Investigation of Deep Spiking Neural Networks Utilizing Gated Schottky Diode as Synaptic Devices

Sung-Tae Lee 1,* and Jong-Ho Bae 2, *

1 Department of Electronic Engineering, Gachon University, Seongnam-si 13120, Gyeonggi-do, Korea
2 School of Electrical Engineering, Kookmin University, Seongbuk-gu, Seoul 02707, Korea
* Correspondence: lst777@gachon.ac.kr (S.-T.L.); jbae@kookmin.ac.kr (J.-H.B.)

Abstract: Deep learning produces a remarkable performance in various applications such as image classification and speech recognition. However, state-of-the-art deep neural networks require a large number of weights and enormous computation power, which results in a bottleneck of efficiency for edge-device applications. To resolve these problems, deep spiking neural networks (DSNNs) have been proposed, given the specialized synapse and neuron hardware. In this work, the hardware neuromorphic system of DSNNs with gated Schottky diodes was investigated. Gated Schottky diodes have a near-linear conductance response, which can easily implement quantized weights in synaptic devices. Based on modeling of synaptic devices, two-layer fully connected neural networks are trained by off-chip learning. The adaptation of a neuron’s threshold is proposed to reduce the accuracy degradation caused by the conversion from analog neural networks (ANNs) to event-driven DSNNs. Using left-justified rate coding as an input encoding method enables low-latency classification. The effect of device variation and noisy images to the classification accuracy is investigated. The time-to-first-spike (TTFS) scheme can significantly reduce power consumption by reducing the number of firing spikes compared to a max-firing scheme.

Keywords: neuromorphic device; in-memory computing; hardware-based neural networks; deep learning; spiking neural networks; off-chip learning; gated Schottky diode; synaptic device

1. Introduction

Deep learning has shown astonishing achievements in a broad range of tasks [1–5], such as image classification and speech recognition. However, recent high-performance deep neural networks require the high-performance computing resources like GPUs [6–11]. The enormous computational requirements limit the application of deep learning in an edge device. The spiking neural network (SNN) is a promising candidate for low-power hardware by mimicking neural network architectures of a human. Recently, deep spiking neural networks (DSNNs) have been actively studied to implement biologically plausible neural networks and large-scale neuromorphic computing platforms based on specialized analog hardware [12–20]. Hardware-based neural network platforms are much more efficient than CPU and GPU in terms of power consumption for running DSNNs and allow asynchronous and distributed event-driven computation, thereby improving scalability and reducing latencies. Furthermore, event-based asynchronous hardware-based neural network systems utilize their computational power on local active parts of the neural networks, and they efficiently save power of the entire neural networks. Therefore, a DSNNs-based neuromorphic system utilizing synapse and neuron devices is a promising candidate as a platform to run large-scale neural networks in real-time applications.

SNNs can be trained by spike-based on-chip learning algorithms, such as backpropagation and spike-timing-dependent plasticity (STDP) [21]. On the other hand, in DSNNs, off-chip learning is used with conventional ANNs followed by a conversion in which SNNs can be converted from conventional ANNs, which can utilize well-developed
training methods and fully trained models of ANNs [22–26]. DSNNs have been a driving factor in the development of many modern computer vision and mobile robots with continuous high-dimensional observation and action spaces [24]. A previous work [22] analyzed the reason of accuracy degradation in detail and proposed weight normalization method to prevent data loss and achieved high inference accuracy close to the inference accuracy of software-based ANNs.

