Design of 2T/Cell and 3T/Cell Nonvolatile Memories with Emerging Ferroelectric FETs

Xueqing Li and Juejian Wu
Tsinghua University

Kai Ni
University of Notre Dame

Sumitha George
The Pennsylvania State University

Kaisheng Ma
Tsinghua University

John Sampson
The Pennsylvania State University

Sumeet Kumar Gupta
Purdue University

Yongpan Liu and Huazhong Yang
Tsinghua University

Suman Datta
University of Notre Dame

Vijaykrishnan Narayanan
The Pennsylvania State University

Editor's note:
This article explores the design of high-density, low-power, and high-speed embedded nonvolatile memory arrays exploiting the unique device characteristics of the emerging ferroelectric FETs.
—Vivek De, Intel Corporation

Random-access embedded nonvolatile memories (NVMs) have become a useful solution to getting rid of static leakage power and maintaining memory states without power in embedded memories [1], [2]. There is already a set of NVM solutions based on phase change random-access memory (PCRAM), resistive random-access memory (ReRAM), ferroelectric random-access memory (FeRAM), and spin-transfer torque magnetic random-access memory (STT-MRAM) [1]. Currently, these solutions still exhibit unsatisfactory features in terms of memory access energy and latency, CMOS process compatibility and scaling affinity, yield and/or endurance, peripheral circuitry complexity, etc. Further NVM exploration from device to architecture is thus critical for better performance and/or flexibility [2].

Digital Object Identifier 10.1109/MDAT.2019.2902094
Date of publication: 27 February 2019; date of current version: 29 May 2019.

Trends with FeFET NVM

Recent emerging ferroelectric FETs (FeFETs) innovated by new material and fabrication technologies enable a new era with appealing CMOS-scaling compatibility, moderate endurance (e.g., 10^{12}), and reduced operating voltage down to 1.5 V [3]–[5]. These have made FeFETs promising for array-style and distributed data storage [3], [4], and also for the emerging neuromorphic computing [6]. Conventional FeFET-based NVM has been based on the one-transistor per cell (1T/cell) structure [7]. The density is high but some disadvantages also exist, i.e., 1) write disturb to unaccessed

Copublished by the IEEE CEDA, IEEE CASS, IEEE SSCS, and TTTC

cells due to the lack of access shielding transistor, 2) supply overheads due to the use of multilevel voltage driving, and 3) write energy overhead due to the need of charging two bitlines. Another 2T/cell NVM array design was proposed in [8], which could isolate the write disturb with the extra write access transistor. However, [8] only modulates the FeFET gate voltage and not concurrently the drain and source voltages for write. This causes a higher, though not necessarily negative, gate-driving voltage range to set the memory state, resulting in lowered energy efficiency.

This article proposes new 2T/cell and 3T/cell NVM arrays, including 1) low-power write access that charges only one bitline (2T and 3T), 2) disturb-free write (3T), and 3) single-supply operation (3T). Simulations show that the 2T and 3T structures, respectively, reduce 50.8% and 30.4% write energy-delay product (EDP) of the prior 2T structure in [8].

Single-FeFET operating mechanisms

Figure 1a and b shows the conceptual FeFET of a fin structure, which is essentially a MOSFET with an extra ferroelectric gate insulator, such as doped hafnium dioxide. By increasing the ferroelectric layer thickness (T_{FE}) and engineering the MOSFET work function, hysteresis appears and may exhibit distinct ON and OFF states at zero gate–source voltage (V_{GS}) based on the ferroelectric material polarization [3]–[11]. Figure 1c shows typical FeFET channel conductance (G_{DS}) versus V_{GS} curves. Polarization switching can be accomplished by setting V_{GS} to set up a voltage across the ferroelectric layer beyond the coercive voltage.

