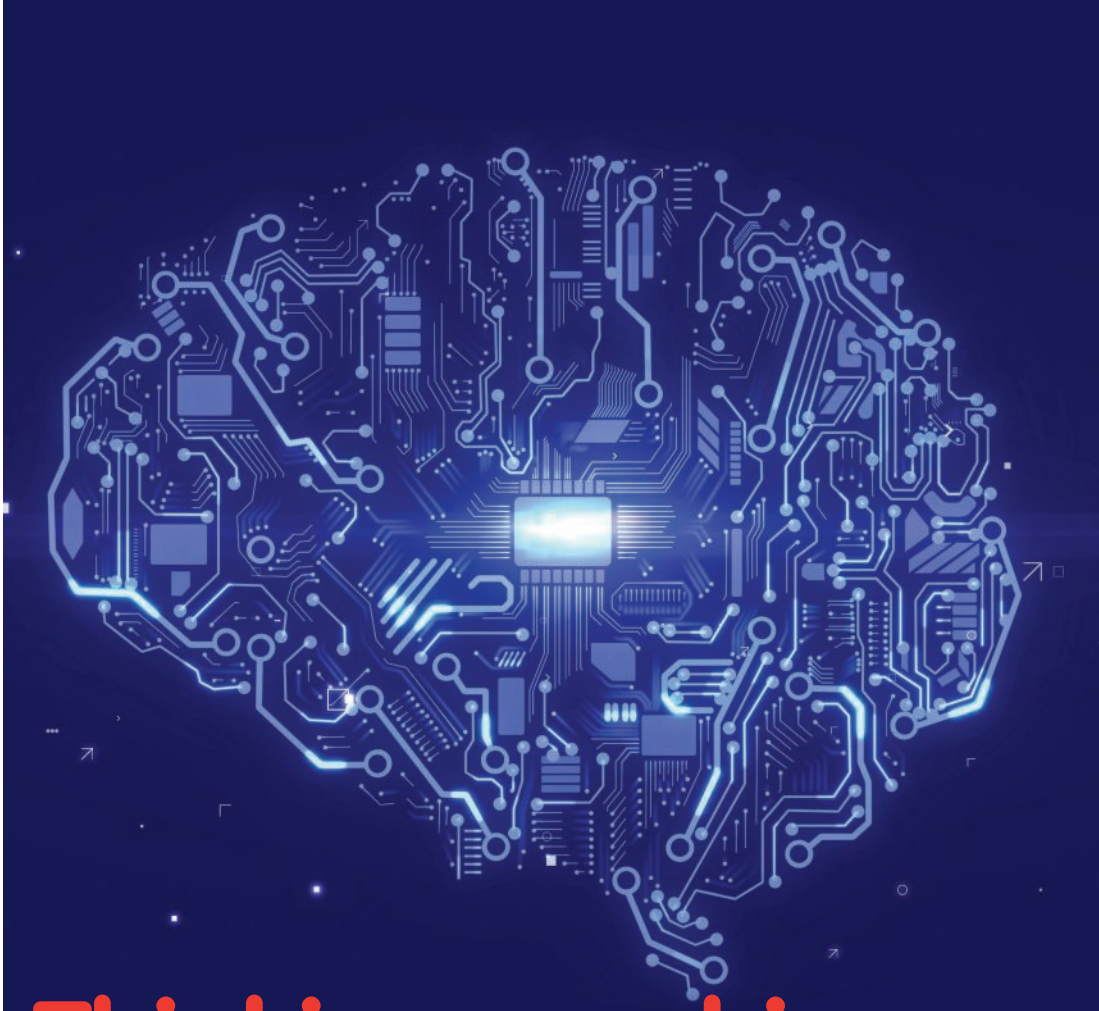

Thinking machines

While artificial neural networks underpin much of the research into AI, could they be leading us up a blind alley? By **Chris Edwards**

Today, artificial neural networks underpin much of the research into artificial intelligence (AI) and machine learning. But they face major issues that may see the whole architecture become a blind alley in the search for efficient AI.

A major problem with artificial neural networks, particularly those used in deep learning, is how power hungry they are. That has meant parcelling the work off to remote cloud servers. But for real-time systems, the finite speed of light presents a difficulty. Denis Dutoit, strategic marketing manager at research institute CEA-Leti, said at a workshop held as part of the VLSI Symposia in Honolulu earlier this year: "We have to process more data at the edge to understand the information that is coming from sensors. If we need a decision made in 1ms or less, the server that does that can't be more than 150km away."

Engineers working in AI based on neural networks are trying to use techniques such as pruning and approximate computing to reduce the workload of edge computers expected to handle AI applications. But machine-learning researchers are all too aware they are a long way from what is possible. It is easy to see why that is because the active human brain gets by on a power budget of just 20W. Although the neural network took its inspiration from the brain's structure, it maybe did not take enough.

The key question is why is the brain so efficient? Some argue that much of what drives the brain's processing lies in the way that neurons process stimulus. Whereas artificial neural networks process arithmetic values and thresholds, the brain relies on trains of electrical spikes that move through the synapses that connect neurons together. Peter van der Made, founder and CTO of start-up Brainchip argues the spiking behaviour can be processed by the simplest of neural models implemented in the company's chips, rejecting even an influential expressive model based on just a few differential equations developed by Eugene Izikevich 15 years ago.

