Computing with Ferroelectric FETs: Devices, Models, Systems, and Applications

Ahmedullah Aziz[1], Evelyn T. Breyer[6], An Chen[2], Xiaoming Chen[3], Suman Datta[4], Sumeet Kumar Gupta[1], Michael Hoffmann[6], Xiaobo Sharon Hu[4], Adrian Ionescu[5], Matthew Jerry[4], Thomas Mikolajick[6,8], Halid Mulaosmanovic[6], Kai Ni[4], Michael Niemier[4], Ian O'Connor[7], Atanu Saha[1], Stefan Slesazeck[6], Sandeep Krishna Thirumala[1], and Xunzhao Yin[4]

[1]Purdue University, USA, Email: {aziz5, guptask, saha26, sthirum}@purdue.edu

[2]Semiconductor Research Corporation, USA, Email: An.Chen@src.org

[3]State Key Laboratory of Computer Architecture, Institute of Computing Technology, Chinese Academy of Sciences, China, Email: chenxiaoming@ict.ac.cn

[4]University of Notre Dame, USA, Email: {sdatta, shu, mjerry, kni, mniemier, xyin1}@nd.edu

[5]Ećole polytechnique feďefale de Lausanne (EPFL), Switzerland, Email: adrian.ionescu@epfl.ch

[6]NaMLab GmbH, Germany, Email: {Evelyn.Breyer, Michael.Hoffmann, Halid.Mulaosmanovic, Thomas.Mikolajick, Stefan.Slesazeck}@namlab.com

[7]Ecole Centrale de Lyon, France, Email: Ian.Oconnor@ec-lyon.fr

[8]IHM TU-Dresden, Germany, Email: thomas.mikolajick@tu-dresden.de

***Abstract*—In this paper, we consider devices, circuits, and systems comprised of transistors with integrated ferroelectrics. Said structures are actively being considered by various semiconductor manufacturers as they can address a large and unique design space. Transistors with integrated ferroelectrics could (i) enable a better switch (i.e., offer steeper subthreshold swings), (ii) are CMOS compatible, (iii) have multiple operating modes (i.e., I-V characteristics can also enable compact, 1-transistor, non-volatile storage elements, as well as analog synaptic behavior), and (iv) have been experimentally demonstrated (i.e., with respect to all of the aforementioned operating modes). These device-level characteristics offer unique opportunities at the circuit, architectural, and system-level, and are considered here from device, circuit/architecture, and foundry-level perspectives.**

I. Introduction

There is obvious interest in finding new ways to preserve performance scaling trends that have historically accompanied Moore's Law-based device scaling. This has motivated research efforts in Asia, Europe, and North America [1]–[3] – all with the end goal of finding the next switch. However, a replacement for the MOSFET has proven to be elusive. As an example, tunneling field effect transistors (or TFETs) were viewed by many [4], [5] as a promising MOSFET alternative that could potentially serve as a "drop-in-replacement" when considering existing CMOS-based, core logic structures (circuits requiring pass gates are a notable exception). However, experimentally, TFETs have either suffered from low I_{on} currents, or have not been able to achieve a desirable subthreshold swing [6]. Moreover, comprehensive benchmarking efforts suggest that when used as just a switch, other devices (e.g., charge-based TFETs, as well as devices that employ other state variables) will not offer substantial improvements over CMOS technology slated for 2018 [7] (Fig. 1).

The current state-of-the-art is perhaps best summarized by an excerpt from a Semiconductor Research Corporation (SRC)

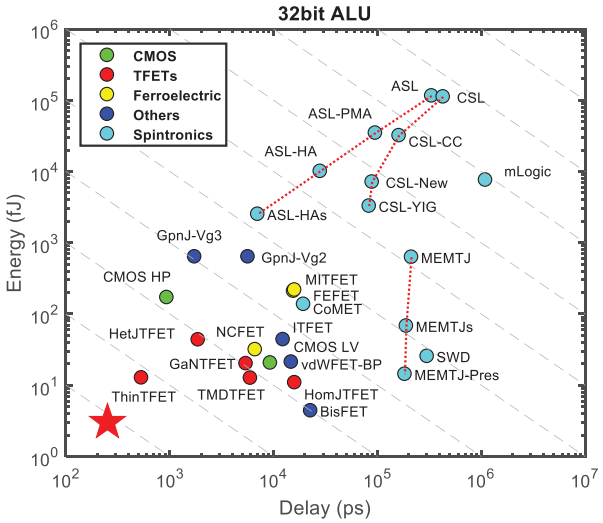


Figure 1. Functional unit-level benchmarking of different logic devices; many device concepts exist only in simulation (from [7] via Naeemi, Pan).

Nanotechnology Research Initiative (NRI) recent call for proposals [8]: *"NRI research has explored a broad spectrum of beyond-CMOS devices for a 'new logic switch' to replace the current CMOS-based transistor ... a* **'better switch'** *has not been found. Comprehensive benchmarking of beyond-CMOS devices ... has revealed little or no advantage of these devices over CMOS for conventional Boolean logic and the von Neumann architecture."* Given the above, many new research programs are either (i) adopting a more system-centric focus [2], [9], [10] and/or (ii) looking to exploit other characteristics associated with emerging devices (besides just better subthreshold swings). Quoting again from [8], *"some devices demonstrate unique characteristics suitable for novel architectures or computing paradigms, e.g., non-volatility in*

978-3-9819263-0-9/DATE18/©2018 EDAA

logic devices, reconfigurability, high computation density." Devices with integrated ferroelectrics are well-positioned to address this space, and are being considered by various semiconductor manufacturers [11], [12].

In this paper, we discuss different types of ferroelectric devices and relevant device physics (Sec. II-A), as well as recent modeling efforts (Sec. II-B). Recent experimental advances with ferroelectric devices/transistors are also discussed (Sec. II-C). We then consider the impact of ferroelectric devices at the circuit, architecture, and application-levels. Application-centric case studies discussed in Sec. III include *design space explorations* of: (i) non-volatile flip-flops and memory structures [13] for non-volatile processors [14], (iii) fine-grained logic-in-memory [15], (iv) content addressable memories [15], and (v) crossbar structures for neuro-inspired computing models [16]. We also consider how ferroelectric devices fare when compared to other device-centric solutions for similar problems (e.g., RRAM, STTRAM, etc.). We conclude by examining FeFET implementations from the perspective of a large-scale manufacturing process – specifically discussing GlobalFoundries-based FeFET implementations, recent memory-centric results, and the potential impact of device endurance. Recent experimental advances/demonstrations of logic-in-memory primitives and FeFET-based synaptic structures [17], [18] are also discussed.

II. Background and Related Work

Here, we present a brief description of the structure and operation of ferroelectric transistors. This sets the stage for a better understanding of the circuits, architectures, and applications discussed in subsequent sections of the paper. We also provide an overview of the modeling approaches for ferroelectric transistors, highlighting the features, assumptions and limitations of each. We conclude with a brief review of the experimental state-of-the-art.

A. Ferroelectric transistor structures and operation

A ferroelectric transistor (FeFET) is structurally similar to a regular bulk MOSFET or FinFET, except that an additional layer of ferroelectric (FE) material is integrated in its gate stack (Fig. 2). The metal layer between the FE and dielectric shown in Figs. 2a, b may or may not be included, and there have been demonstrations of FE transistor structures with [19] and without [20] this layer. Note that some FE materials (e.g. lead zirconium titanate (PZT) [21]) may be incompatible with CMOS processes. However, recent demonstrations of ferroelectricity in hafnium zirconium oxide (HZO) [22] (highly compatible with CMOS processes) has mitigated concerns regarding large-scale demonstrations of FE transistors that might impede industrial-scale realizations.

