This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JXCDC.2021.3130839, IEEE Journal on Exploratory Solid-State Computational Devices and Circuits

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

S.S.Teja Nibhanupudi*, Student Member, IEEE, Siddhartha Raman Sundara Raman*, Mikaël Cassé, Louis Hutin, and Jaydeep P. Kulkarni *Senior Member, IEEE* (*-Equally contributing Authors)

***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 MOSFET's 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 viz. embedded DRAM bitcell and a flip-flop operating at 0.2V, 77K. This paper presents UTBB-SOI device fabrication details and calibrate experimental device characteristics with BSIM compact models. SPICE simulations suggest the feasibility of 3-Transistor gain-cell eDRAM capable of reliably storing three distinct voltage levels (1.5 bits/cell) and exhibiting retention time of the order of 10^4 seconds. 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, UTBB-SOI, Pseudo-Static, eDRAM, Flip-flop, Retention time**

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 ∼77K) 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 sub-threshold characteristics, which lead 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 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-100mV V_T, a wider range of effective work function gate metals are required (below 4.0eV for nMOSFETs and above 5.2eV 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 Box - Silicon on Insulator (UTBB-SOI) transistors [8] [9] have an independent back-gate that can effectively lower the V_T of the transistor. The V_T sensitivity to the back-gate bias (body factor) can be modulated by adjusting thickness of the BOX layer (typically ∼ 10-30nm). 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-100mV 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]. Further, 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 is 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 doesn't experience any contention, unlike a Static Random Access Memory (SRAM) write operation [12]. In

S. S. Teja Nibhanupudi, Siddhartha Raman Sundara Raman, and Jaydeep. P. Kulkarni are with the Department of Electrical and Computer Engineering, University of Texas at Austin, Austin, TX, 78712, USA. Mikaël Cassé, and Louis Hutin are with Université Grenoble Alpes and CEA-Leti Minatec. E-mail: subrahmanya_teja@utexas.edu, jaydeep@austin.utexas.edu

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JXCDC.2021.3130839, IEEE Journal on Exploratory Solid-State Computational Devices and Circuits

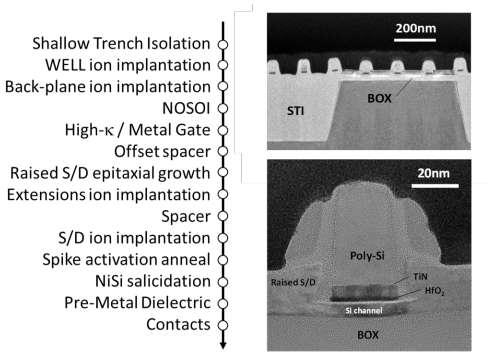


Fig. 1. Left: simplified Front-End-Of-Line process flow. Top right: Transmission Electron Microscopy (TEM) showing transistor gates on isolation (STI) and active areas. Bottom right: TEM cross-sectional close-up view of a transistor after second spacer definition.

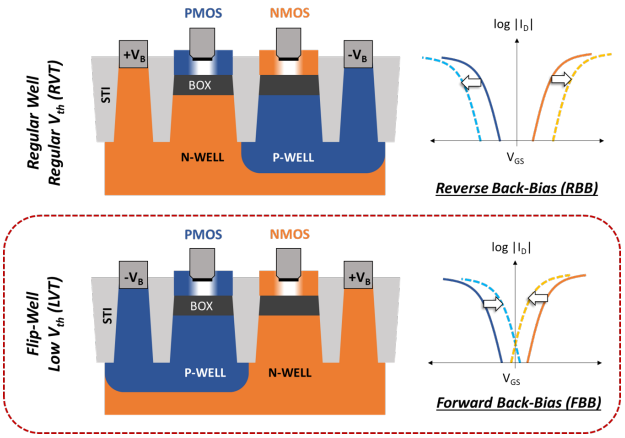


Fig. 2. Top: Regular Well architecture for RVT flavor PMOS and NMOS, with symmetrically applied reverse back-biasing resulting in increased threshold voltages. Bottom: Flip-Well architecture for LVT flavor PMOS and NMOS under symmetrical forward back-bias leading to decreased threshold voltages.

