Phase Transition Material-Assisted Low-Power SRAM Design

S. S. Teja Nibhanupudi, *Graduate Student Member, IEEE*,
Siddhartha Raman Sundara Raman, *Student Member, IEEE*,
and Jaydeep P. Kulkarni, *Senior Member, IEEE*

***Abstract*— The threshold switching properties of the phase transition material (PTM) can be exploited to realize a heterogeneous static random access memory (SRAM) bitcell, which can obviate the need for assist techniques. This unique PTM-SRAM bitcell is designed by placing the PTM in series with the gate of pull-down nMOS transistors. The large insulating state resistance of the PTM device blocks the propagation of read voltage rise during a read operation and noise during retention operation, thereby enhancing read/retention stability. On the flip side, the write-time and write-ability are impacted, which can be improved by tuning the access transistor strength. Overall, the PTM-SRAM bitcell achieves active-V_{MIN} equivalent to baseline SRAM V_{MIN} aided by both read- and write-assist techniques. Furthermore, the read access time of the PTM-SRAM bitcell does not degrade with reducing read V_{MIN} in contrast to other read-assist techniques. For an isoactive-V_{MIN} of 0.52 V, the PTM-SRAM has 20% lower read access time, 35% lower read power, 16% higher write time, and 55% lower write power compared to the baseline SRAM aided by assist techniques. Also, the dynamic retention stability of the PTM-SRAM improves by 11.36× compared to baseline SRAM. Detailed analysis highlighting the sensitivity of PTM parameters on PTM-SRAM performance metrics is also presented.**

***Index Terms*— Active V_{MIN}, phase transition material (PTM), read-stability, retention-stability, static random access memory (SRAM), write-ability.**

I. Introduction

STATIC random access memory (SRAM) technology is the enabler of advanced CMOS logic technology scaling and has significant implications on the continuation of Moore's law. With rapid growth in data-intensive computing, the need for large capacity SRAMs is growing across all product segments, such as high-performance ASICs, machine learning accelerators and battery-powered SoCs. However, SRAM bitcell optimization requires intricate balance to meet the conflicting read-write requirements. Traditionally, SRAM bitcells have been designed with (Pull-up < Access < Pull-down) sizing ratio to ensure successful read and write operations under typical operating conditions. However, such sizing optimizations are more restrictive in FinFET CMOS technologies due to the inherent width quantization effect [1], [2]. This restricts the designers to employ all single fin transistors (Pull-up, Pull-down, and Access) in the bitcell for realizing a high-density memory array. The equally sized transistors in the bitcell can lead to contention both during the read (contention between Pull-down and Access devices) and write (contention between Pull-up and Access devices) operations resulting in a large number of read/write failures. Moreover, the increasing complexity in semiconductor process integration compounded by the reduction in feature sizes has resulted in increased interdie and intradie process variations. These process variations are primarily dominated by random dopant fluctuations [3], line edge roughness [4], and mask edge misalignment that leads to broader distribution of transistor threshold voltages. The threshold voltage (Vt) variations make the SRAM bitcell more vulnerable to read and write failures, especially at lower voltages due to increased sensitivity of circuit parameters to Vt variations. This limits the minimum supply voltage (also called V_{MIN}) scaling for successful SRAM array operation. To overcome these potential failures in the advanced technology nodes, circuit-assist techniques, such as wordline underdrive (WLUD) [5], [6] for read and supply voltage collapse (SVC) [7], [8] and negative bitline (NBL) [9] for write, are being employed. However, the aforementioned circuit-level assist techniques consume a significant area and power and incur a cycle time penalty. Hence, there is a need to minimize the overheads of SRAM-assist techniques by exploring such methods at various levels, including the use of novel materials, device structures, process optimization, and compact circuit-assist techniques.

Over the recent years, threshold switching selector devices fabricated using phase transition material (PTM) [10] have gained attention for application in memory, steep-subthreshold swing transistors [11], and power management circuits [12]. These PTM devices exhibit volatile abrupt switching characteristics accompanied by the abrupt change in resistance [10]. In this article, we leverage this unique property of abrupt PTM switching as a technology assist to improve the read stability

Manuscript received February 14, 2021; revised March 17, 2021; accepted March 18, 2021. Date of publication April 8, 2021; date of current version April 22, 2021. This work was supported in part by the National Science Foundation (NSF) under Grant 1815616 and in part by the Semiconductor Research Corporation (SRC) under Grant 2824.001. The review of this article was arranged by Editor S. Alam. *(Corresponding author: S. S. Teja Nibhanupudi.)*

The authors are with the Department of Electrical and Computer Engineering, The University of Texas at Austin, Austin, TX 78712 USA (e-mail: subrahmanya_teja@utexas.edu; jaydeep@austin.utexas.edu).

Color versions of one or more figures in this article are available at https://doi.org/10.1109/TED.2021.3067849.
Digital Object Identifier 10.1109/TED.2021.3067849

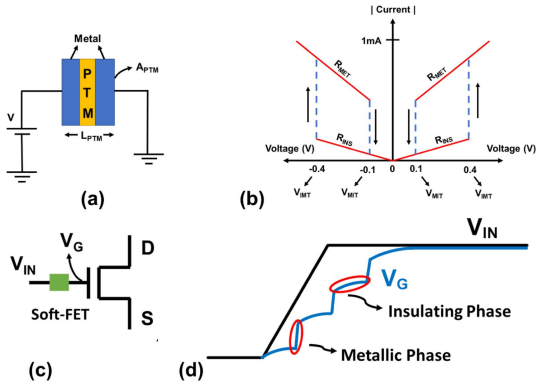

Fig. 1. (a) PTM device structure. (b) PTM *I–V* characteristics. (c) Soft-FET structure. (d) Soft-FET transient response.

and access time of the SRAM bitcell (termed PTM-SRAM in the rest of this article). The read-stability improvement due to PTM assist is leveraged to improve the SRAM write-ability by device Vt optimization. We present a thorough analysis of the proposed PTM-SRAM read/write/retention operation and its impact on the active V_{MIN}.

This article is organized as follows. Section II provides a brief introduction of PTM structure and operation. The proposed PTM-SRAM bitcell configuration and operation are discussed in Section III. Simulation results quantifying the impact of PTM insertion on read/write/retention performance supplemented by design space exploration and layout studies are presented in Section IV. Section V concludes this article with a brief summary and key observations.

