Drain-Erase Scheme in Ferroelectric Field Effect Transistor—Part II: 3-D-NAND Architecture for In-Memory Computing

Panni Wang, *Student Member, IEEE*, Wonbo Shim, Zheng Wang, *Student Member, IEEE*, Jae Hur, Suman Datta, *Fellow, IEEE*, Asif Islam Khan, *Member, IEEE*, and Shimeng Yu, *Senior Member, IEEE*

***Abstract*—Ferroelectric-doped HfO_2-based ferroelectric field-effect transistors (FeFETs) are being actively explored as emerging nonvolatile memory (NVM) devices with the potential for in-memory computing. In this two-part article, we explore the feasibility of the FeFET-based 3-D NAND architecture for both *in situ* training and inference. To address the challenge of erase-by-block in a NAND-like structure, we propose and experimentally demonstrate the drain-erase scheme to enable the individual cell's program/erase/inhibition, which is necessary for individual weight updates in *in situ* training. We described the device characterization of different drain-erase conditions and results in Part I. The array-level design for this drain-erase scheme for both AND-type and NAND-type array is addressed in this Part II. A 3-D vertical channel FeFET array architecture is proposed to accelerate the vector-matrix multiplication (VMM). 3-D timing sequence of the weight update rule is designed and verified through the 3-D-array-level SPICE simulation. Finally, the VMM operation is simulated in a 3-D NAND-like FeFET array.**

***Index Terms*—3-D-NAND, drain-erase, ferroelectric transistor, *in situ* training, vector-matrix multiplication (VMM).**

I. Introduction

DEEPneural networks (DNNs) have made remarkable improvements in intelligent tasks such as image and speech recognition. However, the energy efficiency of DNNs is highly limited by moving the data back and forth between the memory and the processor in von Neumann-based hardware. To overcome this bottleneck, in-memory computing, in which computation is done at the location of the data storage, has been proposed to accelerate the computation. To this end, static random-access memory (SRAM) [1] and emerging nonvolatile memories (NVMs) such as pulse-code modulation (PCM) [2], [3] and resistive random-access memory (RRAM) [4]–[7] have been explored for both *in situ* training and inference. However, these embedded memory technologies typically have megabytes-level capacity, which is insufficient to hold gigabytes-level weights of large-scale DNNs. Alternatively, there are approaches using charge-trap-transistor [8], 2-D NOR Flash [9], 2-D NAND Flash [10], or 3-D AND Flash [11] to implement DNNs leveraging their high density. However, due to the high write voltage and long write latency, Flash-based solutions are only applicable for inference instead of *in situ* training where the weights are frequently updated.

The ferroelectric field-effect transistor (FeFET) is recently proposed as a promising candidate as a multilevel synaptic device for *in situ* training on-chip [12], [13]. The doped HfO_2-based FeFET operates in a similar fashion as Flash with tunable threshold voltage (V_{th}), but its lower write voltage (~3 V) and shorter write latency (~50 ns) [14]–[16] overcome the aforementioned shortcomings of Flash. A four-layer 3-D vertical channel FeFET prototype has been experimentally demonstrated [17]. However, one grand challenge that remains is to use FeFET for *in situ* training, which is the block-erase nature of the NAND array. The weight update rule in DNNs requires each weight's conductance to be independently increased or decreased. This means the conventional substrate-erase scheme is not applicable, as it will erase the entire block.

In this article, we proposed the drain-erase scheme to erase the cell by raising the channel voltage through the drain side. To enable the individual cell program/erase, we experimentally demonstrated the feasibility of the drain-erase scheme, i.e., flipping the ferroelectric domain polarity by applying V_{erase} to the drain to increase the channel potential while grounding the gate, on 22-nm fully depleted silicon-on-insulator (FDSOI) FeFET [15] and 28-nm high-k metal gate (HKMG) FeFET [16] from GLOBALFOUNDRIES in the Part-I article [21].

In this Part-II article, we will focus on the array-level design for the drain-erase scheme. For simplicity, the individual cell operation on a 2-D FeFET array is discussed first, as the proposed 2-D drain-erase scheme could be extended to 3-D

Manuscript received December 22, 2019; accepted January 20, 2020. Date of publication February 14, 2020; date of current version February 26, 2020. This work was supported in part by the NSF under Grant 1810005 and in part by Applications and Systems Driven Center for Energy-Efficient Integrated Nanotechnologies (ASCENT), one of the six Semiconductor Research Corporation (SRC)/Defense Advanced Research Projects Agency (DARPA) Joint University Microelectronics Program (JUMP) Centers. The review of this article was arranged by Editor J. Kang. *(Panni Wang and Wonbo Shim contributed equally to this work.) (Corresponding authors: Asif Islam Khan; Shimeng Yu.)*

Panni Wang, Wonbo Shim, Zheng Wang, Jae Hur, Asif Islam Khan, and Shimeng Yu are with the School of Electrical and Computer Engineering, Georgia Institute of Technology, Atlanta, GA 30326 USA (e-mail: asif.khan@ece.gatech.edu; shimeng.yu@ece.gatech.edu).

