IEEE Journal on Exploratory Solid-State Computational Devices and Circuits

Received 8 October 2021; revised 2 November 2021; accepted 20 November 2021.
Date of publication 25 November 2021; date of current version 23 December 2021.

Digital Object Identifier 10.1109/JXCDC.2021.3130839

Ultra-Low-Voltage UTBB-SOI-Based, Pseudo-Static Storage Circuits for Cryogenic CMOS Applications

S. S. TEJA NIBHANUPUDI [1,*] (Graduate Student Member, IEEE), SIDDHARTHA RAMAN SUNDARA RAMAN [1,*], MIKAËL CASSÉ [2] (Member, IEEE), LOUIS HUTIN [2] (Member, IEEE), and JAYDEEP P. KULKARNI [1] (Senior Member, IEEE)

[1]Department of Electrical and Computer Engineering, The University of Texas at Austin, Austin, TX 78712 USA
[2]CEA-Leti Minatec, Université Grenoble Alpes, 38400 Grenoble, France

CORRESPONDING AUTHORS: S. S. T. NIBHANUPUDI (subrahmanya_teja@utexas.edu) and J. P. KULKARNI (jaydeep@austin.utexas.edu)

This work was supported in part by NSF NASCENT ERC, in part by Intel, and in part by Micron.

*S. S. Teja Nibhanupudi and Siddhartha Raman Sundara Raman contributed equally to this work.

ABSTRACT Operating CMOS circuits at cryogenic temperatures offers advantages of higher mobility, higher ON-current, and better subthreshold characteristics, which can be leveraged to realize high-performance CMOS circuits. However, an ultra-low-voltage operation is necessary to minimize the power consumption and to offset the cooling cost overheads. The MOSFET threshold voltages (Vt) increase at cryogenic temperatures making it challenging to achieve high performance while operating at very low voltage. Ultra-thin body and buried oxide silicon-on-insulator (UTBB-SOI)-based MOSFETs can modulate the transistor threshold voltage using the back-gate bias, unlike conventional FinFETs. This unique UTBB-SOI technology attribute has been leveraged to realize compact pseudo-static storage circuits, namely, embedded dynamic random access memory (DRAM) bitcell and a flip-flop operating at 0.2 V and 77 K. This article presents UTBB-SOI device fabrication details and calibrate experimental device characteristics with BSIM compact models. SPICE simulations suggest the feasibility of three-transistor gain-cell embedded DRAM (eDRAM) capable of reliably storing three distinct voltage levels (1.5 bits/cell) and exhibiting retention time of the order of 10^4 s. Furthermore, a unique pseudo-static flip-flop design is presented, which can lower the clock power by 50%, transistor count by 20%, and static power consumption by 20%.

INDEX TERMS Cryo-CMOS, embedded dynamic random access memory (eDRAM), flip-flop, pseudo-static, retention time, ultra-thin body and buried oxide silicon-on-insulator (UTBB-SOI).

I. INTRODUCTION

THE rapid growth in data-intensive applications has accelerated the need for computing systems having high-density energy-efficient memory combined with high-performance computing capability. The increased short-channel effects in advanced CMOS technology nodes have limited the threshold voltage and active gate length scaling, resulting in the transistor performance not being sufficient to meet the growing demands of high-performance computing applications. Cryogenic operation (temperature ~77 K) has emerged as a technology booster that strikes the right balance between low voltage and high-performance operation [1]–[4]. The low-temperature process provides advantages, such as increased mobility and steeper subthreshold characteristics, which leads to enhanced transistor ON/OFF ratio [1], [5], although one of the limitations for low-temperature operation is the shift in Fermi potential combined with an increase in the bandgap, leading to the increased threshold voltage [5]. In addition, cryo-CMOS requires extreme low voltage operation to keep the cooling cost overhead at a manageable level. Hence, it is vital to identify technologies that enable modulating threshold voltages at such low temperatures to achieve higher performance while operating at ultra-low supply voltage.

CMOS FinFET-based technology is the state-of-the-art transistor technology favored for high-performance computing applications. With the channel undoped in the FinFET transistors (to reduce random dopant fluctuations [6]), the

threshold voltage (V_T) tuning is achieved by work function engineering [7]. Therefore, to achieve sub-100-mV V_T, a wider range of effective work function gate metals is required (below 4.0 eV for nMOSFETs and above 5.2 eV for pMOSFETs) for advanced FinFETs. Such extreme work-function metals need to be extensively researched for their successful integration into high-volume manufacturing of advanced FinFETs and beyond-CMOS devices.

In contrast to FinFET transistors, ultra-thin body and buried oxide (BOX) silicon-on-insulator (UTBB-SOI) transistors [8], [9] have an independent back gate that can effectively lower V_T of the transistor. The V_T sensitivity to the back-gate bias (body factor) can be modulated by adjusting the thickness of the BOX layer (typically, ∼10–30 nm). Furthermore, the backplane well doping (silicon region below the BOX layer) determines the work function of the back gate (n-well work function $<$$p$-well work function), which, in turn, modifies the V_T of the transistor. Therefore, there are multiple V_T tuning knobs available in the UTBB-SOI technology, which can be leveraged to achieve sub-100-mV V_T transistors. The ultra-thin silicon channel ensures superior electrostatics effectively suppressing undesired short-channel effects and significantly reducing the junction leakage. Experimental demonstrations of UTBB-SOI transistors have exhibited comparable performance to FinFET transistors [10]. Furthermore, the V_T variations have been demonstrated to be low in undoped UTBB-SOI transistors [11]. This work leverages the ease of threshold voltage tuning in UTBB-SOI technology using available work-function metals to demonstrate high-performance transistors operating at ultra-low voltage experimentally. Extreme low leakage currents in UTBB-SOI transistors are leveraged to realize compact pseudo-static storage circuits having higher storage density and lower power consumption.