Few are convinced that the brain has given up its secrets so easily and are looking to build more accurate models of its working. "Nobody knows how much detail you can strip away before you strip away something important," says Professor Steve Furber of the University of Manchester and leader of the SpiNNaker programme there.

In early November, about ten years after they produced the formal proposal for its construction, the team led by Furber celebrated the completion of a machine that can deploy a million Arm cores spread across 1200 boards to simulate roughly one percent of the human brain. Each core has the capacity to simulate a thousand neurons and a million synapses – reflecting the ratio of processing to interconnect found in biological brains.

In that decade, the team has learned a lot. SpiNNaker set out largely as a hardware-oriented project: using Arm cores built from asynchronous logic in a massively parallel machine with the depth of logical interconnectivity found in biological brains.

"The [million-core] machine brings new ideas to achieving reliability," Furber says. "At that scale it's not all going to work at the same time, ever. Most of it will: you don't have the situation where 30 per cent of it isn't operational at any one time. It's

more like one per cent or a fraction of one per cent. But you have to know which fraction it is. Developing the redundancy and software framework to do that has been a challenge."

Now part of Europe's Human Brain Project, SpiNNaker is the digital foil to the University of Heidelberg's predominantly analogue BrainScalesS architecture. Some see analogue-domain processing as a possible endpoint of neuromorphic research. Barbara de Salvo, deputy director of science and long-term research at CEA-Leti, says resistive memories have properties that lend themselves to energy-efficient spike-based processing but much depends on what more extensive digital simulations tell researchers.

SpiNNaker's programmability lets researchers explore the many models of neural interactions neuroscience has explored. Thanks to this flexibility, SpiNNaker has spread around the world as an increasing number of researchers have got involved in neuromorphic computing and brain science and found they need digital computers optimised for the task to simulate their models. With many located in European labs, SpiNNaker systems now sit as far afield as Sandia Laboratories in the US and Auckland University of Technology.

"It's intrinsically a very interdisciplinary game," Furber says. "We are computer engineers. We bring expertise in microchip design. But if we are going to build artificial brains we don't have the expertise to do that. The Human Brain Project provides an excellent environment for collaboration."

The software development continues, including techniques to load the enormous quantities of data needed to support simulations with the complexity of a brain of a small mammal. Most jobs today are much smaller, partly because load times outweigh actual experimental time in most situations. A recent change to the software in which the models are generated within the machine from a high-level description. "It's an obvious step but it's non-trivial getting it to run."

Second generation

With the learning from the original SpiNNaker project, the Manchester group has teamed up with colleagues at the University of Dresden to develop a second generation. SpiNNaker-2 provides the opportunity to move to a new process technology – the 22nm FD-SOI process offered by the German university's neighbourhood fab operated by GlobalFoundries.

"With SpiNNaker-1 we were building to a very tight budget. The 130nm technology we used was a mature process even then," Furber says. "We can be more aggressive on the process now, which will give us an order of magnitude improvement in performance."

The FD-SOI process also provides the opportunity to use techniques such as adaptive body biasing to reduce energy consumption.

Although SpiNNaker-2 keeps the Arm processor architecture at its heart, the revised design takes advantage of additional silicon budget to deploy accelerators. "We've worked out what the system spends a lot of its time doing," Furber says. "Learning is still in a rapid state of flux. Everything is evolving. We want to keep the software flexibility of SpiNNaker."

The Arm processor will be supported by dedicated accelerator that the kinds of exponential functions commonly used in mathematical models of neurons as well as a hardware random number generation. Using the Arm Cortex-M4 as the main processor core, there is a greater emphasis on double-precision floating-point arithmetic than before. Although, as Furber points out, biology is noisy and likely to be amenable to computing platforms that use approximate computing, the higher accuracy is needed to support the kinds of models scientists need to run to find out which approximations can be made.

"Often the users are coming from a background of running code on double-precision supercomputers and clusters. When offered SpiNNaker, the first question they ask is will they get the same answer?"

The tape-out target for the first complete SpiNNaker-2 processor, which will contain 144 Arm cores is April 2020. The run-up to that is a series of smaller design projects intended to reduce the risk inherent in a large MPSoC design. The first processor-based test chip, called Santos, was used to test a quad-core design on a 28nm bulk CMOS process. The second, which saw first silicon in recent months, deployed two quad elements on the FD-SOI process. "We are planning one more prototype. That will be to de-risk the memory controller design. Modern DRAM memories have got really scary in their interfaces," Furber says.

The final MPSoC will be powerful enough to, in effect, put the equivalent of an insect brain on a chip. "Something of the scale of an insect brain on a drone that would be an interesting place to be," Furber says, pointing to the EPSRC-funded Brains-on-Board project as an example of the kind of group that might implement such a design. A SpiNNaker-2 machine on the same scale as today's million-core system would support simulations of around 10 per cent of the human brain. "It would take another couple of steps get the whole brain, assuming the aim is to get to that point," Furber says.

With the ability to run many, smaller models at high speed, computer architects and neuroscientists may be able to unlock enough of the secrets before full-sized brain simulations get underway and allow the insights from the software-based SpiNNaker to distill into custom hardware.

"We are computer engineers. We bring expertise in microchip design. But if we are going to build artificial brains we don't have the expertise to do that. The Human Brain Project provides an excellent environment for collaboration. "
Prof. Steve Furber