The sustainability of ever more sophisticated artificial intelligence relies on the continual development of highly energy-efficient and compact computing hardware that mimics the biological neural networks. Recently, the neural firing properties have been widely explored in various spiking neuron devices, which could emerge as the fundamental building blocks of future neuromorphic/in-memory computing hardware. By leveraging the intrinsic device characteristics, the device-based spiking neuron has the potential advantage of a compact circuit area for implementing neural networks with high density and high parallelism. However, a comprehensive benchmark that considers not only the device but also the peripheral circuit necessary for realizing complete neural functions is still lacking. Herein, the recent progress of emerging spiking neuron devices and circuits is reviewed. By implementing peripheral analog circuits for supporting various spiking neuron devices in the in-memory computing architecture, the advantages and challenges in area and energy efficiency are discussed by benchmarking various technologies. A small or even no membrane capacitor, a self-reset property, and a high spiking frequency are highly desirable.

1. Introduction

Deep learning and artificial intelligence (AI) are now leading the way in improving both computing technology and human life. The rapid advance on the neural network topology and learning algorithms has been astonishing in recent years. However, the traditional computing hardware has not been able to keep pace with the high computational demands of AI processing, which has inspired the development of dedicated hardware such as the general-purpose graphic processing unit (GPU), which has inspired the development of dedicated hardware such as the general-purpose graphic processing unit (GPU) and the field-programmable gate array (FPGA). These devices have shown significant performance improvements compared to traditional central processing units (CPUs). However, they still face challenges in terms of energy efficiency, cost, and size.

To address these challenges, neuromorphic computing has emerged as a promising alternative. Neuromorphic computing leverages the principles of biological neural networks to build hardware that mimics both biological synapses and neurons. This approach can potentially offer both high energy efficiency and high performance.

Several functional blocks in the CMOS neuron circuits account for a large fraction of the total chip area because the integration and reset functions for generating spiking signals require a large number of transistors and large-area capacitors. To improve the area efficiency thus the density in the next-generation neuromorphic hardware, several spiking neuron devices have been proposed using varieties of technologies, such as the magnetoresistance memory (MRAM), in-membrane crossbar array and neuron circuit are used for multiplication and accumulation, respectively, which constitute the neural layer as inputs. Many complementary metal–oxide–semiconductor (CMOS)-based neuron circuits have been demonstrated previously. Integrating both memory-based synaptic arrays and CMOS neuron circuits enables a low-power, highly parallel analog neuromorphic system. However, the CMOS neuron circuits account for a large fraction of the total chip area because the integration and reset functions for generating spiking signals require a large number of transistors and large-area capacitors. To improve the area efficiency thus the density in the next-generation neuromorphic hardware, several spiking neuron devices have been proposed using varieties of technologies, such as the MRAM, ferroelectric field-effect transistor (FeFET), threshold-switching (TS) device, silicon-on-insulator MOSFET (SOI-MOSFET), and PCM. Several functional blocks in the CMOS neuron circuits could be replaced using a single device with equivalent functionality by leveraging intrinsic device characteristics. As a result, the area of the overall spiking neuron circuit could be effectively reduced. In addition to biologically inspired analog neuromorphic systems, a mixed analog–digital hardware implementation, the so-called in-memory computing (IMC), for accelerating AI based on the fundamental concepts of both neuromorphic computing (e.g., the synapse and neuron) and deep learning algorithms (e.g., matrix multiplication), has been gaining tremendous attention recently. In the IMC architecture, the memory crossbar array and neuron circuit are used for multiplication and accumulation, respectively, which constitute the
general multiply-accumulate (MAC) operations. The neuron function in the IMC architecture is typically implemented using analog-to-digital converters (ADCs), sense amplifiers (SAs), or integrate-and-fire (IF) neuron circuits with a digital counter. Several IMC-based AI chips have been demonstrated recently with excellent performance.\textsuperscript{[43–50]} More details regarding IMC will be introduced in Section 2.

In this article, we review the recent progress of various types of spiking neuron devices and circuits. Most studies in the literature are focused on the intrinsic device properties rather than complete neuron circuit implementation that is readily integrable with the synaptic array in the IMC architectures. Furthermore, the device operation, such as the types of input and output signals (voltage or current), and the required auxiliary circuit for supporting spike generation vary significantly among devices. Therefore, comparing performance across various spiking neuron circuits remains difficult. This article aims to provide a useful benchmark of device-based spiking neuron circuits and discuss the advantages and challenges in terms of area and energy efficiency compared with the conventional CMOS-based approach. The rest of this article is organized as follows: Section 2 introduces the analog IMC architecture with an emphasis on the peripheral circuit of the synaptic crossbar array. Section 3 provides an overview of prospective spiking neuron devices and their operating principles. Section 4 reports a generalized peripheral analog circuit (GPAC) for supporting various spiking neuron devices in the IMC architecture. The equivalent functionality of the GPAC is designed depending on the specific device characteristics for a fair benchmark. Finally, Section 5 discusses the advantages and challenges of different spiking neuron circuits.