Dynamic random access memory (DRAM) with ultra-low leakage current operating at cryogenic temperature can yield a pseudo-static memory operation that does not require frequent refresh operations. Furthermore, DRAM write operation fundamentally does not experience any contention, unlike a static random access memory (SRAM) write operation [12]. In addition, a gain-cell embedded DRAM (eDRAM) [13] with a dedicated read port offers nondestructive read operation, which can enable multilevel cell (MLC) storage functionality and improve the bitcell storage density. The MLC functionality makes gain-cell eDRAM a viable candidate for high-density, ultra-low-voltage, cryogenic memory technology. In this article, we evaluate the performance of a 3T gain-cell eDRAM for storing three distinct voltage levels in a single gain cell, achieving 1.5-bits/cell functionality.

From the power consumption perspective, the subthreshold leakage and other temperature-dependent leakage currents are lowered significantly at the cryogenic temperature. Hence, the dynamic switching power is the dominant power contributor at the cryogenic conditions and needs to be minimized to keep the cooling cost overheads minimum.

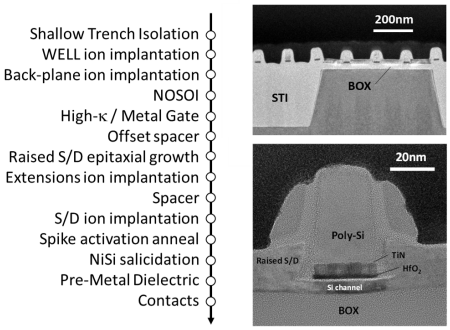


FIGURE 1. Left: simplified front-end-of-line process flow. Top right: transmission electron microscopy (TEM) showing transistor gates on isolation (STI) and active areas. Bottom right: TEM cross-sectional close-up view of a transistor after second spacer definition.

Among various design components contributing to the dynamic power, flip-flops contribute ∼20% of the total dynamic power consumption in modern CPUs, despite adopting aggressive clock gating techniques [14]. This is due to toggling transmission gates and tristate inverter nodes within a flip-flop circuit driven by an active clock signal. In this article, we present a pseudo-static flip-flop design that leverages the intrinsic gate capacitance of an inverter as a flip-flop storage element. It lowers the clocking power by 50%, flip-flop transistor count by 20% with minimal performance impact compared to the conventional flip-flop design.

This article is organized as follows. Section II presents experimental results for the UTBB-SOI-based N and P MOSFETs, along with model calibration to the experimental results. Section III evaluates multilevel, high refresh time eDRAM technology in the presence of process variations. A pseudo-static flip-flop is presented in Section IV, along with the performance, power, and area comparison with the conventional flip-flop.

II. DEVICE MODELING

A. EXPERIMENT

Fully depleted SOI n and pMOSFETs were fabricated in 28-nm ground rules with a gate-first high-κ metal gate process on 300-mm (100) SOI wafers with a BOX thickness of 25 nm (see Fig. 1). The undoped silicon channel thickness is 7 nm after complete processing. Adjacent active areas are separated using shallow trench isolation (STI), and ion-implanted wells are defined to electrically connect the BOX back-interface to substrate plugs defined later in the process. Additional shallower ''back-plane'' doping is performed directly beneath the BOX for static optimization of V_{th} through a fine localized adjustment of the back-gate work function.

Several combinations of device well polarities and substrate biases are possible in this technology, offering extended threshold voltage tunability for high performance or low-power CMOS optimization. The most significant configurations are shown in Fig. 2; we aim to counterbalance the threshold voltage increase at 77 K to achieve high performance at a reduced drain voltage while benefiting from

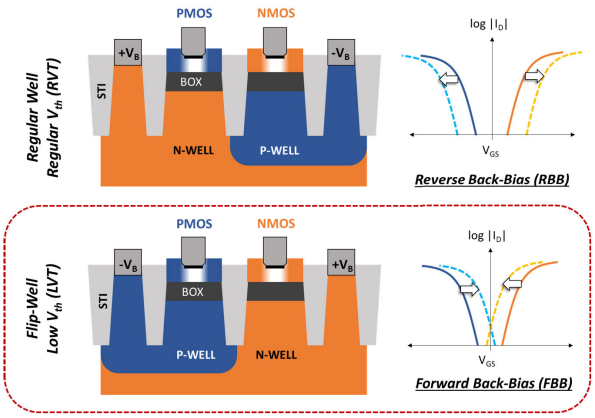


FIGURE 2. Top: regular well architecture for RVT flavor pMOS and nMOS, with symmetrically applied reverse back-biasing resulting in increased threshold voltages. Bottom: flip-well architecture for LVT flavor pMOS and nMOS under symmetrical forward back-bias leading to decreased threshold voltages.

the steeper subthreshold slope, keeping the leakage current low. To this effect, the most suitable configuration is the flip-well architecture with forward back-gate biasing (FBB), i.e., positive (resp. negative) bias on N-WELL (resp. P-WELL) for nMOS (resp. pMOS) [15].

Test dies cleaved from a wafer were mounted on a sample holder in a tabletop lakeshore cryogenic probe station with four adjustable contact needles connected to source/measurement units. The chamber was cooled down under continuous helium flow with temperature regulation between 300 K and 77 K. The data acquisition was performed using a semiconductor parameter analyzer (HP 4156) with a noise floor of 50 fA. Isolated test devices were characterized at 77 K with the necessary forward back-bias to lower their threshold voltage down to sub-100-mV values (V_T is quantified using the constant current $|I_D| = 10^{-7} \times W/L$ criterion).

