Received 14 June 2022; revised 3 September 2022; accepted 3 September 2022. Date of publication 13 September 2022; date of current version 14 October 2022.
The review of this article was arranged by Editor M. Saitoh.

*Digital Object Identifier 10.1109/JEDS.2022.3206317*

# A Stochastic Leaky-Integrate-and-Fire Neuron Model With Floating Gate-Based Technology for Fast and Accurate Population Coding

**AKIRA GODA, CHIHIRO MATSUI, AND KEN TAKEUCHI (Member, IEEE)**

Department of Electrical Engineering and Information Systems, The University of Tokyo, Tokyo 1138656, Japan

CORRESPONDING AUTHOR: A. GODA (e-mail: goda@co-design.t.u-tokyo.ac.jp)

**ABSTRACT** An analytical model has been developed for stochastic leaky-integrate-and-fire (LIF) neurons with floating gate (FG) technology. The stochastic behaviors have been modeled extensively for both individual neurons and populations of neurons. In the FG LIF neurons, the electron injection is governed by the tunneling process through the gate oxide, leading to the exponential distributions of the injection time and inter spike interval (ISI) stochasticity. The concept of the population coding is demonstrated by simulating the stochastic behaviors of the populations of the FG LIF neurons. The ISI stochasticity enables encoding of the input signals to the population outputs. Spike-to-spike stochasticity improves the signal-to-noise ratio of the population outputs. Moreover, the shape of the ISI distribution can be controlled by adjusting the number of electrons to spike (NES). Exponential-like ISI distributions are realized by reducing the NES. With the exponential-like ISI distributions, the population of fast spiking neurons increases significantly (more than 10% of neurons spiking twice faster than the mean ISI), potentially contributing to the fast computation. Finally, step-by-step procedures have been proposed to design the FG LIF neurons exhibiting the desired neuron characteristics including operation voltage (0.5 V to 3 V), leaky time constant ($<1$ $\mu$s to $>10$ ms), ISI mean (in the range of 6 orders of magnitude) and stochasticity ($\sim$0 % to $\sim$60 %) as well as the type of the distribution (exponential-like to Gaussian-like).

**INDEX TERMS** Stochastic neuron, floating gate, population code, inter spike interval, neuron time constant, leaky integrate and fire, spiking neural networks.

## I. INTRODUCTION

SPIKING neural networks (SNNs) have been gaining attention for the brain-inspired, energy efficient and error-tolerant computing [1], [2], [3]. Fig. 1(a) shows the concept of the SNN. Information is encoded in the form of spikes. The event-based nature of the SNN realizes the energy efficient computing as there is no energy consumption between the spikes. In the hardware implementation of SNN chips, synaptic devices and neuron devices are the key elements to realize the high density and low power computation.

Nonvolatile memory (NVM) technologies have been actively investigated for both synaptic devices and neuron devices [4], [5], [6], [7]. For the synaptic devices storing the weights, the high density integration is the key to processing the large amount of the data. NVMs are suitable as the synaptic devices owing to their nonvolatility and excellent device scalability [8], [9].

For the neuron devices, leaky-integrate-and-fire (LIF) is one of the widely accepted models in SNNs. In the SNN hardware, the LIF functions could be realized by CMOS circuits [10], [11]. As an alternative hardware solution for the neuron circuits, NVM applications to the neuron devices have been widely investigated to reduce the circuit area and energy consumption [12], [13], [14], [15]. Fig. 1(b) shows the simplified concept of the LIF neuron circuit with an NVM cell. In this circuit, leaky and integrate functions are built into the single NVM cell as opposed to a large CMOS circuit including resistors and capacitors. There are research and development of volatile memory-based

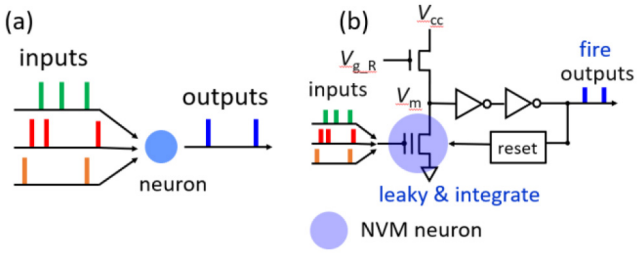



**FIGURE 1.** **(a) Spiking neural network (SNN). (b) Leaky-integrate-and-fire (LIF) neuron circuit with NVM neuron device.**

neurons [16], [17]. With the volatile-based neurons, the nonlinear activation function (such as sigmoid function) can be replicated. With the NVM-based neurons, the neuron can remember the history of the inputs, realizing the LIF function where the information of the previous inputs is stored.

In addition to the area savings and energy reduction, another key interest for NVM-based neurons is the inherent stochasticity originating from NVM device physics [18], [19]. The spiking stochasticity of the neurons could improve computation efficiency and robustness [20], [21], [22], [23]. It was reported that the population of neurons can respond faster than the individual neurons when the spiking stochasticity exists [18], [24]. Also, it was pointed out that the spiking stochasticity contributes to the SNN computation robustness of asynchronous operations [25].

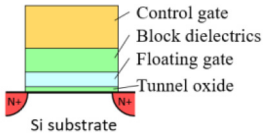
ISI distributions of NVM neurons have been studied from the device physics perspective and the population-based computing perspective. Kornijcuk *et al.* [13] modeled the ISI distribution of FG neurons. In that work, LIF and ISI properties were evaluated by LTspice circuit simulator. The leaky function was achieved by using the thin tunnel oxide. The ISI stochasticity was simulated by injecting the thermal noise and random telegraph noise into the CMOS neuron circuit and devices. The resultant ISI distributions were shown to follow the gamma distributions. Tuma *et al.* [18] experimentally measured ISI distributions of phase-change neurons. The ISI distributions were fitted to Gaussian distributions and were attributed to the stochasticity of the crystal growth in a single neuron as well as inter-neuron variations due to variations in device fabrication. In addition, the benefit of the ISI stochasticity on the population code was demonstrated by simulation.

In our previous work [26], FG LIF neuron characteristics were modeled with the focus on the electron injection physics and statistics. An analytical model was developed for leaky and integrate operations. ISI stochasticity was shown to universally follow the number of electrons to spike.

In this work, the FG LIF neuron model is further extended to develop the full perspective of the individual neuron behaviors. In addition, the stochastic behaviors of the populations of the neurons are simulated. The key advancements in this work include (1) a complete picture of the ISI targeting for FG LIF neurons by engineering device and operation parameters, (2) modeling of the time-variable neuron time constant in FG LIF neurons, (3) demonstration of the population coding function enabled by the ISI stochasticity, (4) demonstration of noise reduction effects in the population coding realized by the spike stochasticity, (5) statistical analysis for fast responding neurons enabled by the exponential ISI distribution and (6) a proposal of step-by-step procedures to design FG LIF neurons to comprehensively target al. neuron characteristics for both mean and stochastic behaviors. By using the relation between the LIF neuron characteristics and device parameters modeled in this work, it becomes possible to precisely optimize the device and operation parameters of FG LIF neurons to realize the desired neuron characteristics including the stochasticity.