Suman Datta is with the Department of Electrical Engineering, University of Notre Dame, Notre Dame, IN 46556 USA (e-mail: sdatta@nd.edu).

Color versions of one or more of the figures in this article are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TED.2020.2969383

with carefully designed timing sequence. The biasing scheme of the 2-D AND and NAND array are both designed to show the individual cell's erase/program with the drain-erase scheme. Then we proposed a 3-D NAND-like FeFET array architecture feasible for both *in situ* training and inference.

II. FeFET Operation Modes

As experimentally demonstrated in the Part-I article [21], there are four different modes for the biasing scheme during the write operation in the FeFET array: drain erase mode, gate program mode, erase-inhibition, and program-inhibition mode. With the characterized conditions in the Part-I article [21], when the FeFET's drain is biased at 3 V, gate/body is grounded, and source is biased at 1.5 V, it could be successfully erased. This is because the drain side voltage and the gate side ground will introduce the gate-induced-drain-leakage (GIDL) effect that occurs when V_{DG} potential and band bending are high enough to generate the electron/holes pairs by valence band to conduct band tunneling. Holes are easier to accumulate in the confined geometry, such as FDSOI or gate-all-around (GAA) vertical channel transistor, thus increasing the surface potential of the channel. The program mode of the cell is done by the gate program, in which the gate is biased at a high voltage (~3 V) and the source/drain/body is grounded. During both program/erase operations, unselected cells will be in the inhibition mode. For the erase-inhibition mode, the potential of the drain side is higher than that of the gate side, but the difference is small, so that it will not be erased. For the program-inhibition mode, the FeFET receives high voltage on its gate, but the drain side voltage is increased so that the voltage difference between the gate and the channel is not large enough to program the cell. Such individual cell operation conditions were characterized in the Part-I article [21]; we will extend further to the analysis of the FeFET array writing scheme in this article.

III. Individual Cell's Erasure/Programming in 2-D FeFET Array

When designing the memory array, all the disturbance issues need to be taken into consideration. The individual cell's program and the erase scheme for the AND array is shown in Fig. 1. The selected cell is cell (1, 1). WL$_1$ is biased at 3 V while its source and drain are grounded by applying ground to the BL$_1$ and SL$_1$. Since cell (1, 2) receives high voltage on its gate, it needs to raise the channel potential by increasing the BL$_2$ and SL$_2$ to 1.5 V. A 1.5-V voltage on cell (2, 2)'s source would be safe without erasing the cell. Therefore, only cell (1, 1) would be programed. The individual cell's erase scheme is shown in Fig. 1(b) for the AND array. Cell (1, 1) is the selected cell to be erased with 3 V on its drain and 0 V on its gate. Cell (2, 1) receives high voltage on its drain. Therefore, WL$_2$ needs to raise to 1.5 V to inhibit it from being erased. For all the other neighboring cells, their gate-to-channel voltage difference would not be enough to flip the polarity of the ferroelectric film. Therefore, only cell (1, 1) would be erased.

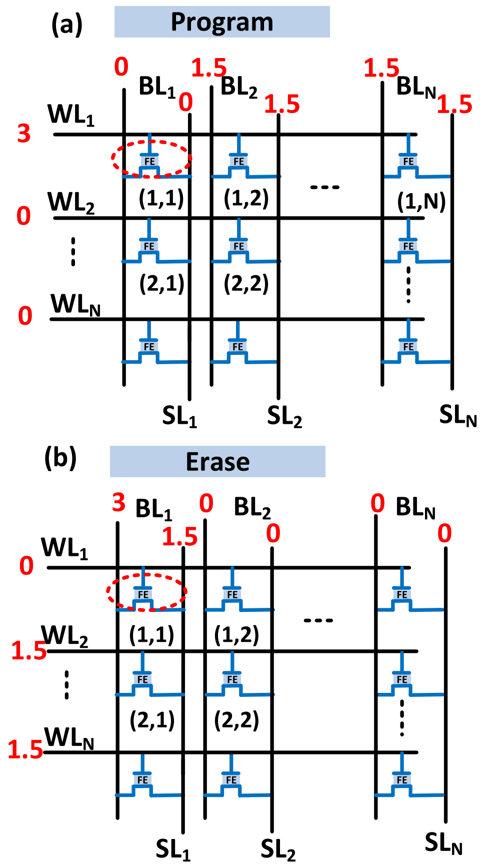

Fig. 1. (a) Individual cell's program and (b) erase scheme in the 2-D AND array with drain-erase scheme. Cell (1, 1) is the selected cell.

