Received 17 August 2023; revised 29 September 2023 and 26 October 2023; accepted 1 November 2023.
Date of publication 7 November 2023; date of current version 5 December 2023.

Digital Object Identifier 10.1109/JXCDC.2023.3330767

Design Exploration of 14 nm FinFET for Energy-Efficient Cryogenic Computing

AMOL D. GAIDHANE[1] (Member, IEEE), RAKSHITH SALIGRAM[2], WRIDDHI CHAKRABORTY[3], SUMAN DATTA[2] (Fellow, IEEE), ARIJIT RAYCHOWDHURY[2] (Fellow, IEEE), and YU CAO[4] (Fellow, IEEE)

[1]School of ECEE, Arizona State University, Tempe, AZ 85287 USA
[2]Department of ECE, Georgia Institute of Technology, Atlanta, GA 30332 USA
[3]Department of Electrical Engineering, University of Notre Dame, Notre Dame, IN 46556 USA
[4]Department of ECE, University of Minnesota, Minneapolis, MN 55455 USA

CORRESPONDING AUTHOR: A. D. GAIDHANE (agaidhan@asu.edu)

This work was supported in part by the Defense Advanced Research Projects Agency (DARPA), USA, through the Low Temperature Logic Technology (LTLT) Program.

ABSTRACT Cryogenic operation of CMOS transistors (i.e., cryo-CMOS) effectively brings an ultrasteep subthreshold slope (SS) and ultralow leakage, enabling high energy efficiency with appropriate tuning of threshold voltage and supply voltage. On the other hand, cryo-CMOS suffers from elevated sensitivity to process and voltage variations. To facilitate early-stage design exploration, we develop predictive BSIM-CMG model cards, which are calibrated with 14 nm TCAD simulation and our experimental FinFET data from 300 to 77 K. These models are scalable with temperatures from 300 K down to 77 K, device engineering and variations. Based on them, we benchmark various circuit examples to illustrate the tremendous potential of cryo-CMOS for energy-efficient computing, in the presence of process variations. For logic circuits, such as a canonical critical path, more than 15× reduction in total energy consumption is demonstrated at 77 K for the iso–Delay condition, compared to the operation at the room temperature (RT). The presence of variations only has a marginal impact on energy efficiency, after threshold voltage and supply voltage are adaptively increased. For static noise margin (SNM), it is consistently improved at 77 K. However, the impact of variations on SNM is much more pronounced than that on logic circuits.

INDEX TERMS Cryogenic temperatures, energy-efficient computing, FinFET, predictive modeling, variations.

I. INTRODUCTION

The Dennard's law of scaling the dimension of CMOS devices and supply voltage (V_{DD}) at the same rate stopped few decades ago due to the Boltzmann tyranny effect [1]. Scaling of threshold voltage (V_{TH}) also stopped in recent technology nodes due to the concern of exacerbated leakage current. In fact, even at constant V_{DD} and V_{TH}, leakage current keeps increasing over generations as the subthreshold slope (SS) becomes larger due to short-channel effects. For large-scale circuits, continuous scaling of CMOS devices without the reduction in supply voltage and threshold voltage significantly increases active power consumption and the density of heat dissipation.

In this context, temperature scaling has been proposed as an alternative approach toward low leakage and high performance in scaled CMOS devices [2]. The leading principle is that the SS linearly decreases with the temperature as given in (1) and thus, the leakage current is exponentially reduced

$$\mathrm{SS}(T) = \left(\frac{\partial \log(I_D)}{\partial V_{\mathrm{GS}}}\right)^{-1} = \ln(10)\frac{nkT}{q} \tag{1}$$

where, n is the SS factor and kT/q is the thermal voltage. For instance, when we reduce the temperature from the room temperature (RT) (300 K) to the cryogenic temperature of 77 K, SS is reduced from 70 to 18 mV/dec for a 14 nm FinFET technology, bringing more than 1000× reduction in the leakage current (I_{OFF}), as shown in Fig. 1(a). Such a dramatic reduction in I_{OFF} offers the engineering space to modulate V_{DD} and V_{TH} in cryo-CMOS design, in order to balance the performance need and the reduction of power consumption. As a demonstration, Fig. 1(a) illustrates an approach to scale down V_{TH} at 77 K to reach the same I_{OFF} as that under the

