Ferroelectrics: From Memory to Computing

Kai Ni

Department of Microsystems Engineering
Rochester Institute of Technology
Rochester, NY 14623
Tel : 585-475-7855
e-mail : kai.ni@rit.edu

Sourav Dutta

Department of Electrical Engineering
University of Notre Dame
Notre Dame, IN 46556
e-mail : sdutta4@nd.edu

Suman Datta

Department of Electrical Engineering
University of Notre Dame
Notre Dame, IN 46556
Tel : 574-631-8835
e-mail : sdatta@nd.edu

Abstract - Research discovery of ferroelectricity in doped hafnium dioxide thin films has ignited tremendous activity in exploration of ferroelectric FETs for a range of applications from low-power logic to embedded non-volatile memory to in-memory compute kernels. In this paper, key milestones in the evolution of Ferroelectric Field Effect Transistors (FeFETs) and the emergence of a versatile ferroelectronic platform are presented. FeFET exhibits superior energy efficiency and high performance as embedded nonvolatile memory. When embedded into logic, such as SRAM or D-flip-flop, nonvolatile processor can be designed, which is critical for intermittent computing with unreliable power. The partial polarization switching in multi-domain ferroelectric can be harnessed to develop analog synaptic weight cell for deep learning accelerators. To further improve the energy-efficiency of computation, ferroelectric in-memory computing hardware primitive is designed, with one prominent example of ferroelectric TCAM. Utilizing the ferroelectric switching dynamics, ferroelectric neuron with intrinsic homeostasis can be realized to enable a unified ferroelectric platform for spiking neural network. From all these developments, ferroelectric emerges as a highly promising platform for various exciting applications.

Key words: ***Ferroelectric, HfO_2, Nonvolatile Memory, Synaptic Weight Cell, In-Memory Computing, Neuron***

I. Introduction

The recent discovery of ferroelectricity in doped HfO_2 has ignited tremendous research activities in its integration into ferroelectric FETs (FeFET) due to its excellent CMOS-compatibility and superior scalability [1]. Meanwhile, applications of FeFET, ranging from low-power logic to embedded non-volatile memory to in-memory compute kernels, have been actively explored. As a result, ferroelectric HfO_2 has become a versatile platform for various exciting applications.

The application of ferroelectrics in high performance logic transistor is motivated by the existence of unstable negative capacitance region in the ferroelectric energy landscape [2]. It is proposed that by connecting the ferroelectric capacitor with a regular capacitor in series, the negative capacitance can be stabilized under certain conditions and the resulting capacitance is greater than individual component. With that assumption, the MOSFET subthreshold slope can break the Boltzmann limit and achieve less than 60mV/dec. This can reduce the operation voltage (hence power consumption) without sacrifice in the driving capability, hence providing a new scaling route for logic transistors. Though abundant steep slope data has been reported [3], [4], it has been shown that they are caused by transient snapback effects during ferroelectric switching, which is undesirable for logic transistors [5], [6]. Therefore, the search for stabilized negative capacitance FET is still an active ongoing work. In this paper, we will focus on the memory and computing applications and not delve into the logic applications in detail.

Unlike other types of nonvolatile memories (NVM), such as flash memory, phase change memory (PCM), resistive memory (RRAM), and spin-transfer-torque magnetic memory (STT-MRAM), etc., where the memory write processes have to be driven by a large conduction current (hence large power consumption), the write in ferroelectric memory is electric-field driven [7]. The superior energy efficiency is a huge advantage and a significant drive for the research and development of ferroelectric memory. Such activities include the successful integration of ferroelectric HfO_2 in 28nm bulk [8] and 22nm silicon-on-insulator (SOI) platform [9]. Therefore, FeFET memory is a highly promising candidate for embedded NVM applications.

For a ferroelectric film with multiple domains, the partial polarization switching, where different portions of domain distribution are switched, can be harnessed to realize multiple intermediate polarization states [10]. Those polarization states will result in different output conductance in FeFET, which can serve as synaptic weight cell to store the neural network weights. By arranging FeFETs in a pseudo-crossbar architecture, the multiply-accumulate operation can be performed in the analog domain, obviating the costly data movement in the conventional Von-Neumann system.