In this work, we investigate the hardware neuromorphic system of DSNNs based on the gated Schottky diode. Note that 2-layer fully connected neural networks are trained for the Modified National Institute of Standards and Technology database (MNIST) image recognition using gradient descent. The trained weights are imported into the hardware system modeled by the characteristics of the fabricated gated Schottky diode used as a synapse device [27,28]. The effective Schottky barrier height of the gated Schottky diode is modulated by the bottom gate bias or program pulses. The Schottky reverse current is exponentially proportional to the amount of stored charge, and the amount of stored charge is logarithmically proportional to the number of applied pulses. These exponential and logarithmic relationships cancel each other out, and this can provide the excellent linearity between the reverse current and the number of applied pulses. There is a trade-off between the endurance and retention characteristics of the devices, which needs to be considered in device design [29,30]. Near-linear conductance response of the gated Schottky diode makes it easy to transfer quantized weights from software to hardware synaptic devices. We propose an adaptation of the neuron’s threshold to reduce the accuracy degradation due to the conversion from ANNs to DSNNs. In addition, we show that the left-justified rate coding method can considerably reduce the number of firing spikes compared to the right-justified rate coding, while achieving the same inference accuracy. The effect of the weight variation and noise in the input images on the classification accuracy is investigated. Finally, we show that the time-to-first-spike (TTFS) scheme can considerably reduce power consumption compared to the max-firing scheme by decreasing the number of spikes.

2. Materials and Methods

The gated Schottky diode (GSD), which is used as synaptic devices, is fabricated [27]. Figure 1a shows a schematic diagram of a fabricated GSD. The minimal unit cell size is $6F^2$ for one GSD device and $12F^2$ for one synaptic device. Two electronic devices are required to represent a unit synapse, because the weights of the unit synapse in neural networks should have positive and negative values for high accuracy. O and S represent the metal electrodes for ohmic-like junction and Schottky junction, respectively. BG_O and BG_S represent the bottom gates under O and S, respectively. The main device fabrication steps are described below, and the detailed fabrication steps are described in the previous work [14]. On top of a 300 nm thick buried SiO$_2$, an $n^+$ poly silicon layer and Si$_3$N$_4$ layer, which are a bottom gate and a sacrificial layer, respectively, were deposited and patterned. A SiO$_2$ layer was thermally grown along the sidewall of the patterned $n^+$ poly silicon to isolate the bottom gates. After that, another $n^+$ poly-Si layer was deposited, followed by chemical mechanical polishing and patterning $n^+$ poly-Si to isolate the bottom gates. Then, the sacrificial Si$_3$N$_4$ was stripped by H$_3$PO$_4$ at 160 °C. A SiO$_2$/Si$_3$N$_4$/SiO$_2$ (12/6/6 nm) charge trap stack was deposited and a 20 nm thick undoped poly silicon active layer was formed. After passivating the device with SiO$_2$, the contact holes for the S, O and bottom gates were formed and Al electrodes were deposited and patterned by thermal evaporation and a lift-off process.

By applying pulse to the bottom gate under the S electrode, the charge stored in the Si$_3$N$_4$ layer is changing, which modulates the Schottky barrier and the conductance of the GSD. As GSD operates as a reverse diode, the output current is represented by the reverse Schottky diode current ($I_R$). Figure 1b represents the conductance ($G$) response of the GSD with respect to the number of applied pulses when $V_{PGM}$ (9 V, 10 µs) is applied to BG_S 10, 15, 35, 64 times in sequence. The conductance response shows excellent repeatability and linearity. Electrons are stored in the Si$_3$N$_4$ layer when $V_{PGM}$ are imposed, which decreases
the Schottky barrier height and increases $I_R$ and $G$, and vice versa. $I_R$ is exponentially proportional to the Schottky barrier height (equivalently the amount of stored charge), and the amount of stored charge is logarithmically proportional to the number of applied pulses. These exponential and logarithmic relationships cancel each other out, and this can provide the excellent linearity between the $I_R$ and the number of applied pulses [27].

