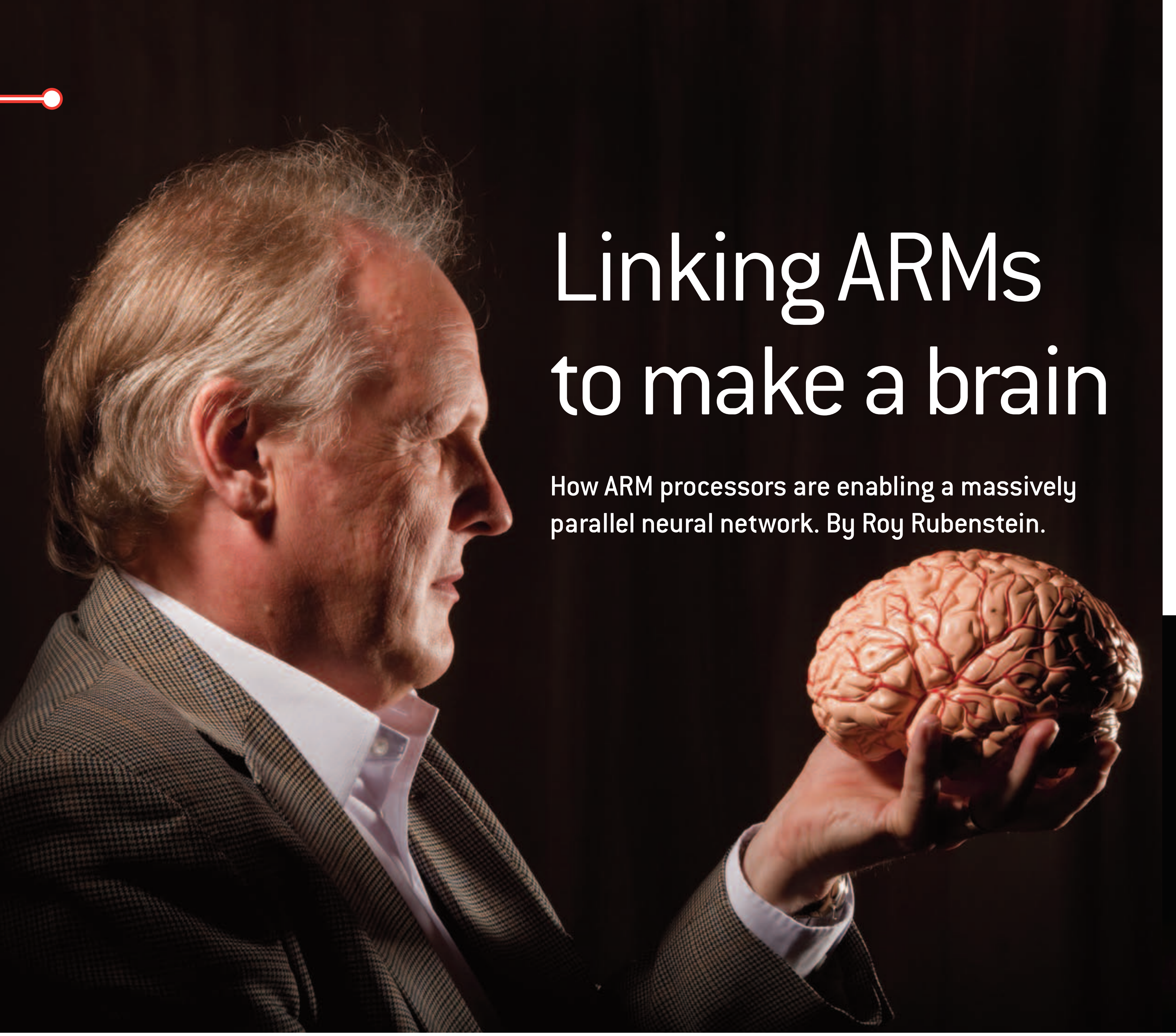
Linking ARMs to make a brain

How ARM processors are enabling a massively parallel neural network. By Roy Rubenstein.



For someone who has spent his working life in computing, Professor Steve Furber has taken his time before alighting on the human brain as a topic for research.

His interest in alternative modes of computing started in 1998 after two decades of chip design and computer science. "I'd spent 20 years in conventional computing and, in those years, computers got an awful lot faster, but there were still things they couldn't do; interesting things," said Furber, ICL Professor of Computer Engineering at the University of Manchester.

