Spiking Neural Networks with Nonidealities from Memristive Silicon Oxide Devices

Viet Cuong Vu*, Anthony Kenyon, Dovydas Joksas, Adnan Mehonic, Daniel J. Mannion, Wing H. Ng
Department of Electronic & Electrical Engineering
University College London (UCL)
London, United Kingdom
*Correspondance Address: viet.vu.14@ucl.ac.uk

***Abstract*—Recent years have seen a rapid surge in the application of artificial neural networks in diverse cognitive settings. The augmented computational demands of these structures have led to an interest in new technologies and paradigms. Of all the artificial neural networks, the spiking neural network (SNN) is notable for its capability to imitate the energy-efficient signalling system in the brain. The memristor presents a promising potential for the integration of SNN into hardware, despite certain non-ideal device properties posing a challenge to its implementation. This study involves the simulation of a SNN model utilizing experimental data on silicon oxide. Particularly, it examines the impact of a non-linear weight update on SNN performance. SNNs were shown to possess tolerance for device non-linearity, while the network can simultaneously maintain a high degree of accuracy. These results provide valuable prior information for future implementation of silicon oxide device-based neuromorphic hardware.**

***Index Terms*—neuromorphic, spiking neural networks, memristive devices, non-linearity**

I. Introduction

Deep Learning (DL) systems have demonstrated exceptional performance in numerous challenging engineering applications [1, 2]. With the rise in system complexity, there is an increased demand for processing capabilities, which are not easily met by resource-constrained processors such as those found in Internet of Things (IoT) edge devices [3]. Memristive In-Memory Computing systems for DL, which conduct computation and storage of recurring operations in the same physical location using advanced memory devices, have the potential to enhance the performance of conventional DL architectures [4]. However, fabricating memristive devices in large quantities is challenging and prohibitively expensive. They are also susceptible to a variety of device non-idealities that must be addressed. Thus, the use of simulation frameworks to model memristive deep learning systems before implementing them at the circuit level is becoming increasingly popular.

Spiking Neural Networks (SNNs) are artificial neural networks (ANNs) that draw on observations from biology, where neurons communicate with each other using spikes transmitted via synapses, with neurons linked by adjustable weight values [6]. It is believed that the energy efficiency of computation in the brain results from the sparsity of low-frequency neuron spikes and the localized approach. However, conventional von Neumann systems are unable to realize the full potential of SNNs' inherent parallelism and asynchrony operations [7]. Neuromorphic hardware, such as memristors or resistive switching memory, is emerging as an efficient synapse block in the construction of future neuromorphic systems. Memristors possess a tunable conductance that directly represents a synaptic weight in biology [8], and a spike signal received from the presynaptic neuron is transferred to the postsynaptic neuron as an electric current or charge proportional to the conductance of the memristor.

In a straightforward crossbar array configuration, the current flowing through all the interconnected synapses is combined in a parallel manner at the post-neuron with remarkable efficiency. Additionally, the memristor has replicated multiple biological phenomena associated with human learning [9]. Despite the advantages of using memristors, some non-ideal effects make implementation in neuromorphic hardware difficult. For instance, variations in device conductance and operation voltage, limited reliability, and non-linear conductance updates can significantly diminish network performance [10]. Additional operation protocols or circuits may be necessary to compensate for these non-idealities [11]. Previous publications have focused on the effects of non-ideal device properties in deep neural networks (DNNs), with fewer studies on SNNs. Some networks were simulated to examine how variations in device properties affect network performance, demonstrating good immunity to device variations in weight updates [12]. This work involves a high-level SNN simulation including a device model to examine the impact of nonlinear conductance updates on SNN performance [13].

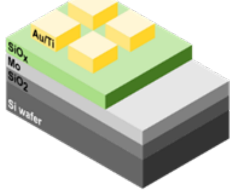

Fig. 1. The device analysed in this article features a metal-insulator-metal structure. The two electrical contacts consist of a molybdenum bottom contact and a gold top contact. To improve the adhesion of the top contact, a 3 nm titanium buffer layer is deposited before the gold deposition [5].

