Chapter 8
Low-Power High-Speed Neuromorphic Integrated Circuits

Ramesh Kumar
National Institute of Technology, Patna, India

Bal Chand Nagar
https://orcid.org/0000-0001-7984-4086
National Institute of Technology, Patna, India

ABSTRACT

The term neuromorphic is the biological activity of neurons. Neuromorphic integrated circuits (ICs) mimic the biological characteristics of the nervous system. The neural networks inspire neuromorphic IC creation in the brain to construct hardware suitable for the current need for high-speed data processing. Memristor is a fourth fundamental element proposed by Leon O. Chua in 1971 and physically fabricated by HP lab in 2008. It is the most suitable candidate for neuromorphic computing. Memristor uses considerable promise in synaptic circuits in neuromorphic systems by providing energy-efficient memory and processing capabilities. Understanding the electrical properties of neuronal membranes allows neurons to retain their resting potential and generate spikes. A neuron circuit provides the underlying mechanisms of the action potentials, including signal production and membrane potential function. This chapter describes the fundamental implementation of neuron models and their applications in spiking neural networks (SNNs). It also highlights challenges in designing neuromorphic ICs.

INTRODUCTION

Neuro Chips are semiconductor-based integrated circuits (ICs) that model human neural structure and function. They achieve this by applying neuroscience principles to developing more efficient computing systems. The Von Neumann-based computing architecture currently used needs to improve regarding real-time learning, energy efficiency, and parallel processing. Instead of these liabilities, neuromorphic IC model structures and behaviors like biological neural networks are suitable for high-speed data processing. Low-power (LP) and high-speed functionality are crucial in neuromorphic chip designs.

DOI: 10.4018/979-8-3693-8084-0.ch008

With such innovations, various applications can be bound, from mobile devices and edge computing to brain-machine interfaces (BMIs) and large-scale artificial intelligence (AI) systems. Traditional computing issues can be realized as the architectures of the central processing unit (CPU) and graphics processing unit (GPU) are so architecturally constrained. A Von Neumann bottleneck in such architectures separates the processing and memory components and demands continuous data transfer between these components, thereby suffering additional power and latency overheads. Such factors become much more challenging, especially in machine learning (ML) applications, where volumes of data are lost and often processed in real-time. Usual systems primarily rely on clocked operations, where each instruction executes in separate time steps. Such a procedure could be more efficient for processing real-time tasks that also involve parallelism and require adaptability on the run; the tasks include sensory processing, pattern recognition, and the decision-making capabilities of biological brains. This article discusses the techniques, components, and innovations toward developing LP, high-speed neuromorphic ICs.

BACKGROUND

To attain low-power (LP) consumption and high speed, neuro chips are employed to implement event-driven processing, non-volatile memory elements, analog computation, and local processing. One of the innovations that matter in *neuron* circuits is their event-driven, or spiking, nature of computation. Biologically, the neurons spike only when the cell receives enough to cross a certain threshold. Therefore, mimicking nature, neuromorphic systems only process information when there is activity, which results in saving lots of power for minimizing unnecessary computations. The spike-based computation enables neuromorphic ICs to operate asynchronously, so signals compute only when they are produced and not according to some fixed clock cycle. This results in efficient system operation with power consumption that is non-compromising on speed (Thakur et al., 2018). Most systems comprise synaptic weights and data, where refreshing is continuously required; thus, the refreshing process significantly contributes to increased power consumption. Many neural circuits incorporate non-volatile memory elements, such as memristors, phase change materials, or resistive random-access memory (ReRAM) (Schuman et al., 2022). These elements store the weight and data after the power is switched off, saving lots of power. The synapse stores data and weight, which results in energy-efficient computation without always requiring power. Non-volatile memory lets a neuromorphic system preserve its present state even after power-down. It saves the need for frequent refresh cycles of the stored data, saving more energy. Neuromorphic circuits typically use analog computation to continuously process signals, which requires fewer transistors for specific operations and facilitates smoother signal transitions, making them more power-efficient than digital processing (Nagar & Paul, 2018). An analog signal can be applied to mimic the continuous variation of synaptic weights or *neuron* membrane potentials, as the tasks of pattern recognition and sensory processing heavily depend on these very parameters in bio-inspired circuits. Applying analog computation combined with digital control mechanisms within neuromorphic ICs can thus balance the benefits drawn from both domains of high-speed processing with reduced power overhead (Yu et al., 2011). Therefore, each *neuron* continues processing the input signal parallel to other *neuron*s in a *neuron*