80-kb Logic Embedded High-K Charge Trap Transistor-Based Multi-Time-Programmable Memory With No Added Process Complexity

Balaji Jayaraman, Derek Leu, Janakiraman Viraraghavan, Alberto Cestero, Ming Yin, John Golz, Rajesh Reddy Tummuru, Ramesh Raghavan, Dan Moy, Thejas Kempanna, Faraz Khan, Toshiaki Kirihata, *Senior Member, IEEE*, and Subramanian S. Iyer, *Fellow, IEEE*

***Abstract*— This paper describes the design and implementation of an 80-kb logic-embedded non-volatile multi-time programmable memory (MTPM) with no added process complexity. Charge trap transistors (CTTs) that exploit charge trapping and de-trapping behavior in high-K dielectric of 32-/22-nm Logic FETs are used as storage elements with logic-compatible programming voltages. A high-gain slew-sense amplifier (SA) is used to efficiently detect the threshold voltage difference (ΔV_{DIF}) between the true and complement FETs in the twin cell. Design-assist techniques including multi-step programming with over-write protection and block write algorithm are used to enhance the programming efficiency without causing a dielectric breakdown. High-temperature stress results show a projected data retention of 10 years at 125 °C with a signal loss of <30% that is margined in while programming, by employing a sense margining logic in the SA. Scalability of CTT has been established by the first demonstration of CTT-based MTPM in 14-nm bulk FinFET technology with read cycle time of 40 ns at 0.7-V VDD.**

***Index Terms*— Charge trap transistor (CTT), embedded memory, FinFET, high-K dielectric (HiK), logic compatible, non-volatile.**

I. Introduction

ADVANCES in silicon interposers, 3-D stacking, high-speed serial communication, and embedded memories have supplemented traditional device scaling by enabling integration of components at the module and chip level. Embedding dynamic-random access memory [1], [2], built with logic-compatible devices, with the processor, for example, has significantly increased on-chip cache memory density with wide IO, short latency, low power, and reduced soft-error rate, as demonstrated by the 24-core 120-MB shared L3 server-class microprocessors (μPs) [3]. But similar integration of non-volatile memory (NVM) into high-performance μPs, high-end application-specified integrated circuit, or foundry technology has not yet been achieved. Logic technology, particularly for high-performance μPs, has been fully optimized for performance, and hence, the slightest process modification or additional masks can be prohibitively expensive. Currently, integration of NVM is limited to low-density chip ID functions using 100% process compatible one-time-programmable memory (OTPM) using electrically blowable electrically programmable fuse (eFUSE) [4]. Future high-end very large scale integration logic chips will require 3-D, on-chip field-programmable logic, intelligent repair, chip personalization, and encryption strategies driving the need for scalable embedded high-density NVM solutions with multiple write function.

This paper describes a dense, secure, and rewritable non-volatile embedded multi-time-programmable-memory (MTPM) [5] implemented and tested across 32- and 22-nm silicon-on-insulator (SOI) and 14-nm Bulk FinFET CMOS technologies. A charge trap transistor (CTT) that exploits charge trapping behavior in high-K metal gate logic transistors is used as a non-volatile storage element with no added process complexity. This paper is organized as follows. Section II summarizes the existing embedded NVM solutions. Section III discusses the CTT technology used in this MTPM. Section IV describes the 32-nm high-K 80-kb CTT memory prototype and related circuit techniques. The discussion in Section V details the hardware results highlighting the use of the proposed circuit techniques, and Section VI demonstrates the scalability of the MTPM technology down to 14-nm FinFETs. Section VII summarizes the key aspects of the proposed technology.

II. Embedded Non-Volatile Memories

Multi-level-cell 3-D NAND flash memories [6], [7] and through-silicon-via technology [8] are expected to enable storage class memory [9], which combines system memory performance with storage class capacity. Other emerging non-volatile memories such as PCRAM [10], FeRAM [11], MRAM [12], and ReRAM [13] target high-performance or low-power non-volatile applications. Some of these memories have been successfully embedded into certain logic technologies with logic compatible but additional process steps.

Manuscript received June 25, 2017; revised September 26, 2017; accepted December 4, 2017. Date of publication January 9, 2018; date of current version February 21, 2018. This paper was approved by Associate Editor Vivek De. *(Corresponding author: Balaji Jayaraman.)*

B. Jayaraman, R. R. Tummuru, R. Raghavan, and T. Kempanna are with GLOBALFOUNDRIES Engineering. Pvt. Ltd., Bangalore 560045, India (e-mail: balaji.jayaraman@globalfoundries.com).

D. Leu, A. Cestero, M. Yin, J. Golz, D. Moy, and T. Kirihata are with GLOBALFOUNDRIES Inc., East Fishkill, NY 12533 USA.

J. Viraraghavan is with the Department of Electrical Engineering, IIT Madras, Chennai 600036, India.

F. Khan is with the Department of Electrical Engineering, University of California, Los Angeles, CA 90095 USA, and also with GLOBALFOUNDRIES Inc., East Fishkill, NY 12533 USA.

S. S. Iyer is with the Department of Electrical Engineering, University of California, Los Angeles, CA 90095 USA.

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/JSSC.2017.2784760

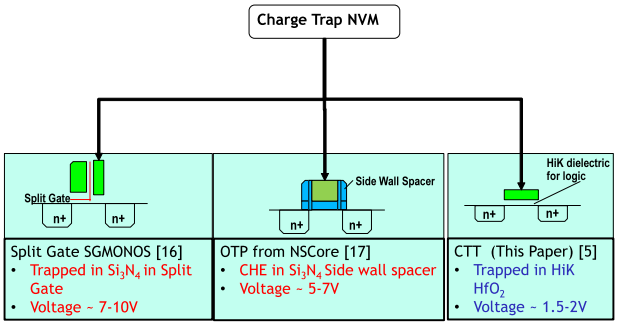


Fig. 1. Charge trap non-volatile memories.

Embedded flash technologies (eFlash) [14] have been successfully used in embedded automotive micro-controller-units. However, the eFlash memories require a dual polysilicon structure to create a floating gate. This results in additional process steps during front-end-of-line (FEOL) processing, which are not readily compatible with high-performance logic technologies optimized for 1-nm gate oxide (GOX) planar/FinFET devices. Embedded MRAM [12] and ReRAM [13] keep the logic technology FEOL process as is, and integrate non-volatile storage material in back EOL (BEOL) process, overcoming the device scaling problem. However, it may be difficult to integrate unique materials such as the magnetic or resistive elements into advanced technology with low-K and dual or triple phase shift process steps.

In general, regardless of the logic process compatibility, the additional process steps in either FEOL or BEOL may change not only the device performance but also the wiring, via resistances and capacitances. A change in any one of these parameters will require re-development of intellectual properties (IPs) as well as several process design kits (PDKs) for the system integration. IP/PDK re-development introduces additional cost and complexity altering the diverse array of third party IP. The additional process for integrating the NVM increases the process cost as high as ∼30%, which is not competitive unless the memory density requirement is large enough. In some cases, additional process tools may be required, further increasing the manufacturing overhead. The most important requirement for embedded NVM development is to keep the logic technology unaltered. It is also important to operate at logic compatible system voltages, because generating the high voltage is expensive, and even otherwise increases the process complexity to support high-voltage devices. Hence, most embedded applications, particularly with advanced logic technologies, use the well-established eFUSE approach [4] that exploits electrical migration or an anti-fuse approach based on oxide break down [15].

The memory described here [5] employs CTT as a rewritable non-volatile storage element. The cell structures and operating voltages of other known charge trap memories are compared against the CTT in Fig. 1. A split-gate (SG) (1.5 transistor) metal-oxide–nitride-oxide–silicon (MONOS) cell has been used in [16] for building a flash memory. The SG-MONOS cell has enhanced data retention with robustness against point defects. Charge trapping in side-wall spacer (typically a nitride layer) in an FET is used in [17] for non-volatile storage. This also exhibits good retention like SG-MONOS as the charge is trapped in a dielectric layer, but makes its scalability to FinFETs challenging. In both these techniques, the charges are trapped through capacitive coupling that includes a reasonably thick oxide tunneling layer to the nitride layer (storage layer) resulting in a high-programming voltage requirement (5–10 V). These technologies require additional process steps to create the rewritable non-volatile elements, which are not desirable. On the contrary, CTT memory traps a charge into the high-K dielectric (HiK) (gate insulator) of the existing logic FET. This results in lowering the programming voltage as low as 2 V with zero process adder. More details on CTT are provided in Section III.