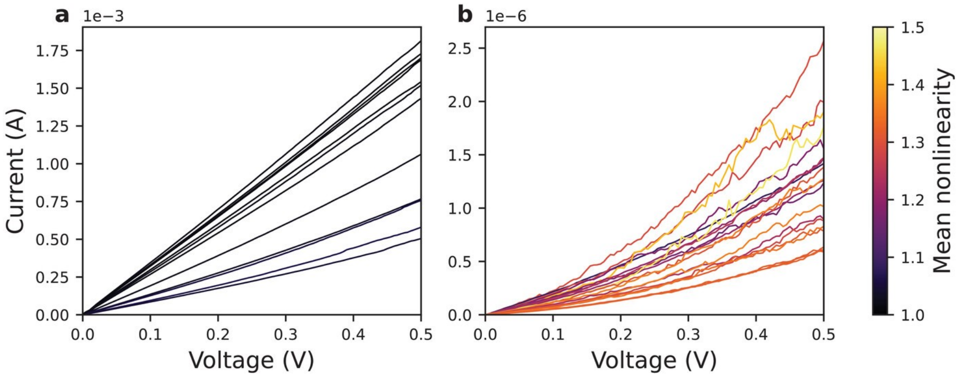

Fig. 2. Experimental data used for this work. I-V sweeps of a SiOx device are presented for two regions: a) low-resistance region with average resistance ranging from 284.6 Ω to 1003 Ω, and b) high-resistance region with average resistance ranging from 366.2 kΩ to 1.295 MΩ. Only the voltage range from 0.0 V to 0.5 V was considered for all curves. The nonlinearity parameter was calculated by dividing the current at 0.5 V by the current at 0.25 V [14].

II. Materials and Method

A. Device Fabrication

The device features a metal-insulator-metal structure, initially designed for binary resistance switching applications, with a bottom metal contact consisting of a 280 nm thick molybdenum film deposited via magnetron sputtering. The insulator layer is a slightly sub-stoichiometric and amorphous silicon oxide, 35 nm thick, deposited via RF magnetron sputtering. The top metal contact is a 115 nm thick gold film deposited via e-beam evaporation through a contact mask, defining the device's active area with a square shape measuring 200 × 200 µm. To enhance adhesion, a 3 nm layer of titanium was deposited before the gold, serving also as a gettering layer to seed the oxide with oxygen vacancies. Electrical characterization, performed using a Keithley 4200A-SCS, involved applying signals to the top electrode (Au) while the bottom electrode (Mo) was grounded. Before stable resistive switching was achieved, the device underwent an initial electroforming step involving a negative voltage sweep, stopped at a current limit of 3 mA. Eighteen voltage sweeps were conducted, ramping from 0.0 V to ±2.5 V and back to 0.0 V using a 3 mA current compliance. Incremental positive sweeps, starting from 0.5 V and increasing by 0.05 V in each run, were applied to achieve a wide range of resistances, repeated until no further resistance change was observed. Experimental data were then used to explore various effects of SiOx technology on SNN.

B. Conventional Framework

Traditionally, spiking neuron models describe neuron properties that generate electrical potentials across their cell membrane. The Leaky Integrate-and-Fire (LIF) model is a widely used spiking neuron model due to its simplicity and computational efficiency [15]. In this model, a neuron emits a spike when its membrane potential reaches a threshold value, then enters a phase of hyperpolarization, preventing a second spike due to a refractory period. Despite the existence of more biologically realistic models, the LIF model remains popular for its balance of simplicity and efficiency.

$$\tau \frac{dU_j(t))}{dt} = \sum_{i=1}^{n} w_{ji} \times n_i(t) - U_j(t) \tag{1}$$

This study conducted a high-level simulation [16]. Input neurons are fully connected to the output neuron via synapses with varying connection strengths. Pre-synaptic spikes generate post-synaptic current based on conductivity and weights, which accumulates at the output neuron nodes, increasing the membrane potential $U(t)$. The LIF model assumes that the potential decays spontaneously with a time constant τ [17]. When input spikes generate a postsynaptic current, $U(t)$ increases, decaying with time constant τ. If $U(t)$ exceeds the threshold, it triggers a post-synaptic spike, resetting $U(t)$ to the resting state for a refractory period.