**TABLE 1.** **Device parameters for FG LIF neuron.**

| Device parameter | Value |
|---|---|
| Block dielectrics thickness | 8nm (EOT) |
| Floating gate thickness | 5nm |
| Tunnel oxide thickness | 1.5-7.0nm |
| Neutral threshold voltage | 0V |




## II. FLOATING GATE-BASED LIF NEURON MODEL

LIF functions include 'leaky', 'integrate', 'fire' and 'reset' operations. In the 'integrate' phase, the neuron receives the input signals from the previous layer. The membrane potential of the neuron gradually develops. When the input signals are absent, the membrane potential decays which is called 'leaky' phase. When the membrane potential reaches the threshold potential, the neuron fires a spike. After the 'fire', the membrane potential is 'reset' to the initial value. The operations of FG LIF neurons can be controlled by the threshold voltage ($V_{TH}$) of the FG cell and the threshold voltage to trigger the spike ($V_{TH\_spike}$). In the FG LIF neurons, 'leaky', 'integrate', 'fire' and 'reset' functions can be reproduced by programming the FG cells (for integrate), data retention decaying $V_{TH}$ (for leaky), reading $V_{TH}$ (to detect $V_{TH\_spike}$ for fire) and erasing the FG cells (to reset the $V_{TH}$ to the initial state). The LIF functions and the corresponding FG neuron operations are shown in Fig. 2.

The device parameters of the FG LIF neurons are summarized in Table 1. A planar-type FG cell acts as the FG LIF neuron. The gate coupling ratio is calculated by considering the thicknesses of the block dielectrics and the tunnel oxide as well as the fringe capacitance of the floating gate.

### A. ISI MODELING

During the integration phase, the input bias is applied to the control gate. Electrons are injected to FG through the tunnel oxide by tunneling mechanisms. Both of Fowler-Nordheim (FN) tunneling and direct tunneling (DT) are considered in the simulation. The dominant injection mechanism is chosen based on the voltage across the tunnel oxide [27].

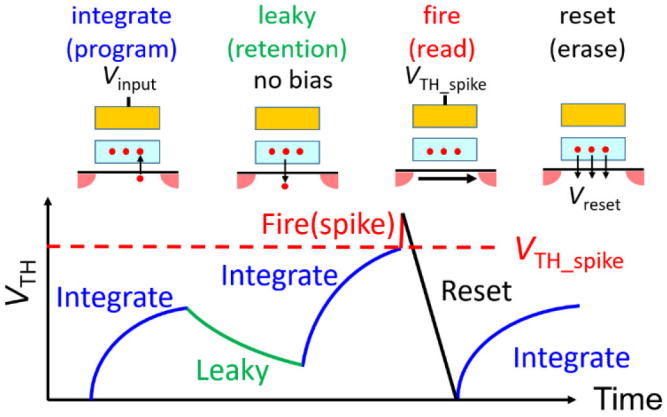



**FIGURE 2.** Leaky-integrate-fire (LIF) operations in FG cell.

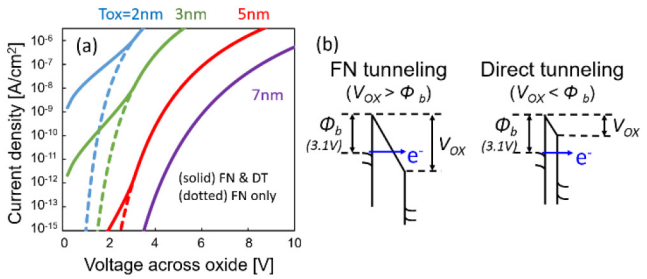



**FIGURE 3.** Electron injection through tunnel oxide. (a) Calculated J-V curves. (b) Band diagrams.

FN tunneling current density is calculated by;

$$J_n = AE_{ox}^2 \mathrm{e}^{-\frac{B}{Eox}} \tag{1}$$

DT current density is calculated by a simplified model in [27];

$$J_n = AE_{ox}^2 \mathrm{e}^{\frac{-B\left[1-\left(1-\frac{V_{OX}}{\Phi \mathrm{b}}\right)^{\frac{3}{2}}\right]}{Eox}} \tag{2}$$

where $J_n$ is the current density, $E_{ox}$ is the electric field across the tunnel oxide, and $V_{ox}$ is the voltage across the tunnel oxide. $\Phi_b$ is the barrier height of the tunnel oxide.

The FN tunneling current (1) is used for $Vox > \Phi_b$ while the DT current (2) is assumed for $Vox < \Phi_b$. Fig. 3(a) shows the simulated tunneling current density. The band diagrams of the FN and DT injections are shown in Fig. 3(b). For the thin oxide at 3 nm or below, the DT injection becomes dominant especially at the low voltages. This result suggests the possibility of the low voltage operation by enabling DT injection with the thin oxide.

The time evolution of $V_{TH}$ for the FG LIF neuron under the constant gate bias is simulated in Fig. 4(a) by calculating the amount of the charges due to the tunneling injections. To produce the LIF neuron function, the $V_{TH}$ needs to be reset when it reaches the predetermined spike threshold voltage ($V_{TH\_spike}$). Fig. 4 (b) shows the simulation results with the reset operations at $V_{TH\_spike} = 1.0$ V. The reset operation is completed in 1 µsec. And $V_{TH}$ is brought back to the initial $V_{TH}$ of 0 V. By adding the reset operations, $V_{TH}$ oscillates over time. The interval between the spikes (the interval between the peaks of $V_{TH}$) is defined as the inter spike interval (ISI).

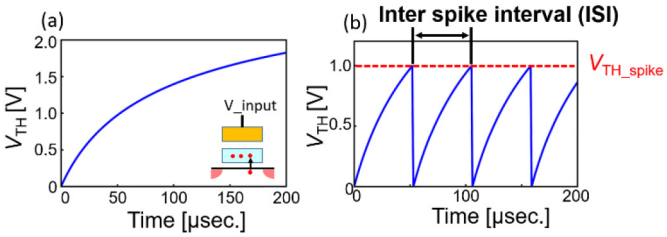



**FIGURE 4.** Modeled time evolution of $V_{TH}$ for FG LIF neuron. (a) Integration without reset. (b) With reset operations at $V_{TH\_spike}$.