To further improve the computation energy-efficiency, it is imperative to move data closer to the computation. The ultimate solution would be to directly perform computation in memory. We are demonstrating one such example, ferroelectric ternary content addressable memory (TCAM) [11], [12]. Such systems exhibit superior performance and energy efficiency. In the following, device characteristics are presented, and each application are explained in detail.

II. Ferroelectric Applications

A. Embedded Nonvolatile Memory

FeFET is a highly promising candidate to replace the existing NOR flash embedded NVM due to its greatly improved performance, e.g., reduced energy consumption, high write speed, and improved endurance. Its memory operation is achieved by setting the ferroelectric polarization

978-1-7281-4123-7/20/$31.00 ©2020 IEEE

pointing towards the semiconductor channel or the gate electrode using the positive or negative gate pulses, respectively, as shown in Fig. 1(a) [13]. The two polarization states will set the FeFET to two different threshold voltages, which can store one bit of information. To date, the reported memory window (the voltage separation of the two threshold voltages) of FeFET has been around 1V-1.5V for ferroelectric HfO_2 thickness around 10 nm, good enough for the sensing circuits to differentiate between the two states.

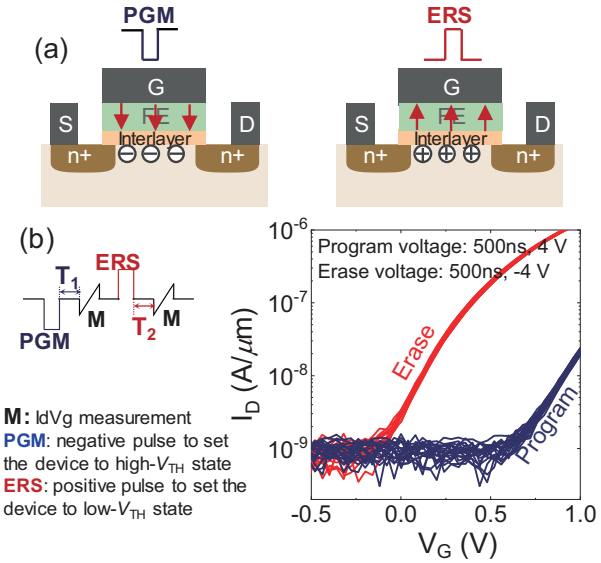

Fig.1. (a) The program/erase operations and the corresponding polarization directions in a FeFET memory device; (b) experimentally measured I_D-V_G characteristics of FeFET after program/erase pulses. Due to the existence of charge trapping induced by the write pulses, the experimentally measured memory window increases with the delay between the measurement and the write pulses due to the release of trapped charge in the gate dielectric [13].

When arranging FeFET devices into a memory array, several additional issues have to be addressed. One is the cell structure. It can be that each FeFET is associated with an additional select transistor (2T cell), which is used to access a target memory device, meanwhile not disturbing the other unselected devices [14]. This, however, scarifies the memory density. Alternatively, single FeFET memory cell (1T cell) can be applied but write disturbs to unselected cells have to be minimized [15]. Inhibition bias schemes, e.g., $V_W/2$ and $V_W/3$ (V_W is the write pulse amplitude) bias schemes, have been explored and experimentally evaluated, which shows its effectiveness in controlling the write disturb in 1T FeFET memory array [15].

Another issue with FeFET array is the write voltage polarity. Current memory operation of FeFET relies on the supply of negative pulse amplitude, which increases the design complexity of the peripheral circuits. To alleviate such restriction, FeFET operation with all positive pulses are highly desirable. For bulk FeFET, to program the device (V_{TH} is set high), a negative potential between the gate and body has to be supplied. Instead of applying negative gate pulses, positive pulses can be applied to the body terminal to program the FeFET. As such, we proposed a column wise body contact for the 1T FeFET array with all positive write pulses, which achieves a balance between the memory density and design complexity [16].