2. Analog IMC Architecture

The circuit functional blocks of the analog IMC architecture for accelerating the fundamental operator of matrix–vector multiplication (MVM) in deep learning algorithms are shown in Figure 1.\textsuperscript{[43–47]} including word/bit/source line (WL/BL/SL) address decoders, voltage drivers, multiplexers (Mux), neuron circuits, and memory crossbar arrays. To perform MVM, the input vectors are converted into digital voltage signals and provided to the WL decoder to turn on different rows of WL voltage drivers. The selected WL voltage drivers then turn on the three-terminal memory cells, e.g., one transistor–one resistor memory cells, by applying an appropriate voltage bias. The voltage bias from the SL voltage driver and the neuron circuit is used for reading out the memory cell current, which is generated by the multiplication of the voltage drop (difference of the bias voltages between the SL driver and neuron circuit) on the selected memory cell and its conductance. The BL current ($I_{BL}$) accumulates all selected memory cell currents on the same BL and flows into the neuron circuits through the BL Mux and decoder. The neuron circuit not only provides a stable BL bias and reads the analog BL current, but also converts the output current signal to the required input voltage signal of the next layer. Generally speaking, a complete neuron circuit contains the functions of the analog bias circuit, ADC, and digital-to-analog converter.

The functional blocks inside the neuron circuit are further explained in Figure 2. First, the bias circuit provides a stable bias voltage on the BL for reading out the $I_{BL}$ that represents the MAC result. The scaling circuit is essential for reducing the current and thus power consumption in the neuron circuit because $I_{BL}$ in IMC is typically large. The $I/V$ conversion circuit is optional depending on the type of input signal required for the next data conversion stage. The data conversion stage, which is categorized into three classes, is mainly responsible for converting the MAC result into a suitable signal required for the next neural layer. A high-precision neuron circuit with a digital output can be realized using multibit ADCs, which are often used for designing deep neural network (DNN) hardware. However, the multibit ADC typically accounts for a large fraction of the overall circuit area and power. Alternatively, the SA is more area- and power-efficient but it is limited to one-bit precision. Thus, the tradeoff between circuit area/power and neuron precision is an important consideration in DNN hardware.\textsuperscript{[51,52]} The CMOS IF circuit could be utilized for converting the MAC results...
into a spike train in the time domain. The spike train could be fed into a digital counter to generate a digital output similar to a multibit ADC. The spike train could also be directly conveyed to the next neural layer in the spiking neural network (SNN) hardware through an appropriate output buffer. The output buffer is used to generate desired waveforms for the SNN input signal. The device-based spiking neuron circuit, the main focus of this article, is an alternative of the CMOS IF circuit to further improve the area efficiency. Both the digital counter and output buffer can also be combined with the device-based spiking neuron for the DNN and SNN hardware, respectively.

3. Spiking Neuron Devices: Operations, Physics, and Characteristics

Figure 3 shows an exemplary circuit implementation of the CMOS IF neuron and the device-based spiking neuron. For the CMOS IF neuron, the $I_{BL}$ is generated from the bias circuit and then used to charge the capacitor. When the voltage on the capacitor, also called the membrane potential, exceeds the reference voltage ($V_{REF}$) of the comparator, the output of the comparator switches from 0 V to $V_{DD}$, which simultaneously turns on the reset circuit and discharges the capacitor. When the voltage on the capacitor resets to a lower state less than $V_{REF}$, the output of the comparator switches back to 0 V to complete one spike generation at the output node. However, the capacitor, comparator, and reset circuit typically account for a large fraction of the circuit area in the overall neuron design. Therefore, a promising solution for reducing circuit area is to utilize intrinsic device characteristics for achieving equivalent circuit functionalities (as represented in the green region of Figure 3). For example, in a magnetic tunnel junction (MTJ), the core component of MRAM cells, the device resistance is stochastically varied depending on the magnitude of the $I_{BL}$ flowing into the MTJ. The voltage spikes are generated at the output without the assistance of an external circuit in an extremely compact configuration. We introduce various promising spiking neuron devices in more detail subsequently.

3.1. Phase-Change Neuron

The PCM has been widely adopted as the synaptic device for neuromorphic computing. The PCM has a metal-chalcogenide material–metal structure. The changeable resistance state of chalcogenides emerges by the phase transition between the resistive amorphous phase and the conductive crystalline phase in the programmable region near the bottom electrode through the local Joule heating effect. This transition process between the high-resistance state (HRS) and low-resistance state (LRS) could be used for voltage or current spike generation, as shown in Figure 4. By applying consecutive voltage pulses on the PCM device, the material phase in the programmable region gradually changes from the amorphous to crystalline phase. When the material phase transition reaches a threshold, the resistance state is switched to the LRS. Additional analog and digital circuits are required for detecting the LRS resistance by

![Figure 3. Comparison between CMOS IF neuron circuit and device-based spiking neuron circuit.](image1)

![Figure 4. Illustration of spike generation process in phase-change neuron devices.](image2)
setting an appropriate firing threshold. When the resistance is higher than the preset threshold, the PCM device is continuously programmed. By contrast, when the resistance is lower than the preset threshold, the reset circuit provides a reset pulse on the PCM device for switching it back to the HRS before another spike could be generated. Furthermore, the PCM spiking neuron exhibits stochasticity because of the intrinsic cycle-to-cycle variation of the resistive switching process. The potential drawback of this PCM neuron is that the additional reset circuit is inevitable because of the lack of the self-reset mechanism.

### 3.2. MTJ Neuron

Several studies have reported using the MRAM as a binary synapse for IMC. Realizing an all-spin IMC system by adopting the MRAM as both the synapse and the neuron is highly desirable from the viewpoint of system integration. Our previous report demonstrated the stochastic neuron firing behavior in an MTJ, as shown in Figure 5a. The LRS (parallel state) and HRS (antiparallel state) of an MTJ are defined by the different magnetization directions between the pinned layer and the free layer. When applying a high-enough voltage or current bias to an MTJ, the magnetization direction in the free layer is randomly switched to the direction parallel (LRS) or antiparallel (HRS) to the fixed layer due to the intrinsic back-hopping oscillation. The back-hopping oscillation originates from the strong Joule heating at the high-current regime that effectively reduces the energy barrier of magnetic anisotropy. The magnetization direction then randomly flips due to thermal fluctuations in two bistable states. Figure 5b shows that the spiking frequency and the duty cycle of the spike train increase with the applied voltage/current. Therefore, the MTJ could be connected to a digital counter to demonstrate an equivalent 4 bit ADC resolution for DNN applications. In contrast, generating spike trains with a bias-dependent duty cycle behaves similarly to a stochastic Poisson neuron for SNN applications.