![Figure 1. (a) A cross-sectional schematic diagram of the fabricated GSD. (b) Conductance response of GSD with the number of pulses when $V_{FGM}$ (9 V, 10 μs) is applied to BGs for 10, 15, 35, 64 times in sequence.](image)

3. Results

In DSNNs, off-chip learning is used with conventional ANNs followed by a conversion in which SNN is converted from conventional ANNs, which can utilize well-developed training methods and fully trained models of ANNs. A previous work [13] analyzed that the reason of accuracy degradation occurred in conversion from ANNs to SNNs in detail and proposed a weight normalization method to prevent data loss and achieved the result close to that of software-based ANNs. In addition, [31] applied a weight normalization method to hardware-based DSNNs. On the other hand, we propose rescaling of the neuron threshold and apply it to the hardware-based DSNNs, which improves the learning accuracy. As the number of neuron devices is much smaller than that of the synapse device, it is easier to adapt the threshold voltage ($V_{th}$) of the neuron device than to adapt the conductance of the synaptic device. The split-gate positive feedback device can be used as a neuron device which can adapt the threshold of neuron [32].

The multilayer neural network of 784-200-10 is trained for the MNIST dataset using a backpropagation algorithm, achieving an inference accuracy of 98.24%. Then, the trained weights are transferred to hardware-based DSNNs of 784-200-10, reflecting conductance response of the GSD device in Figure 1b. The weights of synapse in the system have values from 1 to −1, and the magnitude of input spike is 1 or 0. Figure 2a shows the voltage of neuron membrane ($V_{mem}$) with respect to the timestep as an example. As the MNIST dataset has a resolution of 8-bit, the input spike in DSNNs has 255 timesteps. As shown in Figure 2a, $V_{mem}$ can exceed two times of the $V_{th}$ of a neuron device. However, the neuron device cannot generate two spikes at one timestep, which results in data overflow and decreases the inference accuracy of neural networks. In this case, the threshold of neuron is low and should be increased to prevent overflow of data. On the other hand, when the $V_{th}$ of a neuron device is too high, the $V_{mem}$ of the output neuron cannot reach to the $V_{th}$ until the last input spike is transferred to the output neuron. In this case, the neuron cannot fire, which results in data underflow and decreases the inference accuracy. The $V_{th}$ of the neuron device should be decreased to prevent underflow of data. As shown in Figure 2b, the inference accuracy is low when the threshold of the neuron is too high or too low due to the data underflow and overflow, respectively. Inference accuracy of 98.22% is achieved by optimizing the $V_{th}$ of the neuron device, which is similar to the inference accuracy of software-based ANN.
To convert input value of software-based ANN to time-series spikes of DSNNs, there are two types of encoding methods, which are rate-based coding and time-based coding [33,34]. In addition, there are the max-firing scheme and TTFS scheme in rate-based coding [35,36]. In the max-firing scheme, the neuron which generates the maximum number of firing spikes in the last neuron layer is selected as a result. There are right-justified rate coding (RRC) and left-justified rate coding (LRC) in rate-based coding, as shown in Figure 3 [31]. As the MNIST image has 256 grayscales, the timesteps of rate coding in DSNNs are 255. The intensity of image pixel is proportional to the number of spikes. RRC fills the spikes from the last timestep, but LRC fills the spikes from the first timestep. In RRC, the last input spike should be generated at the last timestep. On the other hand, in LRC, the last input spike can be generated before the last timestep. Therefore, except for the grayscale of 256, the last spike in LRC is generated earlier than the last spike of RRC. Therefore, in LRC, input spikes generated from the pre-synaptic neuron can be reached to the post-synaptic neuron faster than in RRC. Figure 4 shows the inference accuracy of hardware neural networks when RRC and LRC are used. As shown in Figure 4, the LRC method can achieve high inference accuracy faster than the RRC method. In addition, the LRC method can achieve the same final accuracy as the RRC method.

