
# TECHNOLOGY QUARTERLY
## AFTER MOORE'S LAW

### Double, double, toil and trouble

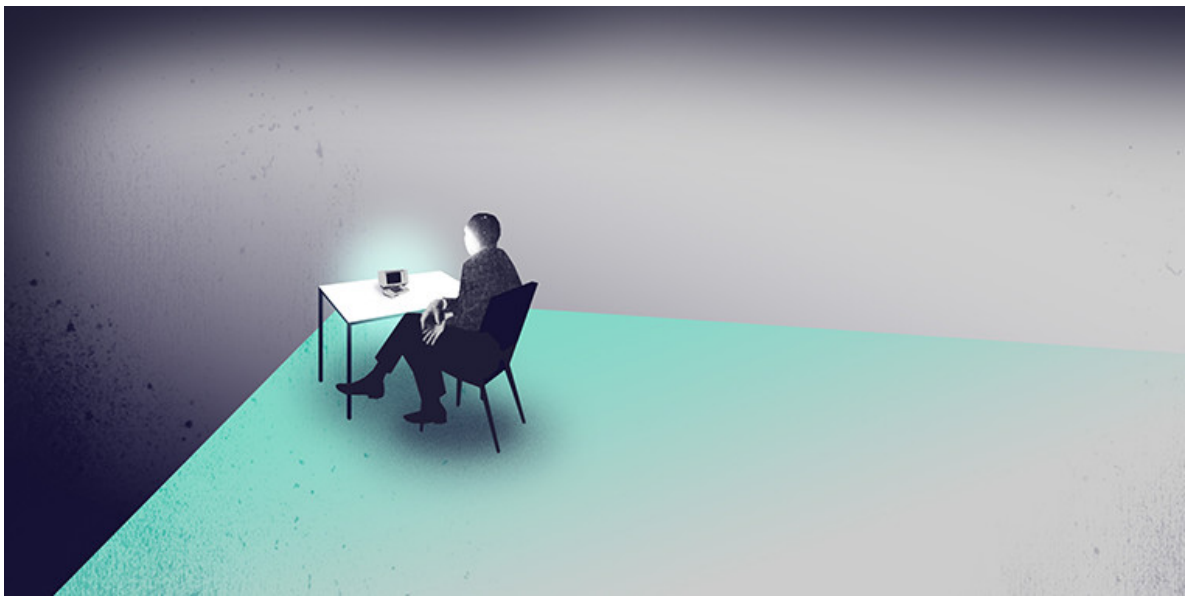
**After a glorious 50 years, Moore's law—which states that computer power doubles every two years at the same cost—is running out of steam. Tim Cross asks what might replace it**

IN 1971 a small company called Intel released the 4004, its first ever microprocessor. The chip, measuring 12 square millimetres, contained 2,300 transistors—tiny electrical switches representing the 1s and 0s that are the basic language of computers. The gap between each transistor was 10,000 nanometres (billionths of a metre) in size, about as big as a red blood cell. The result was a miracle of miniaturisation, but still on something close to a human scale. A child with a decent microscope could have counted the individual transistors of the 4004.

The transistors on the Skylake chips Intel makes today would flummox any such inspection. The chips themselves are ten times the size of the 4004, but at a

light human eyes and microscopes use. If the 4004's transistors were blown up to the height of a person, the Skylake devices would be the size of an ant.

The difference between the 4004 and the Skylake is the difference between computer behemoths that occupy whole basements and stylish little slabs 100,000 times more powerful that slip into a pocket. It is the difference between telephone systems operated circuit by circuit with bulky electromechanical switches and an internet that ceaselessly shuttles data packets around the world in their countless trillions. It is a difference that has changed everything from metal-bashing to foreign policy, from the booking of holidays to the designing of H-bombs.

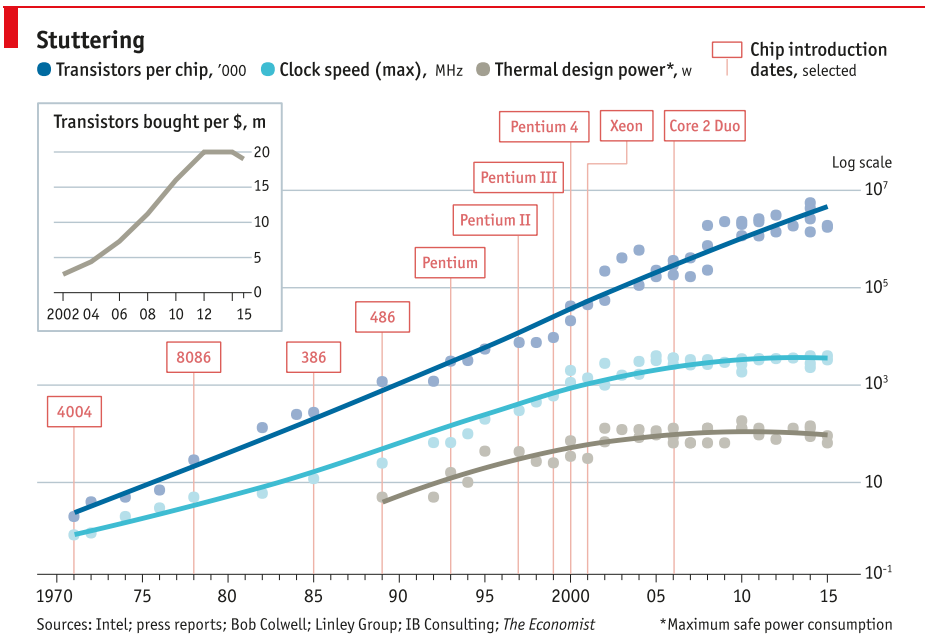


It is also a difference capable of easy mathematical quantification. In 1965 Gordon Moore, who would later become one of the founders of Intel, a chipmaker, wrote a paper noting that the number of electronic components which could be crammed into an integrated circuit was doubling every year. This exponential increase came to be known as Moore's law.

In the 1970s the rate of doubling was reduced to once every two years. Even so, you would have had to be very

all, double something 22 times and you have 4m times more of it, or perhaps something 4m times better. But that is indeed what has happened. Intel does not publish transistor counts for its Skylake chips, but whereas the 4004 had 2,300 of them, the company's Xeon Haswell E-5, launched in 2014, sports over 5 billion, just 22 nm apart.