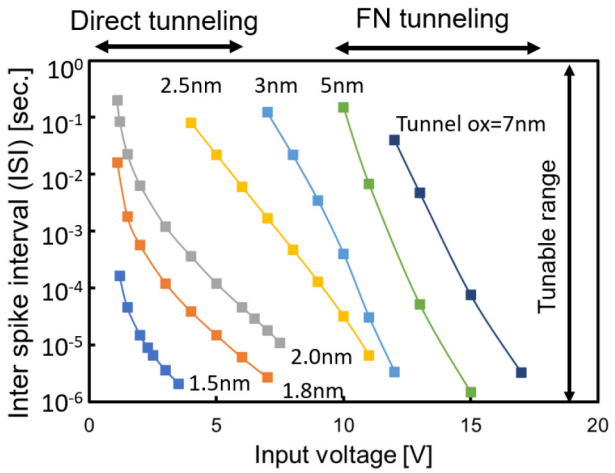



**FIGURE 5.** ISI as a function of input voltage with tunnel oxide thickness as a parameter.

In the FG LIF neuron, ISI is dictated by the programming time of the FG cell under the constant gate voltage. Because of this, ISI is controllable by adjusting the device and operational parameters. Fig. 5 shows the ISI dependence on the input voltage and tunnel oxide thickness. The ISI is tunable over a very wide range of 6 orders of magnitude. By using a thin tunnel oxide (< 3 nm), DT becomes the dominant injection mechanism, and the low voltage operation (~2 V) is realized. Fig. 6 (a) and (b) show how ISI fine tuning can be performed under the DT conditions. By

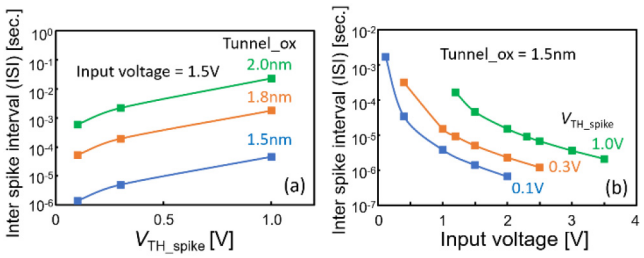



**FIGURE 6.** (a) ISI as a function of $V_{TH\_spike}$ with tunnel oxide thickness as a parameter. (b) ISI as a function of input voltage with $V_{TH\_spike}$ as a parameter.

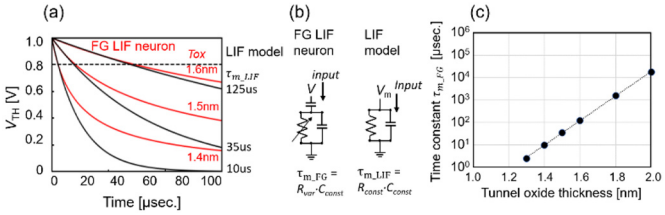



**FIGURE 7. (a) Leaky characteristics comparisons between FG LIF neuron (red) and LIF model (black). (b) Equivalent circuits for FG LIF neuron and LIF model. (c) FG LIF neuron time constant dependence on tunnel oxide thickness.**

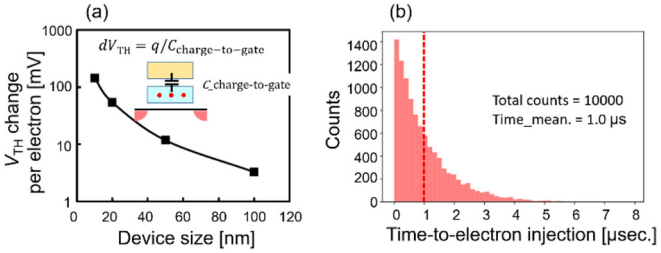



**FIGURE 8. (a) Threshold voltage change ($V_{TH}$ change) per electron. (b) Exponential distribution of time-to-injection of each electron.**

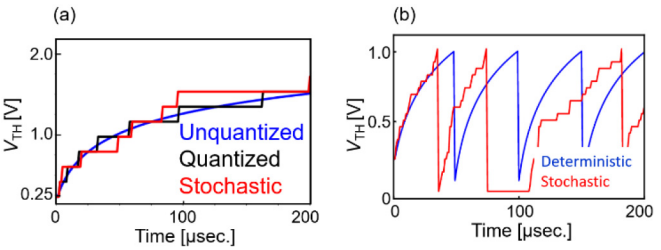



**FIGURE 9. (a) $V_{TH}$ evolution of unquantized (blue), quantized (black) and stochastic (red) injections. (b) ISI dynamics for deterministic injection and stochastic injection.**

adjusting $V_{TH\_spike}$ or input voltage, even under the restricted condition of DT, ISI can be controlled within a range of a few orders of magnitude.

### B. LEAKY CHARACTERISTICS MODELING

Running the FG LIF neuron under the DT condition is critical to realizing the leaky function, too. In Fig. 7, the leaky characteristics of the FG LIF neuron are compared with the widely used LIF neuron model. The key feature of the FG neuron is found to be the variable time constant.

In the common LIF neuron model, the leaky characteristics are defined as;

$$V_{m(t)} = V_{m(t_0)} exp\left(\frac{\tau_{m\_LIF}}{t}\right) \tag{3}$$

where, $V_{\_m(t)}$ is the membrane potential at time t and $\tau_{m\_LIF}$ is the time constant of the neuron which is defined as RC. In Fig. 7, the voltage decay curves for three different $\tau_{m\_LIF}$ are plotted (black curves).

In the FIG neuron, the leakage originates from the DT tunneling current in (2). As seen in the curves in red in Fig. 7(a), the voltage decay of the FG LIF neuron saturates sooner than that of the LIF model. This means that the time constant for FG LIF neurons $\tau_{m\_FG}$ increases over the time. The time constant $\tau_{m\_FG}$ is defined as RC where R is the resistance of the tunnel oxide and C is the capacitance of the FG cell. Due to the exponential dependence of the tunneling current on the electric field in (2), which is the nonohmic conduction, R is variable over the time. The equivalent circuits dictating the time constant for the FG LIF neurons and the LIF model are compared in Fig. 7(b). This variable time constant is a unique characteristic of the FG LIF neuron. The effect of the variable time constant on the SNN computing is not clear at this point and requires further investigation.

The leaky characteristics can be controlled by adjusting the tunnel oxide thickness. Fig. 7(c) shows the leaky time constant as a function of the tunnel oxide thickness. Due to the variable time constant of the FG LIF neuron, $\tau_{m\_FG}$ is defined for the corresponding $\tau_{m\_LIF}$ to match the time to reach $V_{TH}$ of 0.8 V (the horizontal line in Fig. 7(a)). As seen in Fig. 7(c), the leaky time constant of the FG LIF neuron can be targeted over a very wide range (> 3 orders of magnitude) by adjusting the tunnel oxide thickness.

## III. INDIVIDUAL STOCHASTIC FG-LIF NEURON

In this chapter, the spiking stochasticity is analyzed for the individual neuron characteristics.

### A. ORIGIN OF ISI STOCHASTICITY

Due to the nature of the quantum process of the electron tunneling, the $V_{TH}$ evolution of the FG LIF neuron has discrete and stochastic characteristics [28], [29].