Figure 2. (a) The voltage of neuron membrane with respect to the timestep. (b) Inference accuracy with respect to the neuron’s threshold in DSNN.
We estimate the effect of device variation (\(\sigma_G/\mu_G\)) to the inference accuracy, as shown in Figure 5. The device variation which is assumed to follow Gaussian distribution is applied to hardware-based DSNNs. As shown in Figure 5a, in all the cases of \(\sigma_G/\mu_G\), the inference accuracy reaches saturated accuracy during the timestep, which means that DSNNs can perform stable operation even with \(\sigma_G/\mu_G\). On the other hand, saturated inference accuracy decreases as \(\sigma_G/\mu_G\) increases. In Figure 5b, simulation is repeated 20 times for each \(\sigma_G/\mu_G\). The median value of inference accuracy decreases as \(\sigma_G/\mu_G\) increases. The inference accuracy decreases by 1% when the \(\sigma_G/\mu_G\) increases to 0.4.

Figure 4. Comparison of inference accuracy in RRC and LRC. The inset shows the final inference accuracy of ANN, LRC and RRC.
Input noise which follows Gaussian distribution with zero mean and standard deviation ($\sigma_{\text{noise}}$) is added to the input image of MNIST dataset to investigate the effect of input noise to the inference accuracy. Figure 6 shows the inference accuracy of neural network with respect to various standard deviation of input noise. It is worth noticing that the inference accuracy does not saturate with the timestep as standard deviation of noise increases. On the other hand, the inference accuracy reaches saturated accuracy when the device variation exists. Therefore, the effect of input noise is more detrimental to the inference accuracy than the effect of device variation.

In addition, the effect of the number of input spike to the inference accuracy is investigated, as shown in Figure 7. The number of input spikes represents the resolution of input data. The MNIST dataset has 8-bit resolution, which corresponds to the 255 timesteps in DSNNs. As the number of timesteps decreases from 255, the data loss occurs in the conversion from MNIST image in ANN to the input spikes in DSNN. Therefore, inference accuracy decreases as the resolution of input data decreases from 255. The inference accuracy decreases by only 0.1% as resolution of input data decreases from 255 to 75. However, inference accuracy significantly decreases when the number of input spikes decreases from 75 to 50. Therefore, the resolution of input spikes should be greater than 75 steps to achieve sufficiently high inference accuracy.

Figure 5. (a) Inference accuracy with increasing $\sigma_{G}/\mu_{G}$ from 0 to 1. (b) Final inference accuracy as a function of device variation ($\sigma_{G}/\mu_{G}$).

Figure 6. Inference accuracy over inference time of DSNN for various standard deviations of input noise ($\sigma_{\text{noise}}$).
As described above, there are two types of schemes, which are the max-firing scheme and the time-to-first-spike (TTFS) scheme in rate-based coding. Up to now, we have investigated DSNNs based on the gated Schottky diode using the max-firing scheme. In the max-firing scheme, neuron which generates the maximum number of firing spikes in the last neuron layer is selected as a result. Therefore, the DSNN system needs counter circuits to add the number of firing spikes. In addition, the system counts the number of firing spikes until the last timestep in the last neuron layer. On the other hand, in the TTFS scheme, the neuron that generates the spike first is selected as a result, which does not need a counter. In addition, when the firing neuron first appears in the DSNN system, then the system does not enter further input data and the firing spike is not generated anymore. Therefore, it can significantly reduce the number of firing spikes and power consumption compared to the max-firing scheme. Figure 8a shows the inference accuracy of a software-based neural network and hardware-based neural network when the max-firing scheme and TTFS schemes are used. As shown in Figure 8a, the inference accuracy of 98.21% is achieved in the TTFS scheme used, which is comparable to the inference accuracy of the max-firing scheme. On the other hand, the TTFS scheme can considerably reduce the number of firing spikes than the max-firing scheme, as shown in Figure 8b.