Nowadays, FeFET operation voltages has been limited to be ~4 V for ferroelectric with thickness of about 10 nm. This is due to the inefficiency of the voltage division between the ferroelectric and the underling interlayer and semiconductor layers [17], [18]. Due to the high field in the interlayer, a significant amount of charge trapping happens in the ferroelectric layer, causing degradation in the memory window and endurance [13]. This can be alleviated by using high-κ interlayer [13], but challenging to implement practically. To overcome this limitation, we proposed a novel device architecture, called ferroelectric metal FET (FeMFET) [18], where the ferroelectric capacitor is integrated at the back end and connected with a front end MOSFET through a via, shown in Fig. 2 (a). This provides an additional design freedom of optimizing the area ratio between the ferroelectric and MOSFET, and hence the capacitor divider, as illustrated in Fig.2 (b). Logic compatible operation voltage (<1.8V) has been demonstrated (Fig. 2 (c-d)). Because the ferroelectric is physically separated from the semiconductor channel, the charge trapping is eliminated, thereby showing great endurance properties.

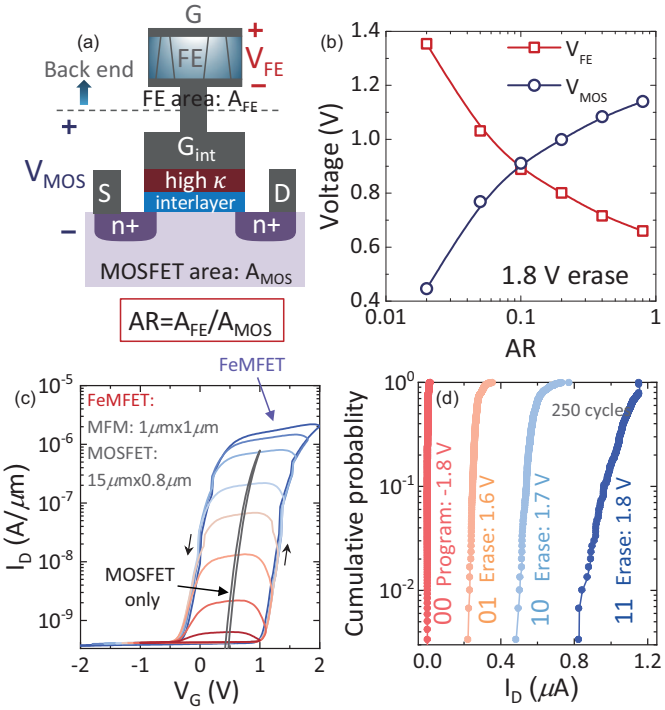

Fig. 2. (a) The ferroelectric metal FET structure. The ferroelectric capacitor is placed at the back end and connected with the gate of an underling MOSFET with a via; (b) simulated voltage drop across the ferroelectric layer can be increased by optimizing of the capacitor divider through area ratio tuning; (c) the DC I_D-V_G characteristics of FeMFET showing 1.5V hysteresis window; (d) the distribution of four levels in FeMFET with logic compatible voltage levels [18].

A lot of other exciting research activities have been conducted in FeFET memory, such as integration of ferroelectric on other important substrates, e.g., Ge [19], InGaAs [20], 2D materials [21], metal-oxide thin film [22], etc., and improving its reliability (endurance and charge

trapping [23]). Another active research area of FeFET is to enable the design of nonvolatile processor, which includes embedding FeFET in the design of SRAM and D-flip flop [24], [25], [26]. This enables the processor to quickly backup and restore its system state, without much power and latency penalty. Such capability is very important for intermittent computing in IoT edge devices with unreliable or harvested power supply, where frequent backup and restore operations are necessary.

One serious issue still remains to be solved, i.e. device-to-device variation [27]. Distribution of two V_{TH} states starts to overlap with the scaling of device area, suggesting memory window collapse. It is caused by the domain inhomogeneity and switching stochasticity, and get exacerbated by the reduced number of domains. At present, it is one of the bottlenecks to realize large FeFET memory arrays.

