4

Using FeFETs as Resistive Synapses in Crossbar-based Analog MAC Accelerating Units

Lei Zhang[1,2], David Borggreve[1], Frank Vanselow[1], Ralf Brederlow[2]

[1] Fraunhofer EMFT, Germany
[2] Technical University of Munich, Germany

Abstract

Emerging non-volatile memories (eNVMs) face problems such as insufficient R_{OFF}/R_{ON}-ratio and limited memory operating window that significantly deteriorate the precision of multiply-accumulate computations (MACs), the core computation of artificial intelligence algorithms, using crossbar-based analogue resistive compute-in-memory (CIM) structures. Properly selecting between single-ended and pseudo-differential structures is the fundamental for the most efficient use of the advantages of a particular eNVM, where, e.g., ferroelectric field-effect-transistors (FeFETs) have a large R_{OFF}/R_{ON}-ratio as a great advantage but a significant variability between devices due to the current technology maturity. By investigating and modelling both structures, the results demonstrate that the pseudo-differential structure requires a larger combined operating window from eNVM cells. The reason relies on a statistically enlarged state variation with an increasing number of input channels in the pseudo-differential structure, while the difference between the means of memory's state distributions remains unchanged. Compared to pseudo-differential structures, single-ended structures require a much higher R_{OFF}/R_{ON}-ratio from resistance-switching memories, while the requirement for process variation can be relaxed. The results indicate that FeFETs can be well suited to single-ended crossbar-based structures. However, the considerable state variation of FeFETs makes the applications of FeFETs as resistive synapses hard suited into practice. After investigating existing methods, a gate-cascaded synapse with a higher R_{OFF}/R_{ON}-ratio and a significantly enlarged operating window is proposed. This article discusses boundary conditions for using eNVMs such as FeFETs in crossbar-based analogue MAC accelerating units from a circuit design perspective.

Keywords: Ferroelectric Field-Effect Transistor, Compute-In-Memory, Resistive Synapses, Multiply-Accumulate Computations, Analog MAC Accelerator, Dot-Product Accelerator, Emerging Non-Volatile Memories, Crossbar

4.1 Introduction and Background

Emerging non-volatile memory-based CIM is attracting widespread interest in the field of integrated circuit (IC) design on account of its great potential for enabling a highly parallel analogue (or multi-bits) computation to accelerate MACs in artificial intelligence algorithms sharply [1]. The FeFET, one of the eNVMs, has been studied and implemented for accelerating MACs using its programmable switching property [2–5], where its threshold voltage can be programmed by adapting the polarization of the ferroelectric layer on the top of the transistor's gate, as illustrated in Figure 4.1. Like using other resistance-switching eNVMs (e.g., ReRAM, OxRAM) for crossbar-based MAC accelerators, FeFETs must fulfil requirements such as a reasonably large R_{OFF}/R_{ON}-ratio, and a sufficient operating window to allow an analogue (multi-bits) computation [6].

Unfortunately, according to the current technology maturity but also the fact that the techniques with smaller sizing dimensions have often a more significant variation, using eNVMs with a minimal size likes FeFETs in a resistive crossbar-based accelerator must face a considerable process variation as shown in Figure 4.1(c), which causes an insufficient operating window, and consequently, leads to an unpromising inference computation precision. Due to this fact, implementing accelerators with either binary states (On or Off) [7] or few bits [8] become an intermediate step towards to fully analogue computation, and significant power efficiencies of 532 TOP/W and over 10000TOPS/W for binary computations are achieved for particular use cases, respectively. However, further improving the efficiency and accuracy of FeFET-based accelerators requires knowledges of the fundamental design challenges of crossbar-based MAC accelerators.