II. Phase Transition Material and Soft-FET

A. Phase Transition Material

PTMs are a subset of the transition metal oxides that undergo abrupt insulator to metal transition under the influence of an external electric field. These materials can be utilized to fabricate two-terminal devices [PTM sandwiched in between metal electrodes, as shown in Fig. 1(a)], which can be utilized for realizing unique circuit operation. Of these materials, devices fabricated using vanadium dioxide [13] and niobium dioxide [14] have been experimentally demonstrated. The two-terminal PTM device exhibits hysteretic *I–V* characteristics, as depicted in Fig. 1(b). As evident from the graph, the PTM device resides in an insulating state and transitions to a metallic state when the voltage across its terminals exceeds the switching threshold (V_{IMT}). Then, the PTM device remains in the metallic state until the voltage across the terminals drops below the V_{MIT} threshold. At this point, the reverse phase transition is triggered, and the PTM device moves into an insulating state. The time taken by the PTM device to complete phase transition (insulating to metallic and vice versa) is referred to as the intrinsic switching time (T_{PTM}). Furthermore, the PTM device exhibits identical current profile for negative voltage bias [13], [14].

The PTM device resistance in the insulating state and the metallic state has been experimentally observed to scale similar to an ordinary resistor [15], [16]. Therefore, the resistance in the insulating and metallic state can be modeled as

$$R_{INS} = \rho_{INS} * \frac{L_{PTM}}{A_{PTM}}$$
$$R_{MET} = \rho_{MET} * \frac{L_{PTM}}{A_{PTM}}$$

where ρ_{INS} and ρ_{MET} correspond to the resistivity of the device in insulating, metallic state, respectively, and L_{PTM} and A_{PTM} correspond to the thickness and the area of the cross section of the device, respectively. In addition, the switching threshold voltages increase with thickness of the device and remain unaffected by the area of the device [16]. Both these observations suggest that a constant current density is required to trigger the phase transition [16]. Therefore, the switching thresholds can modeled as

$$V_{IMT} = J_{CIMT} * A_{PTM} * \rho_{INS} * \frac{L_{PTM}}{A_{PTM}} = J_{CIMT} * \rho_{INS} * L_{PTM}$$
$$V_{MIT} = J_{CMIT} * A_{PTM} * \rho_{MET} * \frac{L_{PTM}}{A_{PTM}} = J_{CMIT} * \rho_{MET} * L_{PTM}$$

where J_{CIMT} and J_{CMIT} correspond to the current density required to trigger phase transition from insulating to metallic state and vice versa.

B. Soft-FET

The two-terminal PTM device when placed in series with the gate of the transistor, as shown in Fig. 1(c), is referred to as Soft-FET [12]. The Soft-FET has unique transient characteristics as the PTM device and the gate capacitor mimic an *RC* circuit. As the voltage V_{IN} increases, the current starts charging the gate capacitor through the PTM device residing in an insulating state. The large insulating state resistance ($R_{INS} \sim$ MΩ) causes the gate capacitor to be charged slowly due to the large ($R_{INS} * C_{Gate}$) time constant. Therefore, the V_G node voltage increases at slower rate compared to V_{IN} resulting in increasing ($V_{IN} - V_G$) voltage difference across the PTM device. An abrupt insulator-to-metal transition is triggered when this voltage difference exceeds the V_{IMT} threshold. The PTM device now transitions into metallic state, and the corresponding time constant ($R_{MET} * C_{Gate}$) is small. Therefore, the gate capacitor is charged with this small time constant, and the V_G node follows the V_{IN} node voltage. Consequently, the voltage across the PTM ($V_{IN} - V_G$) also reduces sharply, and when this voltage drops below the V_{MIT} threshold, the phase transition from metallic to insulating state occurs. Again, the time constant of the circuit increases to $R_{INS} * C_{Gate}$, leading to slow charging of the V_G node and eventual phase transition to the metallic phase. These subsequent transitions between slow and quick charging cycles result in staircase waveform, as depicted in Fig. 1(d). The lateral portions of the V_G staircase waveform correspond to the insulating state of the PTM device, and the vertical portions correspond to the metallic state of the PTM device. Similarly, when the V_{IN} voltage ramps down, the V_G node voltage exhibits a downward staircase waveform. We utilize the concept of Soft-FET in the proposed PTM-SRAM configuration, as explained in Section III.

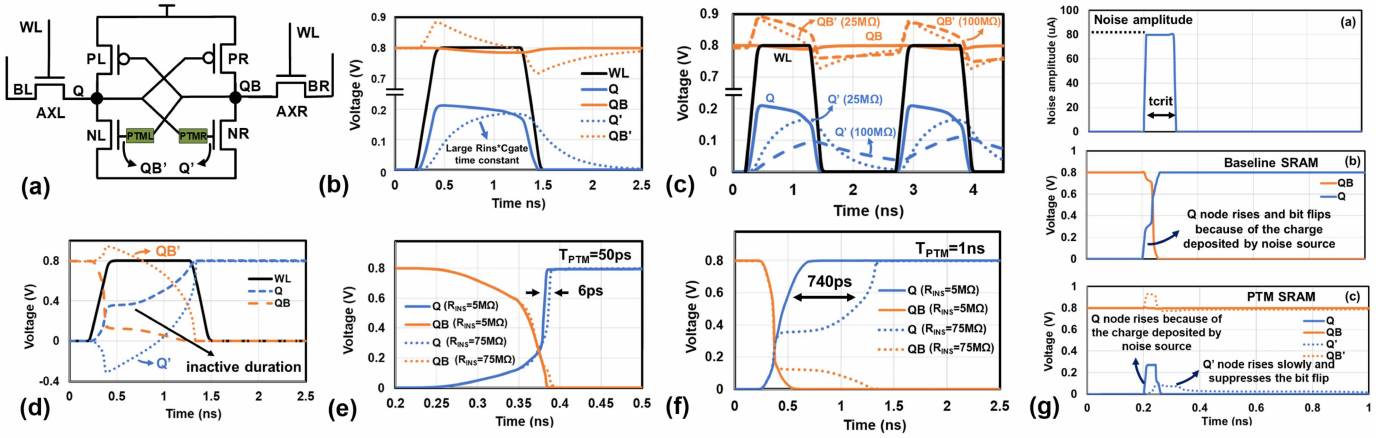

Fig. 2. PTM-SRAM: (a) bitcell configuration, (b) read operation timing waveform, (c) read operation timing waveform showing two subsequent cycles for $R_{INS} = 25$ MΩ and 100 MΩ, (d) write operation timing waveform, (e) write operation with $T_{PTM} = 50$ ps and $R_{INS} = 5$ and 75 MΩ, (f) write operation with $T_{PTM} = 1$ ns and $R_{INS} = 5$ and 75 MΩ, and (g) retention operation timing waveform for baseline SRAM and PTM-SRAM.