III. Charge Trap Transistor Technology

High-K gate stacks, widely used in advanced nano-scale CMOS nodes and beyond, exhibit charge trapping and de-trapping behavior due to oxygen vacancies, Fig. 2(a), in the Hafnium oxide layer [18]. The retention behavior of immobilized charge in an insulating oxide layer is expected to show lower sensitivity to leakage defects than a conventional floating gate memory, where charge is stored on the conductive floating gate electrode, and hence, could cause a total charge loss from the electrode with the defect. The proposed CTT memory uses the HiK dielectric of the transistor as a charge trap storage element which results in 100% logic compatible process with no mask adder, while enabling ∼2-V voltage programming since the charge is trapped into HiK dielectric layer. Programming, Fig. 2(b), is achieved by electron injection into the gate dielectric, with an elevated gate voltage pulse (VGATE or wordline (WL) voltage) of 2 V and a high drain bias (VSL or source line (SL) voltage) of 1.5 V. This enables efficient and stable trapping of electrons in the HfO_2/interfacial layer, resulting in a stable threshold voltage (V_{TH}) shift. The typical V_{TH} shift (V_{THSFT}) in 32 nm is ∼200 mV with a ∼10-ms programming pulse and can be increased to ∼300 mV with a ∼100-ms pulse without causing catastrophic oxide breakdown. Unlike the positive bias temperature instability V_{THSFT} effect [19], life time >10 years at 125 °C can be realized as long as the right programming voltages are applied during the relatively short programming pulses [20], [21]. This programming range provides sufficient margin against inherent V_{TH} fluctuations (ΔV_{TH}) and retention loss to enable product designs that utilize typical repair solutions with reasonable overhead.

The memory element is erased, Fig. 2(b), by applying a VGATE of −1 V while holding VSL at ∼2 V. The trapped electrons get de-trapped by this reverse bias of −3 V. A small residual charge typically remains following de-trapping, such that V_{TH} may be higher than its native state resulting in hysteresis. This hysteresis, though indicative of an upper limit on the number of write–erase cycles, can be overcome with the implementation of the initialization technique [20] during the first few programming cycles as explained in Section V.

In order to use the CTT as an NVM, the programming has to be controlled so that the (V_{THSFT}) in the logic transistor lies in the "safe" zone, which is bounded above by GOX breakdown

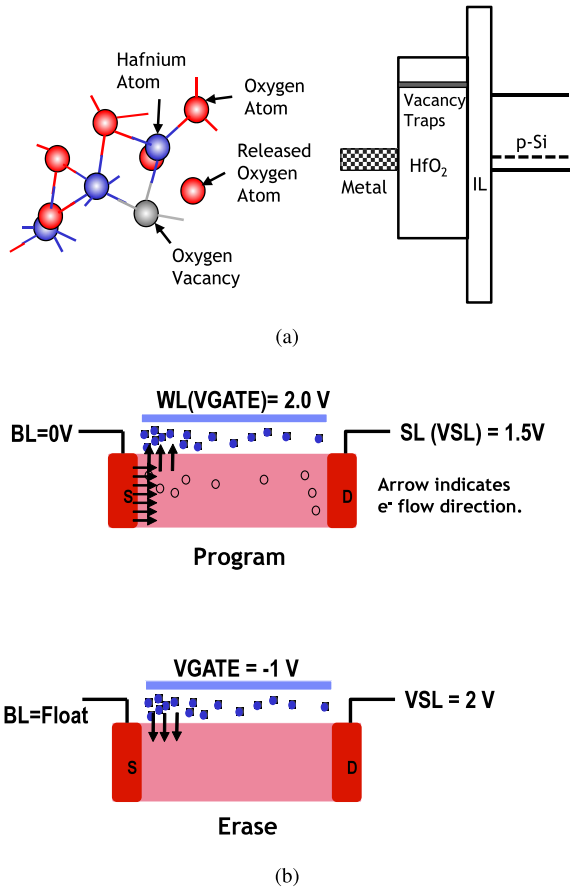


Fig. 2. CTT technology. (a) Charge trapping mechanism: intrinsic oxygen vacancies in HfO_2 [18]. (b) Program and erase operations on a HiK logic NMOS transistor.

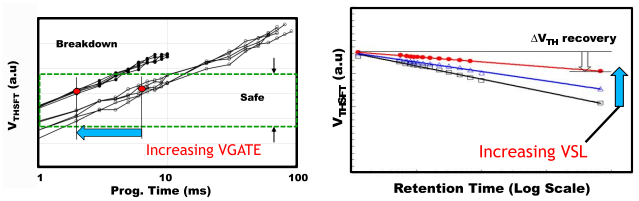


Fig. 3. CTT programming. (a) Increasing WL voltage improves programming time. (b) Increasing SL voltage improves retention.

and below by the ability of the sense amplifier (SA) to reliably sense the data. Fig. 3(a) plots the (V_{THSFT}) versus programming time demonstrating linear increase of (V_{THSFT}) with log time which eventually leads to dielectric breakdown, marking the boundary for the safe zone. Further, it also demonstrates how an increase in VGATE leads to significant speedup in programming time, providing a design knob to improve the programming speed. Another key figure of merit for NVMs is the retention time. Fig. 3(b) plots the (V_{THSFT}) over time for various VSLs. As expected, a fraction of the trapped charges get de-trapped over the life time of the product. However, an increase in VSL improves the retention over the life time of the product providing a second design knob to improve programming efficiency. Thus, circuit techniques that control the combination of VGATE, VSL, and programming time is a key to make the CTT a successful NVM technology.

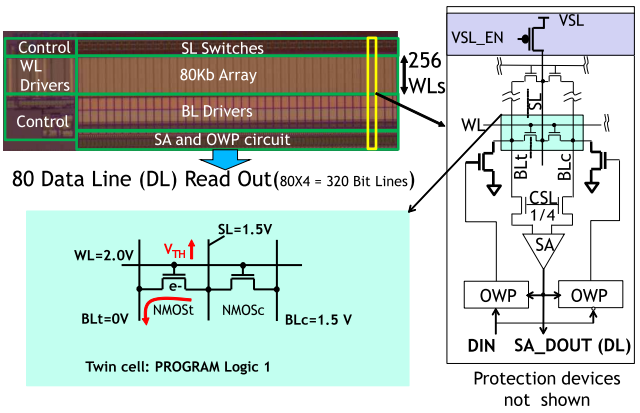


Fig. 4. 80-kb twin cell macro architecture.

IV. 32-nm 80-Kb CTT MTPM Prototype

In this section, we discuss the details of our CTT MTPM prototype and their design-assisted techniques. Twin cell array architecture with elevated SL, multi-step programming, block write algorithm, signal margin test-mode, and high-gain slew SA (SSA) has been developed.

A. Architecture

Fig. 4 shows the chip micro-photograph of the MTPM macro prototype implemented in 32-nm SOI Technology. It consists of an 80-kb memory array, WL drivers, a programming circuit [bitline (BL) drivers], SAs with over-write-protection (SA and OWP circuits), SL switch blocks, and their control circuits. The memory array uses an NOR-type NMOS array structure similar to a conventional NOR flash memory, except with the floating gate device replaced by a standard 1.2-nm thin GOX high-performance NMOS device. The channel dimensions of 208 nm width and 40 nm length provide adequate charge trapping efficiency. To maximize signal margin, the memory cell uses a twin cell approach consisting of true (NMOSt) and complement (NMOSc) transistors shown in Fig. 4. Storing "0" and "1" data is realized by trapping charge in the NMOSc and NMOSt devices, respectively, such that a V_{TH} difference between the pair (ΔV_{DIF}) can be sensed by high-gain SSA (discussed later) through a true/complement BL pair (BLt and BLc). This approach minimizes sensitivity to V_{TH} fluctuations across lot, wafer, and die. Although the twin cell layout consumes more area, the overall cell size is still attractive with a footprint of 0.109 μm^2.

The memory array is organized as 256 rows and 320 columns, resulting in 80 Kb density. The array is controlled by WL, SL, BLt, and BLc. Cells in each row are coupled to the corresponding WL that runs horizontally, in parallel to the gate poly (poly conductor). All NMOSt and NMOSc devices in a column are coupled to their respective BLt and BLc lines that run orthogonally to the WL. The column select switch (CSL) selects one of four BL pairs to couple to the SA, thus providing an 80-bit data-line (DL) read out. The SLs run parallel to the BLs, and couples to the SL switch at the top of the 80-kb array. A switch is provided per DL segment (four BLs) such that the SLs, only in selected segments, are raised to VSL and VDD during programming

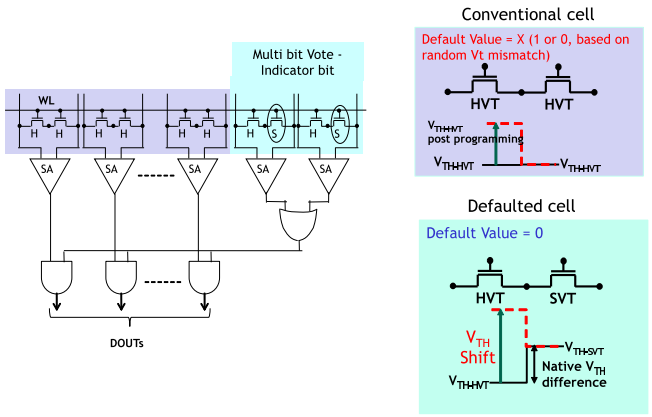


Fig. 5. Default bit emulation.