Moore's law is not a law in the sense of, say, Newton's laws of motion. But Intel, which has for decades been the leading maker of microprocessors, and the rest of the industry turned it into a self-fulfilling prophecy.




That fulfilment was made possible largely because transistors have the unusual quality of getting better as they get smaller; a small transistor can be turned on and off with less power and at greater speeds than a larger one. This meant that you could use more and faster transistors without needing more power or generating more waste heat, and thus that chips could get bigger as well as better.

**approaching a fundamental limit of smallness: the atom**

have for decades spent heavily on R&D, and the facilities—"fabs"—in which the chips have been made have become much more expensive. But each time transistors shrank, and the chips made out of them became faster and more capable, the market for them grew, allowing the makers to recoup their R&D costs and reinvest in yet more research to make their products still tinier. The demise of this virtuous circle has been predicted many times. "There's a law about Moore's law," jokes Peter Lee, a vice-president at Microsoft Research: "The number of people predicting the death of Moore's law doubles every two years." But now the computer industry is increasingly aware that the jig will soon be up. For some time, making transistors smaller has no longer been making them more energy-efficient; as a result, the operating speed of high-end chips has been on a plateau since the mid-2000s (see chart). And while the benefits of making things smaller have been decreasing, the costs have been rising. This is in large part because the components are approaching a fundamental limit of smallness: the atom. A Skylake transistor is around 100 atoms across, and the fewer atoms you have, the harder it becomes to store and manipulate electronic 1s and 0s. Smaller transistors now need trickier designs and extra materials. And as chips get harder to make, fabs get ever more expensive. Handel Jones, the CEO of International Business Strategies, reckons that a fab for state-of-the-art microprocessors now costs around $7 billion. He thinks that by the time the industry produces 5nm chips (which at past rates of progress might be in the early 2020s), this could rise to over $16 billion, or nearly a third of Intel's current annual revenue. In 2015 that revenue, at $55.4 billion, was only 2% more than in 2011. Such slow increases in revenue and big increases in cost seem to point to an obvious conclusion. "From an economic standpoint, Moore's law is over," says Linley Gwennap,

The pace of advance has been slowing for a while. Marc Snir, a supercomputing expert at Argonne National Laboratory, Illinois, points out that the industry's International Technology Roadmap for Semiconductors, a collaborative document that tries to forecast the near future of chipmaking, has been over-optimistic for a decade. Promised manufacturing innovations have proved more difficult than expected, arriving years late or not at all.

Brian Krzanich, Intel's boss, has publicly admitted that the firm's rate of progress has slowed. Intel has a biennial "tick-tock" strategy: in one year it will bring out a chip featuring smaller transistors ("tick"); the following year it tweaks that chip's design ("tock") and prepares to shrink the transistors again in the following year. But when its first 14nm chips, codenamed Broadwell, ticked their way to market in 2014 they were nearly a year behind schedule. The tick to 10nm that was meant to follow the tock of the Skylakes has slipped too; Intel has said such products will not now arrive until 2017. Analysts reckon that because of technological problems the company is now on a "tick-tock-tock" cycle. Other big chipmakers have had similar problems.

Moore's law has not hit a brick wall. Chipmakers are spending billions on new designs and materials that may make transistors amenable to a bit more shrinkage and allow another few turns of the exponential crank. They are also exploring ways in which performance can be improved with customised designs and cleverer programming. In the past the relentless doubling and redoubling of computing power meant there was less of an incentive to experiment with other sorts of improvement.

## Try a different route

certain calculations much faster than any classical computer could ever hope to do. Another is to emulate biological brains, which perform impressive feats using very little energy. Yet another is to diffuse computer power rather than concentrating it, spreading the ability to calculate and communicate across an ever greater range of everyday objects in the nascent internet of things. Moore's law provided an unprecedented combination of blistering progress and certainty about the near future. As that certainty wanes, the effects could be felt far beyond the chipmakers faced with new challenges and costs. In a world where so many things—from the cruising speed of airliners to the median wage—seem to change little from decade to decade, the exponential growth in computing power underlies the future plans of technology providers working on everything from augmented-reality headsets to self-driving cars. More important, it has come to stand in the imagination for progress itself. If something like it cannot be salvaged, the world would look a grimmer place. At the same time, some see benefits in a less predictable world that gives all sorts of new computing technologies an opportunity to come into their own. "The end of Moore's law could be an inflection point," says Microsoft's Dr Lee. "It's full of challenges—but it's also a chance to strike out in different directions, and to really shake things up."

## More Moore: The incredible shrinking transistor

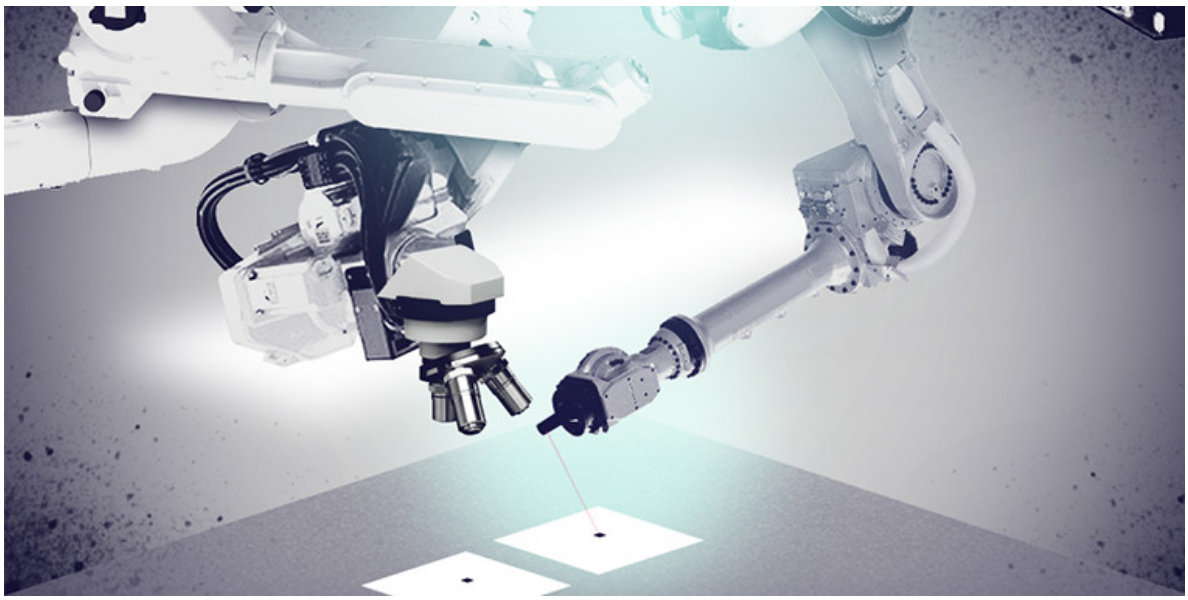
### New sorts of transistors can eke out a few more iterations of Moore's law, but they will get increasingly expensive

THANKS to the exponential power of Moore's law, the electronic components that run modern computers vastly outnumber all the leaves on the Earth's trees. Chris Mack, a chipmaking expert, working from a previous estimate by VLSI Research, an analysis firm, reckons that perhaps 400 billion billion (4x1020) transistors were churned out

small: millions could fit on the full stop at the end of this sentence.