B. BSIMIMG MODEL CALIBRATION

Fig. 3(a) shows the I_d–V_{gs} characteristics of an LVT nMOS device (flipped well) for $V_{\rm DS} = 0.2$ V operating at 77 K. The dimensions for the device are: gate length $L = 100$ nm, channel width $W = 2$ μm, silicon layer thickness $T_{\rm Si} = 7$ nm, front gate EOT $= 3.7$ nm, and BOX thickness $= 25$ nm. Fig. 3(a) also shows the characteristics for different back-gate biases ranging from 0 to 3 V. The gate current is below the noise floor (50 fA) for the entire range of gate voltage biases indicating the extreme low gate leakage. The experimental data are calibrated to the BSIMIMG (version 102.9.2) compact model for SPICE circuit simulations. The compact model device specifications, such as channel length (100 nm), width (2 μm), box thickness (25 nm), and back-gate work function, are kept the same as the fabricated device. The compact model parameters, such as NBODY (channel doping), U0 (low-field mobility), UTL (mobility temperature co-efficient), and K0 (lateral nonuniform doping), are optimized to obtain a good fit with experimental data as seen from Fig. 3(a). Similarly, Fig. 3(b)

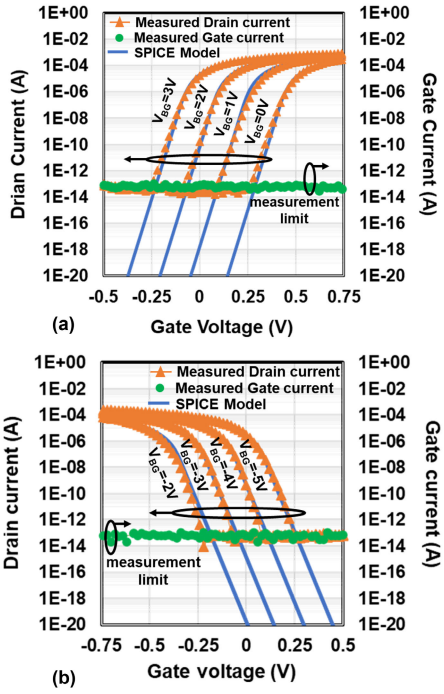


FIGURE 3. (a) Measured and SPICE simulated characteristics for various forward back biasing conditions on an LVT nMOS device at 77 K with channel width $W = 2$ μm, gate length $L = 100$ nm, and front-gate equivalent oxide thickness EOT = 3.7 nm. The applied drain-to-source voltage is 200 mV. The gate current (right axis) remains below the noise floor of the measurement equipment across the gate voltage sweeping range. (b) Measured and SPICE simulated characteristics for various forward back biasing conditions on an LVT pMOS device at 77 K with the exact device dimensions as nMOS and source-to-drain voltage of 0.2 V.

shows the model calibrated to experimental data for the LVT pMOS device. The calibrated models are used for circuit simulations in Sections III and IV.

C. TCAD MODEL CALIBRATION

As mentioned in Section II-A, the lowest current level detected by the measurement setup is limited to 50 fA. To reliably estimate the current below this limit, the transistor characteristics are simulated using a multidimensional device simulator, such as Sentaurus TCAD (technology computer-aided design). TCAD analysis would account for possible leakage mechanisms, such as gate induced drain leakage (GIDL), band-to-band tunneling (BTBT), gate tunneling leakage, and junction leakage [16]. Fig. 4(a) shows the cross section of the transistor model in TCAD, which adopts the experimental device dimensions. The simulations employ the Philips Unified Mobility (PhuMob) model coupled with the Lombardi thin-layer mobility model to accurately capture the carrier transport inside the transistor at 77 K lattice temperature. The gate tunneling current is simulated by activating both the Fowler Nordheim tunneling and direct tunneling models [16]. The transistor channel doping is optimized to obtain a good fit with the experimental

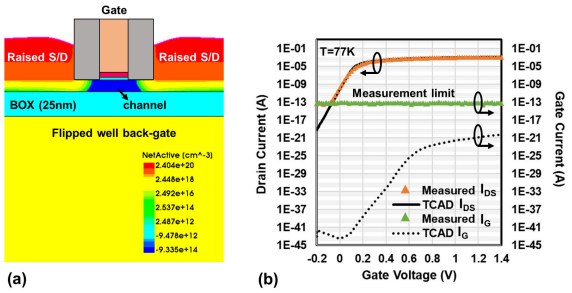


FIGURE 4. (a) UTBB-SOI device TCAD cross section. (b) TCAD simulated I_d–V_g characteristics for LVT nMOS transistor at 77 K temperature.

device characteristics, as shown in Fig. 4(b). The simulated drain-to-source current closely traces the experimental data above the noise floor of 50 fA. The thicker gate oxide (∼3.7 nm) limits the gate leakage current below 10^{-20} A across the entire range of gate voltage biases. The simulations also highlight that the junction leakage current is negligible due to the reduced junction area in the SOI technology. This component of leakage is higher in transistors with bulk substrate connections.

D. BACK BIASING FLEXIBILITY

There are some practical constraints on boundaries for the back-bias V_B in the integration route described above. Independent control of adjacent *P*- and *N*-WELL electrostatic potentials can be compromised if the diode that they form is placed under forward bias, setting the condition $V_{\text{P-WELL}} - V_{\text{N-WELL}} < 0.6$ V. This is the main reason why, in general, positive (resp. negative) biases are applied to the *N*-WELL (resp. *P*-WELL), making the flip-well configuration naturally amenable to V_{th} lowering by FBB.

On the other hand, reverse breakdown of the diode should also be avoided, setting the condition $V_{\text{N-WELL}} - V_{\text{P-WELL}} < 6$ V. In the case of symmetrical biasing, such as described in Fig. 2, this would translate to $V_p < 3$ V, a constraint that needs to be transgressed to reach sub-100-mV threshold voltage on devices with more aggressive front-gate EOT (1.5 nm). One way of circumventing this issue is to improve the body factor by decreasing the BOX thickness. Another could be to resort to a dual-STI structure, effectively separating adjacent wells of opposite polarity by deeper trenches while allowing substrate plugs to remain connected to their wells beneath shallower trenches.