The interplay between the FE material with the underlying transistor capacitance creates different modes of operation in an FeFET. In a broad sense, based on the transfer characteristics, FeFETs can operate in two different modes: a *non-volatile* mode and a *steep switching* mode. The steep switching mode can be hysteretic or non-hysteretic. In principle, for either

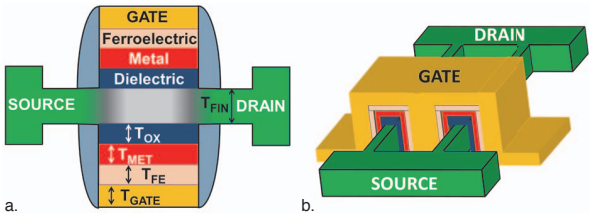


Figure 2. **(a)** cross-section of a FeFET; **(b)** 3D view of a FeFET.

operating mode, the basic device structure is the same. The relative capacitance of the FE and the transistor determines the specific mode that an FeFET exhibits.

1) Non-volatile mode: Non-volatile PZT-based FeFETs have been explored for several decades [23]. The emergence of HZO-based FeFETs has led to additional studies of the non-volatile operating mode – not only from the perspective of device operation, but also in conjunction with application-centric and/or memory cell designs [14]. The FE material can be considered in the context of hysteretic polarization versus voltage (P versus V_{FE}) (Fig. 3a).When placed in series with the gate of a transistor, the hysteretic window of P versus V_{GS} (Fig. 3b) is reduced due the effect of the capacitance of the MOS structure of the FET, and the associated depolarization fields [24]. Nevertheless, with a sufficiently thick FE, the hysteretic behavior is preserved, and can be observed in the I_D-V_{GS} transfer characteristics of a device (Figs. 3c, d). This corresponds to the non-volatile, hysteretic mode of the FeFET. In this mode of operation, at $V_{GS} = 0$ V (i.e., when the supply voltage is turned off), the FeFET exhibits bi-stable states which correspond to positive and negative polarization retention in the FE layer. Depending on the polarization, the FeFET exhibits high resistance ($P < 0$ for an n-type FeFET and $P > 0$ for a p-type FeFET), or low resistance ($P > 0$ for an n-type FeFET and $P < 0$ for a p-type FeFET). (Again, see Figs. 3c, d.) Thus, non-volatility is embedded inside the transistor provided the FE layer is sufficiently thick [14], [24]. Alternatively, if the thickness of the FE is lower than a critical level (e.g., 7 nm in Fig. 3c), the FeFET loses its non-volatility as the hysteresis in P versus V_{GS} is reduced. However, the volatile version of the FeFET can exhibit a different mode of operation – steep switching mode (discussed below).

2) Steep switching mode: The steep switching mode of an FeFET was proposed conceptually in 2008 by Salahuddin and Datta [25]. They envisioned switching the polarization of the FE layer following an S-shaped trajectory (Fig. 4a). The "snap-back" region of the S-shaped, P versus V_{FE} curve exhibits a negative differential capacitance (dQ/dV) because of its negative slope. This region in the polarization landscape is physically unstable as it is energetically unfavorable [25]. However, when in series with a positive capacitance, the negative differential capacitance of the FE can manifest itself as long as the overall capacitance is still positive, which stabilizes the FE operation in the negative capacitance region. Having a negative capacitance in series with a positive one in the gate stack has interesting implications. First, the voltage

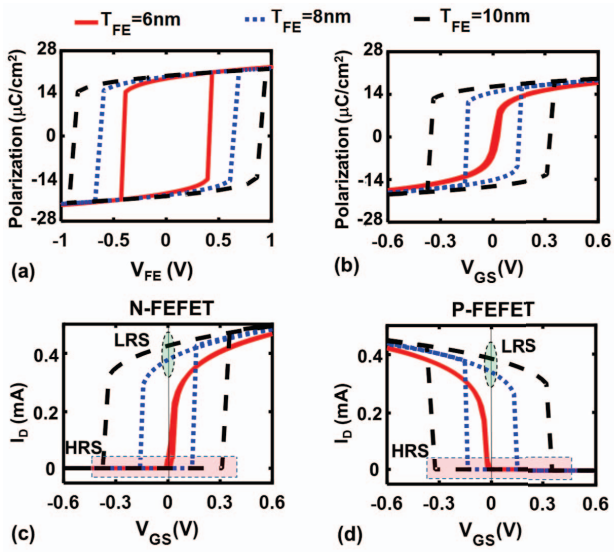


Figure 3. **(a)** Polarization versus voltage characteristics of a standalone ferroelectric material; **(b)** polarization versus gate voltage of a FeFET; **(c)** N-type FeFET; **(d)** P-type FeFET.

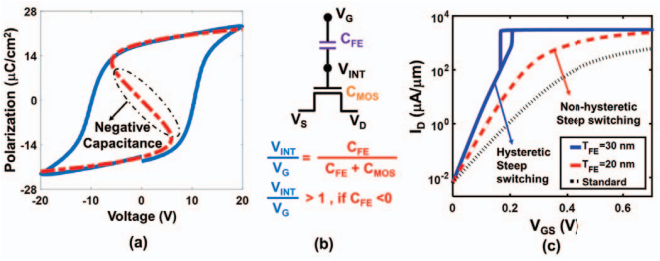


Figure 4. **(a)** Polarization-voltage characteristics of a ferroelectric material showing the envisioned S-shaped transition; **(b)** voltage gain in a NCFET; **(c)** transfer characteristics of steep switching NCFETs.

division between the positive and negative capacitors leads to amplification in the internal gate voltage of the transistor (Fig. 4b). This amplification in surface potential leads to enhanced drain current (I_D), and also allows the transistor to achieve below 60 mV/dec (Boltzmann limit) subthreshold swing [25] (hence the name steep switching). This form of FeFET is more commonly referred to as a negative capacitance transistor (NCFET). An important dissimilarity between the non-volatile FeFET and the steep switching NCFET is that in the former, complete polarization switching (positive to negative) occurs in the FE. The latter experiences only partial polarization switching and hence is expected to be faster. NCFETs are being studied as they could enable low power logic operation. (While other emerging transistor technologies [26], [27] also promise steep switching operation, FeFETs may ultimately be more desirable owing to their close resemblance to memory-centric transistor structures.)

Note that in certain scenarios, there can be hysteresis in the transfer characteristics of a NCFET (Fig. 4c). More specifically, there may be hysteresis, but not across $V_{GS} = 0$ V (Fig. 4c). Hysteresis may occur on either side of the $V_{GS} = 0$ axis and hence the device operation is volatile. In other words, at $V_{GS} = 0$, the FE layer has one fixed polarization value. Volatile hysteretic behavior arises from having two possible stable solutions when $V_{GS} \neq 0$. (This arises from the non-linearity of the transistor capacitance, and its interaction with the negative FE capacitance.) This condition exists when the thickness of the FE in a NCFET exceeds a critical level and is typically observed when the capacitance of the transistor is highly non-linear. The relative capacitance of the underlying transistor with respect to the FE layer determines if the hysteresis will be volatile or non-volatile. For instance, it has been shown that both of these modes can be achieved by changing the thickness of the FE [28].

B. Device Modeling

With ever-growing interest in FE transistor technology, it is vital to model the behaviors of different versions of the FeFETs to develop and evaluate future designs. Significant modeling efforts have been directed towards predicting the behavior of this technology. Below, we provide a brief review.

To understand the complex interaction between the FE material in the gate stack, and the underlying FET (baseline FET) capacitance, device level modeling of FeFETs can provide very important insights. In this regard, transistor equations (Poisson's equation and the drift-diffusion current continuity equations) must be solved self-consistently with the Landau-Khalatnikov (L-K) equation of the FE. The approach for modeling a FeFET strongly depends on the absence or presence of the internal metal layer (in between the FE and dielectric). To understand the impact of this internal metal layer, we first consider the multi-domain L-K equation:

$$E = \alpha P + \beta P^3 + \gamma P^5 - [\frac{1}{2} K_P \frac{d^2 P}{dx^2}] \quad (1)$$

In Eq. 1, P is polarization, E is electric field and α, β, γ are Landau coefficients. The term K_P is called the domain interaction parameter. Interestingly, having an interlayer metal screens out the non-uniformity in the electric field within the FE layer (the metal layer acts as an equipotential surface). That implies an equal polarization in all the FE domains. As a result, the term in parenthesis in Eq. 1 – $(\frac{1}{2})K_P(\frac{d^2P}{dx^2})$ – becomes zero and can be treated as a single domain L-K equation. A 3D TCAD model of FE-FinFET at the 10 nm node has been demonstrated in [29], which assumes an interlayer metal between the FE and dielectric in the gate stack. In this work, Poisson's equation and the drift-diffusion current continuity equations for the baseline FinFET have been solved self-consistently with the single domain L-K equation.