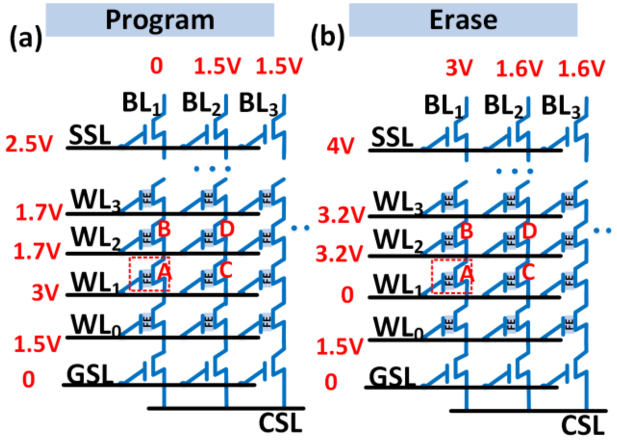

Fig. 2. (a) Individual cell's program and (b) erase scheme in the 2-D NAND array with drain-erase scheme. Cell A is the selected cell.

The individual cell's program and erase scheme for the NAND array is shown in Fig. 2. The NAND array is different from the AND array in that the NAND array needs select transistors on the top and bottom of the string. FeFETs in the same column are connected in series, forming a string. Select transistors are connected to the top and bottom of the string to isolate from the bitline (BL) and the common source line (CSL), respectively. In the AND array, the voltage is directly applying to the gate–source, and to the drain of the transistors. However, in the NAND only the gate voltage could be directly applied by the word line (WL). The source and drain voltages need to be given from the BL, and passing through the select transistor with an appropriate passing voltage.

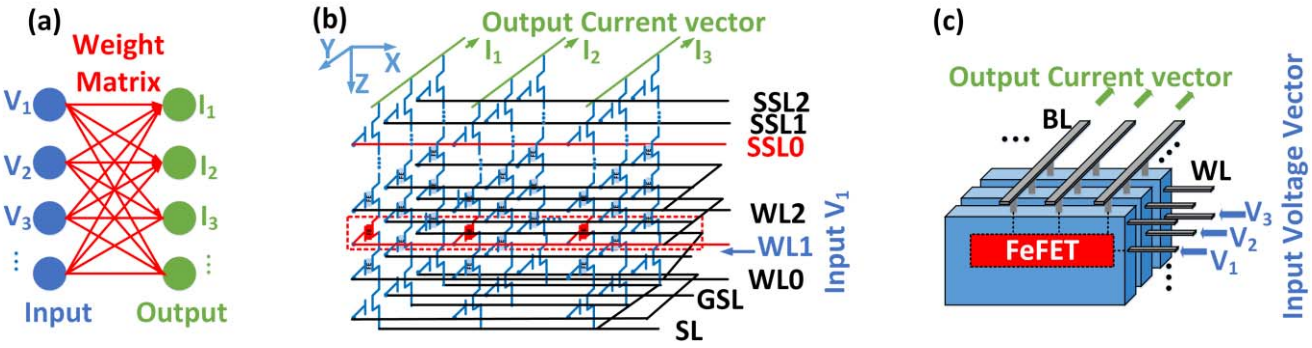

Fig. 3. (a) Weight matrix between two layers in a neural network. (b) Circuit diagram and bias scheme of 3-D NAND-like FeFET array for VMM operation within one block. The weights are mapped to the multilevel channel conductance of the FeFETs that are connected in the same *xy* plane. (c) Schematic of 3-D FeFET array consisting of multiple blocks. Input voltage vector is applied to WLs of the selected layer in different blocks from the *x*-direction. The weighted sum is computed by reading out currents along BLs that are shared among blocks in the *y*-direction. VMM is done in a layer-by-layer fashion.

As shown in Fig. 2(a), Cell A is the selected cell to be programed, and its WL is biased at a programming voltage (3 V) and the selected BL_1 is grounded. Cell C shares the same WL with Cell A. To inhibit Cell C from being programed, the drain of Cell C should be boosted to $V_{\text{inhibit}} = 1.5$ V from the unselected BL_2. All the upper FeFET's gates are biased at 1.7 V larger than ($V_{\text{inhibit}} + V_{\text{th}}$) to pass the BL_2's 1.5 V drain to Cell C. All the lower FeFETs' gates are biased at 1.5 V to prevent the lower layer cells to be programed. Other cells would not have enough voltage difference to be disturbed.

As shown in Fig. 2(b), Cell A is the selected cell to be erased by applying 3 V to the drain while its gate is grounded. All the upper FeFET's gates are biased at 3.2 V larger than ($V_{\text{erase}} + V_{\text{th}}$) to pass BL_1's 3 V to the drain of Cell A. All the lower FeFETs' gates are biased at 1.5 V to prevent the lower layer cells to be programed or erased. The ground select line (GSL) needs be closed so that the source of Cell A would be floating. Then the source voltage could be precharged to 1.5 V first and remain at 1.5 V during the erase operation, which is a key parameter in achieving successfully erase, as discussed in the Part-I article [21].