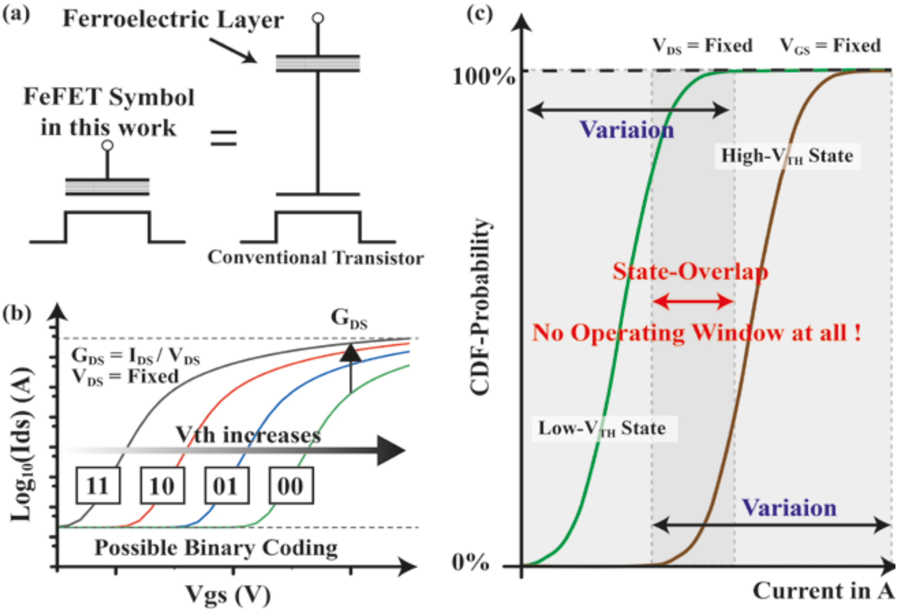

Figure 4.1 (a) shows FeFETs' abstract structure (modified from [9]), where a ferroelectric layer is placed at the top of the transistor's gate. The threshold voltage of FeFETs can be programmed by adapting the polarity of the ferroelectric layer and coded as shown in (b). (c) illustrates possible cumulative distribution functions (CDFs) of real FeFET's current in High-/Low-V_{TH} states, where a state-overlap happens, and the operating window vanishes.

This article shows how to select an optimal structure out of single-ended and pseudo-differential read-out schemes for a particular eNVM. Furthermore, this article discusses the scenarios using FeFETs for two above mentioned structures and how to deal with a limited FeFET's operating window in synapse design.

4.2 Requirements of Crossbar Structure on eNVMs

Figure 4.2 shows single-ended (a) and pseudo-differential (b) structures of analogue crossbar-based MAC accelerating units, where the resistive synapses in the pseudo-differential structure is realized by two resistance-switching devices with oppositely programmed states instead of utilizing a single device in single-ended structure. The computations performed using the single-ended, and pseudo-differential structures can be written as

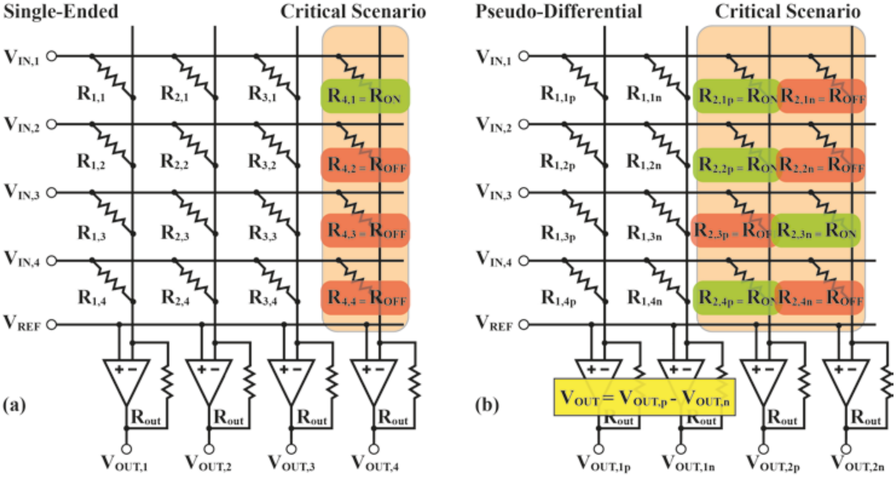

Figure 4.2 Implementations of analogue MAC accelerating units using single-ended (a) and pseudo-differential (b) structures are shown.

Equations (4.1a) and (4.1b), respectively.

$$V_{OUT,\ x} = \sum_{y=0}^{\infty} (V_{REF} - V_{IN,x}) \cdot \frac{R_{out}}{R_{x,\ y}}, \tag{4.1a}$$

$$V_{OUT,\ x} = \sum_{y=0}^{\infty} (V_{REF} - V_{IN,x}) \cdot R_{out} \cdot (\frac{1}{R_{x,\ y_p}} - \frac{1}{R_{x,\ y_n}}), \tag{4.1b}$$

where x and y represent the output channel and input channel as shown in Figure 4.2. The critical scenario happens for single-ended structures if only one resistive synapse is programmed to low-resistance state (LRS) R_{ON} and the others are programmed to high-resistance state (HRS) R_{OFF}. In order to ensure that the output voltage is still can be distinguished, the following condition has to be met:

$$(N-1) \cdot (V_{IN,max} - V_{REF}) \, / \, R_{OFF} \ll (V_{IN,max} - V_{REF}) \, / \, R_{ON}, \tag{4.2}$$

where N is the total number of input channels. The total on- and off-currents in the single-ended structure can be represented by Equations (4.3a) and (4.3b).

$$I_{ON(a)} = (V_{IN,max} - V_{REF}) \, / \, R_{ON} \tag{4.3a}$$

$$I_{OFF(a)} = (N-1) \cdot (V_{IN,max} - V_{REF}) \, / \, R_{OFF} \tag{4.3b}$$

By simplifying Equation (1.2), the final requirement for on-/off-resistance can be given:

$$\frac{R_{\mathrm{OFF}}}{R_{ON}} \gg (N-1)\,. \tag{4.4}$$

The pseudo-differential structure has a different critical scenario, where (N/2 + 1) synapses are positively programmed, and others are negatively programmed. Its total on-current $\mathrm{I}_{\mathrm{ON(b)}}$, and off-current $\mathrm{I}_{\mathrm{OFF(b)}}$ can be represented as:

$$I_{ON(b)} = (V_{IN,max} - \ V_{REF}) \cdot \{(N/2\ +1)/R_{ON} + (N/2-1)/R_{OFF}\} \tag{4.5a}$$

$$I_{OFF(b)} = (V_{IN,max} - \ V_{REF}) \cdot \{(N/2\ +1)/R_{OFF} + (N/2-1)/R_{ON}\} \tag{4.5b}$$