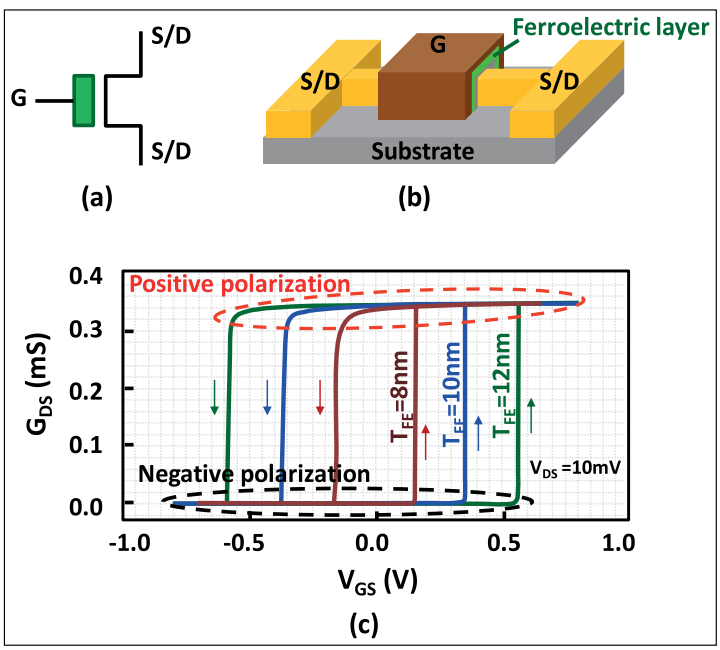

Figure 1. FeFET concepts. (a) N-type symbol. (b) Fin-structure FeFET device. (c) Typical hysteretic G_{DS}–V_{GS} (single-domain model) [8].

Read

The two typical nonvolatile G_{DS} states can show more than four orders of magnitude difference, leading to low-cost sensing schemes [4], [9]. The sharp transitioning also improves the noise margin. These advantages come from the unique FeFET features: 1) the settling-down transition behavior in the energy landscape as a passive amplification for V_{MOS} and 2) the internal MOSFET gain to the sensed I_{DS}.

Write

Different from memory devices such as ReRAM and STT-RAM, no static direct current (DC) current is consumed in FeFETs as V_{DS} can be 0 V. This feature provides higher energy efficiency when compared with the ferroelectric (Fe)-NAND flash memory, the prevention of injecting/ejecting electrons into/from the floating gate avoids the use of a much higher voltage [12].

Modeling

This article uses the calibrated FeFET model [11]. The polarization switching speed is modeled with a kinetic coefficient ρ. This article sets ρ typically between 0.05 and 0.25 as in recent works [9], [10].

Proposed 2T FeFET-based NVM

Circuit, and OFF and idle modes

Figure 2 shows the proposed 2T NVM scheme that has one bitline BL and two wordlines (WLW and WLR). In the power off mode, the bitlines and wordlines are grounded, and the FeFET stays with G_{DS} being high or low with V_{GS} = 0 V. With the power turned on, cells in rows, which are not being read or written, are in the idle mode with WLW voltage set to about $V_{DD}/2$ and WLR to GND. Note that V_{BL} could vary from GND to V_{DD} while other rows are being read or written, leading to V_{GS} between – $V_{DD}/2$ and $V_{DD}/2$. To prevent the idle-mode FeFET polarization state being flipped, the stable hysteresis region should cover this range through device-circuit codesign methods like tuning T_{FE}.

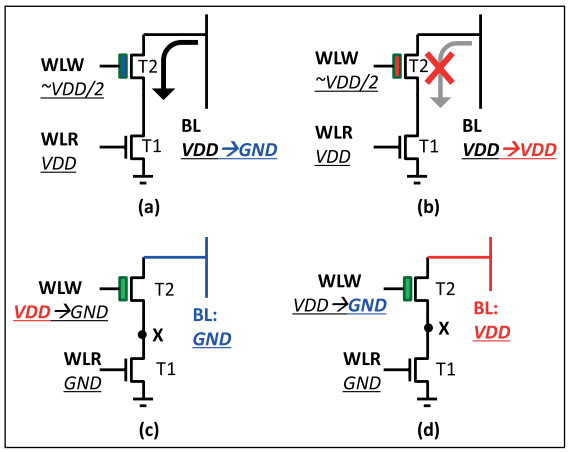

Figure 2. Proposed 2T/cell. (a) and (b) Voltage-mode read. (c) and (d) Voltage-mode write.