addition, a gain-cell embedded DRAM (eDRAM) [13] with a dedicated read port offers non-destructive read operation, which can enable multi-level 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 paper, 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 sub-threshold 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. 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 tri-state inverter nodes within a flip-flop circuit driven by an active clock signal. In this paper, 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 paper 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 multi-level, high refresh time embedded DRAM 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.

experimentally

II. Device Modeling

A. Experiment

Fully-Depleted SOI N- and PMOSFETs were fabricated in 28nm ground rules with a gate-first high-κ metal gate process on 300mm (100) SOI wafers with a buried oxide (BOX) thickness of 25nm (Fig.1). The undoped silicon channel thickness is 7nm 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 a 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 77K to achieve high performance at a reduced drain voltage while benefiting from the steeper sub-threshold 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 and 77K. 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 77K with the necessary forward back-bias to lower their threshold voltage down to sub-100mV 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 Id-Vgs characteristics of an LVT NMOS device (flipped well) for V_{DS}=0.2V operating at 77K. The dimensions for the device are - gate length L=100nm, channel

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JXCDC.2021.3130839, IEEE Journal on Exploratory Solid-State Computational Devices and Circuits

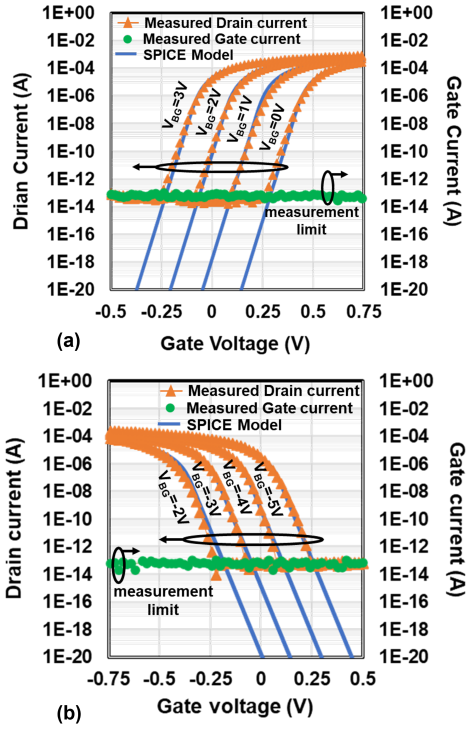


Fig. 3. (a) Measured and SPICE simulated characteristics for various Forward Back Biasing conditions on an LVT NMOS device at 77K with channel width W=2µm, gate length L=100nm, front-gate equivalent oxide thickness EOT=3.7nm. The applied drain-to-source voltage is 200mV. The gate current (right axis) 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 77K with the exact device dimensions as NMOS and source to drain voltage of 0.2V

width W=2µm, silicon layer thickness T_{Si}=7nm, front gate EOT=3.7nm, BOX thickness=25nm. Fig.3(a) also shows the characteristics for different back-gate biases ranging from 0V-3V. The gate current is below the noise floor (50fA) for the entire range of gate voltage biases indicating the extreme low gate leakage. The experimental data is calibrated to the BSIMIMG (version 102.9.2) compact model for SPICE circuit simulations. The compact model device specifications such as channel length (100nm), width (2μm), box thickness (25nm), 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), K0 (lateral non-uniform doping) are optimized to obtain a good fit with experimental data as seen from Fig.3(a). Similarly, Fig.3(b) shows the model calibrated to experimental data for the LVT PMOS device. The calibrated models are used for circuit simulations in section III-IV.