The assumption of having an interlayer metal may justify the use of the single domain L-K equation in [29], as the metal makes the electric field across the FE uniform. However, this assumption has some limitations. For polycrystalline FE materials, having an equipotential internal metal layer does not mandate equal polarization in all domains. Variation in polarization axes and charge trapping may contribute to creating non-uniformity in the FE polarization. Moreover, it has been reported that the internal metal layer might destabilize the negative capacitance effect [30], [31]. But, if the FE

is fabricated directly on top of a dielectric/oxide layer (i.e. without an inter-layer metal [20]) such issues are mitigated. However, the non-uniform electric field in the channel also makes the electric field in the ferroelectric non-uniform. Due to the non-uniformity in the electric field, FE polarization varies along the gate length, which in turn leads to domain formation in the FE layer of the FeFET. It is important to note that in reality even in a single crystal FE sample, domain formation is a typical phenomenon. To analyze the effect of domain formation in the FE layer in FeFETs, a 2-D self-consistent model utilizing the multi-domain L-K equation within the NEGF (Non-Equilibrium Green's Function) framework has been developed [32]. The model captures the signatures of negative capacitance of the FE in the FeFET characteristics, i.e., negative DIBL, negative output conductance, etc. According to the analysis in [32], it is important to have higher domain interaction (higher K_P value) to obtain the benefit from negative capacitance of the FE layer in the FeFET characteristics.

To perform more rigorous circuit analysis and architectural design-space explorations based on FeFET devices, a circuit compatible model is needed. A SPICE model for FeFETs based on the time-dependent single-domain L-K equation (solved self-consistently with the transistor equations) has been proposed in [21]. It considers the presence of (i) an interlayer metal in the gate stack, and (ii) the depolarization fields due to non-ideal contacts. Implementation of the model entirely in SPICE enables efficient monitoring of the intermediate variables such as polarization, interlayer metal potential, and depolarization fields during circuit operation.

A Verilog-A model of an FeFET in [33] is also based on a self-consistent solution of the single-domain L-K and transistor equations. Verilog-A based implementations makes the model simulator independent and therefore can be run in different interfaces/compilers [34]. A distributed charge model for FeFETs (BSIM-CMG) has also been presented in [35]. It can capture the characteristics of FeFETs with and without an interlayer metal considering the lumped and distributed nature of channel charges respectively. To characterize the multi-domain effect of FE, a compact model [36] for FeFETs has been proposed considering multiple FE domain structures that can be thermally activated. To analyze the dynamics of the electric polarization and to calculate the thermal activation rate, L-K theory has been used in this model. The compact model in [37] uses an explicit expression for the channel current in a bulk NCFET, that considers the spatial variation of FE polarization in the longitudinal direction.

C. Experimental Progress

HfO_2 based FeFETs have recently received great interest for application in nonvolatile memory (NVM) [1]. As noted above, unlike conventional perovskite based ferroelectrics, HfO_2 is CMOS compatible and scalable to film thickness in the nanometer range. Indeed, HfO_2 has been successfully integrated at the 28 nm technology node, which suggests that this technology is highly promising for NVM [38].

Compared with existing current-driven NVM technologies (e.g., flash, phase change memory (PCM), and STT-MRAM), FeFET-based storage is electric field driven, which improves energy efficiency. Existing HfO_2 based FeFETs employ a write pulse of approximately 5V with a 100 ns pulse width, and achieve a memory window of about 1V – which is only about half of the theoretical value $2 \times E_C \times t_{FE}$ [12], [39]–[41]. This is because an applied voltage is divided between the ferroelectric and the underlying interlayer and semiconductor, which causes the ferroelectric to operate on a non-saturated inner polarization-voltage loop – which reduces the memory window. Potential strategies to improve the performance include the usage of a high κ interlayer or increasing the interlayer to FE area ratio [42].

One of the challenges of HfO_2 based FeFETs is limited endurance (approximately 10^5 cycles) [40], [42] caused by dielectric breakdown or charge trapping, which is related to the high coercive field of HfO_2. A large electric field is needed to flip polarization, which in turn results in a high interlayer electric field. This high electric field stress in the interlayer facilitates dielectric breakdown and charge trapping. Therefore, it is necessary to reduce electric field in the interlayer and charge injection. Potential solutions include increasing interlayer thickness, or using a high κ interlayer [43]. Endurance cycles of up to 10^7 have been shown for FeFETs with a 2 nm SiO_2 interlayer [44] and 10^{12} cycles was demonstrated for FeFETs with a 3 nm SiO_2 interlayer [45]. Larger-scale array results will be discussed further in Sec. IV.

While a more detailed discussion is beyond the scope of this paper, for recent work regarding the experimental state-of-the-art with NCFETs, we refer the reader to [20], [46], [47]

III. FeFET Circuits and Architectures

We now discuss how different operating modes associated with transistors with integrated ferroelectrics might ultimately impact application-level tasks and performance. When possible, case studies benchmark FE devices and circuits against other functional equivalents.

A. Non-volatile flip-flops

FeFET-based non-volatile flip-flops have been proposed/designed to back up processor state in the event of power failures, for power gating, etc. [24]. Data is stored via the FE layer in a given FeFET. A representative design appears in Fig. 5. While a more detailed description of device operation can be found in [24], in brief, during a backup operation, the flip-flop's output (Q) is connected to the gate of the FeFET. (During normal operation, the pass transistor NQ may be turned off in order to isolate the flip-flop output (Q) from the gate capacitance of the FeFET.) During normal operation or a restore operation, the FeFET gate is driven to V_{DD} via serially connected transistors PB and PR, which in turn are controlled by the backup (BKP) and reset (RST) signals, respectively. The remaining terminal of the FeFET is controlled by the signal BKP OR $RSTR$, which can be shared among multiple flip-flops.

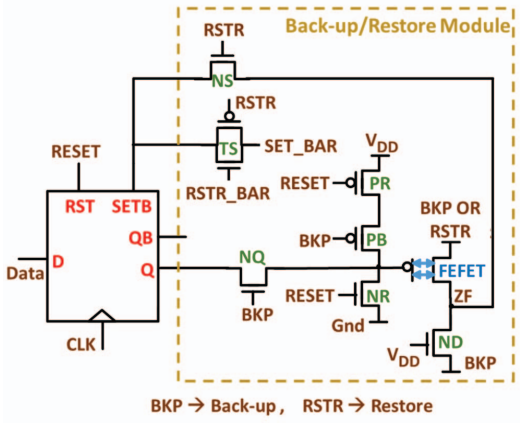


Figure 5. FeFET-based non-volatile flip-flop.

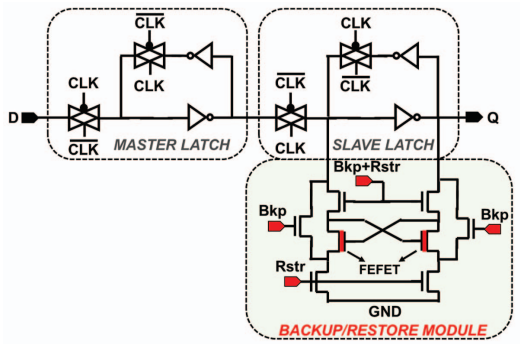


Figure 6. Another FeFET-based non-volatile flip-flop design.