Read

The proposed 2T array supports random rowwise reads in both voltage-sensing and current-sensing modes. Figure 2a and b shows the voltage-sensing read operation with precharged BL. V_{WLW} is set to about $V_{DD}/2$ to keep the FeFET polarization unchanged. V_{WLR} is set to V_{DD} to turn on NMOS T1, and V_{BL} may remain high if T2 is in the OFF state, or discharge quickly otherwise. Such a difference could be conveniently sensed with a voltage amplifier sensing BL. It is noted that, for some memory devices such as ReRAM, voltage-sensing read can be possible but challenging if the OFF-state current is high. In this regard, the low OFF-state current of FeFETs is intrinsically superior to high OFF-state current NVM devices.

If the current-sensing read is adopted, the bitline voltage should be fixed, and the bitline current flowing through the cell in the selected row is sensed by a current-sensing amplifier. For large arrays, current-sensing read is helpful to reduce the latency by avoiding charging and discharging the large bitline capacitance. The capability of supporting both sensing modes provides more design flexibility.

Write

Figure 2c and d shows the proposed write setup. V_{BL} is set to GND and V_{DD} to write "0" and "1." V_{WLR} is set to GND to turn off NMOS T1. V_{WLW} is set to V_{DD} in the first phase and then GND in the second phase. Such a two-phase write enables writing different bit values in the same wordline.

Writing "0" and "1" occurs during the first and second phases, respectively. If the T2 polarization state was positive, that is, ON, before writing "1," the internal node X would have been charged to V_{DD} by BL through T2. Then, in the second phase, T2 biasing V_{GS} = –VDD triggers switching to negative polarization.

Transient simulation and analysis

One concern may be the initial state before a read or write operation. Demonstrably, this 2T scheme could handle the remnant charges left at the gate or internal nodes after a previous read or write operation. One key is that WLW stays at ~$V_{DD}/2$ before a subsequent read to prevent unwanted polarization switching when a read occurs after a "0" write.

Figure 3 shows the SPICE transients, including the polarization status and the voltage at the internal node X, T_{FE} is 8 nm. ρ is set to 0.25 as an example. V_{DD} is 0.6 V. Figure 3 shows all of the above-mentioned operations. It is noted that some coupled voltage glitches could be observed at node X but do not affect the operation functionalities.

Proposed 3T FeFET-based NVM

Circuit and OFF/idle/read modes

Figure 4a shows the 3T/cell configured in the voltage-sensing read mode. In power-OFF and idle modes, all wordlines and bitlines could be safely grounded. This 3T NVM could also reuse the bitline for higher density, as shown in Figure 4b. Despite the

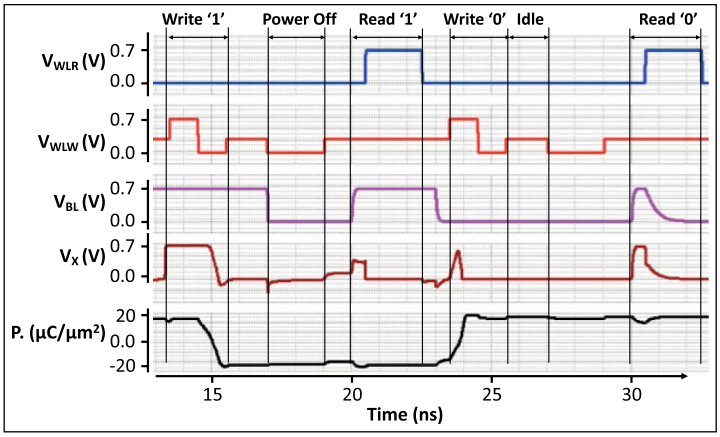

Figure 3. Proposed 2T/cell: transient waveforms. The bottom is the polarization.

