Design of Variation-Tolerant 1F-1T Memory Array for Neuromorphic Computing

Masud S K Rana 1, Sunanda Thunder 1, Franz Müller 1, Nellie Laleni 1, Yannick Raffel 1, Maximilian Lederer 1, Luca Pirro 1, Talha Chohan 1, Jing-Hua Hsuen 1, Tian-Li Wu 1, Konrad Seidel 1, Thomas Kaempfe 1, Sourav De 1, and Bhaswar Chakrabarti 2

1Affiliation not available
2Indian Institute of Technology Madras

October 30, 2023

Abstract

This letter proposes a memory cell, denoted by 1F-1T, consisting of a ferroelectric field-effect transistor (Fe-FET) cascoded with another current-limiting transistor (T). The transistor reduces the impact of drain current (Id) variations by limiting the on-state current in FeFET, denoted by 1F. We have fabricated 28nm high-k-meta-gate (HKMG) based FeFETs, and the experimental data is used to model and simulate single-cell and memory arrays. The simulation shows significant improvement in bit-line (BL) current (IBL) variation for 1F-1T memory array. Finally, the system-level neuromorphic simulation with 1F-1T synapses shows an inference accuracy of 97.6% for MNIST hand-written digits with multi-layer perceptron (MLP) neural networks.
Design of Variation-Tolerant 1F-1T Memory Array for Neuromorphic Computing

Masud Rana Sk, Sunanda Thunder, Franz Müller, Nellie Laleni, Yannick Raffel, Maximilian Lederer, Luca Pirro, Talha Chohan, Jing-Hua Hsuen, Tian-Li Wu, Konrad Seidel, Thomas Kämpfe, Sourav De, Bhaswar Chakrabarti

Abstract—This letter proposes a memory cell, denoted by 1F-1T, consisting of a ferroelectric field-effect transistor (FeFET) cascoded with another current-limiting transistor (T). The transistor reduces the impact of drain current ($I_D$) variations by limiting the on-state current in FeFET, denoted by $I_F$. We have fabricated 28nm high-k-meta-gate (HKMG) based FeFETs, and the experimental data is used to model and simulate single-cell and memory arrays. The simulation shows significant improvement in bit-line (BL) current ($I_{BL}$) variation for 1F-1T memory array. Finally, the system-level neuromorphic simulation with 1F-1T synapses shows an inference accuracy of 97.6% for MNIST hand-written digits with multi-layer perceptron (MLP) neural networks.

Index Terms—FeFET, 1F-1T, Neuromorphic, CIM, HfO$_2$.

I. INTRODUCTION

DEEP neural networks (DNN) play a significant role in performing many data-intensive computing tasks such as speech recognition, motion detection, computer vision, and natural language processing. Training DNNs with enormous amounts of data, generated from the internet and real-time devices lead to high energy and latency costs. Recently, in-memory computing (IMC) with emerging non-volatile memory (eNVM) technologies are being researched to alleviate this issue [1]–[4]. The basic step of IMC is vector-matrix multiplication (VMM). Many eNVMs such as resistive random access memory (ReRAM) [5]–[7], phase change memory (PCM) [8], [9], and FeFETs [10]–[13] have been investigated in recent years.

Amidst such a plethora of eNVMs, FeFET is a promising one due to field-based operation, low power consumption, fast switching, high on-current ($I_{ON}$) to off-current ($I_{OFF}$) ratio ($I_{ON}/I_{OFF}$), excellent linearity and bidirectional programmability, good endurance, and compatibility with the Complementary Metal-Oxide Semiconductor (CMOS) technology [14]–[19]. However, the primary obstacle in implementing FeFET-based computing systems lies in the inherent stochasticity of FeFET devices. One of the primary reasons for this stochasticity lies in the polycrystalline nature of HfO$_2$-based ferroelectric thin films. The presence of charge traps at the Ferroelectric/interlayer interface and within the Ferroelectric film can lead to asymmetrical conductive response and large device-to-device variations [20]–[22]. Abundant efforts have been made to minimize the effects of such non-idealities both from the devices and from the circuits perspective [23]–[28]. It is imperative that device-to-device variations in the drain currents of a FeFET can adversely affect the performance of a synaptic core, resulting in significant degradations in training and inference accuracy. Our previous work demonstrated how a series-resistor connected to the drain terminal of the FeFETs could reduce variations by limiting the current in the Low-Voltage-Threshold (LVT) state [29]. However, the fabrication of resistors in a standard CMOS process adds complexity to the macro-design because of the enormous size of such resistors. Poly-silicon resistors have the highest resistance density, but it suffers from a larger mismatch, requiring 100 $\mu$m to achieve 1Mohm.