In conventional ANNs for multi-class classification, the neuron with the highest activation predicts the class. For SNNs, different methods interpret output spikes, such as rate coding and latency coding [18]. This study utilized rate coding, where input is converted into a spike frequency. The goal is for the correct neuron class to emit the most spikes during the simulation run. Training involves weight updates from backpropagation through time (BPTT), resembling spike-timing dependent plasticity (STDP) learning curves [19]. STDP adjusts synaptic weights based on the timing of pre- and postsynaptic spikes [20]. Memristors can implement this learning rule [21], with weight changes modulated by the time difference between spikes [21]. The largest weight changes occur when one neuron reaches the threshold while the other is at the reset voltage, creating a learning window that increases rapidly and then decays slowly[19].

C. Non-idealities

Memristor crossbars are typically modelled as structures that compute linear dot products, with only activation functions introducing nonlinearities [22]. From previous works, the output operation of the synaptic layers is modified to reflect the nonidealities [14]. The proposed method is achieved by calculating the outputs, $y_j \in \mathbf{y} \in \mathbb{R}^{1\times N}$ using the inputs $x_i \in \mathbf{x} \in \mathbb{R}^{1\times M}$, weights $w_{ij} \in \mathbf{w} \in \mathbb{R}^{M\times N}$, a

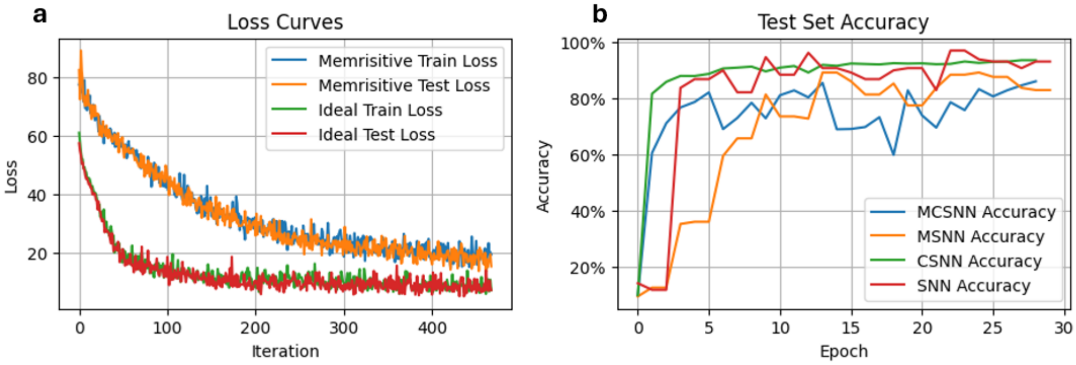

Fig. 3. Training results for standard and memristive schemes when exposed to I-V nonlinearities. a) The loss curves are noisy due to the tracking of losses at every iteration, rather than averaging across multiple iterations. b) Comparison of different spiking neural network performance when implemented with I-V nonlinearities from a single training run. The networks successfully learned MNIST features during training in both models.

nonlinear activation function f, the memristor's non-ohmic IV behaviour function g, as shown in equation (2).

$$y_j = f\left(\sum_{i=1}^{M} g\left(x_i, w_{ij}\right)\right) \tag{2}$$

During inference, the variables x, w, and y are mapped onto voltages, conductances, and currents, respectively (3), using the scaling factors k_v and k_I, where k_G is the conductance scaling factor and k_V is determined before training.

$$V = k_v x \tag{3}$$

$$y = \frac{I}{k_I} = \frac{I}{k_V k_G} \tag{4}$$

$$k_G = \frac{G_{on} - G_{off}}{max|w|} \tag{5}$$

To represent both positive and negative weights, pairs of conductances $G+$ and $G-$ are respectively added to the 'positive' and 'negative' bit lines of crossbar arrays (6), which can be advantageous for mitigating the effects of stuck devices [23]. The output currents of the negative bit lines are then subtracted from the output currents of the positive bit lines, which is known as the differential pair architecture [24].

$$G\pm = G_{avg} \pm \frac{k_G w}{2} \tag{6}$$

$$G_{avg} = \frac{G_{off} + G_{on}}{2} \tag{7}$$

In the context of memristive applications, it is commonly accepted that perfectly ohmic devices are those for which the ratio of current (I) to voltage (V) remains constant. Deviations from this behavior in memristors are typically characterized by considering a pair of points on the I-V curve [25]. For instance, it may be possible to define a nonlinearity parameter γ [26] that were extracted from experimental data of a SiOx RRAM device "Fig. 2" as expressed in equation (8), where V_{ref} is introduced at $0.25V$ and G is the conductance parameter.

$$\gamma \equiv \frac{f(2V_{ref})}{f(V_{ref})}; I = f(V) \tag{8}$$

$$I \equiv V_{ref} G \left(\frac{V}{V_{ref}}\right)^{log_2 \gamma} \tag{9}$$

Silicon oxide devices are capable of resistance switching, as previously studied [5]. To analyze I-V nonlinearity and achieve a wide range of resistance states, incremental positive sweeps were used to gradually reset the device from the low-resistance state (LRS) to the high-resistance state (HRS).