This MTJ neuron requires neither the large membrane capacitor nor the reset circuit because of its inherent stochastic oscillation property. The potential drawback of this MTJ neuron is that a large input current is required for spike generation, which increases the power consumption of the neuron device.

### 3.3. FeFET Neuron Cells

The FeFET acts as a three-terminal synapse where the transistor channel conductance is modulated by the dipole switching of the ferroelectric gate insulator. Recent studies also showed that it could also act as a neuron cell for mapping an all-ferroelectric hardware neural network. Figure 6 shows the schematic and operation of two types of FeFET neuron cells. In Figure 6a, the FeFET neuron cell is comprised of one FeFET, one capacitor, and one n-channel field-effect transistor (nFET). Different from the typical bistable FeFET used for the memory and synapse, the FeFET neuron cell is designed to be only stable at LRS when a sufficient gate-source bias \( V_{GS} \) is applied. The capacitor is charged when the \( V_{GF} \) bias applied to the gate of the FeFET is high, and the FeFET is switched to the LRS. The FeFET is switched from the LRS to the HRS once the output voltage \( V_{out} \) increases, and the \( V_{GS} \) of the FeFET decreases below the coercive voltage for sustaining the LRS. Therefore, \( V_{out} \) gradually decreases via the discharging current path of the nFET, which is biased at \( V_{GM} \). When the \( V_{GS} \) of the FeFET increases above the coercive voltage, the FeFET switches back to LRS again. As a result, the output voltage spikes are periodically generated as
the FeFET switches between LRS and HRS. The magnitudes of $V_{GF}$ and $V_{GM}$ are used to control the speed of charging and discharging, thus modifying the spiking frequency. The potential drawback of this FeFET neuron cell is that it requires a large-area membrane capacitor (8 nF), possibly limited by the dynamic of FeFET response. Another implementation of the FeFET neuron cell without using a capacitor is shown in Figure 6b. Consecutive input voltage pulses applied to the gate of the FeFET are used for gradually adjusting the resistance of the FeFET, similar to the integration process. The resistance is abruptly switched to LRS after a period of time, and the spike train is generated at the output following the input pulses. By exploiting the intrinsic poor retention in a leaky FeFET, the FeFET gradually switches back to the HRS without intentional reset when the input voltage pulses are not applied. Furthermore, by applying an optional gate voltage bias to turn on the nFET, negative $V_{inh}$ is provided to the gate of the FeFET for inhibiting neural spiking in an inhibitory neuron. Compared with the previous FeFET neuron cell (Figure 6a), this neuron cell requires no capacitor. However, its slow reset process relying on the retention loss might limit the neuron response time.

### 3.4. TS Neuron

The TS devices literally perform abrupt resistive switching in a volatile fashion based on various mechanisms, including insulator-to-metal (IMT) transition,[33–39] ionic threshold transition (OTS),[33] and electrochemical metallization (ECM) that induces volatile formation of metal filaments.[33,34] Figure 7a shows the schematic current–voltage ($I$–$V$) characteristic of a TS device, where the device switches to LRS when applying a bias exceeding the threshold voltage ($V_{th}$) and returns to HRS when the applied bias is lower than the hold voltage ($V_{hold}$). The experimental data of different types of TS devices are shown in Figure 7b–d for a NbO$_x$-based IMT device, B–Te-based OTS device, and Ag/HfO$_2$-based ECM TS device, respectively.[33] To exploit the voltage-dependent volatile switching as a TS neuron, an additional capacitor and a resistor are typically required as a leaky integrate-and-fire (LIF) neuron, as shown in Figure 7e.[33–35,38,39] The current spike train generated from the synaptic memory array charges the capacitor and increases $V_{TS}$ gradually (region 1). The capacitor is slightly discharged via the resistive circuit path where the TS device is at HRS during the off period of the current pulse (region 2). Once sufficient input currents...
are integrated and $V_{TS}$ reaches the $V_{th}$ of the TS device (region 3), the TS device switches to LRS and $V_{out}$ is high in a firing process. This also triggers the rapid discharge and reset process of $V_{TS}$ (region 4). Once $V_{TS}$ is lower than $V_{hold}$, the TS device switches back to HRS and $V_{out}$ is low again. This iterative process produces voltage spikes at the output node. Another approach shows that combining the characteristic of the nFET load line with the cycle-to-cycle variation of the TS device can perform stochastic neuron oscillation, as shown in Figure 7f.\[36,37\] The nFET is a voltage-controlled resistor adjusted by $V_{GS}$. As $V_{GS}$ increases, the voltage drop on the TS device gradually increases as the resistance of the nFET becomes smaller. When the voltage drop on the TS device reaches its $V_{th}$, the TS device is switched from HRS to LRS, and $V_{out}$ is switched to high. At this time, the voltage on the LRS TS device is reduced below $V_{hold}$. The TS device is switched back to HRS and $V_{out}$ is switched back to low again in a self-reset process. The choice of $V_{GS}$ (i.e., resistor loading) is critical for spike generation. Furthermore, the spike generation is stochastic in nature due to the intrinsic cycle-to-cycle variation during TS. Compared with the TS neuron shown in Figure 7e, the advantage is that it does not require a capacitor for firing, further reducing the neuron circuit area. However, the range of the voltage-controlled resistor capable of spike generation is limited and requires precise control and tuning. An optional integrator circuit including a capacitor and resistor could be connected to the gate of an nFET for mimicking neuron integration and the leaky behavior in an LIF neuron instead of a stochastic neuron.\[36,37\]