4. Conclusions

In this work, we investigated the DSNN using the gated Schottky diode as synaptic devices. We proposed an adaptation of a neuron’s threshold to reduce the degradation of inference accuracy due to the conversion from ANNs to SNNs. The LRC method can achieve high inference accuracy faster than the RRC method, while achieving the same final accuracy to the RRC method. The effect of device variation and input noise to the
DSNN system was investigated. The inference accuracy decreases only by 1% when the device variation increases to 0.4. The effect of input noise is more detrimental to the inference accuracy than the effect of device variation. In addition, the resolution of input spikes should be greater than 75 steps to achieve sufficiently high inference accuracy. Finally, the TTFS scheme can considerably reduce the number of firing spikes compared to the max-firing scheme, while achieving comparable inference accuracy to the max-firing scheme. The main limitation of this work is that the proposed scheme is verified by neural network simulation. The proposed scheme in this work can be validated by implementing a neuromorphic chip for deep spiking neural networks. In future work, the neuromorphic chip will be implemented utilizing synaptic devices and neuron devices where the proposed method in this work can be applied and verified.

Author Contributions: Methodology, J.-H.B. and S.-T.L.; formal analysis, J.-H.B. and S.-T.L.; writing—original draft preparation, S.-T.L.; writing—review and editing, J.-H.B. and S.-T.L.; project administration, J.-H.B. and S.-T.L.; funding acquisition, J.-H.B. and S.-T.L. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported in part by the Gachon University research fund of 2021 (GCU-202110160001), and in part by the National Research Foundation (NRF) grant funded by the Korean Government (MSIT) under Grant 2016R1A5A1012966.

Data Availability Statement: Not applicable.

Conflicts of Interest: The authors declare no conflict of interest.

References
1. LeCun, Y. Deep learning. Nature 2015, 521, 436–444. [CrossRef] [PubMed]
2. Szegedy, C. Going deeper with convolutions. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 1–9.
3. Kwon, H.; Kang, S.; Park, W.; Park, J.; Lee, Y. Deep learning based pre-screening method for depression with imagery frontal eeg channels. In Proceedings of the 2019 International Conference on Information and Communication Technology Convergence (ICTC), Jeju, Korea, 16–18 October 2019.
4. Joo, M.; Park, A.; Kim, K.; Son, W.-J.; Lee, H.S.; Lim, G.; Lee, J.; Lee, D.H.; An, J.; Kim, J.H.; et al. A deep learning model for cell growth inhibition IC50 prediction and its application for gastric cancer patients. Int. J. Mol. Sci. 2019, 20, 6276. [CrossRef] [PubMed]
5. Seo, J.W.; Lim, S.H.; Jeong, J.G.; Kim, Y.J.; Kim, K.G.; Jeon, J.Y. A deep learning algorithm for automated measurement of vertebral body compression from X-ray images. Sci. Rep. 2021, 11, 13732. [CrossRef]
6. Schmidhuber, J. Deep Learning in Neural Networks: An Overview. Neural Netw. 2015, 61, 85–117. [CrossRef] [PubMed]
7. Livne, M.; Rieger, J.; Aydin, O.U.; Taha, A.A.; Akay, E.M.; Kossen, T.; Sabesky, J.; Kelleher, J.D.; Hildebrand, K.; Frey, D.; et al. A U-Net Deep Learning Framework for High Performance Vessel Segmentation in Patients with Cerebrovascular Disease. Front. Neurosci. 2019, 13, 97. [CrossRef] [PubMed]
8. Komar, M.; Yakobchuk, P.; Golovko, V.; Dorosh, V.; Sachenko, A. Deep neural network for image recognition based on the Caffe framework. In Proceedings of the 2018 IEEE Second International Conference on Data Stream Mining & Processing (DSMP), Lviv, Ukraine, 21–25 August 2018; pp. 102–106.
9. Li, K. Recurrent neural network language model adaptation for conversational speech recognition. Interspeech 2018, 2018, 3373–3377.
10. Burr, G.W.; Shelby, R.M.; Sebastian, A.; Kim, S.; Kim, S.; Sidler, S.; Virwani, K.; Ishii, M.; Narayanan, P.; Fumarola, A.; et al. Neuromorphic computing using non-volatile memory. Adv. Phys. X 2017, 2, 89–124. [CrossRef]
11. Yu, S. Neuro-Inspired Computing with Emerging Nonvolatile Memories. Proc. IEEE 2018, 106, 260–285. [CrossRef]
12. Lee, S.-T.; Kim, H.; Bae, J.-H.; Yoo, H.; Choi, N.Y.; Kwon, D.; Lim, S.; Park, B.-G.; Lee, J.-H. High-Density and Highly-Reliable Binary Neural Networks Using NAND Flash Memory Cells as Synaptic Devices. In Proceedings of the 2019 IEEE International Electron Devices Meeting (IEDM), San Francisco, CA, USA, 7–11 December 2019; pp. 1–4. [CrossRef]
13. Seongjae, C. Volatile and Nonvolatile Memory Devices for Neuromorphic and Processing-in-memory Applications. J. Semicond. Technol. Sci. 2022, 22, 30–46.
14. Lee, S.-T.; Lee, J.-H. Neuromorphic Computing Using NAND Flash Memory Architecture With Pulse Width Modulation Scheme. Front. Neurosci. 2020, 14, 571292. [CrossRef]
15. Rosselló, J.L.; Canals, V.; Morro, A.; Oliver, A. Hardware Implementation of Stochastic Spiking Neural Networks. Int. J. Neural Syst. 2012, 22, 1250014. [CrossRef] [PubMed]
16. Lee, S.-T.; Lim, S.; Choi, N.Y.; Bae, J.-H.; Kwon, D.; Park, B.-G.; Lee, J.-H. Operation Scheme of Multi-Layer Neural Networks Using NAND Flash Memory as High-Density Synaptic Devices. *IEEE J. Electron Devices Soc.* **2019**, *7*, 1085–1093. [CrossRef]