IV. 3-D FeFET Array for In-Memory Computing

To accommodate the high demands for memory storage in DNNs, we proposed a 3-D vertical channel NAND-like ferroelectric transistor (FeFET) array architecture feasible for both *in situ* training and inference. Fig. 3 shows the circuit schematic of a 3-D NAND-like FeFET array architecture. The top and bottom layers are string select transistors and ground select transistors. The gates of string select transistors in the same row (x-direction) are connected to the same string select line (SSL). All the gates of the bottom layers are connected by the GSL. The middle layers are all vertical channel FeFETs, forming pillars in the z-direction. In each block, all the pillars in the y-direction share the same BL, while all the gates of the FeFETs in the same layer (xy plane) are connected to the same WL at the edge of the plane. The weights in the neural network can be mapped to the multilevel channel conductance of the FeFETs that are connected in the same xy plane. Sharing all the gates at the same layer in one block will allow only one WL input. However, multiple inputs are required in the neural network, which could be solved by combining the WLs among different blocks in the y-direction. As shown in Fig. 3(c), the BLs among different blocks are connected, while the WLs are independent among blocks. This 3-D array design is the same as the commercial 3-D V-NAND Flash memory [18], [19], and is thus compatible with the mature 3-D NAND fabrication technology with minor changes such as replacing the charge-trapping dielectric with ferroelectric-doped HfO_2.

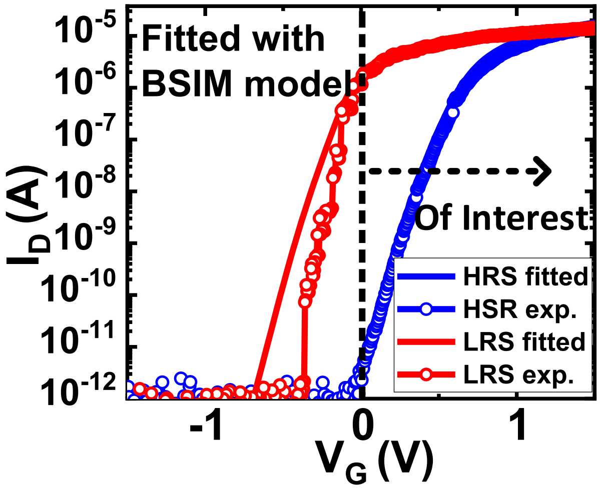

Fig. 4. FeFET's $I_D - V_G$ curve fit with the modified BSIM model. $I_D - V_G$ fit well when $V_G > 0$, and in our proposed schemes, FeFET always sees a positive or zero gate voltage.

When performing vector-matrix multiplication (VMM), the input vector is applied to WLs of multiple blocks from the x-direction to activate one layer, and BL currents are summed up along the y-direction from multiple blocks as the output. In this way, the weight matrix is represented as the transistor's channel conductance, while tuning the transistor's threshold voltage (V_{th}) by erasing and programming the cell could effectively update the weight.

V. Simulation on 3-D FeFET Array

To evaluate the feasibility of the drain-erase scheme to the 3-D NAND-like FeFET array, the BSIM model was modified to fit the experimental programed/erased FeFET's I_D–V_G curve (Fig. 4) and then it was used for SPICE simulation for the

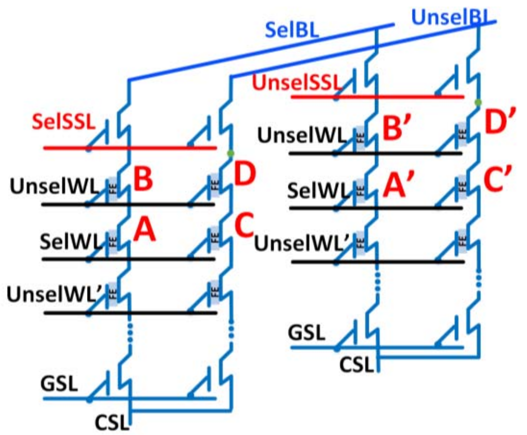

Fig. 5. 3-D NAND-like FeFET array bias scheme for an individual cell's erase/program scheme. Cell A is the selected cell.

