
SUPER 2025
8 + 5
= 12
POWER
ERRORS

SPENCER WILSON

To make computers better, we need to make them worse, says **Paul Marks**

Let's cut them some slack

KRISHNA PALEM'S computers won't win any awards for accuracy. Most of the time they can't even add up correctly. For them, 2 + 2 might as well be 5. But don't be fooled by the wobbly arithmetic. Palem is making machines that could represent a new dawn for computing.

Inaccuracy is not something we typically associate with computers. Since Alan Turing laid down their ground rules in the 1930s, computers have been sticklers for precision, built on the principle of following step-by-step instructions in an exact and reproducible manner. They should not make errors.

But maybe we should cut them some slack. Letting computers make mistakes could be the best way to unlock the next wave of smart devices and prevent high-performance computing hitting a wall. It would allow us to run complex simulations that are beyond today's supercomputers – models that better predict climate change, help us design more efficient cars and aircraft, and reveal the secrets of galaxy formation. They may even unlock the biggest mystery of all, by letting us simulate the human brain.

Until now, we have had to accept a trade-off between performance and energy efficiency: a computer can be either fast or low-powered, but not both. This not only means that more powerful smartphones need better batteries, but also that supercomputers are energy guzzlers. Next-generation "exaflop" machines, which are capable of 10^{18} operations a second, could consume as much as 100 megawatts, the output of a small power station. So the race is on to make computers do more with less.

One way is simply to reduce the amount of time spent executing code – the less time taken, the less power used. For programmers, this means looking for ways to get the desired result more quickly. Take the classic travelling-salesman problem of finding the shortest route around a group of cities. It's notoriously tough to solve, given that the number of possible routes shoots up exponentially with the number of cities. Palem, a computer scientist at Rice University in Houston, Texas, says that coders often settle for a route that they estimate to be about half as good as the best, because to do better would use up too much computer time. A more recent version of this approach is to use a machine-learning algorithm to arrive at an approximate result for a given piece of code. This rough answer – like a back-of-the-envelope calculation – can then be used each time the program runs instead of executing the original section of code itself.

But saving energy by cutting corners in software only goes so far. To really save power, you need to change the way the hardware works. Computers can save vast amounts of energy simply by not operating all their transistors at full power all the time, but as we'll soon see, this means sacrificing accuracy. Palem's team is hobbling computers so that they get their sums wrong in an acceptable way. "What we are proposing is to alter the computer itself to give you cheaper but slightly less accurate answers," says Palem. Take any algorithm that you think does a good job, and he will solve it inexactly with a different physical system under the hood.

Standard computer chips use a sliver of silicon called a channel to act as a switch that can flip between on (1) and off (0). The switching is controlled by a gate that stops a current flowing through the channel until you apply a voltage. Then the gate opens like a sluice in a dam, letting current through. ›

But it's finicky. This complementary metal-oxide semiconductor (CMOS) technology only works well when it has a reliable 5-volt power supply. Start to lower that and the channel becomes unstable – sometimes switching, sometimes not.

In 2003, Palem, then at the Georgia Institute of Technology in Atlanta, saw trouble coming. It was clear that the ability of the electronics industry to continue doubling the number of transistors on a chip every 18 to 24 months – a miniaturisation trend known as Moore's law – was coming to an end. Miniaturisation was introducing errors at the chip level. This was largely due to overheating and interference, or crosstalk, between the densely packed transistors. "It was quite likely ultra-small devices would become quite unstable," says Palem. Power was now the critical issue. What if you could harness instabilities in some way that would also save energy?

Palem's answer was to design a probabilistic version of CMOS technology that was deliberately unstable. His team built digital circuits in which the most significant bits – those representing values that need to be accurate – get a regular 5-volt supply, but the least significant bits get 1 volt. "The significant bits are running at a proper, well-behaved voltage, but the least significant get really slack," says Palem. As many as half the bits representing a number can be hobbled like this.

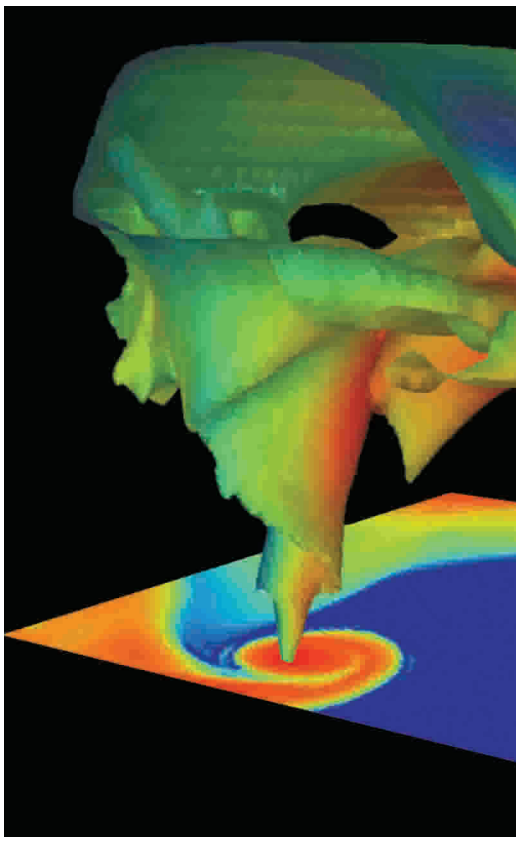
This means that Palem's version of an adder, a common logic circuit that simply adds two numbers, doesn't work with the usual precision (see "Missing bits", below). "When it adds two numbers, it gives an answer that is reasonably good but not exact," he says. "But it is much cheaper in terms of energy use."