A transistor is a sort of switch. To turn it on, a voltage is applied to its gate, which allows the current to flow through the channel between the transistor's source and drain (see first diagram). When no current flows, the transistor is off. The on-off states represent the 1s and 0s that are the fundamental language of computers.

The silicon from which these switches are made is a semiconductor, meaning that its electrical properties are halfway between those of a conductor (in which current can flow easily) and an insulator (in which it cannot). The electrical characteristics of a semiconductor can be tweaked, either by a process called "doping", in which the material is spiced with atoms of other elements, such as arsenic or boron, or by the application of an electrical field.



In a silicon transistor, the channel will be doped with one material and the source and drain with another. Doping alters the amount of energy required for any charge to flow through a semiconductor, so where two differently doped materials abut each other, current cannot flow. But when the device is switched on, the electric field from the

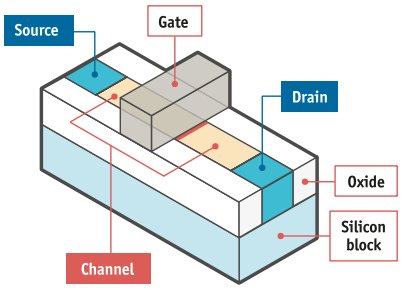
flow through.

For a long time that basic design worked better and better as transistors became ever smaller. But at truly tiny scales it begins to break down. In modern transistors the source and drain are very close together, of the order of 20nm. That causes the channel to leak, with a residual current flowing even when the device is meant to be off, wasting power and generating unwanted heat.

Heat from this and other sources causes serious problems. Many modern chips must either run below their maximum speeds or even periodically switch parts of themselves off to avoid overheating, which limits their performance.

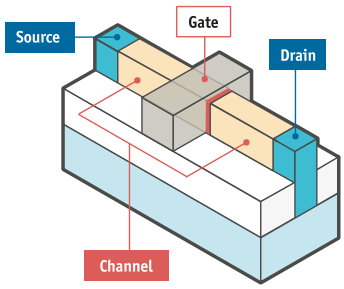
**Better by design**

**Standard transistor**




A transistor is a switch. Ordinarily, current cannot flow. When a voltage is applied to the **gate**, the **channel** becomes conductive, current flows from the **source** to the **drain**, and the transistor switches on.

**finFET transistor**




A finFET transistor raises the **channel** above the block of silicon upon which the device sits. That allows the **gate** to wrap around three sides of the **channel**, improving its electrical properties.

Source: *The Economist*

Chipmakers are trying various methods to avoid this. One of them, called strained silicon, which was introduced by Intel in 2004, involves stretching the atoms of the silicon crystal further apart than normal, which lubricates the passage of charge carriers through the channel, reducing the heat generated.

In another technique, first adopted in 2007, metal oxides are used to combat the effects of tunnelling, a quantum

the other side without ever passing through the intervening space. Developing more such esoteric techniques may allow chipmakers to go on shrinking transistors for a little longer, but not much.

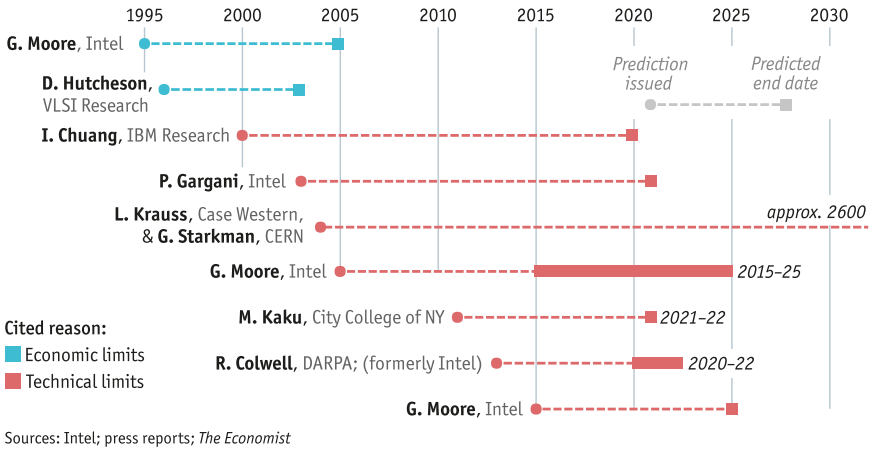
## The 3D effect

Beyond that, two broad changes will be needed. First, the design of the transistor will have to be changed radically. Second, the industry will have to find a replacement for silicon, the electrical properties of which have already been pushed to their limits.

**The design of the transistor will have to be changed radically**

One solution to the problem of leaking current is to redesign the channel and the gate. Conventionally, transistors have been flat, but in 2012 Intel added a third dimension to its products. To enable it to build chips with features just 22nm apart, it switched to transistors known as "finFET", which feature a channel that sticks up from the surface of the chip. The gate is then wrapped around the channel's three exposed sides (see second diagram), which gives it much better control over what takes place inside the channel. These new transistors are trickier to make, but they switch 37% faster than old ones of the same size and consume only half as much power.

The next logical step, says Mr Snir of Argonne National Laboratory, is "gate-all-around" transistors, in which the channel is surrounded by its gate on all four sides. That offers maximum control, but it adds extra steps to the manufacturing process, since the gate must now be built in multiple sections. Big chipmakers such as Samsung have said that it might take gate-all-around transistors to build chips with features 5nm apart, a stage that Samsung and other makers expect to be reached by the early 2020s.