C. TCAD model calibration

As mentioned in sub-section IIA, the lowest current level detected by the measurement setup is limited to 50fA. To reliably estimate the current below this limit, the transistor characteristics are simulated using a multidimensional device

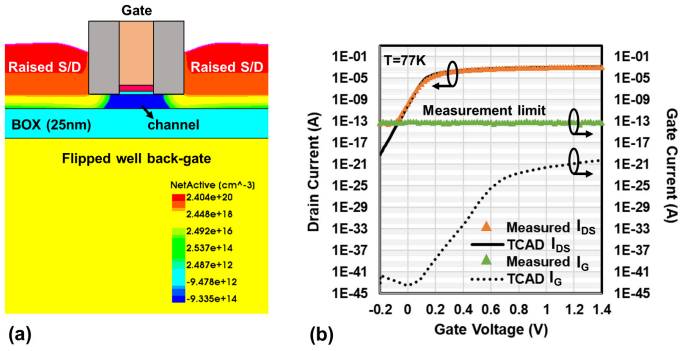


Fig. 4. (a) UTBB-SOI device TCAD cross-section (b) TCAD simulated IdVg characteristics for LVT NMOS transistor at 77K temperature.

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 Philips Unified mobility (PhuMob) model coupled with the Lombardi Thin-layer mobility model to accurately capture the carrier transport inside the transistor at 77K 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 device characteristics as shown in Fig.4(b). The simulated drain to source current closely traces the experimental data above the noise floor of 50fA. The thicker gate oxide (∼3.7nm) limits the gate leakage current below 10^{-20}A across the entire range of gate voltage biases. The simulations also highlight that 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 $V_{PWELL} - V_{NWELL} < 0.6V$. 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 $V_{NWELL} - V_{PWELL} < 6V$. In the case of symmetrical biasing such as described in Fig.2, this would translate to $V_B < 3V$, a constraint that we had to transgress to reach sub-100mV threshold voltages on some devices with more aggressive front-gate EOT (1.5nm). 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

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JXCDC.2021.3130839, IEEE Journal on Exploratory Solid-State Computational Devices and Circuits

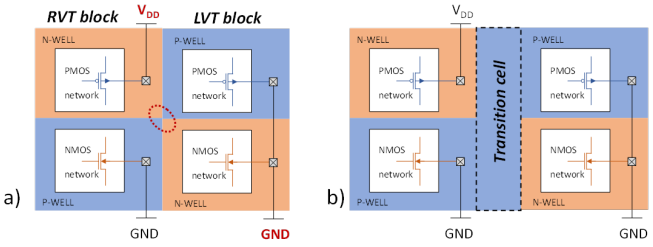


Fig. 5. (a) Direct abutment of an RVT block (Regular Well) with an LVT block (Flip Well), leading to an undesirable singularity point. The indicated body biases correspond to the reference conditions (V_B=0V). The contact between two N-WELL regions biased at different values is particularly problematic.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

wells of opposite polarity by deeper trenches while allowing substrate plugs to remain connected to their wells beneath shallower trenches.

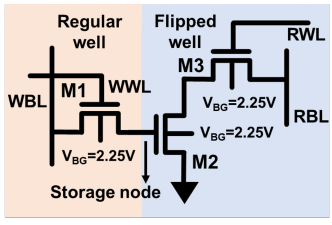
Mixing and matching V_{th} flavors in adjacent blocks may also cause singularity points and well continuity issues, as exemplified in Fig.5 These can be mitigated by the use of transition cells and a Deep N-WELL implantation level. Note that the NMOS-only bit cell studied in the next section (Fig.6)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