B. Synaptic Weight Cell for Deep Learning Accelerator

The partial polarization switching in ferroelectric can be harnessed to design synaptic weight cell for deep learning accelerators. The ideal weight cell needs to have linear and symmetric weight update characteristics, so that it relieves the burden of the peripheral circuits design. Depending on the application of the accelerator (supporting training or just inference), the bit resolution requirements of the weight cell are different, with higher resolution for the training and much relaxed resolution for the inference purposes [28].

We have explored various pulse schemes to access intermediate states, shown in Fig. 3[10]. With identical pulses, the channel conductance is observed to increase (potentiation) and decrease (depression) rapidly with applied pulse number. There are only 6 to 7 intermediate states accessible as the same pulses are applied and only a portion of domains with the coercive field (E_C) below the pulse amplitude gets switched. This can be improved by modulating the pulse width as this allows domains with longer time constant to switch with the increase of pulse number. The amplitude modulation pulse scheme produces the largest amount of intermediate states (32 states) and symmetric response. With increase of pulse number, domains with different E_C are sampled by the increasing pulse amplitude, thus exhibiting the largest amount of intermediate states. These results highlight the potential of applying FeFET as synaptic weight cell.

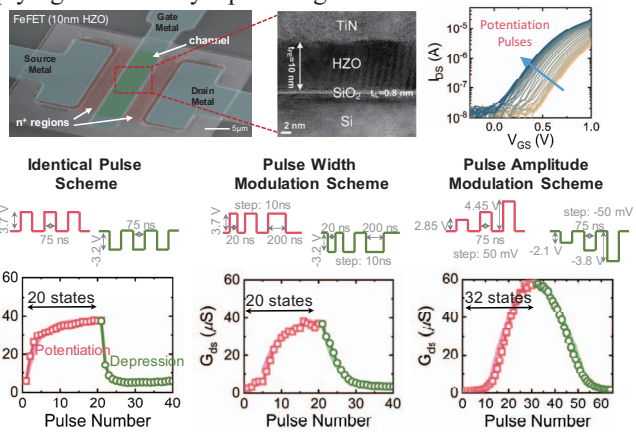

Fig. 3. Top-view SEM and cross-sectional TEM image of the FeFET for the analog synapse applications. Three pulse modulation schemes are applied to access the intermediate states (identical, pulse width modulation, and pulse amplitude modulation). Limited number of states are present for the identical pulse scheme. It is improved by the pulse width modulation scheme. Pulse amplitude schemes exhibit 5bits resolution and symmetric response. The I_D-V_G characteristics of FeFET under potentiation shows parallel shift of the device V_{TH} [10].

To further improve the bit resolution, linearity of weight update and serve the training and inference simultaneously, we proposed a novel hybrid ferroelectric weight cell, shown in Fig.3 (a) [29]. The key idea is to store the most significant bits (MSB) of the weight as the nonvolatile polarization states and the least significant bits (LSB) as the volatile gate voltage of the FeFET. The LSB is tuned by small charging/discharging pulses through the supporting PMOS/NMOS transistor, respectively. Once the gate voltage reaches the boundary defined by the current MSB, then the polarization states are switched to the next MSB level. In doing this, 6 bits per cell is demonstrated and close to ideal weight update characteristics are realized, as shown in Fig.3 (c). Another advantage of this hybrid weight cell is that for the training purposes, both the MSBs and LSBs can be used for high precision and only MSBs are used for the inference applications. Since the training need frequent weight update, the volatility of LSBs can be tolerated. At the same time, the inference needs nonvolatile weight storage, which can be supported by the polarization states. Therefore, the hybrid weight cell provides a promising solution to the deep learning accelerators.