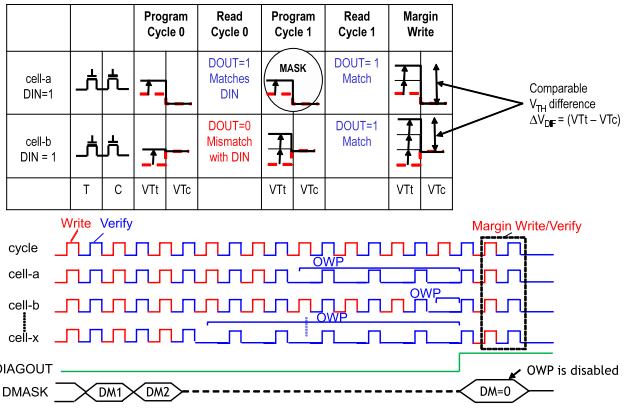


Fig. 6. Circuit assist—OWP.

and read, respectively, while grounding them in unselected DL segments. This selective SL approach with the switch reduces the V_{THSFT} recovery effect due to the SL voltage disturb for unselected cells. WL drivers include a voltage switch to select between a boosted voltage (VPP = 2.0 V) and a main voltage (VDD = 1 V) in programming and read modes, respectively.

One of the key applications of NVM is to enable redundancy in volatile memory arrays such as SRAM/DRAM. Conventional NVMs such as eFUSE and anti-fuse have a defaulted state when unprogrammed [4], and redundancy in the volatile memory array is exercised only when the NVM is programmed. However, the state of each cell for this CTT-based memory is random prior to programming or the following reset due to the random initial V_{TH} state of the transistors in the twin cell. In order to make it easy to use, the MTPM emulates a default state, without programming, using an indicator bit per 78 bits, Fig. 5, which is realized using twin cell with different V_{TH} flavors: high V_{TH} (HVT) and super high V_{TH} (SVT). In the right side of Fig. 5, we have shown a conventional HVT-HVT cell versus a defaulted HVT-SVT cell. For the conventional cell, in its native state (unprogrammed), the data that are read out could be random (either 0 or 1, and hence, shown as $\times$ (unknown)) based on the initial Vt mismatch between the true and complement FETs. For the defaulted HVT-SVT cell though, because of inherent Vt delta (about 90 mV, approximately 4.5-Vt sigma) prior to programming, we would read a 0 by default. The default bit gates each bit on the WL to read all zero. The HVT N-type field effect transistor (NFET) can be programmed to have a larger V_{TH} than the SVT device in the first programming cycle. Since the signal of the indicator bit is halved due to the V_{TH} asymmetry, multiple cells are used for the indicator bit.

This chip used an OR logic to improve the margin for "1" for default cell as shown in Fig. 5. The expected default state of 0 is known at manufacturing time and any row that does not read the default state of "0" will be fixed with redundancy. When we wish to indicate the programmed state, only one of the default cells need to read a logic high post programming since we are using the OR logic. We have implemented OR function for two bits as an indicator per 80 bits for 32-nm chip (78 conventional cells +2 defaulted cells), resulting in $\sim$2.5% area penalty. Alternatively, we can select multiple of the bits and couple them to the same SA to emulate a larger cell (larger width). Since V_{TH} sigma goes inversely as square root of area, a 4$\times$ sized defaulted cell effectively experiences only 1/2 the variation, and hence, introducing a similar signal margin compared to the twin cell is more than sufficient.

B. Multi-Step Programming Method With Over-Write Protection

Prior to programming, the required amount of V_{THSFT} varies from cell to cell because of random V_{TH} mismatch of the twin NMOS pair. Subsequent programming cycles may have to deal with similar cell-to-cell variability due to erase hysteresis. Moreover, increasing the V_{THSFT} too much not only increases the risk of the device breakdown, but also makes it difficult to erase for subsequent programming. We now discuss how we can introduce the required amount of V_{THSFT} across the entire array taking into consideration all the above-mentioned factors.

Fig. 6 shows an example where the native V_{TH} configuration of two cells *a* and *b* is such that cell-a has a higher V_{TH} on the true side (T) than the complement (C), and vice versa for cell-b. The initial V_{TH} differences could be due to random within-die V_{TH} variations, or hysteresis due to previous program-erase cycles. For improved write speed, it is necessary to program multiple cells (both in this example) at the same time. Here, we wish to program both cells to store a logic 1, which requires us to increase the V_{TH} of NMOSt. Programming NMOSt in both cells to the same extent may either cause an insufficient V_{TH} shift in cell-b (with an initially lower V_{TH}), or can potentially damage the dielectric of cell-a (with an initially higher V_{TH}). In order to overcome the problem, the programming is realized by using multiple short write cycles. In each cycle, the MTPM reads the cell to verify it against the required input bit and disables subsequent write operations if the verification result matches, thus avoiding a condition of over-programming. This method of protecting sufficiently programmed bits is called OWP. Subsequent programming is progressively enabled for fewer and fewer bits every cycle until all bits have been successfully programmed. The signal DIAGOUT goes high when all cells have been successfully programmed, as seen in Fig. 6. We have included a maximum time for the write cycle. This may cause

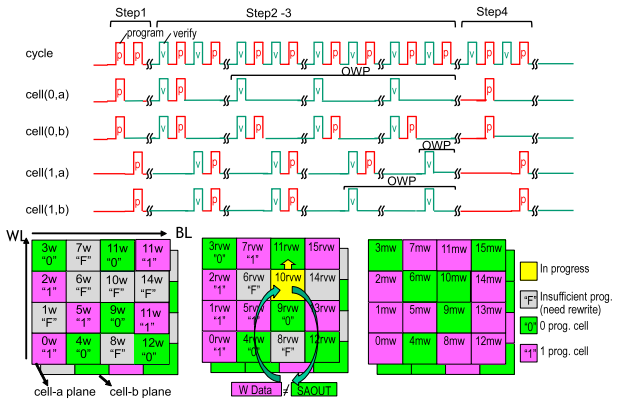


Fig. 7. Block write algorithm.

a small fraction of the bits to be left under-programmed after a pre-defined maximum number of programming cycles due to defects. These defects should be repaired with error correction codes (ECC). The current design does not include ECC logic to generate the DIAGOUT. However, by including ECC logic before computing the DIAGOUT logic, the system performs normally, by giving the write command and waiting asynchronously until DIAGOUT goes high. The macro supports a data mask (DM) function for masking the specific bits that have been sufficiently programmed, during subsequent programming steps. A final margin write, by disabling OWP, is performed to ensure that sufficient signal margin has been introduced. During the margin write operation, all the DM bits are made 0, as seen in Fig. 6, in order to ensure that all the cells are written. Actual circuit implementation, Fig. 4, is realized by holding the data from the SA and feeding back for comparison in the subsequent program operation, through the OWP circuit, without having tester-controlled DMASK signals. The SA is used as a latch avoiding area penalty for storing the data bit. The chip includes the function to bypass the comparison for defective cells. This technique not only ensures that the devices operate in the safe zone, but also causes similar V_{TH} difference (ΔV_{DIF}) to be introduced in all cells irrespective of the native V_{TH} state of the twin cells. This OWP programming method realizes 80-bit parallel programming (or 1/DL), without causing a cell dielectric damage while satisfying a sufficient V_{TH} shift. Thus, OWP avoids unintentional over-programming of previously programmed bits while programming the under-programmed bits.

C. Block Write Algorithm

In the aforementioned multi-step program-verify algorithm, verifying immediately after the write may cause a false OWP, as transient shallow trap electrons (unusable for steady state V_{TH} shift) tend to de-trap slowly. A block write algorithm, Fig. 7, has been developed to overcome this limitation while also considering the impact of thermal and voltage disturb, described next, during programming.

Programming one cell can affect adjacent cells through shared diffusion and shared SL. This disturb mechanism can alter the V_{TH} of adjacent cells post programming, and hence, should be accounted for appropriately. Thermal disturb occurs along the WL direction through the shared diffusion. When a cell is being programmed, cells connected to the adjacent WLs on the same BL see an increase in temperature thus inadvertently altering the neighboring cells V_{TH}. Voltage disturb occurs along the column direction through the shared SL in the DL segment (four BLs). When a cell is being programmed, cells on the remaining three columns, on the same DL, experience a reverse drain (source)—gate bias since the shared SL is at 1.5 V and the inactive WLs are grounded. This again leads to a partial erase of the adjacent cells. The voltage disturb phenomenon is experienced only by cells on the same DL since a SL switch is used to isolate remaining cells on the other DLs. This voltage disturb may cause address pattern sensitivities. The goal of the block write algorithm is to make sure that the programmed cells see the same voltage disturb while reducing the thermal disturb pattern sensitivities. The key is to write data in a block as opposed to a single row or column. A four-step block write algorithm, Fig. 7, is introduced.