In this work, we demonstrate the efficacy of 1F-1T structure in overcoming this issue. We have considered 4×4 arrays to evaluate the effects of this 1F-1T architecture. We show that such a configuration prevents the accumulation of variations over bit-line current ($I_{BL}$) and also reduces the impact of voltage swing across word lines (WL), bit lines (BL), and select lines (SL). We benchmark the system-level performance of a neuromorphic circuit with a 1F-1T synaptic array. Our simulations indicate significant improvements in inference accuracy compared to the network using 1F synapses.

II. EXPERIMENTS

A. Characterization of 28nm HKMG FeFET

![Fig. (1) (a) Schematic and (b) TEM image of a 28nm HKMG-based FeFET fabricated at GlobalFoundries. [12], [13]](image)

The experiment began with fabricating the crossbar arrays and memory cell structures on 300mm wafers using the 28nm HKMG technology at GlobalFoundries. Figure 1a shows the schematic of a HKMG FeFET, and Figure 1b shows the transmission electron microscopic (TEM) image. On the application of voltage pulses to the gate terminal, the dipoles of the
ferroelectric layer align themselves according to the polarity of the pulse, affecting the surface charge density of the channel and, correspondingly, the threshold voltage ($V_{th}$). A positive pulse at the gate terminal of n-type FeFET programs the device at a lower $V_{th}$ state (LVT), and the negative pulse programs the device at a higher $V_{th}$ state (HVT). The devices were programmed and erased by applying 500ns pulses of amplitude 4.5V and -5V, respectively, to the gate terminal of FeFET, with the drain and source terminals grounded. Before READ-WRITE operations, the devices were subjected to wake-up cycling (100 times) by 4.5V and -5V pulses of 500ns.

B. Design of 28nm HKMG Based Memory Array

In this work, we have considered 4×4 memory arrays to evaluate the effects of variability in the FeFETs. Figure.2 demonstrates the two different architectures we investigate. Note that program and erase conditions for the single devices were adopted for simulating the characteristics of the arrays. Figure.2a shows the architecture with one FeFET device per synaptic cell (1F architecture). Here, the word lines (WL) are connected to the gate terminals of the FeFETs in a row. Each cell can be accessed by controlling the corresponding WL and the select line (SL). Bit-line (BL) current ($I_{BL}$) defines the state of the cell. Our proposed schematic of the 1F-1T synaptic array is shown in Figure.2b. Here, each synaptic cell consists of an extra transistor which acts as a current limiter. To access a particular cell, a bias voltage is applied to the gate terminal of the transistor. FeFET conductance is controlled by tuning the WL and SL voltages. Table (1) summarizes the operating voltages for simulating the 1F and 1F-1T arrays. The row-wise write operation was conducted by applying 4.5V at the word line (WL) and 1.8V at the Bias line. The cell-wise read operations were conducted by biasing BL to 0.1V, WL at 0-1.5V, and Bias line at 0.3V. While writing/reading operations are performed on a particular row, all other rows are inhibited by applying -0.3V to WLs and 0V to the Bias lines. Similarly, columns are disabled by controlling the SL voltages so that the total gate-to-source ($V_{GS}$) are below the threshold voltages of the FeFETs. The operating voltages are given in Table(1) for the write, read, and inhibition operations in the arrays.

<table>
<thead>
<tr>
<th></th>
<th>WL</th>
<th>BL</th>
<th>SL</th>
<th>Bias</th>
</tr>
</thead>
<tbody>
<tr>
<td>Write</td>
<td>4.5V</td>
<td>0V</td>
<td>0V</td>
<td>1.8V</td>
</tr>
<tr>
<td>Read</td>
<td>0-1.5V</td>
<td>0.1V</td>
<td>0V</td>
<td>0.3V</td>
</tr>
<tr>
<td>WL/Bias Inhibit</td>
<td>-0.3V</td>
<td>x</td>
<td>x</td>
<td>0V</td>
</tr>
</tbody>
</table>

C. Neural Network Simulation

Finally, the impact of variations on the system-level performance, especially for neuromorphic applications, was evaluated by the Neurosim platform [30]. Experimentally calibrated $I_d$ values with the statistics of variations were used to simulate a multilevel perception (MLP) neural network (NN) for the MNIST dataset. The neural network’s architecture is illustrated in Figure.3. In this work, offline training has been considered for neural networks, where synaptic weight coefficients are pre-trained in software without considering hardware-related variation. Then the weights are encoded into the circuitry. This method is more energy efficient but less noise-tolerant as the synaptic weight cannot be modified during the training process. After offline training, the synaptic weights, in terms of channel conductance values of FeFETs, were updated on the hardware using a single-shot programming pulse. The synaptic weights were normalized between the minimum value ($W_{min}$) of -1 and the maximum value ($W_{max}$) of 1. The $I_{OFF}$ of the FeFET was mapped to $W_{min}$, and the $I_{ON}$ was mapped to $W_{max}$. The FeFET-based synaptic core shown in Figure.3, is used to carry out the VMM operation. The output of the VMM is directly digitized using a current-to-digital converter [30].