Prof Furber became interested in computing while undertaking a doctorate in aerodynamics at Cambridge University. He joined a fledgling student society dubbed the Cambridge Processing Unit Group. "A bunch of students that liked to build computers for fun," he said.

Involvement in the society brought him to the attention of Chris Curry and Hermann Hauser (see box), cofounders of Acorn Computers, a pioneering UK desktop company. And it was while at Acorn that Prof Furber led the design of the 32bit ARM microprocessor.

Prof Furber could not have known how hugely successful the ARM processor would prove. More than 20billion ARM cores have now been shipped and this year, according to Hauser, the value of all ARM based chips will exceed Intel's chip sales. "It not just the number of processors that outpaces Intel, but the value of these chips is also bigger than Intel's sales," said Hauser.

Prof Furber's ARM design achievement continues to be recognised. In 2010, he became one of three 2010 Millennium Technology Prize laureates, a Finnish award for life enhancing innovations.

ARM has had a recurring role in Prof Furber's work. He and his team developed a clockless version of the ARM core as part of research into asynchronous chip design. Now, ARM is the centrepiece of a £5million Engineering and Physical Sciences Research Council (EPSRC) funded project to develop a massively parallel neural network computer. The project, led by Prof Furber and his team, also involves the Universities of Cambridge, Sheffield and Southampton.

His interest in the brain began during research into associative memory.

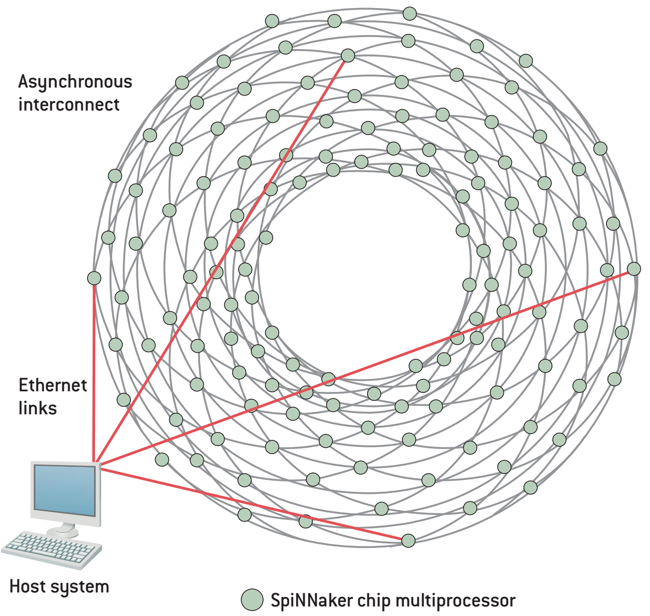
"I liked using associative memory – I used it on the ARM for caches – but it's very brittle," said Prof Furber. "If you give it the things it knows, it recognises them perfectly; if you give it something it doesn't quite know, it fails totally."

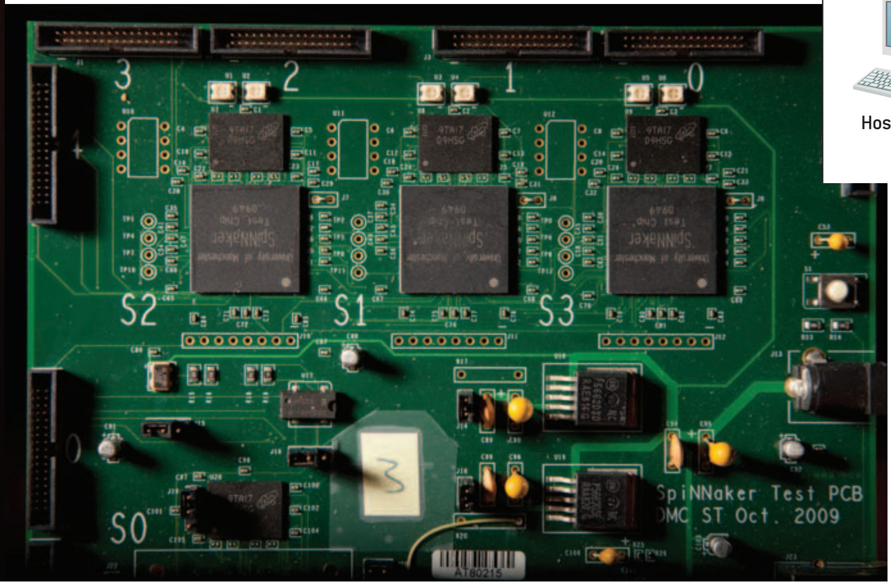
What he found was that whichever approach he took to 'soften' the memory, he kept reinventing neural networks. "In the end I threw in the towel: 'blow this, what I'm interested in is neural networks'," he said.