Another possibility is to integrate FeFETs inside the flip-flop architecture (Fig. 6) [48]. This design uses 2 FeFETs. During back-up operation, one storage node of the slave latch is used to control V_{GS} of a FeFET, and the other controls source/drain bias. Depending on flip-flop state, V_{GS} of the FeFET is either negative or positive. The other FeFET stores a complementary polarization. During restore, the differential FeFETs drive the storage nodes of the slave latch. Thus, this implementation uses FeFETs in conjunction with the cross-coupled inverters of the flip-flop to restore state.

The power and speed of the FeFET-based, NV flip-flop has been benchmarked against a FE capacitor-based solution [49]. FeFET-based designs exhibits 40% to 50% lower delay across a range of V_{DD} values when compared to the FE capacitor-based flip-flop. Energy dissipation is 27% to 40% lower with the FeFET-based design. (This is largely due to the fact that a single FeFET can retain information, while four FE capacitors are required.) For restore operations, delay is comparable when comparing FeFET and FE capacitor-based solutions. However, the latter design requires a sense amplifier to distinguish between states in the FE capacitor, while the FeFET-based design can exploit the orders of magnitude differences between states for a restore operation.

B. FeFET Memory Cells

FeFET-based memory structures have been considered [50], and are appealing given the potential for non-volatile retention

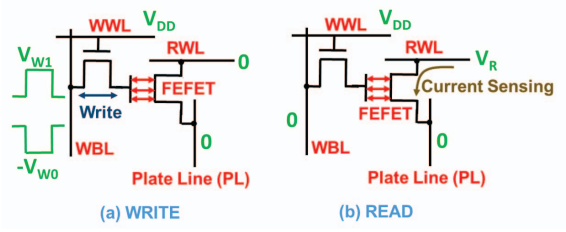


Figure 7. FeFET-based memory cells **(a)** write and **(b)** read.

capability based on FE polarization, high I_{ON}/I_{OFF} ratios, and low voltage operation. Besides the realization of single FeFET cell-based approaches (see Sec. IV), 2T Fe-FET-based memory cells have been proposed (see Fig. 7). Owing to a third terminal, a cell can have separate read/write paths, which allows for read/write operations to be simultaneously optimized. More details about cell operation can be found in [50]. In the context of a NV memory array, write and read energies are 3.1X and 55.4X better (respectively) than a FeRAM-based backup at iso-write delay [50].

C. FeFET-based, fine-grained logic in memory

While there is obvious interest in a "better switch", even if improved transistors evolve, and traditional core scaling continues, one must still supply each core with data to process. Recent work [51] suggests that in order for future microprocessors to match traditional Moore's Law performance scaling trends, 58W of a 65W power budget would be allocated for just data transfer. However, if the distance between logic and storage can be reduced by 10X, 90% of a 65W power budget could be devoted to computation. "Near data processing" [52] and processing-in-memory (PIM) prototypes have been heavily pursued since the 1990s (e.g., [53], [54]), and have gained more momentum with the recent advent of 3D integration. Indeed, system-level analysis of Micron's hybrid memory cube [55], [56], the N3XT project [57], etc. suggest that 10-1000X improvements in performance/energy are possible.

In contrast to "coarse-grained" efforts (i.e., with logic and memory on separate dies), "fine-grained" logic-in-memory (LiM) structures could also bring data closer to processing elements by (i) leveraging local, NV storage to preserve system state, and (ii) integrating NV storage elements with the logic itself. Work described in Sec. III-A is an example of the former. As an example of the latter, [58] proposed that magnetic tunnel junctions (MTJs) be integrated with MOS transistors to store data words that might be repeatedly used over the course of a given computation, e.g., for a sum-of-absolute differences (SAD) calculation commonly used in compression, motion detection, etc.

FeFET-based structures are also well-suited for this space. As a representative example, FeFET-based dynamic logic (DL)-style LiM circuit structures have been developed [15] (Fig. 8). One of the two inputs is stored locally by leveraging FeFET non-volatility. FeFETs (along with associated access transistors) can be distributed in the pull-down network (i.e., with other N-channel devices) and can serve as both a logic switch and an NV storage element. The inputs Y and $\overline{Y}$ are

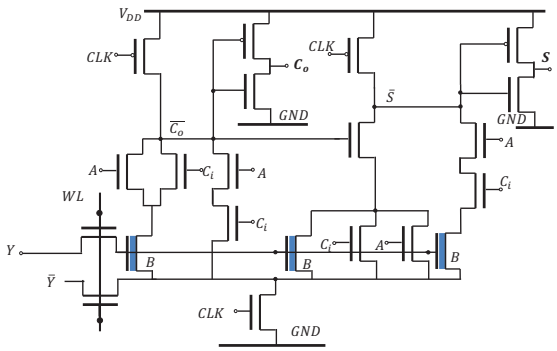


Figure 8. General structure of FeFET-based dynamic logic circuits.

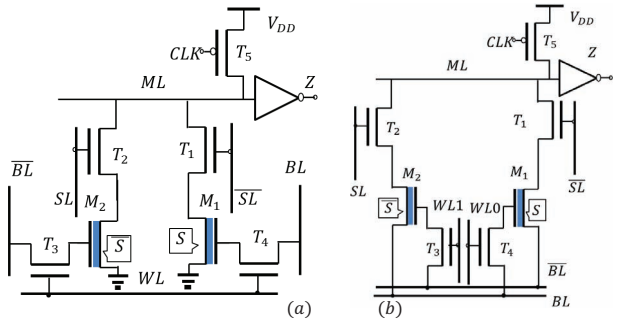


Figure 9. Two FeFET TCAM cells **(a)** with WS1; **(b)** with WS2.

set to have either a positive or negative gate-source voltage for the FeFET to change its state to '1' or '0' respectively, thus achieving NV bit storage based on device hysteresis (albeit at the expense of an access transistor).

The FeFET-based DL 1-bit full adder (FA) in Fig. 8 is similar to a conventional DL FA, but the transistors associated with input B are replaced by the FeFET-based NV memory elements. As memory elements store the same bit, the access transistor can be shared by the three FeFETs. Due to reduced transistor count and the DL-style employed, this FeFET-based NV LiM FA achieves better dynamic power efficiency as well as delay than other NV LiM FAs. Notably, the FeFET-based DL FA has a nearly identical delay and dynamic power when compared to a CMOS implementation at the same technology node [15]. (This is somewhat expected as the designs have similar circuit topologies.) That said, the FeFET-based approach simultaneously offers the benefit of local, NV storage, which has application-level utility as noted above. Furthermore, additional benefits of the FeFET-based approach are observed when comparing designs to functional equivalents based on other emerging technologies. Notably, the area-delay-power product of a FeFET-based adder is 19X/84X better than that of an MTJ/FTJ-based approach (where area is assumed to be proportional to device count) [15].

D. FeFET-based content addressable memories

Recent work also suggests that FeFET devices are well-suited for realizing content addressable memories (CAMs) and ternary content addressable memories (TCAMs) [15], [59]. TCAMs perform parallel searches for a given piece of data against a table of stored data, and return information as to whether a match occurred. TCAMs have obvious utility in networking hardware/applications, i.e., in routers, database search applications, and associative memories. More recently, [60] has also proposed using TCAMs for more energy efficient, in-memory data processing by reducing the amount of redundant data associated with traditional von-Neumann processing, as well as to efficiently realize various neuro-inspired computing models [61].

FeFET-based TCAM designs were originally proposed in [15]. Here, we discuss two designs that employ different writing schemes (WS1 and WS2 – with and without a negative supply respectively). The design in Fig. 9a (WS1) [59] consists of two parallel FeFETs connected to a matchline (ML) via two transistors. In addition to storing complementary bits, the two FeFETs can also both store logic '0' which represents the "don't care" state. In the cell schematic, the transistors M_1/T_1 and M_2/T_2 serve as two pull down paths for the ML. The inputs to the transistors T_1 and T_2 – SL and $\overline{SL}$ – together with the memory state stored in M_1 and M_2 (S and $\overline{S}$) determine whether the pull down paths are on or off, and provide an XNOR output $\overline{S \oplus SL}$ at the ML.