The required on-/off-resistance can be derived from required on-/off-current as

$$I_{OFF(b)} \ll \ I_{ON(b)}, \tag{4.6}$$

so that

$$\frac{R_{\mathrm{OFF}}}{R_{ON}} \gg 1. \tag{4.7}$$

Equation (4.4) and (4.7) indicate that the pseudo-differential structure has a much relaxing requirement on the $\mathrm{R}_{\mathrm{OFF}}/\mathrm{R}_{\mathrm{ON}}$-ratio. However, the process variation can more easily make the computation with the pseudo-differential structure fail.

Considering the resistance variation of eNVMs as

$$X_{ON} {\sim} \mathbb{N}\left(\mu_{ON},\ \sigma_{ON}^2\right) \text{ and } X_{OFF} {\sim} \mathbb{N}(\mu_{OFF}, \sigma_{OFF}^2), \tag{4.8}$$

and assuming that the resistance variation is independent from devices to devices (joint normally distributed), the distributions of the total resistance for on-/off-current in the single-ended structure can be written as

$$Y_{RON(a)} {\sim} \mathbb{N}(\mu_{ON},\ \sigma_{ON}^2) \tag{4.9a}$$

$$Y_{ROFF(a)} {\sim}\ \mathbb{N}(\mu_{OFF}/\,(N-1)\,, \sigma_{OFF}^2/(N-1)^2). \tag{4.9b}$$

Considering the 3s-variation of eNVMs and assuming no existing state overlap, the relationship between the distribution of total on-/off-resistance can be written as

$$\mu_{ON} + 3 \cdot\ \sigma_{ON} \ll\ \mu_{OFF} - 3 \cdot\ \sigma_{\mathrm{OFF}}. \tag{4.10}$$

For a successful computation, the total resistance for on-/off-current in the single-ended structure must fulfil the relationship as expressed following:

$$\mu_{ON} + 3 \cdot \ \sigma_{ON} \ll \ \mu_{OFF} / \left(N - 1 \right) - 3 \cdot \sigma_{OFF} / \left(N - 1 \right). \tag{4.11}$$

It is obvious that the condition of Equation (4.11) will be met if Equations (4.4) and (4.10) can be simultaneously fulfilled. In terms of the process variation, the pseudo-differential structure faces a more serious situation. Deriving a concrete analytical solution for the distribution of total on-/off-resistance in pseudo-differential structures requires lots of efforts, however, still their rough relationship can be checked by making following assumptions:

$$\sigma_{ON}/\mu_{ON} = \ \sigma_{OFF}/\mu_{OFF} \tag{4.12}$$

$$b = \ R_{OFF}/R_{ON} = \ {}_{OFF}/\mu_{ON}. \tag{4.13}$$

Then, the distributions of total on-/off-resistance can be written as

$$Y_{RON(b)} {\sim} \mathbb{N} \left(\frac{b \cdot \mu_{ON}}{b \cdot (N/2+1) + (N/2-1)}, \left(\frac{b \cdot \sigma_{ON}}{b \cdot (N/2+1) + (N/2-1)} \right)^2 \right), \tag{4.14}$$

$$Y_{ROFF(b)} {\sim} \mathbb{N} \left(\frac{b \cdot \mu_{ON}}{b \cdot (N/2-1) + (N/2+1)}, \left(\frac{b \cdot \sigma_{ON}}{b \cdot (N/2-1) + (N/2+1)} \right)^2 \right). \tag{4.15}$$

A similar condition likes Equation(1.11) for the pseudo-differential structure can be written as

$$\begin{aligned} &\left(\frac{b \cdot \mu_{ON}}{b \cdot (N/2-1) + (N/2+1)} - \frac{b \cdot \mu_{ON}}{b \cdot (N/2+1) + (N/2-1)} \right) \\ &\gg 3 \cdot \left(\frac{b \cdot \sigma_{ON}}{b \cdot (N/2+1) + (N/2-1)} + \frac{b \cdot \sigma_{ON}}{b \cdot (N/2-1) + (N/2+1)} \right). \end{aligned} \tag{4.16}$$

By simplifying Equation (4.16), the condition for the pseudo-differential structure can be finally expressed as

$$\sigma_{ON} \ll \frac{2 \cdot (b-1)}{3 \cdot N \cdot (b+1)} \cdot \mu_{ON}. \tag{4.17}$$