Spread that over billions of transistors and you have a significant power saving. The trick is to choose applications for which the least significant bits don't matter too much: for example, using a large range of numbers to represent the colour of a pixel. In one experiment, Palem and his colleagues built a digital video decoder that interpreted the least significant bits in an imprecise way when converting pixel data into screen colours. They found that human viewers perceived very little loss in image quality. "The human eye averages a lot of things out," says Palem. "Think about how we see illusions. The brain does a lot of work to compensate."

Encouraged by that success, the Rice University researchers have moved on to another application involving the senses: hearing aids. Their initial tests show that inexact digital processing in a hearing aid can halve power consumption while reducing intelligibility by only 5 per cent. The results suggest that we could use such techniques to slash the power used by smartphones and personal computers, given that these are basically audiovisual devices.

Tim Palmer, a climate physicist at the University of Oxford, sees even greater potential. He thinks that computers based on Palem's ideas could be the answer to what is presently an intractable problem: how to improve the accuracy of climate predictions for the next century without waiting years for a new generation of supercomputers.

"The crucial question about climate change centres on the role of clouds," says Palmer, in terms of whether they amplify or dampen the effects of global warming. "You can't really answer that question with any great confidence unless you can simulate cloud systems directly." And right now, it's not clear how to do that.



NOMAD/NATIONAL GEOGRAPHIC CREATIVE

Modelling the crucial role of clouds in climate change is too costly with today's computers

Today's supercomputers don't have the brawn to do it, and their successors, expected in the next decade or so, will just be too energy-hungry. "Based on current estimates, the amount of power needed for such a machine is going to be around 100 megawatts," says Palmer, five to 10 times what today's top supercomputers use. Assuming they don't just melt, running them could prove prohibitively costly.

"Doing 20 calculations inexactly may be more useful than 10 done exactly"

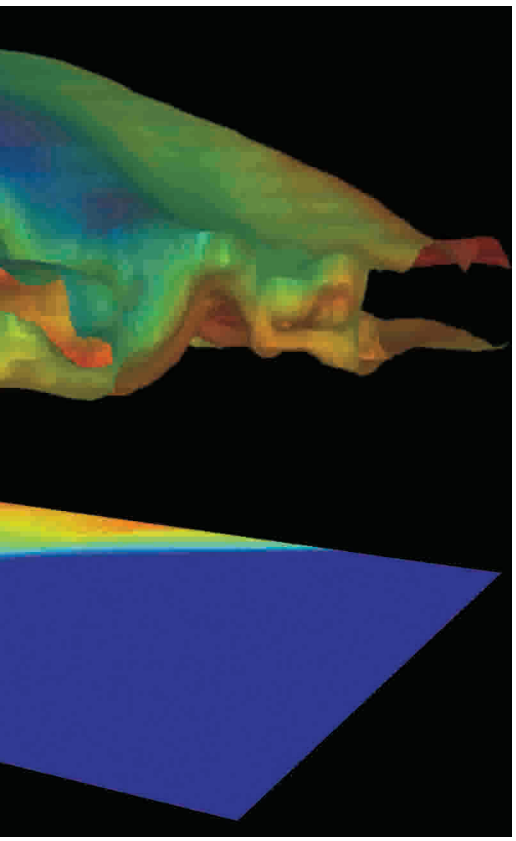
Supercomputers burn so much power because they are generally optimised for computing with 64-bit-long numbers. In principle, this gives greater accuracy. But climate models involve millions of variables, simulating complex interacting factors such as winds, convection, temperatures, air pressures, ocean temperatures and salinity. The result, says Palmer, is that they have too much power-draining data to crunch. What's needed, he says, is for different variables to be

MISSING BITS

Krishna Palem is building an inexact computer that saves power by relaxing its precision. With this set-up, 8 + 5 could equal a range of values.

Here's why: in Palem's system, transistors representing the least significant digits in a number are deliberately run at a lower than ideal voltage. This makes them unstable, prone to flipping from 1 to 0 or 0 to 1. In a 16-bit number, the eight least significant bits could be incorrect. "As many as half the bits can be flipping," says Palem.

As a simplified example, consider adding two 4-bit numbers. In binary, 8 is 1000 and 5 is 0101, and adding them should give 1101, or 13. But if the two least significant – rightmost – bits can flip, then the result could become 1100 (which is 12), 1110 (14) or 1111 (15). It is possible that the inputs, 8 and 5, might get corrupted too: 1000 (8) could turn into 1001 (9), 1010 (10) or 1011 (11); and 0101 (5) into 0100 (4), 0110 (6) or 0111 (7). Finally, the result of adding any two of these numbers could become corrupted as well, leading to a range of inexact answers.



represented in data strings of varying length, depending on their importance to the model.