Mixing and matching V_{th} flavors in adjacent blocks may also cause singularity points and well continuity issues, as exemplified in Fig. 5 These can be mitigated by the use of transition cells and a deep *N*-WELL implantation level. Note that the nMOS-only bit cell studied in Section III (see Fig. 6) is not affected by these risks.

III. MULTILEVEL PSEUDO-STATIC MEMORY BITCELL

The UTBB-SOI transistors operating at 77 K have a steep subthreshold slope (∼25 mV/dec), significantly reducing the drain-to-source leakage current. Operating at low voltages

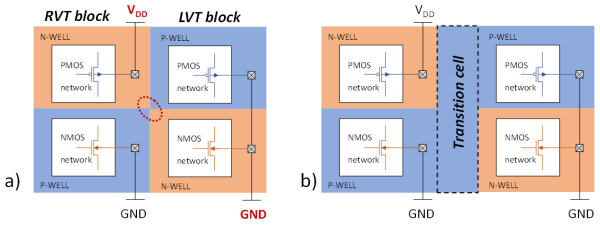
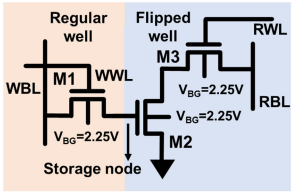


FIGURE 5. (a) Direct abutment of an RVT block (regular well) with an LVT block (flip well), leading to an undesirable singularity point. The indicated body biases correspond to the reference conditions ($V_B = 0$ V). The contact between two *N*-WELL regions biased at different values is particularly problematic. (b) Use of a P-WELL-based transition cell to avoid the singularity.



	Read	Write	Retention
WWL	-0.2V	0.4V	-0.2V
WBL	X	0.2/0V	X
RWL	0.4V	0V	0V
RBL	0.2V	X	0.2V

(b)

FIGURE 6. (a) Schematic of 3T eDRAM bitcell using UTBB-SOI FETs. (b) Operating conditions of the eDRAM bitcell.

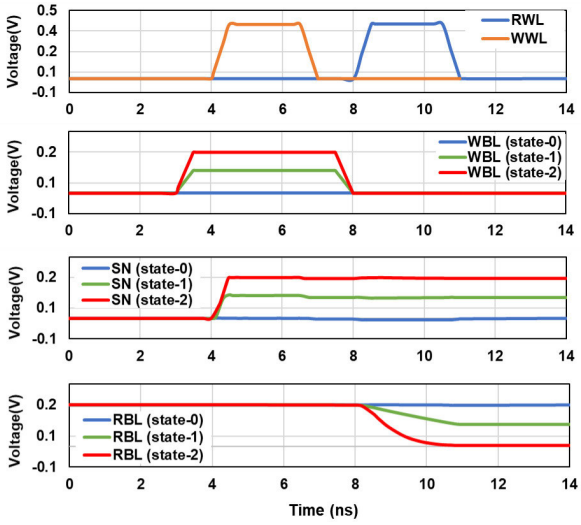


FIGURE 7. Timing diagram highlighting the write followed by a read operation for the three-level pseudo-static memory bitcell.

(200 mV) further reduces electric field-induced leakage components, as demonstrated in Section II-C. Overall, the reduced leakage current can be leveraged to realize a pseudo-static, high-density eDRAM bitcell with a long retention time. This section evaluates the performance of multilevel, pseudo-static eDRAM bitcell designed using UTBB-SOI transistors operating at an ultra-low-voltage and cryogenic temperature conditions.

A. BITCELL OPERATION

Fig. 6(a) shows the schematic of the eDRAM bitcell. The bitcell is designed using three nMOSFET transistors with

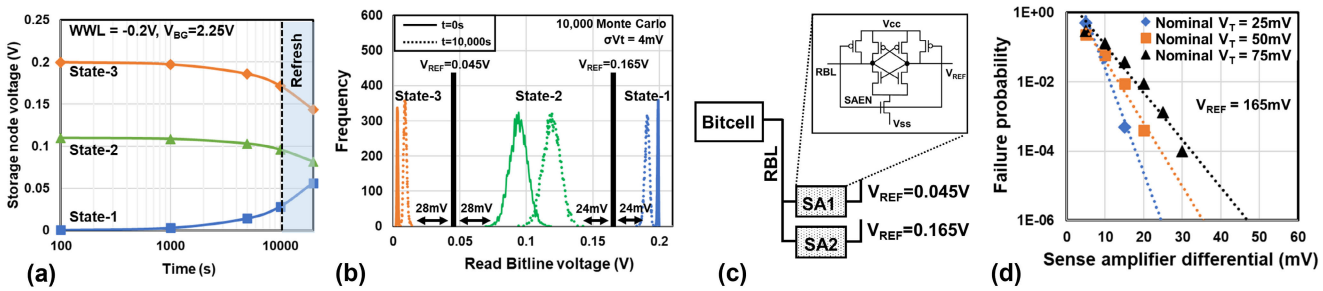


FIGURE 8. (a) SN variation with time for the three states. (b) RBL voltage distribution in the presence of variation for $t = 0$ s and $t = 10\,000$ s. (c) Sensing scheme to resolve the three bits. Inset: schematic of the conventional latch-type sense amplifier. (d) Failure probability of sense amplifier in resolving state-1 and state-2.