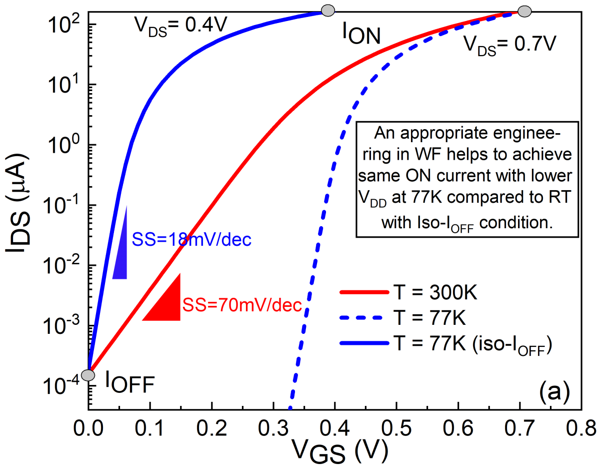


FIGURE 1. I_{DS}–V_{GS} **characteristic a 14 nm n-FinFET device at the RT and at the cryogenic temperature (77 K). For 77 K, appropriate scaling in** V_{DD} **and** V_{TH} **helps achieve the same ON current for the iso–**I_{OFF} **condition at lower supply voltages.**

RT (i.e., iso–I_{OFF}); meanwhile, V_{DD} can be tuned down in proportion to keep the same ON current. During this practice, cryo-CMOS design is able to achieve the same switching speed and leakage as those under RT, and much lower active power consumption which is proportional to V_{DD}^2 [2], [3].

Various studies have been conducted to explore the design benefit of cryo-CMOS design. An experimental study [4] at cryogenic temperatures on the 10 nm FinFET-based benchmark circuits demonstrated more than 4× power reduction at the same logic speed or 50% increase in logic speed at the same power condition. The steeper subthreshold characteristics at the cryogenic temperature also improves static noise margin (SNM) of the inverter and the SRAM cell at lower V_{DD} compared to the RT, as demonstrated in [5] and [4]. Further, the power, delay, and reliability of 5 nm FinFET SRAM circuits has been studied at deep cryogenic temperatures [6]. A recent theoretical study demonstrated 4× improvement in performance-per-watt at the processor level (Arm Cortex-A53) at cryogenic temperatures [7]. The improvement in performance per watt becomes even better at the system level, where recent study demonstrated 12× at the CPU level and 16× at the system level [8], through circuit and architecture cooptimization. Furthermore, thermal conductance of silicon increases under cryogenic temperatures, improving the heat dissipation and reducing self-heating effects toward higher integration density [4], [7], [9].

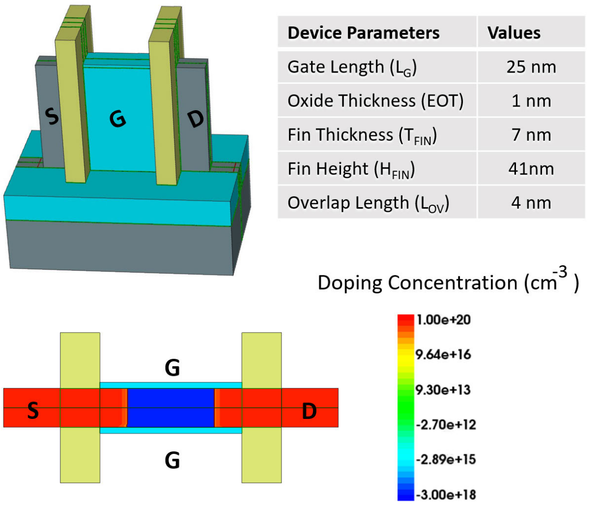
On the other hand, due to the steeper subthreshold characteristics at cryotemperatures, the CMOS device becomes highly sensitive to bias voltages and process variations [10]. The OFF current of CMOS device varies 10 000× at cryotemperatures compared to 10× at the RT for the same amount of global variations in metal gate work-function [11]. The similar sensitivity in the OFF current is demonstrated using our model card developed in Section II. Therefore, it is critical to consider such elevated sensitivity in joint device-design optimization for cryo-CMOS, in order to maximize the benefits.