Sources: Intel; press reports; *The Economist*

Beyond that, more exotic solutions may be needed. One idea is to take advantage of the quantum tunnelling that is such an annoyance for conventional transistors, and that will only get worse as transistors shrink further. It is possible, by applying electrical fields, to control the rate at which tunnelling happens. A low rate of leakage would correspond to a 0; a high rate to a 1. The first experimental tunnelling transistor was demonstrated by a team at IBM in 2004. Since then researchers have been working to commercialise them.

In 2015 a team led by Kaustav Banerjee, of the University of California, reported in Nature that they had built a tunnelling transistor with a working voltage of just 0.1, far below the 0.7V of devices now in use, which means much less heat. But there is more work to be done before tunnelling transistors become viable, says Greg Yeric of ARM, a British designer of microchips: for now they do not yet switch on and off quickly enough to allow them to be used for fast chips. Jim Greer and his colleagues at Ireland's Tyndall Institute are working on another idea. Their device, called a junctionless nanowire transistor (JNT), aims to help with another problem of building at tiny scales: getting the doping right. "These days you're talking about [doping] a very small amount of silicon indeed. You'll soon be at the point where even one or two misplaced dopant atoms could drastically alter the behaviour of your transistor," says Dr Greer.

JNTs, just 3nm across, out of one sort of uniformly doped silicon. Normally that would result in a wire rather than a switch: a device that is uniformly conductive and cannot be turned off. But at these tiny scales the electrical influence of the gate penetrates right through the wire, so the gate alone can prevent current flowing when the transistor is switched off.

Whereas a conventional transistor works by building an electrical bridge between a source and a drain that are otherwise insulated, Dr Greer's device works the other way: more like a hose in which the gate acts to stop the current from flowing. "This is true nanotechnology," he says. "Our device only works at these sorts of scales. The big advantage is you don't have to worry about manufacturing these fiddly junctions."

## Material difference

Chipmakers are also experimenting with materials beyond silicon. Last year a research alliance including Samsung, Global Foundries, IBM and State University New York unveiled a microchip made with components 7nm apart, a technology that is not expected to be in consumers' hands until 2018 at the earliest. It used the same finFET design as the present generation of chips, with slight modifications, but although most of the device was built from the usual silicon, around half of its transistors had channels made from a silicon-germanium (SiGe) alloy.

This was chosen because it is, in some ways, a better conductor than silicon. Once again, that means lower power usage and allows the transistor to switch on and off more quickly, boosting the speed of the chip. But it is not a panacea, says Heike Riel, the director of the physical-sciences department at IBM Research. Modern chips are built from two types of transistor. One is designed to conduct electrons, which carry a negative charge. The other sort is designed to conduct "holes",

as if they were positively charged electrons. And although SiGe excels at transporting holes, it is rather less good at moving electrons than silicon is.

Future paths to higher performance along these lines will probably require both SiGe and another compound that moves electrons even better than silicon. The materials with the most favourable electrical properties are alloys of elements such as indium, gallium and arsenide, collectively known as III-V materials after their location in the periodic table.

The trouble is that these materials do not mix easily with silicon. The spacing between the atoms in their crystal lattices is different from that in silicon, so adding a layer of them to the silicon substrate from which all chips are made causes stress that can have the effect of cracking the chip.

The best-known alternative is graphene, a single-atom-thick (and hence two-dimensional) form of carbon. Graphene conducts electrons and holes very well. The difficulty is making it stop. Researchers have tried to get around this by doping, squashing or squeezing graphene, or applying electric fields to change its electrical properties. Some progress has been made: the University of Manchester reported a working graphene transistor in 2008; a team led by Guanxiong Liu at the University of California built devices using a property of the material called “negative resistance” in 2013. But the main impact of graphene, says Dr Yeric, has been to spur interest in other two-dimensional materials. “Graphene sort of unlocked the box,” he says. “Now we’re looking at things like sheets of molybdenum disulphide, or black phosphorous, or phosphorous-boron compounds.” Crucially, all of those, like silicon, can easily be switched on and off.

ticking along for another five or six years, by which time the transistors may be 5nm apart. But beyond that "we're running out of ways to stave off the need for something really radical."

His favoured candidate for that is something called "spintronics". Whereas electronics uses the charge of an electron to represent information, spintronics uses "spin", another intrinsic property of electrons that is related to the concept of rotational energy an object possesses. Usefully, spin comes in two varieties, up and down, which can be used to represent 1 and 0. And the computing industry has some experience with spintronics already: it is used in hard drives, for instance.

Research into spintronic transistors has been going on for more than 15 years, but none has yet made it into production. Appealingly, the voltage needed to drive them is tiny: 10-20 millivolts, hundreds of times lower than for a conventional transistor, which would solve the heat problem at a stroke. But that brings design problems of its own, says Dr Yeric. With such minute voltages, distinguishing a 1 or a 0 from electrical noise becomes tricky.

"It's relatively easy to build a fancy new transistor in the lab," says Linley Gwennap, the analyst. "But in order to replace what we're doing today, you need to be able to put billions on a chip, at a reasonable cost, with high reliability and almost no defects. I hate to say never, but it is very difficult." That makes it all the more important to pursue other ways of making better computers.

## New designs: Taking it to another dimension

### How to get more out of existing transistors

STRICTLY speaking, Moore's law is about the ever greater number of electronic components that can be

getting better. As transistors become harder and harder to shrink, computing firms are starting to look at making better use of the transistors they already have. “Managers in the past wouldn’t want to invest a lot in intensive design,” says Greg Yeric at ARM. “I think that’s going to start shifting.”

One way is to make the existing chips work harder. Computer chips have a master clock; every time it ticks, the transistors within switch on or off. The faster the clock, the faster the chip can carry out instructions. Increasing clock rates has been the main way of making chips faster over the past 40 years. But since the middle of the past decade clock rates have barely budged.

Chipmakers have responded by using the extra transistors that came with shrinking to duplicate a chip’s existing circuitry. Such “multi-core” chips are, in effect, several processors in one, the idea being that lashing several slower chips together might give better results than relying on a single speedy one. Most modern desktop chips feature four, eight or even 16 cores.

But, as the industry has discovered, multi-core chips rapidly hit limits. “The consensus was that if we could keep doing that, if we could go to chips with 1,000 cores,

chips, programmers have to break down tasks into smaller chunks that can be worked on simultaneously. “It turns out that’s really hard,” says Dr Burger. Indeed, for some mathematical tasks it is impossible.