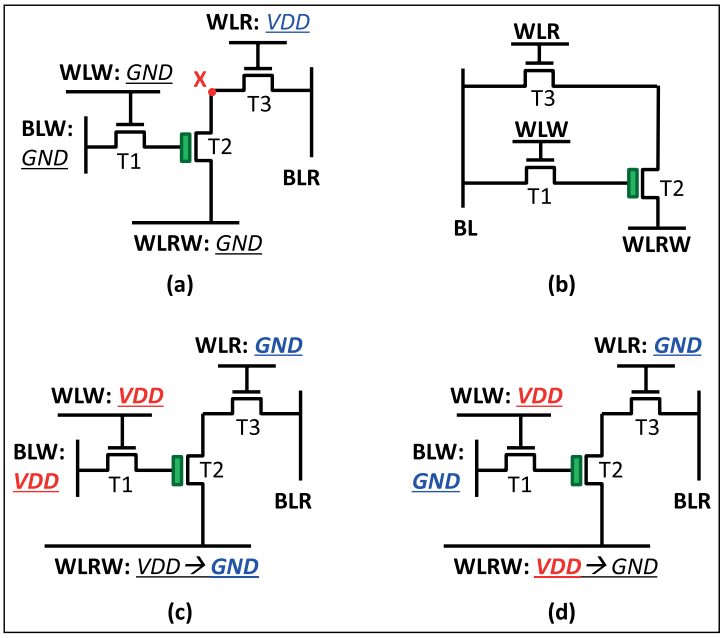

Figure 4. Proposed 3T/cell NVM. (a) Two-bitline design. (b) Single-bitline design. (c) Write "0." (d) Write 1."

circuit structure's similarity to the conventional 3T CMOS-embedded dynamic memory, the proposed 3T FeFET-based designs are fundamentally different, in both the data storage mechanism and the memory access method, as will be discussed in the following sections.

Both voltage-sensing and current-sensing read modes are supported. With voltage-sensing, T3 is turned on, and WLRW is set to GND. The precharged V_{BLR} will remain almost unchanged with an OFF-state T2 or will drop quickly to GND with an ON-state T2. Similar to the existing voltage-mode sensing schemes, voltage thresholding of V_{BLR} could provide the sensing result. In the current-mode sensing scheme, V_{BLR} is fixed and T3 is turned on. Thus, the current delivered by the cell could be sensed at the bitline, providing another meaningful option for energy-delay optimization in a larger memory array.

Write

The rowwise write operation for the 3T topology is shown in Figure 4c and d. T3 is turned off, and T1 is turned on. BLW is set to GND and V_{DD} to write "1" and "0." A two-phase voltage setting for WLRW is adopted, changing from V_{DD} to GND as shown in Figure 4c and d. Write operation for "1" and "0" occurs in the V_{DD} phase (with $V_{GS} = -V_{DD}$) and GND phase (with $V_{GS} = V_{DD}$), respectively. The write theory is similar to that of the 2T topology, which is discussed in the "Proposed 3T FEFET-based NVM" section. If the two bitlines are merged into one as shown in Figure 4b, the new bitline BL inherits the BLW setting as shown in Figure 4a. In this 3T design, the remnant charges of a read or write operation do not affect the functionality of a subsequent read or write operation, nor change the ferroelectric polarization to an incorrect state.

Benchmarking

Simulation settings

In the simulations, a 50-fF parasitic capacitor is assumed for each bitline. By default, the kinetic coefficient ρ is set to 0.1, and T_{FE} is 10.5 nm. The FeFETs are modeled with the 10-nm FeFET model, as was used in [11], and are calibrated with lead zirconate titanate (PZT) ferroelectric material, as was used in [8]–[10]. The MOSFETs, including the one embedded in the FeFET, are 10-nm predictive technology model (PTM) FinFETs.

Write performance evaluation

For a fair comparison, no negative supply is used. For the prior 2T NVM work in [8], the use of negative supply voltage of $-V_{DD}$ is mitigated by equally shifting up all supply and biasing voltages by V_{DD}. Note that the supply voltage range is $2V_{DD}$ in [8] regardless of whether a negative supply is applied. The proposed 2T design is also evaluated with T_{FE} = 12 nm for a wider FeFET hysteresis window to extend the supply voltage operation range to be similar to that of the 3T design.

Write energy and latency definition

The write energy is the average energy consumed to write "1" and "0" from a different prior state. The write operation latency covers a different period of time between the proposed 2T, 3T, and the prior 2T designs. For the proposed 2T and 3T designs that adopt two-phase write, it is the sum of latency in writing "0" and "1." For the prior 2T design in [8], the write operation latency is defined as the maximum latency to write "0" and write "1."