To facilitate such joint design exploration, we develop predictive model cards for cryogenic FinFETs in this work and calibrate them with available data at 14 nm. Leveraging these predictive models, we benchmark the benefit of low power and the challenge of robustness in various cryogenic circuits. Our major contributions are.

1) *Predictive Model Cards for Cryogenic Operations:* We develop predictive model cards which are based on cryogenic device physics and calibrated with 14 nm FinFET measurement data and TCAD simulations. These predictive model cards are scalable with temperatures, and also scalable with device engineering (e.g., V_{DD} and V_{TH} tuning) and major process variations. It is worth noting that [6] has created model cards for the 5 nm FinFET technology node. However, our work stands out in that we have expanded the flexibility of these model cards to predict the behavior of CMOS logic circuits at cryogenic temperatures. This is accomplished by fine-tuning V_{TH} and V_{DD}, specifically for the applications requiring extremely low power.
2) *Energy Analysis for Cryo–CMOS Design:* By reducing V_{TH} down to 75 mV and V_{DD} to 0.16 V, we demonstrate 11× − 15× reduction in energy consumption at 77 K compared to that under the RT, at the same logic switching speed (i.e., iso−Delay). The exact gain depends on the switching activity and the amount of variations. Although process variations increase the leakage power, their impact can be mitigated by adaptively tuning up V_{DD} and V_{TH}. In [8], significant energy enhancements of 12× to 16× are demonstrated at both the CPU and system levels through cooptimization of logic and memory circuits. Additionally, we observe energy improvements in logic circuits within the same range, up to 15×. However, our work highlights that achieving such substantial energy improvements comes with a trade-off due to the high sensitivity to variations in V_{TH} or V_{DD}.
3) *Circuit Robustness Analysis:* Circuit robustness is evaluated by the SNM of inverters. SNM of the inverter at the iso–I_{OFF} condition degrades significantly at same V_{DD} as the RT. However, proper tuning of V_{DD} and V_{TH} is able to boost SNM at lower V_{DD}. Yet, SNM experiences a larger amount of variability under process variations at cryotemperatures.

II. MODEL DEVELOPMENT AND VALIDATION

To realize the full benefit of the cryogenic operation, appropriate scaling of V_{TH} and V_{DD} is necessary, as presented in Fig. 1(a). While device researchers are actively investigating the fabrication techniques to reduce V_{TH} [e.g., negative gate capacitance and gate work function (WF)], compact cryo-CMOS models must be scalable with such device engineering and correctly predict device characteristics for early



Device Parameters	Values
Gate Length (L_G)	25 nm
Oxide Thickness (EOT)	1 nm
Fin Thickness (T_{FIN})	7 nm
Fin Height (H_{FIN})	41nm
Overlap Length (L_{OV})	4 nm

FIGURE 2. **Three-dimensional TCAD device structure, device parameters, and its cross-sectional view of a n-FinFET device.**