two flavors of UTBB-SOI transistors–low-V_T transistor and high-V_T transistor. The low-V_T transistor ($V_T = 75$ mV) is implemented using flipped well configuration, and the high-V_T transistor ($V_T = 150$ mV) is implemented using regular well configuration holding the back-gate bias at 2.25 V. Keeping the back-gate bias constant across all the transistors within the bitcell avoids any integration constraints (diode forward-bias or reverse-bias), as discussed in Section II-D. The high-V_T transistor is employed as the write port transistor (M1) to reduce leakage current. The low-V_T transistors are used as the read port transistor (M2) and the read access transistor (M3) to increase the bitline swing during read-out. Having a dedicated read port enables read-disturb-free operation, facilitating MLC storage on the eDRAM bitcell. Data are written into the bitcell by asserting the write wordline (WWL) and biasing the write bitline (WBL) to the desired voltage. The charge stored on the gate electrode of the read-port transistor (M2) is depleted gradually by the extremely low leakage current of M1 and M2 transistors. This allows storing multiple voltage levels on the bitcell storage capacitance. The eDRAM bitcell is utilized to store three-states/cell—state-1 (0 V), state-2 (0.11 V), and state-3 (0.2 V). During a write operation, the WWL signal is boosted to 0.4 V (to overcome the V_T drop of the M1 transistor) with the WBL biased to a desired voltage (0, 0.11, or 0.2 V). During a retention phase, lowering the WWL signal to -0.2 V lowers the leakage current to 10^{-6} fA, which increases the retention time. The read operation begins by precharging the read bitline (RBL) to Vcc (0.2 V) followed by asserting the read wordline (RWL). The bitline capacitance (assumed to be 30 fF in this study) discharges depending on the charge stored at the storage node (SN). The read pulse duration is assumed to be 2 ns in this study. Fig. 6(b) summarizes the voltages applied to various control signals during read, write, and retention modes of operation.

Fig. 7 shows the timing waveform of eDRAM bitcell during write and read operation for the three storage states. The SN is programmed to 0, 0.11, and 0.2 V for state-1, state-2, and state-3, respectively. The RBL voltage does not discharge for state-1. The RBL voltage discharges to 0.1 and 0 V for state-2 and state-3, respectively. This difference in the bitline voltage is resolved by a sense amplifier to determine the state stored in the bitcell.

B. BITCELL PERFORMANCE

The subthreshold leakage of the wordline access transistor (M1) reaches below 10^{-21} A when the gate (WWL) is biased to -0.2 V. Similarly, the gate leakage of the M2 transistor is negligible compared to the subthreshold leakage [$I_G < 10^{-32}$ A at $V_{GATE} = 0.2$ V, as seen from Fig. 4(b)]. Such low leakage current paths ensure that the charge is retained on the SN for ~10 000 s, essentially making the bitcell pseudo-static in nature. Fig. 8(a) shows the retention characteristics of the SN for the three states. To capture the worst case leakage, the WBL is biased at 0 V when the bitcell is programmed to either state-2 or state-3. For state-1 programming, the WBL is biased at 0.2 V. The leakage current is lower in state-2 since the voltage difference between SN and WBL nodes is only 0.11 V compared to 0.2 V for state-1 or state-3. The states are very stable until 1000 s and start to degrade after that. The separation between state-1 and state-2 reduces to 65 mV at 10 000 s and tends to collapse at 25 000 s [not shown in Fig. 8(a)].

The effect of transistor V_T variation is captured by statistical Monte Carlo (MC) analysis of the bitcell; 10 000 run MC simulations (assuming σ-$V_T = 5\%$ of nominal V_T) are performed on the read operation. Fig. 8(b) shows the RBL voltage distribution for the three storage states. The RBL voltage is measured at the end of the read cycle for each state. State-1 and state-3 have very narrow distributions (σ-RBL < 1 mV), whereas state-2 has wider distribution (σ-RBL~8 mV). This behavior is observed since the state-2 storage voltage is within the high trans-conductance region of the transistor. Therefore, the voltage level of state-2 has been carefully chosen after thorough optimizations to ensure sufficient separation of RBL voltage levels for accurate sensing. Fig. 8(b) also plots the RBL voltage distribution when the read operation is performed 10 000 s after the write operation. The mean of each distribution shifts due to the degradation of voltage levels at the SN. The RBL voltage distribution for state-2/state-3 shifts to the right as the SN node discharges and reduces the drive strength (overdrive voltage) of the M2 transistor. Similarly, the distribution shifts to the left for state-1 as the SN node charges, thereby increasing the drive strength (overdrive voltage) of the M2 transistor. At 10 000 s, the RBL separation between state-1 and state-2 reduces to 48 mV and between state-2 and state-3

reduces to 56 mV. Therefore, the sense amplifier needs to reliably resolve the bits with the input differential voltage of 24 mV, as shown in Fig. 8(b).

C. SENSE AMPLIFIER OPERATION

The three states of the bitcell can be resolved by adopting a sensing scheme, as shown by the schematic in Fig. 8(c). The RBL is connected to two latch-type sense amplifiers (SA1 and SA2) with different reference voltages. The reference voltages for SA1 and SA2 are chosen between the voltage levels of the states, as shown in Fig. 8(b). The output of SA1 and SA2 at the end of the sensing operation provides information about the stored state. For example, both SA1 and SA2 outputs will be "0" for state-1. Similarly, the outputs of SA1 and SA2 will be "1" and "0" for state-2 and "1" and "1" for state-3, respectively.

The reference voltages for the sense amplifier are chosen based on the RBL voltage distribution at $t = 10\,000$ s. This approach ensures that the sense amplifier can reliably resolve the bits under worst case retention and V_T variation conditions. Furthermore, we also consider the transistor V_T variations within the sense amplifier and capture the impact on sensing margin through 10 000 run MC simulations. Fig. 8(d) shows the variation of SA failure probability with increasing SA differential. The reference voltage is held at 165 mV for this study since state-1 and state-2 collapse faster toward each other, thereby accounting for the worst case sense amplifier input scenario. Fig. 8(d) also shows the failure probability for SA designed using different V_T flavors. The SA designed using higher V_T exhibits higher failure probability due to the transistor's smaller ON-current. The trend lines from failure statistics are extrapolated to the failure probability of 10^{-6} to quantify the minimum SA differential required to meet one SA failure in the 1-Mb array target. The SA designed using an ultra-low-V_T transistor ($V_T = 25$ mV) achieves a minimum SA differential of 23 mV. This meets the requirements needed to resolve all the three states reliably under worst case retention ($t = 10\,000$ s) and variations conditions.