Equation (4.17) indicates that increasing the input channels requires reducing the device process variation to keep computation precision unchanged even if the R_{OFF}/R_{ON} is sufficiently large. For easy comparison, Equation (4.10) can be re-written using same assumptions as

$$\sigma_{ON} \ll \frac{(b - N + 1)}{3 \cdot (b + N - 1)} \cdot \mu_{ON}. \tag{4.18}$$

Note that conclusions made from Equations (4.17) and (4.18) are based on some very optimistic assumptions like Equations (4.12) and (4.13) that may vary from the reality. To verify those conclusions, a numerical analysis for the total on-/off-current in both structures is made, and the result is shown in Figure 4.3. This result identifies the above mathematical derivation that increasing R_{OFF}/R_{ON} yields a better computation precision in single-ended structures even if the process variation is significant. For the pseudo-differential structure, ensuring that the device has less variation is the precondition for a good computation precision instead of seeking for a large R_{OFF}/R_{ON}. The requirements given by single-ended and pseudo-differential structures on eNVMs are listed in Table 4.1.

FeFETs have a very high R_{OFF}/R_{ON}-ratio because their switching property is as the same as conventional transistors, but also suffer from the significant process variation due to the current technology maturity. According to those properties, the single-ended structure is a better fit for the design with FeFETs. However, simply using FeFETs in a single-ended structure can still deteriorate computation precision since the state overlap exists, as shown

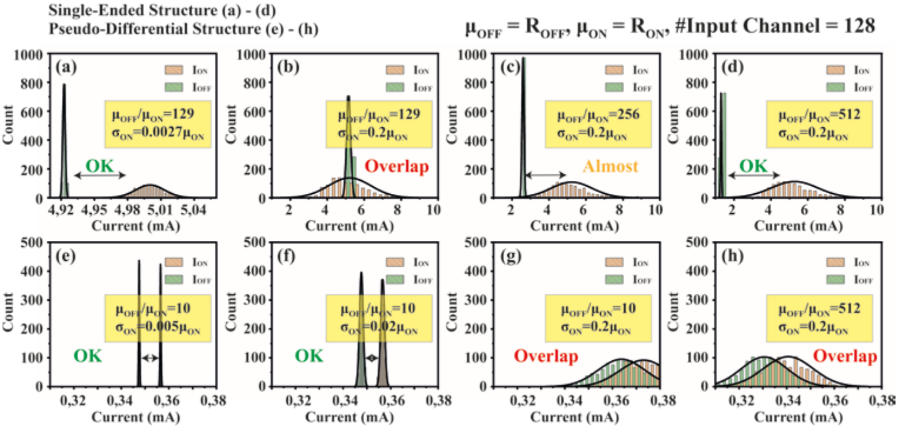

Figure 4.3 The numerical analysis indicates that the R_{OFF}/R_{ON} plays a dominant role for the computation precision in the single-ended structure, where the inherent device process variation is more important for the pseudo-differential structure.

Table 4.1 Comparison between single-ended and pseudo-differential structures

Requirements	Single-Ended	Pseudo-Differential
R_{OFF}/R_{ON}	$\gg (N-1)$	$\gg 1$
Process variation	$\sigma_{ON} \ll \frac{(b-N+1)}{3\cdot(b+N-1)} \cdot \mu_{ON}$	$\sigma_{ON} \ll \frac{2\cdot(b-1)}{3\cdot N\cdot(b+1)} \cdot \mu_{ON}.$
Area	Small	Large
FeFET	suitable	Not suitable

in Figure 4.1 (c). In order to prevent computational precision loss, a proper synapse design should be derived.

4.3 Synapse Design

A good synapse design should have a large R_{OFF}/R_{ON}-ratio and a sufficient operating window without a huge area overhead. This chapter reviews the existing circuit techniques, which could be applied to the synapse design, proposes a gate-cascade technique for improving the synapse's operating window, and shows achieving a better trade-off by combining various circuit techniques.

4.3.1 Conventional Design

Figure 4.4 (a) shows the simplest FeFET synapse that consists of a single FeFET M_{F1} and an access transistor M_a. Its characteristic is the same as conventional transistors' but with an adjustable threshold voltage. However, a slight operating-points shift, or the process variation can result in noticeable current changes for different states, as demonstrated by the case (a) in

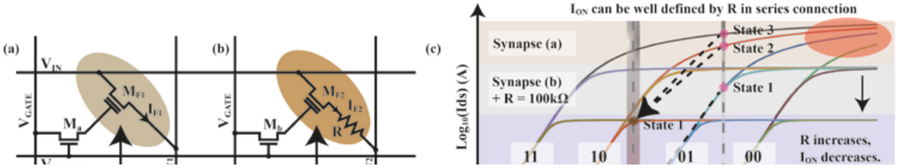

Figure 4.4 Two conventional FeFET synapses are shown, where synapse (b) has an additional current-limiting resistor in the series connection compared to the stand-alone FeFET synapse (a). Both synapses can be activated by connecting a certain gate-voltage using access transistors M_a and M_b, respectively. (c) shows the characteristics of synapses (a) and (b), where a large series resistor enlarges the threshold voltage range of individual states by scarifying the number of available states.