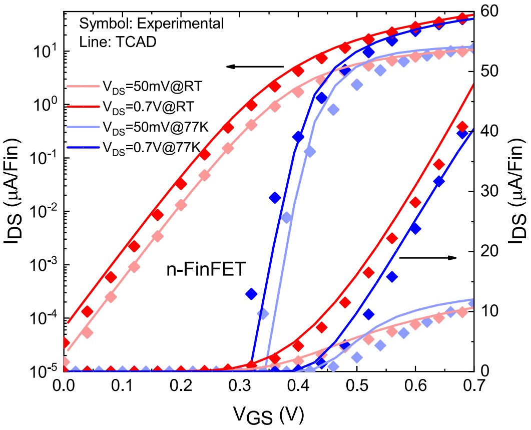


FIGURE 3. **Calibration of the TCAD model with our 14 nm experimental n-FinFET data at both RT and 77 K.**

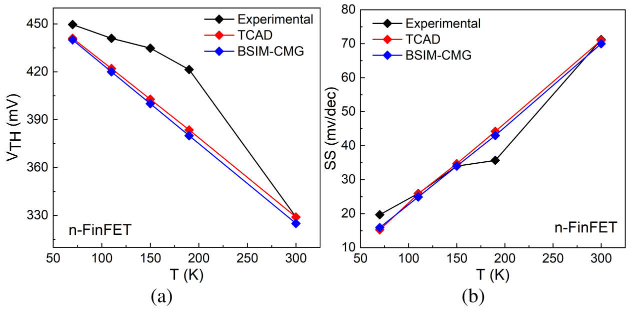


FIGURE 4. **Validation of the cryo-CMOS model for (a) threshold voltage (V_{TH}) at different temperatures and (b) SS at different temperatures.**

design practice. We rely on cryogenic device physics, TCAD simulation, and measured FinFET data at 14 nm to develop this set of predictive models in this section.

A. TCAD SETUP

A 3-D TCAD device structure of n-FinFET and its cross-sectional view along with the doping concentration are depicted in Fig. 2. Fig. 2 also lists the nominal device parameters for the 14 nm FinFET device to build the TCAD model in Sentaurus 3-D TCAD tool [13]. We utilize drift-diffusion transport along with the Fermi–Dirac statistics for an accurate analysis of the device. In addition, we adopt the band-gap narrowing model, the Philips unified mobility model, SRH recombination, and velocity saturation models to capture the accurate behavior of the device. In order to accurately represent the linear trends in parameters such as SS, $V_{\rm TH}$, mobility, and velocity saturation, we have fine-tuned the temperature-dependent parameters of these models. Specifically, within the subthreshold regime, the linear relationships in SS and $V_{\rm TH}$ are inherently captured through the thermal voltage. In the linear regime, we have adjusted the temperature-dependent parameters of the Caughey-Thomas model to account for velocity saturation effects and utilized the Philips Unified Mobility Model to address mobility-related effects. To incorporate the nonlinearity in subthreshold regime, we can add the band-tail model and temperature dependency in interface traps to capture nonlinearity in the subthreshold regime to work at deep cryogenic temperatures. We calibrate our TCAD model with measurement data from the 14 nm n-FinFET at both RT and 77 K by adjusting the parameters of default geometry and physical models for both low and high $V_{\rm DS}$, as shown in Fig. 3.

B. CRYO-CMOS MODELING

For compact modeling of cryogenic FinFET devices, the SS, threshold voltage ($V_{\rm TH}$), average effective mobility ($\mu_{\rm eff}$), and saturation velocity ($V_{\rm SAT}$) are the primary parameters that have a significant dependency on the temperature. The change in threshold voltage and SS are almost linear when the operating temperature scales down from RT to 50 K [14]. Such a linear behavior can be embedded into the parameters in the temperature-dependent model in BSIM-CMG [15].

