electronics

MDPI

Article

A 5.67 ENOB Vector Matrix Multiplier with Charge Storage FET Cells and Non-Linearity Compensation Techniques

Jin-Young Hwang †, Young-Taek Ryu † and Kee-Won Kwon *

Department of Electrical and Computer Engineering, Sungkyunkwan University, Suwon 16419, Korea
* Correspondence: keewkwon@skku.edu; Tel.: +81-31-299-4583
† Theses authors contributed equally to this work.

Abstract: In this paper, we provide a thorough analysis and enhancement techniques of the linearity between the input voltage and output current in charge storage field effect transistor (FET) cells for a vector–matrix multiplier array in neural networks. A planar floating gate FET cell revealed superior linearity, because of boosting the floating gate using a drain voltage through capacitive coupling. If the coupling capacitance is extended by up to half of the gate capacitance, the coefficient of determination for linear regression is easily greater than 99.5%. However, the linearity of the charge trap FET, which keeps electrons in the insulating gate dielectric, must be compensated by either boosting the drain voltage, using a non-linear input driver, or supplying a quadratic current through an auxiliary path in the cell. Drain voltage boosting is limitedly effective over a small input range, while the auxiliary current path shows a coefficient of determination greater than 99.5% over a 500 mV input range. If the cell area matters, the charge trap FET with a diode connected FET as an auxiliary current path revealed the best performance, with an effective number of bits of 5.67, in a 21.3 F^2 cell area.

Keywords: neural network; in-memory computing; charge trap FET; linearity; floating gate FET; planar EEPROM; extended coupling capacitor

check for updates

Citation: Hwang, J.-Y.; Ryu, Y.-T.; Kwon, K.-W. A 5.67 ENOB Vector Matrix Multiplier with Charge Storage FET Cells and Non-Linearity Compensation Techniques. *Electronics* **2022**, *11*, 2911. https://doi.org/10.3390/electronics11182911

Academic Editors: Jongsun Kim and Hyunchol Shin

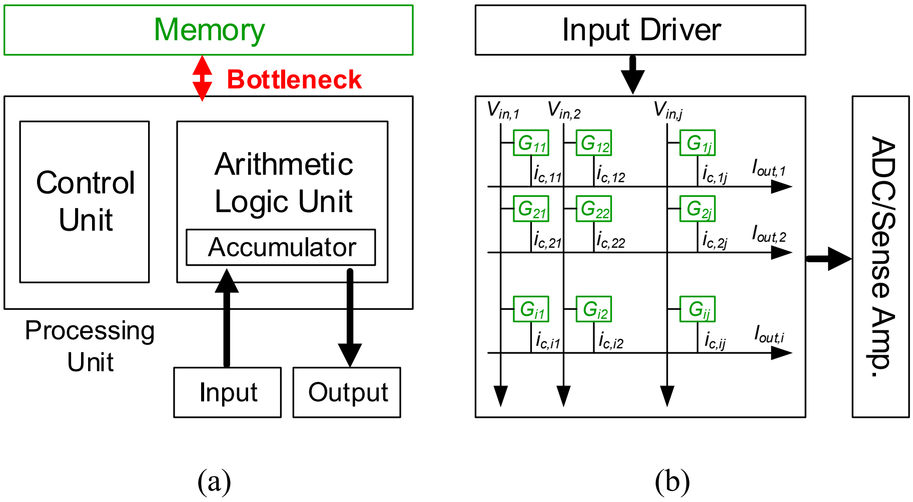
Received: 16 August 2022
Accepted: 8 September 2022
Published: 14 September 2022

1. Introduction

To comply with the emerging processing technologies of smart cities, autonomous driving, and AI robots, a tremendous amount of data needs to be efficiently processed [1–5]. However, the existing von Neumann computing structures shown in Figure 1a have limitations, because distinct areas are separated for dedicated roles [1], which causes heavy traffic in data transmission, inducing a significant efficiency drop and vast power consumption [2]. To overcome these problems, there is a growing interest in "compute-in-memory" (CIM) that mimics the human brain neural network. In CIM, the power efficiency and decision fidelity are significantly improved if multi-level data are processed in parallel and linearity of output is secured [3]. In the implementation of artificial intelligence, with constructed with secured linearity, neural network structures are suitable for in-memory computing in a cross-bar structure, as shown in Figure 1b [4–6]. The special features of the cross-bar structure are that each cell maintains the weight data in the form of cell conductance (G_{ij}); the cell current ($i_{c,ij}$) then appears as the product of the analog input voltage ($V_{in,j}$) and the cell conductance. Each row accumulates the cell current from cells within the row to generate the output current ($I_{out,i}$), which is the result of the multiply-and-accumulate (MAC) calculation, as shown in Equation (1) [7].

$$y_i = \sum_j \left(w_{ij} \cdot x_j\right), \quad I_{out,i} = \sum_j \left(G_{ij} \cdot V_{in,j}\right) \tag{1}$$

The MAC calculation in a cross-bar array is the most frequent and important operation of the vector–matrix multiplier (VMM) of neural networks [8]. The memory element in a neural network VMM has to be the multiplication operator and storage device for the weight factor. If a cell can store multiple levels of states, the MAC performance is enhanced and the power consumption is consequently reduced. The multi-level cells in a VMM array are expected to provide linear characteristics for both the inference and learning modes of the neural network [9]. The memory cell is read during the inference step, while the memory state is modified during the learning step. Other characteristics required of the multi-level cells for the inference and learning steps are summarized in Table 1.



(a) (b)

Figure 1. Schematic architectures of (**a**) von-Neumann and (**b**) in-memory computing.

Table 1. Requirements for a multi-level memory cell used in a neural network.