A fundamental issue with neural network research is that the workings of the brain remain unknown. This led Prof Furber to consider how he could bring computing engineering skills to address the problem. He began with neural science literature and conversations with fellow academics active in neural network research. "The neural science literature is vast; finding bits that matter to the approach you want to take is hard," he said. He admits the research, which he started in 1998, has been a long and slow process.

Prof Furber identified that few people had tackled very large neural networks and he set about exploring how electronics could be used to scale such networks. One flash of insight occurred in 2005, when he realised that people didn't know how much of the biological detail of the neuron – the processing and transmitter cell that makes up the brain – is essential for information processing and how much is to do with its evolution to achieve fundamental tasks, like finding energy and staying alive. "If you cast your neuron into hardware, you have taken an irreversible decision that might be wrong," said Prof Furber. "Those bottom level bits of biology have to be in software, they have to be modifiable."

Fig 1: The SpiNNaker system





This insight proved the final step that has led to the SpiNNaker architecture, Prof Furber's massively parallel neural network.

SpiNNaker architecture

Neurons communicate by transmitting electrical spikes – asynchronous events that encapsulate information in the firing patterns. The SpiNNaker architecture is designed to model such spiking neurons.

The key challenge in building a very large neural model is the communication, says Prof Furber. In the human brain, enormous resources are dedicated to communication: 100billion (10^{11}) neurons, linked by a quadrillion (10^{15}) connections. Since recreating such a vast number of connections electronically is not practical, the design team has exploited the fact that electronic communication is far faster than the biological equivalent. The spikes are thus encapsulated as packets, which are whizzed across links. "We make sure they get to everywhere they need to go in zero biological time, which means much less than a millisecond," said Prof Furber.

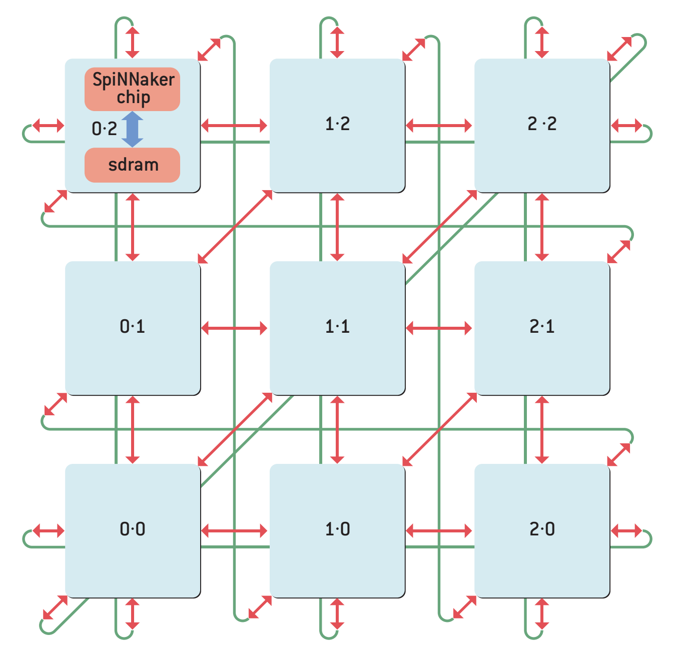
The SpiNNaker architecture comprises a 2d mesh of SoCs arranged as a toroid. Each chip comprises 18 ARM cores and a communication network on chip (NOC). One of the ARM cores is assigned as a monitor processor and one is kept as a spare, with the remaining 16 modelling neurons.

Each ARM core will model 1000 neurons and Prof Furber hopes to scale the architecture to 64,000 SoCs, such that the full size ARM based neural network will process 1billion neurons, equivalent to 1% of the human brain. "We are going to run real time up to a large scale," said Prof Furber.