Another approach is to specialise. The most widely used chips, such as Intel’s Core line or those based on ARM’s Cortex design (found in almost every smartphone on the planet) are generalists, which makes them flexible. That comes at a price: they can do a bit of everything but excel at nothing. Tweaking hardware to make it better at dealing with specific mathematical tasks “can get you something like a 100- to 1,000-fold performance improvement over some general solution”, says Bob Colwell, who helped design Intel’s Pentium chips.

When Moore’s law was doubling performance every couple of years at no cost anyway, there was little incentive to customise processing this way. But now that transistors are not necessarily getting faster and cheaper all the time, those tradeoffs are changing.

## Something special

That was Sean Mitchell’s thinking when, a decade ago, he co-founded a company called Movidius. The firm designs chips for use in computer vision, a booming field with applications in everything from robotics to self-driving cars to augmented reality. Movidius has since raised nearly $90m in funding.

“When we looked at the general-purpose chips out there,” says Dr Mitchell, “we found that they were very inefficient.” So Dr Mitchell and his co-founders set about designing their own specialised microprocessor.

**Designing new chips takes years and can cost tens or even hundreds of millions of dollars**

“We’ve got to process high-resolution images, each containing millions of pixels, and coming in at 60, 90 or even 120 frames per second,” he says. By tweaking the

mathematics of visual processing while leaving out any of the extraneous logic that would allow a general-purpose chip to perform other tasks—Movidius's Myriad 2 chip can crunch huge amounts of visual information but use less than a watt of power (which is about 20% of the consumption of the chips in smartphones and only about 1% of those in desktop computers). In January the firm announced a deal with Google.

Custom-built chips are already in use in other parts of the computing industry. The best-known examples are the graphics chips used to improve the visuals of video games, designed by firms such as Nvidia and AMD and first marketed to consumers in the mid-1990s. Intel's newer Pentium chips also come with built-in specialised logic for tasks such as decoding video. But there are downsides.

Designing new chips takes years and can cost tens or even hundreds of millions of dollars. Specialised chips are also harder to program than general-purpose ones. And, by their very nature, they improve performance only on certain tasks.

A better target for specialised logic, at least at first, might be data centres, the vast computing warehouses that power the servers running the internet.

Because of the sheer volume of information they process, data centres will always be able to find a use for a chip that can do only one thing, but do it very well.

With that in mind, Microsoft, one of the world's biggest software firms and providers of cloud-computing services, is venturing into the chip-design business. In 2014 it announced a new device called Catapult that uses a special kind of chip called a field-programmable gate array (FPGA), the configuration of which can be reshaped at will. FPGAs offer a useful compromise between specialisation and flexibility, says Dr Burger, who led the

software." When one task is finished, an FPGA can be reconfigured for another job in less than a second.

The chips are already in use with Bing, Microsoft's search engine, and the company says this has doubled the number of queries a server can process in a given time. There are plenty of other potential applications, says Peter Lee, Dr Burger's boss at Microsoft. FPGAs excel when one specific algorithm has to be applied over and over again to torrents of data. One idea is to use Catapult to encrypt data flowing between computers to keep them secure. Another possibility is to put it to work on voice- and image-recognition jobs for cloud-connected smartphones.

The technology is not new, but until now there was little reason to use it. What is new is that "the cloud is growing at an incredible rate," says Dr Burger. "And now that Moore's law is slowing down, that makes it much harder to add enough computing capacity to keep up. So these sorts of post-Moore projects start to make economic sense."

At the IBM research lab on the shores of Lake Zurich, ambitions are set even higher. On a table in one of the labs sits a chip connected by thin hoses to a flask of purple-black liquid. Patrick Ruch, who works in IBM's Advanced Thermal Packaging group, sees this liquid as the key to a fundamental redesign of data centres. He and his colleagues think they can shrink a modern supercomputer of the sort that occupies a warehouse into a volume about the size of a cardboard box—by making better use of the third dimension.

Brain scan: Bruno Michel

**IBM's head of advanced micro-integration reckons biology holds the key to more energy-efficient chips**

Leaving aside innovations like finned transistors (see previous article), modern chips are essentially flat. But a number of companies, including IBM, are now working on stacking chips on top of each other, like flats in a tower block, to allow designers to pack more transistors into a given area. Samsung already sells storage systems made from vertically stacked flash memory. Last year Intel and Micron, a big memory-manufacturer, announced a new memory technology called 3D Xpoint that also uses stacking.

IBM's researchers are working on something slightly different: chip stacks in which slices of memory are sandwiched between slices of processing logic. That would allow engineers to pack a huge amount of computing into a tiny volume, as well as offering big performance benefits. A traditional computer's main memory is housed several centimetres from its processor. At silicon speeds, a centimetre is a vast distance. Sending signals across such distances also wastes energy. Moving the memory inside the chip cuts those distances from centimetres to micrometres, allowing it to shuttle data around more quickly. But there are two big problems

blowing hot air out of the server racks emit a constant roar. As more layers are added, the volume inside the chip, where the heat is generated, grows faster than the outside area from which it can be removed.

The second problem is getting electricity in. Chips communicate with the outside world via hundreds of metal "pins" on their undersides. Modern chips are so power-hungry that up to 80% of these pins are reserved for transporting electricity, leaving only a few to get data in and out. In 3D those constraints multiply, as the same number of pins must serve a much more complicated chip.

IBM hopes to kill two birds with one stone by fitting its 3D chips with miniscule internal plumbing. Microfluidic channels will carry cooling liquid into the heart of the chip, removing heat from its entire volume at once. The firm has already tested the liquid-cooling technology with conventional, flat chips. The micro fluidic system could ultimately remove around a kilowatt of heat—about the same as the output of one bar of an electric heater—from a cubic centimetre of volume, says Bruno Michel, the head of the group (see Brain scan, previous page).

But the liquid will do more than cool the chips: it will deliver energy as well. Inspired by his background in biology, Dr Michel has dubbed the liquid "electronic blood". If he can pull it off, it will do for computer chips what biological blood does for bodies: provide energy and regulate the temperature at the same time. Dr Michel's idea is a variant of a flow battery, in which power is provided by two liquids that, meeting on either side of a membrane, produce electricity.

Flow batteries are fairly well understood. The electricity industry has been studying them as a way to store intermittent power from renewable energy sources. Dr Michel's system is still many years away from commercial

are connected flickers into life—without a plug or a wire in sight.