	Requirements
Inference	Linearity of states Memory retention Operation range or number of states
Training	Endurance for overwriting Programming/erase step control Programming speed Program disturbance of half-selected cells

Emerging non-volatile memory devices, including resistant random access memory (RRAM), phase change memory (PCM), charge trap transistors (CTTs), ferroelectric field effective transistors (FeFETs), and floating-gate (FG) FETs, have been proposed as promising candidates for neuromorphic cell memory. Table 2 compares the characteristics of non-volatile memory devices for neuromorphic VMM applications [10]. It is difficult to achieve linear multiple states of conductance with RRAM and PCM, and modification of the conductance is neither symmetric nor reproducible, because the program and erase functions behave differently. Alternatively, charge storage FETs, such as CTT and FG FET, improve the linearity and symmetry in multiple levels of conductance states when accompanied by an incremental step pulse program (ISPP) technique [11]. Although the linearity in multiple weight levels is acceptable in charge storage FET devices, the linearity between the input voltage ($V_{in,i}$) and cell current ($i_{c,ij}$) is not inherent. The linearity quickly deteriorates as the input voltage increases because of the secondary effects of FETs, which limits the narrow input swing range of the operation.

Table 2. Characteristics of the candidates for neuromorphic cell devices.

Device	RRAM	PCM	CTT	FeFET	FG FET
# of terminals	2	2	3	3	3
Cell size (F^2)	4–12	4–12	4–12	6–20	9–12
Write Power	High	High	Low	Low	Medium
Speed	Medium	Fast	Medium	Fast	Slow
Device Variation	High	High	Medium	High	Low
Endurance(Cycles)	10^4–10^{11}	<10^5	<10^5	<10^4	<10^5
No. of States	~128	~64	~256	32–64	~100
Linearity	Bad	Bad	Medium	Medium	Good

In this paper, we provide a systematic analysis of the linearity between the input voltage and cell current in FET-based neuromorphic VMM cells, and suggest noble cell structures that provide improved linearity over an extended input voltage range. The diminishing current because of the quadratic term of the cell current is compensated for by either boosting the bias on the FET terminals or adding an auxiliary current path in the cell, depending on the type of FET.

2. Linearity Improvement in VMM Cells Based on Floating Gate FETs

2.1. Array Architecture and Cell Candidates

Figure 2a shows the full architecture of VMM using charge storage FET cells. Although the word line driver supplies a constant potential to all gate nodes (WL) of FETs during the inference step, each cell may have a different conductance if the threshold voltage of the cell FET is altered. The drain current of a cell is determined by the cell conductance and input voltage on the drain node (BL). The merged drain current from cells that are attached to a WL enter the inference circuit to prioritize the correlation between the input vector and output vector under the weight matrix. All rows simultaneously drive the output current to perform a flash operation of VMM. The threshold voltage of the FET cells is updated row by row, by applying a large electric field between the selected WL and BLs. Unselected WLs are tied to ground, to avoid unwanted disturbance of the stored states during the training operation of the adjacent rows of the array.

The threshold of FET cells is simply modified by changing the amount of trapped charges between the gate node and channel of the FET, similarly to flash memory cells. Traditionally, flash memory writes/erases data by injecting/removing electrons into/from the conductive floating gate, respectively. However, charge trapped FETs, where the charges are stored in the insulating medium of the gate dielectric, have recently replaced floating gate FETs [12]. Commercial flash memory products are currently capable of maintaining multiple levels of electron storage in a cell, by changing the amount of trapped charges in the floating gate or gate dielectric. Figure 2b shows examples of charge storage FET cells for a neuromorphic cross-bar array. In the charge trap FET, shown on the left, electrons are stored in the gate oxide as fixed charges. As shown in Equation (2), the cell current is expressed by the equation of the drain current of a MOSFET with threshold voltage shifted by ΔV_T. μ_n and C_{ox} are the mobilities of the electron and gate capacitance, respectively. The threshold shifts reversely, but linearly, in proportion to the number of trapped electrons, N, as $\Delta V_T = -Nq/C_{ox}$, where q is the magnitude of charges of an electron. Weight-Current is inherently linear. The cell current is not a simple product of input voltage, V_{DS}, and weight, ΔV_T. Offset, $(V_{GS} - V_{th})$, and a second order term, $\frac{1}{2}V_{DS}^2$, are also factors.

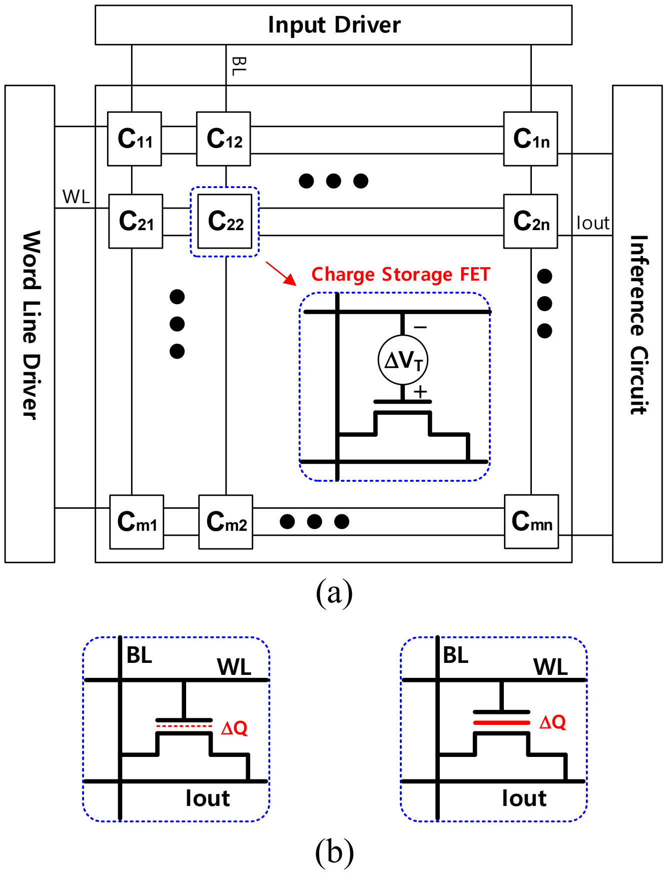


Figure 2. (**a**) Architecture of the VMM using a charge storage FET cell and (**b**) cell examples.