Fig. 8(a) shows the device scaling effects on $dV_{TH}$ per electron. The $dV_{TH}$ per electron can be as large as ~200 mV with the device dimension of ~10 nm because of the very small FG-to-CG capacitance, leading the quantization of the $V_{TH}$ evolution. In addition, the time-to-electron injection is stochastic and follows the exponential distribution [28], [30]. This introduces the stochastic behaviors of the $V_{TH}$ evolution.

Fig. 8(b) shows the simulation results of the time-to-injection distribution with a mean value of 1 μs. A total of 10,000 injection events are simulated. In the exponential distribution, it should be noted that the peak of the time-to-injection is much shorter than the mean value. The consequence of this feature will be further discussed in the Section IV.

After including the quantization and stochasticity effects, the $V_{TH}$ evolution is simulated in Fig. 9(a). The smooth curve in blue is from the analytical (deterministic) model. With the quantization effect, the discrete $V_{TH}$ shift is observed (black). By adding the stochasticity to the injection process (red), the time-to-injection shows fluctuations to shorter or longer than the mean injection time.

The ISI characteristics of the stochastic electron injection are simulated (the red curve in Fig. 9(b)). The injection time

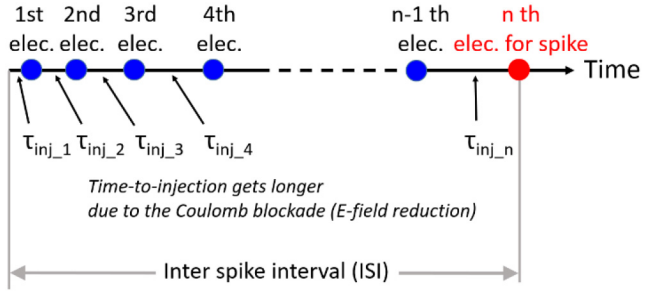



**FIGURE 10.** Process to spike of FG LIF neurons dictated by the accumulative Poisson injections.

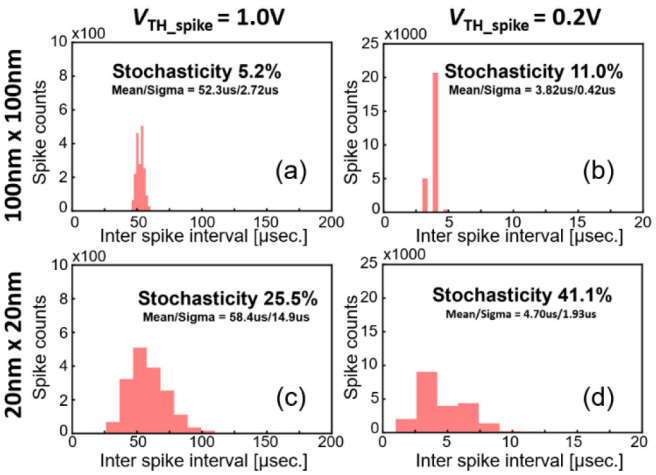



**FIGURE 11.** ISI histograms for various $V_{TH\_spike}$ and device sizes [26].

fluctuates for each single electron injection, resulting in the stochasticity of ISI. Fig. 10 illustrates the relation between the electron injection stochasticity and ISI stochasticity.

When there are n number of electrons injected to reach $V_{TH\_spike}$, the distribution function of a given x (total injection time, ISI) is given by;

$$f(x) = \sum_{i=1}^{n} (1/\tau \text{inj}\_i) \cdot \exp\ (-x/\tau \text{inj}\_\text{i}\ ) \qquad (4)$$

where $\tau_{\text{inj\_i}}$ is the mean time of the ith electron injection. It should be noted that the $\tau_{\text{inj\_i}}$ increases after each electron injection due to the Coulomb blockade because the injected electron reduces the electric field across the tunnel oxide [30].

### B. TARGETING ISI STOCHASTICITY

Fig. 11 shows the simulation results of the ISI distribution of ~200 spikes with varying the device size and $V_{TH\_spike}$. The ISI stochasticity is defined as ISI_sigma/ISI_mean in percentage obtained by the Gaussian fitting of the ISI distributions. Larger ISI stochasticity correlates with a wider ISI distribution. As seen in Fig. 11, the ISI stochasticity increases for a smaller device size and smaller $V_{TH\_spike}$, which means a fewer number of electrons to spike (NES) [26].

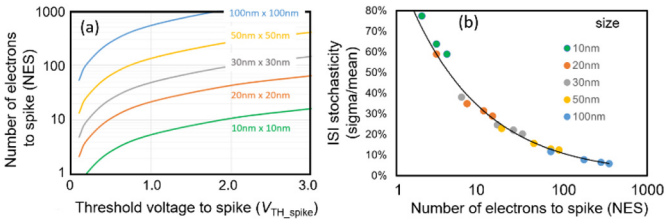



**FIGURE 12.** (a) Number of electrons to spike (NES) for various device size and $V_{TH\_spike}$. (b) ISI stochasticity as a function of NES. Device size and $V_{TH\_spike}$ are varied.

As discussed in Fig. 10 and (4), the ISI is the total time of multiple stochastic electron injections. When more electrons are injected, the injection time fluctuations are averaged out, reducing the stochasticity of the total injection time. Therefore, fewer NESs lead to larger stochasticity due to the lack of the averaging effect. From this consideration, it can be stated that the controlling NES is a key to controlling the ISI stochasticity.

Fig. 12(a) shows the calculated NES for various $V_{TH\_spike}$ and device sizes. NES can be controlled in a very wide range from only a few electrons to more than a thousand electrons. Fig. 12 (b) shows the simulated ISI stochasticity normalized by NES. The universal relationship between ISI and NES is confirmed.

These simulations reveal that the FG LIF neuron has the tunability of ISI stochasticity to a desired value by controlling the NES by targeting $V_{TH\_spike}$ and device size.

## IV. POPILATIONS OF STOCHASTIC FG-LIF NEURONS

In the previous chapter, the ISI stochasticity was discussed by focusing on the single neuron behaviors. In this chapter, the neuron stochasticity is discussed as the behavior of populations of the neurons.

### A. FAST RESPONDING NEURONS

As discussed earlier, the ISI stochasticity is closely tied with the number of electrons to spike (NES). For a single electron injection, the ISI distribution exactly follows the exponential distribution. As the NES increases, the ISI distribution approaches a Gaussian-like distribution because of the convolution of the multiple exponential distributions with various mean time-to-injection (shown in (4)).

In Fig. 13, ISI histograms are simulated for two extreme cases. One is for a very few electrons to spike (NES = 3). This is the case when the device is small and the $V_{TH\_spike}$ is set low. The ISI histogram follows an exponential-like distribution (Fig. 13(a)). The other case is for many electrons to spike (NES = 350). This is the case for a large device and high $V_{TH\_spike}$. The Gaussian-like ISI distribution is seen (Fig. 13(b)).