1) *Initial Write:* Program multiple cells in a pre-determined block sequentially (across WLs and BLs in a block).
2) *Read-Verify-Write:* Sequentially verify the written bits, and rewrite only those bits that fail to read the expected value.
3) *Read-Verify-Write-Loop:* Repeat step 2 process "*m*" times.
4) *Margin Write:* Perform a final margin write.

An example of a block write algorithm (timing diagram and address map) is shown in Fig. 7. For simplicity, the following discussion assumes that the pre-determined block size is two planes of 4 WLs $\times$ 4 BLs, where the two cells having the same address number (0 to 15) from plane-a for DLa and plane-b for DLb are programmed in parallel. Step 1 starts with an address 0, executing a write operation in cell(0, a) (cell for address 0 and DLa, or plane-a), and the cell(0, b) (cell for address 0 and DLb, or plane-b), which is as indicated by "0w" in address map. Subsequently, the address is incremented, executing a write operation in cell(1,a) (cell for address 1 and DLa, or plane-a), and the cell(1,b) (cell for address 1 and DLb, or plane-b), which is as indicated by "1w" in address map. This initial write cycle operation is sequentially performed on all other cells across the four WLs and four BLs in each plane-a and plane-b (or for DLa and DLb) as a block. In step 2, while following the step 1 addressing and parallel write sequence, we perform a read-verify-write step indicated by "0rvw" starting from the cell (0,a) and cell (0,b). The write is performed only if the read data does not match the intended write data. This is repeated till the entire block has been programmed correctly. A block is chosen to be four BLs (one DL) wide because they share the same SL switch where programming one cell can voltage-disturb other cells. The neighboring DLs have their own SL switches. There would still be a pattern-dependent thermal disturb, but creating a loop in a block allows us to verify/characterize the effect of disturb, except for the cells at the block boundary along the row. Thus, employing a block write algorithm helps in reducing the disturb-related signal margin issue. As shown in the timing waveform, the number

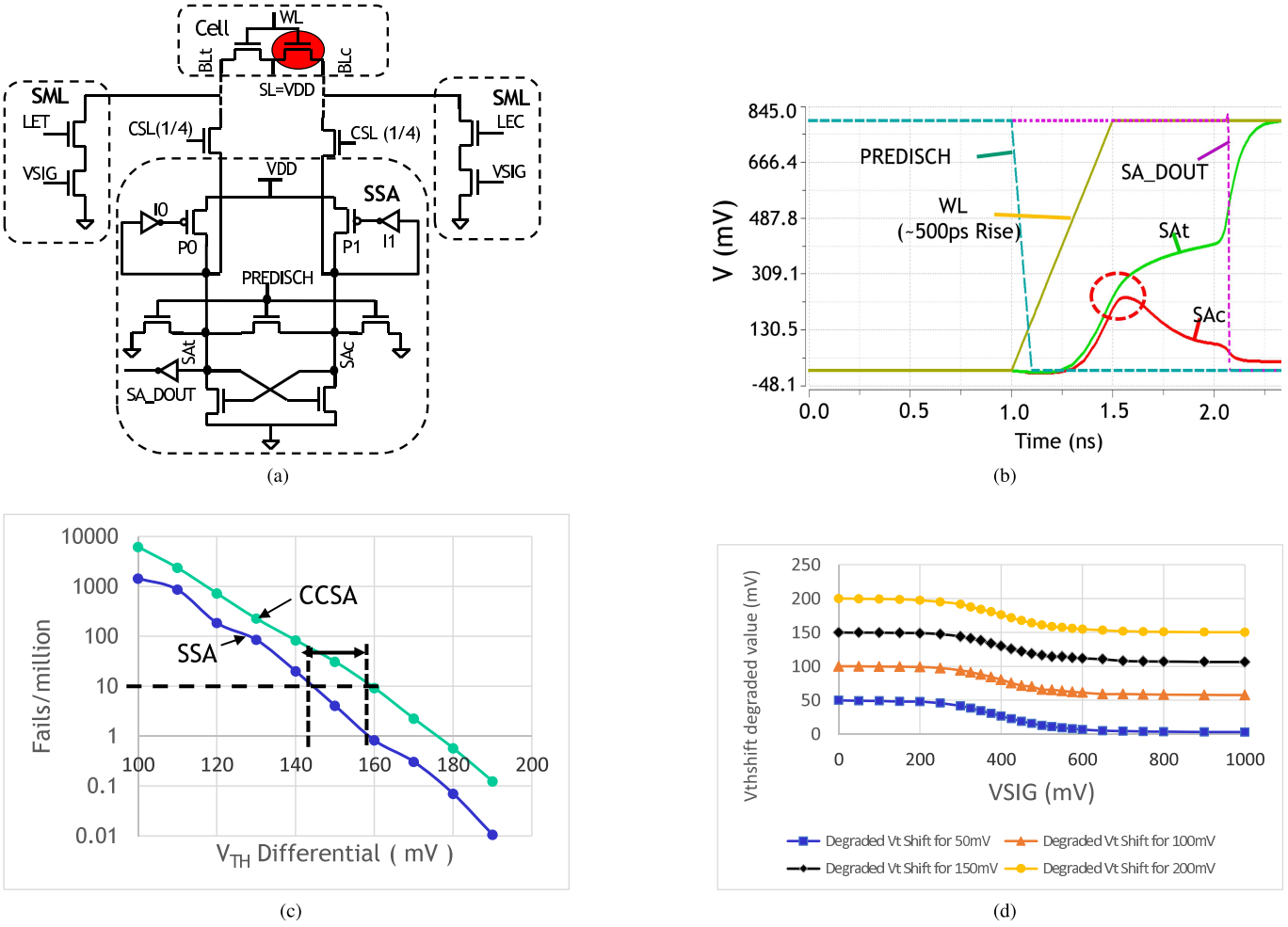


Fig. 8. SSA. (a) Circuit schematic. (b) Timing waveform. Increasing SL voltage improves retention. (c) SSA versus CCSA fail count. (d) SML logic simulation.

of the actual write cycles varies depending on the internally controlled OWP function by the results of the read verification. Finally, in step 4, a margin write as indicated by "mw" is unconditionally performed on all cells to ensure sufficient margin for read. With process variations, we could have a cell that matches the data input to be written even before programming. This is not necessarily bad, as this could result in high retention, especially if the inherent threshold voltage delta between the true and complement FETs is sufficiently large. The mw that is carried out in step 4, along with the signal margin that is built in using SA margining logic (SML) discussed in the next section, ensures sufficient V_{TH} delta to be incorporated during programming.

We could have a false pass during each of the intermediate steps of the multi-step programming. There are four kinds of false passes: defect driven, shallow-trap driven, insufficient margin driven, and noise driven. The defective cells must be repaired using redundancy during manufacturing. Further, the defective cells must be masked using the DM so that DIAGOUT signal works as expected, as explained in section IV-B. The shallow-trap-related false passes are avoided by using block write-verify model, as this allows for a long

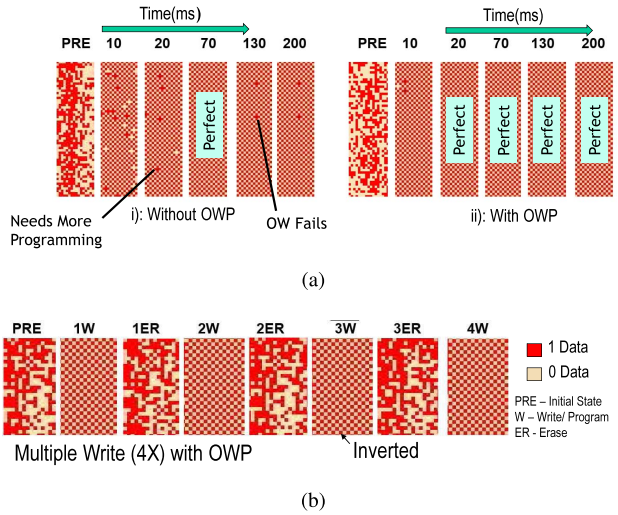


Fig. 9. 32-nm hardware results. (a) OWP effectiveness. (i) Without OWP. (ii) With OWP. (b) Multi-time programming.

pause between write and verify operation. A block size satisfying ~70-ms interval was used to allow shallow trap to de-trap into the channel, which should be optimized while considering