$$I_{D,FET} = \mu_n C_{ox}\left[(V_{GS} - V_{th} + \Delta V_T)V_{DS} - \frac{1}{2}V_{DS}^2\right] \tag{2}$$

If the cell is replaced with a floating gate FET, as shown in the right example in Figure 2b, the current equation is modified because of the capacitive voltage division and the coupling between the floating gate and drain node, as shown in Equation (3), where C_{FD} and C_{TOT} are the FG-drain coupling capacitance and total capacitance observed from the FG, respectively. The nonlinearity because of the secondary effect is mitigated by the capacitive coupling.

$$I_{D,C_{GD}} = \mu_n C_{ox}\left[(V_{GS} - V_{th} + \Delta V_T)V_{DS} - \left(\frac{1}{2} - \frac{C_{FD}}{C_{TOT}}\right)V_{DS}^2\right] \tag{3}$$

As C_{FD}/C_{TOT} is much smaller than 0.5 in the floating gate FET, the quadratic term in the current equation can be further reduced if additional coupling capacitance between FG and drain, C_{FDX}, is inserted, as shown in Equation (4):

$$I_{D,\ C_{GDX}} = \mu_n C_{ox}\left[(V_{GS} - V_{th} + \Delta V_T)V_{DS} - \left(\frac{1}{2} - \frac{C_{FD} + C_{FDX}}{C_{TOT} + C_{FDX}}\right)V_{DS}^2\right] \tag{4}$$

2.2. Floating Gate FET with Adjustable Coupling Capacitor

Electrically erasable and programmable read only memory (EEPROM) is a typical FG FET cell, where cell-by-cell control of electron trapping is possible. Figure 3a shows the cross-section view of common EEPROM cells with a stacked gate structure. The potential of the FG is controlled by the capacitive division of the control gate (CG) voltage. The sizes of the coupling capacitance (C_C) and tunneling capacitance (C_T) are fixed when the fabrication technology is selected. Different structures of FG FET can be made using only a single layer gate, as shown in Figure 3b. The sizes of the coupling and tunneling capacitances are determined by overlapping areas of the floating gate and underlying wells, such that the ratio of the two capacitances is adjustable [13]. In the single poly planar FG FET, the control gate is buried as an n-well. Moreover, the data line (DL) enables cell-by-cell control of both program and erase operations. The vestigial size of the tunneling capacitance in the single poly FG FET can increase the ratio of the coupling to the tunneling capacitance without increasing the overall area of the cell too much. Another advantage of a single poly FG FET is that it can be fabricated using the standard CMOS process.

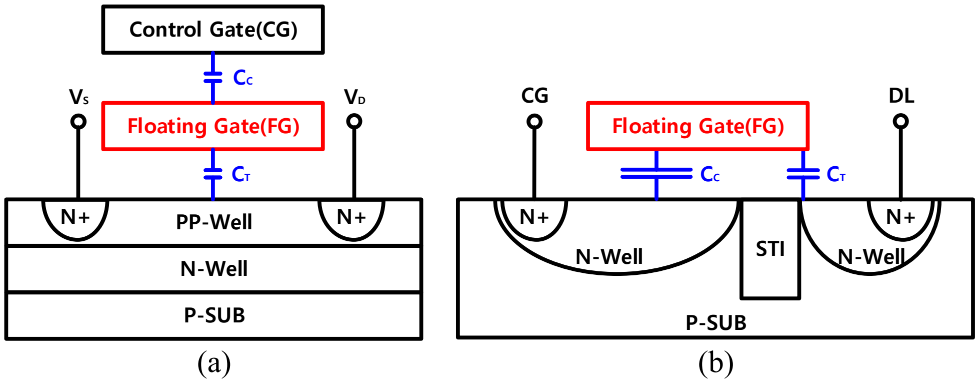


Figure 3. Cross-section view of EEPROM cells with a (**a**) stacked gate and (**b**) planar gate.

Our study starts with a linearity analysis of a VMM cell comprising a single poly FG FET, as shown in Figure 4. As the size ratio of the coupling to tunneling capacitors is set to 20, the voltage across the WL and DL mostly drops at the tunneling dielectric, enabling low voltage programming. Efficient programming at reduced voltage improves the update-after-learning speed of the neural network and saves on the power consumed during the training step. The floating gate in a single poly FG FET is further extended to a MOSFET to drive the cell current, depending on the trapped charges during the inference cycle. The advantages and disadvantages of stacked and planar EEPROMs are summarized in Table 3. A planar FG FET gains an adjustable capacitance ratio and room for an additional capacitor (C_{FDX}), to improve the linearity of the VMM calculation by sacrificing the cell area. A perfect linear relationship between the drain current and drain voltage is achieved when $(C_{FD} + C_{FDX})/(C_{TOT} + C_{FDX})$ is 0.5.