Figure 4 (c). Adding a resistor in series with the FeFET, as demonstrated in Figure 4.4 (b), results in a well-defined on-current, which can be estimated using the linearized transistor equation in the triode-region as following:

$$I_{F2} = \left(\frac{\partial I_{DS}}{\partial V_{DS}} + \frac{1}{R} \right) (V_{OUT2} - V_{IN}) \tag{4.20a}$$

so that

$$I_{F2} = (K_n(V_{GS} - V_{TH} - V_{DS} + 1/R))V_{DS}, \tag{4.20b}$$

where $\mathrm{K_n}$ and $\mathrm{V_{TH}}$ are transistors' transconductance parameter, and threshold voltage. $\mathrm{V_{GS}}$, $\mathrm{V_{DS}}$ are $\mathrm{V_{GATE}}$ and ($\mathrm{V_{OUT2}}$ – $\mathrm{V_{IN}}$) in Figure 4(b), respectively. Considering conditions of:

$$K_n \left(V_{GS} - V_{TH} - V_{DS} \right) > 0 \tag{4.21a}$$

$$K_n \left(V_{GS} - V_{TH} - V_{DS} \right) \ll 1/R, \tag{4.21b}$$

The on-current of synapse (b) is well-defined as

$$I_{F2} \approx \frac{V_{OUT2} - V_{IN}}{R}. \tag{4.22}$$

Additionally, the synapse (b)'s characteristic in the saturation region remains the same as conventional transistors. However, FeFET enters earlier into the triode-region depending on the resistor's value because the drain-source voltage is reduced by voltage drop over the resistor, as revealed by the case (b) in Figure 4.4(c).

Compared to the synapse (a), synapse (b) has a higher robustness against the operating-point shifts since it defines the on-current better. After reducing four states of the case (a) to the case (b) with only two states, the impact of the process variation the on-current is reduced, where the threshold voltage's variation between state 11 and 10 (01 and 00) always results in a well-defined on(off)-current if $\mathrm{V_{GS}}$ is selected between transfer curves of state 10 and 01, as illustrated in Figure 4.4(c). Nevertheless, if the process variation is as significant as shown in Figure 4.1(c), the state overlap cannot be eliminated using synapse (a) and (b) with a single FeFET. A conventional way to yield more stable devices against process variation is connecting multiple FeFETs in series to form a relatively larger FeFET, where a large area overhead may be caused by a large amount of FeFETs in series needed.

4.3.2 Gate-Cascaded FeFETs

Inspired by analysis for single-ended and pseudo-differential structures, the thought was made for enlarging the FeFET's operating window faster than devices variation. Figure 4.5(a) shows a possible implementation, the gate cascaded FeFET. A tiny leakage current I_1 flows through M_1 when gate voltage V_G increases. It enables that the voltage V_X rises with V_G, and correspondingly, FeFET M_2 is turned-on by rising V_X. Due to the diode-connection of M_1 and a very tiny drain-source current I_1, M_1 conducts in the sub-threshold region, and I_1 can be expressed as

$$I_1 = I_s\exp(2 - \frac{V_{TH1}}{nU_T})\exp(\frac{V_{GS1}}{nU_T}), \tag{4.23}$$

where n and I_S are the process-dependent sub-threshold factor and specific current, respectively. U_T represents the thermal voltage, V_GS1 and V_TH1 denote the gate-source voltage and threshold voltage of M_1. By solving Equation (4.23), V_GS1 can be written as

$$V_{GS1} = V_{TH1} + \ln\left(\frac{I_1}{I_s}\right) nU_T - 2nU_T = V_{TH1} + V_C, \tag{4.24}$$

where $\mathrm{V}_\mathrm{C}$represents the sum of the second and third terms of Equation (4.24). Because only I_1 changes very slightly and any other parameters are process-specific, $\mathrm{V}_\mathrm{C}$is approximately constant. Therefore, V_GS2-V_TH2 can be written as

$$V_{GS2} - V_{TH2} = V_G - V_C - 2V_{TH0} - \triangle V_{TH1} - \triangle V_{TH2}, \tag{4.25}$$

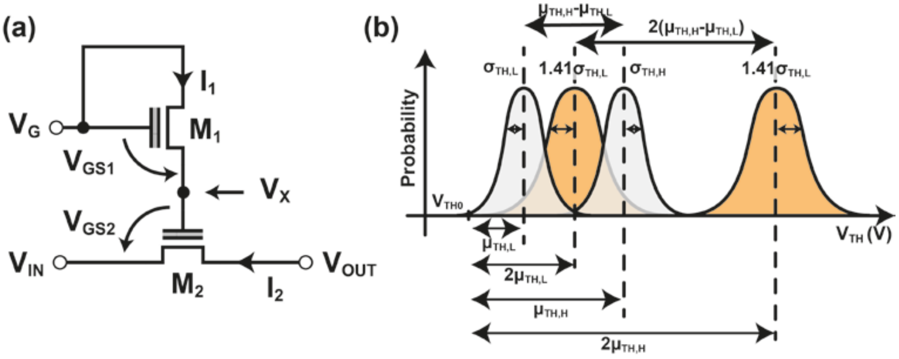

Figure 4.5 The proposed gate-cascaded FeFET synapse, where a diode-connecting FeFET is connected to the gate of another FeFET, is shown in (a). Its statistical distribution is shown in (b), that the distance between threshold voltages doubles and the variation of the state overlap.