III. RESULTS AND DISCUSSION

Figure.4a shows the READ operation conducted 2 seconds after WRITE by applying a voltage ramp with a step size of 100mV. The READ current’s probability density function (PDF) is shown in Figure.4b. We observe that device-to-device variation is more at the LVT state than the HVT state. The experimentally measured READ-WRITE of a single FeFET
Fig. (4) (a) Transfer characteristics of fabricated devices after write operation with 500ns pulses. (b) The PDF of the $I_d$ for HVT and LVT states at 0.7V read voltage. (c) The bit line ($I_{BL}$) current vs. word line voltage ($V_{WL}$) simulation in-memory array. (d) The PDFs of the $I_{BL}$ show overlapping currents for eight-bit lines, which makes the quantization process difficult.

was calibrated with the model [31]–[33] for the FeFETs to evaluate the characteristics of a memory array. The mean and standard deviation of $I_d$ for HVT and LVT states were used for statistical variation modelling of the devices. Figure.4c shows the current ensemble through bit line ($I_{BL}$) or the current through eight different columns of arrays concerning voltage applied at the word line ($V_{WL}$). The PDFs of $I_{BL}$ show that as the number of transistors increases in the column, the distribution curves start to overlap, making the multiply-and-accumulate (MAC) operation futile for neuromorphic applications as shown in Figure.4d. Next, we propose an alternative synaptic cell consisting of a ferroelectric memory transistor cascaded with another logic transistor (1F-1T) to circumvent this issue. Since the variations mostly affect the LVT state, we can easily address the issue by limiting the ON current. The logic transistor acts as a current limiter like the resistive element demonstrated in our previous works [13]. Figure.5a shows the transfer characteristics of the proposed devices with the same pulsing scheme as a single FeFET cell. It is observed that the ratio of drain currents at HVT and LVT is low compared to the 1F cell. The 1F-1T structure was modelled by adopting the FeFET model mentioned above in our previous work along with the BSIM-4 model [31], [33], [34].

The variation statistics obtained from experimental data were used to evaluate the impact of the current limiting transistor on the variation of $I_{BL}$ through Monte Carlo simulation. The variation for the logic transistor cells was not considered as they are insignificant compared to the FeFET devices. After device modelling, the simulation and modelling of the 4×4 array with 1F-1T memory devices were conducted. Figure.5b shows non-overlapping MAC operation among eight columns. Therefore, integrating the current limiter transistor with each FeFET device reduces the variation significantly in the $I_{BL}$. The PDF of the MAC operation with 1F-1T structures in Figure.5c shows non-overlapping distinguishable PDFs of $I_{BL}$ among eight columns with four cells in each column.

Fig. (5) (a) Transfer characteristics of the 1F-1T synaptic cell after write operation with 500ns pulses (b) Simulation of multiply-and-accumulate (MAC) operation on 1F-1T array shows non-overlapping $I_{BL}$ characteristics. (c) The PDF of improved $I_{BL}$ characteristics for eight different columns.

Fig. (6) (a) Inference accuracy and (b) READ energy obtained via system level simulation of neural networks corroborates the superiority of 1F-1T synapses over 1F synapses.

Figure.6a shows the inference accuracy achieved for offline training. 1F-1T exhibit a clear accuracy advantage compared to 1F-based synapses. It achieves 97.6% accuracy, whereas the software benchmark is 98.99%. Further, it is observed that total leakage power is almost the same for both subarrays. But total read energy is less for the 1F-1T-based synaptic core than the 1F-based core due to the lower ON current of FeFET. The total read energy of 1F and 1F-1T based cores are 450 $\mu$J and 17.6 $\mu$J, respectively shown in Figure.6b. So, the MLP-based NN with our proposed 1F-1T synaptic core offers excellent immunity towards device variations and offers excellent accuracy and low energy consumption.

IV. CONCLUSION

We have proposed a novel 1F-1T memory cell for neuromorphic applications in this work. The operations are evaluated by characterizing the 28nm HKMG FeFET arrays. The cascoded 1T cell acts as a current limiter and reduces the variation in channel current compared to 1FeFET cell-based memory array. We have also demonstrated the impact of device variation on system-level performance in training an MLP-NN. The 1F-1T array-level simulation depicts that such an analog weight
cell-based pseudo-crossbar array accelerates the system-level performance for building neuromorphic hardware systems.

REFERENCES