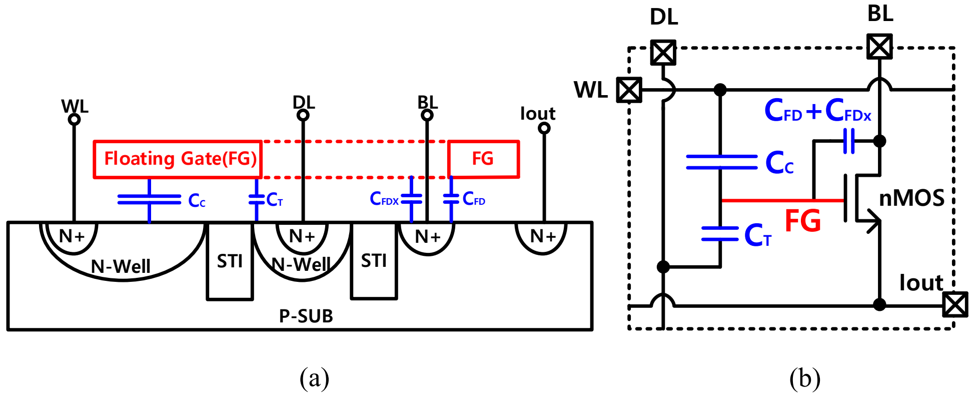


Figure 4. (**a**) Cross-section view and (**b**) equivalent circuit schematic for a single poly FG FET cell for a neuromorphic cross-bar array.

Table 3. Comparison of stacked and planar EEPROMs.

	Stacked EEPROM	**Planar EEPROM**
Capacitor Size Ratio	Fixed	Adjustable
Additional Capacitor	Hard to Add	Easy to Add
Relative Cell Area	×1	×8 (w/o C_{FDX})
CMOS Compatible	No	Yes
Program and Erase Voltage	~14 V	~7 V

2.3. Linearity Analysis of Planar FG FET Cells

Using a physics-based TCAD tool, a planar EEOROM cell, as shown in Figure 4a, was constructed and simulated. The equivalent circuit diagram of the planar EEPROM is described in Figure 4b. The programming indicates electron injection into the floating gate by applying a positive voltage on the WL node, while maintaining the DL node at the ground. Electron accumulation in the floating gate induces an increase in the threshold voltage and reduces the conductance of the MOSFET at a given gate voltage. The programming operation corresponds to the expected depression operation in the learning step of the neuromorphic processors. Alternatively, the erase operation extracts electrons from the floating gate by applying an opposite electric field. The reduced amount of electrons in the floating gate increases the conductance of the MOSFET that is expected in the potentiation process in the learning step of the neural network. Adaptive feedback control of the program/erase voltage is used to keep a constant step size of electron injection/removal, respectively; independently of the precedent program or erase history [14].

A constant voltage, V_{read}, is applied to all WLs during the inference operation, so that each cell produces a cell current proportional to the product of the input voltage on the BL node and the conductance of the charge storage FET. The drain current from cells that share the source node of FET is accumulated to form a dot product of the conductance matrix and input vector.

Figure 5 shows the simulation results of the drain current of the planar FG FET cells, with (3C-1T) and without (3C-1T) extended overlap capacitance (C_{FDX}), as a function of drain (BL) voltage. The size of C_{FDX} is half of C_C in this study. The drain current of the charge trap FET is also included as a reference. At first glance, the linearity is remarkably improved by C_{FDX}. For quantitative analysis, the cell current is fit to a polynomial, $i_c = C_0 + C_1V^1 + C_2V^2 + C_3V^3 + C_4V^4$. The regression analysis results are listed in Table 4 using coefficients of a polynomial. As expected in Equation (4), the quadratic term is diminished by less than 25%. A reduction in the linear coefficient, C_1, is induced by the capacitive voltage division with the extended overlap capacitance. Curves in different colors in Figure 5 indicate the drain current of a cell with different numbers of electrons in the floating gate or gate dielectric. Cells after electron injection ($-\Delta$Q) and electron removal (+ΔQ) have nearly the same tendency of linearity improvement with the extended overlap capacitance.

Since the FG-drain coupling in FG FET increases the potential of the FG, the FET turns to the saturation region at a higher drain voltage as the capacitive coupling increases. The linear behavior of the drain current is extended to a wider swing range of the drain voltage. Therefore, the linearity did not deteriorate much up to a of drain voltage of 500 mV.

Unfortunately, current EEPROM devices transition quickly from the conventional floating gate FET to the charge trap FET. As the charges are stored in the insulating medium, such as a gate dielectric in the charge trap FET, the capacitive coupling between the drain and charge trap medium is neither considerable nor adjustable. There are no reasons to maintain the planar EEPROM structure by consuming a large area in a high density array of a neural network. Next, the area-efficient charge trap FET structure and different approaches are used to improve the linearity between the drain voltage and drain current. The following section suggests different ways of improving the linearity of the VMM cells in the charge trap FET.

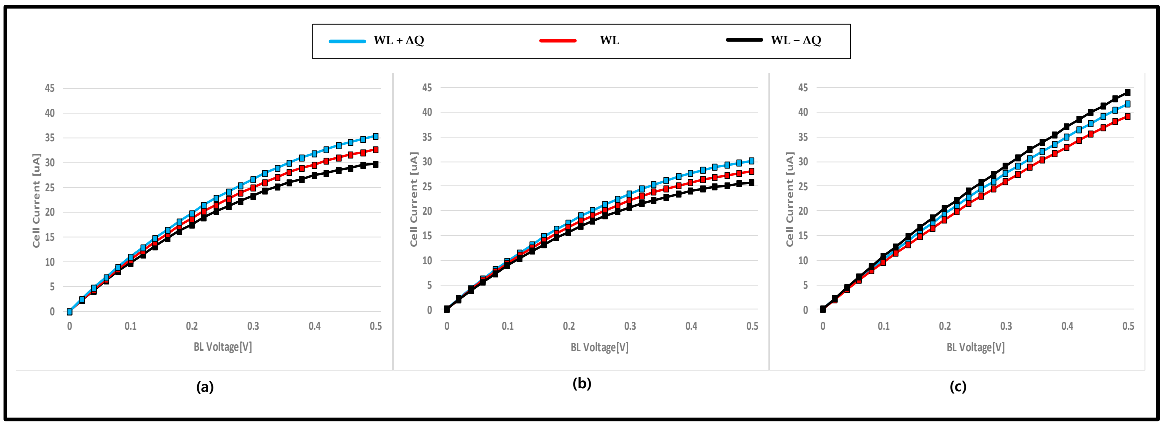


Figure 5. Drain current as a function of BL voltage for (**a**) CTT, (**b**) 2C-1T, and (**c**) 3C-1T cells.