The structure in Fig. 9a can be modified to support WS2, where the complementary bitlines (BL and $\overline{BL}$) are used for writing. Per Fig. 9b, the TCAM cell still employs transistors M_1/T_1 and M_2/T_2 as in Fig. 9a. However, they now serve as the pull down paths, from the ML to BL and $\overline{BL}$, instead of to ground. The buffer transistors driving BL and $\overline{BL}$ serve as the ground for the discharging current, and the negative voltage required for the design in Fig. 9a is eliminated.

FeFET-based TCAM designs have been considered in the context of array-based architectures, and benchmarked against other memory technologies/TCAM cells. (As described in [59], the array consists of TCAM core cells, input buffers and drivers, the output sense amplifier, the clock signal, and the output encoder.) When compared to CMOS functional equivalents, FeFET-based TCAMs have similar latency and energy as they have similar capacitances at the matchlines and search lines. However, FeFET-based designs are (i) denser (CMOS designs typically employ a 12T, NAND-type cell or a 16T NOR-type cell, while the FeFET approach employs a 4T-2 FeFET based cell), and (ii) non-volatile. When compared to ReRAM and MTJ-based TCAMs, FeFET-based designs have energy delay products (EDPs) that are 1.7X and 149X better respectively. (Note that per [59], ReRAM-based designs may offer some advantages with respect to array density.)

E. FeFET-based crossbar design for BCNNs

Recently, convolution neural networks (CNNs) have achieved great success in machine learning applications including image classification, natural language processing, etc. To tackle challenges associated with memory capacity, energy, etc., binary neural networks (NNs) [62]–[64] have recently been proposed. In binary NNs, the weights and/or activations are binarized to ± 1. Such an approximation significantly

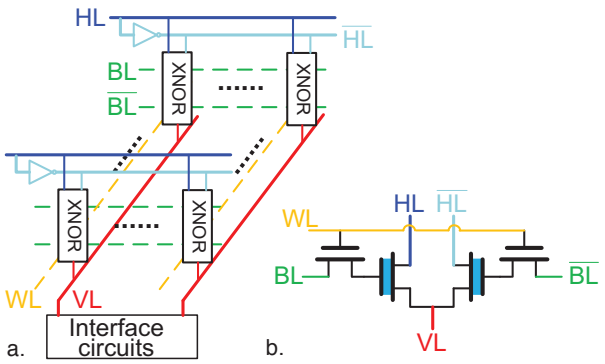


Figure 10. **(a)** FeFET crossbar and **(b)** XNOR cell.

reduces the energy, memory usage, and execution time, while still providing acceptable accuracy.

To further reduce energy and computation time, researchers have also considered the impact of emerging technologies. For example, RRAM-based crossbars can also perform analog multiplications, and are promising candidates for energy-efficient CNN accelerators. RRAM-based crossbars have been used for both conventional and binary NNs (e.g., [65]–[68] from just 2017). However, RRAM designs still face challenges such as sneak paths, high write energy, device variations, etc.

Unlike RRAMs, FeFETs are three-terminal devices. The three-terminal structure enables better control of both write and read power. A lower read voltage at the gate terminal can reduce I_{ON}, and in turn, read power can also be reduced. During programming, we can set V_{DS} to zero. This does not affect the programming operation, but I_{DS} is reduced to almost zero during programming. Write power is only consumed when charging the FE layer capacitance, which is much lower than that caused by the write current of RRAMs. A FeFET crossbar structure (from [16]) is shown in Fig. 10a. Each cell is a FeFET-based XNOR gate and is connected to a horizontal line (HL), its inverse ($\overline{\text{HL}}$), a vertical line (VL), a word line (WL), a bit line (BL), and its inverse ($\overline{\text{BL}}$).

The basic cell in a FeFET-based crossbar array (Fig. 10b) consists of two FeFETs and two access transistors. The two FeFETs (storing complementary bits) store one weight bit. One bit of the input and its complementary bit are applied to HL and $\overline{\text{HL}}$. The cell performs the XNOR operation between the input bit and the weight bit stored in the two FeFETs.

FeFET-based crossbar designs have been simulated using the HSPICE FeFET model from [21]. The 10 nm FinFET PTM (`tfin=8nm`, `hfin=21nm`, `nfin=1`) [69] is adopted for all MOSFET devices. The crossbar array size is 64×64. FeFET designs are compared with RRAM and CMOS equivalents. When compared with two RRAM-based designs, a FeFET-based approach enables write power reductions of 5600X and 395X. Read power can be reduced by 4.1X and 3.1X. (That said, read latency is 8% higher with the FeFET approach.) For a CMOS-based design, we can also set HLs and $\overline{\text{HL}}$s to zero when programming. Thus, write power can be extremely low due to the low programming power of SRAMs. The read power of the CMOS-based design is also the lowest because a MOSFET has a lower I_{DS} than a FeFET given the same width. However, the read delay in a CMOS design is the highest. Overall, the FeFET-based design is the best in terms of read power-delay product (PDP), and represents a promising future direction to explore and benchmark [16].

IV. Advances with FeFET Fabrication

We now consider FeFET-based devices and design at scale, and other recent experimental advances with FE-based devices.

A. Large-scale integration

The realization of complex functionality enabled by combining logic and memory at fine granularities requires the integration of both logic and memory devices in one common CMOS manufacturing process. Moreover, a competitive and flexible solution requires the compatibility of the novel memory technology with existing transistors in the respective base technology. Put another way, the electrical parameters of logic transistors (that are fixed in the respective process design kits) should not be altered by the extension of the technology. Furthermore, operating conditions of the memory devices (in terms of operation voltages and currents) should be readily facilitated by the adoption of the existing device suite.

Given the above context, the successful implementation of ferroelectric HfO_2-based FeFETs in a 28 nm gate-first super low power (28SLP) CMOS technology platform (first demonstrated in 2016 [12]) represents a significant step toward a broad application of FeFET device concepts/their unique features. More specifically, the HfO_2-based FeFET is constructed in a low-cost double-high-k manufacturing process by adopting just two additional DUV structural masks – and the electrical baseline properties of the logic transistors were not impacted by the additional processing steps. Moreover, additional processing steps did not increase the defect level as was demonstrated on a matured high-volume product. (The implemented FeFET module did not affect the D0-limited yield.) As such, the seamless integration of this novel device concept into state-of-the art CMOS technologies is possible. FeFET memory technology adds only negligible overhead to the CMOS baseline and, as the write operation is purely field-based and not current driven, low power writes are possible.

The successfully developed 28 nm SLP integration scheme enables a direct transfer of the FeFET module into more advanced technologies such as the 22FDX$_{TM}$ platform. Memory windows enabled by a 1.5 V threshold voltage shift via adoption of programming and erase voltages in the range of only 2-3 V were recently demonstrated in aggressively scaled FeFET cells with a footprint of just 0.025 μm^2 [41]. Thus, ferroelectric HfO_2 enables a scalable and CMOS compatible eNVM that keeps pace with the scaling demands of leading-edge logic technologies (Fig. 11). Further opportunities for FeFET operation/applications are offered by the back-bias option as a unique feature of the fully depleted SOI. Back bias enables an additional degree of freedom in the analog design of basic LiM building blocks, which enables the design of novel reconfigurable logic concepts (see below) [70].