III. PTM-Assisted SRAM Bitcell

The proposed SRAM bitcell consists of two PTM devices placed at the gate of pull-down transistors, as shown in Fig. 2(a). These PTM devices form the Soft-FET configuration explained above. The storage nodes are represented by Q and QB, and the corresponding gate terminals of the pull-down transistors are represented by Q' and QB′, respectively. For the illustration purpose, we consider the node Q stores "0," and the node QB stores "1."

A. Read Operation

During the read operation, both the bitlines are precharged to Vcc (supply voltage), and the wordline is asserted. The node Q (storing "0") experiences a voltage rise due to the voltage divider action between the AXL (access) transistor and the NL (pull-down) transistor, as depicted in Fig. 2(b). If this voltage rise exceeds the trip point of the coupled inverter (PR-NR inverter), then the bit flips, resulting in a read failure.

In the proposed PTM-SRAM bitcell configuration, the presence of the PTM device hinders the propagation of the Q node voltage rise to the pull-down gate terminal (Q') of the other inverter. The PTM device parameters are optimized such that V_{IMT} threshold is higher than the peak voltage rise at the "0" storing node during a read operation, therefore restricting the PTM device to remain in the insulating state (large R_{INS}) throughout the read operation. This unique behavior of PTM devices enables improved read stability through the following mechanisms.

1) The rising voltage on the node Q capacitively couples onto the node QB′ (gate–drain capacitance) as the PTM device in insulating state incurs high resistance between QB and QB′ nodes. This results in boosted QB′ node voltage, which lowers the ON-resistance of the NL transistor, thereby lowering the voltage rise on node Q.
2) Furthermore, the node Q' does not follow node Q instantly. Instead, Q' is charged slowly with the large $R_{INS} * C_{Gate}$ time constant (PTMR in the insulating state), as shown in Fig. 2(b). This ensures that the NR transistor is not turned on, and the node QB is held close to "1."
3) PTML in the insulating state now effectively shields the node QB′ from any small droop on the node QB. This ensures that that the NL transistor remains strongly turned-on and lowering the voltage rise on node Q.

It is worth noting that the PTM devices partially isolate the pull-down transistors (NL and NR) from the storage nodes (Q and QB). This breaks the positive feedback loop of the cross-coupled inverters, thereby preventing a possible bit-flip/read-failure. Furthermore, even if the phase transition is triggered in the PTM device (in the event of variations or noise coupled to raised Q node), the inherent switching time of the PTM device temporarily shields the node Q' from experiencing the voltage rise. Thus, the read stability of the SRAM bitcell can be significantly enhanced using PTM device-level assist without requiring the circuit-level assist techniques. Increasing R_{INS} increases the $R_{INS} * C_{Gate}$ time constant, and the node Q' is charged very slowly, thereby improving the read stability. However, the node Q' also discharges very slowly toward the end of the read cycle and may extend into the next cycle, as highlighted in Fig. 2(c), with two scenarios of $R_{INS} = 25$ MΩ and 100 MΩ. Hence, the value of R_{INS} needs to be optimized for improved read stability, as well as ensuring correct back-to-back, high-speed read operations on the same bitcell.

B. Write Operation

The SRAM write operation begins by driving the bitlines (BL to "1" and BR to "0") depending on data polarity and asserting the wordline. The node QB begins to discharge through the AXR transistor and the BR bitline. However, the node QB′ does not discharge until PTML transitions into metallic phase, (T_{PTM} after QB′-QB > V_{IMT}), restricting the NL transistor to remain turned on and, thereby, hindering the charging of node Q. Similarly, the node Q' does not charge until PTMR transitions into metallic phase (T_{PTM} after $Q - Q' > V_{IMT}$), restricting the NR transistor to remain turned-off without aiding in the discharge of node QB, thereby delaying the write completion process. However, the pull-up transistors, PL and PR, are turned on/off immediately with the changes on QB and Q nodes, respectively. These opposing effects result in increased contention with PL and NL transistors turned on at the same time. On the other hand, both PR and NR transistors remain turned off at the same

time. This results in some inactive durations where the storage node values remain unchanged, as highlighted in Fig. 2(d). This inactive durations is sensitive to R_{INS} and switching time T_{PTM}. For smaller values of T_{PTM} ($T_{PTM} \sim$ SRAM write time), the write time is nearly independent of R_{INS}, as shown in Fig. 2(e). However if T_{PTM} is large ($T_{PTM} >$ SRAM write time), the R_{INS} value determines the inactive duration and, therefore, the write completion time, as shown in Fig. 2(f).

C. Retention Operation

The presence of PTM devices significantly boosts the dynamic retention stability by hindering the propagation of the transient noise spikes. The PTM with inherently high insulating resistance (R_{INS}) delays the propagation of the spike due to the large $R_{INS} * C_{Gate}$ time constant. Hence, the noise spikes that cannot deposit sufficient charge to trigger the insulator–metal transition are completely filtered out, thereby protecting the storage nodes. Fig. 2(g) shows that, for an identical noise spike, the baseline SRAM bit flips, whereas the PTM-SRAM retains the stored values. Here, the noise disturbance is modeled as an injected current pulse, which accurately reflects the nonlinear dynamic nature of the cell compared to voltage pulse [17]. Although noise spikes with large amplitude can trigger insulator–metal transition, the PTM device requires finite intrinsic switching time to complete the phase transition. Hence, the noise spikes with pulsewidth much smaller than the T_{PTM} are compensated for by the restoring currents before propagating to the other storage node. In addition, the presence of two PTM devices breaks the positive feedback loop by isolating the Q and QB nodes from the Q' and QB′ nodes, respectively, further enhancing the retention stability of the SRAM bitcell. Thus, the inclusion of the PTM device within the SRAM bitcell improves the resiliency of the bitcell to noise-induced either by coupling, crosstalk, or radiation strike. Therefore, the PTM-SRAM can exhibit better soft-error performance compared to the baseline SRAM.

IV. Results and Analysis

A. Simulation Setup

To estimate the effectiveness of the proposed PTM-assisted SRAM, an extensive SPICE simulation-based design space exploration is performed using open-source 7-nm FinFET predictive technology model [18] and a phenomenological Verilog-A model for the PTM device [19]. The intrinsic switching time (T_{PTM}) is an important parameter in determining the performance of the PTM-SRAM. The reported T_{PTM} has a wide range from 50 ps at scaled nanodimensions [20] to experimental demonstrations, showing 1-ns switching time [21] due to differences in measurement setup, patterning methods, and so on. Hence, a detailed sensitivity analysis for both values of T_{PTM} is performed to cover this wide range of T_{PTM}. Other model parameters used for the PTM device simulation are adopted from [22] and are listed in Table I.