Table 4. Polynomial coefficients of the drain current in CTT, 2C-1T, and 3C-1T FET cells.

Cell Type	C_0	C_1	C_2	C_3	C_4
CTT	−0.00245	2.5039	−0.0465	0.0002	-5×10^{-7}
2C-1T	−0.00221	2.2599	−0.0440	0.0002	-8×10^{-7}
3C-1T	−0.00227	2.3024	−0.0104	0.0002	-1×10^{-6}

3. Linearity Improvement in VMM Cells Based on a Charge Trap FET

3.1. Pre-Distorted BL Converter

The drain voltage in the floating gate FET improves the floating gate potential via the coupling capacitance, resulting in enhanced conductance. The increased conductance compensates the negative quadratic term in the cell current equation. However, the charge trap FET has negligible coupling between the drain node and trap sites in the dielectric film. Thus, it is difficult to compensate for the non-linearity by enhancing the FET conductance. Alternatively, the drain voltage is adjusted.

Figure 6 shows a schematic illustration of a method for linearity improvement of the current behavior of a charge trap FET cell. The cell has the same structure as a traditional EEPROM. The drain voltage is boosted using a pre-distorted BL driver, to compensate for the attenuating current increment of the drain voltage. After careful observation of the convex behavior of the current response of the non-linear charge trap FET cell, as shown in Figure 5a, a BL driver transforms the input to drain voltage in a concave way, similarly to a parabolic response, which is similar to the pre-emphasis technique in the communication circuit, with which the transmitter driver inversely distorts the signal as much as it is expected to be attenuated while traveling to the receiver [15]. As mentioned earlier, the cell is as small as the EEPROM cell; the area overhead for linearity enhancement in the FG FET is placed with circuits in the BL driver, which is shared by all cells in a column. The sequential processing of a non-linear BL driver with a concave transfer function and a non-linear cell with a convex current behavior is expected to cancel each other, resulting in a linear relation between the input voltage and drain current.

The simplest parabolic transfer function is available in the drain current of a MOSFET in the saturation region. A pre-distorted BL driver was designed using a voltage-dependent current source followed by an analog buffer, as shown in Figure 6. Through extensive trial-and-error optimization, the inverted transfer curve of a MOSFET was obtained in the pre-distorted BL driver with a resistance of 3 kΩ and a nMOS size that was the same as the cell transistors. Before entering the pre-distorted BL driver, the input voltage was shifted upwards by the threshold voltage, to turn on the BL driver. The level shifter is not included in the circuit schematic.

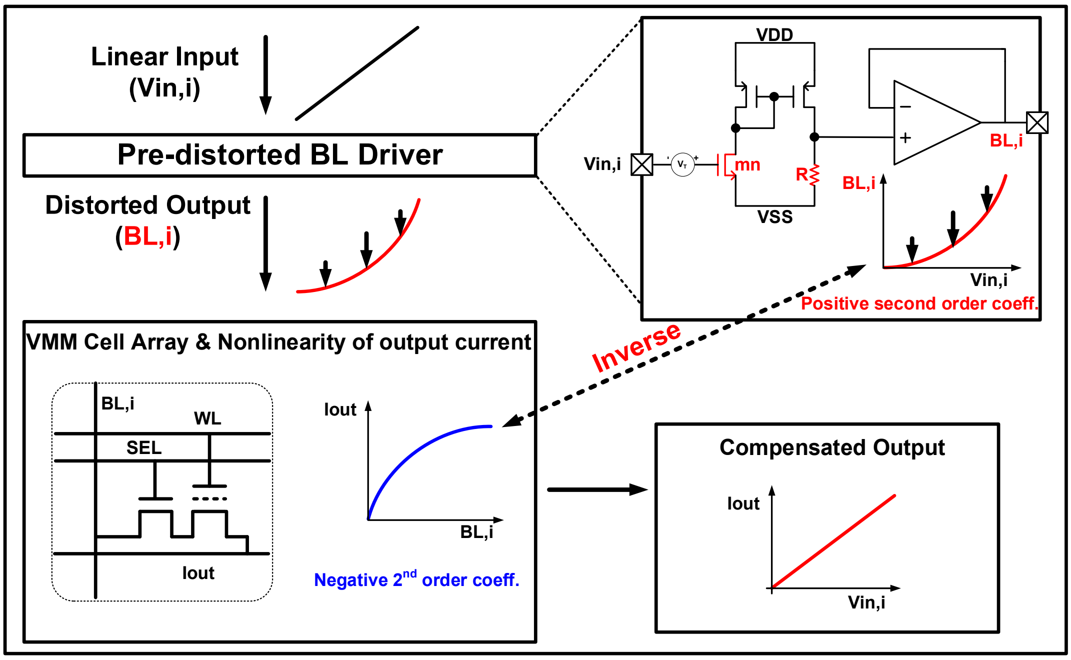


Figure 6. Linearity improvement of the current behavior using a pre-distorted BL driver.