As seen in Fig. 13(a), the key feature of the exponential-like distribution is the asymmetry where the peak of the ISI distribution is shifted toward a shorter time than the ISI_mean. Consequently, when there are multiple neurons

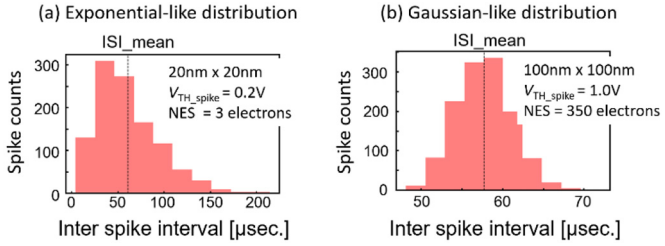



**FIGURE 13.** ISI histograms for (a) exponential-like distribution with few electrons to spike and (b) Gaussian-like distribution with many electrons to spike.

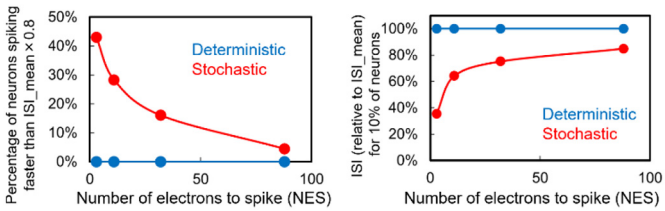



**FIGURE 14.** Few electron effects on the fast-spiking neurons. (a) Percentage of neurons spiking faster than ISI_mean x 0.8. (b) ISI of the fast 10% of neurons.

in a system, the majority of the neurons spike faster than ISI_mean. This subset of the fast-spiking neurons can represent the characteristics of the entire population of the neurons [22]. Hence the population of the neurons can respond faster than the individual neurons that respond with ISI_mean.

The characteristics of the fast-spiking neurons are further investigated from both the quantity and speed perspectives. Fig. 14 (a) shows the percentage of the neurons spiking faster than 0.8x of ISI_mean. All of the deterministic neurons spike exactly at the ISI_mean, so 0% of neurons spike at faster than ISI_mean x0.8 (blue in Fig. 14(a)).

With respect to the stochastic neurons (red in Fig. 14(a)), a subset of the neurons spike at a faster rate due to the electron injection stochasticity. When the NES is large, the ISI distribution of the stochastic neurons is Gaussian-like with the tight sigma, therefore, a very small percentage of the neurons spike faster than the ISI_mean x0.8. On the other hand, with a small NES such as fewer than ten, approximately 30% of the neurons spike faster than the ISI_mean x0.8. due to the exponential-like asymmetric distribution.

Fig. 14 (b) shows the speed of the fast-spiking neuron sub-set at 10% of the entire population (fast 10% neurons). For the deterministic neurons (blue in Fig. 14(b)), all neurons respond at ISI_mean (=100% of ISI_mean). For the stochastic neurons (red in Fig. 14(b)), the 10% fast neurons spike faster than ISI_mean. When NES is large, the response time of the fast 10% neurons is about 80% of the ISI_mean. This means that the populations of neurons response about 20% faster than the single neuron can response. When NES is scaled to less than 10 electrons, the response time of the fast 10% neurons can be significantly faster than the single neuron response time, such as shorter than the 50% of ISI_mean.

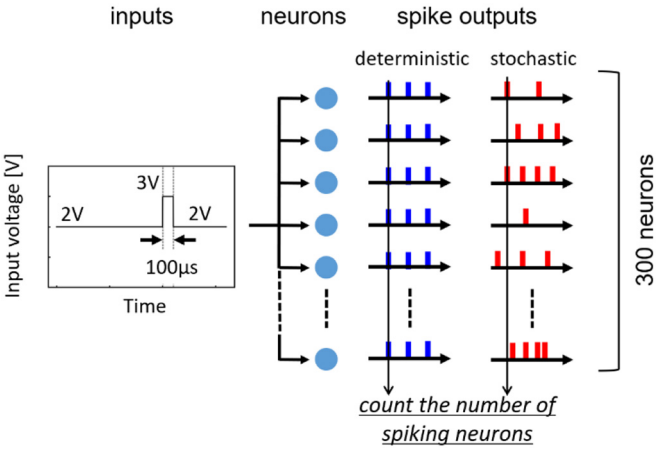



**FIGURE 15.** Experimental set up for population-based coding.

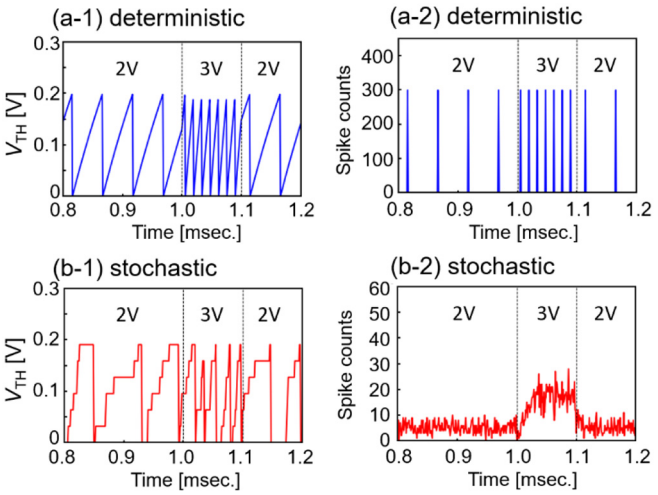



**FIGURE 16.** $_{TH}$ evolution of single neuron and population of spikes. (a-1) deterministic, single neuron, (a-2) deterministic, populations of neurons, (b-1) stochastic, single neuron, (b-2) stochastic, populations of neurons, 30nm x 30nm, $V_{TH\_spike}$ = 0.2V.

These results demonstrate the proof of the concept that the fast response of the populations can be realized by the exponential-like asymmetric distribution with aggressively scaled NES.

### *B. POPULATION-BASED CODING*

The population code is one of the coding techniques for SNNs [31], [32], [33]. Signals are encoded as numbers of spiking neurons at each time slice. Fig. 15 is a diagram describing the simulation setup in this work. The input waveform consists of the 2 V base bias with the 3 V pulse of 100 μs duration. This input waveform is applied to a group of 300 neurons and the number of spiking neurons is counted at each time slice. Two types of the neuron characteristics are simulated. One is a deterministic neuron (blue) and the other is a stochastic neuron (red).