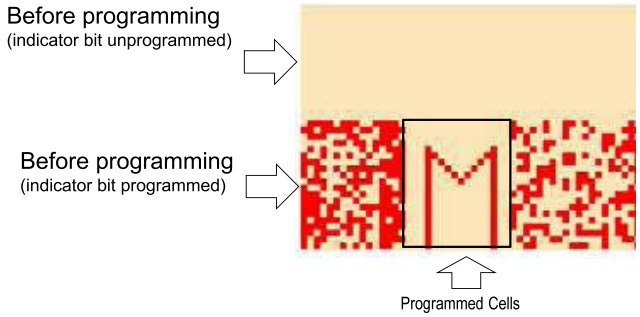


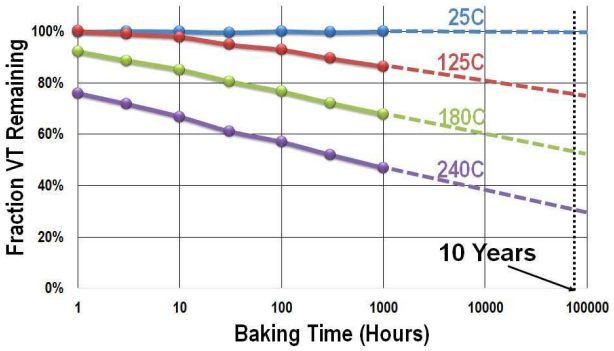
Fig. 10. 32-nm hardware results—default bit.

the block size requirement. The false pass arising due to noise in one cycle will fail in the subsequent programming cycles, thus allowing the corresponding cells to get programmed if this is caused by insufficient programming. Any remaining false pass cells should be repaired using ECC.

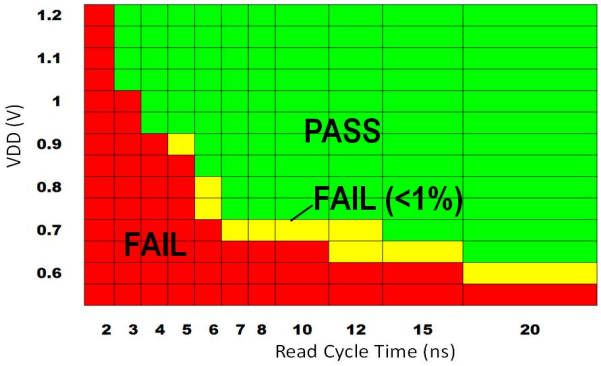
D. Slew Sense Amplifier

A high-gain SSA, Fig. 8(a), is used to sense the V_{TH} difference between the twin cells. The SL is switched to VDD and the CSLs are enabled, connecting BLs (BLt and BLc) to SA nodes (SAt and SAc), respectively. As shown in Fig. 8(b), SA nodes are pre-discharged to ground (GND) and left floating while simultaneously ramping the WL to VDD (1 V) to enable charging of SAt and SAc through the twin cells. As the WL ramps up, the NMOS with lower V_{TH} turns on in saturation, while the pair device is still in sub-threshold region. This results in an abrupt increase in current, and hence, in voltage on one side (SAt or SAc), turning on the pull up PMOS P0/P1 through inverters I0/I1, respectively, while the cross-coupled NMOS stack drives the other side low. The SAt value is inverted to obtain the SA output. The key advantage of this sensing scheme is that the charging current is directly determined by the cell and it exploits the order of magnitude difference between sub-threshold current and ON current to improve V_{TH} sensing. For the same footprint and failure rate (10 per million), Monte Carlo simulation for 32-nm bit-cells with random local variations, Fig. 8(c), shows that the SSA can sense ~10% less V_{THSFT} compared with the cross-coupled SA (CCSA) [22].

The programmed V_{TH} shift can reduce over the life time of the product (~10 years) and needs to be emulated while qualifying the chip. SML, as shown in Fig. 8(a), is used on BLt and BLc to emulate this V_{TH} loss. An NFET controlled by an analog pad VSIG is used to provide a discharge path to GND opposing the charge up by the cell and is enabled on the un-programmed side using leakage enable true (LET)/leakage enable complement (LEC) control signals to emulate V_{TH} degradation. Further, VSIG bias can be adjusted to factor in appropriate margin into the cell at the time of programming. As observed from Fig. 8(d), simulations on 32 nm predict that up to 50 mV of Vt shift can be margined by applying a suitable VSIG bias. Thus, once the cell is programmed, the data is read out with suitable VSIG bias applied, in the multi-step programming with OWP flow. This ensures that sufficient margin is incorporated in all the programmed cells, so as to have adequate delta V_{TH} at the projected end of life.



(a)



(b)

Fig. 11. 32-nm hardware results. (a) High-temperature storage stress results. (b) Voltage shmoo test results.

V. Hardware Results

Fig. 9(a) shows the bitmap change for the multi-step verification checker-board (CKB) pattern write on 32-nm hardware for two cases: 1) without OWP and 2) with OWP. The step size was 10 ms and the verification write cycle was repeated up to 20 cycles using 4-kb block write pattern. Without OWP, Fig. 9(a)(i), a perfect bitmap is achieved at 70 ms and shows overwrite fails (OW fails) for multi-step writes >130 ms as some cells enter the V_{TH} breakdown zone. However, Fig. 9(a)(ii), with OWP circuit shows a perfect bitmap at 20 ms and no fails there on, with OWP successfully kicking in to protect all the cells. Though these are two different chips, the fact that a perfect bitmap was achieved much earlier with OWP as opposed to without it, is in general true across all chips. With OWP, as the number of cells to be programmed keeps reducing with successive programming cycles, the overall programming VSL current, and hence, the IR drop on the VSL grid reduces. This in turn increases the effective VDS seen by the cells that are getting programmed, thus increasing the Vt shift per programming cycle, and enabling more efficient programming for cells that actually require more V_{TH}. However, it is also necessary to control the target of the WL and SL voltages depending on the temperature and process to ensure that the FET is operated within the safe operating window, requiring tunable voltage generators for the actual product. These dc generator circuits use IO voltage supply and can be designed using the

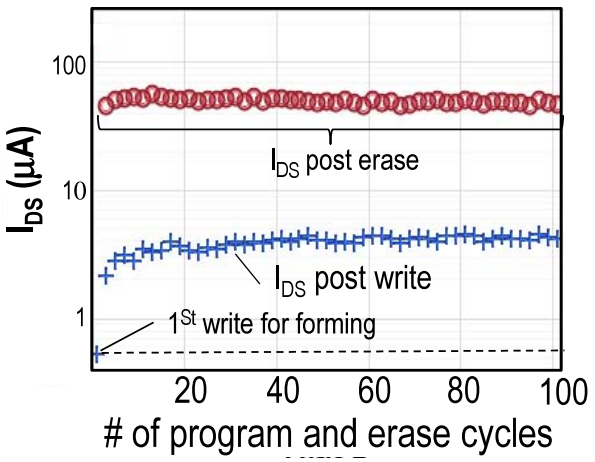


Fig. 12. Multi-time programming in 22 nm.

thick-oxide IO devices available for logic technology, resulting in zero process adder to the existing logic technology while maintaining system compatibility. The bitmap in Fig. 9(b) on 32-nm hardware shows 4× write functionality of a 4-kb CKB pattern erased and written four times with an inverted CKB pattern achieved in the third write cycle. Each CKB program is achieved using the multi-step write with OWP.

Bitmap in Fig. 10 shows the default bit in action. Before programming, the entire sub-array space reads a logic 0, as indicated by the beige squares, since the indicator bits have not yet been programmed, and hence, default to a zero. Post programming the indicator bit, the region reads random zeros and ones (squares) while the intended region can be programmed accordingly, to read the character "M" in this case. High-temperature (25 °C, 125 °C, 180 °C, and 240 °C) bake tests, Fig. 11(a), for 1000 h using 32-kb module show a projected 10 year V_{TH} degradation of <30% at 125 °C. The activation energy (E_a) was extracted using the conventional Arrhenius model as shown in the following equations. Using the life times at 70%, the extracted E_a is ∼1.35 eV

$$t = A \exp\left[\frac{E_a}{k_B}\left(\frac{1}{T(K)}\right)\right]$$

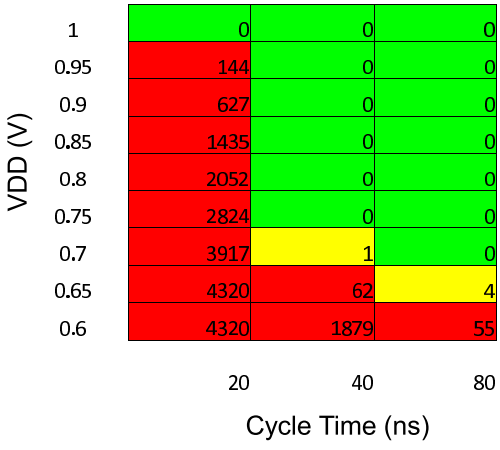
$$\ln(t) = \left(E_a \times \frac{1}{k_B T(K)}\right) + C$$