IV. PSEUDO-STATIC FLIP-FLOP

A. FLIP-FLOP CIRCUIT DESIGN

As shown in Fig. 9(a), conventional flip-flop designs comprise primary and secondary latches that utilize tristated inverters connected in the feedback path. These tristated inverters contribute to the clock load, thereby resulting in increased dynamic power consumption. At cryogenic temperatures, the extremely low leakage of the pass-gate transistor can be leveraged to realize pseudo-static flip-flop without the tristated inverter, as shown in Fig. 9(b). However, a periodic refresh operation is required due to the dynamic nature of the SN, which is realized by the refresh MUX that selects Q during a refresh operation. The proposed pseudo-static flip-flop design has 20% fewer transistors and consumes 50% lower clock power than the conventional flip-flop design despite the added refresh logic.

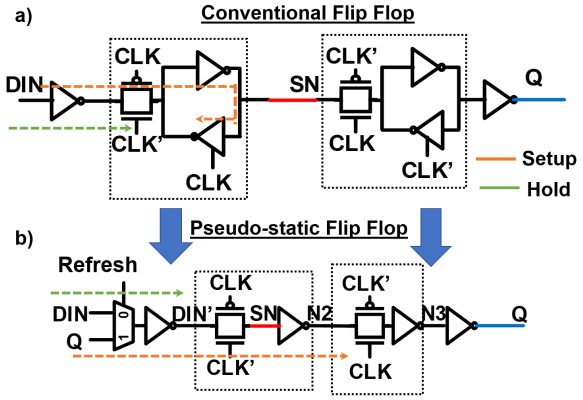


FIGURE 9. (a) Conventional flip-flop with tristate cross-coupled inverter storing the data. (b) Proposed pseudo-static flip-flop with gate capacitance of the inverter as the SN with datapath for setup and hold analysis marked.

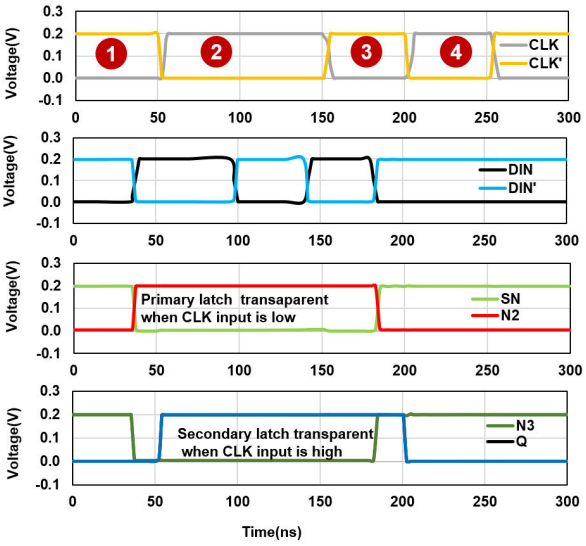


FIGURE 10. Timing diagram for a positive edge-triggered flip-flop when in nonrefresh mode. The primary latch is transparent during Phases 1 and 3 and the secondary latch during Phases 2 and 4, respectively, with phases marked in red circles.

B. FLIP-FLOP OPERATION

Fig. 10 shows the timing diagram of the positive edge-triggered pseudo-static flip-flop during a normal mode of operation (i.e., nonrefresh operation). The primary latch is transparent when the inverted CLK signal is high (during Phase 1 and Phase 3). This allows DIN′ to be transferred onto the SN. Here, the gate capacitance of the inverter is utilized as the SN. The secondary latch is sensitive during the positive half of the clock cycle, and the data stored on the SN node are transferred onto the Q node.

C. REFRESH OPERATION

Conventional flip-flop designs are static due to cross-coupled inverter pairs that preserve the SN values. In the case of the pseudo-static flip-flop, the voltage at SN is subject to leakage due to subthreshold conduction of the transmission-gate transistors. Therefore, the charge at SN needs to be

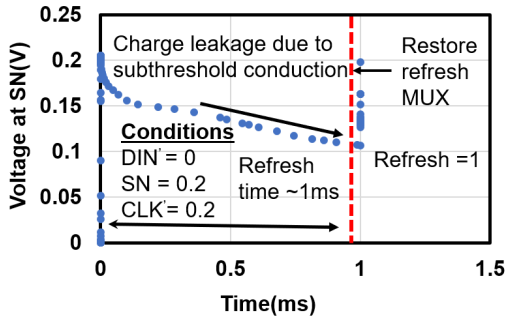


FIGURE 11. Voltage at SN drops down to 0.1 V in 1 ms, assuming the worst case scenario of DIN′ pulled low, CLK′ node driven high, and SN is high. This voltage drop is restored using a refresh MUX.

restored periodically to restore the flip-flop contents. The restore operation is performed by feeding the flip flop's output (Q) back as an input to a 2:1 MUX controlled by a "Refresh" signal. This refresh operation is very infrequent, and the refresh time is around 1 ms, as shown in Fig. 11. Here, the refresh time is quantified as the time required for the voltage at the N2 node to change by Vcc/2. This analysis is performed considering the worst case leakage scenario, i.e., DIN′ held at "0" when SN is charged to "1" and vice versa. Since the refresh time interval is orders of magnitude larger than the operating clock cycle period (MHz–GHz clock frequency), the power and latency overhead due to a pseudo-static flip-flop refresh operation can be significantly amortized.