Since PTM device switching is a time-dependent phenomenon, the conventional static noise margin (SNM) used for the SRAM stability analysis would not capture the PTM transient effects. Hence, a dynamic stability analysis is performed considering bitline and PTM device RC effects to assess

TABLE I
PTM Device Model Parameters

Description	Parameter	Value
Current density (INS state)	J_{CIMT}	$1.87x10^2$(A/cm^2)
Current density (MET state)	J_{CMIT}	$5.1x10^3$(A/cm^2)
Resistivity (INS state)	ρ_{INS}	80 (Ω.cm)
Resistivity in (MET state)	ρ_{MET}	$5x10^{-4}$(Ω.cm)
Intrinsic switching time	T_{PTM}	50ps-1ns
PTM device thickness	t	20-40 (nm)
PTM device width	w	40-100 (nm)

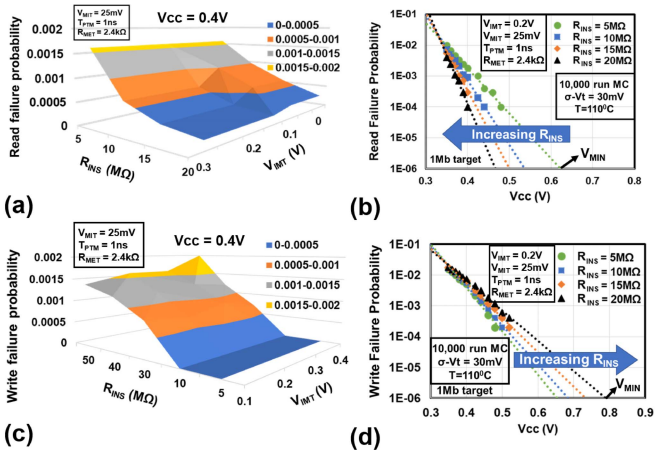

Fig. 3. (a) PTM-SRAM bitcell read failure sensitivity to V_{IMT} and R_{INS}. (b) PTM-SRAM read failure probability variation with Vcc. Trendlines extended for estimating V_{MIN}. (c) PTM-SRAM write failure sensitivity to R_{INS} and V_{IMT}. (d) PTM-SRAM write failure probability variation with Vcc.

the effectiveness of the PTM-SRAM. The transient read/write operation is performed with statistical variations, and the storage nodes are monitored to detect read/write failures. The variations are included by performing a 10 000 sample point Monte Carlo simulations, assuming transistor threshold voltage variation of 30 mV (one-sigma) at an operating temperature of 110 °C. In addition, three-sigma variation of 20% in all PTM parameters (V_{IMT}, V_{MIT}, R_{INS}, R_{MET}, and T_{PTM}) is included in the Monte Carlo analysis to take into account the PTM device-to-device and cycle-to-cycle variations. For this dynamic stability analysis, the bitline capacitance is assumed to be 100 fF for 256 bits/column and modulated as the number of bitcells/column varies. The frequency of operation is scaled by following the voltage-frequency trend of a five-stage ring oscillator delay. At 0.8 V, Fmax is set to 400 MHz and lowered for lower voltages. The wordline pulsewidth is a half-clock cycle adopting a synchronous SRAM operation.

B. Read Analysis

As explained in Section III-A, the PTM-SRAM exhibits enhanced read stability by blocking V_{READ} (V_{READ} is the voltage rise on the "0" storing node during a read operation when wordline is activated) from affecting the other side inverter (storing a "1") during a read operation. Fig. 3(a) highlights the impact of V_{IMT} and R_{INS} on the read failure probability (P_{FAIL} = number of failures/number of MC sample points) of the PTM-SRAM. The read failure probability reduces as V_{IMT} increases and saturates when V_{IMT} is greater than the V_{READ}. In this article, the PTM-SRAM is designed

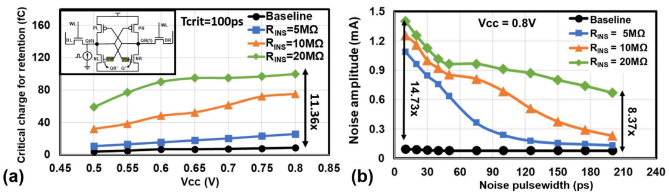

Fig. 4. (a) Critical charge variation with Vcc. Inset: PTM-SRAM bitcell with noise injected as current pulse at storage node "Q" for dynamic retention stability analysis. (b) Current pulse amplitude variation with noise pulsewidth.

with ($V_{IMT} > V_{READ}$) such that the PTM remains in insulating state throughout the read operation. Therefore, the read operation of PTM-SRAM is sensitive only to R_{INS}. As R_{INS} increases, the read failure probability reduces. Fig. 3(b) shows the statistical analysis (Monte Carlo simulations) for the read failure probability for different values of R_{INS} as a function of supply voltage. These Monte Carlo simulation trend lines are extrapolated to estimate the minimum operating voltage (read $-$ V_{MIN}). The P_{FAIL} target is chosen to be 10^{-6} to represent one failure in 1-Mb array size. As evident from Fig. 3(b), the read-V_{MIN} of the PTM-SRAM reduces with increasing R_{INS}.

C. Write Analysis

As explained in Section III-B and depicted by Fig. 2(e) and (f), the PTM-SRAM write operation is sensitive to V_{IMT}, R_{INS}, and T_{PTM} values. Fig. 3(c) plots the write failure probability (P_{FAIL} = number of failures/number of MC sample points) as a function of R_{INS} and V_{IMT} while keeping T_{PTM} at 1 ns to be worst case of the switching time range. It is observed that the write failure probability increases rapidly with increasing R_{INS}, whereas V_{IMT} shows smaller impact on the write failures. The write failure probability is plotted as a function of operating voltage in Fig. 3(d) for varying R_{INS} values while keeping T_{PTM} to be 1 ns. The trendlines from failure statistics are extrapolated to a failure probability of 10^{-6} to quantify the minimum operating voltage (write-V_{MIN}) to meet a 1-Mb array target. The write V_{MIN} of the PTM-SRAM increases with an increase in R_{INS}. This trend is opposite to the read V_{MIN} trend observed in Section IV-B. Therefore, R_{INS} of the PTM device needs to be carefully optimized to ensure both low read and write V_{MIN}.