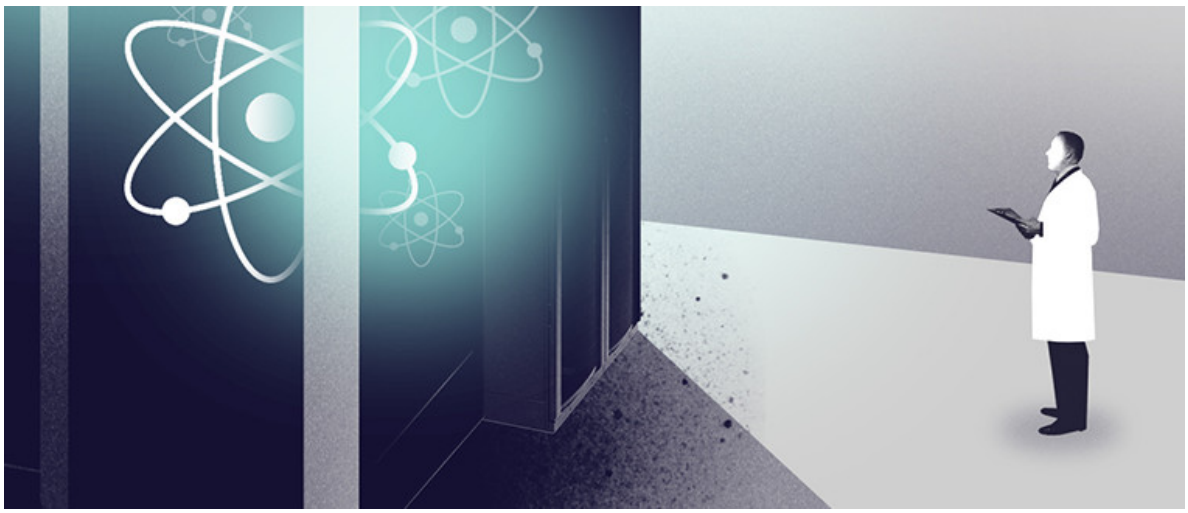
## Quantum computing: Harnessing weirdness

### Quantum computers could offer a giant leap in speed—but only for certain applications

THE D-Wave 2X is a black box, 3.3 metres to a side, that looks a bit like a shorter, squatter version of the enigmatic monoliths from the film "2001: A Space Odyssey". Its insides, too, are intriguing. Most of the space, says Colin Williams, D-Wave's director of business development, is given over to a liquid-helium refrigeration system designed to cool it to 0.015 Kelvin, only a shade above the lowest temperature that is physically possible. Magnetic shielding protects the chip at the machine's heart from ripples and fluctuations in the Earth's magnetic field.

Such high-tech coddling is necessary because the D-Wave 2X is no ordinary machine; it is one of the world's first commercially available quantum computers. In fact, it is not a full-blown computer in the conventional sense of the word, for it is limited to one particular area of mathematics: finding the lowest value of complicated functions. But that specialism can be rather useful, especially in engineering. D-Wave's client list so far includes Google, NASA and Lockheed Martin, a big American weapons-maker.

D-Wave's machine has caused controversy, especially among other quantum-computing researchers. For a while academics in the field even questioned that the firm had built a true quantum machine. Those arguments were settled in 2014, in D-Wave's favour. But it is still not clear whether the machine is indeed faster than its non-quantum rivals.



D-Wave, based in Canada, is only one of many firms in the quantum-computing business. And whereas its machine is highly specialised, academics have also been trying to build more general ones that could attack any problem. In recent years they have been joined by some of the computing industry's biggest guns, such as Hewlett-Packard, Microsoft, IBM and Google.

Quantum computing is a fundamentally different way of manipulating information. It could offer a huge speed advantage for some mathematical problems that still stump ordinary machines—and would continue to stump them even if Moore's law were to carry on indefinitely. It is also often misunderstood and sometimes overhyped. That is partly because the field itself is so new that its theoretical underpinnings are still a work in progress. There are some tasks at which quantum machines will be unambiguously faster than the best non-quantum sort. But for a lot of others the advantage is less clear. "In many cases we don't know whether a given quantum algorithm will be faster than the best-known classical one," says Scott Aaronson, a computer scientist at the Massachusetts Institute of Technology. A working quantum computer would be a boon—but no one is sure how much of one.

The basic unit of classical computing is the bit, the smallest possible chunk of information. A bit can can take

and perform all sorts of mathematical manipulations upon them. But classical machines can deal with just a handful of those bit-strings at a time. And although some of them can now crunch through billions of strings every second, some problems are so complex that even the latest computers cannot keep pace. Finding the prime factors of a big number is one example: the difficulty of the problem increases exponentially as the number in question gets bigger. Each tick of Moore's law, in other words, enables the factoring of only slightly larger numbers. And finding prime factors forms the mathematical backbone of much of the cryptography that protects data as they scoot around the internet, precisely because it is hard.

Quantum bits, or qubits, behave differently, thanks to two counterintuitive quantum phenomena. The first is "superposition", a state of inherent uncertainty that allows particles to exist in a mixture of states at the same time. For instance, a quantum particle, rather than having a specific location, merely has a certain chance of appearing in any one place.

## Wait for it

### A pipeline of new technologies to prolong Moore's magic

THE world's IT firms spend huge amounts on research and development. In 2015 they occupied three of the top five places in the list of biggest R&D spenders compiled by PricewaterhouseCoopers, a consultancy. Samsung, Intel and Microsoft, the three largest, alone shelled out $37 billion between them. Many of

magic of Moore's law. Here are a few promising ideas.

**Optical communication:** the use of light instead of electricity to communicate between computers, and even within chips. This should cut energy use and boost performance *Hewlett-Packard, Massachusetts Institute of Technology*.

**Better memory technologies:** building new kinds of fast, dense, cheap memory to ease one bottleneck in computer performance *Intel, Micron*.

**Quantum-well transistors:** the use of quantum phenomena to alter the behaviour of electrical-charge carriers in a transistor to boost its performance, enabling extra iterations of Moore's law, increased speed and lower power consumption *Intel*.

**Developing new chips and new software** to automate the writing of code for machines built from clusters of specialised chips. This has proved especially difficult *Soft Machines*.

**Approximate computing:** making computers' internal representation of numbers less precise to reduce the numbers of bits per calculation and thus save energy; and allowing computers to make random small mistakes in calculations that cancel each other out

also save energy *University of Washington, Microsoft*.