$$E_a = \text{slope of } \ln(t) \text{ vs } \frac{1}{k_B T(K)} \text{ curve}$$

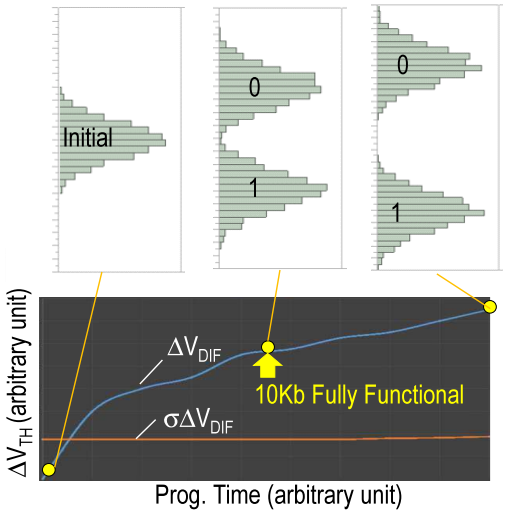
where

$$k_B = \text{Boltzmann's constant } = 8.617 \times 10^{-5} \text{ eVK}^{-1}.$$

The MTPM read cycle shmoo for 32-nm hardware, Fig. 11(b), shows <10-ns read cycle at 1 V, and is functional down to 0.65 V at 20 ns. Similar to 32 nm, CTT-based MTPM macro was realized and characterized in 22-nm SOI technology as well. Fig. 12 shows the hardware results for multiple program-erase cycles in 22-nm node using an appropriate initialization technique [20]. Note that the first few write cycles use a relatively "strong" write condition, achievable through longer programming pulse or slightly elevated programming voltage. Subsequent erases and writes are not as "strong" and this indicates that V_{THSFT} (I_{DS} shown in the graph) saturates after 10 program/erase cycles resulting in significantly smaller hysteresis, demonstrating a feasibility to improve the endurance to as high as 100.



(a)

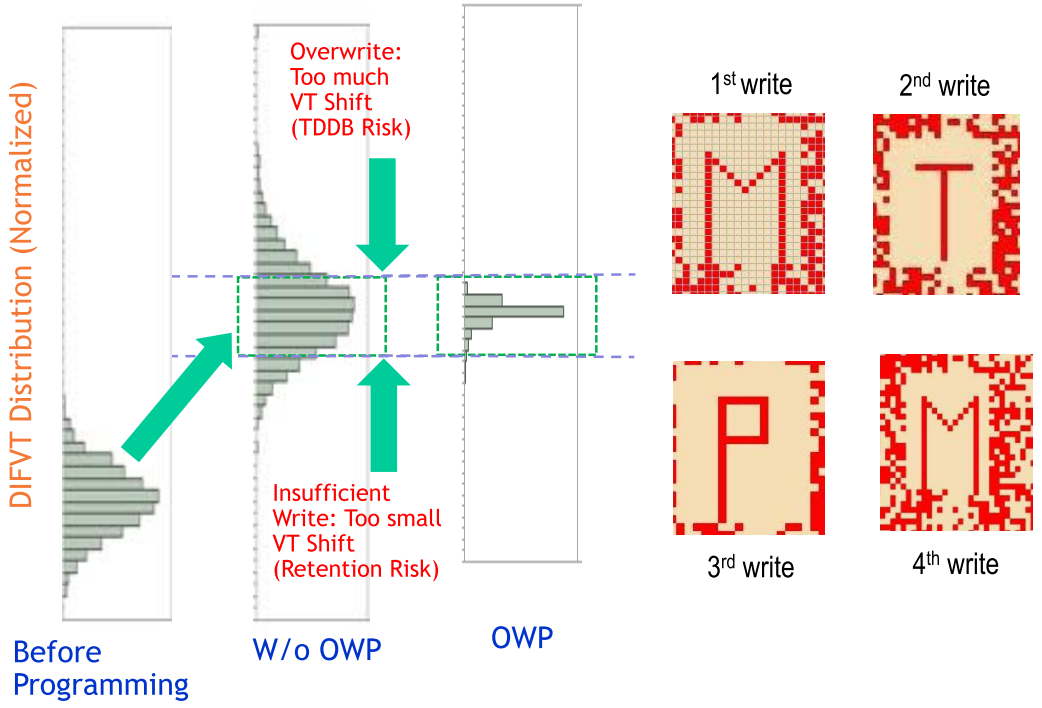


(b)

Fig. 13. Hardware results in 14-nm FinFET bulk technology. (a) Voltage shmoo results. (b) Twin cell V_{TH} difference (ΔV_{DIF}) versus programming time, without OWP.

VI. Scalability—CTT in 14-nm Bulk FinFET Technology

The MTPM with SSA has been shown to scale well with technology down to 14-nm FinFET in bulk technology. Voltage shmoo in Fig. 13(a) shows that the macro is functional down to 0.7 V at ∼40-ns cycle time. The monotonic increase in the difference in the threshold voltages (ΔV_{DIF} between the two cells in the twin cell pair) versus the programming time without OWP is shown in Fig. 13(b). The initial distribution of ΔV_{DIF} is symmetrically centered around zero indicating random V_{TH} variation. Using CKB pattern without OWP, half the array is then programmed to read a zero and the other half a one. This pushes the mean of ΔV_{DIF} on either side of the origin and any further programming pushes it further away. However, it is important to note that the standard deviation of ΔV_{DIF} remains constant throughout.



(a) (b)

Fig. 14. OWP and multi-time programming in 14-nm FinFET bulk technology. (a) OWP introduces comparable V_{THSFT}. (b) Multi-time programming.

TABLE I

SUMMARY OF CTT-MTPM MACROS REALIZED IN 32-nm SOI, 22-nm SOI, AND 14-nm FINFET BULK TECHNOLOGIES

Technology	32nm SOI	22nm SOI	14nm FinFET bulk
Cell	0.109μm^2 with 1.2nm Gox NMOS	0.144μm^2 with 1nm Gox NMOS	0.1411μm^2 with FIN NMOS
Macro density	80Kb	64Kb	40Kb
Density/mm^2	$\sim$2Mb/mm^2	$\sim$2.5Mb/mm^2	$\sim$1.3Mb/mm^2
Activation energy	$\sim$1.35eV	$\sim$2.4eV	$\sim$1.6eV
Design-assist circuits	Yes	No	Yes
Number of cells / BL	256	256	128
Sacrificial WLs and BLs	Yes	No	Yes
Redundant WLs and BLs	Yes	No	Yes

OWP circuit causes similar ΔV_{DIF} across all twin cells irrespective of the initial random V_{TH} variation. This is equivalent to tightening the distribution of ΔV_{DIF} as is shown in Fig. 14(a). Without OWP, the mean shifts with programming, keeping the distribution as is and thereby pushing some cells into the breakdown zone. However, with OWP, the mean shifts but distribution tightens by keeping all cells in the safe zone. The 4$\times$ MTPM functionality is demonstrated in Fig. 14(b) with the characters "M," "T," "P," and "M" being successively written in the same sub-array. This is the first successful demonstration of MTPM using 14-nm FinFET technology. The extracted E_a was $\sim$1.6 eV.

Table I summarizes the key features of the MTPM across technologies. The development started with the 22-nm SOI test chip, which used 0.144-μm^2 cell with a vanilla array for feasibility demonstration, resulting in a density of 2.5 Mb/mm^2. The second prototype used 32-nm SOI technology with 0.109-μm^2 cell, and included design assist circuits such as OWP, SL switch, and redundancy for product feasibility demonstration, resulting in 2 Mb/mm^2 due to their overhead. The most recently developed 14-nm chip used 0.1411-μm^2 cell and included design assist circuits similar to 32-nm prototype. In order to compensate for the low-power technology device and minimize the technology risk for this early technology demonstration using FIN technology, the BL length is halved, resulting in 1.3 Mb/mm^2 due to the SL switch and BL driver overhead. The E_a is a strong function of the device gate stack and the device dimension [21]. Due to the difference in technology and device dimensions, the characterized activation energies for 32 and 22 nm are $\sim$1.35 and $\sim$2.4 eV, respectively. For 14-nm FinFET, the E_a is about $\sim$1.6 eV. The macros realized in 32- and 14-nm FinFET technologies include sacrificial cells that are used for program/read testing during manufacturing to detect WL/BL fails. The faulty BLs/WLs are replaced with redundant BLs/WLs provided in the macro.

VII. CONCLUSION

Embedded NVM based on charge trapping in HiK GOX of logic FETs was demonstrated in 32-, 22-nm SOI, and 14-nm Bulk FinFET logic technologies with no mask adder. Charge trapping was shown to be achievable with logic process compatible voltages. A multi-step block write with OWP was

described and shown to successfully protect all cells in the array. The SSA is shown to be 10× more sensitive than CCSA making it more suitable for this technology. This is the first demonstration of multi-time programming in 14-nm FIN bulk process. The proposed NVM is ∼30× denser than eFUSE for OTPM. Future scope for this memory includes endurance improvement by employing initialization technique to make it a viable option for varied applications including redundancy in 3-D memories and firmware for micro-controllers that need enhanced hardware security.