Figure 7 shows the transfer function of the BL driver and the resultant drain current as a function of the input voltage. Although the inverse transfer function of the BL driver was obtained, the overall transfer from the input voltage to the output current was not linear. Linearity was maintained at an input voltage of 150 mV or less. The limited input swing range was attributed to the increased drain, which pushes the FET into the saturation region. Therefore, the convex non-linearity of the cell transistor is insufficiently compensated by boosting the drain voltage over a wide input range. As the input voltage increases, the output current response quickly loses linearity, because of the lack of headroom beyond the saturation boundary.

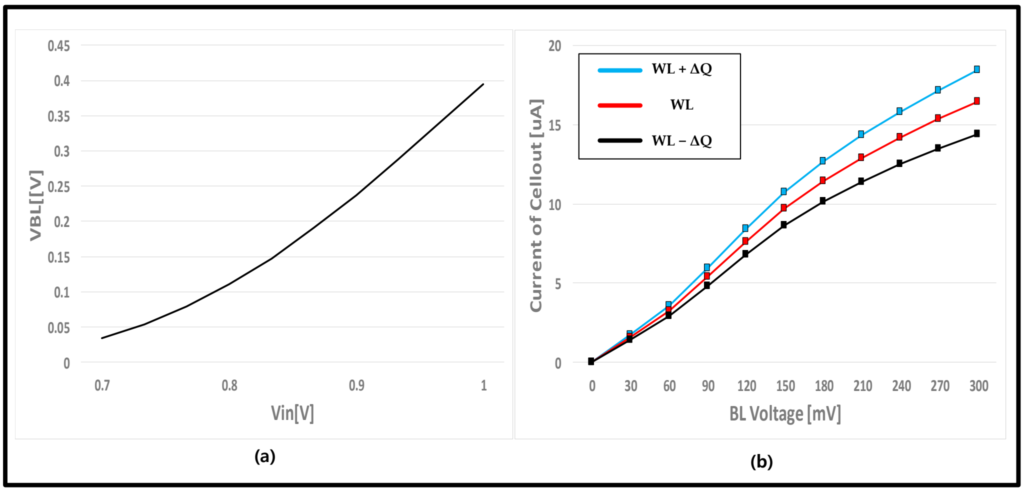


Figure 7. (**a**) Voltage transfer curve of the pre-distorted BL driver and (**b**) the CTT cell current.

3.2. Auxiliary Input Driver and Current Path

Sequential processing of concave and convex transfer functions multiplies the two functions. The multiplication function was inefficient in the compensation of nonlinearity over a wide operation range. The summation of the convex and concave transfer functions may also have cancelled each other. The positive quadratic term of the additional current path efficiently annihilates the negative quadratic term of the charge trap FET current over a wide operation range. Figure 8 shows the cell structure with a quadratic auxiliary current path. The auxiliary current path comprises a diode-connected FET (mn1) whose current response is pure quadratic, as shown in Equation (5).

$$I_{comp} = \frac{1}{2}\mu_n C_{ox}\frac{W}{L}(V_{in,aux} - V_{th})^2 \quad (5)$$

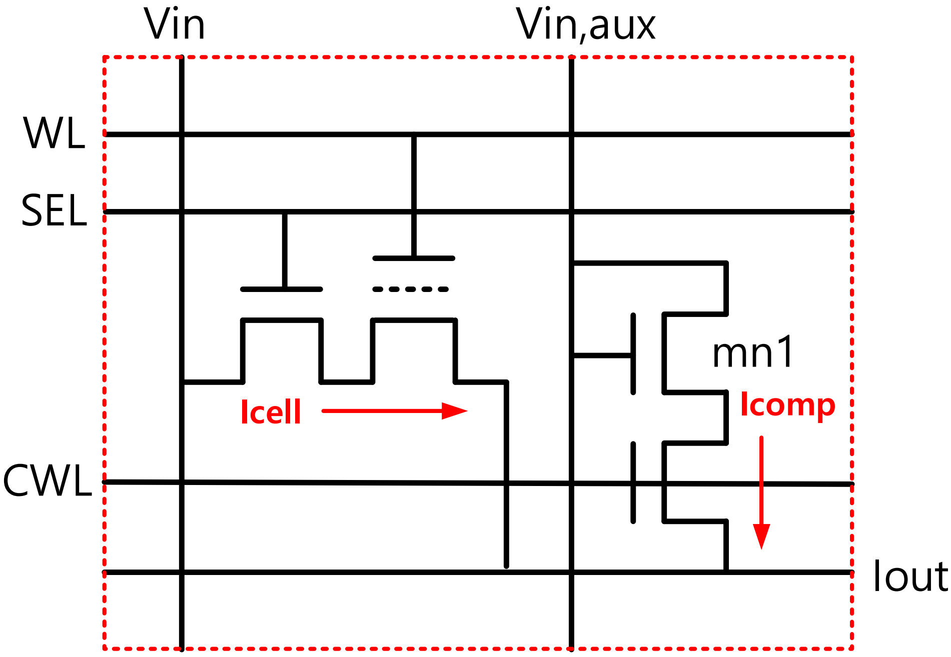


Figure 8. Proposed charge trap FET cell with auxiliary current path.