**Neuromorphic computing:** developing devices loosely modelled on the tangled, densely linked bundles of neurons that process information in animal brains. This may cut energy use and prove useful for pattern recognition and other *Qualcomm*.

**Carbon nanotube transistors:** these rolled-up sheets of graphene promise low power consumption and high speed, as graphene does. Unike graphene, they can also be switched off easily. But they have proved difficult to mass-produce *IBM, Stanford University*.

In computing terms, this means that a qubit, rather than being a certain 1 or a certain 0, exists as a mixture of both. The second quantum phenomenon, "entanglement", binds together the destiny of a quantity of different particles, so that what happens to one of them will immediately affect the others. That allows a quantum computer to manipulate all of its qubits at the same time.

The upshot is a machine that can represent—and process—vast amounts of data at once. A 300-qubit machine, for instance, could represent $2^{300}$ different strings of 1s and 0s at the same time, a number roughly equivalent to the number of atoms in the visible universe. And because the qubits are entangled, it is possible to manipulate all those numbers simultaneously.

Yet building qubits is hard. Superpositions are delicate things: the slightest puff of heat, or a passing

That is why D-Wave's machine—and all other quantum computers—have to be so carefully isolated from outside influences. Still, progress has been quick: in 2012 the record for maintaining a quantum superposition without the use of silicon stood at two seconds; by last year it had risen to six hours.

Another problem is what to build the qubits out of. Academics at the universities of Oxford and Maryland, among others, favour tickling tightly confined ions with laser beams. Hewlett-Packard, building on its expertise in optics, thinks that photons—the fundamental particles of light—hold the key. Microsoft is pursuing a technology that is exotic even by the standards of quantum computing, involving quasi-particles called anyons. Like those "holes" in a semiconductor, anyons are not real particles, but a mathematically useful way of describing phenomena that behave as if they were. Microsoft is currently far behind any of its competitors, but hopes eventually to come up with more elegantly designed and much less error-prone machines than the rest.

Probably the leading approach, used by Google, D-Wave and IBM, is to represent qubits as currents flowing through superconducting wires (which offer no electrical resistance at all). The presence or absence of a current—or alternatively, whether it is circulating clockwise or anti-clockwise—stands for a 1 or a 0. What makes this attractive is that the required circuits are relatively easy to etch into silicon, using manufacturing techniques with which the industry is already familar. And superconducting circuits are becoming more robust, too.

Last year a team led by John Martinis, a quantum physicist working at Google, published a paper describing a system of nine superconducting qubits in which four could be examined without collapsing the other five, allowing the researchers to check for, and correct,

Using a quantum computer is hard, too. In order to get the computer to answer the question put to it, its operator must measure the state of its qubits. That causes them to collapse out of their superposed state so that the result of the calculation can be read. And if the measurement is done the wrong way, the computer will spit out just one of its zillions of possible states, and almost certainly the wrong one. "You will have built the world's most expensive random-number generator," says Dr Aaronson.

**A quantum computer can represent—and process—vast amounts of data at once**

For a quantum algorithm to work, the machine must be manipulated in such a way that the probability of obtaining the right answer is continually reinforced while the chances of getting a wrong answer are suppressed. One of the first useful algorithms for this purpose was published in 1994 by Peter Shor, a mathematician; it is designed to solve the prime-factorising problem explained above. Dr Aaronson points out that alongside error correction of the sort that Dr Martinis has pioneered, Dr Shor's algorithm was one of the crucial advances which persuaded researchers that quantum computers were more than just a theoretical curiosity. Since then more such algorithms have been discovered. Some are known to be faster than their best-known classical rivals; others have yet to prove their speed advantage.

## A cryptographer's dream

That leaves open the question of what, exactly, a quantum computer would be good for. Matthias Troyer, of the Swiss Federal Institute of Technology in Zurich, has spent the past four years conducting a rigorous search for a "killer app" for quantum computers. One commonly cited, and exciting, application is in codebreaking. Dr Shor's algorithm would allow a quantum computer to make short work of most modern cryptographic codes.

long suspected: that America's National Security Agency (NSA) was working on quantum computers for exactly that reason. Last August the NSA recommended that the American government begin switching to new codes that are potentially less susceptible to quantum attack. The hope is that this will pre-empt any damage before a working quantum computer is built.

Another potential killer app is artificial intelligence (AI). Firms such as Google, Facebook and Baidu, China's biggest search engine, are already putting significant sums into computers than can teach themselves to understand human voices, identify objects in images, interpret medical scans and so on. Such AI programs must be trained before they can be deployed. For a face-recognition algorithm, for instance, that means showing it thousands of images. The computer has to learn which of these are faces and which are not, and perhaps which picture shows a specific face and which not, and come up with a rule that efficiently transforms the input of an image into a correct identification.

Ordinary computers can already perform all these tasks, but D-Wave's machine is meant to be much faster. In 2013 Google and NASA put one of them into their newly established Quantum AI Lab to see whether the machine could provide a speed boost. The practical value of this would be immense, but Dr Troyer says the answer is not yet clear.

In his view, the best use for quantum computers could be in simulating quantum mechanics itself, specifically the complicated dance of electrons that is chemistry. With conventional computers, that is fiendishly difficult. The 2013 Nobel prize for chemistry was awarded for the development of simplified models that can be run on classical computers. But, says Dr Troyer, "for complex molecules, the existing [models] are not good enough." His team reckoned that a mixed approach, combining

times of hundreds of years. Over the past three years, though, the researchers have refined their algorithms to the point where a simulation could be run in hundreds of seconds instead.

It may not be as exciting as AI or code-breaking, but being able to simulate quantum processes accurately could revolutionise all sorts of industrial chemistry. The potential applications Dr Troyer lists include better catalysts, improved engine design, a better understanding of biological molecules and improving things like the Haber process, which produces the bulk of the world's fertilisers. All of those are worthwhile goals that no amount of conventional computing power seems likely to achieve.

## What comes next: Horses for courses

### The end of Moore's law will make the computer industry a much more complicated place

WHEN Moore's law was in its pomp, life was simple. Computers got better in predictable ways and at a predictable rate. As the metronome begins to falter, the computer industry will become a more complicated place. Things like clever design and cunning programming are useful, says Bob Colwell, the Pentium chip designer, "but a collection of one-off ideas can't make up for the lack of an underlying exponential."