ACKNOWLEDGMENT

The authors would like to thank R. Kilker for SA design, and the GLOBALFOUNDRIES 14LPP technology development team and the IBM 22- and 14-nm silicon-on-insulator technology development teams.

REFERENCES

[1] T. Kirihata, "High performance embedded dynamic random access memory in nano-scale technologies," in *CMOS Processors and Memories*, K. Iniewski, Ed. Dordrecht, The Netherlands: Springer, 2010, ch. 10, pp. 295–336.

[2] G. Fredeman *et al.*, "A 14 nm 1.1 Mb embedded DRAM macro with 1 ns access," *IEEE J. Solid-State Circuits*, vol. 51, no. 1, pp. 230–239, Jan. 2016.

[3] E. Fluhr *et al.*, "POWER9: A processor family optimized for cognitive computing with 25 Gb/s accelerator links and 16 Gb/s PCIe Gen4," in *IEEE Int. Solid-State Circuits Conf. (ISSCC) Dig. Tech. Papers*, Feb. 2017, pp. 50–51.

[4] N. Robson *et al.*, "Electrically programmable fuse (eFUSE): From memory redundancy to autonomic chips," in *Proc. IEEE Custom Integr. Circuits Conf. (CICC)*, Sep. 2007, pp. 799–804.

[5] J. Viraraghavan *et al.*, "80 Kb 10 ns read cycle logic embedded high-*k* charge trap multi-time-programmable memory scalable to 14 nm FIN with no added process complexity," in *IEEE Symp. VLSI Circuits Dig. Tech. Papers*, Jun. 2016, pp. 1–2, doi: 10.1109/VLSIC.2016.7573462.

[6] R. Yamashita *et al.*, "A 512 Gb 3b/cell flash memory on 64-word-line-layer BiCS technology," in *IEEE Int. Solid-State Circuits Conf. (ISSCC) Dig. Tech. Papers*, Feb. 2017, pp. 196–197.

[7] C. Kim *et al.*, "A 512 Gb 3b/cell 64-stacked WL 3D V-NAND flash memory," in *IEEE Int. Solid-State Circuits Conf. (ISSCC) Dig. Tech. Papers*, Feb. 2017, pp. 202–203.

[8] S. S. Iyer, T. Kirihata, and J. E. Barth, "Three dimensional integration—Considerations for memory applications," in *Proc. IEEE Custom Integr. Circuits Conf. (CICC)*, Sep. 2011, pp. 1–7.

[9] K. Takeuchi, "Scaling challenges of NAND flash memory and hybrid memory system with storage class memory & NAND flash memory," in *Proc. IEEE Custom Integr. Circuits Conf. (CICC)*, Sep. 2013, pp. 1–6.

[10] D. Cai *et al.*, "An 8-Mb phase-change random access memory chip based on a resistor-on-via-stacked-plug storage cell," *IEEE Electron Device Lett.*, vol. 33, no. 9, pp. 1270–1272, Sep. 2012.

[11] H. Shiga *et al.*, "A 1.6 GB/s DDR2 128 Mb chain FeRAM with scalable octal bitline and sensing schemes," in *IEEE Int. Solid-State Circuits Conf. (ISSCC) Dig. Tech. Papers*, Feb. 2009, pp. 464–465, 465a.

[12] M. Jefremow *et al.*, "Time-differential sense amplifier for sub-80 mV bitline voltage embedded STT-MRAM in 40 nm CMOS," in *IEEE Int. Solid-State Circuits Conf. (ISSCC) Dig. Tech. Papers*, Feb. 2013, pp. 216–217, doi: 10.1109/ISSCC.2013.6487706.

[13] M. Ueki *et al.*, "Low-power embedded ReRAM technology for IoT applications," in *IEEE Symp. VLSI Tech. Dig. Tech. Papers*, Jun. 2015, pp. T108–T109.

[14] H. Kojima *et al.*, "Embedded flash on 90 nm logic technology & beyond for FPGAs," in *IEDM Tech. Dig.*, Dec. 2007, p. 677–680.

[15] S.-Y. Chou, Y.-S. Chen, J.-H. Chang, Y.-D. Chih, and T.-Y. J. Chang, "A 10 nm 32 Kb low-voltage logic-compatible anti-fuse one-time-programmable memory with anti-tampering sensing scheme," in *IEEE Int. Solid-State Circuits Conf. (ISSCC) Dig. Tech. Papers*, Feb. 2017, pp. 200–201.

[16] Y. Taito *et al.*, "A 28 nm embedded SG-MONOS flash macro for automotive achieving 200 MHz read operation and 2.0 MB/S write throughput at Ti, of 170°C," in *IEEE Int. Solid-State Circuits Conf. (ISSCC) Dig. Tech. Papers*, Feb. 2015, pp. 132–133.

[17] K. Noda, "Using hot carrier injection for embedded non-volatile memory," NSCore, Inc., Fukuoka, Japan, White Paper WhitePaper_081002, Oct. 2008. [Online]. Available: http://www.nscore.com/images/WhitePaper_081002.pdf

[18] C. Kothandaraman *et al.*, "Oxygen vacancy traps in Hi-K/Metal gate technologies and their potential for embedded memory applications," in *Proc. IEEE Rel. Phys. Symp. (IRPS)*, Apr. 2015, pp. MY.2.1–MY.2.4.

[19] S. Zafar *et al.*, "A comparative study of NBTI and PBTI (charge trapping) in SiO_2/HfO_2 stacks with FUSI, TiN, Re gates," in *IEEE Symp. VLSI Technol., Dig. Tech. Papers*, Jun. 2006, pp. 23–25.

[20] F. Khan, E. Cartier, J. C. S. Woo, and S. S. Iyer, "Charge trap transistor (CTT): An embedded fully logic-compatible multiple-time programmable non-volatile memory element for high-*k*-metal-gate CMOS technologies," *IEEE Electron Device Lett.*, vol. 38, no. 1, pp. 44–47, Jan. 2017.

[21] F. Khan, E. Cartier, C. Kothandaraman, J. C. Scott, J. C. S. Woo, and S. S. Iyer, "The impact of self-heating on charge trapping in high-*k*-metal-gate nFETs," *IEEE Electron Device Lett.*, vol. 37, no. 1, pp. 88–91, Jan. 2016.

[22] A. Hajimiri and R. Heald, "Design issues in cross-coupled inverter sense amplifier," in *Proc. IEEE Int. Symp. Circuits Syst. (ISCAS)*, vol. 2. May 1998, pp. 149–152.

Balaji Jayaraman received the B.E. degree in electronics and communication engineering from Osmania University, Hyderabad, India, in 2004, and the Ph.D. degree from the Electrical Communication Engineering Department, Indian Institute of Science, Bangalore, India, in 2011.

He joined the IBM Semiconductor Research and Development Center, Bangalore, India, in 2010, where he was involved in 22-nm embedded DRAM health-of-the-line characterization and yield analysis. Since 2015, he has been with GLOBALFOUNDRIES, Bangalore, where he is involved in embedded non-volatile memory design. His current research interests include non-volatile memories, 3-D memory design, and silicon photonics.

Derek Leu received the B.E. degree from the State University of New York, Stony Brook, NY, USA, in 2006.

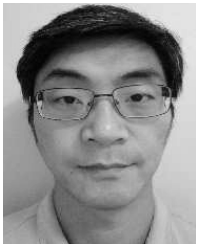
He joined the IBM Microelectronics Division, East Fishkill, NY, USA, in 2006, where he was involved in embedded DRAM and non-volatile memory design development. He is currently with Advanced Silicon and Packaging Department, GLOBALFOUNDRIES, East Fishkill, NY, USA, where he is involved in embedded memory design development.

Janakiraman Viraraghavan received the Ph.D. degree in very large scale integration (VLSI) from the Electrical Communication Department, Indian Institute of Science, Bangalore, India, in 2010.

He joined the Semiconductor Research and Development Center, IBM India Pvt. Ltd., Bangalore, where he was involved in the design of embedded DRAM in 14 nm for early technology qualification. He was involved in embedded non-volatile memories at GLOBALFOUNDRIES, Bangalore, India. He joined the Electrical Engineering Department, IIT Madras, Chennai, India, as an Assistant Professor, in 2016. His current research interests include statistical analysis in VLSI and ASIC implementation of machine learning algorithms on hardware.

Alberto Cestero received the B.S.E.E. and M.S.E.E. degrees in electronics from the University of Puerto Rico Mayaguez campus, Mayagüez, Puerto Rico, where his thesis included the development of a performance simulator of k-ary n cube parallel processor networks.