3.5. SOI-MOSFET Neuron

A three-terminal SOI-MOSFET is typically used for suppressing the source-drain leakage current in the scaled CMOS technology. It was also explored for one-transistor DRAM (1T-DRAM) applications by utilizing the intrinsic floating body effect.\[53\] Recently, an SOI-MOSFET was reported for mimicking the neuron functionality.\[40] The device structure and operating principles are shown in Figure 8. By applying positive constant voltages to the drain and gate while grounding the source, holes are generated due to the impact ionization induced by the strong lateral electric field and accumulate in the body. As a result, the body energy potential and thus the energy barrier between the body and source are further reduced. The rise of the drain current ($I_{DS}$) further increases the hole generation. This positive feedback process produces an abrupt $I_{DS}$ increase. The excessive amount of holes accumulated in the body eventually leaks away through the body–source barrier. To recover $I_{DS}$ back to its original state for the next current spike generation, a reset circuit is required for interrupting the drain bias after $I_{DS}$ exceeds a current threshold. The advantage of this SOI-MOSFET neuron is that no large-area capacitor is required for integration. However, the additional reset circuit is essential. Han et al. reported a single SOI-MOSFET neuron cell without an additional reset circuit and capacitor.\[41\] A constant input current flows into the drain terminal of an off-state SOI-MOSFET (i.e., gate bias is negative) to charge the gate–drain overlapped capacitor ($C_{GD}$). The gradually increased drain voltage would instantly switch the device to its on-state after reaching the intrinsic device latch voltage. After discharging the $C_{GD}$, the transistor then switches back to the original off-state and is ready for the next spike generation. However, due to the small parasitic $C_{GD}$, the input current that this device can handle is potentially too low for practical use. An additional external capacitor for integration might be unavoidable.

3.6. Comparison of Spiking Neuron Devices

After an overview of various spiking neuron devices, we summarize their unique features in Table 1. The MTJ neuron demonstrates most of the desired features for a compact neuron cell except for its relatively high input current level,\[26\] which potentially increases the power consumption of the neuron device. The device is triggered by the current input and generates voltage output. The current input and voltage output are directly compatible with the analog IMC architecture, as shown in Figure 1. The PCM neuron\[42\] requires both the reset circuit and an extremely large reset voltage (6 V) for spike generation. The SOI-MOSFET neuron\[40\] shows a high firing rate of 20 MHz but requires a relatively high $V_{DD}$ (2.8 V) and an additional reset circuit. The other SOI-MOSFET neuron that directly integrates the drain current requires no external capacitor and reset circuit,\[41\] but it requires negative gate bias and the allowed input current level is low. For the FeFET neuron with an integrated capacitor,\[30,31\] the additional capacitor and external gate voltage for controlling the reset
Table 1. Performance comparison between various spiking neuron devices.

| Device type | MTJ | PCM | SOI-MOSFET | FeFET | IMT (NbO₂/VO₂) | OTS (ionic TS) | ECM (Ag-based) |
|-------------|-----|-----|------------|-------|----------------|---------------|---------------|
| Neuron type | IF\([26,27]\) | IF\([26]\) | IF\([26]\) | IF\([28]\) | IF\([29]\) | IF\([30]\) | IF\([31]\) |
| Cell input/output | I/V | V/V | I/I | I/V | I/V | I/V | I/V |
| Max. input I/V level | 0.9 mA | 6 V | 2.8 V | 10 nA | 0.3 V | 1.8 V | 40 µA |
| External voltage supply | No | No | No | Vₓ C | Vₓ C | Vᵢ₉ (optional) | No |
| Max. firing rate | >100 kHz | 50 kHz | 20 MHz | 1 kHz | 14 kHz | <50 Hz | 200 kHz |
| Membrane capacitor | No | No | No | No | Yes | (8 nF) | No |
| Reset circuit | Yes | Yes | Yes | No | No | No | No |
| Stochasticity | Yes | Yes | Yes | No | No | No | No |

4. Performance Benchmark of Spiking Neurons in IMC Architecture

In this section, various neuron cells are benchmarked using the IMC architecture for evaluating the overall performance of the complete neuron circuit. In this work focusing on the spiking neurons, we assume the input signal is encoded in the time-domain spike train by using either the spike rate\([26]\) or spike duty cycle\([27]\). The spike train of a voltage pulse is then applied to the WLs of memory crossbar arrays. The neuron circuit reads the Iₓₛₗₚ and generates the output signal (see Figure 1 and 2). In the following discussion, we will first design a GPAC that is compatible with different spiking neuron cells. Although the circuit diagram of the GPAC is similar, the circuit specifications are adjusted to satisfy the different input requirements among various devices. Furthermore, we also designed a conventional CMOS IF circuit for comparison.