The communication NOC on each chip ensures that neurons, in the form of packets, get distributed as required. Each 40bit packet comprises an 8bit system management field and a 32bit unique neuron identifier. The NOC's router sends the source neuron packet to all the processors modelling neurons to which it is connected. These are typically on the same chip, but can reside on other chips. A packet is routed every clock cycle.

A host computer is used to set up the neural network. The host looks at the neural network program that is to be run and allocates neurons to each core and chip. It then computes the neural network's initial weights – parameters that define the network's workings – loads them and broadcasts a 'go' signal to start the network. This is fine for certain sized neural networks, says Prof Furber, but as these grow in size, it becomes

Fig 2: The SpiNNaker mesh detail



increasingly difficult to find a host computer big enough to do the configuration. Prof Furber plans to use the SoC devices themselves to do such large system configurations by making use of self timing techniques.

Managing the power consumption of such a massively parallel computer has also been an area of focus. The SoC itself is being implemented using a 130nm cmos process. While such a mature process has the disadvantage of higher dynamic power consumption, compared to the latest 32 and 45nm nodes, the issue of leakage power is avoided. In turn, the architecture's software model is event driven, such that processors are asleep when no interrupts occur. Interrupt events include an incoming neuron spike, a direct memory access transfer completing or a 1ms timer. The timer is used to drive the real time neuron model dynamics.

Speed is also traded for lower power with regard the 'lump of memory' used alongside each SoC. Each SoC needs a large memory store to hold the neural network's vast synaptic data and low power sdram, designed for the mobile handset market, has been chosen.

Prototype SpiNNaker chips based on two ARM cores have been made and tested by several neural network users. "A cheaper design on which to make mistakes," is how Prof Furber puts it. These chips have proved the riskier parts of the design and have given the design team confidence with regard the full SoC design. Meanwhile, the 18 core design is now running as a simulation and its physical layout is largely complete. "We'll have first samples in the fourth quarter this year," said Prof Furber.

Prof Furber plans to use the samples to build a 50 chip system by year end, which he hopes will be quickly followed by a 500 chip system. "We are going up in factors of ten," said Prof Furber. "The next step will be a 5000 node system, where we'll have some fairly serious power and thermal engineering to do. I hope to see that towards the end of 2011." The full machine, with 50,000 to 60,000 nodes, is expected in two years' time.

One of the first applications to be mapped onto the computer will be the work of psychologists at Manchester University. They have developed a neural network that learns to read by converting text to speech. Once the neutral network is trained to read, the psychologists can damage its various components selectively to reproduce symptoms of the brain damaged patients they see in the clinic.

"At the moment, because of limitations in their machines, they are limited to fairly simple vocabularies," said Prof Furber. "Our project is to take this to the next step, so they can get it richer, more detailed and with larger vocabularies." The aim is to develop tailored therapies for patients to help them recovery the ability to read.

Another area of interest is robotics. A neural network is a real time system with input and input, Prof Furber explained. "We'd like to start connecting SpiNNaker machines where the input is the robot's sensors and the outputs are the robot's actuators." The SpiNNaker architecture will not need to be fitted directly to the robot, a wireless network can be used instead.

While Prof Furber is delighted with the work's progress, he points out that their efforts have been focused on chip design. "I don't think we have contributed anything to neural networks at this stage," Prof Furber concluded. "However, I'll be disappointed if, in a year from now, that is still the case."

Hermann Hauser on Prof Steve Furber

Hermann Hauser, one of the founders of Acorn Computers, interviewed Steve Furber three years before he joined the company full time. Hauser describes Prof Furber as one of the smartest people he has met. "I have had the privilege of meeting many extraordinarily gifted people and living here in Cambridge, you do meet them now and then," he said. "One the greatest jobs I have had was leading the R&D department at Acorn Computers. It became clear in our design meetings that Steve just came out with outstandingly brilliant solutions to difficult problems." Hauser, cofounder of venture capital firm Amadeus Capital Partners, is keeping an entrepreneurial eye on Prof Furber's latest work. For him, the cleverness resides not in the scale of the ARM based neural network architecture, but in the way in which the ARM cores interact and the network that links them. "Nobody can predict the effect it will have," said Hauser, who describes the hardware as a 'sand pit' for neural scientists to explore highly parallel systems. "There is potential in the way the demonstrator works that one can build a computer that can do certain things others cannot," he said. Tasks like scene analysis, picture recognition and sophisticated database searches. "The sort of things humans are very good at and computers are not," he concluded.