D. PERFORMANCE ANALYSIS

The performance of flip flops can be quantified in terms of the setup time, hold time, and Clk→Q delay metrics. The setup time requirement arises due to finite delay for the data to traverse the primary latch before the arrival of the clock at the secondary latch's transmission gate. The setup time for the pseudo-static flip-flop is quantified by measuring the time delay for data arrival at the second transmission gate in the presence of process variations. Fig. 12(a) presents setup time variation for the pseudo-static flip-flop and conventional flip-flop in the presence of process variations by performing 10 000 run MC simulations assuming 1σ-V_T of 4 mV. In the pseudo-static flip-flop design, the data must traverse a MUX path (designed using transmission gates) → inverter → transmission gate → inverter before reaching the N2 node. On the contrary, the worst case datapath in conventional design is inverter → transmission gate → the cross-coupled inverter pair to reach the N2 node. The delay of the transmission gate MUX is slightly higher compared to the cross-coupled inverter pair. This results in a conventional flip-flop design having a shorter setup time of 38 than 41 ps in the proposed design.

In case of hold time, the data at the primary latch's transmission gate's input must be stable even after the clock edge arrives to account for the finite delay in turning off the primary latch. Thus, the hold analysis is performed at the input of the primary latch transmission gate (DIN′) node. In the conventional flip-flop design, the hold time, i.e., the difference between the clock arrival and the data arrival time, is 0 because the data and clock paths have one inverter delay. On the contrary, the hold time in the pseudo-static flip-flop design is negative because the datapath has to traverse a MUX and an inverter. In contrast, the clock signal traverses a single inverter, resulting in a lesser clock path delay. Fig. 12(b) shows the hold time comparison between the conventional and proposed flip-flops in the presence of process variations.

The Clk→Q delay is an essential metric in high-frequency designs that employ several flip-flop stages for computation. Increased Clk→Q delay on the launch flip-flop results in an increased datapath delay for the capture flip-flop, thereby limiting the maximum operating frequency. In conventional flip-flop and the proposed flip-flop, the data must traverse a transmission gate and two inverters to reach Q, thus having similar Clk→Q delay. Fig. 12(c) shows that the Clk→Q delay for the conventional and proposed flip-flops has an almost equal distribution around 31 ps, thus having minimal performance difference.

TABLE 1. Summary of performance, power, and area comparison between conventional and proposed flip-flops (Vcc = 200 mV and T = 77 K).

Parameter	Conventional F/F	Proposed Pseudo static F/F
Setup Time (ps)	38	41
Hold Time (ps)	0	-11
Clk→Q Delay (ps)	31	31
Transistor Count	20	16
Norm. Clock Power	1	0.5
Norm. Retention Power	1	0.8

E. POWER ANALYSIS

Clocking power is a significant component of the total dynamic power contributing around 30% in the case of a single-bit flip-flop design and about 50% in the case of multibit flip-flop designs [17]. The low operating voltage using UTBB-SOI helps in lowering the clock tree dynamic power. Furthermore, the pseudo-static flip-flop design lowers the clock dynamic power consumption due to the reduced clock load. The clock load power in the proposed technique is reduced by at least 50% compared to the conventional flip-flop design because of the reduction in the gate capacitance of the clock network by 2× (four transistors connected to the CLK in conventional versus two transistors connected to the CLK in proposed design). The cost of inversion of the clock network can be amortized by sharing the inverter across multiple flip-flops and does not contribute to power increase at the flip-flop level. Table 1 shows the normalized clock power (conventional/proposed clock power) comparison between the conventional and pseudo-static flip-flop designs.

The power dissipated in the flip-flop when the clock is turned off is a characteristic measure of the retention power. Conventional flip-flops have additional static leakage power associated with the cross-coupled inverter pairs. This leakage component is eliminated by using the capacitor as an SN, reducing leakage power by around 20% compared to the

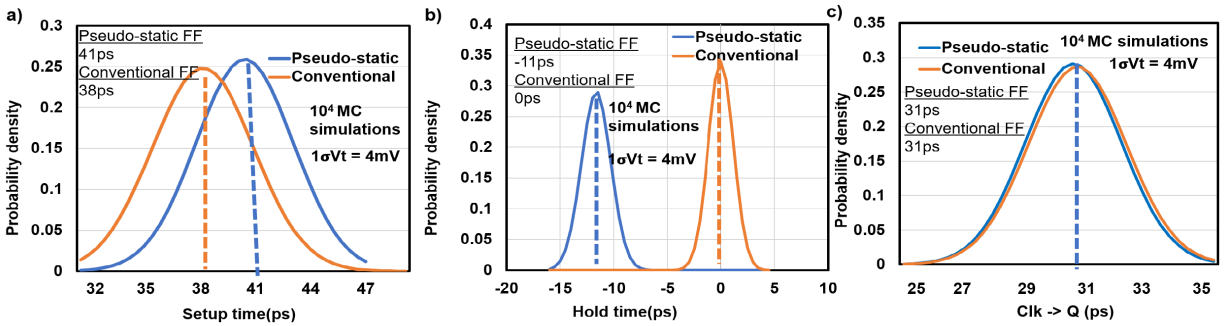


FIGURE 12. Statistical simulations showing (a) setup time, (b) hold time, and (C) Clk-Q delay distribution for the conventional flip-flop and the pseudo-static flip-flop operating at Vcc = 200 mV and T = 77 K.

conventional flop. The same analysis can be extended when the clock is ON, leading to lesser active power because of the symmetrical nature of primary and secondary latches.

F. AREA ANALYSIS

The pseudo-static flip-flop design has a lesser transistor count compared to a conventional flip-flop design. This can be attributed to eliminating the tristated inverters in the feedback path of primary and secondary latches. For the pseudo-static flip-flop design, refresh logic is implemented using a compact transmission-gate MUX (as opposed to OR-AND-Invert(OAI)22-based implementation) to lower the area overhead. Overall, the pseudo-static flip-flop reduces the transistor count from 20 to 16 (see Table 1), assuming the clocked inverter can be shared across multiple flip-flops.

