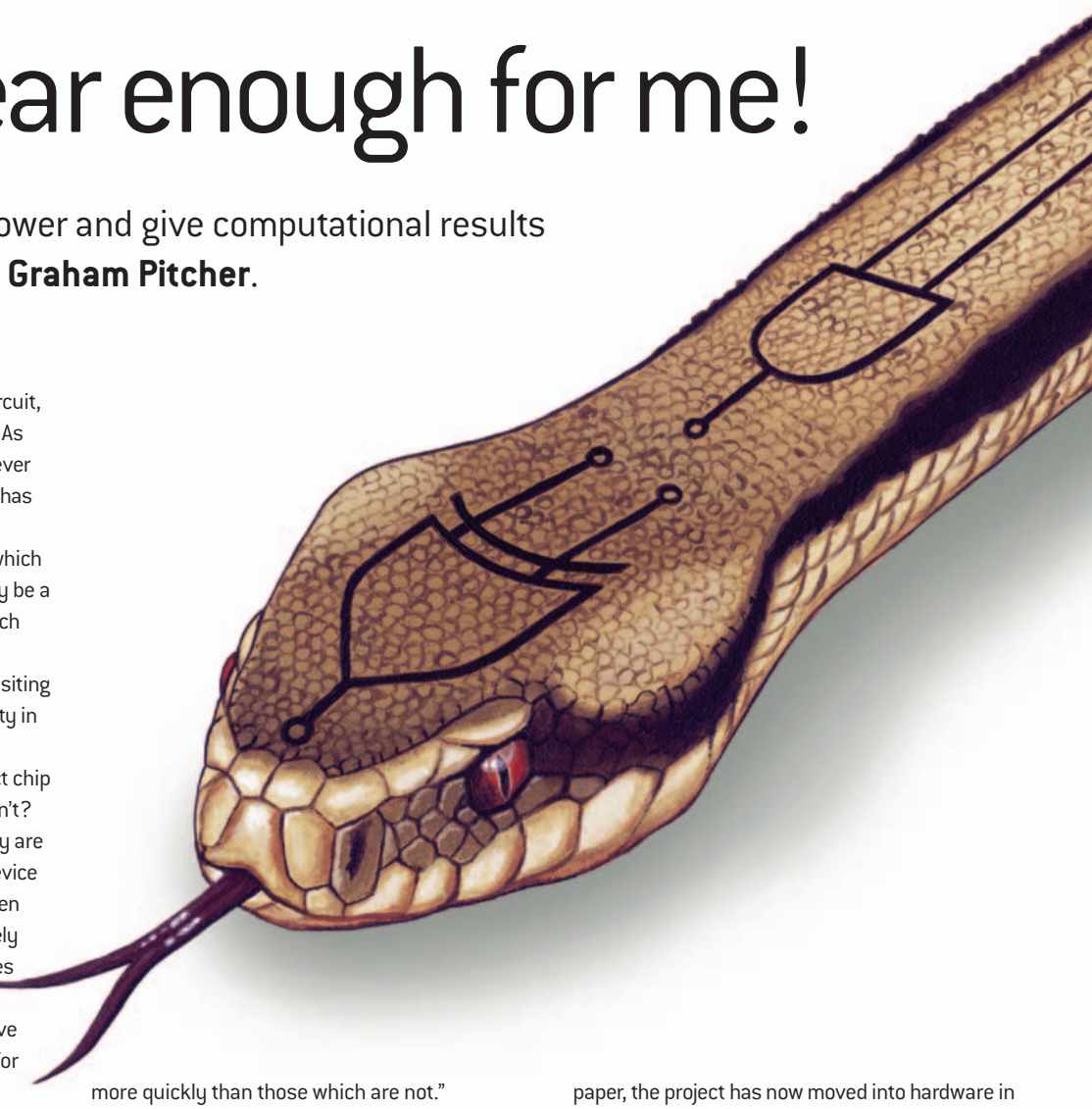
That's near enough for me!

Inexact chips may save power and give computational results which are 'acceptable'. By **Graham Pitcher**.



Since the invention of the integrated circuit, the industry has striven for accuracy. As manufacturing technology moves to ever smaller dimensions, that need for accuracy has grown in importance.

However, there is one school of thought which takes the view that inaccuracy might actually be a useful concept. To prove the theory, a research group led by Krishna Palem, a professor of computing at Rice University in the US and visiting professor at Nanyang Technological University in Singapore, has created inexact chips.

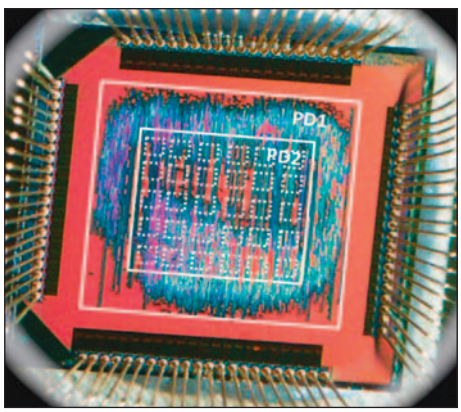
What is the difference between an inexact chip and one designed to be correct, but which isn't? Prof Palem said most people like to think they are correct in their designs. "But any physical device will have a tiny chance of being incorrect, even though the designer has tried to be completely accurate all the time. Our approach introduces inexactness through the use of probability.

