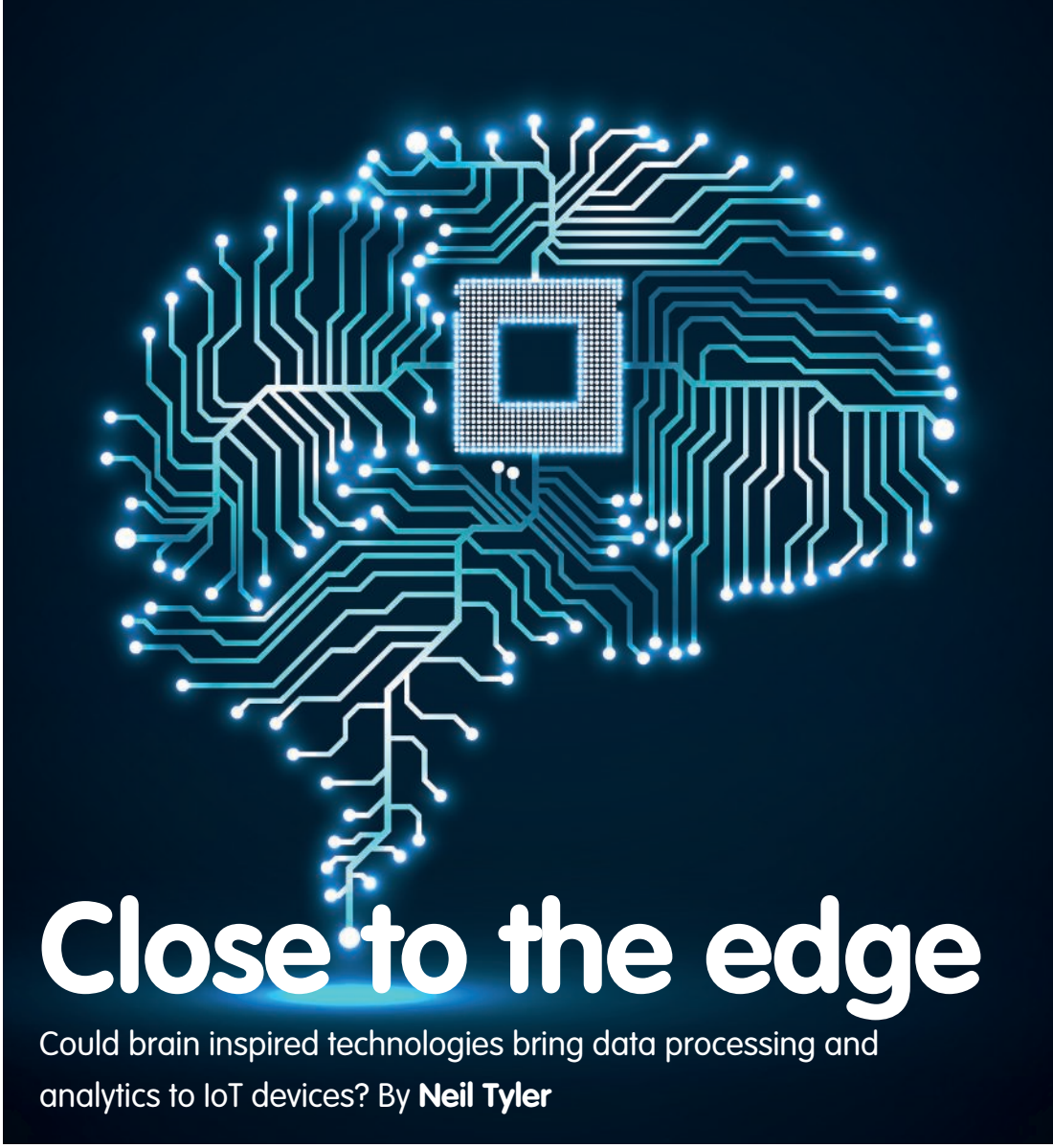

Close to the edge

Could brain inspired technologies bring data processing and analytics to IoT devices? By **Neil Tyler**

"We are entering a new era where artificial intelligence (AI) systems are helping to shape the future world," said CEA-Leti's chief scientist, Barbara De Salvo.

Speaking earlier this year, she described a number of emerging technologies, such as neuromorphic hardware and ultra-low power microdevices, that are helping to create a radically new, digital communication architecture for the Internet of Things (IoT) with analytics processing taking place at the edge and at the end devices instead of in the Cloud

"With billions of easy-access and low-cost connected devices, the world has entered the era of hyper-connectivity, enabling people and machines to interact in a symbiotic way with both the physical and cyber worlds," De Salvo said. "AI is at the centre of this revolution."

Speaking at ISSCC 2018, De Salvo said that this architecture will include human-brain inspired hardware, coupled with computing paradigms and algorithms that "will allow for distributed intelligence over the whole IoT network, all-the-way down to ultralow-power end-devices."

There's a growing consensus that the potential efficiencies of processing data at the edge, rather than at distant data centres or in the Cloud, will be significant, but reaching that long-term goal will be a challenge. For example, IoT battery-powered devices lack the processing power to analyse data and a power source to support data processing.

Transformative approaches will be needed to "address the enduring power-efficiency issues of traditional computing architectures," and De Salvo called for a "holistic research approach to the development of low-power architectures inspired by the human brain, where process development and integration, circuit design, system architecture and learning algorithms are optimised."

De Salvo contended that optimised neuromorphic hardware provided a highly promising solution for future ultralow-power cognitive systems that could extend well beyond the IoT.