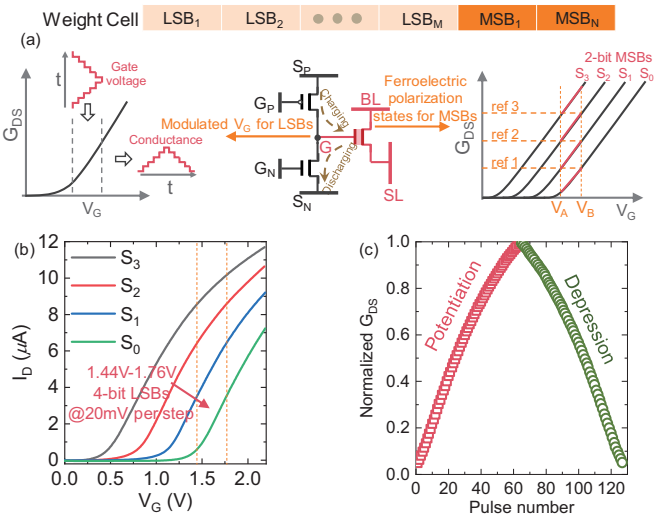

Fig. 1. (a) Hybrid weight cell based on FeFET. The weight cell is composed of one PMOS and one NMOS transistor to charge/discharge the gate voltage of FeFET, which serves as the LSBs of the weight. The polarization states store the MSBs of the weight; (b) Four V_{TH} levels correspond to 2 bits of MSBs. The blue shaded region is the range where LSBs are defined; (c) The simulated conductance tuning characteristics show almost ideal characteristics [29].

C. In-Memory Computing Kernel

To improve the computation performance and energy efficiency, it is critical to move data storage location closer to the computing unit. One solution would be to directly perform computation within the memory itself, thus eliminating the

need for data transfer. We have implemented several such computing systems using FeFET. One such system is the ferroelectric ternary content addressable memory (TCAM) [11], [12]. In TCAM, memory information is the input and the memory location which stores the matched information is returned as the output. It has been widely applied in high-performance network routers for fast package forwarding. We have demonstrated, to date, the most compact TCAM cell using two FeFETs, as shown in Fig. 2. When the stored information and the query matches, both FeFETs have very small leakage current to discharge the match line (ML). However, when the mismatch happens, a large discharge current flows through the low-V_{TH} FeFET and discharges the ML. Therefore, by sensing the discharge current through the ML, the TCAM cell function can be realized.

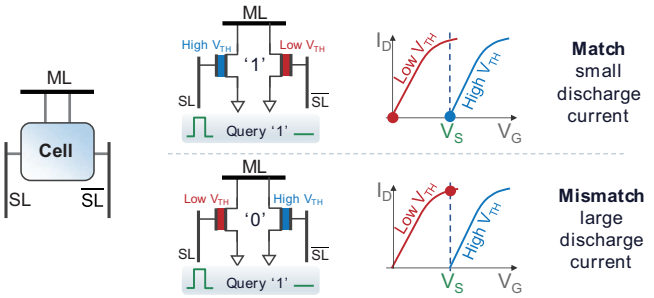

Fig. 2. An ultra-compact TCAM cell composed of two FeFETs. By encoding the stored information as the configuration of the two FeFET V_{TH} and the query information as the search pulses, two FeFETs can form a TCAM cell. When the query information matches the stored information, negligible leakage current flow through both FeFETs; while when mismatch happens, a large discharge current flows through the low-V_{TH} FeFET to discharge the match line.

When arranging multiple TCAM cells into a word (contains N cells sharing the same ML), the total discharge current through the ML would be proportional to the number of mismatched bits. We have demonstrated such functionality in a small 1x6 prototype TCAM array, showing the successful detection of the degree of mismatch between the stored information and query (Fig. 3) [12]. This capability of detecting the degree of match directly on the ML at the memory location greatly reduces the energy (60x) and latency (2700x) of the memory search operation compared with the GPU approach. We have applied this kernel for the exciting application of one-shot learning, where neural network learns with only one example per class, and achieved almost iso-accuracy compared with the GPU approach [12]. Therefore, ferroelectric TCAM is a highly promising candidate for such applications.

Other types of in-memory computing systems have also been proposed, for example one that can perform logic and arithmetic operations directly in FeFET memory, e.g., NOT, AND, etc. [30]. Another representative example is to harness the physical dynamics of polarization switching in ferroelectric for computation [31]. Under the excitation of identical pulses, the ferroelectric polarization will accumulate. Such accumulation dynamics have been applied to realize a hardware primitive for statistical correlation detection, which is widely used in signal processing and event detection. Such computational memory is highly promising to achieve superior performance and energy efficiency.