Chipmakers are starting to accommodate such needs. Nvidia has launched a graphics processor unit, the TX1, that is capable of "mixed-precision" processing, allowing software to switch between 16 and 32-bit operation as it runs. But Palmer wants to see Palem's inexact chips adopted too. "If we can reduce the number of bits that you need to do calculations, that would have an enormous impact on energy consumption," he says. Palmer and his colleagues are talking to supercomputer makers like IBM and Cray about developing a new breed of energy-efficient hybrid machines that allow varying levels of accuracy, and that may even adopt Palem's strategy. And Palem's team is working on mixed-precision computing with the US government's Argonne National Laboratory in Illinois and the European Centre for Medium-Range Weather Forecasts in Reading, UK.

The pay-offs could be huge. Today's climate models tackle Earth's atmosphere by breaking it into regions roughly 100 kilometres square and a kilometre high. Palmer thinks inexact computing would get this down to cubes a kilometre across – detailed enough to model individual clouds.

"Doing 20 calculations inexactly could be much more useful than 10 done exactly," says Palmer. This is because at 100-kilometre scales, the simulation is a crude reflection of reality. The computations may be accurate, but the model is not. Cutting precision to get a finer-grained model would actually give you greater accuracy overall. "It is more valuable to have an inexact answer to an exact equation than an exact answer to an inexact equation," he says. "With the exact equation, I'm really describing the physics of clouds."

Degrees of accuracy

You can't just give up on accuracy across the board, of course. "There is no doubt that if you represent all of the variables in the climate model with just 16 bits rather than 64 bits it would be a disaster: it would fail very quickly," says Palmer. The challenge is choosing which parts can be treated more crudely than others.

Researchers are attacking the problem from several different angles. Mostly, it comes down to devising ways to specify thresholds of accuracy in code so that programmers can say when and where errors are acceptable. The software then computes inexactly only in parts that have been designated safe.

Approximation is not the answer for everyone, however. Rashid Mansoor is a London-based computer scientist and entrepreneur who invented Adbrain, an algorithm that tracks millions of web users as they move between devices. He is now looking at ways to speed up computing done in the cloud. But Mansoor sees inexactness as a last resort. "We don't yield approximate results for the sake of speeding up computation," he says. "We'd consider it to be cheating."

Even so, Stan Posey, who heads Nvidia's high-performance computing team in Santa Clara, California, sees a host of applications for which inexact computing will make a big difference, including accident investigations. After the Columbia space-shuttle disaster of 2003, caused by a chunk of insulating foam breaking off and making a hole in a wing, Posey and his colleagues spent countless hours simulating scenarios that could have led to this happening. He thinks inexact computing would now allow a number of simulations an order of magnitude higher within the same time frame. Scenarios that are worth looking into more closely can then be pursued with higher accuracy.

Some think inexact simulations could ultimately help us understand the brain. Supercomputers like IBM's Blue Gene are being used to model neurological functions in the Human Brain Project, for example. But there is a huge discrepancy in power consumption between the brain and a supercomputer, says Palmer (see "Power hungry"). "A supercomputer needs megawatts of power, yet a human brain runs on the power of a light bulb." What could account for this?

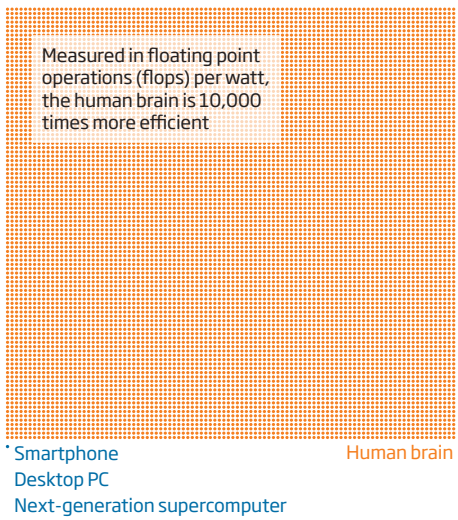
Palmer and colleagues at the University of Sussex in Brighton, UK, are exploring whether random electrical fluctuations might provide probabilistic signals in the brain. His theory is that this is what lets it do so much with so little power. Indeed, the brain could be the perfect example of inexact computing, shaped by pressure to keep energy consumption down.

What's clear is that to make computers better, we need to make them worse. Palmer is convinced that partly abandoning Turing's concept of how a computer should work is the way forward if we are to discover the true risks we face from global warming. "It could be the difference between climate change being a relatively manageable problem and one that will be an existential problem for humanity."

And if approximate computing seems a shaky foundation on which to build the future of computing, it's worth remembering that computers are always dealing with the abstract. "All computing is approximation," says Posey. Some are just more approximate than others. ■

Power hungry

Despite huge advances in performance, conventional **computers** are no match for the **human brain** when it comes to efficiency



Paul Marks is a technology journalist based in London