where V_{TH0} is the threshold voltage of conventional transistors, ΔV_{TH1} and ΔV_{TH2} represent the threshold voltage changes applied to conventional transistors by the ferroelectric layer. If both FeFETs are simultaneously programmed to the same state (ΔV_{TH1} = ΔV_{TH2}), the operating window V_{Δ}, which is defined as voltage difference $\Delta(V_{GS2}$ - $V_{TH2})$ between high- and low threshold voltage state (HVT and LVT), can be written as

$$V_{\triangle} = \triangle\left(V_{GS2} - V_{TH2}\right) = 2(\triangle V_{TH,HVT} - \triangle V_{TH,LVT}), \tag{4.26}$$

where is twice as conventional synapses.

Considering that the ΔV_{TH}-variation has approximately a normal distribution, the distribution of ΔV_{TH1} and ΔV_{TH2} are referred to X and Y, where

$$X{\sim}\mathbb{N}\left(\mu_{\triangle TH1},\ \sigma^2_{\triangle TH1}\right) \tag{4.27a}$$

$$Y{\sim}\mathbb{N}\left(\mu_{\triangle TH2},\ \sigma^2_{\triangle TH2}\right). \tag{4.27b}$$

Two FeFETs in a circuit should have identical distributions, and they are independent of each other, which means that they are jointly normal. The distribution U of (ΔV_{TH1} + ΔV_{TH2}) with (ΔV_{TH1} = ΔV_{TH2}) can be written as

$$U = X + Y \tag{4.28a}$$

$$U{\sim}\mathbb{N}\left(2\mu_{\triangle TH},\ 2\sigma^2_{\triangle TH}\right) \tag{4.28b}$$

with

$$\mu_{TH} = \mu_{TH1} = \mu_{TH2} \tag{4.29a}$$

$$\sigma_{TH} = \sigma_{TH1} = \sigma_{TH2}. \tag{4.29b}$$

Assuming a 3s-variation, the operating window for the conventional synapse $V_{\Delta conv}$ and the gate-cascaded FeFET $V_{\Delta prop}$ can be derived as

$$V_{\triangle conv} = (\mu_{HVT} - \mu_{LVT}) - 6(\sigma_{HVT} - \sigma_{HVT}) \tag{4.30a}$$

$$V_{\triangle prop} = (\mu_{HVT} - \mu_{LVT}) - 3\sqrt{2}(\sigma_{HVT} - \sigma_{HVT}). \tag{4.30b}$$

If no overlap between two states is expected, the operating window must be positive (V_{Δ}>0). The conventional synapse (CONV.) and the gate-cascaded synapses (PROP.) operate correctly if the following conditions are fulfilled.

$$CONV.: (\mu_{HVT} - \mu_{LVT}) > 6(\sigma_{HVT} - \sigma_{HVT}) \tag{4.31a}$$

$$PROP.: (\mu_{HVT} - \mu_{LVT}) > 3\sqrt{2}(\sigma_{HVT} - \sigma_{HVT}) \tag{4.31b}$$

According to Equations (4.31a) and (4.31b), synapses with gate-cascaded FeFETs has a 1.4 times relaxed requirement for the process variation compared to the conventional synapses, which is shown in Figure 4.5(b). Extending a single gate-cascaded synapse to N-stage gate-cascaded synapses (N>0), as shown in Figure 4.6(a) and (b), the improvement A can be written as:

$$A = \ (N+1)/\sqrt{N+1}. \tag{4.32}$$

The derivative of the improvement can be written as

$$A' = \frac{1}{\sqrt{(4N+4)}}, \tag{4.33}$$

which indicates that the improvement of the operating window slows down with a continuously increasing number of gate-cascaded stages.

4.3.3 Exploration Results

Figure 4.6 (c) shows the drain-source current curve of FeFETs without gate-cascade, with one- and two-stage gate-cascade. By increasing the number of gate-cascaded stages, the operating window, and the voltage difference between states are enhanced with the same gate voltage V_G. Figure 4.6(d) compares the conventional synapses with 3 FeFETs in series and with 3-stage gate-cascaded FeFETs. The conventional design has a slightly improved operating window compared to a single FeFET that has no operating window at all. The design with gate-cascaded FeFETs has an operating window up to 12.1 times larger than the operating window with 3 FeFETs in series. The I_{ON}/I_{OFF}-ratio, which is exactly equal to R_{OFF}/R_{ON}-ratio, and the operating window are enhanced approximately 2.67 times and 12.1 times compared to the conventional design, respectively.