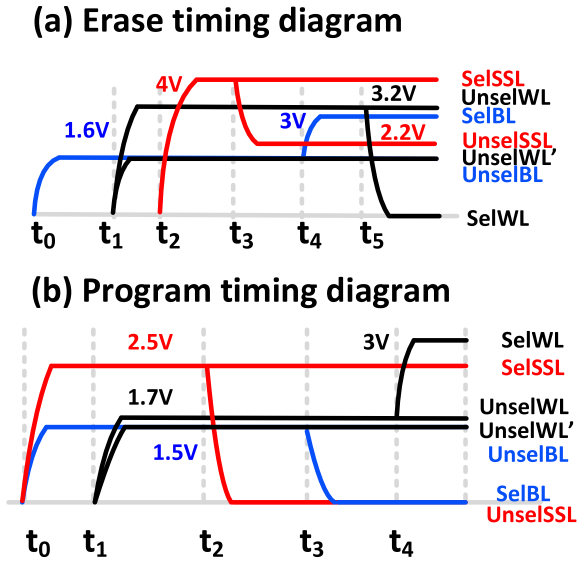

Fig. 6. (a) 3-D FeFET array timing diagram for Cell A′s erase. (b) 3-D FeFET array timing diagram for Cell A′s program.

3-D array. I_D–V_G is fit well when $V_G > 0$, and in our proposed schemes, FeFET always sees a positive or zero gate voltage. Therefore, the model is accurate in the regions of interest. To illustrate the timing diagram, the naming of each line and node involved in the 3-D programming/erase scheme is marked in Fig. 5.

It should be noted that for all the SPICE simulations, all the cell states (either programed state or erased state) do not change during the simulation, as our model does not capture the actual switching. We only build the model to extract the node voltage and calculate the gate-to-channel voltage difference to evaluate whether the cell will be programed, erased, or disturbed with proper biasing.

A. 3-D-NAND FeFET Individual Cell's Erase Scheme

To erase Cell A, SelBL should be 3 V, UnselWL will be biased at 3.2 V to pass the 3 V through, and SelWL should be 0 V. In a vertical *xz* plane, Cell B, Cell C, and Cell D's write disturbance situation is similar to a 2-D NAND array, as discussed in Fig. 2 in Section III. In the 3-D NAND array, the conditions of Cell A′, Cell B′ , and Cell D′ need to be considered, since the cells in the same *xy* plane share the same WL as the gate. Cell A′ has the same gate voltage as Cell A, and the corresponding BL for Cell A′ is the same as that of Cell A. Cell A′ receives 0 V at its gate, therefore, the UnselSSL must be off to prevent Cell A′ from being erased. Thus, during the erase operation, the top SSL transistors and the bottom GSL transistors of pillar A′–B′ and pillar C′–D′ will be closed. The channel potential of pillar A′–B′ and pillar C′–D′ depends mainly on its initial voltage before the SSL turns off, which should be boosted to V_{inhibit} before the erase operation to prevent Cell A′/C′ from being erased and to prevent Cell B′/D′ from being programed.

Considering the write disturbance, the erase sequence should be as follows [Fig. 6(a)].

1) Setting BLs at t_0: all the BLs are raised to 1.6 V to inhibit unselected cells from being programed in the latter timing period.
2) Setting WLs at t_1: upper unselected WLs are raised up to 3.2 V to make sure the BL erase voltage can be transferred to the drain side of the selected cell. Lower unselected WLs are raised to 1.6 V to avoid a lower cell from being either programed or erased. Since the SSLs are not turned on at this time, the channel of the cells will be coupled to its gate voltage, and high WL voltage will not program the cell.
3) Setting SSLs at t_2: all the SSLs are raised to 4 V to transfer the BL voltage to the channel string and unselected SSL string (A′–D′) channels are discharged to 1.6 V. Since the unselected cell's channel voltage is coupled to the high-gate voltage in t_1, if it is not discharged and remains high as t_1, the gate-to-channel voltage difference between Cell A′ and Cell C′ would be around −3 V at t_5 and erase disturbance occurs.
4) Unselected SSLs clamping at t_3: reducing the voltage of unselected SSLs can avoid transferring the erase BL 3 V to the neighbor cell (Cell A′) that shares the same WL and BL with the selected cell during the erase operation.
5) Selected BL setup at t_4: the erase voltage 3 V is applied only to the selected BL.
6) Erase operation at t_5: grounding the selected WL to erase Cell A only.

To validate this scheme, a SPICE circuit simulation was performed with a 3-D netlist for the array in transient mode to check the node voltage in each timing point (Fig. 7). The voltage waveform in Fig. 7(b) proves that only Cell A′s drain is ∼3 V and the source is ∼1.5 V. According to the above simulation, Cell A could be successfully erased. This simulation also verifies that all the other cells will not be disturbed due to insufficient node voltage differences.