III. Multi-level Pseudo-Static Memory Bitcell

The UTBB-SOI transistors operating at 77K have a steep sub-threshold slope (∼25mV/dec), significantly reducing the drain-to-source leakage current. Operating at low voltages (200mV) further reduces electric field induced leakage components, as demonstrated in section IIC. Overall, the reduced leakage current can be leveraged to realize a pseudo-static, high density embedded DRAM (eDRAM) bitcell with a long retention time. This sub-section evaluates the performance of multi-level, 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 two flavors of UTBB-SOI transistors - low-V_T transistor and high-V_T transistor. The low-V_T transistor (V_T=75mV) is implemented using flipped well configuration, and the high-V_T transistor (V_T=150mV) is implemented using regular well configuration holding the back-gate bias at 2.25V. Keeping the back-gate bias constant across all the transistors within the bitcell avoids any integration constraints (diode forward-bias or



	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

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

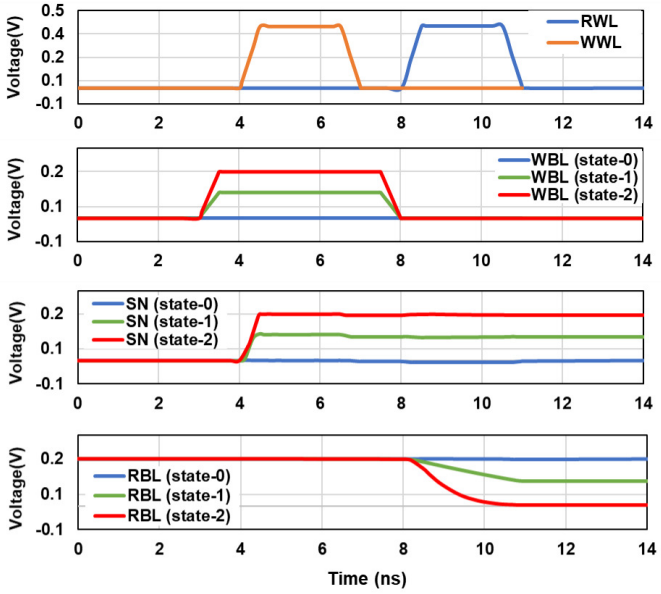


Fig. 7. Timing diagram highlighting the write followed by a read operation for the three-level pseudo-static memory bitcell

reverse-bias) discussed in section IIC. 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 multi-level cell storage on the eDRAM bitcell. Data is 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 3-states/cell - state-1 (0V), state-2 (0.11V) and state-3 (0.2V). During a write operation, the write wordline (WWL) signal is boosted to 0.4V (to overcome the V_T drop of the M1 transistor) with the write bitline (WBL) biased to a desired voltage (0V, 0.11V, or 0.2V). During a retention phase, lowering the WWL signal to -0.2V lowers the leakage current to 10^{-6}fA, which increases the retention time. The read operation begins by pre-charging the read bitline (RBL) to Vcc (0.2V) followed by asserting the read wordline (RWL). The bitline capacitance (assumed to be 30fF in this study) discharges depending on the charge stored at the storage node. The read pulse duration is assumed to be 2ns 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 storage node is programmed to 0V, 0.11V and 0.2V for state-1, state-2 and state-3 respectively. The RBL voltage does not discharge for state-1. The RBL voltage discharges to 0.1V and 0V for state-2 and state-3, respectively. This difference in the

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JXCDC.2021.3130839, IEEE Journal on Exploratory Solid-State Computational Devices and Circuits

