugal: a smartphone that packs more computing power than was available to entire nations in 1971 can last a day or more on a single battery charge. But its most famous effect has been to make computers faster. By 2050, when Moore’s Law will be ancient history, engineers will have to make use of a string of other tricks if they are to keep computers getting faster.

One trick is better programming. The breakneck pace of Moore’s Law has in the past left software firms with little time to streamline their products. The fact that their customers would be buying faster machines every few years weakened the incentive even further. As Moore’s Law winds down, the famously short product cycles of the computing industry may start to lengthen, giving programmers more time to polish their work. Another is to design chips that trade general mathematical prowess for more specialised hardware. Modern chips are starting to feature specialised circuits designed to speed up common tasks, such as decompressing a film, performing the complex calculations required for encryption, or drawing the complicated 3D graphics used in video games. As computers spread into all sorts of other products, such specialised silicon will be very useful. Self-driving cars, for instance, will increasingly make use of machine vision, in which computers learn to interpret images from the real world.

Another idea is to try to keep Moore’s Law going by moving it into the third dimension. Modern chips are essentially flat, but

researchers are toying with chips that stack their components on top of each other. Building up would allow their designers to keep cramming in more components. IBM reckons 3D chips could allow designers to shrink a supercomputer that currently fills a building to something the size of a shoebox. But making it work will require some fundamental design changes. Modern chips already run hot, requiring beefy heat sinks and fans to keep them cool. A 3D chip would be even worse, for the surface area available to remove heat would grow much more slowly than the volume that generates it. IBM's shoebox supercomputer would therefore require liquid cooling. Microscopic channels would be drilled into each chip, allowing cooling liquid to flow through.



Gordon Moore: a self-fulfilling prophecy

There are more exotic ideas, too. Quantum computing proposes to use the counter-intuitive rules of quantum mechanics to build machines that can solve certain types of mathematical problem far more quickly than any conventional computer, no matter how fast or high-tech (for many other problems, though, a quantum machine would offer no advantage). But, like 3D chips, quantum computers need specialised care and feeding. For a quantum computer to work, its internals must be sealed off from the outside world. Quantum computers must be chilled with liquid helium to within a hair's breadth of absolute zero, and protected by sophisticated shielding, for even the smallest pulse of heat or stray electromagnetic wave could ruin the delicate quantum states that such machines rely on.

Each of these prospective improvements, though, is limited: either the gains are a one-off, or they apply only to certain sorts of calculations. And, unlike the glory days of Moore's Law, it is not clear how well any of this translates to consumer products. Few people would want a cryogenically cooled quantum PC or smartphone, after all. Ditto liquid cooling, which is heavy, messy and complicated. Even building specialised logic for a given task is worthwhile only if it will be regularly used. But all three technologies will work well in data centres, where they will help to power another big trend of the next few decades. Traditionally, a computer has been a box on your desk or in your pocket. In the future, the increasingly ubiquitous connectivity provided by the internet and the mobile-phone network will allow a great deal of computing power to be hidden away in data centres, with customers using it as and when needed. Computing will become a utility that is tapped on demand, like electricity or water. The ability to remove the hardware that does the computational heavy lifting from the hunk of plastic with which users interact – known as “cloud computing” – will be one of the most important ways for the industry to blunt the impact of the demise of Moore's Law. Unlike a smartphone or a PC, which can only grow so large, data centres can be made more powerful simply by building them bigger. As the world's demand for computing continues to grow, an increasing proportion of it will take place in shadowy warehouses hundreds of miles from the users who are being served.

This is already beginning to happen. Take an app such as Siri, Apple's voice-powered personal assistant. Decoding human speech and working out the intent behind an instruction such as, “Siri, find me some Indian restaurants nearby”, requires more computing power than an iPhone has available. Instead, the phone simply records its user's voice and forwards the information to a beefier computer in one of Apple's data centres. Once that remote computer has figured out an appropriate response, it sends the information back to the iPhone. The same

model can be applied to much more than just smartphones. Chips have already made their way into things not normally thought of as computers, from cars to medical implants to televisions and kettles, and the process is accelerating. Dubbed the “internet of things” (IoT), the idea is to embed computing into almost every conceivable object.

Smart clothes will use a home network to tell a washing machine what settings to use; smart paving slabs will monitor pedestrian traffic in cities and give governments forensically detailed maps of air pollution. But for the IoT to reach its full potential will require some way to make sense of the torrents of data that billions of embedded chips will throw off. The IoT chips themselves will not be up to the task: the chip embedded in a smart paving slab, for instance, will have to be as cheap as possible, and very frugal with its power: since connecting individual paving stones to the electricity network is impractical, such chips will have to scavenge energy from heat, footfalls or even ambient electromagnetic radiation. Much effort is going into improving the energy efficiency of computers, for several reasons: consumers want their smartphones to have longer battery life; the IoT will require computers to be deployed in places where mains power is not available; and the sheer amount of computing going on is already consuming something like 2% of the world's electricity generation.

User interfaces are another area ripe for improvement, for today's technology is ancient. Keyboards are a direct descendant of mechanical typewriters. The mouse was first demonstrated in 1968, as were the “graphical user interfaces”, such as Windows or iOS, which have replaced the arcane text symbols of early computers with friendly icons and windows. Cern, Europe's particle-physics laboratory, pioneered touchscreens in the 1970s. Siri may leave your phone and become omnipresent: artificial intelligence and cloud computing could allow virtually any machine to be controlled simply by talking to it. Samsung already makes a voice-controlled television. Technologies such as gesture tracking and gaze tracking, currently being pioneered for

“Siri may leave your phone and become omnipresent: you will be able to control virtually any machine simply by talking to it”

virtual-reality video games, may also prove useful. Augmented reality (AR), a close cousin of virtual reality that involves laying computer-generated information over the top of the real world, will begin to blend

the virtual and the real. Google is working on electronic contact lenses that could perform AR functions.

Moore's Law cannot go on for ever. But as it fades, it will fade in importance. It mattered a lot when your computer was confined to a box on your desk, and when computers were too slow to perform many desirable tasks. It gave a gigantic global industry a master metronome, and a future without it will see computing progress become harder, more fitful and more irregular. But progress will still happen. The computer of 2050 will be a system of tiny chips embedded in everything from your kitchen counter to your car. Most of them will have access to vast amounts of computing power delivered wirelessly, through the internet, and you will interact with them by speaking to the room. Trillions of tiny chips will be scattered through every corner of the physical environment, making a world more comprehensible and more monitored than ever before. Moore's Law may soon be over. The computing revolution is not.

A longer version of this article first appeared in The Guardian. Adapted from Megatech: Technology in 2050, edited by Daniel Franklin, published by Economist Books at £15.