4.1. Design of GPAC

A GPAC includes a bias circuit, scaling circuit, and offset circuit, as shown in Figure 9. First, the Iₓₛₗₚ is read out through the bias circuit by applying an appropriate read voltage to memory cells on the same BL. The output current of the bias circuit is scaled and shifted to the detection range of the specific spiking neuron cell. Both the scaling and offset circuits are the nFET current mirrors. \(I_{SNM} (I_{scaling} + I_{offset})\) is then fed into the spiking neuron module (SNM) that includes a spiking neuron cell and auxiliary components. The choice of these three SNM designs, also shown in Figure 9, depends on the spiking neuron cell. SNM-I is suitable for two-terminal neuron cells that are directly driven by current inputs, such as an MTJ neuron\([26,27]\) and PCM neuron\([42]\). In the MTJ neuron, spikes are generated at \(V_{out}\). When a suitable current flows through the device and induces back-hopping oscillation,\([26,27]\) Liang et al. have reported a complete MTJ spiking neuron circuit design previously\([28]\). SNM-II is suitable for spiking neuron cells triggered by voltage input, such as the SOI-MOSFET neuron\([40]\). FeFET neuron cells\([36,37]\) and capacitor-less TS neuron\([36,37]\). A resistor is used for converting \(I_{SNM}\) to \(V_{SNM}\) (I/V conversion in Figure 2) as the input of the SNM-II. \(V_{SNM}\) could be adjusted to an appropriate bias point for spike generation using either the scaling circuit and offset circuit or the resistor. However, a reset circuit is still required for the SOI-MOSFET neuron\([40]\). An exemplary reset circuit realized using the fully analog approach is shown in Figure 9b. First, the current \(I_{SOI}\) from the SOI-MOSFET neuron flows into the transimpedance amplifier, and the converted voltage \(V_{DP}\) is compared with the firing threshold voltage \(V_{th}\). As \(I_{SOI}\) gradually increases over time due to hole accumulation and \(V_{DP}\) reaches \(V_{th}\), the comparator sends a trigger voltage to tune off the top pFET for resetting the energy band state of the SOI-MOSFET by cutting off the pFET. As a result, a spike generates at \(V_{out}\). SNM-III is suitable for voltage-driven TS neuron cells.\([33–35,38,39]\) \(I_{SNM}\) is mirrored and charges the membrane capacitor. When the \(V_{out}\) reaches the \(V_{th}\) of the TS device (switched from HRS to LRS), the capacitor is discharged through the TS device for generating voltage spikes. For all the aforementioned approaches, an optional voltage buffer could be connected to the neuron circuit at \(V_{out}\). For shifting the voltage level to an adequate voltage level for driving the WL transistors of the next layer.

The complete GPAC design is shown in Figure 10a. The bias circuit is composed of transistors M1 to M7. M1 is used for providing the BL readout voltage \(V_{BL}\), which is mirrored at the source of M3. M7 is used for balancing the current flowing through M3 and thus the current flowing through M5 equals Iₓₛₗₚ. Iₓₛₗₚ is then copied to M6 as Iₛₗₚ. Iₛₗₚ then flows into...
the scaling circuit (current mirror) for scaling down the current to $I_{\text{scaling}}$. An offset current circuit provides a constant current $I_{\text{offset}}$ through a cascode current mirror. Eventually, $I_{\text{SNM}}$ ($I_{\text{scaling}} + I_{\text{offset}}$) is fed into the three different modules as mentioned in Figure 9a. We also show a CMOS IF neuron circuit based on an identical bias circuit for comparison, as shown in Figure 10b. Figure 10c shows the simulation results of the CMOS IF neuron circuit with different firing rates. A high firing rate of 2 GHz is produced by applying $I_{BL} = 3.5$ mA without additional scaling and offset circuits.

4.2. Performance Benchmarking for Various Neuron Circuits

We selected several promising neuron devices and integrated them with the proposed neuron circuit for performance evaluations, as shown in Table 2. The neuron devices we evaluated in this study include the current-input MTJ with stochastic oscillation, the voltage-input SOI-MOSFET with an additional reset circuit, and the voltage-input FeFET neuron cell with a membrane capacitor. Among the various TS neurons in Table 1, we evaluated two promising TS neurons: the IMT-based NbO$_2$ device with a small membrane capacitor and extremely high firing rate$^{[19]}$ and another IMT-based capacitor-less VO$_2$ device with a compact size. The MTJ, SOI-MOSFET, FeFET, NbO$_2$ TS, and VO$_2$ TS neurons are integrated with SNM-I, SNM-II, SNM-2, SNM-3, and SNM-2, respectively. We did not include the PCM-based neuron cell, current-driven SOI-MOSFET neuron cell, and capacitor-less FeFET neuron due to their respective aforementioned limitations. The 65 nm CMOS using the predictive technology model was applied for circuit simulation. $I_{BL}$ is assumed to range from 0.1 to 3.5 mA, which is based on our previous report of the neural
network training results on CIFAR-10\textsuperscript{[16,26]} in a binary MRAM array. In our simulation, we adopted the equivalent BL resistor (28.57–1 kΩ) with a BL readout voltage of 0.1 V from the bias circuit to emulate the $I_{\text{BL}}$ generated from the synaptic array. The transistor size of the GPAC was tuned to match the input current/voltage requirement of specific neuron devices.