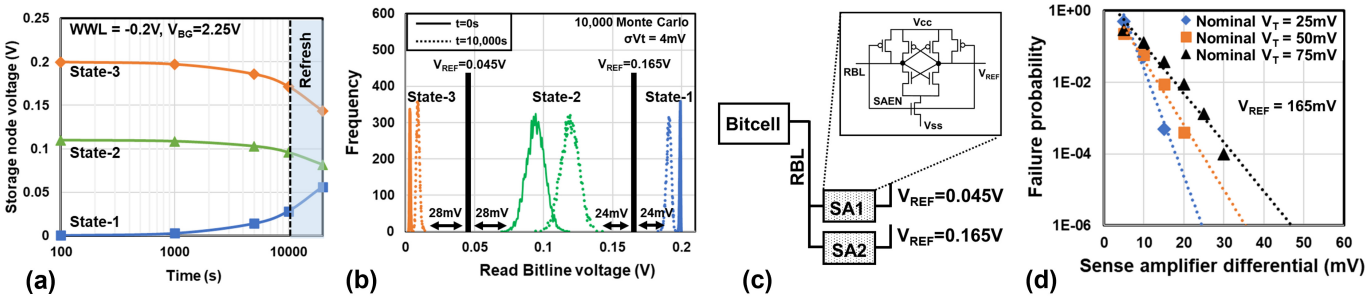


Fig. 8. (a) Storage node variation with time for the three states (b) Read bitline voltage distribution in the presence of variation for t=0 and t=10,000s (c) Sensing scheme to resolve the three bits. Inset shows the schematic of the conventional latch-type sense amplifier. (d) Failure probability of sense amplifier in resolving state-1 and state-2.

bitline voltage is resolved by a sense amplifier to determine the state stored in the bitcell.

B. Bitcell performance

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

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 read bitline (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 < 1mV) whereas state-2 has wider distribution (σ-RBL ∼ 8mV). 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 10,000s after the write operation. The mean of each distribution shifts due to the degradation of voltage levels at the storage node. The RBL voltage distribution for state-2/state-3 shifts to the right as the SN node discharges and reduces the drive strength (over-drive 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 (over-drive voltage) of the M2 transistor. At 10,000s, the RBL separation between state-1 and state-2 reduces to 48mV, and between state-2 and state-3 reduces to 56mV. Therefore the sense amplifier needs to reliably resolve the bits with the input differential voltage of 24mV, 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 level of the states as shown by Fig.8(b). The output of SA1, SA2 at the end of the sensing operation provides information about the stored state. For example, both SA1, SA2 outputs will be '0' for state-1. Similarly the outputs of SA1, SA2 will be '1', '0' for state-2 and '1', '1' for state-3 respectively.

The reference voltages for the sense amplifier are chosen based on the RBL voltage distribution at t=10,000s. This approach ensures that the sense amplifier can reliably resolve the bits under worst-case retention and V_T variation conditions. Further, we also consider the transistor V_T variations within the sense amplifier and capture the impact on sensing margin through 10,000 run Monte-Carlo simulations. Fig.8(d) shows the variation of SA failure probability with increasing SA differential. The reference voltage is held at 165mV for this study since state-1 and state-2 collapse faster towards 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 transistors' 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 1Mb array target. The SA designed using an ultra-low V_T transistor (V_T=25mV) achieves a minimum SA differential of 23mV. This meets the requirements needed to resolve all the three states reliably under worst-case retention (t=10,000s) and variations conditions.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JXCDC.2021.3130839, IEEE Journal on Exploratory Solid-State Computational Devices and Circuits

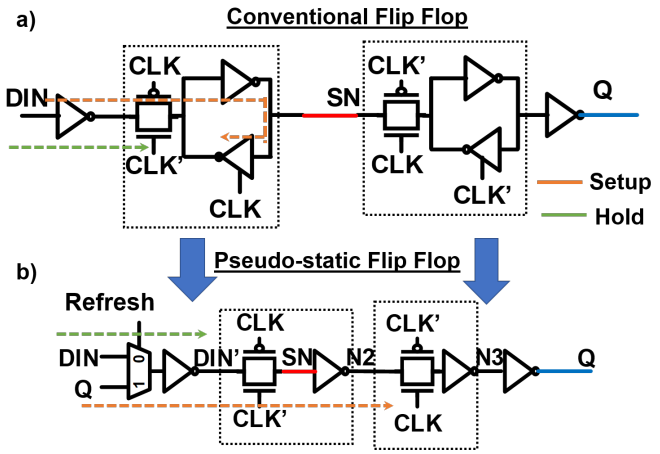


Fig. 9. (a) Conventional Flip-Flop with tri-state cross-coupled inverter storing the data (b) Proposed pseudo-static Flip-Flop with gate capacitance of the inverter as the storage node with datapath for setup and hold analysis marked