Table 4.2 Relative Performance Comparison

	Single FeFET	3 FeFETs in series	3-Stage Gate-Cascaded FeFET
# of FeFETs	1	3	3
I_{ON}/I_{OFF}	N/A	α	2.67 α
I_{ON}/I_{OFF} with process variation	N/A	β	26900 β
Operating Window	<0	γ	12.1 γ

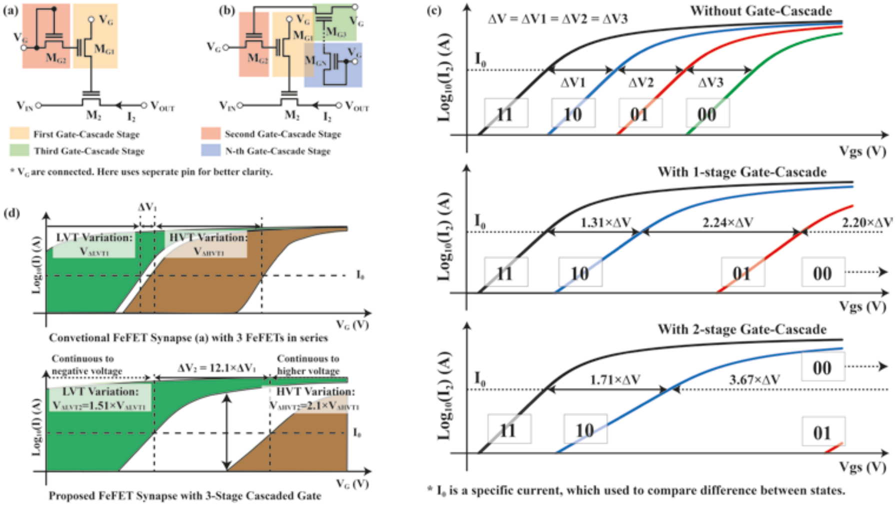

Figure 4.6 (a) and (b) show a two-stage and a N-stage gate-cascaded FeFET synapse, respectively. (c) shows the change of their characteristics, where the voltage difference between states is enlarged. (d) demonstrate the characteristic of a conventional synapse with three serially connected FeFET and a three-stage gate cascaded FeFET. The gate-cascaded FeFET achieved 12.1 times larger operating window than conventional design.

Nevertheless, the drawbacks that the gate-cascaded FeFET brings need to be pointed out:

1. Programming FeFETs requires a particularly high voltage applied to FeFETs. The more gate-cascaded stages are used, the more access high-voltage transistors are required, which occupy the most area in the design as the design example shown in Figure 4.7(a) and (b).
2. Shifting threshold voltage to a very high value does not yield much. On the one hand, the improvement slows down with an increasing number of cascaded stages, according to Equation (4.33). On the other hand, the real gate voltage cannot achieve a very high potential. An optimal number of stages is highly technology dependent.

Since drawbacks listed above, combing different methods in a right manner will result into an optimal design point. A design example is shown in Figure 4.7(a), where

- $M_{3,L}$and $M_{3,R}$ play the role of resistors to limit the current,
- M_1 and M_2 are serially connected FeFETs for reducing the process variation slightly,

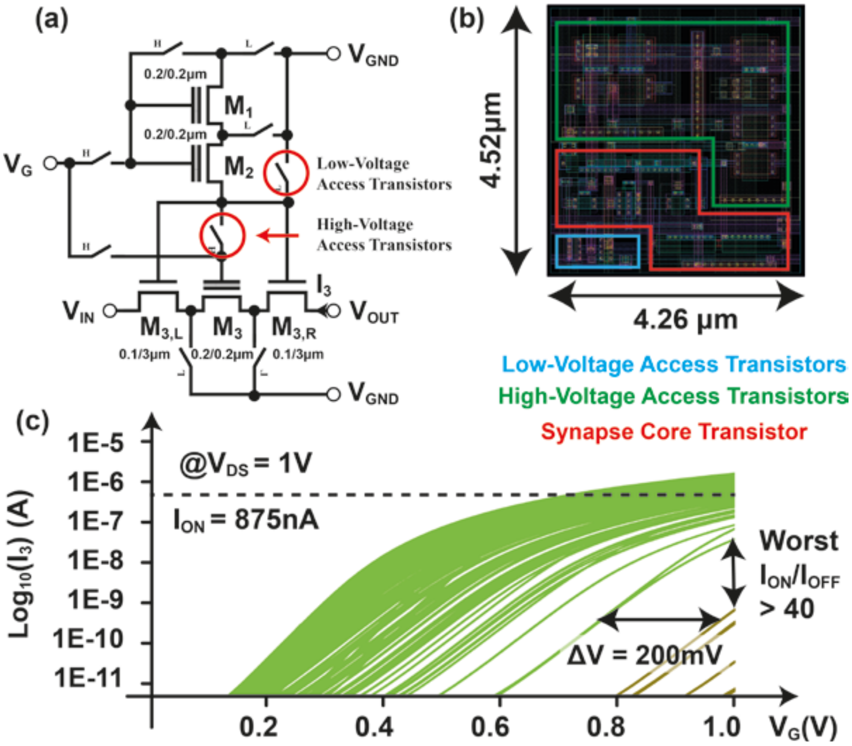

Figure 4.7 design example, which combine the proposed and conventional techniques, is shown in (a). (b) displays the layout of this design example. (c) indicates that a up to 200mV operating window is achieved using 1-stage gate-cascade.