D. Retention Analysis

As explained in Section III-C, the PTM-SRAM improves the retention stability of the bitcell. The dynamic retention stability analysis is performed by deasserting the wordline (WL = 0) and injecting noise at the bitcell nodes. The noise is injected from a transient current source connected to one of the storage nodes (assumed as Q in this article) in the form of a pulse waveform, as shown in the inset of Fig. 4(a). The amount of charge (from the current source) required to induce a bit-flip in the SRAM bitcell is referred to as critical charge. The critical charge provides a measure of bitcell dynamic retention stability [17]. This critical charge is quantified as a function of operating voltage for the baseline bitcell and the PTM-SRAM bitcell with R_{INS} =5,10, and 20 MΩ values [see Fig. 4(a)]. A current pulse with a duration of 100 ps is assumed

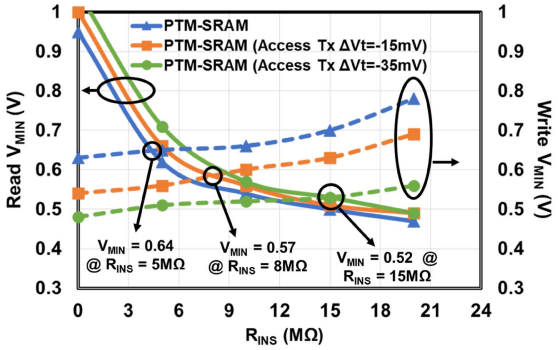

Fig. 5. PTM-SRAM read and write V_{MIN} plotted as a function of R_{INS}.

for this analysis. The PTM-assisted SRAM bitcells exhibit 2.95×–11.36× (for R_{INS} =5−20 MΩ) critical charge improvement compared to the baseline bitcell. To capture the effect of pulsewidth variation on the dynamic retention stability, the noise amplitude of the pulse required to induce a bit-flip is evaluated. It is observed that the noise amplitude required to induce bit-flip is 1.68×–8.37× (for R_{INS} = 5–20 MΩ) higher and 11.47×–14.73× (for R_{INS} = 5–20 MΩ) higher than baseline bitcell for 10-ps pulsewidth and 200-ps pulsewidth, respectively, as shown in Fig. 4(b). Overall, the PTM-assisted SRAM improves the dynamic retention stability of the SRAM bitcell.

E. R_{INS} Impact on V_{MIN}

As mentioned in Section IV-C, the R_{INS} value needs to be optimized to achieve simultaneously low read and write V_{MIN} for PTM-SRAM. Fig. 5 demonstrates the trade-off between read-V_{MIN} and write-V_{MIN} with varying R_{INS} values. As explained in Sections III-A and III-B, the read stability (read-V_{MIN}) improves, and the write ability (write-V_{MIN}) degrades with increasing R_{INS}. The point of crossover is the optimal R_{INS} value where both read and write V_{MIN}'s are the lowest. As evident from Fig. 5, the PTM-SRAM can achieve an optimized active V_{MIN} of 0.64 V when R_{INS} = 5 MΩ (crossover point of blue traces). To further improve the active V_{MIN}, the threshold voltage of the access transistors (AXL/AXRΔVt) can be reduced to make the bitcell write favorable, and the R_{INS} can be increased to further reduce the read V_{MIN}, as shown in Fig. 5. The PTM-SRAM with this write-skewed bitcell (access transistors with 35 mV lower threshold, $\Delta Vt = -35$ mV) can achieve a V_{MIN} of 0.52 V when R_{INS} is increased to 15 MΩ (crossover point of green traces).

F. Comparison With Baseline SRAM

The optimized PTM-SRAM bitcell (access transistors with 35 mV lower threshold voltage, $\Delta Vt = -35$ mV) obtained from the analysis of Section IV-E (active V_{MIN} = 0.52 V) is compared with the baseline SRAM. The baseline SRAM designed using the 7-nm predictive technology model [18] has a read V_{MIN} of 0.92 V. The read V_{MIN} could be improved by employing the read-assist technique, such as WLUD, wherein the wordline voltage is lowered to suppress the read voltage rise (V_{READ}). The read V_{MIN} reduces from 0.92 to 0.42 V by increasing the underdrive voltage from 0% to 30%

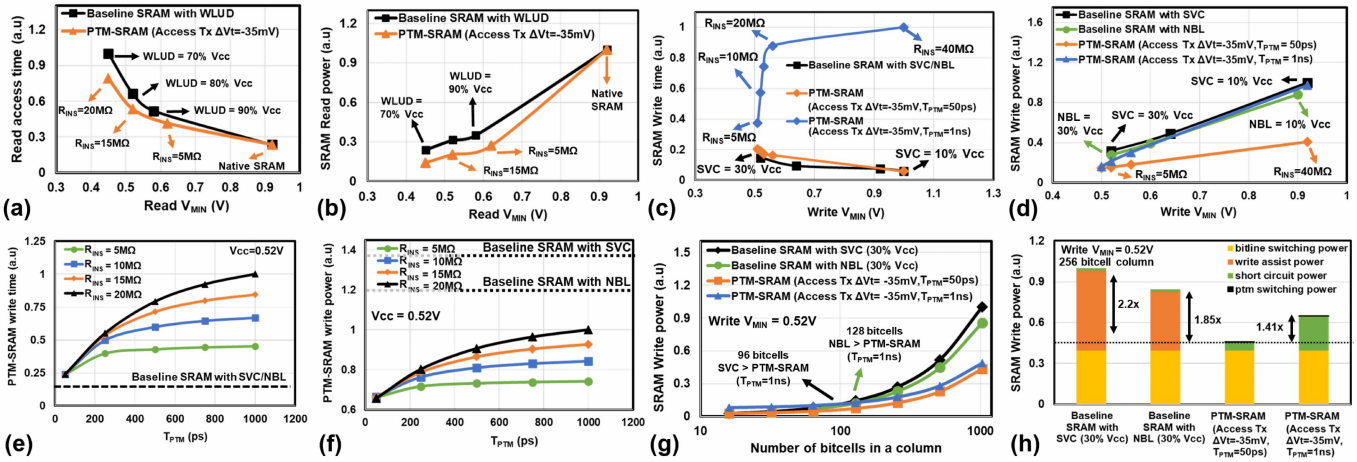

Fig. 6. (a) Read access time sensitivity to read V_{MIN} for baseline SRAM with WLUD and PTM-SRAM. (b) Read power variation with read V_{MIN}. (c) Write time variation with write V_{MIN} for baseline SRAM with SVC/NBL and PTM-SRAM. (d) Write power variation with write V_{MIN}. (e) PTM-SRAM write time variation with T_{PTM}. (f) PTM-SRAM write power variation with T_{PTM}. (g) Iso-V_{MIN} comparison of SRAM write power variation with size of the array. (h) Write power distribution comparison for iso-V_{MIN} bitcells.