17. Sagar, S.; Mohanan, K.U.; Cho, S.; Majewski, L.A.; Das, B.C. Emulation of synaptic functions with low voltage organic mem transistor for hardware oriented neuromorphic computing. *Sci. Rep.* **2022**, *12*, 3808. [CrossRef]

18. Mohanan, K.U.; Cho, S.; Park, B.-G. Medium-Temperature-Oxidized GeOx Resistive-Switching Random-Access Memory and Its Applicability in Processing-in-Memory Computing. *Nanoscale Res. Lett.* **2022**, *17*, 63. [CrossRef]

19. Lee, S.T.; Lim, S.; Choi, N.; Bae, J.H.; Kim, C.H.; Lee, S.; Lee, D.H.; Lee, T.; Chung, S.; Lee, J.H.; et al. Neuromorphic technology based on charge storage memory devices. In *Proceedings of the 2018 IEEE Symposium on VLSI Technology*, Honolulu, HI, USA, 18–22 June 2018.

20. Shim, W.; Seo, J.-S.; Yu, S. Two-step write–verify scheme and impact of the read noise in multilevel RRAM-based inference engine. *Semicond. Sci. Technol.* **2020**, *35*, 115026. [CrossRef]

21. Lee, J.H.; Delbruck, T.; Pfeiffer, M. Training Deep Spiking Neural Networks Using Backpropagation. *Front. Neurosci.* **2016**, *10*, 508. [CrossRef]

22. Diehl, P.U.; Neil, D.; Binas, J.; Cook, M.; Liu, S.C.; Pfeiffer, M. Fast-classifying, high-accuracy spiking deep networks through weight and threshold balancing. In *Proceedings of the 2015 International Joint Conference on Neural Networks (IJCNN)*, Killarney, Ireland, 12–17 July 2015; pp. 1–8.