As in the pre-distorted BL driver, the auxiliary path functions when the input voltage is greater than the threshold voltage. Therefore, the input driver for the auxiliary input, $V_{in,aux}$, is a simple shift of V_{in} by V_{th}. The total cell current is expressed as in Equation (6). The quadratic term is completely eliminated when $V_{in,aux} = (V_{in} + V_{th})$. The CMW switch was added to block the current flow through the compensation path during programming and erase operations, to include a weight update step.

$$I_{CELL} = \mu_n C_{ox}\frac{\mathrm{W}}{\mathrm{L}}(V_{GS} - V_{th} + \Delta V_T)V_{in} - \frac{1}{2}\mu_n C_{ox}\frac{\mathrm{W}}{\mathrm{L}}\left[V_{in}^2 - (V_{in,aux} - V_{th})^2\right] \quad (6)$$

The quadratic current response over a wide range and the linearity compensation performance were verified, and the results are shown in Figure 9. The positive quadratic current response through the auxiliary path was verified and the linearity of the total cell current was comparable with that of the floating gate FET with extended coupling capacitor. Both exhibited a coefficient of determination greater than 99.5%. However, a large leakage current was observed near the zero input voltage via the auxiliary path, which is related to the sub-threshold current of the diode connected FET. Fortunately, the same amount of sub-threshold leakage is included in all rows, such that the prioritization and activation is not affected much by the constant error component of the subthreshold leakage. The area penalty for the auxiliary current path was 47% of the charge trap FET cell. However, the cell area was smaller than the 3C-1T cell, which is the only architecture that can provide a comparable linearity over a wide input range.

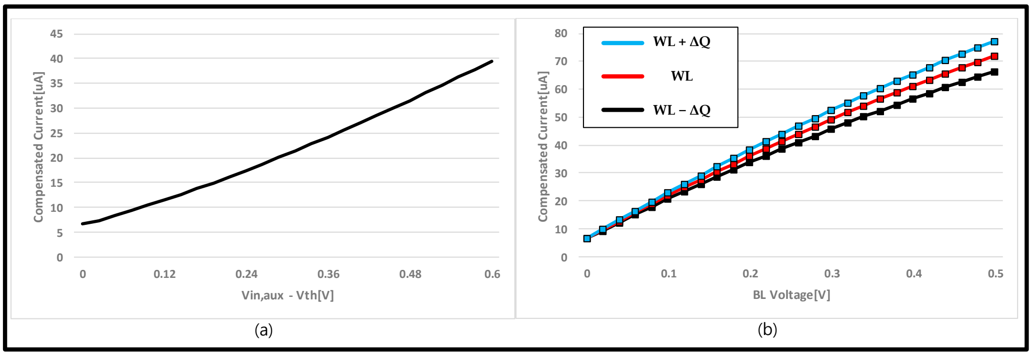


Figure 9. (**a**) Current response of the auxiliary current path and (**b**) total cell current.

4. Comparative Study

Cell layouts are compared in Figure 10. Although the planar floating gate FETs exhibited an inherently good linearity, they occupy a very large area when compared with a group of charge trap FETs. Cells comprising 2C-1T and 3C-1T are as large as 102 F^2 and 153 F^2, respectively, where F is the minimum feature size. The charge trap FET cell with pre-distorted BL driver is as small as a EEPROM cell, which is 29 F^2. The charge trap FET with a compensation current path requires 42.6 F^2 for a cell. The size of the input drivers is not included, because the impact on the overall integrated circuit is negligibly small, as all of the cells in a column share a driver.

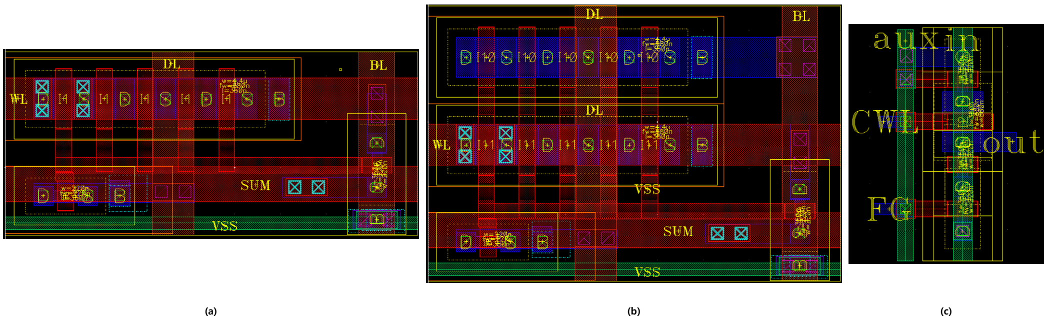


Figure 10. Layout of the unit cells in (**a**) 2C-1T, (**b**) 3C-1T, and (**c**) CTT with an auxiliary path.

Table 5 compares the key performance parameters of the four VMM cells. The linearity was evaluated in three different ways. The C1/C2 ratio in the polynomial regression was used to estimate the capability for quadratic term elimination. The coefficient of determination of the linear regression was adopted to check the overall trend of straightness of the output current. The signal-to-noise ratio (SNR) and effective number of bits (ENOB) are standard metrics for linearity analysis. In the calculation of SNR and ENOB, the response of the output current of the cells after injection and removal of the trapped electrons are also included. Without considering the area penalty, the 3C-1T cell can achieve the best linearity if an unlimited size of extended coupling capacitor is available. All linearity comparisons were conducted under an input swing of 300 mV. The normalized area was defined as a figure of merit, in which the cell area was divided by the ENOB, which represents the area occupancy for 1-bit processing of the neural network. The charge trap FET cell with an auxiliary current path had the best overall performance as a VMM cell for the neural network.

Table 5. Key performance parameters of the charge storage FET cells.