The simulation results are shown in Fig. 16 for both individual neuron behaviors (a-1 and b-1) and the neuron population behaviors (a-2 and b-2). For the individual neuron

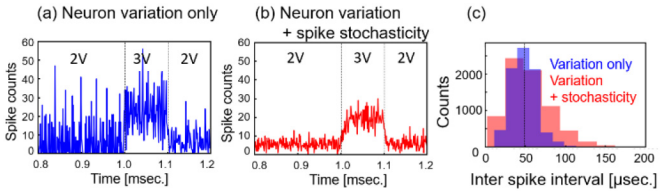



**FIGURE 17.** Variation and stochasticity effects on spike population. NES=3. (Size=30nm×30nm, $V_{TH\_spike}$ = 200mV). (a) Neuron-to-neuron variation only ($V_{TH\_sigma}$ = 50mV). (b) Variation (neuron-to-neuron) + stochasticity (spike-to-spike). (c) ISI histograms for 2V input.

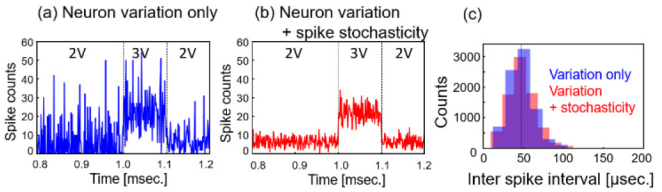



**FIGURE 18.** Variation and stochasticity effects on spike population. NES=70. (Size=100nm×100nm, $V_{TH\_spike}$ = 200mV). (a) Neuron-to-neuron variation only ($V_{TH\_sigma}$ = 50mV). (b) Variation (neuron-to-neuron) + stochasticity (spike-to-spike). (c) ISI histogram for 2V input.

behaviors, one neuron was randomly sampled, while all 300 neurons are included in the neuron population behaviors.

The deterministic neuron shows the constant ISI (Fig. 16 a-1) with a given input bias. Since all neurons have the exact same ISI, all of the 300 neurons spike simultaneously (Fig. 16 a-2). Therefore, the change in the input bias cannot be detected in the number of spiking neurons. It can be detected as the modulation of ISI instead.

On the other hand, with the stochastic neurons (Fig. 16 b-1), the timing of the spikes is distributed among the spikes. As a result, the change in the input bias can be detected as the number of spiking neurons (Fig, 16 b-2), thus enabling the population coding.

Next, the effects from different types of stochasticity are analyzed. There are two types of spiking neuron stochasticity. One is the neuron-to-neuron variation originating from the manufacturing process variability. The other is the spike-to-spike stochasticity in the same neuron originating from the electron injection stochasticity. NES is set to 3 with a 30 nm device size and 0.2 V $V_{\mathrm{TH\_spike}}$. Fig. 17 (a) shows the evolution of the number of spiking neurons when only the neuron-to-neuron variation is considered. For the FG LIF neuron device variation, the Gaussian distribution of 50 mV $V_{\mathrm{TH}}$ sigma is considered. By introducing the neuron-to-neuron variability, the change in the input can be detected by the populations of the spiking neurons. However, the baseline noise is significant and potentially degrades the detection capability. Fig. 17 (b) shows the case with the electron injection stochasticity (spike-to-spike stochasticity) in addition to the neuron-to-neuron variability of the 50 mV $V_{\mathrm{TH}}$ sigma. As seen in the simulation results, the baseline noise is dramatically reduced by introducing the spike-to-spike stochasticity. Fig. 17(c) shows ISI histograms with neuron-to-neuron variation only and with neuron-to-neuron variation and spike-to-spike stochasticity. ISI stochasticity increases by adding the spike stochasticity.

To further segment the effect between the type of the stochasticity and the degree of the stochasticity further, the large NES of 70 is evaluated in Fig. 18. Fig. 18 (a) corresponds with the case with neuron variation, and Fig. 18 (b) includes neuron variation and spike stochasticity. The noise reduction effect is observed with the spike stochasticity even with the large NES case. Due to the large NES, the degree of the total ISI stochasticity is almost unchanged by introducing the spike stochasticity (Fig. 18 (c)). This means that the noise reduction effect originates from the type of the stochasticity as opposed to the degree of the stochasticity.

These results can be understood as follows. When only the neuron-to-neuron variability is introduced, each neuron still spikes at a constant interval. Therefore, the number of spiking populations increases when the timing is in accordance with the common multiples of ISIs from many neurons. This introduces the periodic peaks of the baseline noise under the constant input bias. In the contrast, with the spike-to-spike stochasticity, the ISI of each neuron is not defined, so there is no peak of the spikes due to the common multiples of ISIs. As a result, the baseline noise is dramatically suppressed. Based on this mechanism, the spike-to-spike stochasticity due to the stochastic electron injection is expected to improve the computing accuracy in the population coding by reducing the baseline noise.

In conclusion, in the population coding, the speed of the response is dictated by the degree of the stochasticity which is further accelerated by the exponential-like ISI distributions. The noise reduction depends on the type of the stochasticity, where the spike-to-spike stochasticity plays the dominant role.

## V. STEP-BY-STEP DEVICE DESIGN FOR FG LIF NEURON

Based on the dependencies of the neuron characteristics on the device and operation parameters revealed in this work, the step-by-step design procedures of the stochastic FG LIF neuron device are proposed (Fig. 19). In the proposed procedures, the device parameters are set first as these are tied with the hardware manufacturing while the operation parameters can be set later because these parameters can be flexibly set by the software.

In the first step, the device size is set based on the process capability. The planar FG cell can be scaled down all the way below 20 nm [34]. By using the relation given in the Fig. 8(a), the number of electrons for the desired amount of the shift of $V_{\mathrm{TH}}$ can be obtained. When a smaller device size is chosen, the stochasticity becomes large due to fewer number of electrons. If the stochasticity is not needed, a large device or multiple small devices jointed together can be used.

In the second step, the tunnel oxide thickness should be determined based on the desired neuron time constant. Fig. 7 (c) provides the relation between the tunnel oxide thickness and the neuron time constant. The thinner tunnel

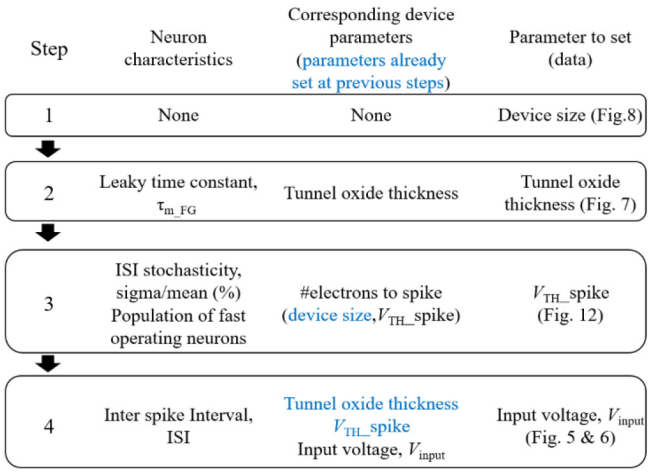



**FIGURE 19. Step-by-step procedures for designing FG neuron device and operation conditions.**

**TABLE 2. Tunability of neuron characteristics for FG LIF neuron.**

| Neuron characteristics | Tunable range | Data |
|---|---|---|
| Operation voltage | 0.5V~3V | Fig. 5&6 |
| Number of electrons to spike (NES) | <1 ~ >1000 electrons | Fig. 12 |
| 'Leaky' time constant | <1μsec. ~ >10msec. | Fig. 7 |
| ISI_mean | μsec. ~ sec. | Fig. 5&6 |
| ISI_stochasticity | ~0% to ~60% (sigma/mean) | Fig. 12 |
| ISI distribution | Exponential-like to Gaussian-like | Fig.13 |

oxide realizes a short time constant for both 'integrate' and 'leaky' functions.