V. CONCLUSION

This article presents an experimental demonstration of UTBB-SOI-based transistors operating at cryogenic temperatures. The flexible V_T tuning capability of the UTBB-SOI technology has been leveraged to realize transistors with sub-100-mV threshold voltage capable of operating at an ultra-low voltage of 0.2 V. Device measurements have been calibrated with SPICE models for enabling circuit simulations. Extreme low leakage at cryogenic temperature has been leveraged to design pseudo-static memory bitcells. The 3T gain-cell eDRAM having a considerable retention time of 10 000 s with a potential of storing three levels in a single bitcell has been presented. Read analysis in the presence of process variations is performed to determine the feasibility of reading out multiple levels. A pseudo-static flip-flop utilizing gate capacitance of an inverter as the SN has been presented. The proposed flip-flop has reduced bitcell area, reduced dynamic power compared to a conventional flip-flop. Setup, hold, and Clk-Q delay analyses have been performed in the presence of process variations to provide an insight into the timing impact.

ACKNOWLEDGMENT

The authors would like to thank Prof. Tsu-Jae King Liu from the University of California at Berkeley for helpful discussions. They would also like to thank Rishabh Sehgal and Sirish Oruganti for proofreading the manuscript.

REFERENCES

[1] R. Saligram, D. Prasad, D. Pietromonaco, A. Raychowdhury, and B. Cline, "A 64-bit arm CPU at cryogenic temperatures: Design technology co-optimization for power and performance," in *Proc. IEEE Custom Integr. Circuits Conf. (CICC)*, Apr. 2021, pp. 1–2.

[2] I. Byun, D. Min, G.-H. Lee, S. Na, and J. Kim, "CryoCore: A fast and dense processor architecture for cryogenic computing," in *Proc. ACM/IEEE 47th Annu. Int. Symp. Comput. Archit. (ISCA)*, May/Jun. 2020, pp. 335–348.

[3] J. C. Bardin *et al.*, "Design and characterization of a 28-nm bulk-CMOS cryogenic quantum controller dissipating less than 2 mW at 3 K," *IEEE J. Solid-State Circuits*, vol. 54, no. 11, pp. 3043–3060, Nov. 2019.

[4] F. Sebastiano *et al.*, "Cryogenic CMOS interfaces for quantum devices," in *Proc. 7th IEEE Int. Workshop Adv. Sensors Interfaces (IWASI)*, Jun. 2017, pp. 59–62.

[5] A. Beckers, F. Jazaeri, A. Ruffino, C. Bruschini, A. Baschirotto, and C. Enz, "Cryogenic characterization of 28 nm bulk CMOS technology for quantum computing," in *Proc. 47th Eur. Solid-State Device Res. Conf. (ESSDERC)*, Sep. 2017, pp. 62–65.

[6] C.-H. Lin *et al.*, "Channel doping impact on FinFETs for 22 nm and beyond," in *Proc. Symp. VLSI Technol. (VLSIT)*, Jun. 2012, pp. 15–16.

[7] S. Hung, "Multi-Vt engineering and gate performance control for advanced FinFET architecture," in *IEDM Tech. Dig.*, 2017.

[8] T. A. Karatsori *et al.*, "Analytical compact model for lightly doped nanoscale ultrathin-body and box SOI MOSFETs with back-gate control," *IEEE Trans. Electron Devices*, vol. 62, no. 10, pp. 3117–3124, Oct. 2015.

[9] M. Cassé and G. Ghibaudo, *Low Temperature Characterization and Modeling of FDSOI Transistors for Cryo CMOS Applications*. Rijeka, Croatia: InTech, 2021.

[10] W.-T. Chang, C.-T. Shih, J.-L. Wu, S.-W. Lin, L.-G. Cin, and W.-K. Yeh, "Back-biasing to performance and reliability evaluation of UTBB FDSOI, bulk FinFETs, and SOI FinFETs," *IEEE Trans. Nanotechnol.*, vol. 17, no. 1, pp. 36–40, May 2017.

[11] O. Weber *et al.*, "High immunity to threshold voltage variability in undoped ultra-thin FDSOI MOSFETs and its physical understanding," in *IEDM Tech. Dig.*, Dec. 2008, pp. 1–4.

[12] G.-H. Lee, S. Na, I. Byun, D. Min, and J. Kim, "CryoGuard: A near refresh-free robust DRAM design for cryogenic computing," in *Proc. ACM/IEEE 48th Annu. Int. Symp. Comput. Archit. (ISCA)*, Jun. 2021, pp. 637–650.

[13] E. Garzon, Y. Greenblatt, O. Harel, M. Lanuzza, and A. Teman, "Gain-cell embedded DRAM under cryogenic operation—A first study," *IEEE Trans. Very Large Scale Integr. (VLSI) Syst.*, vol. 29, no. 7, pp. 1319–1324, Jul. 2021.

[14] T. Singh *et al.*, "2.1 Zen 2: The AMD 7 nm energy-efficient high-performance x86–64 microprocessor core," in *IEEE Int. Solid-State Circuits Conf. (ISSCC) Dig. Tech. Papers*, Feb. 2020, pp. 42–44.

[15] B. C. Paz *et al.*, "Variability evaluation of 28 nm FD-SOI technology at cryogenic temperatures down to 100 mK for quantum computing," in *Proc. IEEE Symp. VLSI Technol.*, Jun. 2020, pp. 1–2.

[16] J. G. Fossum and V. P. Trivedi, *Fundamentals Ultra-Thin-Body MOSFETs FinFETs*. Cambridge, U.K.: Cambridge Univ. Press, 2013.

[17] M. K. Gowan, L. L. Biro, and D. B. Jackson, "Power considerations in the design of the Alpha 21264 microprocessor," in *Proc. 35th Annu. Conf. Design Automat. Conf. (DAC)*, 1998, pp. 726–731.