Regarding the supply voltage, the $V_{\text{DD}}$ of all GPACs was fixed at 1.2 V except for the capacitor-less IMT neuron circuit, where 2 V was used because the minimum $V_{\text{SNM}}$ for oscillation is 1.8 V\textsuperscript{[36]} The SOI-MOSFET neuron, FeFET neuron, and capacitor-less neuron require an additional power supply of 2.8, 0.4, and 0.7 V, respectively, for supporting the operation of different SNMs. Additional supply voltages are known to increase the complexity and area of circuit design. Among the SNMs, the MTJ neuron (one MTJ) and capacitor-less TS neuron (one TS + one transistor) show the simplest structure. The SOI-MOSFET neuron requires an additional reset circuit that occupies a large circuit area. The FeFET neuron includes a large capacitor (0.8 nF)\textsuperscript{[30,31]} and the capacitor-based TS neuron requires a more competitive capacitor size (0.1 pF)\textsuperscript{[39]} The circuit areas of the GPAC for different neurons are comparable except for the capacitor-less TS neuron. This is because the use of higher $V_{\text{DD}}$ (2 V) helps to reduce the size of transistors for handling high $I_{\text{BL}}$ (M5 and M6 in Figure 10a). The capacitor is assumed to occupy 650 μm$^2$ pF$^{-1}$ in the 65 nm CMOS\textsuperscript{[56]} Thus, the capacitor (8 nF) of a FeFET neuron dominates the overall neuron circuit area. The total neuron circuit area is estimated using only the GPAC and capacitor areas in this study for simplicity. The significant area occupied by the reset circuit of the SOI-MOSFET neuron is not included here. The capacitor-less TS neuron and the MTJ neuron achieve 42 and 18 times area saving compared with the CMOS IF neuron, suggesting the outstanding advantage of a device-based spiking neuron in area efficiency. The energy consumption per spike generation in a single neuron device was evaluated using the reported firing rate and the input voltage/current levels. Even though the average power of the FeFET neuron cell is low, its energy per spike is relatively large because of the low firing rate. In our design, the GPAC, especially the bias circuit part along the path of high $I_{\text{BL}}$ (0.1–3.5 mA), dominates the overall power consumption of

Figure 10. a) Complete circuit design of GPAC. b) Circuit design of traditional CMOS IF neuron connected to the bias circuit in (a). c) Simulation results of CMOS IF neuron circuit showing a high firing rate of 2 GHz with $I_{\text{BL}} = 3.5$ mA.
the entire neuron circuit. The use of high $I_{RI}$ is rational because most energy is consumed by a large number of computing operations within the synaptic array rather than the data conversion in the neuron circuit. The overhead of data conversion could then be shared to boost energy efficiency per computing operation. Therefore, the energy per spike for the entire neuron circuit is dominated by the GPAC power consumption and the intrinsic firing rate of the neuron device. Generally speaking, the CMOS IF neuron consumes less energy due to its higher firing rate compared with the device-based spiking neurons. Therefore, the high firing rate is a critical requirement for future energy-efficient spiking neurons.

5. Future Perspectives

The device-based spiking neuron circuits discussed herein have shown a clear advantage over the CMOS neuron circuits in area efficiency, although the energy efficiency could be further improved by increasing the firing rate up to gigahertz, which is comparable to that in the CMOS counterpart. The other advantage of the higher firing rate is to accelerate the operating speed in hardware, one of the main motivations for IMC architectures and artificial neural networks. However, the gigahertz target could be challenging because the firing rate depends strongly on the intrinsic device properties, and most of the devices discussed in this study are based on conventional memory technologies that typically show a switching speed of tens of nanoseconds. Unlike the memory technology where the switching speed is compromised due to the requirement of a large energy barrier for long retention, the neuron device could be optimized solely for the high firing rate with a sufficiently low energy barrier. Trading retention with speed was successfully demonstrated in MRAM for last-level cache applications at a sub-10 ns switching speed.[57] Note that the requirement on the high firing rate is to minimize the DC energy consumption in the GPAC with high $I_{RI}$. By contrast, the firing rate in biological neurons is less than 100 Hz, indicating that high energy efficiency is also possible with an extremely low firing rate. Innovations are required to minimize DC power consumption in neuron circuits and improve energy efficiency in low-power hardware operating at a much lower frequency when speed is not the primary concern. The much lower operating voltage (<100 mV) for biological neurons is also beneficial for high energy efficiency and should be explored in future neuron circuit designs.

Because the development of the device-based spiking neuron circuits is at its early stage, scalability, variability, and reliability of these technologies are seldom discussed but are of great importance for practical use in the future. The area of the complete neuron circuit is dominated by the GPAC rather than the neuron device itself. Therefore, the scaling of CMOS technology is highly beneficial. A projected 100 times shrinkage in the area from the 65 to 5 nm node would result in a neuron circuit area of less than micrometer squared based on the most compact technology such as MTJ and TS devices. Such compact neuron circuits would allow a pitch-matched design with the synaptic array for achieving maximum parallelism. Scaling the neuron device itself could affect variability and reliability. For example, the relation of scalability and reliability of PCM has been extensively studied.[58-60] Many material properties of PCM, including the crystallization/melting temperature, thermal conductivity, and threshold voltage, were found to be dependent on device scaling. An ultimately scaled PCM cell (∼1.2 nm) using carbon nanotubes as the memory electrode maintained appropriate memory performance.[61] The bulk switching of the phase-change material allows tight distributions of memory states.[62] Scaled MTJ devices show degraded performance and variability due to the edge damage during MTJ etching when the MTJ size scales below 30 nm.[63,64] Therefore, process optimization is required to repair the damage.[65] Although the MTJ is expected to have unlimited endurance,[66] time-dependent dielectric breakdown in the MgO barrier layer limits the endurance of the MTJ in practice.[67] As for the FeFET, the limited number of ferroelectric domains due to device area scaling aggravates device-to-device variation.[68]
Endurance properties in HfO$_2$-based ferroelectric materials have been discussed. Interface engineering by plasma treatment and high-k interfacial layer insertion were proposed to improve device reliability. Finally, relatively fewer studies have focused on the scalability and reliability of TS devices. However, thermal instability and $V_{th}$ drift are concerns for real applications. In general, because the scalability of the neuron device itself is less critical for the area of neuron circuits, a more relaxed cell size could be used for achieving better variability and reliability in neuron devices.