As presented in (1), we use a linear model for SS to incorporate its temperature dependency. The temperature dependency in threshold voltage [$V_{\rm TH}(T)$] of a MOSFET transistor can be modeled as [15]

$$V_{\rm TH}(T) = V_{\rm TH}(T_0) + \alpha_T\left(\frac{T}{T_0} - 1\right) \tag{2}$$

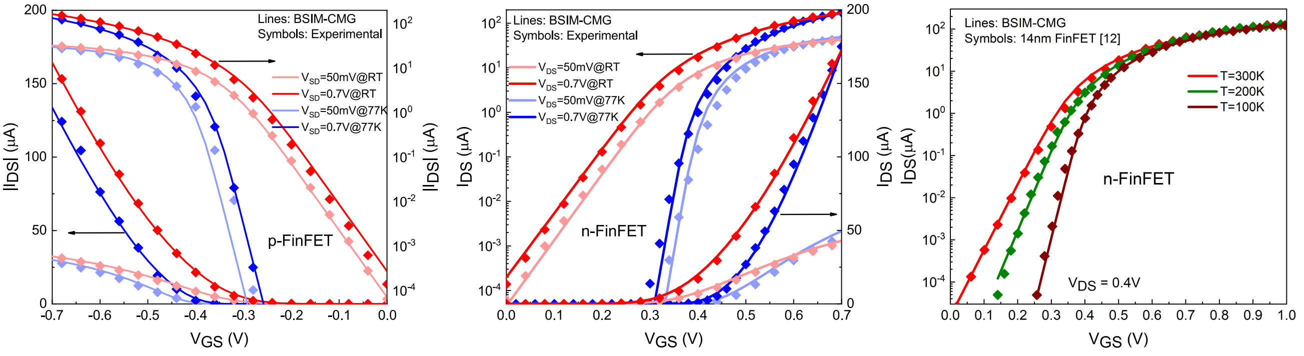
where, α_T is a negative temperature coefficient of the threshold voltage. The average effective mobility in bulk MOSFETs almost linearly increases with the decrease in the operating temperature [16], which can be modeled as

$$\mu_{\rm eff}(T) = \mu_{\rm eff}(T_0)(T/T_0)^{\alpha} \tag{3}$$

where, α is a temperature coefficient of the mobility. Further, the temperature dependency in saturation velocity of the carrier can be modeled as [17]

$$V_{\rm SAT}(T) = \frac{V_{\rm SAT}(T_0)}{(1-A)+A(T/T_0)} \tag{4}$$

where, A is a temperature coefficient of saturation velocity.



(a) (b) (c)

FIGURE 5. Validation of the cryo-CMOS model with our experimental (a) 14 nm p-FinFET and (b) 14 nm n-FinFET at both RT and 77 K for V_{DS} = 50 mV and V_{DS} = 0.7 V cases. (c) Validation of 14 nm IBM FinFET data [12] using a single model card at various temperatures for V_{DS} = 0.4 V.

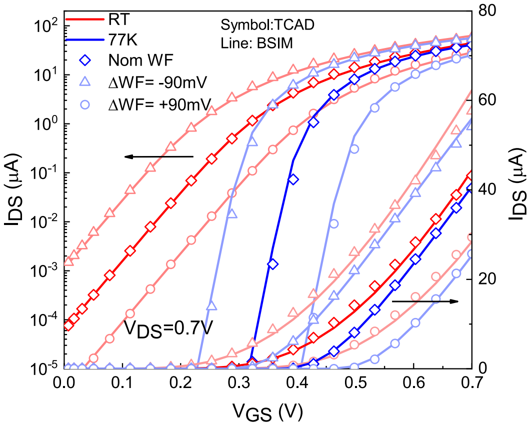


FIGURE 6. Validation of cryo-CMOS models with TCAD, under 2% variation of metal gate WF, i.e., ±90 mV.