Progress will become less predictable, narrower and less rapid than the industry has been used to. "As Moore's law slows down, we are being forced to make tough choices between the three key metrics of power, performance and cost," says Greg Yeric, the chip designer at ARM. "Not all end uses will be best served by one particular answer."

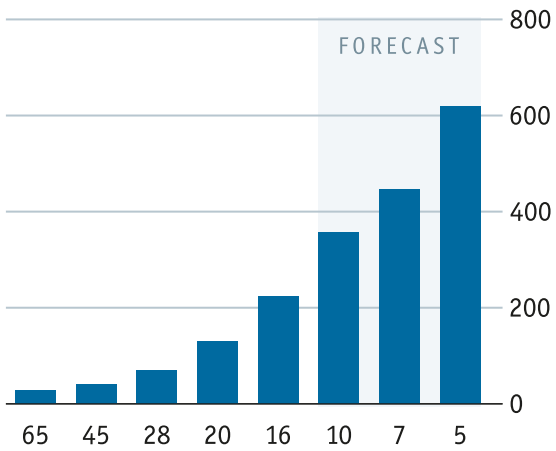
And as computers become ever more integrated into everyday life, the definition of progress will change.

Argonne National Laboratory. “They’re in it to produce useful products, and to make money.”

Moore’s law has moved computers from entire basements to desks to laps and hence to pockets. The industry is hoping that they will now carry on to everything from clothes to smart homes to self-driving cars. Many of those applications demand things other than raw performance. “I think we will see a lot of creativity unleashed over next decade,” says Linley Gwennap, the Silicon Valley analyst. “We’ll see performance improved in different ways, and existing tech used in new ways.”

**This can’t go on**

Design cost by chip component size in nm, $m




Source: IB Consulting

Mr Gwennap points to the smartphone as an example of the kind of innovation that might serve as a model for the computing industry. Only four years after the iPhone first launched, in 2011, smartphone sales outstripped those of conventional PCs. Smartphones would never have been possible without Moore's law. But although the small, powerful, frugal chips at their hearts are necessary, they are not sufficient. The appeal of smartphones lies not just in their performance but in their light, thin and rugged design and their modest power consumption. To achieve this, Apple has been heavily involved in the design of the iPhone's chips.

And they do more than crunch numbers. Besides their microprocessors, smartphones contain tiny versions of other components such as accelerometers, GPS receivers, radios and cameras. That combination of computing power, portability and sensor capacity allows smartphones to interact with the world and with their users in ways that no desktop computer ever could.

Virtual reality (VR) is another example. This year the computer industry will make another attempt at getting this off the ground, after a previous effort in the 1990s. Firms such as Oculus, an American startup bought by Facebook, Sony, which manufactures the PlayStation console, and HTC, a Taiwanese electronics firm, all plan to launch virtual-reality headsets to revolutionise everything from films and video games to architecture and engineering.

A certain amount of computing power is necessary to produce convincing graphics for VR users, but users will settle for far less than photo-realism. The most important thing, say the manufacturers, is to build fast, accurate sensors that can keep track of where a user's head is pointing, so that the picture shown by the goggles can be updated correctly. If the sensors are inaccurate, the user will feel "VR sickness", an unpleasant sensation closely

The biggest market of all is expected to be the "internet of things"—in which cheap chips and sensors will be attached to everything, from fridges that order food or washing machines that ask clothes for laundering instructions to paving slabs in cities to monitor traffic or pollution. Gartner, a computing consultancy, reckons that by 2020 the number of connected devices in the world could run to 21 billion.

## Never mind the quality, feel the bulk

The processors needed to make the internet of things happen will need to be as cheap as possible, says Dr Yeric. They will have to be highly energy-efficient, and ideally able to dispense with batteries, harvesting energy from their surroundings, perhaps in the form of vibrations or ambient electromagnetic waves. They will need to be able to communicate, both with each other and with the internet at large, using tiny amounts of power and in an extremely crowded radio spectrum. What they will not need is the latest high-tech specification. "I suspect most of the chips that power the internet of things will be built on much older, cheaper production lines," says Dr Yeric.

Churning out untold numbers of low-cost chips to turn dumb objects into smart ones will be a big, if unglamorous, business. At the same time, though, the vast amount of data thrown off by the internet of things will boost demand for the sort of cutting-edge chips that firms such as Intel specialise in. According to Dr Yeric, "if we really do get sensors everywhere, you could see a single engineering company—say Rolls Royce [a British manufacturer of turbines and jet engines]—having to deal with more data than the whole of YouTube does today."

Increasingly, though, those chips will sit not in desktops but in the data centres that make up the rapidly growing computing "cloud". The firms involved keep their financial cards very close to their chests, but making

that cloud computing grew by 30% last year and will keep on expanding at that rate at least until 2018.

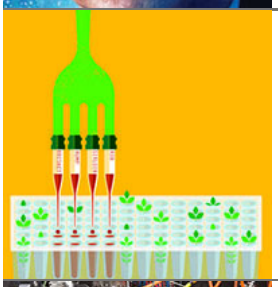
The scramble for that market could upend the industry's familiar structure. Big companies that crunch a lot of numbers, such as Facebook and Amazon, already design their own data centres, but they buy most of their hardware off the shelf from firms such as Intel and Cisco, which makes routers and networking equipment. Microsoft, a software giant, has started designing chips of its own. Given the rapid growth in the size of the market for cloud computing, other software firms may soon follow.

The twilight of Moore's law, then, will bring change, disorder and plenty of creative destruction. An industry that used to rely on steady improvements in a handful of devices will splinter. Software firms may begin to dabble in hardware; hardware makers will have to tailor their offerings more closely to their customers' increasingly diverse needs. But, says Dr Colwell, remember that consumers do not care about Moore's law per se: "Most of the people who buy computers don't even know what a transistor does." They simply want the products they buy to keep getting ever better and more useful. In the past, that meant mostly going for exponential growth in speed. That road is beginning to run out. But there will still be plenty of other ways to make better computers.

Read more from the print edition »

**Monitoring nuclear weapons: The nuke detectives**
New ways to detect covert nuclear weapons are being developed, which could help inspectors monitor Iran's nuclear deal

needed for travel to other planets



## Green food: Silicon Valley gets a taste for food

Tech startups are moving into the food business to make sustainable versions of meat and dairy products from plants

## Nuclear fusion: A big bet on small

An American company thinks it can have a commercial reactor ready and working within a decade

## The connected car: Smartphones on wheels

The way cars are made, bought and driven is changing with mobile communications