IV. Pseudo-static Flip Flop

A. Flip-flop circuit design

As shown in Fig (a), conventional flip-flop designs comprise primary and secondary latches that utilize tri-stated inverters connected in the feedback path, as shown in Fig 9(a). These tri-stated 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 tri-stated inverter, as shown in Fig.9(b). However, a periodic refresh operation is required due to the dynamic nature of the storage node, 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.

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.non-refresh 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 storage node (SN). Here, the gate capacitance of the inverter is utilized as the storage node. The secondary latch is sensitive during the positive half of the clock cycle, and the data stored on the SN node is transferred onto the Q node.

C. Refresh operation

Conventional flip flop designs are static due to cross-coupled inverter pairs that preserve the storage node 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 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

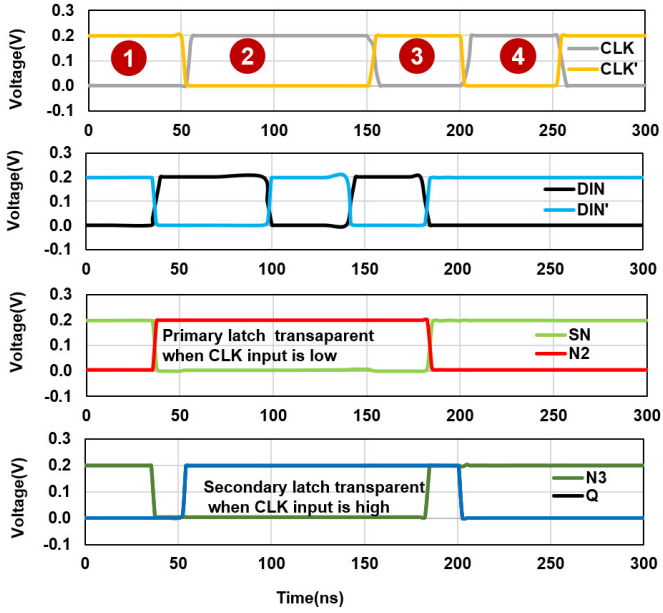


Fig. 10. The timing diagram for a positive edge-triggered flip-flop when in non-refresh mode. The primary latch is transparent during Phase 1 and 3 and the secondary latch during Phase 2 and 4, respectively, with phases marked in red circles.

the refresh time is around 1ms, as shown by 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

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

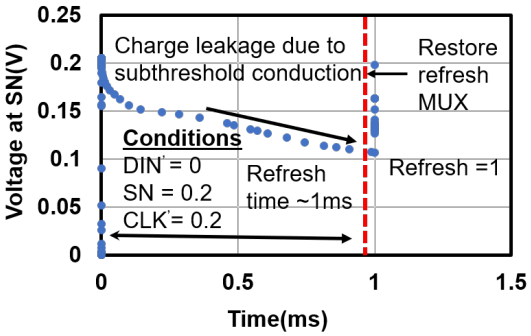


Fig. 11. The voltage at SN drops down to 0.1V in 1ms, assuming the worst-case scenario of DIN' pulled low, CLK' node driven high, and SN is high. This voltage drop is restored using a refresh MUX

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JXCDC.2021.3130839, IEEE Journal on Exploratory Solid-State Computational Devices and Circuits