Figure 5a shows the simulated write energy per cell versus the write latency. In addition to low write energy and latency of the proposed designs, a few observations are manifest: First, a higher supply voltage leads

to less latency and more energy consumption. Meanwhile, thanks to the no-DC-current write, not more than 2% of the average write energy is consumed to switch the polarization. Second, the proposed 2T design with a higher T_{FE} of 12 nm has higher write latency than that of 10.5-nm T_{FE}, even under the same supply voltage. This is because of a different ferroelectric energy landscape and wider hysteresis width. Third, the proposed 3T design with 10.5-nm T_{FE} has the write energy-latency curve that lies in between the two curve segments of the proposed 2T design. In the higher voltage segment, it has lower latency than the proposed 2T design with T_{FE} = 12 nm, mainly because of an intrinsically faster FeFET. At the lower voltage segment, the proposed 3T design has more write latency than the proposed 2T design with T_{FE} = 10.5 nm, mainly because of the write mechanism differences: before WLW is effectively triggered, the proposed 2T design can use a prior read operation and BL is preset to set the desired voltage for both drain and source of the FeFET, as shown in Figure 2; whereas the proposed 3T design has neither the source nor drain of the FeFET preset (after a read operation, the internal node X in Figure 4a is desired to be GND but charged to V_{DD}, and the internal node X in Figure 4b is desired to be V_{DD} but was discharged to GND). Lastly, the proposed 2T and 3T designs have higher energy efficiency than the prior 2T design in write operations. This is mainly because of the prior 2T designs that require a doubled voltage range. In addition, it also results in higher voltage stress on the write-access transistor, which may induce a stability problem and also a power problem as the drain–body or the source–body interface diode may be turned on with a high voltage across it.

It is important to note that, the write latency is based on ρ set as 0.1. The latency as a function of the kinetic coefficient ρ is provided in Figure 5c. The write energy is almost constant, while the write latency shows a strong linear function of ρ. This is consistent with the FeFET model and prior reported results [9]. Further device design and fabrication efforts are expected to further improve the write speed.

Read performance evaluation

Here, the voltage-sensing read operations are evaluated. In this article, the read operation energy is the average energy consumed to read "1" and "0." As reading "1" does not noticeably change the bitline voltage, the read latency is actually equal to the latency for

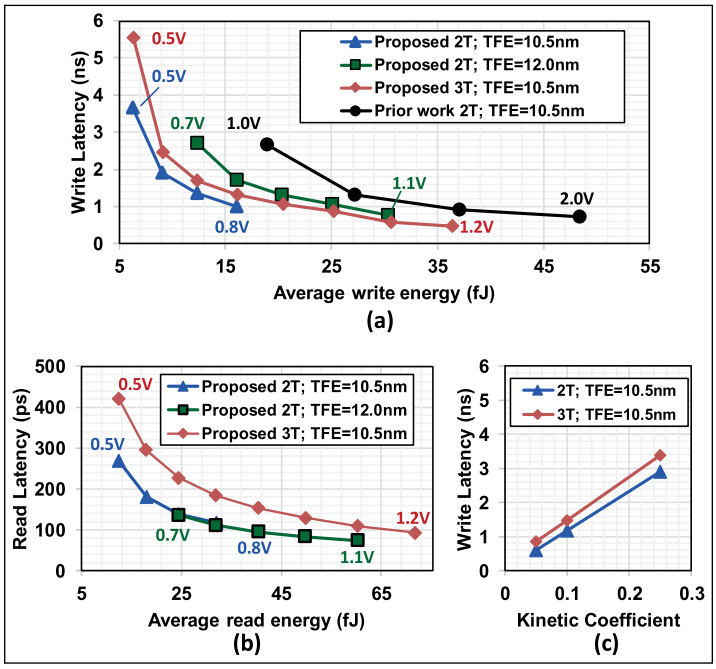

Figure 5. Performance evaluations. (a) Write energy latency. (b) Energy latency for voltage-sensing read. (c) Impact of the kinetic coefficient on write latency.

the read of "0," which is defined as the delay from the effective wordline-triggering point (with 50% voltage change) to the point when the read bitline voltage has reduced by 150 mV. Practically, a 150-mV voltage difference could be sensed with a voltage-mode amplifier without requiring a high gain.