B. 3-D-NAND FeFET Individual Cell's Program Scheme

To program the selected Cell A, its WL should be biased at 3 V and its channel should be 0 V by grounding the select BL and turning on all the upper passing transistors. Therefore, Cells C, A′ , and C′ all receive a 3 V gate voltage. To prevent these cells from being programed, their source

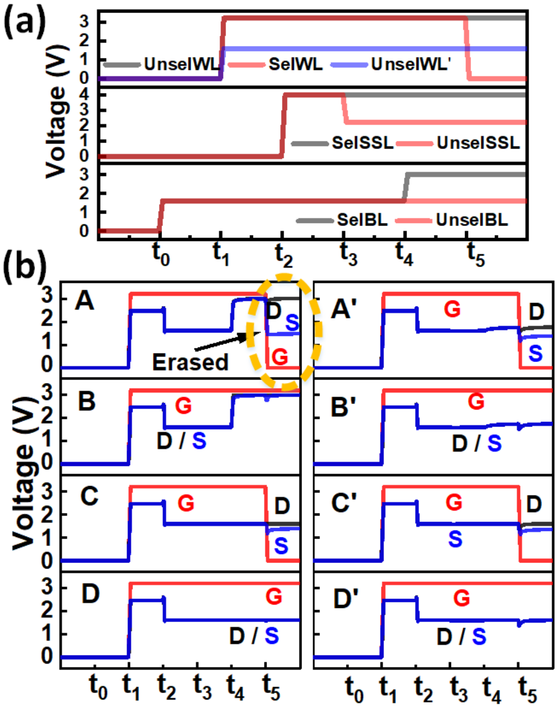

Fig. 7. SPICE transient simulation of the 3-D FeFET array erase scheme showing (a) WL/SSL/BL setup and (b) source and drain node voltage of different cells marked in Fig. 5; only Cell A is erased.

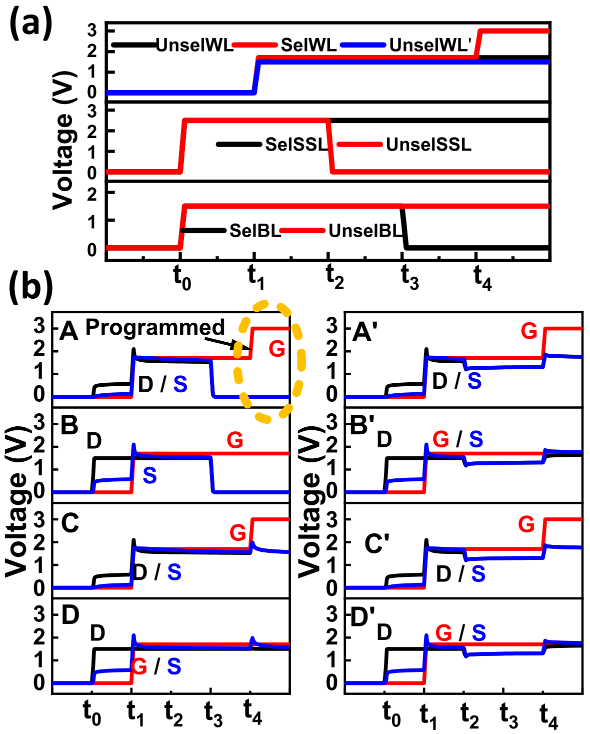

Fig. 8. SPICE transient simulation of the 3-D FeFET array program scheme showing (a) WL/SSL/BL setup and (b) source and drain node voltage of different cells marked in Fig. 5; only Cell A is programed.

and drain should be biased at V_{inhibit}. For Cell C, V_{inhibit} could be applied by unselected BL and passing through upper passing transistors. However, the unselected SSLs should be turned off to prevent the selected BL voltage (0 V) being passed to the drain of Cell A′, thereby programming Cell A′. Therefore, the source and the drain voltage of A′ and C′ need to be precharged to V_{inhibit} through activating all the SSLs and setting the BL at V_{inhibit}.

As shown in Fig. 6(b), considering the write disturbance, the programming sequence should be as follows: (1) BLs/SSLs setup at t_0: all the SSLs are raised to 2.5 V and BLs are raised to 1.5 V to provide enough drain inhibition voltage for unselected cells during programming operation; (2) turning on the WLs at t_1: all the channels are charged to the BL voltage 1.5 V with WLs and SSLs turned on so that Cells A′, C, and C′ will not be programed at t_4; (3) turning off unselected SSLs at t_2: since the selected BL will be grounded during programming, turning off unselected SSLs avoids transfer of the 0 V selected BL voltage to Cell A′, otherwise Cell A′ will be programed; (4) grounding the selected BL at t_3, and unselected BLs remain high to avoid Cell C from being programed; and (5) raising the selected WL to program voltage at t_4 to program Cell A.

Similarly, 3-D array-level SPICE simulation was performed to validate that only Cell A receives 3 V at the gate and 0 V at the drain and source for effective gate programming, while the other cells will not be disturbed (Fig. 8).