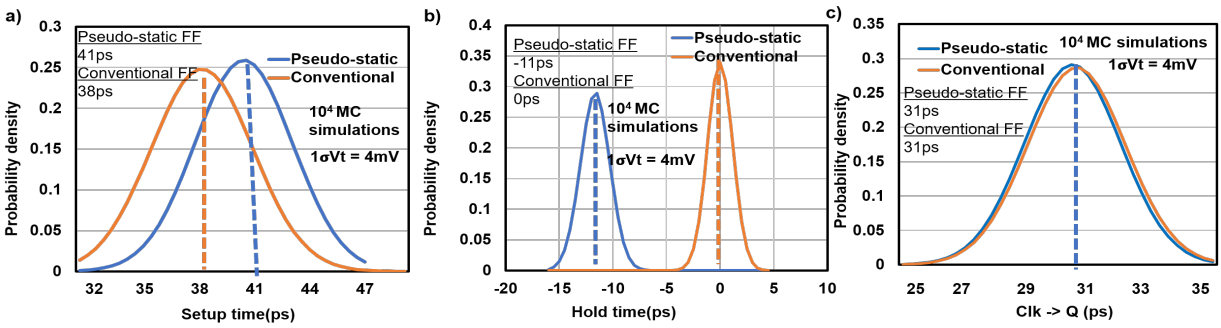


Fig. 12. Statistical simulations showing (a) Setup time (b) Hold time (C) Clk-Q delay distribution for the conventional flip-flop and the pseudo-static flip-flop operating at Vcc=200mV and T=77K

time variation for the pseudo-static flip flop and conventional flip-flop in the presence of process variations by performing 10,000 run Monte-Carlo simulations assuming 1σ-V_T of 4mV. In the pseudo-static flip-flop design, the data must traverse a MUX path (designed using transmission gates) $\rightarrow$ inverter $\rightarrow$ Transmission Gate $\rightarrow$ Inverter before reaching the N2 node. On the contrary, the worst-case datapath in conventional design is Inverter $\rightarrow$ Transmission Gate $\rightarrow$ 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 38ps than 41ps 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 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.

Clk$\rightarrow$Q delay is an essential metric in high-frequency designs which employ several flip-flop stages for computation. Increased Clk$\rightarrow$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$\rightarrow$Q delay. Fig.12 (c) shows that the Clk$\rightarrow$Q delay for the conventional and proposed flip-flops have an almost equal distribution around 31ps, thus having minimal performance difference.

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 multi-bit 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 owing 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 2x (4 transistors connected to the CLK in conventional vs. 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 I 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 a storage node, reducing leakage power by around 20% compared to the 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 tri-stated 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

TABLE I
SUMMARY OF PERFORMANCE, POWER, AREA COMPARISON BETWEEN CONVENTIONAL AND PROPOSED FLIP FLOP (VCC=200MV, T=77K)

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

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JXCDC.2021.3130839, IEEE Journal on Exploratory Solid-State Computational Devices and Circuits

overhead. Overall, the pseudo-static flip-flop reduces the transistor count from 20 to 16 (Table I), 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-100mV threshold voltage capable of operating at an ultra-low voltage of 0.2V. 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. 3T gain-cell embedded DRAM having a considerable retention time of 10,000 seconds 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 storage node 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.

VI. Acknowledgments

The authors would like to thank Prof. Tsu-Jae King Liu from the University of California, Berkeley, for helpful discussions. The authors 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 *2021 IEEE Custom Integrated Circuits Conference (CICC)*, pp. 1–2, 2021.

[2] I. Byun, D. Min, G.-h. Lee, S. Na, and J. Kim, "Cryocore: A fast and dense processor architecture for cryogenic computing," in *2020 ACM/IEEE 47th Annual International Symposium on Computer Architecture (ISCA)*, pp. 335–348, 2020.

[3] J. C. Bardin, E. Jeffrey, E. Lucero, T. Huang, S. Das, D. T. Sank, O. Naaman, A. E. Megrant, R. Barends, T. White, *et al.*, "Design and characterization of a 28-nm bulk-cmos cryogenic quantum controller dissipating less than 2 mw at 3 k," *IEEE Journal of Solid-State Circuits*, vol. 54, no. 11, pp. 3043–3060, 2019.