Next, in the third step, $V_{TH\_spike}$ is set for a desired degree of the ISI stochasticity. As shown in Fig. 12(b), the stochasticity is a universal function of NES which is dictated by the combination of the device size and $V_{TH\_spike}$. Given that the device size cannot change after the manufacturing, $V_{TH\_spike}$ is the key parameter to adjust the ISI stochasticity post manufacturing. By integrating $V_{TH}$ change per electron given at the first step and $V_{TH\_spike}$ set at this step, NES can be fine-tuned to realize the desired ISI stochasticity.

Finally, in step four, the input voltage ($V_{input}$) is determined for the 'integrate' operation. Given that the tunnel oxide thickness is already set at the previous steps, the ISI_mean can be set by the input voltage with the relation shown in Fig. 5 and Fig. 6. By adjusting the input voltage, the time constant can be independently set between the 'integrate' operation and 'leaky' operation. This is because the 'leaky' time constant is a solo function of the tunnel oxide thickness while the 'integrate' time constant is determined by a combination of the tunnel oxide thickness and the input voltage.

The tunable ranges of the neuron characteristics are summarized in Table 2. In the proposed FG LIF neurons, all of the key characteristics (operation voltage, NES, time constant, ISI mean and stochasticity) are tunable over very wide ranges.

## VI. FUTURE CHALLENGES

While the FG LIF neurons have the great advantages in the tunability of the neuron characteristics as well as the device scalability, there are several challenges to overcome.

The tunnel oxide reliability is one of the device concerns. Since the neuron is a switching device, electron injections and emissions occur at every computing operation, leading to many program/erase cycles to the FG neuron devices. The ultra-thin tunnel oxide combined with the very low voltage operation is expected to relieve the oxide degradation.

The peripheral circuits controlling the FG LIF neurons should be another challenge. While the FG LIF neurons replace the large CMOS circuits performing the LIF operation, the new additional circuits are required to control the FG LIF neuron operations such as program, erase and read. It's critical that the entire neuron circuits are kept small enough to realize the scaling advantage of the FG LIF neurons.

From the manufacturing perspective, the isolated patterning of the FG LIF neuron devices would rise the challenge. Compared to the synaptic array, the neuron circuit tends to have less dense layout. As a result, the FG LIF neurons can be placed in a relatively isolated manner. This would cause the challenge in the patterning especially for the aggressively scaled device.

For the further model enhancement, while the electron injection stochasticity is focused in this work, there is an interest to understand the interactions with other cell noises such as random telegraph signal (RTN). Integrating various other noises is required for the more complete model.

These challenges need to be overcome in the future.

## VII. CONCLUSION

An analytical model for the stochastic FG LIF neuron has been developed. The wide range of tunability of the neuron characteristics is shown and the step-by-step procedures to design the device and operation conditions are proposed. The ISI stochasticity originates from the tunneling electron injection statistics governed by the number of electrons to spike. The spike response of the neuron population becomes faster owing to the exponential time distribution of the scaled FG LIF neurons. In the population coding, the proof of the concept is demonstrated for the signal-to-noise ratio enhancement by introducing the spike stochasticity. This work supports device and operation optimization of the FG LIF neurons. In addition, this work contributes to developing insights of inherent stochasticity of NVM devices which could enable the energy efficient neuromorphic computing.

## REFERENCES

[1] W. Maass, "Networks of spiking neurons: The third generation of neural network models," *Neural Netw.*, vol. 10, no. 9, pp. 1659–1671, 1997, doi: 10.1016/S0893-6080(97)00011-7.

[2] M. V Debole *et al.*, "TrueNorth: Accelerating from zero to 64 million neurons in 10 years," *Computer*, vol. 52, no. 5, pp. 20–29, May 2019, doi: 10.1109/MC.2019.2903009.