	2C-1T	3C-1T	BL Driver	Aux. Path
C1/C2	−51.36	−95.93	−53.84	−168.21
R^2	0.945	0.995	0.989	0.996
SNR (dB)	30.05	34.14	31.08	35.89
ENOB (bit)	4.7	5.37	4.87	5.67
Cell Area (F^2)	102	153	29	42.6
Cell Area/ENOB (F^2/bit)	21.7	28.49	5.95	7.51
V_{in} swing range (mV)	>500	>500	~150	~500

5. Conclusions

After a thorough analysis of the linearity of drain current with respect to the drain voltage of the FETs, two different approaches of linearity improvement for charge storage FET cells in a neural network were identified. When a conductive floating gate node is used for charge storage, the strong coupling between the floating gate and drain node improves the linearity, because of the loose bootstrapping of the effective gate potential. The coefficient of determination for linear regression easily exceeds 99.5%, because the extended FG-drain coupling exceeds half of the FG-gate coupling. If charges are trapped in the insulating medium, such as the gate dielectric, the linearity has to be improved by any means, to add positive quadratic current. Drain voltage boosting is not recommended, because it narrows the input swing range, although boosting could improve the linearity at a small input voltage. Therefore, an additional current path with a positive quadratic current response is effective for linearity enhancement over a wide input swing range. The best linearity was achieved in a charge trap FET with an auxiliary current path, resulting in a 5.67 ENOB with 42.6 F^2 cell area.

Author Contributions: Data collection and analysis, J.-Y.H. and Y.-T.R.; writing, J.-Y.H. and K.-W.K.; Supervise, K.-W.K. All authors have read and agreed to the published version of the manuscript.

Funding: This research was supported by National R&D Program through the National Research Foundation of Korea (NRF) funded by Ministry of Science and ICT (No. 2020M3H2A1076786). EDA tool was, in part, supported by the IC Design Education Center (IDEC), Korea.

Data Availability Statement: Not applicable.

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

References

1. Choi, H.S.; Park, Y.J.; Lee, J.H.; Yoon Kim, Y. 3-D Synapse Array Architecture Based on Charge-Trap Flash Memory for Neuromorphic Application. *Electronics* **2020**, *9*, 57. [CrossRef]
2. Chiu, Y.Y.; Shirota, R. Technique for Profiling the Cycling-Induced Oxide Trapped Charge in NAND Flash Memories. *Electronics* **2021**, *10*, 2492. [CrossRef]
3. Geoffrey, W.B.; 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.
4. Choi, H.-S.; Wee, D.-H.; Kim, H.; Kim, S.; Ryoo, K.-C.; Park, B.-G.; Kim, Y. 3-D Floating-Gate Synapse Array With Spike-Time-Dependent Plasticity. *IEEE Trans. Electron Devices* **2017**, *65*, 101–107. [CrossRef]
5. Lee, K.S.; Chun, J.H.; Kwon, K.W. A low power CMOS compatible embedded EEPROM for passive RFID tag. *Microelectron. J.* **2010**, *41*, 662–668. [CrossRef]
6. Shin, D.; Yoo, H.-J. The Heterogeneous Deep Neural Network Processor with a Non-von Neumann Architecture. *Proc. IEEE* **2020**, *108*, 1245–1260. [CrossRef]
7. Park, K.H.; Cho, M.H.; Jeon, Y.D.; Lee, J.H. Design of Analog and Digital Hybrid MAC Circuit for Artificial Neural Networks. In Proceedings of the 2019 International Conference on Electronics, Information, and Communication (ICEIC), Auckland, New Zealand, 22–25 January 2019.
8. Sze, V.; Chen, Y.-H.; Yang, T.-J.; Emer, J.S. Efficient Processing of Deep Neural Networks: A Tutorial and Survey. *Proc. IEEE* **2017**, *105*, 2295–2329. [CrossRef]

9. Kim, Y.-H.; Choi, J.-M.; Woo, J.-J.; Park, E.-J.; Kim, S.-W.; Kwon, K.-W. A 16x16 Programmable Analog Vector Matrix Multiplier using CMOS compatible Floating gate device. In Proceedings of the 2019 International Conference on Electronics, Information, and Communication (ICEIC), Auckland, New Zealand, 22–25 January 2019.
10. Choi, J.-M.; Kwon, D.-W.; Woo, J.-J.; Park, E.-J.; Kwon, K.-W. Implementation of an On-Chip Learning Neural Network IC Using Highly Linear Charge Trap Device. *IEEE Trans. Circuits Syst. I Regul. Pap.* **2021**, *68*, 2863–2875. [CrossRef]
11. Suh, K.D.; Suh, B.-H.; Lim, Y.-H.; Kim, J.-K.; Choi, Y.-J.; Koh, Y.-N.; Lee, S.-S.; Kwon, S.-C.; Choi, B.-S.; Yum, J.-S.; et al. A 3.3 V 32 Mb NAND flash memory with incremental step pulse programming scheme. *IEEE J. Solid-State Circuits* **1995**, *30*, 1149–1156.
12. Ted Chang, S.D.T. PMOS Single-Poly Non-Volatile Memory Structure. U.S. Patent 5,761,121, 2 June 1998.
13. Na, K.-Y.; Kim, Y.-S. Novel Single Polysilicon EEPROM Cell With Dual Work Function Floating Gate. *IEEE Electron Device Lett.* **2007**, *28*, 151–153. [CrossRef]
14. Choi, J.-M.; Park, E.-J.; Woo, J.-J.; Kwon, K.-W. A Highly Linear Neuromorphic Synaptic Device Based on Regulated Charge Trap/Detrap. *IEEE Electron Device Lett.* **2019**, *40*, 1848–1851. [CrossRef]
15. Hwang, J.Y.; Kwon, K.W. A Non-linear Input Converter Inversely Pre-distorted Against Nonlinear Behavior of FG-based Neuromorphic Synaptic Devices. In Proceedings of the 2021 18th International SoC Design Conference (ISOCC), Jeju, Korea, 6–9 October 2021. [CrossRef]