Vcc, as shown in Fig. 6(a). However, lowering the wordline voltage increases the time taken to develop bitline differential (assumed to be 100 mV) and, thereby, degrades the read access time compared to the baseline SRAM. On the contrary, the PTM-SRAM read access time remains unaltered compared to baseline SRAM. Therefore, for an isoread-V_{MIN} condition, the PTM-SRAM read access time is lower than baseline SRAM with WLUD, as shown in Fig. 6(a). In addition to the degraded read access time, the WLUD technique also consumes additional power to generate lower voltages from a single rail power supply [5]. Fig. 6(b) plots the read power with change in read-V_{MIN} considering isobitline discharge. The PTM-SRAM read power is lower than baseline SRAM with WLUD read power. Overall, for an isoread-V_{MIN} of 0.52 V, the PTM-SRAM exhibits 20% lower read access time and 35% lower read power compared to the baseline SRAM with WLUD.

Similar to the read operation, write-assist circuitry is required to lower the write V_{MIN}. Especially, the need for write-assist becomes critical in the bitline-interleaved architecture where the SRAM assisted by WLUD technique degrades the write V_{MIN} (read half-select problem) due to the lowering wordline voltage. Bitline interleaving is commonly employed in advanced technology nodes to enhance the soft error tolerance of the SRAM arrays. In a such case, the write V_{MIN} is improved by employing write-assist-technique along the SRAM column, such as SVC [8] or NBL [9]. In the SVC-assist technique, the bitcell Vcc is held at a lower voltage for the duration of wordline pulsewidth, while, in the NBL-assist techniques, the bitline (of the side writing "0") is driven to a negative voltage to increase the access transistor overdrive voltage. Both these techniques aid the access transistors in overcoming the contention from the cross-coupled inverters and, thereby, resulting in successful write operation (lower write-V_{MIN}). The write V_{MIN} reduces from 0.95 to 0.52 V by increasing SVC (collapse) voltage from 0% to 30% Vcc. Similarly, increasing the NBL voltage from 0% to 30% Vcc reduces the write V_{MIN} from 0.95 to 0.52 V. On the other hand, the PTM-SRAM write V_{MIN} reduces by reducing the R_{INS} value. Fig. 6(c) plots the write time variation with write V_{MIN} for baseline SRAM with SVC/NBL and PTM-SRAM. Here, the write-time is quantified as the duration between 50% wordline rise to the bitcell storage node reaching 10% (for write-0) or 90% (for write-1) of the supply voltage. For the PTM-SRAM, unlike the read operation, the write operation is dependent on the PTM intrinsic switching time (T_{PTM}). The PTM-SRAM write analysis is performed incorporating a wide range of T_{PTM} values (50 ps and 1 ns). The write time for PTM-SRAM with $T_{PTM} = 50$ ps is independent of the R_{INS} value and, therefore, follows a similar trend as baseline SRAM using SVC but marginally higher. The write time for PTM-SRAM with $T_{PTM} = 1$ ns reduces with reducing R_{INS} values (which lowers the write-V_{MIN}), as evident from Fig. 6(c). For the write-power comparison, the SVC assist consumes additional power due to the charging/discharging of the large Vcc line capacitance, and the NBL assist consumes additional power to generate negative voltages. On the other hand, the PTM-SRAM consumes additional power due to the short-circuit current in the cross-coupled inverter pair during a write operation. The write power variation with write V_{MIN} is compared for both baseline SRAM and PTM-SRAM in Fig. 6(d). The PTM-SRAM with $T_{PTM} = 50$ ps consumes significantly lower power compared to baseline SRAM assisted by SVC or NBL. The PTM-SRAM with $T_{PTM} = 1$ ns consumes marginally lower power than baseline SRAM but closely follows the trend. Overall, for an isowrite-V_{MIN} of 0.52 V, the PTM-SRAM ($T_{PTM} = 50$ ps) incurs 16% higher write time with 55% (and 45%) lower write power compared to the baseline SRAM with SVC (and NBL).

Fig. 6(e) shows the impact of T_{PTM} and R_{INS} on the write time of the PTM-SRAM when operated at 0.52 V. The PTM-SRAM write time increases with increasing T_{PTM} consistent with observation in Fig. 2(e) and (f). However, the exact write time is determined by both R_{INS} and T_{PTM}. For smaller R_{INS} values, as T_{PTM} increases, the write time increases but tends to saturate quickly. However, when the R_{INS} value is very large (i.e., when $R_{INS} = 20$ MΩ), PTM-SRAM write time follows the T_{PTM} trend. Fig. 6(f) shows the sensitivity of PTM-SRAM write power to R_{INS} and T_{PTM} values. Since the write power is determined by the inactive durations

TABLE II
PERFORMANCE COMPARISON OF PTM-SRAM WITH BASELINE SRAM

SRAM configuration		Active V_{MIN}	Read time	Read power	Write time	Write power	Max. operating frequency (Vcc = 0.52V) $f = \frac{1}{2 * max\,(read, write\ time)}$
PTM-SRAM (Access Tx ΔVt=-35mV) (R_{INS} = 15MΩ)	T_{PTM}=50ps	0.52V	-20%	-35%	+16%	-55%	+23%
	T_{PTM}=1ns				+500%	-35%	+15%

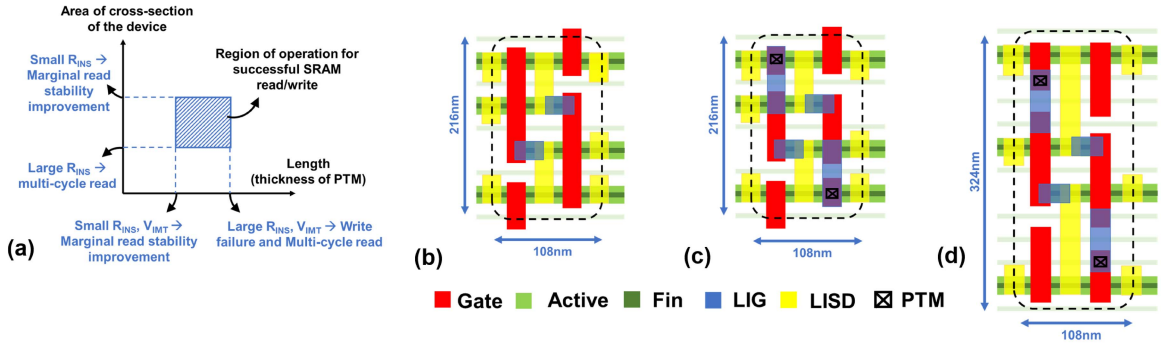

Fig. 7. (a) Illustration highlighting the region of successful operation of PTM-SRAM. (b) 7-nm baseline SRAM layout. (c) PTM-SRAM layout with contact-over-gate. (d) PTM-SRAM layout without contact-over-gate.