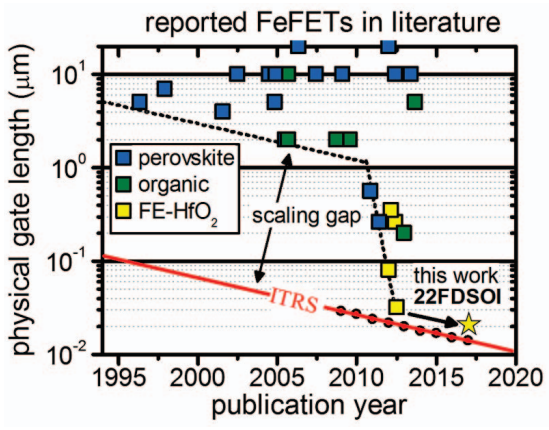
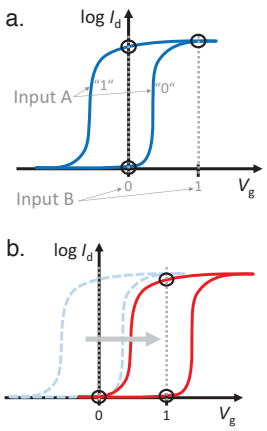


Figure 11. Physical gate length scaling of FeFET compared to the eNVM logic platforms.



Input A (Vth)	Input B (Vg)	Out (I_d)	Out (V_{out})
0	0	0	1
0	1	1	0
1	0	1	0
1	1	1	0

Input A (Vth)	Input B (Vg)	Out (I_d)	Out (V_{out})
0	0	0	1
0	1	0	1
1	0	0	1
1	1	1	0

Figure 12. Concept of a single-FeFET based sequential logic gate realizing reconfigurable **(a)** OR and **(b)** AND operations.

Various test structures – ranging from (i) single devices, (ii) 64kBit passive arrays, and (iii) up to large prototype memory arrays with memory sizes of up to 32Mb – have been manufactured. Fully functional 64kBit memory arrays were demonstrated by successfully writing a low-VT checkerboard after block erase. Moreover, the large array data based on $0.036\mu m^2$ 32Mb cell array results [41] reveal the successful realization of a disruptive embedded non-volatile memory (eNVM) process where the individual memory cell is constructed by a single FeFET device. As an FeFET cell can be erased via the source and drain regions [71], no dedicated bulk area is required for embedded FeFET cells, and the need for on-chip generation of negative voltages can be avoided.

B. Experimental Demonstrations

Using the memory array results discussed above as context, we now consider prospects for adopting existing, integrated FeFET technology toward other application-level ends. Results to date suggests that cycling endurance in integrated devices is in the range of up to 10^5 programming and erase cycles, which is certainly sufficient for the realization of embedded storage solutions. However, for fine-grained logic-in-memory circuits based on the available Generation-1 FeFETs, more careful study/analysis is needed. Logic operation (that includes programming and erasing steps under full circuit speed) would not be supported assuming an envisaged circuit lifetime of 10 years. That said, [44] suggests that higher endurance at the cell-level may be achievable via a gate-last process integration scheme. Given the above, two different mechanisms for adopting fully integrated FeFET devices are considered.

1) Reconfigurable AND/OR gates: For some applications, the state of a memory cell may only need to be changed occasionally. Similar to the work discussed in Sec. III-C, an FeFET can be adopted as a local storage element that enhances/complements computational tasks in a traditional von Neumann architecture. In [70], a single FeFET is employed in conjunction with a sequential logic operation. The internal polarization state of a FeFET is used as one input of the logic gate (input A), i.e., FeFET non-volatility is exploited to store one input. A high V_{th} or low V_{th} state corresponds to a logical zero or one, respectively. The second input is represented by the gate voltage V_g (input B) that is applied in a second step of a sequential logic operation.

A low or high voltage level of V_g represents a logical zero or one, respectively (Fig. 12a). As such, the drain current I_d is representative of a logic OR based on these two inputs. Moreover, a logic AND is possible by tuning the operating point of the FeFET by adjusting the back bias voltage (Fig. 12b). To realize a logic gate with voltage output (V_{out}), a pull-up device must be added in series to the drain terminal. For example, the PMOS pull-up transistor can operate in the subthreshold region to act as an adjustable resistor. The resulting V_{out} shows NAND or NOR behavior, respectively, resulting in an inverted output signal when compared to I_d. Thus, to implement the logic gate, only 2 transistors are needed, resulting in a very compact circuit layout. This work demonstrates fine-grained integration of logic and memory functionality, and represents a promising approach for the realization of future LiM hardware solutions and "normally-off"-computing. This may be especially useful when memory data is programmed infrequently but is used often, e.g., in field programmable gate arrays, CAMs, or digital filters.

2) FeFET-based synapse: Another approach for circumventing limitations associated with limited cycling endurance is gradual switching over a large number of programming steps. In this sense, FeFET technology may be well-suited to serve as a solid state synapse for neuromorphic computing models. Along these lines, (i) a single FeFET (integrated in a 28nm HKMG technology) and (ii) a resistive element (connected in series) was presented in [18]. In [18], the gradual and non-volatile switching of the ferroelectric hafnium oxide [72] is exploited to continuously tune the transistor channel's conductivity in an effort to mimic the synaptic weight update over a large amount of subsequent switching pulses). Devices having a comparably large channel length and width of 500 nm were employed (that exhibit a comparatively large amount of FE domains in the FE gate oxide). External voltage pulses can be applied either at the gate and/or source/drain/bulk

terminals in order to switch the polarization of the ferroelectric in a non-volatile manner, and subsequently tune the channel conductivity. By applying progressively increasing gate pulses, the multi-domain device is gradually brought from an initially high-VT state into a low-VT state. This circuit also exhibited spike time dependent plasticity (STDP).

FeFET-based synapse have also been considered by Jerry, et al. [17]. As additional motivation, note that the expanding utilization of neural networks in tasks such as image recognition and speech-to-text translation motivates the design of hardware systems capable of running networks at lower power and latency for both inference and *training*. Within the current CMOS framework, as network sizes continue to expand, weight values are increasingly required to be stored in off chip memory such as DRAM – where the energy consumption and training time of the neural networks can become limited by the off-chip memory access bottleneck. For a fully connected deep neural network, significant acceleration in training can be achieved by minimizing data movement by utilizing on-chip storage and performing the computation and weight updates at the same node. This entails the development of a crossbar compatible analog synaptic memory where the weight values are stored as the conductance of the synaptic memory element. In order to be capable of accelerating neural networks over the current CMOS framework, such a device should exhibit 1V, 1 nanosecond potentiation and depression programming pulses, a symmetric and linear conductance response with $\geq$32 conductance states ($\geq$5-bit), and a G_{max}/G_{min} ratio of $>$10 [18], [73], [74].

Again, by gradually tuning the remnant polarization of the ferroelectric gate oxide through successive weak SET/RESET pulses, one can induce gradual shifts in the transistor threshold voltage, and in turn creates a steady modulation of the transistor channel conductance (for a fixed gate voltage). Thus far, 5-bit (32 level) devices has been demonstrated. However, a non-identical pulse scheme is needed, which would require increased peripheral circuitry and latency to determine the update pulse for each device. However, the high-speed electric-field controlled switching of the ferroelectric (75ns [17]), and large G_{max}/G_{min} ratio of FET remain enticing characteristics.

V. Acknowledgments

Datta, Gupta, Hu, and Niemier were supported in part by (i) the center for Low Energy Systems Technology (LEAST), one of six SRC STARnet centers sponsored by MARCO and DARPA, (ii) the National Science Foundation under grant 1640081, and (iii) the Nanoelectronics Research Corporation (NERC), a wholly-owned subsidiary of the SRC, through Extremely Energy Efficient Collective Electronics (EXCEL), an SRC-NRI Nanoelectronics Research Initiative under Research Task ID 2698.004.They also acknowledge SRC GRC. NaMLab was supported by the European Fund for regional Development EFRD and the Free State of Saxony, Europe supports Saxony. They gratefully acknowledge support by GLOBALFOUNDRIES Fab1 LLC and Co. KG, Dresden, Germany. We also thank A. Naeemi and C. Pan.

References

[1] Semiconductor Research Corporation, "STARnet Research," https://www.src.org/program/starnet/, 2017.