- M_1, M_2 and M_3 build a one-stage gate cascade for generally enhancing operating window.

The performance of the design example is shown in Figure 4.7(c). Depending on the need, the operating window can further be enhanced by either connecting more FeFETs in series or applying more gate-cascaded stages.

4.4 Conclusion

This article mainly reviews circuit aspects, such as select of the best readout structure and the design of resistive synapses, for using FeFETs in a crossbar-based analogue MAC accelerating unit. Both analytical and numerical analyses indicate that FeFETs have a better fit to the single-ended structure, which requires a high R_{OFF}/R_{ON}-ratio but has relaxing requirements for the process variation. Those considerations can be transferred to other types of eNVMs. Furthermore, this article compares three ways of using FeFETs as resistive synapses, and the result indicates that only combining different methods can lead into a high R_{OFF}/R_{ON}-ratio and non-overlapped states without a significant area overhead. For implementing an entire accelerator, other design aspects, such as programming algorithms, parasitic effects, design of efficient data-converter and so on, need to be considered. However, this article gives readers an essential guidance how to start using FeFETs or other eNVMs, for crossbar-based analogue MAC accelerators.

Acknowledgements

This work is supported through the project TEMPO. TEMPO has received funding from the ECSEL Joint Undertaking (JU) under grant No.826655. The JU receives support from the European Union's (EU's) Horizon 2020 research and innovation program and France, Belgium, Germany, Netherlands, Portugal, Spain, Switzerland. TEMPO has also received funding from the German Federal Ministry of Education and Research (BMBF) under Grant No.16ESE0407. This work reflects only the author's view, JU, EU and BMBF are not responsible for any use that may be made of the information it contains.

References

[1] W. Haensch, T. Gokmen, and R. Puri, "The next generation of deep learning hardware: Analog Computing", *Proceeding of the IEEE*, vol. 107, no. 1, pp. 108-122, 2019.

[2] M. Jerry, P.-Y. Chen, J. Zhang, P. Sharma, K. Ni, S. Yu and S. Datta, "Ferroelectric fet analog synapse for acceleration of deep neural network training", in *2017 IEEE International Electron Devices Meeting (IEDM)*, pp. 6.2.1-6.2.4, 2017.

[3] M. Jerry, S. Dutta, A. Kazemi, K. Ni, J. Zhang, P.-Y. Chen, P. Sharma, S. Yu, X. S. Hu, M. Niemier, and S. Datta, "A ferroelectric field effect transistor based synapse weight cell", *Journal of Physics D: Applied Physics*, vol. 51, no. 43, august 2018.

[4] S. Oh, H. Hwang, and I. K. Yoo, "Ferroelectric materials for neuromorphic computing", *APL Materials*, vol. 7, no. 9, p.091-109, 2019.

[5] N. E. Miller, Z. Wang, S. Dash, A. I. Khan, and S. Mukhopadhyay, “Characterization of drain current variations in fefets for pim-based dnn accelerator”, in *2021 IEEE 3rd International Conference on Artificial Intelligence Circuits and Systems (AICAS)*, pp. 1-4, 2021.

[6] P.-Y. Chen, X. Peng, and S. Yu, “NeuroSim+: An integrated device-to-algorithm framework for benchmarking synaptic devices and array architectures,” in *2017 IEEE International Electron Devices Meeting (IEDM)*, pp.6.1.1-6.1.4, 2017.

[7] D. Bankman, L. Yang, B. Moons, M. Verhelst, and B. Murmann, “An always-on 3.8µJ 86% cifar-10 mixed-signal binary cnn processor with all memory on chip in 28-nm process”, *IEEE Journal of Solid-State Circuits*, vol. 54, no. 1, pp. 158-172, 2019.

[8] T. Soliman, F. Müller, T. Kirchner, T. Hoffmann, H. Ganem, E. Karimov, T. Ali, M. Lederer, C. Sudarshan, T. Kämpfe, A. Guntoro, and N. Wehn, “Ultra-low power flexible precision fefet based analog in-memory computing”, in 2020 *IEEE International Electron Devices Meeting (IEDM)*, pp. 29.2.1-29.2.4, 2020.

[9] A. Aziz, E. T. Breyer, A. Chen, X. Chen, S. Datta, S. K. Gupta, M. Hoff-mann, X. S. Hu, A. Ionescu, M. Jerry, T. Mikolajick, H. Mulaosmanovic,K. Ni, M. Niemier, I. O’Connor, A. Saha, S. Slesazeck, S. K. Thirumala,and X. Yin, “Computing with ferroelectric fets: Devices, models, systems,and applications,” in *2018 Design, Automation, Test in Europe Conference Exhibition (DATE)*, pp. 1289–1298, 2018.