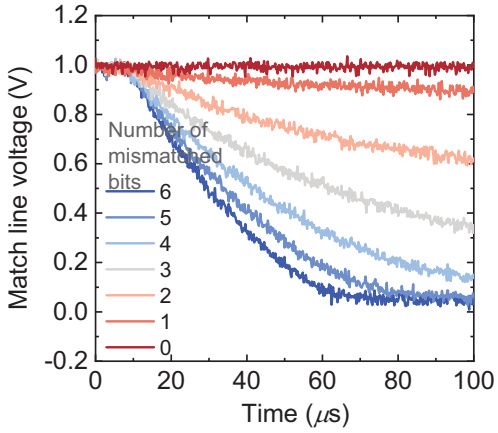

Fig. 3. Experimental demonstration of the TCAM array functionality of detecting the degree of mismatch between the stored information and query on a small 1x6 prototype array. The match line discharge rate increases with the number of mismatched bits and it follows a linear relationship, suggesting the capability of TCAM in performing such computation [12].

D. Quasi Leaky Integrate & Fire Neuron

Biologically inspired spiking neural network (SNN), where the information is communicated through neurons with sparse discrete spiking event, is a promising high performance and energy-efficient network architecture for deep learning applications [32]. To implement a SNN in hardware, neuron, synapse, and the learning mechanisms are the key elements. The local Hebbian learning rule through spike time dependent plasticity (STDP) employed in biological neurons has also been successfully demonstrated in ferroelectric synapse. So, there is a strong motivation of realizing ferroelectric neuron so that the whole SNN can be implemented in a single unified platform.

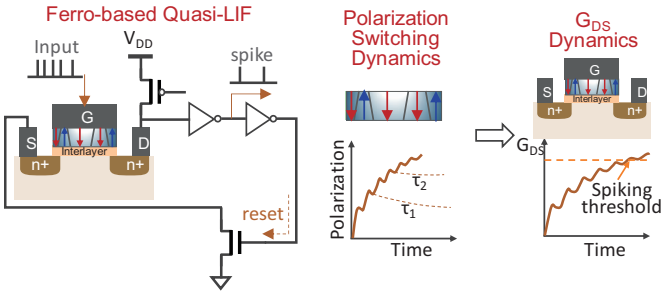

Fig. 4. The ferroelectric quasi-leaky integrate and fire neuron exhibits variable relaxation time constant τ. Specially the leak rate decreases with an increase in the polarization state. When the polarization hits a threshold, the neuron will generate a spike [33].

In a biological neuron, homeostatic regulation is employed to maintain a target level of neuronal excitability. Translating such biologically plausible mechanism into spiking neural networks ensures their stability in an unsupervised learning environment. We have proposed to utilize a unique polarization dynamic in a multi-domain ferroelectric film in a FeFET to realize a bio-plausible neuron model with built-in

homeostasis, as shown in Fig. 4 [33]. Contrary to a traditional leaky integrate and fire (LIF) neuron that employs a fixed time-constant τ for integrate and leak the membrane potential, ferroelectric neuron exhibits a variable τ. Specifically, the leak rate decreases (τ increases) with an increase in the polarization state (aka membrane potential). Such an increase in τ results in a strong activation of homeostasis via adaptive firing threshold, thereby decreasing the overall network firing rate. We can interpret this as "intrinsic homeostasis" wherein the spiking activity of the neuron gets modulated by its own internal membrane state in a natural way.