[2] Semiconductor Research Corporation, "JUMP Research Announcement," https://www.src.org/compete/s201617/, 2017.

[3] "H2020-ICT-2017-1, ICT-31-2017-RIA; 3eFERRO: Energy Efficient Embedded Non-volatile Memory Logic based on Ferroelectric Hf(Zr)O2," 2017.

[4] T. N. Theis *et al.*, "In Quest of the Next Switch: Prospects for Greatly Reduced Power Dissipation in a Successor to the Silicon Field-Effect Transistor," *Proc. of the IEEE*, vol. 98, no. 12, pp. 2005–14, Dec 2010.

[5] A. C. Seabaugh *et al.*, "Low-Voltage Tunnel Transistors for Beyond CMOS Logic," *Proceedings of the IEEE*, vol. 98, no. 12, pp. 2095–2110, Dec 2010.

[6] A. Seabaugh *et al.*, "Steep slope transistors: Tunnel FETs and beyond," in *2016 46th European Solid-State Device Research Conference (ESSDERC)*, Sept 2016, pp. 349–351.

[7] C. Pan *et al.*, "Beyond-CMOS device benchmarking for boolean and non-boolean logic applications," *arXiv preprint arXiv:1711.04295*, 2017.

[8] Semiconductor Research Corporation, "nanoelectronic COomputing REsearch (nCORE)," https://www.src.org/compete/ncore/, 2017, [Online; accessed 06-September-2017].

[9] "H2020 FET Flagship Project: Human Brain Project," https://www.humanbrainproject.eu/en/, 2017.

[10] "NeuRAM Cube: NEUral computing aRchitectures in Advanced Monolithic 3D-VLSI nano-technologies," http://www.neuram3.eu/, 2017.

[11] S. Salahuddin, "Salahuddin NCFET Consortium," http://www.techdesignforums.com/blog/2016/04/07/intel-tsmc-globalfoundries-post-finfet-chenming-hu-uc-berkeley/, 2017.

[12] M. Trentzsch *et al.*, "A 28nm HKMG super low power embedded nvm technology based on ferroelectric FETs," in *2016 IEEE International Electron Devices Meeting (IEDM)*, Dec 2016, pp. 11.5.1–11.5.4.

[13] X. Li *et al.*, "Design of nonvolatile SRAM with ferroelectric FETs for energy-efficient backup and restore," *IEEE Transactions on Electron Devices*, vol. 64, no. 7, pp. 3037–3040, July 2017.

[14] X. Li *et al.*, "Enabling energy-efficient nonvolatile computing with negative capacitance FET," *IEEE Transactions on Electron Devices*, vol. 64, no. 8, pp. 3452–3458, Aug 2017.

[15] X. Yin *et al.*, "Exploiting ferroelectric FETs for low-power non-volatile logic-in-memory circuits," in *Computer-Aided Design (ICCAD), 2016 IEEE/ACM International Conference on*. IEEE, 2016, pp. 1–8.

[16] X. Chen *et al.*, "Design and optimization of FeFET-based crossbars for binary convolution neural network," in *Design Automation and Test in Europe (DATE)*, March 2018.

[17] M. Jerry *et al.*, "Ferroelectric FET analog synapse for acceleration of deep neural network training," in *2017 IEEE International Electron Devices Meeting (IEDM)*, Dec 2017.

[18] H. Mulaosmanovic *et al.*, "Novel ferroelectric FET based synapse for neuromorphic systems," in *2017 Symposium on VLSI Technology*, June 2017, pp. T176–T177.

[19] K. S. Li *et al.*, "Sub-60mv-swing negative-capacitance FinFET without hysteresis," in *2015 IEEE International Electron Devices Meeting (IEDM)*, Dec 2015, pp. 22.6.1–22.6.4.

[20] P. Sharma *et al.*, "Impact of total and partial dipole switching on the switching slope of gate-last negative capacitance FETs with ferroelectric hafnium zirconium oxide gate stack," in *2017 Symposium on VLSI Technology*, June 2017, pp. T154–T155.

[21] A. Aziz *et al.*, "Physics-based circuit-compatible SPICE model for ferroelectric transistors," *IEEE Electron Device Letters*, vol. 37, no. 6, pp. 805–808, June 2016.

[22] M. H. Lee *et al.*, "Prospects for ferroelectric HfZrOx FETs with experimentally CET=0.98nm, SSfor=42mv/dec, SSrev=28mv/dec, switch-off $<$0.2v, and hysteresis-free strategies," in *2015 IEEE International Electron Devices Meeting (IEDM)*, Dec 2015, pp. 22.5.1–22.5.4.

[23] Y. Katoh *et al.*, "Non-volatile fcg (ferroelectric-capacitor and transistor-gate connection) memory cell with non-destructive read-out operation," in *1996 Symposium on VLSI Technology Digest of Technical Papers*, June 1996, pp. 56–57.

[24] D. Wang *et al.*, "Ferroelectric transistor based non-volatile flip-flop," in *International Symposium on Low Power Electronics and Design*, ser. ISLPED '16. New York, NY, USA: ACM, 2016, pp. 10–15.

[25] S. Salahuddin *et al.*, "Use of negative capacitance to provide voltage amplification for low power nanoscale devices," *Nano Letters*, vol. 8, no. 2, pp. 405–410, 2008.

[26] J. Frougier *et al.*, "Phase-transition-FET exhibiting steep switching slope of 8mv/decade and 36% enhanced on current," in *2016 IEEE Symposium on VLSI Technology*, June 2016, pp. 1–2.

[27] N. Bagga *et al.*, "Demonstration of a novel two source region tunnel FET," *IEEE T. on Electron Devices*, vol. 64, no. 12, pp. 5256–5262, Dec 2017.

[28] A. I. Khan *et al.*, "Ferroelectric negative capacitance MOSFET: Capacitance tuning amp; antiferroelectric operation," in *2011 International Electron Devices Meeting*, Dec 2011, pp. 11.3.1–11.3.4.

[29] H. Ota *et al.*, "Fully coupled 3-D device simulation of negative capacitance FinFETs for sub 10 nm integration," in *2016 IEEE International Electron Devices Meeting (IEDM)*, Dec 2016, pp. 12.4.1–12.4.4.

[30] M. Hoffmann *et al.*, "Modeling and design considerations for negative capacitance field-effect transistors," in *2017 Joint International EUROSOI Workshop and International Conference on Ultimate Integration on Silicon (EUROSOI-ULIS)*, April 2017, pp. 1–4.

[31] A. I. Khan *et al.*, "Work function engineering for performance improvement in leaky negative capacitance FETs," *IEEE Electron Device Letters*, vol. 38, no. 9, pp. 1335–1338, Sept 2017.

[32] A. K. Saha *et al.*, "Ferroelectric transistor model based on self-consistent solution of 2D poissons, non-equilibrium greens function and multi-domain landau khalatnikov equations," in *2017 IEEE International Electron Devices Meeting (IEDM)*, Dec 2017.

[33] M. A. Alam *et al.*, "Physics-based compact models for insulated-gate field-effect biosensors, landau-transistors, and thin-film solar cells," in *Custom Integrated Circuits Conference (CICC)*, Sept 2015, pp. 1–8.

[34] C. C. McAndrew *et al.*, "Best practices for compact modeling in verilog-A," *IEEE Journal of the Electron Devices Society*, vol. 3, no. 5, pp. 383–396, Sept 2015.

[35] J. P. Duarte *et al.*, "Compact models of negative-capacitance FinFETs: Lumped and distributed charge models," in *2016 IEEE International Electron Devices Meeting (IEDM)*, Dec 2016, pp. 30.5.1–30.5.4.

[36] H. Asai *et al.*, "Compact model of ferroelectric-gate field-effect transistor for circuit simulation based on multidomain landau–kalathnikov theory," *Jap. J. of Appl. Phys.*, vol. 56, no. 4S, p. 04CE07, 2017.