"Emerging technologies such as advanced CMOS, 3D technologies, emerging resistive memories, and silicon photonics, coupled with novel brain-inspired paradigms, such as spike-coding and spike-time-dependent-plasticity, have extraordinary potential to provide intelligent features in hardware, replicating the way knowledge is created and processed in the human brain," she said.

De Salvo added that work looking at how the brain operates, was helping researchers to better understand the emergence of connectionism, novel neuroimaging techniques and the functioning of neural networks, "all of which may provide models for brain-inspired technologies."

She noted that the convergence of miniaturisation, wireless connectivity, increased data-storage capacity, and data analytics, was helping to position the IoT at the epicentre of profound social, business and political changes.

She pointed to significant gains in the performance and applications of machine learning, driven by vast data storage in images, videos, audio and text files. These gains have been essential to the dramatic improvement of learning/training approaches and algorithms, as well as the increased computational power of computers. This includes parallel computing for neural network processing, which has compensated for the slowing down of Moore's Law below the 10nm node.

According to De Salva, deep learning is the most popular machine-learning field.

"Today, for tasks such as image or speech recognition, machine-learning applications are equalling

Credit: vladystock/stock.adobe.com

or even surpassing expert human performance," she said. "Other tasks considered as extremely difficult in the past, such as natural language comprehension or complex games, have also been successfully tackled."

Future applications will require even more analysis, understanding of the environment and intelligence, and machine-learning algorithms will require even more computing power to become pervasive.

Intelligence to the Edge

"Bringing intelligence to the edge or to end-devices means doing useful processing of the data as close to the collection point as possible, and allowing systems to make some operational decisions locally, even semi-autonomously," De Salvo explained.

Controlling real-time distance learning locally will be essential for many applications, whether that's landing drones or navigating driverless cars and De Salva said any delays caused by having to send data to the Cloud could lead to disastrous results.

"Privacy will also require that key data doesn't leave the user's device, while transmission of high-level information, generated by local neural-network algorithms, will have to be authorised," she said.

De Salva warned that the use of millions of cameras for example, would require data to be locally analysed, as sending it to the cloud would likely result in bandwidth issues and communication costs.

"We need new concepts and technologies that can bring AI closer to the edge and end-devices," she explained.

"The primary design goal in distributed applications covering several levels of hierarchy, is to find a global optimum between performance and energy consumption," De Salvo continued. "This requires a holistic research approach, where the technology stack is redesigned."

According to De Salvo that process is underway and companies are addressing embedded applications by developing specialised edge platforms that can execute machine-learning algorithms on embedded hardware.

She noted impressive power improvements (down to a few watts) by exploiting Moore's Law and by using hardware-software co-optimisation.

To optimise energy efficiency, research has focused on hardware designs using Convolutional Neural Network (CNN) accelerators, De Salvo noted. Off-chip storage devices, such as DRAMs, significantly increase power consumption, but mobile apps using low-power programmable deep-learning accelerators can consume less than 300µW.

Power requirements

Bringing intelligence into low-power IoT-connected end-devices that support applications such as habitat and medical monitoring, will be significantly more difficult than traditional networked mobile devices at the edge, according to De Salvo.

"Most connected end devices are wireless sensor nodes containing microcontrollers, wireless transceivers, sensors and actuators," she said. "The power requirement for these systems is critical – less than 100µW for normal workloads – as these devices often operate using energy-harvesting sources or a single battery over several years."

De Salvo said scientists inspired by the human brain are now pursuing radically different approaches to neuromorphic systems.

"They are implementing bio-inspired architectures in optimised neuromorphic hardware to provide direct one-to-one mapping between the hardware and the learning algorithm running on it," she said.

These architectures include spike coding, which encodes neuron values as pulses or spikes rather than analogue or digital values, and spike-timing-dependent-plasticity, a bio-inspired algorithm that enables unsupervised learning.

The human brain's intelligence and efficiency are strongly linked to its extremely dense 3D interconnectivity, there are approximately 10,000 synapses per neuron, and billions of neurons in the human brain cortex.

"The hierarchical structure in the cortex follows specific patterns, through vertical arrangements or µcolumns, where local data flow on subcortical specialised structures, and laminar interconnections, which foster inter-area communications build the hierarchy.

"Based on these considerations, it is clear that emerging 3D technologies, such as through-silicon vias and 3D monolithic integration, also called CoolCube, will be a key enabler for efficient neuromorphic hardware," she said.

Outlining silicon technologies that will be vital in creating brain-inspired hardware, De Salvo cited resistive memories or ReRAM, Fully Depleted Silicon on Insulator and silicon photonics.

"Thanks to its suitability for low-power design, FDSOI technology is a great candidate for neuromorphic hardware," she said.

In deep-learning architectures, high-performance reconfigurable digital processors based on 28nm FDSOI have already shown power consumption in the range of 50mW, a level of power efficiency achieved by introducing optimised data-movement strategy and exploiting FDSOI back-biasing strategies.

De Salvo noted that a large-scale multi-core neuromorphic processor called Dynap-SEL, based on 28nm FDSOI, had also been demonstrated.

"New materials to interface devices with living cells and tissues, new design architectures for lowering power consumption, data extraction and management at the system level, as well as secured communications are the next domains that I expect will experience intense development in the years ahead," De Salvo concluded.

"We've entered a period where there have been leaps in technology and the provision of more benefits, including other ways of connecting."
Barbara De Salvo