23. Diehl, P.U.; Zarrella, G.; Cassidy, A.; Pedroni, B.U.; Neftci, E. Conversion of artificial recurrent neural networks to spiking neural networks for low-power neuromorphic hardware. In *Proceedings of the 2016 IEEE International Conference on Rebooting Computing (ICRC)*, San Diego, CA, USA, 17–19 October 2016; pp. 1–8. [CrossRef]

24. Yamazaki, K.; Vo-Ho, V.K.; Bulsara, D.; Le, N. Spiking Neural Networks and Their Applications: A Review. *Brain Sci.* **2022**, *12*, 863. [CrossRef]

25. Wu, J.; Chua, Y.; Zhang, M.; Yang, Q.; Li, G.; Li, H. Deep Spiking Neural Network with Spike Count based Learning Rule. In *Proceedings of the 2019 International Joint Conference on Neural Networks (IJCNN)*, Budapest, Hungary, 14–19 July 2019; pp. 1–6. [CrossRef]

26. Wu, J.; Yilmaz, E.; Zhang, M.; Li, H.; Tan, K.C. Deep Spiking Neural Networks for Large Vocabulary Automatic Speech Recognition. *Front. Neurosci.* **2020**, *14*, 199. [CrossRef] [PubMed]

27. Bae, J.-H.; Lim, S.; Park, B.-G.; Lee, J.-H. High-Density and Near-Linear Synaptic Device Based on a Reconfigurable Gated Schottky Diode. *IEEE Electron Device Lett.* **2017**, *38*, 1153–1156. [CrossRef]

28. Lee, S.-T.; Lim, S.; Bae, J.-H.; Kwon, D.; Kim, H.-S.; Park, B.-G.; Lee, J.-H. Pruning for Hardware-Based Deep Spiking Neural Networks Using Gated Schottky Diode as Synaptic Devices. *J. Nanosci. Nanotechnol.* **2020**, *20*, 6603–6608. [CrossRef]

29. Xia, Z.; Kim, D.S.; Jeong, N.; Kim, Y.-G.; Kim, J.-H.; Park, Y.-K.; Chung, C.; Lee, H.; Han, J. Comprehensive modeling of NAND flash memory reliability: Endurance and data retention. In *Proceedings of the 2012 IEEE International Reliability Physics Symposium (IRPS)*, Anaheim, CA, USA, 15–19 April 2012; pp. 5.1–5.4. [CrossRef]

30. Wann, H.; Hu, C. High-endurance ultra-thin tunnel oxide in MONOS device structure for dynamic memory application. *IEEE Electron Device Lett.* **1995**, *16*, 491–493. [CrossRef]

31. Hwang, S.; Kim, H.; Park, J.; Kwon, M.-W.; Baek, M.-H.; Lee, J.-J.; Park, B.-G. System-Level Simulation of Hardware Spiking Neural Network Based on Synaptic Transistors and &F Neuron Circuits. *IEEE Electron Device Lett.* **2018**, *39*, 1441–1444. [CrossRef]

32. Lee, S.-T.; Woo, S.Y.; Lee, J.-H. Low-Power Binary Neuron Circuit with Adjustable Threshold for Binary Neural Networks Using NAND Flash Memory. *IEEE Access* **2020**, *8*, 153334–153340. [CrossRef]

33. Kiselev, M. Rate coding vs. temporal coding—is optimum between? In *Proceedings of the 2016 International Joint Conference on Neural Networks (IJCNN)*, Vancouver, BC, Canada, 24–29 July 2016; pp. 1355–1359.

34. Han, B.; Roy, K. Deep Spiking Neural Network: Energy Efficiency Through Time Based Coding. In *European Conference on Computer Vision*; Springer: Cham, Switzerland, 2020; pp. 388–404. [CrossRef]

35. Park, S.; Kim, S.; Na, B.; Yoon, S. T2FSNN: Deep Spiking Neural Networks with Time-to-first-spike Coding. In *Proceedings of the 2020 57th ACM/IEEE Design Automation Conference (DAC)*, San Francisco, CA, USA, 20–24 July 2020; pp. 1–6. [CrossRef]

36. Göltz, J.; Baumbach, A.; Billaudelle, S.; Kungl, A.F.; Breitwieser, O.; Meier, K.; Schemmel, J.; Kriener, L.; Petrovici, M.A. Fast and deep neuromorphic learning with first-spike coding. In *Proceedings of the Neuro-Inspired Computational Elements Workshop*, Heidelberg, Germany, 17–20 March 2020.