Figure 5b shows the simulated read energy per cell versus read latency. Thanks to the ultralow OFF-state current and high ON-state current, ultralow read energy and latency are achieved. More conclusions could also be reached: First, most read energy is consumed in precharging the bitline capacitance. For read, precharging the bitline is required for both "0" and "1." In contrast, for write, precharging of the bitline is required only for writing "1" in 2T and 3T topologies. Consequently, more energy is consumed by read than write. Second, the read latency reduces as the supply voltage increases. This is because the read access transistors have lower resistance and discharge the bitline faster. For the proposed 2T design with different T_{FE}, the read latency difference is not quite significant at 0.7 and 0.8 V. This is because the discharging current is mainly limited by the read access MOSFET and that the FeFET is not the bottleneck element. Third, the proposed 2T design has relatively lower read latency than the 3T

design. This is because 1) the 2T design is biased at $V_{DD}/2$ at the FeFET gate while the proposed 3T design is biased with GND at the FeFET gate and 2) the 2T design always has $V_{GS} = V_{DD}$ for the access NMOS while the 3T design has $V_{GS} = V_{DD}$ only at the beginning point for the read access transistor T3.

As mentioned in the previous sections, it is also feasible to use a lower voltage to precharge the bitline for read to reduce the read energy consumption as long as the sensing scheme is not the bottleneck.

Benchmarking summary and more discussions

Table 1 summarizes the benchmarking results, where the comparison on the memory access performance, density, additional supply voltage requirement is shown. It is also interesting to consider the OFF-to-ready energy, which is the energy needed to wake up the NVM array from a completely OFF state to the idle state ready for read and write. For both the prior and the proposed 2T designs, considering some idle-state biasing voltage settings are nonzero, for example, $2V_{DD}$ for the prior work in [8] and $\sim V_{DD}/2$ in the proposed 2T array, the memory controller has to raise their voltage level accordingly and, thus, consumes extra energy. Such nonzero wake-up energy also exists in other NVM designs such as the access-device-free crosspoint structure ReRAM array whose bitlines and wordlines are biased at $\sim V_{DD}/2$ levels.

Meanwhile, the 3T design can operate with the highest supply voltage range and does not require an extra supply voltage of $V_{DD}/2$ or $2V_{DD}$, making it a good fit in scenarios when multiple supplies are not available. From a reliability perspective, both the proposed 3T and 2T designs can eliminate write disturb. In conclusion, providing both 3T and 2T designs enables a broader and more flexible optimization space.

Table 1. Comparisons between FeFET-based NVM designs.

Specifications	Prior 2T [8]	Proposed 2T	Proposed 3T
Transistor number per cell	2	2	3
Low voltage operation	No	Yes	Yes
Write energy (fJ) @min(EDP)	37	16	36
Write speed (ns) @min(EDP)	0.9	1.0	0.47
Energy delay product (fJ*ns)	34.6	16.0	17.0
OFF-to-Ready energy	4x	1x	0
Additional supply voltage	2VDD	~½ VDD	Not Needed

Future work

FeFET devices are promising to design embedded NVM. At the device perspective, future work on variation analysis, endurance improvement, and experimental demonstration is encouraged. At circuit and architecture levels, harnessing the logic-NVM fusion of FeFETs in a fashion of memory-centric computing is promising for future computing paradigms.

THIS ARTICLE HAS explored 2T/cell and 3T/cell NVM array designs using ferroelectric FETs. By harnessing the unique FeFET device characteristics with proper operation schemes, low-power, high-density, and high-speed embedded NVM solutions could be achieved. ■

Acknowledgment

We thank Prof. Sharon Hu, Prof. Sayeef Salahuddin, Prof. Asif Khan, and Prof. Peter Asbeck for device support and useful discussions. This work was supported in part by NSFC under Grant 61674094 and Grant 61720106013, in part by NSF Expedition, in part by SRC Centers (LEAST, ASCENT, and CRISP), and in part by BNRist and the Beijing Innovation Center for Future Chips.

■ References

[1] Y. Xie, *Emerging Memory Technologies: Design Architecture and Applications*, Springer, 2014.

[2] Y. Liu et al., "Ambient energy harvesting nonvolatile processors: From circuit to system," in *Proc. 52nd ACM/EDAC/IEEE Design Autom. Conf.*, 2015, pp. 1–6.

[3] S. Dünkel et al., "A FeFET based super-low-power ultra-fast embedded NVM technology for 22nm FDSOI and beyond," presented at the IEEE International Electron Devices Meeting, 2017.