[7] S. Hung, "Multi-vt engineering and gate performance control for advanced finfet architecture," *Short Course 1: Boosting Performance,Ensuring Reliability, Managing Variation in sub-5nm CMOS, IEDM.*, 2017.

[4] F. Sebastiano, H. A. Homulle, J. P. van Dijk, R. M. Incandela, B. Patra, M. Mehrpoo, M. Babaie, A. Vladimirescu, and E. Charbon, "Cryogenic cmos interfaces for quantum devices," in *2017 7th IEEE International Workshop on Advances in Sensors and Interfaces (IWASI)*, pp. 59–62, IEEE, 2017.

[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 *2017 47th European Solid-State Device Research Conference (ESSDERC)*, pp. 62–65, 2017.

[6] C.-H. Lin, R. Kambhampati, R. J. Miller, T. B. Hook, A. Bryant, W. Haensch, P. Oldiges, I. Lauer, T. Yamashita, V. Basker, T. Standaert, K. Rim, E. Leobandung, H. Bu, and M. Khare, "Channel doping impact on finfets for 22nm and beyond," in *2012 Symposium on VLSI Technology (VLSIT)*, pp. 15–16, 2012.

[8] T. A. Karatsori, A. Tsormpatzoglou, C. G. Theodorou, E. G. Ioannidis, S. Haendler, N. Planes, G. Ghibaudo, and C. A. Dimitriadis, "Analytical compact model for lightly doped nanoscale ultrathin-body and box soi mosfets with back-gate control," *IEEE Transactions on Electron Devices*, vol. 62, no. 10, pp. 3117–3124, 2015.

[9] M. Cassé and G. Ghibaudo, "Low temperature characterization and modeling of fdsoi transistors for cryo cmos applications," 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 Transactions on Nanotechnology*, vol. 17, no. 1, pp. 36–40, 2017.

[11] O. Weber, O. Faynot, F. Andrieu, C. Buj-Dufournet, F. Allain, P. Scheiblin, J. Foucher, N. Daval, D. Lafond, L. Tosti, *et al.*, "High immunity to threshold voltage variability in undoped ultra-thin fdsoi mosfets and its physical understanding," in *2008 IEEE International Electron Devices Meeting*, pp. 1–4, IEEE, 2008.

[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 *2021 ACM/IEEE 48th Annual International Symposium on Computer Architecture (ISCA)*, pp. 637–650, IEEE, 2021.

[13] E. Garzón, Y. Greenblatt, O. Harel, M. Lanuzza, and A. Teman, "Gain-cell embedded dram under cryogenic operation—a first study," *IEEE Transactions on Very Large Scale Integration (VLSI) Systems*, vol. 29, no. 7, pp. 1319–1324, 2021.

[14] T. Singh, S. Rangarajan, D. John, R. Schreiber, S. Oliver, R. Seahra, and A. Schaefer, "2.1 zen 2: The amd 7nm energy-efficient high-performance x86-64 microprocessor core," in *2020 IEEE International Solid- State Circuits Conference - (ISSCC)*, pp. 42–44, 2020.

[15] B. C. Paz, L. Le Guevel, M. Cassé, G. Billiot, G. Pillonnet, A. Jansen, R. Maurand, S. Haendler, A. Juge, E. Vincent, *et al.*, "Variability evaluation of 28nm fd-soi technology at cryogenic temperatures down to 100mk for quantum computing," in *2020 IEEE Symposium on VLSI Technology*, pp. 1–2, IEEE, 2020.

[16] J. G. Fossum and V. P. Trivedi, *Fundamentals of Ultra-thin-body MOSFETs and FinFETs*. Cambridge University Press, 2013.

[17] M. Gowan, L. Biro, and D. Jackson, "Power considerations in the design of the alpha 21264 microprocessor," in *Proceedings 1998 Design and Automation Conference. 35th DAC. (Cat. No.98CH36175)*, pp. 726–731, 1998.