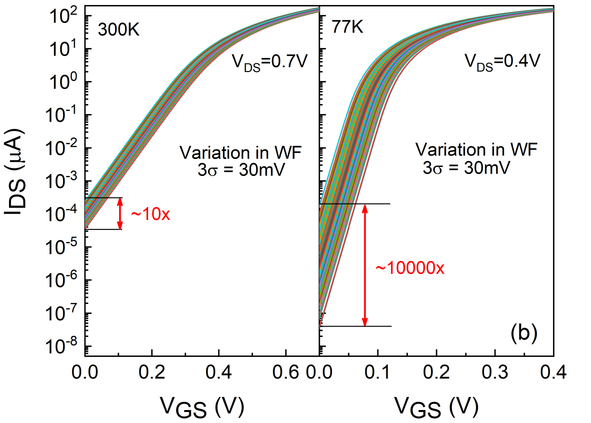


FIGURE 7. Monte Carlo simulations for 14 nm n-FinFET at the RT and at 77 K under gate WF variations (3σ = 30 mV). I_{OFF} is much more sensitive to the variations at 77 K compared to RT.

Such approximately linear behaviors are used in our predictive model to benchmark the design with cryogenic CMOS devices. However, the trends become nonlinear when the operating temperature further scales down below 50 K due to the band tail, interface traps, and other effects [14], [18].

C. MODEL CARDS DEVELOPMENT AND VALIDATION

By embedding these equations into BSIM-CMG and tuning corresponding model parameters, we prepare a set of 14 nm FinFET model cards that are continuous with the change of temperatures.

The comparison of threshold voltage (V_{TH}) and SS are demonstrated in Fig. 4(a) and (b), respectively. Note that, we calculate the threshold voltage using the constant current method, i.e., 100 nA × (W/L). For compact modeling in the range of temperatures down to 77 K, we can ignore the secondary effects, such as interface traps and the band tail, which could produce nonlinearity in V_{TH} and SS with the temperature. Depending on the specific fabrication technology, these effects may be present in the experimental results and thus, nonlinearity is observed below 200 K, as shown in Fig. 4(a) and (b). The nonlinearity highly differ from one technology to other. Therefore, we follow the linear behavior as a generic model for design benchmarking, which is also consistent with TCAD simulations.

Fig. 5(a) and (b) presents the validation of full IV characteristics for both n/p-FinFET at the RT and at cryogenic temperatures, i.e., 77 K, for our 14 nm FinFET experimental data. The model is further validated for 14 nm IBM FinFET data [12] at different temperatures, as shown in Fig. 5(c). We adjust the parameters in the temperature-dependent model parameters, such as SS, threshold voltage, effective mobility, and velocity saturation effect, such that both ON and OFF characteristics of the device are captured accurately at different temperatures using a single set of model cards. The obtained range for $U_{\text{eff}}(T)$ varies from 4×10^2 to 2.5×10^3 cm^2/Vs when we reduce operating temperatures. Further, the range of $V_{\text{SAT}}(T)$ varies between 1×10^7 and 1.5×10^7 cm/s.

We further validate our generic cryo-CMOS model cards under excessive variations in metal gate workfunction (i.e., 90 mV), as shown in Fig. 6. The model prediction well matches TCAD simulations at both RT and 77 K.

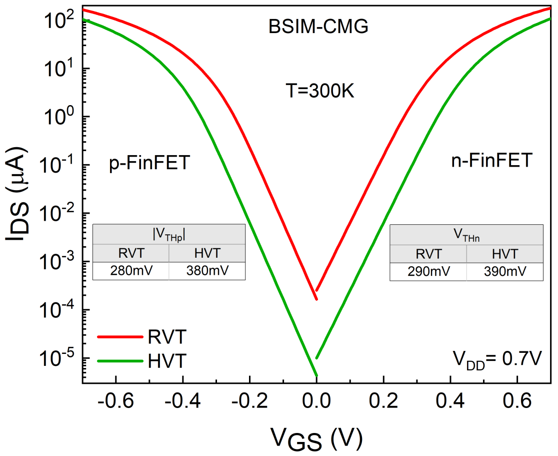


FIGURE 8. I_{DS}–V_{GS} **characteristics of n-FinFET and p-FinFET for both regular** V_{TH} **(RVT) and high** V_{TH} **(HVT) in our predictive model cards.**