He was a Board In-Circuit and a Functional Test Engineer with Storagetek, Ponce, Puerto Rico. He joined IBM, East Fishkill, NY, USA, in 2001. He is currently a Staff Product Test and a Characterization Engineer for GLOBALFOUNDRIES, East Fishkill. His current research interest includes test development and design verification of high-speed embedded DRAM, array device monitors, and eFUSE technology.



Ming Yin received the B.S.E.E. degree from the Huazhong University of Science and Technology, Wuhan, China, the M.S.E.E. degree from the University of Science and Technology of China, Hefei, China, and the Electrical Engineer degree from the University of Utah, Salt Lake, UT, USA.

He joined IBM Microelectronics, East Fishkill, NY, USA, in 2006, where he was the Circuit Designer for the development of embedded DRAM. He was involved in caches and customer macros for SPARC server chips at Sun Microsystems, Sunnyvale, CA, USA. He is currently with GLOBALFOUNDRIES, East Fishkill, NY, USA, where he is involved in embedded non-volatile memory and silicon photonic transceiver.

John Golz received the B.A. degree in physics from the University of California at Berkeley, Berkeley, CA, USA, in 1989, and the M.S. degree in electrical engineering from Yale University, New Haven, CT, USA, in 1992.

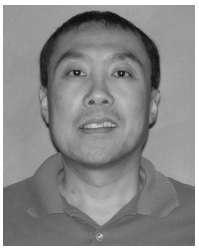
He joined IBM Microelectronics, East Fishkill, NY, USA, in 1996, where he was involved in the development of standalone, embedded, and 3-D stacked DRAM memories. He has been with GLOBALFOUNDRIES, East Fishkill, NY, USA, developing solutions including 2.5-D interposer and 3-D chip stacking for high-bandwidth memory and low-voltage embedded non-volatile memories for high-performance logic.

Rajesh Reddy Tummuru received the Diploma degree in electronics from Nettur Technical Training Foundation, Bangalore, India, in 2006, and the M.S. degree in microelectronics from the Birla Institute of Technology and Science, Bangalore, India, in 2011.

He joined Texas Instruments, Bangalore, where he was involved in the physical design of chips for high-speed data converter products. He joined the Microelectronics Division, IBM, Bangalore, in 2011, and he is involved in the physical design of test chips for process development in advanced deep-submicrometer processes. He has also done physical design for 3-D staking chips. He is also involved in the lithography evaluation of printability for the structures in eDRAM design in the emerging technologies. He has beenwith GLOBALFOUNDRIES, Bangalore, since 2015, where he is involved in silicon photonics technology development, packaging, and physical design of MTPM test chips.

Ramesh Raghavan received the B.Tech. degree in electrical and electronics engineering from Jawaharlal Nehru Technological University, Hyderabad, India, in 2005, and the M.Tech. degree in information and communication technology from the Dhirubhai Ambani Institute of Information and Communication Technology, Gandhinagar, India, in 2008.

He joined the Memory Products Division, Cypress Semiconductor Corporation, Bangalore, India, in 2008, as an Applications Engineer, where he was involved in failure analysis and debug of standalone Async SRAM, FIFO and dual port SRAM memories, and t a Circuit Designer focusing on standalone Async SRAMs. He joined the IBM Semiconductor Research and Development Center, Bangalore, India, in 2013, where he was involved in physical design of eDRAM and SRAM test chips in 10- and 14-nm FINFET technologies until 2014. He is currently with GLOBALFOUNDRIES, Bangalore, where he is involved in circuit and physical design of embedded non-volatile memories. He holds five patents in embedded non-volatile memory circuits. His current research interests include the design of embedded memory circuits and 3-D embedded memory design.



Dan Moy received the B.S. degree in physics from the Massachusetts Institute of Technology, Cambridge, MA, USA, and the M.S. and Ph.D. degrees in solid state physics from the University of Illinois at Champaign–Urbana, Champaign, IL, USA.

He has worked on semiconductors at Intel, IBM Research, and IBM Microelectronics. He is currently with GLOBALFOUNDRIES, East Fishkill, NY, USA, where he is involved in OTP and MTP technologies for embedded memories. He has authored or co-authored more than 25 papers, and co-invented over 50 patents. He has contributed to developments in silicide, metal CMP, BEOL scaling, lithography, DRAM, silicon-on-insulator, and OTP memories.

Thejas Kempanna received the bachelor's degree in telecommunication engineering from the BMS College of Engineering, Bangalore, India, in 2004, the M.Tech. degree in electronics from the Sir. M. Visvesvaraya Institute of Technology, Bangalore, in 2006, and the Ph. D. degree from the Indian Institute of Science, Bangalore, in 2015.

He joined the IBM Semiconductor Research and Development Center, Bangalore, in 2013, where he was involved in device characterization and health-of-the-line of advanced technology nodes, namely, 22 and 14 nm. He was a part of the IBM eMemory Team in 2014 and was involved in glue logic design for memory circuits. Since 2015, he has been a Custom Design Engineer with the Embedded Memory Team, GLOBALFOUNDRIES, Bangalore. He is involved in circuit custom designs, I/O logic design, Verification, and physical design for embedded memory and for 3-D stacking.

Dr. Kempanna was a part of the Smart Detect WSN Team which received the Best Paper and Demo Award at WISARD 2010. He was a recipient of the Best Paper Award at the "Institute of Smart Structures and Systems" International Conference held in 2012 for his paper titled "Fringe Field Junctionless as a Sensitive Displacement Sensor."

Faraz Khan received the B.S. and M.S. degrees in electrical engineering from Rutgers University, New Brunswick, NJ, USA. He is currently pursuing the Ph.D. degree in electrical engineering from the University of California, Los Angeles, CA, USA.

He was a Research Scientist with IBM Semiconductor Research and Development Center, Bangalore, where he was involved in 32-, 22-, and 14-nm SOI CMOS technology development. He is a member of Technical Staff at GLOBALFOUNDRIES, East Fishkill, NY, USA, where he is leading the development of charge trap transistors, an embedded fully logic-compatible multiple-time programmable non-volatile memory element for advanced HKMG CMOS technology nodes. His current research interest includes emerging memory technologies.

Toshiaki Kirihata (SM'99) received the B.S. and M.S. degrees in precision engineering from Shinshu University, Nagano, Japan, in 1984 and 1986, respectively.

In 1986, he joined the Tokyo Research Laboratory, IBM Research, Tokyo, Japan, and he was involved in high-speed DRAM design development. In 1996, he transferred to T. J. Watson Research Center, IBM Research, where he was involved in research and development for high-density DRAMs. In 2000, he joined the IBM Semiconductor Research and Development Center, East Fishkill, NY, USA, as a Design Manager, where he was involved in the development of high-performance embedded DRAMs and embedded 3-D memories. He is currently with GLOBALFOUNDRIES, East Fishkill, NY, USA, where he manages the design department for advanced silicon and packaging division. He authored the books *CMOS Processors and Memories* and *Circuit for Emerging Applications* contributing to the chapters on eDRAM and intrinsic chip ID using eDRAM, respectively. His current research interests include high-performance embedded DRAM, embedded non-volatile memory, 3-D memory, and hardware security.

Mr. Kirihata was the IEEE CICC Committee Member on memory, and a Memory Session Chair in the year 2014. He presented papers at the ISSCC 1998, 1999, 2001, and 2004 conferences. He was a recipient of the Lewis Winner Outstanding Paper Award on the ISSCC paper entitled "A 500MHz Random Cycle, 1.5-ns Latency, SOI Embedded DRAM macro Featuring a Three Transistor Micro Sense Amplifier."

Subramanian S. Iyer (F'95) is Distinguished Professor and holds the Charles P. Reames Endowed Chair in the Electrical Engineering Department, University of California, Los Angeles, CA, USA, and the Director of the Center for Heterogeneous Integration and Performance. His key technical contributions have been the development of the world's first SiGe base HBT, salicide, electrical fuses, embedded DRAM, and 45-nm technology node. He also was among the first to commercialize bonded SOI for CMOS applications through a startup called SiBond LLC. He has authored over 300 papers and holds over 70 patents. His current research interests include advanced packaging and 3-D integration for system-level scaling and new integration and computing paradigms, as well as the long-term semiconductor and packaging roadmap for logic, memory, and other devices including hardware security and supply chain integrity.

Dr. Iyer was a recipient of several Outstanding Technical Achievements and Corporate Awards at IBM. He is an APS Fellow, IBM Fellow, and a Distinguished Lecturer of the IEEE EDS as well as its treasurer and a member of the Board of Governors of the IEEE EPS. He is also a fellow of the National Academy of Inventors. He is a Distinguished Alumnus of IIT Bombay and also a recipient of the IEEE Daniel Noble Medal for emerging technologies in 2012.