[4] K. Ni et al., "SoC logic compatible multi-bit FeMFET weight cell for neuromorphic applications," presented at the 2018 IEEE International Electron Devices Meeting, Dec. 2018, pp. 296–299.

[5] Y.-C. Chiu et al., "One-transistor ferroelectric versatile memory: Strained-gate engineering for realizing energy efficient switching and fast negative-capacitance operation," in *Proc. 2016 IEEE Symp. VLSI Technol.*, Jun. 2016, pp. 1–2.

[6] K. Ni, B. Grisafe, W. Chakraborty, and A. K. Saha, "In-memory computing primitive for sensor data fusion in 28 nm HKMG FeFET technology," presented at the IEEE International Electron Devices Meeting, 2018, pp. 364–367.
[7] K. Ni et al., "Write disturb in ferroelectric FETs and its implication for 1T-FeFET AND memory arrays," *IEEE Electron Device Lett.*, vol. 39, no. 11, pp. 1656–1659, 2018.
[8] S. George et al., "Nonvolatile memory design based on ferroelectric FETs," in *Proc. 2016 53nd ACM/EDAC/IEEE Design Autom. Conf.*, Austin, TX, 2016, pp. 1–6.
[9] X. Li et al., "Enabling energy-efficient nonvolatile computing with negative capacitance FET," *IEEE Trans. Electron Devices*, vol. 64, no. 8, pp. 3452–3458, Aug. 2017.
[10] X. Li et al., "Advancing nonvolatile computing with nonvolatile NCFET latches and flip-flops," *IEEE Trans. Circ. Syst. I*, vol. 64, no. 11, pp. 2907–2919, Nov. 2017.
[11] A. Aziz et al., "Physics-based circuit-compatible SPICE model for ferroelectric transistors," *IEEE Electron Device Lett.*, vol. 37, no. 6, pp. 805–808, Jun. 2016.
[12] T. Hatanaka et al., "Ferroelectric (Fe)-NAND flash memory with batch write algorithm and smart data store to the nonvolatile page buffer for data center application high-speed and highly reliable enterprise solid-state drives," *IEEE J. Solid State Circ.,* vol. 45, no. 10, pp. 2156–2164, Oct. 2010.

Xueqing Li is an Assistant Professor in the Department of Electronic Engineering, Tsinghua University, Beijing, China. Li has a BS and a PhD from the Department of Electronic Engineering, Tsinghua University. He is a Member of the IEEE.

Juejian Wu is a Student in the Department of Electronic Engineering, Tsinghua University, Beijing, China.

Kai Ni is a Post-Doctoral Research Associate at the University of Notre Dame, Notre Dame, IN.

Sumitha George is currently pursuing the PhD degree from The Pennsylvania State University, University Park, PA. George has a BTech in electronics and communication from the University of Kerala, Thiruvananthapuram, India, and an MTech from the IIT Delhi, New Delhi, India.

Kaisheng Ma is an Assistant Professor at Tsinghua University, Beijing, China. Ma has a PhD with The Pennsylvania State University, University Park, PA.

John Sampson is an Assistant Professor in the Department of Computer Science and Engineering, The Pennsylvania State University, University Park, PA. He is a Member of the IEEE.

Sumeet Kumar Gupta is an Assistant Professor of electrical and computer engineering at Purdue University, West Lafayette, IN. He is a Member of the IEEE.

Yongpan Liu is an Associate Professor at Tsinghua University, Beijing, China. He is a Senior Member of the IEEE.

Huazhong Yang is a Full Professor at Tsinghua University, Beijing, China. Yang has a BS in microelectronics and an MS and a PhD in electronic science and technology from Tsinghua University. He is a Senior Member of the IEEE.

Suman Datta is the Stinson Chair Professor of Nanotechnology, University of Notre Dame, Notre Dame, IN. He is a Fellow of the IEEE.

Vijaykrishnan Narayanan is a Robert Noll Chair of Computer Science and Engineering and Electrical Engineering with The Pennsylvania State University, University Park, PA. He is a Fellow of the IEEE.

■ Direct questions and comments about this article to Xueqing Li, Tsinghua University, Beijing 100084, China; xueqingli@tsinghua.edu.cn.