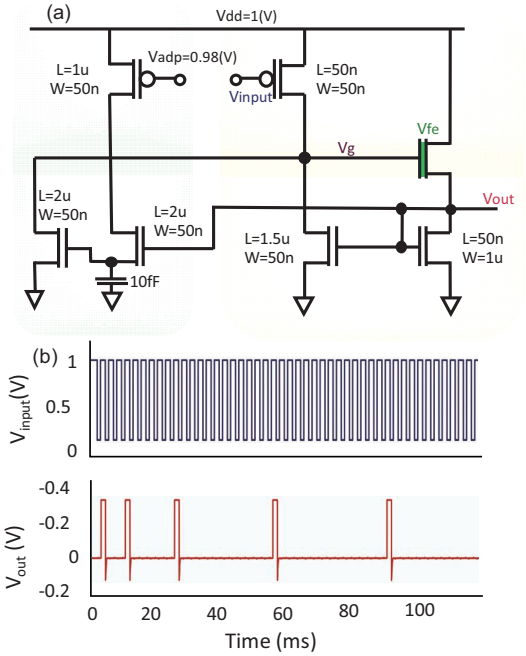

Fig. 5. The complete neuron circuit includes the FeFET and also 6 other supporting transistors. It exhibits the homeostasis characteristic, similar to the biological neuron.

The complete neuron circuit is presented in Fig. 5, showing the desired homeostasis can be implemented in ferroelectric neuron. With this neuron, the network level simulations with the ferroelectric neuron model exhibits a 2.3x reduction in firing rate compared to traditional LIF neuron while maintaining iso-accuracy of 84-85% across varying network sizes. Such an energy-efficient hardware for spiking neuron can enable ultra-low power data processing in energy constrained environments suitable for edge-intelligence.

III. Conclusions

In summary, various exciting applications has been enabled and developed on the ferroelectric HfO_2 platform. The applications encompass from the high-performance logic transistor, to embedded nonvolatile memory, to analog synaptic weight cell, to in-memory computing, and to the spiking neural network implementation. This list of ferroelectric applications is, by no means, exhaustive and it will only keep expanding as a matter of fact. Therefore, this versatile ferroelectronic platform is a great complement to the current electronics.

All these development efforts are just scratching the surface of the whole ferroelectronic platform. Further investigations are still indispensable to make ferroelectric a practical technology, such as solving the challenges of endurance, variation, and multi-level cell [34] etc. It is our hope that this summary provides a guide to the exciting ferroelectric developments.

References

[1] J. Müller et al., "Ferroelectric hafnium oxide: A CMOS compatible and highly scalable approach to future ferroelectric memories," *IEDM Tech. Dig*., 280-283, 2013.

[2] S. Salahuddin and S. Datta, "Use of negative capacitance to provide voltage amplification for low power nanoscale devices," *Nano Lett.* vol. 8, no. 2, pp. 405–410, Feb. 2008.

[3] Z. Krivokapic et al., "14nm ferroelectric FinFET technology with steep subthreshold slope for ultra-low power applications," *IEDM Tech. Dig*., 357-360, 2017.

[4] M. H. Lee et al., "Physical thickness 1.x nm ferroelectric $HfZrO_x$ negative capacitance FETs," *IEDM Tech. Dig*., 306-309, 2016.

[5] B. Obradovic et al., "Ferroelectric switching delay as cause of negative capacitance and the implications to NCFETs," *Symposium on VLSI Technology*, 51-52, 2018.

[6] P. Sharma et al., "Time resolved measurement of negative capacitance," *IEEE Electron Dev. Lett.* 272-275, 2017.

[7] S. Salahuddin et al., "The era of hyper-scaling in electronics," *Nature Electronics*, 442-450, 2018.

[8] M. Trentzsch et al., "A 28 nm HKMG super low power embedded NVM technology based on ferroelectric FETs," *IEDM Tech. Dig*., 294-297, 2016.

[9] S. Dunkel et al., ""A FeFET based super-low-power ultra-fast embedded NVM technology for 22 nm FDSOI and beyond," *IEDM Tech. Dig*., 485-488, 2017.

[10] M. Jerry et al., "Ferroelectric FET analog synapse for acceleration of deep neural network training," *IEDM Tech. Dig*., 139-142, 2017.

[11] X. Yin et al., "An ultra-dense 2FeFET TCAM design based on a multi-domain FeFET model," *IEEE TCAS II*, 1577-1581, 2018.

[12] K. Ni et al., "Ferroelectric ternary content-addressable memory for one-shot learning," *Nature Electronics*, 521-529, 2019.