Furthermore, the potential for variability in device-to-device (D2D) communication due to programming inaccuracies was also considered. During the mapping of conductances, it is possible that the values may differ from those intended. In some memristors, these resistive deviations are modelled using a lognormal distribution. This can be incorporated into the training process by introducing random variations in the values in each iteration. In this case, the values are drawn from a random lognormal distribution. For the purposes of lognormal modelling, the standard deviation of the natural logarithm of resistances was linearly interpolated from a set of values of $0.25R_{off}$ and $0.25R_{on}$, representing uniform behaviours across different device.

III. Results and Discussion

A. Training Setup

A two-layer memrisitive SNN (MSNN) was designed to examine the impact of non-linear device properties on the MNIST handwritten dataset converted to a spike train. The network was fed from the 28x28 image and trained on 60,000 samples with 10,000 samples used for testing. The network learns representative features in the input samples through updating the synaptic weights. The weight conductance values were initially generated using a uniform random distribution. To evaluate the performance of non-ideality-aware training on complex tasks, memrisitive convolutional spiking neural networks (MCSNN) were trained with the assumption that their convolutional layers would be implemented digitally.

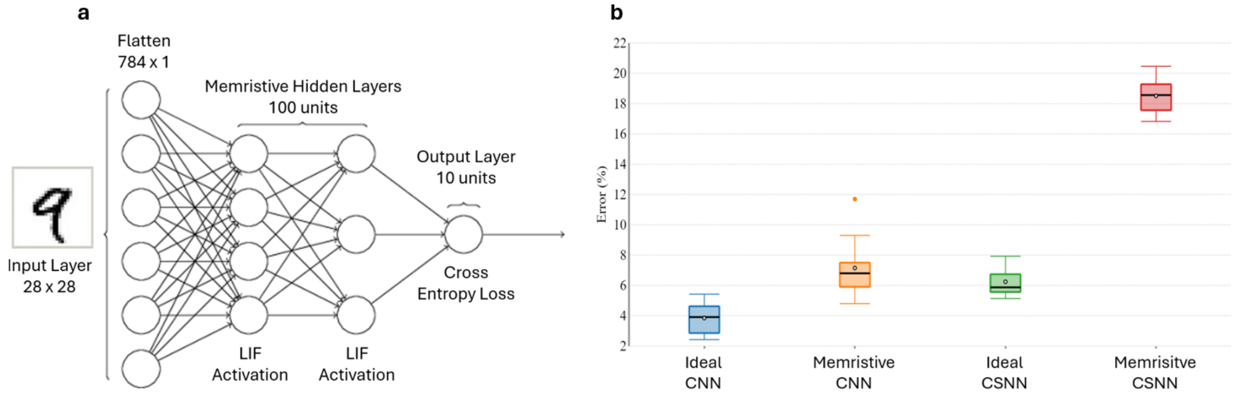

Fig. 4. Description of structure and operation of memristive SNN. a) Abstract SNN representation for MNIST pattern recognition with input, hidden memristive layers, and output layer. b) Impact of memrisitve nonidealities on performance of standard and spiking CNN models under the same training scheme.

Their fully connected layers were trained using memristive crossbar arrays that suffer from high I-V nonlinearity.

The MSNN model in “Fig. 4a” is fully connected and consists of a layer with 100 hidden neurons and LIF activation function, followed by another layer before the output layer with 10 neurons. The MSCNN architecture includes a convolutional layer with 12 output filters, a 5 × 5 kernel size, and an LIF activation function. This is followed by a pooling layer with a 2 × 2 pool size. The architecture then includes another convolutional layer with 64 output filters, a 5 × 5 kernel size, and an LIF activation function, followed by another pooling layer with a 2 × 2 pool size. Finally, the architecture includes a memristive fully connected layer with 10 output neurons and an LIF activation function. The cross entropy loss function from Torch [16] automatically generates a loss at the output and handles the softmax of the output layer.

