Article:

Menon, P.S. orcid.org/0000-0003-4693-7934, Hatta, S.F.W.M. orcid.org/0000-0001-9519-073X and de Souza, M.M. orcid.org/0000-0002-7804-7154 (2023) After 75 Years of the transistor: an age of neuromorphic computing [women in electronic devices]. IEEE Electron Devices Magazine, 1 (2). pp. 57-58. ISSN 2832-7683

https://doi.org/10.1109/med.2023.3292619

Post 75 years of the transistor: An age of neuromorphic computing

PS Menon, SFWM Hatta, MM de Souza

Aligned to the IEEE's mission to foster diversity, equity and inclusion (DEI) in technological innovation and excellence, the Electron Devices Society (EDS) has committed to an ambitious target to increase the percentage of women in EDS by 2% annually over the next 5 years. The aim is to create a vibrant and inclusive community of both women and men collectively using their diverse talents to innovate for the benefit of humanity. In this spirit, this news column is contributed by WiEDS (Women in EDS).

Transistors are considered the workhorse of the microelectronics revolution that led to a transformative impact on society. 75 years since the birth of the transistor (1947-2022), however, brings home the fact that traditional Von Neumann computers, which are based on the sequential execution of instructions fetched from memory, now face a bottleneck in terms of computing power. They struggle with certain types of tasks that the human brain can perform effortlessly, such as pattern recognition, sensory and parallel processing. The brain operates at time scales of milliseconds and spends more energy on synaptic transmission rather than information processing. Neuromorphic computing, on the other hand, is an evolution of artificial intelligence (AI) that draws inspiration from brain-inspired architectures and algorithms to design efficient and intelligent computing systems that are more advanced, yet, more energy efficient than Von Neumann machines. The term "neuromorphic" refers to broadly two classes of systems: the first refers to generic spike based neural networks for the exploration of large-scale computational neuroscience models. The second consists of hybrid electronic neural architectures that rely on CMOS circuits and memristors for edge applications such as the Internet of Things (IOT) which aims to connect millions of sensors into a seamless and intelligent environment. Neuromorphic computing (in both forms) has many broad applications in robotics, brain-machine interfaces, cognitive computing, image and speech processing and wearable devices, in applications that are pivotal to almost every aspect of society, such as transport, healthcare, medical diagnosis, energy, new materials discovery and defence.

In the case of hybrid neural architectures, synapses are emulated by memristor devices whose conductance varies in proportion to the applied signal. Synapses exhibit plasticity which refers to their capability to change conductance based on the history of applied signals. Unlike traditional binary computation, synapses are engineered to exhibit analog behaviour. Typical memristors include Phase change Memory (PCM), Resistive RAM (RERAM), Magnetic RAM (MRAM), ferroelectric RAM (FERAM). In liquid environments, electrolyte gated transistors based on ionogels, polymer electrolytes, ionic liquids or even organic electrolytes such as chitosan, gelatine and fish collagen have been demonstrated. Amorphous oxide-based memristors are also promising candidates due to their low process temperature and higher reliability in comparison to organic memristors.

The memristors enable a densely interconnected large-scale network of artificial neurons. Neurons undertake the task of generating appropriate electrical signals (thresholding) based on the sum of currents from the memristor array and transmitting them to other neurons. A simplest neuron may be classed as a perceptron, giving a binary output of either a '0' or a '1'. In general, neurons tend to be more complex continuously differentiable functions, that output continuous values between '0' and '1', such as the Rectified Linear Unit 'ReLu'. Because the currents in a memristor array can be summed in rows and columns, arrays can automatically undertake addition by Kirchoff's law and multiplication by Ohm's law, which is referred to as "compute-in-memory". Synapses hold the weights of the training algorithm and can be constantly updated based on the error in the output, in a process known as supervised learning. Networks may be unidirectional (feedforward) or cyclic (recurrent). Typically, neuromorphic neural networks use spikes to minimise power consumption. More recent approaches use delay systems to enable compatibility with biological time scales such as for example gesture recognition or robotics or cardiac monitoring.

The most energy consuming component of the neural network is the Analogue to Digital converter (consuming up to 60% of the energy of the system). Therefore, although array-based computing holds great promise, its most significant challenge remains the carbon footprint for training and implementation, sometimes done on two different platforms, that increases the error. Moreover, device non-linearities and other resistive drops in the network from line resistances and sources/sinks, high cost of writing in resistive memories, and low on/off ratio in magnetic memories, process variation due and filamentary operation, reliability and endurance, result in loss of accuracy when used in floating point calculations. To date, analogue computations using memristor arrays can be done with increasing inaccuracy from ~ 4-11 bits, or higher, unlike their Von Neumann counterparts. Although neuromorphic computing is bright and promising, it's the challenge of simultaneously delivering an energy efficiency, compatible with that of the human brain, and accuracy compatible with that of a Von Neumann computer, that will remain a driver in the decades to come.