"Most building blocks used in circuits have very specific functionality," he continued. "For example, take an adder block; it will add two numbers and give you an exact result – if you asked it to add 8 and 5, it would give 13 as the result. An inexact chip relaxes that, so if you asked an inexact adder to add 8 and 5, the answer might be 12 one time, 15 the next or it might give 13. It all depends upon how you introduce the inexactness."

Prof Palem's interest in inexactness has built over the last decade. The project was started in 2002 as a conceptual framework to explore the use of inherently incorrect hardware. The work was initially published in 2003 and a US patent was granted in 2005.

"The idea came when I was developing a contribution to a Feynman Lecture on Computation," he said. "I was looking to explore the link between thermodynamics – the cost of producing a piece of information – and the fact that randomised software algorithms tend to run more quickly than those which are not."

The idea is that an inexact chip could be computationally more effective, performing operations more quickly and saving power, although the final result may not be 100% accurate. While the theory has been proven on paper, the project has now moved into hardware in an attempt to validate the theory experimentally.



The work is driven, in part, by the ever smaller features of electronic circuits. Prof Palem believes there will be a fundamental limit 'at some time in the future' where devices will become inherently random because of the decreasing feature size. "We don't have randomness in devices today, but we may in the future," he added.

In this respect, Prof Palem's work resonates with that being undertaken by Prof Asen Asenov at Glasgow University on statistical variability at ever smaller feature sizes. "In our approach," said Prof Palem, "we have developed a model which can be used to take advantage of imperfections."

Prof Palem continued on the theme of decreasing feature size when explaining two ways in which to induce inexactness in chip designs – probabilistic or using an input drawn from a probability distribution. "Either approach

CREDIT: Rice University/CSEM/NTU

Images produced using traditional processing elements (left), inexact processing hardware with a relative error of 0.54% (middle) and a relative error of 7.58% (right). The right hand image is about 15 times more efficient in terms of speed, space and energy than the pristine image (left).

decides which data stays and which gets vacated – something we call pruning."

He said that, in the future, transistors will be so small that they will be vulnerable to background imperfections. Expanding, he used the analogy of trees.

"Today, transistors are like oak trees," he offered. "But at some time in the future, they will be more like saplings and vulnerable to effects such as the wind. Unlike fully grown trees, these saplings won't be standing upright all the time.

"If you think of the shadows cast by the saplings, they will move all the time – and that's what we need when designing inexact chips."

But the analogy can also be applied to full grown trees. "If there's a group of trees, some may not be serving their purpose, so you could cut them down or prune them." Pruning is all about cutting out seldom used parts of a design. Tests in 2011 showed 'pruned' chips were twice as fast, used half as much energy and were half the size of their exact counterparts.

Rice graduate student Avinash Lingamneni noted: "We showed that pruning could cut energy demands by 3.5 times with chips that deviated from the correct value by an average of 0.25%. When we factored in size and speed gains, these chips were 7.5 times more efficient than regular chips. Chips that got wrong answers with a deviation of about 8% were up to 15 times more efficient."

Despite the apparently radical approach of inexactness, Prof Palem's concept doesn't need a new design approach. "Inexact chips can be designed using standard eda tools because of our 'pruning' approach and users don't have to change their design flow. What does need to change is for the designer to decide where there is the opportunity to take a more relaxed approach to what is correct."

In a paper outlining the approach, Prof Palem says conventional design automation tools map application primitives to logic gates using algorithms rooted in Boolean logic. "We have studied the probabilistic counterpart of this logic; probabilistic Boolean logic. This logic is composed of Boolean logic primitives with an associated probability of correctness. Well formed probabilistic Boolean formulae – like their deterministic counterparts – can be constructed from the Boolean constants, Boolean variables and probabilistic Boolean operators."

He suggested that the verification effort for a complex SoC could be an example of where a more relaxed approach could be applied. As chips get larger, verification becomes the dominant part of the design effort. Relaxing the rules to some extent would speed the design process and may well result in a device which performs in an acceptable fashion.

Prof Palem: "If you asked an inexact adder to add 8 and 5, the answer might be 12 one time, 15 the next or it might give 13."

The test chips which the project has created were designed by CSEM in Switzerland. "It's a halfway house between industry and academia," said Prof Palem. One of the first examples was a device which featured 25 adders in a 5 x 5 matrix, produced on a 180nm process. Since then, the project has created more complex devices. On the agenda is a fast Fourier transform chip on a 65nm cmos process. "In parallel," he said, "we are building the infrastructure to create graphics processors and hearing aid electronics."

These two examples may seem strange, but have been picked for a purpose – essentially, they don't need the results of their computations to be spot on. "We are creating graphics processors which are glitchy and which reproduce errors," said Prof Palem. "The results might not be of the same quality as another graphics processor, but that isn't important." The reason? Our brains and eyes are trained to deal with this and can create better images than they see. The hearing aid electronics take advantage of the same approach: the brain will compensate for some reduction in sound quality.

Prof Palem said his group's work has been supported from the start by Intel. "It has been actively involved in the project as it developed," he noted. "And ideas based on inexact computing are beginning to surface elsewhere in the industry."

After a decade of development, the inexact processor could be close to commercialisation. "I hope that we will have commercially available devices – such as the hearing aid electronics – available sometime in 2013," he concluded.