(short-circuit current), the write power follows similar trend as write time. For an isowrite-V_{MIN} of 0.52 V ($R_{INS} = 15$ MΩ), the PTM-SRAM write power is lower than baseline SRAM write power (with SVC or NBL assist) even with the largest $T_{PTM} = 1$ ns.

Since the SVC assist involves charging/discharging the bitcell Vcc, the baseline SRAM write power increases with an increasing number of bitcells in the column (increasing Vcc line capacitance and bitline capacitance). Similarly, the power consumed by NBL assist also increases with an increasing number of bitcells in the column due to the higher coupling capacitance required to generate a negative voltage on bitline. On the other hand, the PTM-SRAM write power increases only due to the increase in bitline switching power. Therefore, the write power for baseline SRAM with SVC and NBL assists exceeds the PTM-SRAM even with $T_{PTM} = 1$ ns beyond 96 and 128 bitcells per column, respectively, as shown in Fig. 6(g), indicating PTM-SRAM write-power benefits for large capacity arrays supporting large number of bits/bitline. The distribution of write power among various components is highlighted in Fig. 6(h). Apart from bitline switching power, the dominant component is the write-assist power for baseline SRAMs (for both SVC and NBL assists) and short-circuit power for PTM-SRAMs, respectively. The write-assist power is higher than the PTM-SRAM short-circuit power since it involves dynamic switching of the large capacitances (bitcell Vcc and bitline). Table II summarizes the comparison of the PTM-SRAM metrics with the baseline SRAM. For an isoactive-V_{MIN} of 0.52 V, the PTM-SRAM ($T_{PTM} = 50$ ps) consumes 20% lower read access time, 35% lower read power, 16% higher write time, and 45%–55% lower write power compared to the baseline SRAM. For the 7-nm PDK [18] used in this study, the maximum operating frequency of the SRAM array is limited by the read access time. Therefore, the PTM-SRAM maximum operating frequency improved by 23% and 15% for $T_{PTM} = 50$ ps and $T_{PTM} = 1$ ns cases, respectively.

G. Design Recommendations

Based on the observations from Sections IV-B and IV-C, certain design recommendations to realize an efficient PTM-SRAM bitcell with optimal performance are given as follows. The key PTM device parameters that determine the tradeoff between read and write are R_{INS} and V_{IMT}. These values are, in turn, determined by the thickness and the cross-sectional area of the device. Therefore, to achieve the desired R_{INS} and V_{IMT} values, the physical dimensions of the PTM device need to be scaled accordingly. Fig. 7(a) illustrates the PTM device dimension range for achieving optimal SRAM read/write characteristics. The thickness of the PTM device cannot be too large or too small. Increased thickness of the PTM device increases R_{INS} and V_{IMT} values, which would affect the write-ability (large V_{IMT}) and read operation extending to subsequent cycles (large R_{INS}). While very thin PTM layer would lead to smaller R_{INS}, V_{IMT} values result in lower read-stability improvement. Similarly, the device cross-sectional area needs to within certain upper/lower limits. The very small area would lead to increased R_{INS} resulting in overlapping multicycle read, and a very large area would lead to decreased R_{INS} lowering read-V_{MIN} benefits. Exact design constraints for other SRAM bitcell sizing (112 or 122) may vary depending on the transistor strength and PTM parameters; however, the trends would likely remain the same.

H. Layout Studies

The 6T SRAM thin cell layout is carefully optimized for achieving high bitcell density. Any minor modification due to PTM integration in the proposed PTM-SRAM bitcell could incur a bitcell area penalty degrading the bit density. Detailed

layout analysis is performed using open-source ASAP 7-nm FinFET Process Design Kit (PDK) [23], making sure that all spacing and DRC rules are satisfied. The PTM device can be integrated by depositing the PTM layer in the via hole that connects the gate electrode to LIG (local interconnect gate) of pull-down nMOS devices. However, the common gate connection driving both pull-up and pull-down transistors needs to be split (since the PTM device is only connected to the pull-down transistors) and can be connected to LIG to create the required via hole, as shown in Fig. 7(d), resulting in 1.5× larger bitcell area. However, if the FinFET fabrication technology supports contact over active gate [24], then the PTM-SRAM layout could be designed without any area penalty, as depicted in Fig. 7(c).

I. Impact of PTM Endurance

The endurance and reliability of the PTM device are critical aspects that need to be accounted for realizing the PTM-assisted SRAM arrays. Although the experimental demonstration of the PTM device has exhibited endurance up to 10^9 cycles [25], an extensive study into the factors impacting the PTM reliability needs to be conducted. An insight into the failure mechanism can help designers develop compensation circuits in the event of PTM reliability failure. For example, if the PTM device always fails in the low resistance state (similar to oxide breakdown), then the SRAM bitcell would still be functional but without PTM acting as V_{MIN}-assist mechanism and may incur higher read failures once the number of read accesses exceeds the PTM endurance levels. On the other hand, if the PTM device fails in the high resistance state, then the compensation circuits to boost the write performance can be incorporated.

V. Conclusion

This article presents a PTM-assisted SRAM bitcell design incorporating the PTM at the gate terminal of the pull-down nMOS transistor. The proposed PTM-SRAM enhances the read-stability and retention stability at the cost of slightly degraded write-ability. However, tweaking the access transistor strength helps in compensating this loss and, thereby, resulting in an SRAM array with lower active V_{MIN} than baseline SRAM. Overall, the optimized PTM-SRAM bitcell can achieve an active V_{MIN} of 0.52 V, which is equal to the baseline SRAM V_{MIN} assisted by both WLUD and SVC techniques. For an iso-V_{MIN} of 0.52 V, the PTM-SRAM has 20% lower read access time, 35% lower read power, 16% higher write time, and 55% lower write power compared to baseline SRAM. Therefore, PTM-SRAM enables low-power operation and obviates the need for circuit-assist techniques. Statistical analysis describing the various factors (PTM parameters and transistor Vt) impacting the SRAM bitcell operation are presented. Design recommendations for achieving optimal read, write, and retention operations are also discussed.