6. Conclusion

The emerging spiking neuron devices and circuits are reviewed and benchmarked in this work. The area and energy efficiency are evaluated based on the proposed GPAC and IMC architecture. A small or even no membrane capacitor, a self-reset property, and a high spiking frequency are highly desirable. The MTJ and TS neurons appear to be two promising candidates. Their circuit area variation, and temporal variation/degradation should be critically evaluated in the future for a viable device-based neuron technology to compete with the more mature CMOS neuron circuit.

Acknowledgements

This work was supported by the Ministry of Science and Technology of Taiwan under grant: MOST 109-11027-E-009-004-MY3, MOST 109-2634-F-009-022, and MOST 109-2639-E-009-001. It was also in part supported by the Industrial Technology Research Institute. The authors are grateful to Ms. Wei-Ching Kuo for editing assistance in preparing the manuscript.

Conflict of Interest

The authors declare no conflict of interest.

Keywords

in-memory computing, neuromorphic computing, spiking neurons

Received: January 12, 2021
Revised: March 13, 2021
Published online:

[1] Y. LeCun, Y. Bengio, G. Hinton, Nature 2015, 521, 7553.
[2] Y. Lecun, L. Bottou, Y. Bengio, P. Haffner, Proc. IEEE 1998, 86, 2278.
[3] A. Krizhevsky, I. Sutskever, G. Hinton, Commun. ACM 2017, 60, 84.
[4] D. Rumelhart, G. Hinton, R. Williams, Nature 1986, 323, 533.
[5] J. Nickolls, W. Dally, IEEE Micro. 2010, 30, 56.
[6] N. P. Jouppi, C. Young, N. Patil, D. Patterson, G. Agrawal, R. Bajwa, S. Bates, S. Bhatia, N. Boden, A. Borchers, R. Boyle, P. Cantin, C. Chao, C. Clark, J. Coriell, M. Daley, M. Dau, J. Dean, B. Gelb, T. V. Ghaemmaghami, R. Gottipati, W. Guiland, R. Hagmann, C. R. Ho, D. Hogberg, J. Hu, R. Hundt, D. Hurt, J. Ibarz, A. Jaffey, et al., in ACM/IEEE 44th Annual Int. Symp. Comput. Archit. (ISCA), ACM/IEEE, Tokyo, Canada 2017, p. 7.
[7] D. Silver, A. Huang, C. Maddison, A. Guez, L. Sifre, G. van den Driessche, J. Schrittwieser, I. Antonoglou, V. Panneershelvam, M. Lanctot, S. Dieleman, D. Grewe, J. Nham, N. Kalchbrenner, I. Sutskever, T. Lillicrap, M. Leach, K. Kavukcuoglu, T. Graepel, D. Hassabis, Nature 2016, 529, 484.
[8] M. Zidan, J. Strachan, W. Lu, Nat. Electron. 2018, 1, 22.
[9] P. Merolla, J. Arthur, R. Alvarez-Icaza, A. Cassidy, J. Sawada, F. Akopyan, B. Jackson, N. Imam, C. Guo, Y. Nakamura, B. Brezzo, I. Vo, S. Esser, R. Appuswamy, B. Tabo, A. Amir, M. Flickner, W. Risk, R. Manohar, D. Modha, Science 2014, 345, 668.
[10] S. H. Jo, T. Chang, I. Ebong, B. B. Bhadviya, P. Mazumder, W. Lu, Nano Lett. 2010, 10, 1297.
[11] S. Yu, Y. Wu, R. Jayasimh, D. Kuzum, H.-S. P. Wong, IEEE Trans. Electron Devices 2011, 58, 2729.
[12] C.-C. Chang, P.-C. Chen, B. Hudec, P.-T. Liu, T.-H. Hou, in 2018 IEEE Int. Electron Devices Meet. (IEDM), IEEE, Piscataway, NJ, 2018, p. 356.
[13] G. Burr, R. Shelby, S. Sidler, C. di Nolfo, J. Jang, I. Boybat, R. Shenoy, P. Narayanan, K. Virwani, E. Giacometti, B. Kurdi, H. Hwang, IEEE Trans. Electron Devices 2015, 62, 3498.
[14] S. Ambrogio, P. Narayanan, H. Tasi, R. Shelby, I. Boybat, C. di Nolfo, S. Sidler, M. Giordano, M. Bodini, N. Farinha, B. Killeen, C. Cheng, J. Yiaoud, G. Bur, Nature 2018, 558, 60.
[15] A. Vincent, J. Larroque, N. Locatelli, B. Romdhane, O. Bichler, C. Gamrat, W. Zhao, J. Klein, S. Galdin-Retailleau, D. Querlioz, IEEE Trans. Biomed. Circuits Syst. 2015, 9, 166.
[16] C.-C. Chang, M.-H. Wu, J.-W. Lin, C.-H. Li, V. Parmar, H.-Y. Lee, J.-H. Wei, S.-S. Sheu, M. Suri, T.-S. Chang, T.-H. Hou, in 2019 ACM/IEEE 56th Design Automation Conf. (DAC), ACM/IEEE, Las Vegas, NV 2019, p. 1.
[17] X. Sun, P. Wang, K. Ni, S. Datta, S. Yu, in 2018 IEEE Int. Electron Devices Meet. (IEDM), IEEE, Piscataway, NJ 2018, p. 55.
[18] M. Jerry, P. Y. Chen, J. Zhang, P. Sharma, K. Ni, S. Yu, S. Datta, in 2017 IEEE Int. Electron Devices Meet. (IEDM), IEEE, Piscataway, NJ 2017, p. 139.
[19] C. Zamarreño-Ramos, L. Camuñas-Mesa, J. Pérez-Carrasco, T. Masquelin, T. Serrano-Gotarredona, B. Linares-Barranco, Front. Neurosci. 2011, 5, 25.
[20] G. Indiveri, B. Linares-Barranco, T. Hamilton, A. Schaik, R. Etienne-Cummings, T. Delbruck, S. Liu, P. Dudek, P. Häfliger, S. Renaud, J. Schemmel, G. Cauwenberghs, J. Arthur, K. Hynna, F. Folowosele, S. Saighi, T. Serrano-Gotarredona, J. Wijekoon, Y. Wang, K. Boahen, Front. Neurosci. 2011, 5, 73.
[21] S. A. Aamir, P. Müller, A. Hartel, J. Schemmel, K. Meier, in 2016 42nd European Solid-State Circuits Conf., IEEE, Lausanne, Switzerland 2016, p. 71.
[22] I. Sourikopoulos, S. Hedayat, C. Loyez, F. Danneville, V. Hoel, E. Mercier, A. Cappy, Front. Neurosci. 2017, 11, 23.
[23] B. Yan, Q. Yang, W. Chen, K. Chang, J. Su, C. Hsu, S. Li, H. Lee, S. Sheu, M. Ho, Q. Wu, M. Chang, Y. Chen, H. Li, in 2019 IEEE Symp. on VLSI Technology (VLSI-T), IEEE, Piscataway, NJ 2019, p. T86.
[24] S. Park, H. Kim, M. Choo, J. Noh, A. Sheri, S. Jung, K. Seo, J. Park, S. Kim, W. Lee, J. Shin, D. Lee, G. Choi, J. Woo, E. Cha, J. Jang, C. Park, M. Jeon, B. Lee, B. H. Lee, H. Hwang, in 2012 IEEE Int. Electron Devices Meet. (IEDM), IEEE, Piscataway, NJ 2012, p. 231.
[25] M. Prezioso, M. R. Mahmoodi, F. Merrikh Bayat, H. Nili, H. Kim, A. Vincent, D. B. Strukov, Nat. Commun. 2018, 9, 5311.
R. Boschke, R. van Bentum, U. Schröder, T. Mikolajick, in 2013 IEEE Int. Electron Devices Meet. (IEDM), IEEE, Piscataway, NJ 2013, p. 280.