B. Learning and Classification

The training results are presented to explore the effect of high I-V nonlinearity. Training curves for MSNNs with the MNIST dataset and exposed to I-V nonlinearities are shown in “Fig. 3a”. The blue training curve, which assumes the presence of memristive non-ideal effects, closely follows the orange test curve, with both reaching loss values of approximately 20. This plot provides SNN and MSNN curves to aid in comprehending the distinctions between ideal and non-ideality-aware training. It is observed that the ideal curves (red/green) are distinct from the memristive curves (blue/orange) with lower loss values of around 10. Moreover, the global minimum of the ideal curves is reached early in the training, at 100 iterations, compared to the memristive training, which reaches it at 400 iterations. “Fig. 3b” presents accuracy comparisons between SNN and CSNN models for a single run. The ideal SNN (red) and CSNN (green) models both quickly converged with accuracies above 90%. The non-ideality-aware models, MCSNN (blue) and MSNN (orange), achieved accuracies above 85%, with MSNN taking longer to learn the convolutional variant after 10 epochs. To address the high variability of non-idealities, which were nondeterministic, five networks were trained for each configuration. The results are summarised in Table I. Under the previous training setup, additional architectural comparison between standard and spiking CNN’s are provided in “Fig. 4b”, showing non-ideality effects from memristors have degraded the performance of both implementations. In summary, when the SNNs are configured with non-ideality-aware training, their performance degrades to reflect the physical implementation with memristive silicon oxide. However, these networks still retain a high degree of accuracy.

TABLE I
MODEL ACCURACY FOR DIFFERENT SNN CONFIGURATIONS

Model Accuracy	Ideal implementations	Memristive non-idealities
Fully-Connected	$91.73 \pm 2.97\%$	$79.54 \pm 3.86\%$
Convolutional	$93.42 \pm 1.63\%$	$81.21 \pm 3.19\%$

C. Limitations and Future Works

Accurately modeling non-idealities for ex-situ training of memristors is a significant challenge due to the variability among devices. For example, the device-to-device variability of silicon oxide memristors means that the behavior of any individual device may not be perfectly representative of others. In real-world scenarios, various non-idealities may be encountered, and if training only accounts for I-V nonlinearities, memristive spiking neural networks (MSNNs) may still encounter stuck devices when deployed. To improve the performance of SNNs trained with silicon oxide non-ideality, it is necessary to include the memristor current transient by implementing homeostasis properties within artificial synapses to regulate excessive firing, similar to biological systems. This empirical model, influenced by the physical model of the device, requires assessment of the reliability of MSNNs employing non-ideality-aware training. Additional research is needed to validate the physical implementations of these configurations and construct more comprehensive models.

IV. CONCLUSION

The aim of this work is to develop a new training scheme that accounts for memristive non-ideality in order to accurately reflect the performance of SNNs. The networks consist of silicon oxide memristive layer with LIF, and able to achieve accuracy over 85% for MNIST dataset. The study emphasizes the importance of accounting for non-ideal device behavior during training, which was previously not explored for memrisitive silicon oxide devices. Neglecting to do so may lead to unreliable training performance as an indicator of network performance during inference. The study shows how experimental data and training techniques can address I-V nonlinearity, which has not been previously addressed during exsitu SNN training. The method presented here highlights the important first step of better modelling and accounting for silicon oxide nonidealities. This work therefore provides the foundation for nonidealities-aware implementations of SNN, especially using emerging silicon oxide devices.

Acknowledgment

VCV acknowledges the EPSRC for funding via project [2576207].

References

[1] Lijuan Liu, Yanping Wang, and Wanle Chi. "Image recognition technology based on machine learning". In: *IEEE Access* (2020).

[2] Young Hoon Jung et al. "Flexible piezoelectric acoustic sensors and machine learning for speech processing". In: *Advanced Materials* 32.35 (2020), p. 1904020.

[3] Mehmet Demirci. "A survey of machine learning applications for energy-efficient resource management in cloud computing environments". In: *2015 IEEE 14th international conference on machine learning and applications (ICMLA)*. IEEE. 2015, pp. 1185–1190.