References

[1] J. Gu, J. Keane, S. Sapatnekar, and C. Kim, "Width quantization aware FinFET circuit design," in *Proc. IEEE Custom Integr. Circuits Conf.*, Sep. 2006, pp. 337–340.

[2] M. Jurczak, N. Collaert, A. Veloso, T. Hoffmann, and S. Biesemans, "Review of FINFET technology," in *Proc. IEEE Int. SOI Conf.*, Oct. 2009, pp. 1–4.

[3] M.-H. Chiang, J.-N. Lin, K. Kim, and C.-T. Chuang, "Random dopant fluctuation in limited-width FinFET technologies," *IEEE Trans. Electron Devices*, vol. 54, no. 8, pp. 2055–2060, Aug. 2007.

[4] E. Baravelli, M. Jurczak, N. Speciale, K. De Meyer, and A. Dixit, "Impact of LER and random dopant fluctuations on FinFET matching performance," *IEEE Trans. Nanotechnol.*, vol. 7, no. 3, pp. 291–298, May 2008.

[5] H. Nho *et al.*, "A 32 nm high-K metal gate SRAM with adaptive dynamic stability enhancement for low-voltage operation," in *IEEE Int. Solid-State Circuits Conf. (ISSCC) Dig. Tech. Papers*, Feb. 2010, pp. 346–347.

[6] E. Karl *et al.*, "A 4.6 GHz 162 Mb SRAM design in 22 nm tri-gate CMOS technology with integrated active VMIN-enhancing assist circuitry," in *IEEE Int. Solid-State Circuits Conf. (ISSCC) Dig. Tech. Papers*, Feb. 2012, pp. 230–232.

[7] Y. Wang *et al.*, "Dynamic behavior of SRAM data retention and a novel transient voltage collapse technique for 0.6 V 32 nm LP SRAM," in *IEDM Tech. Dig.*, Dec. 2011, pp. 1–32.

[8] E. Karl *et al.*, "17.1 A 0.6V 1.5 GHz 84Mb SRAM design in 14 nm FinFET CMOS technology," in *IEEE Int. Solid-State Circuits Conf. (ISSCC) Dig. Tech. Papers*, Feb. 2015, pp. 1–3.

[9] J. Chang *et al.*, "12.1 A 7 nm 256 Mb SRAM in high-K metal-gate FinFET technology with write-assist circuitry for low-VMIN applications," in *IEEE Int. Solid-State Circuits Conf. (ISSCC) Dig. Tech. Papers*, Feb. 2017, pp. 206–207.

[10] J. Park, T. Hadamek, A. B. Posadas, E. Cha, A. A. Demkov, and H. Hwang, "Multi-layered $NiO_y/NbO_x/NiO_y$ fast drift-free threshold switch with high I_{on}/I_{off} ratio for selector application," *Sci. Rep.*, vol. 7, no. 1, Dec. 2017, Art. no. 4068.

[11] N. Shukla *et al.*, "A steep-slope transistor based on abrupt electronic phase transition," *Nature Commun.*, vol. 6, no. 1, Nov. 2015, Art. no. 7812.

[12] S. Teja and J. P. Kulkarni, "Soft-FET: Phase transition material assisted Soft switching field effect transistor for supply voltage droop mitigation," in *Proc. 55th ACM/ESDA/IEEE Design Autom. Conf. (DAC)*, Jun. 2018, pp. 1–6.

[13] K. Kosuge, "The phase transition in VO_2," *J. Phys. Soc. Jpn.*, vol. 22, no. 2, pp. 551–557, 1967.

[14] T. Sakata, K. Sakata, and I. Nishida, "Study of phase transition in NbO_2," *Phys. Status Solidi*, vol. 20, no. 2, pp. K155–K157, 1967.

[15] Y. Zhou, X. Chen, C. Ko, Z. Yang, C. Mouli, and S. Ramanathan, "Voltage-triggered ultrafast phase transition in vanadium dioxide switches," *IEEE Electron Device Lett.*, vol. 34, no. 2, pp. 220–222, Feb. 2013.

[16] E. Cha, J. Park, J. Woo, D. Lee, A. Prakash, and H. Hwang, "Comprehensive scaling study of NbO_2 insulator-metal-transition selector for cross point array application," *Appl. Phys. Lett.*, vol. 108, no. 15, Apr. 2016, Art. no. 153502.

[17] W. Dong, P. Li, and G. M. Huang, "SRAM dynamic stability: Theory, variability and analysis," in *Proc. IEEE/ACM Int. Conf. Computer-Aided Design*, Nov. 2008, pp. 378–385.

[18] S. Sinha, G. Yeric, V. Chandra, B. Cline, and Y. Cao, "Exploring sub-20 nm FinFET design with predictive technology models," in *Proc. DAC Design Automat. Conf.* IEEE, 2012, pp. 283–288.

[19] W.-Y. Tsai *et al.*, "Enabling new computation paradigms with HyperFET-an emerging device," *IEEE Trans. Multi-Scale Comput. Syst.*, vol. 2, no. 1, pp. 30–48, Jan. 2016.

[20] S. Srinivasa *et al.*, "Correlated material enhanced SRAMs with robust low power operation," *IEEE Trans. Electron Devices*, vol. 63, no. 12, pp. 4744–4752, Dec. 2016.

[21] M. Jerry, N. Shukla, H. Paik, D. G. Schlom, and S. Datta, "Dynamics of electrically driven sub-nanosecond switching in vanadium dioxide," in *Proc. IEEE Silicon Nanoelectron. Workshop (SNW)*, Jun. 2016, pp. 26–27.

[22] A. Aziz, N. Shukla, S. Datta, and S. K. Gupta, "Steep switching hybrid phase transition FETs (hyper-FET) for low power applications: A device-circuit co-design perspective—Part II," *IEEE Trans. Electron Devices*, vol. 64, no. 3, pp. 1358–1365, Jan. 2017.

[23] L. T. Clark *et al.*, "ASAP7: A 7-nm finFET predictive process design kit," *Microelectron. J.*, vol. 53, pp. 105–115, Jul. 2016.

[24] C. Auth *et al.*, "A 10 nm high performance and low-power CMOS technology featuring 3rd generation FinFET transistors, self-aligned quad patterning, contact over active gate and cobalt local interconnects," in *IEDM Tech. Dig.*, Dec. 2017, p. 29.

[25] J. Frougier *et al.*, "Phase-transition-FET exhibiting steep switching slope of 8 mV/decade and 36% enhanced ON current," in *Proc. IEEE Symp. VLSI Technol.*, Jun. 2016, pp. 1–2.