C. 3-D-NAND FeFET Vector-Matrix-Multiplication

After the program/erase operation, 3-D NAND-like array could be used for VMM in a layer-by-layer computation mode similar to the 3-D NAND Flash-based design in [20]. As shown in Fig. 3(b), during the VMM operation, the source lines (SLs) receive the same read voltage and the BLs are grounded. SSL will select the *xz* plane that has current passing through the BLs. All the WLs in the other unselected layers are biased at a higher read pass voltage such that the unselected transistors will act as pass-transistors, independent of the value of their *V*th. The input voltage or ground is then applied to the WLs in the selected layer among different blocks according to the input vector pattern. The current in each pillar depends mainly on the FeFET channel conductance in the selected layer. Then all the drain currents of the selected cells [red cells in Fig. 3(b)] among different blocks will be summed up along BLs to represent the weighted sum results in the corresponding columns.

The limiting factor of the VMM accuracy is the series channel resistance in passing transistors along the pillar. To test the weighted sum accuracy, the number of ON-states in each column in an *xy* plane varies, while all the other unselected cells are in the OFF-state. Simulated BL current from a 128(BLs) × 128(blocks) × (4-layer or 8-layer) array is shown in Fig. 9. The BL current is in a good linear relationship with the number of ON-state cells. However, compared to a single layer (2-D case), read-out current is reduced due to the voltage drop on passing transistors in the 3-D array; the reduced BL current could be compensated by the peripheral circuitry. Scalability toward large-scale 1024 × 1024 × (8-layer) 3-D array is explored in Fig. 10, showing that the *xy* plane scalability is good while the limiting factor is still the number of vertical layers.

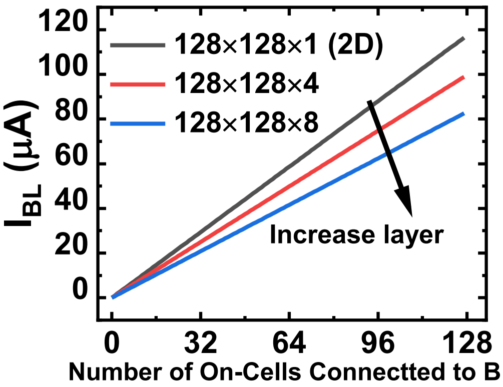

Fig. 9. Simulated BL current output in 3-D FeFET array during VMM as a function of the number of ON-state cells connected to BLs for a 128(BLs) × 128(blocks) × 8(layers) 3-D NAND array.

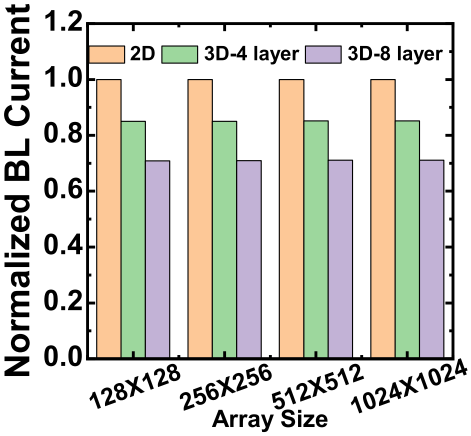

Fig. 10. Simulated weighted sum current (normalized to one-layer 2-D case) of 3-D FeFET array with various array sizes.

VI. Conclusion

With the individual cell operation mode that we characterized in the Part-I article [21], the biasing scheme of both 2-D AND and 2-D NAND array were designed to show an individual cell's erase/program with the drain-erase scheme. Then, we proposed a 3-D NAND-like FeFET array architecture feasible for both *in situ* training and inference. With the extracted BSIM model of FeFET and the specially designed 3-D timing sequence, the program/erase (without disturb) is successfully demonstrated through 3-D array-level SPICE simulations. VMM operations were also simulated, showing scalability to a large *xy* plane with a potential limit on the number of vertical layers. This article provided the design guidelines of engineering FeFET for in-memory computing.

References

[1] X. Si *et al.*, "24.5 a twin-8T SRAM computation-in-memory macro for multiple-bit CNN-based machine learning," in *IEEE Int. Solid-State Circuits Conf. (ISSCC) Dig. Tech. Papers*, Feb. 2019, pp. 396–398.

[2] G. Burr *et al.*, "Experimental demonstration and tolerancing of a large-scale neural network (165 000 synapses), using phase-change memory as the synaptic weight element," in *IEDM Tech. Dig.*, Dec. 2014, pp. 29.5.1–29.5.4, doi: 10.1109/IEDM.2014.7047135.

[3] W. Kim *et al.*, "Confined PCM-based analog synaptic devices offering low resistance-drift and 1000 programmable states for deep learning," in *Proc. Symp. VLSI Technol.*, Jun. 2019, pp. T66–T67, doi: 10.23919/vlsit.2019.8776551.

[4] M. Prezioso, F. Merrikh-Bayat, B. D. Hoskins, G. C. Adam, K. K. Likharev, and D. B. Strukov, "Training and operation of an integrated neuromorphic network based on metal-oxide memristors," *Nature*, vol. 521, no. 7550, pp. 61–64, May 2015, doi: 10.1038/nature14441.