[13] K. Ni et al., "Critical role of interlayer in $Hf_{0.5}Zr_{0.5}O_2$ ferroelectric FET nonvolatile memory performance," *IEEE Tran. Electron Dev.,* vol. 65, no. 6, 2641-2649, 2018.

[14] X. Li et al., "Design of 2T/cell and 3T/cell nonvolatile memories with emerging ferroelectric FETs," *IEEE Design & Test,* vol. 36, no. 3, 39-45, 2019.

[15] K. Ni et al., "Write disturb in ferroelectric FETs and its implication for 1T-FeFET and memory arrays," *IEEE Electron Dev. Lett.* vol. 39, no. 11, 1656-1659, 2018.

[16] D. Reis et al., "Design and analysis of an ultra-dense, low-leakage and fast FeFET-based random access memory array," *IEEE JxCDC,* 2019.

[17] K. Ni et al., "A circuit compatible accurate compact model for ferroelectric FETs," *Symposium on VLSI Technology,* 131-132, 2018.

[18] K. Ni et al., "SoC logic compatible multi-bit FeMFET weight cell for neuromorphic applications," *IEDM Tech. Dig*., 296-299, 2018.

[19] P. D. Lomenzo et al., "Ferroelectric Si-doped HfO_2 device properties on highly doped Germanium," *IEEE Electron Dev. Lett.* vol. 36, no. 8, 766-768, 2015.

[20] Q. H. Luc et al., "First experimental demonstration of negative capacitance InGaAs MOSFETs with $Hf_{0.5}Zr_{0.5}O_2$ ferroelectric gate stack," *Symposium on VLSI Technology*, 47-48, 2018.

[21] F. A. McGuire et al., "Sustained sub-60mV/decade switching via the negative capacitance effect in MoS_2 transistors," *Nano Lett.*

vol. 17, no. 8, 4801-4806, 2017.
[22] F. Mo et al., "Experimental demonstration of ferroelectric HfO_2 FET with ultrathin-body IGZO for high-density and low-power memory applications," *Symposium on VLSI Technology*, 42-43, 2019.
[23] J. Muller et al., "High endurance strategies for hafnium oxide based ferroelectric field effect transistor," *Nonvolatile Memory Technology Symposium*, 2016.
[24] X. Li et al., "Lowering area overheads for FeFET-based energy efficient nonvolatile flip-flops," *IEEE Tran. Electron Dev.,* vol. 65, no. 6, 2670-2674, 2018.
[25] X. Li et al., "Design of nonvolatile SRAM with ferroelectric FETs for energy-efficient backup and restore," *IEEE Tran. Electron Dev.,* vol. 64, no. 7, 3037-3040, 2017.
[26] X. Li et al., "Advancing nonvolatile computing with nonvolatile NCFET latches and flip-flops," *IEEE TCAS I*, vol. 64, no. 11 2907-2919, 2017.
[27] K. Ni et al., "Fundamental understanding and control of device-to-device variation in deeply scaled ferroelectric FETs," *Symposium on VLSI Technology*, 40-41, 2019.
[28] P. Y. Chen et al., "NeuroSim+: an integrated device-to-algorithm framework for benchmarking synaptic devices and array architectures," *IEDM Tech. Dig*., 135-138, 2017.
[29] X. Sun et al., "Exploiting hybrid precision for training and inference: a 2T-1FeFET based analog synaptic weight cell," *IEDM Tech. Dig*., 55-58, 2018.
[30] D. Reis, et al., "Computing in memory with FeFETs," *ISLPED*, 2018.
[31] K. Ni et al., "In-memory computing primitive for sensor data fusion in 28nm HKMG FeFET technology," *IEDM Tech. Dig*., 364-367, 2018.
[32] P. U. Diehl et al., "Unsupervised learning of digit recognition using spike-timing-dependent plasticity," *Front. Comput. NeuroSci.,* 2015.
[33] S. Dutta et al., "Biologically plausible ferroelectric quasi-leaky integrate and fire neuron," *Symposium on VLSI Technology*, 140-141, 2019.
[34] K. Ni et al., "A novel ferroelectric superlattice based multi-level cell non-volatile memory," *IEDM Tech. Dig.*, 2019

.