[3] M. Davies *et al.*, "Loihi: A neuromorphic manycore processor with on-chip learning," *IEEE Micro*, vol. 38, no. 1, pp. 82–99, Jan./Feb. 2018, doi: 10.1109/MM.2018.112130359.
[4] S. Yu, "Neuro-inspired computing with emerging nonvolatile memorys," *Proc. IEEE*, vol. 106, no. 2, pp. 260–285, Feb. 2018, doi: 10.1109/JPROC.2018.2790840.
[5] J. Zhu, T. Zhang, Y. Yang, and R. Huang, "A comprehensive review on emerging artificial neuromorphic devices," *Appl. Phys. Rev.*, vol. 7, no. 1, 2020, Art. no. 11312, doi: 10.1063/1.5118217.
[6] I. Chakraborty, A. Jaiswal, A. K. Saha, S. K. Gupta, and K. Roy, "Pathways to efficient neuromorphic computing with non-volatile memory technologies," *Appl. Phys. Rev.*, vol. 7, no. 2, pp. 1–30, 2020, doi: 10.1063/1.5113536.
[7] J. Gupta and D. Koppad, "A survey on memristor and CMOS based spiking neural networks," in *Proc. 2nd Int. Conf. Inven. Res. Comput. Appl. (ICIRCA)*, 2020, pp. 1052–1058, doi: 10.1109/ICIRCA48905.2020.9183111.
[8] M. Kim *et al.*, "A 3D NAND flash ready 8-bit convolutional neural network core demonstrated in a standard logic process," in *Int. Electron Devices Meeting Tech. Dig. (IEDM)*, 2019, pp. 923–926, doi: 10.1109/IEDM19573.2019.8993574.
[9] H. T. Lue *et al.*, "Optimal design methods to transform 3D NAND flash into a high-density, high-bandwidth and low-power nonvolatile computing in memory (nvCIM) accelerator for deep-learning neural networks (DNN)," in *Int. Electron Devices Meeting Tech. Dig. (IEDM)*, 2019, pp. 915–918, doi: 10.1109/IEDM19573.2019.8993652.
[10] X. Wu, V. Saxena, K. Zhu, and S. Balagopal, "A CMOS spiking neuron for brain-inspired neural networks with resistive synapses and in situ learning," *IEEE Trans. Circuits Syst. II, Exp. Briefs*, vol. 62, no. 11, pp. 1088–1092, Nov. 2015, doi: 10.1109/TCSII.2015.2456372.
[11] J. M. Cruz-albrecht, M. W. Yung, and N. Srinivasa, "Energy-efficient neuron, synapse and stdp integrated circuits," *IEEE Trans. Biomed. Circuits Syst.*, vol. 6, no. 3, pp. 246–256, Jun. 2012, doi: 10.1109/TBCAS.2011.2174152.
[12] W. H. Brigner *et al.*, "Shape-based magnetic domain wall drift for an artificial spintronic leaky integrate-and-fire neuron," *IEEE Trans. Electron Devices*, vol. 66, no. 11, pp. 4970–4975, Nov. 2019, doi: 10.1109/TED.2019.2938952.
[13] V. Kornijcuk *et al.*, "Leaky integrate-and-fire neuron circuit based on floating-gate integrator," *Front. Neurosci.*, vol. 10, pp. 1–16, May 2016, doi: 10.3389/fnins.2016.00212.
[14] X. Wang *et al.*, "A novel RRAM-based adaptive-threshold LIF neuron circuit for high recognition accuracy," in *Proc. Int. Symp. VLSI Technol. Syst. Appl. (VLSI-TSA)*, 2018, pp. 1–2, doi: 10.1109/VLSI-TSA.2018.8403854.
[15] M.-H. Wu *et al.*, "Extremely compact integrate-and-fire STT-MRAM neuron: A pathway toward all-spin artificial deep neural network," in *Proc. Symp. VLSI Technol.*, 2019, pp. T34–T35, doi: 10.23919/VLSIT.2019.8776569.
[16] C. Chen *et al.*, "A photoelectric spiking neuron for visual depth perception," *Adv. Mater.*, vol. 34, no. 20, pp. 1–9, 2022, doi: 10.1002/adma.202201895.
[17] H. Mao *et al.*, "A spiking stochastic neuron based on stacked InGaZnO memristors," *Adv. Electron. Mater.*, vol. 8, no. 2, pp. 1–7, 2022, doi: 10.1002/aelm.202100918.
[18] T. Tuma, A. Pantazi, M. Le Gallo, A. Sebastian, and E. Eleftheriou, "Stochastic phase-change neurons," *Nat. Nanotechnol.*, vol. 11, no. 8, pp. 693–699, 2016, doi: 10.1038/nnano.2016.70.
[19] A. Agrawal *et al.*, "Revisiting stochastic computing in the era of nanoscale nonvolatile technologies," *IEEE Trans. Very Large Scale Integr. (VLSI) Syst.*, vol. 28, no. 12, pp. 2481–2494, Dec. 2020, doi: 10.1109/TVLSI.2020.2991679.
[20] W. Maass, "Noise as a resource for computation and learning in networks of spiking neurons," *Proc. IEEE*, vol. 102, no. 5, pp. 860–880, May 2014, doi: 10.1109/JPROC.2014.2310593.
[21] M. D. McDonnell and L. M. Ward, "The benefits of noise in neural systems: Bridging theory and experiment," *Nat. Rev. Neurosci.*, vol. 12, no. 7, pp. 415–425, 2011, doi: 10.1038/nrn3061.
[22] T. Tchumatchenko, A. Malyshev, F. Wolf, and M. Volgushev, "Ultrafast population encoding by cortical neurons," *J. Neurosci.*, vol. 31, no. 34, pp. 12171–12179, 2011, doi: 10.1523/JNEUROSCI.2182–11.2011.
[23] B. B. Averbeck, P. E. Latham, and A. Pouget, "Neural correlations, population coding and computation," *Nature Rev. Neurosci.*, vol. 7, pp. 358–366, May 2006, doi: 10.1038/nrn1888.
[24] M. C. W. Van Rossum, G. G. Turrigiano, and S. B. Nelson, "Fast propagation of firing rates through layered networks of noisy neurons," *J. Neurosci.*, vol. 22, no. 5, pp. 1956–1966, 2002, doi: 10.1523/jneurosci.22–05-01956.2002.
[25] D. S. Jeong, "Tutorial: Neuromorphic spiking neural networks for temporal learning," *J. Appl. Phys.*, vol. 124, no. 15, 2018, Art. no. 152002, doi: 10.1063/1.5042243.
[26] A. Goda, C. Matsui, and K. Takeuchi, "Inter spike interval and stochasticity engineering of floating gate technology-based neurons for spiking neural network hardware," in *Proc. 6th IEEE Electron Devices Technol. Manuf. Conf. (EDTM)*, 2022, pp. 129–131, doi: 10.1109/EDTM53872.2022.9798349.
[27] K. F. Schuegraf and C. Hu, "Hole injection $SiO_2$ breakdown model for very low voltage lifetime extrapolation," *IEEE Trans. Electron Devices*, vol. 41, no. 5, pp. 761–767, May 1994, doi: 10.1109/16.285029.
[28] C. M. Compagnoni, R. Gusmeroli, A. S. Spinelli, and A. Visconti, "Analytical model for the electron-injection statistics during programming of nanoscale NAND Flash memories," *IEEE Trans. Electron Devices*, vol. 55, no. 11, pp. 3192–3199, Nov. 2008, doi: 10.1109/TED.2008.2003332.
[29] G. Molas *et al.*, "Impact of few electron phenomena on floating-gate memory reliability," in *Int. Electron Devices Meet. (IEDM) Tech. Dig.*, 2004, pp. 877–880, doi: 10.1109/iedm.2004.1419320.
[30] K. Yano *et al.*, "Single-electron memory for giga-to-tera bit storage," *Proc. IEEE*, vol. 87, no. 4, pp. 633–651, Apr. 1999, doi: 10.1109/5.752519.
[31] D. Auge, J. Hille, E. Mueller, and A. Knoll, "A survey of encoding techniques for signal processing in spiking neural networks," *Neural Process. Lett.*, vol. 53, no. 6, pp. 4693–4710, 2021, doi: 10.1007/s11063–021-10562–2.
[32] P. Alexandre, D. Peter, and Z. Richard, "Information processing with population codes," *Nat. Rev. Neurosci.*, vol. 1, pp. 125–132, Nov. 2000, [Online]. Available: www.nature.com/reviews/neuro
[33] H. Fang, Y. Zeng, and F. Zhao, "Brain inspired sequences production by spiking neural networks with reward-modulated STDP," *Front. Comput. Neurosci.*, vol. 15, pp. 1–13, Feb. 2021, doi: 10.3389/fncom.2021.612041.
[34] S. Lee *et al.*, "A 128Gb 2b/cell NAND flash memory in 14nm technology with tPROG = 640µs and 800MB/s I/O rate," in *Proc. IEEE Int. Solid-State Circuits Conf. (ISSCC)*, 2016, pp. 138–139, doi: 10.1109/ISSCC.2016.7417945.