[4] Adnan Mehonic and Dovydas Joksas. "Emerging Nonvolatile Memories for Machine Learning". In: *arXiv preprint arXiv:2308.03659* (2023).

[5] Adnan Mehonic et al. "Intrinsic resistance switching in amorphous silicon oxide for high performance SiOx ReRAM devices". In: *Microelectronic Engineering* 178 (2017), pp. 98–103.

[6] Samanwoy Ghosh-Dastidar and Hojjat Adeli. "Spiking neural networks". In: *International journal of neural systems* 19.04 (2009), pp. 295–308.

[7] Doo Seok Jeong et al. "Memristors for energy-efficient new computing paradigms". In: *Advanced Electronic Materials* 2.9 (2016), p. 1600090.

[8] Ting Chang, Yuchao Yang, and Wei Lu. "Building neuromorphic circuits with memristive devices". In: *IEEE Circuits and Systems Magazine* 13.2 (2013), pp. 56–73.

[9] Peng Yao et al. "Fully hardware-implemented memristor convolutional neural network". In: *Nature* 577.7792 (2020), pp. 641–646.

[10] Jacopo Frascaroli et al. "Evidence of soft bound behaviour in analogue memristive devices for neuromorphic computing". In: *Scientific reports* 8.1 (2018), p. 7178.

[11] Can Li et al. "Efficient and self-adaptive in-situ learning in multilayer memristor neural networks". In: *Nature communications* 9.1 (2018), p. 2385.

[12] Sung Yun Woo et al. "Implementation of homeostasis functionality in neuron circuit using double-gate device for spiking neural network". In: *Solid-State Electronics* 165 (2020), p. 107741.

[13] Sridhar Chandrasekaran et al. "Improving linearity by introducing Al in HfO2 as a memristor synapse device". In: *Nanotechnology* 30.44 (2019), p. 445205.

[14] Dovydas Joksas et al. "Nonideality-Aware Training for Accurate and Robust Low-Power Memristive Neural Networks". In: *Advanced Science* 9.17 (2022), p. 2105784.

[15] Wulfram Gerstner and Werner M Kistler. *Spiking neuron models: Single neurons, populations, plasticity*. Cambridge university press, 2002.

[16] Jason K Eshraghian et al. "Training spiking neural networks using lessons from deep learning". In: *Proceedings of the IEEE* (2023).

[17] Nicolas Brunel and Simone Sergi. "Firing frequency of leaky intergrate-and-fire neurons with synaptic current dynamics". In: *Journal of theoretical Biology* 195.1 (1998), pp. 87–95.

[18] Filip Ponulak and Andrzej Kasinski. "Introduction to spiking neural networks: Information processing, learning and applications." In: *Acta neurobiologiae experimentalis* 71.4 (2011), pp. 409–433.

[19] Guo-qiang Bi and Mu-ming Poo. "Synaptic modifications in cultured hippocampal neurons: dependence on spike timing, synaptic strength, and postsynaptic cell type". In: *Journal of neuroscience* 18.24 (1998), pp. 10464–10472.

[20] Daniel E Feldman. "The spike-timing dependence of plasticity". In: *Neuron* 75.4 (2012), pp. 556–571.

[21] Gabriel Maranhão and Janaina Gonçalves Guimarães. "Low-power hybrid memristor-CMOS spiking neuromorphic STDP learning system". In: *IET Circuits, Devices & Systems* 15.3 (2021), pp. 237–250.

[22] J Joshua Yang, Dmitri B Strukov, and Duncan R Stewart. "Memristive devices for computing". In: *Nature nanotechnology* 8.1 (2013), pp. 13–24.

[23] Can Li et al. "Analogue signal and image processing with large memristor crossbars". In: *Nature electronics* 1.1 (2018), pp. 52–59.

[24] D Ielmini, Z Wang, and Y Liu. "Brain-inspired computing via memory device physics". In: *APL Materials* 9.5 (2021).

[25] Changhyuck Sung et al. "Effect of conductance linearity and multi-level cell characteristics of TaOx-based synapse device on pattern recognition accuracy of neuromorphic system". In: *Nanotechnology* 29.11 (2018), p. 115203.

[26] Florian Lentz et al. "Current Compliance-Dependent Nonlinearity in TiO ReRAM". In: *IEEE electron device letters* 34.8 (2013), pp. 996–998.