[37] G. Pahwa *et al.*, "Compact model for ferroelectric negative capacitance transistor with MFIS structure," *IEEE Transactions on Electron Devices*, vol. 64, no. 3, pp. 1366–1374, March 2017.

[38] J. Müller *et al.*, "Ferroelectricity in HfO2 enables nonvolatile data storage in 28 nm HKMG," in *2012 Symposium on VLSI Technology (VLSIT)*, June 2012, pp. 25–26.

[39] J. Müller *et al.*, "Ferroelectric hafnium oxide: A CMOS-compatible and highly scalable approach to future ferroelectric memories," in *IEDM*, Dec 2013, pp. 10.8.1–10.8.4.

[40] H. Mulaosmanovic *et al.*, "Evidence of single domain switching in hafnium oxide based FeFETs: Enabler for multi-level FeFET memory cells," in *2015 IEEE International Electron Devices Meeting (IEDM)*, Dec 2015, pp. 26.8.1–26.8.3.

[41] S. Dünkel *et al.*, "A FeFET based ultra-low-power ultra-fast embedded nvm technology for 22nm FDSOI and beyond," in *2017 IEEE International Electron Devices Meeting (IEDM)*, Dec 2017.

[42] J. Müller *et al.*, "High endurance strategies for hafnium oxide based ferroelectric field effect transistor," in *2016 16th Non-Volatile Memory Technology Symposium (NVMTS)*, Oct 2016, pp. 1–7.

[43] "Ferroelectric hafnium oxide a game changer to FRAM?" in *2014 14th Annual Non-Volatile Memory Technology Symposium (NVMTS)*, Oct 2014, pp. 1–7.

[44] K. Chatterjee *et al.*, "Self-aligned, gate last, FDSOI, ferroelectric gate memory device with 5.5-nm hf0.8zr0.2o2, high endurance and breakdown recovery," *IEEE Electron Device Letters*, vol. 38, no. 10, pp. 1379–1382, Oct 2017.

[45] C. H. Cheng *et al.*, "Low-leakage-current dram-like memory using a one-transistor ferroelectric MOSFET with a hf-based gate dielectric," *IEEE Electron Device Letters*, vol. 35, no. 1, pp. 138–140, Jan 2014.

[46] A. I. Khan *et al.*, "Negative capacitance in short-channel FinFETs externally connected to an epitaxial ferroelectric capacitor," *IEEE Electron Device Letters*, vol. 37, no. 1, pp. 111–114, Jan 2016.

[47] J. Jo *et al.*, "Negative capacitance field effect transistor with hysteresis-free sub-60-mV/decade switching," *IEEE Electron Device Letters*, vol. 37, no. 3, pp. 245–248, March 2016.

[48] X. Li *et al.*, "Advancing nonvolatile computing with nonvolatile ncfet latches and flip-flops," *IEEE Transactions on Circuits and Systems I: Regular Papers*, vol. 64, no. 11, pp. 2907–2919, Nov 2017.

[49] H. Kimura *et al.*, "Highly reliable non-volatile logic circuit technology and its application," in *Multiple-Valued Logic (ISMVL), 2013 IEEE 43rd International Symposium on*. IEEE, 2013, pp. 212–218.

[50] S. George *et al.*, "Nonvolatile memory design based on ferroelectric FETs," in *Proceedings of the 53rd Annual Design Automation Conference*. ACM, 2016, p. 118.

[51] S. Borkar *et al.*, "The future of microprocessors," *Commun. ACM*, vol. 54, no. 5, pp. 67–77, May 2011.

[52] R. Balasubramonian *et al.*, "Near-data processing: Insights from a micro-46 workshop," *IEEE Micro*, vol. 34, no. 4, pp. 36–42, July 2014.

[53] C. E. Kozyrakis *et al.*, "Scalable processors in the billion-transistor era: Iram," *Computer*, vol. 30, no. 9, pp. 75–78, Sep 1997.

[54] J. Draper *et al.*, "The architecture of the DIVA processing-in-memory chip," in *Proceedings of the 16th International Conference on Supercomputing*, ser. ICS '02. New York, NY, USA: ACM, 2002, pp. 14–25. [Online]. Available: http://doi.acm.org/10.1145/514191.514197

[55] J. T. Pawlowski, "Hybrid memory cube (HMC)," in *HotChips 23*, 2011.

[56] S. H. Pugsley *et al.*, "NDC: Analyzing the impact of 3d-stacked memory+logic devices on mapreduce workloads," in *2014 ISPASS*, March 2014, pp. 190–200.

[57] M. M. S. Aly *et al.*, "Energy-efficient abundant-data computing: The N3XT 1,000x," *Computer*, vol. 48, no. 12, pp. 24–33, Dec 2015.

[58] A. Mochizuki *et al.*, "TMR-based logic-in-memory circuit for low-power vlsi*," *IEICE Trans. Fundam. Electron. Commun. Comput. Sci.*, vol. E88-A, no. 6, pp. 1408–1415, 2005.

[59] X. Yin *et al.*, "Design and benchmarking of ferroelectric FET based TCAM," in *2017 Design, Automation & Test in Europe Conference & Exhibition (DATE)*. IEEE, 2017, pp. 1444–1449.

[60] T. Kohonen, *Associative memory: A system-theoretical approach*. Springer Science & Business Media, 2012, vol. 17.

[61] M. Imani *et al.*, "Exploring hyperdimensional associative memory," in *2017 IEEE International Symposium on High Performance Computer Architecture (HPCA)*, Feb 2017, pp. 445–456.

[62] M. Courbariaux *et al.*, "BinaryConnect: Training Deep Neural Networks with Binary Weights During Propagations," in *NIPS*, 2015, pp. 3123–31.

[63] M. Courbariaux *et al.*, "Binarized Neural Networks: Training Deep Neural Networks with Weights and Activations Constrained to +1 or -1," *arXiv preprint arXiv:1602.02830*, 2016.

[64] M. Rastegari *et al.*, "XNOR-Net: ImageNet Classification Using Binary Convolutional Neural Networks," in *ECCV*, 2016, pp. 525–542.

[65] M. Cheng *et al.*, "TIME: A Training-in-memory Architecture for Memristor-based Deep Neural Networks," in *DAC*, 2017, pp. 26:1–26:6.

[66] P. Yao *et al.*, "Face classification using electronic synapses," *Nature Communications*, vol. 8, 2017.

[67] L. Ni *et al.*, "Distributed In-Memory Computing on Binary RRAM Crossbar," *ACM JETC*, vol. 13, no. 3, pp. 36:1–36:18, Mar. 2017.

[68] T. Tang *et al.*, "Binary convolutional neural network on RRAM," in *ASP-DAC*, Jan 2017, pp. 782–787.

[69] "Predictive Technology Model." [Online]. Available: http://ptm.asu.edu/

[70] E. T. Breyer *et al.*, "Reconfigurable NAND/NOR logic gates in 28 nm HKMG and 22 nm fd-soi FeFET technology," in *2017 IEEE International Electron Devices Meeting (IEDM)*, Dec 2017.

[71] S. Müller *et al.*, "Correlation between the macroscopic ferroelectric material properties of si: HfO2 and the statistics of 28 nm FeFET memory arrays," *Ferroelectrics*, vol. 497, no. 1, pp. 42–51, 2016.

[72] H. Mulaosmanovic *et al.*, "Switching kinetics in nanoscale hafnium oxide based ferroelectric field-effect transistors," *ACS Applied Materials & Interfaces*, vol. 9, no. 4, pp. 3792–3798, 2017.

[73] T. Gokmen *et al.*, "Acceleration of deep neural network training with resistive cross-point devices: design considerations," *Frontiers in neuroscience*, vol. 10, 2016.

[74] S. Yu *et al.*, "Scaling-up resistive synaptic arrays for neuro-inspired architecture: Challenges and prospect," in *2015 IEEE International Electron Devices Meeting (IEDM)*, Dec 2015, pp. 17.3.1–17.3.4.