[70] A. J. Tan, M. Pešić, L. Larcher, Y.-H. Liao, L.-C. Wang, J.-H. Bae, C. Hu, S. Salahuddin, in 2020 IEEE Symp. on VLSI Technology (VLSI-T), IEEE, Piscataway, NJ 2020, p. 1.

[71] K.-Y. Chen, P.-H. Chen, Y.-H. Wu, in 2017 IEEE Symp. on VLSI Technology (VLSI-T), IEEE, Piscataway, NJ 2017, p. T84.

[72] T. Ali, P. Polakowski, S. Riedel, T. Büttner, T. Kämpfe, M. Rudolph, B. Pätzold, K. Seidel, D. Lühr, R. Hoffmann, M. Czernohorsky, K. Kühnel, P. Steinke, J. Calvo, K. Zimmermann, J. Müller, IEEE Trans. Electron Devices 2018, 65, 3769.

[73] C.-Y. Chan, K.-Y. Chen, H.-K. Peng, Y.-H. Wu, in 2020 IEEE Symp. on VLSI Technology (VLSI-T), IEEE, Piscataway, NJ 2020, p. 1.

[74] A. Velea, K. Opsomer, W. Devulder, J. Dumortier, J. Fan, C. Detavernier, M. Jurczak, B. Coboreanu, Sci. Rep. 2017, 7, 8103.

[75] J. Yoo, I. Karpov, S. Lee, J. Jung, H. S. Kim, H. Hwang, IEEE Electron Device Lett. 2020, 41, 191.

Fu-Xiang Liang is currently pursuing his Ph.D. degree in the Department of Electrical Engineering, University of Notre Dame, Notre Dame, IN, USA. He received his M.S. degree in electrical engineering from National Cheng-Kung University, Tainan, Taiwan, in 2018. From 2017 to 2018, he joined IMEC, Leuven, Belgium, as an internship student. From 2019 to 2020, he joined the Department of Electronics Engineering, National Chiao Tung University, Hsinchu, Taiwan, as a research assistant. His research interest focuses on nonconventional device and circuit designs for energy-efficient computing applications.

I-Ting Wang received her Ph.D. degree in electronics engineering from National Chiao Tung University (NCTU), Hsinchu, Taiwan, in 2018. She joined the Institute of Microelectronics (IME), A*STAR, Singapore, as a research scientist from 2018 to 2020. She is currently a postdoctoral researcher with the Institute of Electronics, National Yang Ming Chiao Tung University (NYCU), Hsinchu, Taiwan. Her research interests include memristors for neuromorphic computing applications and the 3D high-density emerging memory technology.

Tuo-Hung Hou received his Ph.D. degree in electrical and computer engineering from Cornell University in 2008. In 2000, he joined Taiwan Semiconductor Manufacturing Company (TSMC). From 2001 to 2003, he was also a TSMC assignee at International SEMATECH. Since 2008, he joined the Department of Electronics Engineering, National Chiao Tung University (NCTU) (as National Yang Ming Chiao Tung University (NYCU) in 2021), where he is currently a Distinguished Professor and Associate Vice President for Research & Development. His research interests include the emerging nonvolatile memory, electronic synaptic device and neuromorphic computing systems, and heterogeneous integration of low-dimensional nanomaterials.