[5] W. Wu *et al.*, "A methodology to improve linearity of analog RRAM for neuromorphic computing," in *Proc. IEEE Symp. VLSI Technol.*, Jun. 2018, pp. 103–104.

[6] F. Cai *et al.*, "A fully integrated reprogrammable memristor–CMOS system for efficient multiply–accumulate operations," *Nature Electron.*, vol. 2, no. 7, pp. 290–299, Jul. 2019.

[7] C. Li *et al.*, "Efficient and self-adaptive *in-situ* learning in multilayer memristor neural networks," *Nature Commun.*, vol. 9, no. 1, 2018, Art. no. 2385.

[8] X. Gu, Z. Wan, and S. S. Iyer, "Charge-trap transistors for CMOS-only analog memory," *IEEE Trans. Electron Devices*, vol. 66, no. 10, pp. 4183–4187, Oct. 2019.

[9] X. Guo *et al.*, "Fast, energy-efficient, robust, and reproducible mixed-signal neuromorphic classifier based on embedded NOR flash memory technology," in *IEDM Tech. Dig.*, Dec. 2017, pp. 6.5.1–6.5.4.

[10] Y.-Y. Lin *et al.*, "A novel voltage-accumulation vector-matrix multiplication architecture using resistor-shunted floating gate flash memory device for low-power and high-density neural network applications," in *IEDM Tech. Dig.*, Dec. 2018, pp. 39–42.

[11] H.-T. Lue, W. Chen, H.-S. Chang, K.-C. Wang, and C.-Y. Lu, "A novel 3D and-type NVM architecture capable of high-density, low-power in-memory sum-of-product computation for artificial intelligence application," in *Proc. IEEE Symp. VLSI Technol.*, Jun. 2018, pp. 177–178.

[12] M. Jerry *et al.*, "Ferroelectric FET analog synapse for acceleration of deep neural network training," in *IEDM Tech. Dig.*, San Francisco, CA, USA, Dec. 2017, pp. 6.2.1–6.2.4, doi: 10.1109/IEDM.2017.8268338.

[13] X. Sun, P. Wang, K. Ni, S. Datta, and S. Yu, "Exploiting hybrid precision for training and inference: A 2T-1FEFET based analog synaptic weight cell," in *IEDM Tech. Dig.*, Dec. 2018, pp. 3.1.1–3.1.4, doi: 10.1109 394/IEDM.2018.8614611.

[14] H. Mulaosmanovic, E. T. Breyer, T. Mikolajick, and S. Slesazeck, "Ferroelectric FETs with 20-nm-thick HfO_2 layer for large memory window and high performance," *IEEE Trans. Electron Devices*, vol. 66, no. 9, pp. 3828–3833, Sep. 2019, doi: 10.1109/TED.2019.2930749.

[15] S. Dunkel *et al.*, "A FeFET based super-low-power ultra-fast embedded NVM technology for 22nm FDSOI and beyond," in *IEDM Tech. Dig.*, Dec. 2017, pp. 19.7.1–19.7.4, doi: 10.1109/IEDM.2017.8268425.

[16] M. Trentzsch *et al.*, "A 28nm HKMG super low power embedded NVM technology based on ferroelectric FETs," in *IEDM Tech. Dig.*, Dec. 2016, pp. 11.5.1–11.5.4, doi: 10.1109/IEDM.2016.7838397.

[17] K. Florent *et al.*, "Vertical ferroelectric HfO_2 FET based on 3-D NAND architecture: Towards dense low-power memory," in *IEDM Tech. Dig.*, Dec. 2018, pp. 2.5.1–2.5.4..

[18] W. Kim *et al.*, "Multi-layered vertical gate NAND flash overcoming stacking limit for terabit density storage," in *Proc. IEEE Symp. VLSI Technol.*, Jun. 2009, pp. 188–189.

[19] D. Kang *et al.*, "A 512 Gb 3-bit/cell 3D 6th-generation V-NAND flash memory with 82 MB/s write throughput and 1.2 Gb/s interface," in *IEEE Int. Solid-State Circuits Conf. (ISSCC) Dig. Tech. Papers*, 2019, pp. 216–218.

[20] P. Wang *et al.*, "Three-dimensional NAND flash for vector-matrix multiplication," in *IEEE Trans. Very Large Scale Integr. (VLSI) Syst.*, vol. 27, no. 4, pp. 988–991, Apr. 2019, doi: 10.1109/TVLSI.2018.2882194.

[21] P. Wang *et al.*, "Drain–erase scheme in ferroelectric field-effect transistor—Part I: Device characterization," *IEEE Trans. Electron Devices*, vol. 67, no. 3, pp. 955–961, Mar. 2020.