Using these calibrated cryo-CMOS model cards, we demonstrate the sensitivity of I_{OFF} to process variations, using Monte Carlo SPICE simulations. Fig. 7 shows the I_{DS}–V_{GS} characteristic of n-FinFET at the RT and 77 K at $V_{\mathrm{DD}} = 0.7$ V and $V_{\mathrm{DD}} = 0.4$ V, respectively. The sensitivity to the variation at 77 K is dramatically elevated compared to the RT case. With 3σ variation of V_{TH} at 30 mV, I_{OFF} suffers 10 000× variability, compared to 10× variability at the RT. The result is consistent with that presented in [11].

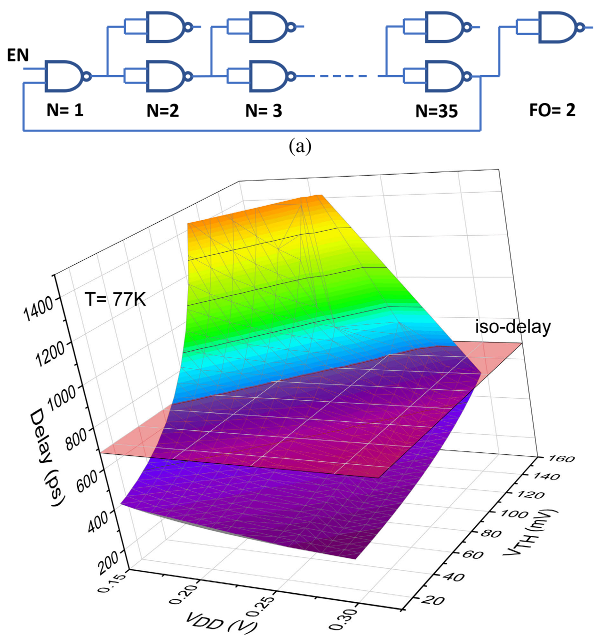
Finally, we adjust the value of V_{TH} in the generic cryo-CMOS for a low power design. We increase the threshold voltage of the n-FinFET ($V_{\mathrm{TH_n}}$) and p-FinFET ($V_{\mathrm{TH_p}}$) at the RT from 290 to 390 mV and from 280 to 380 mV, respectively, to keep the OFF current in the order of nA/μm which is desirable for HVT transistors for low-power logic applications. The drain current characteristics for both regular V_{TH} (RVT) and high V_{TH} (HVT) at the RT are shown in Fig. 8.

III. CIRCUIT PERFORMANCE AT ISO–DELAY CONDITION

Leveraging the newly developed cryo-CMOS model, we explore low-power design at 77 K, under the delay constraint (i.e., iso–Delay). Our results demonstrate that with appropriate device engineering, as well as V_{DD} and V_{TH} scaling, cryogenic design effectively reduces the energy consumption by more than 10× in representative circuit examples.

A. ENERGY ANALYSIS ON ROs

We perform the energy analysis on a 35-stage NAND RO with fan-out (FO) of 2. The schematic of the benchmark circuit is shown in Fig. 9(a). In the NAND gate, the number of fins (NFINs) are six and four for n-FinFET and p-FinFET, respectively. At the RT, the period of this 35-stage RO at RT is around 680 ps. Next, we tune V_{TH} of n/p-FinFETs (i.e., through WF) at 77 K for each supply voltage to achieve the same delay as that of the RT case (i.e., 680 ps) as shown in Fig. 9(b). The values of V_{TH} at each V_{DD} for the iso–Delay condition are listed in the table in Fig. 9.



(a)

(b)

$V_{\mathrm{DD}}(V)$	$V_{\mathrm{THn}}(mV)$	$\lvert V_{\mathrm{THp}}\rvert(mV)$
0.15	67	69
0.16	75	77
0.17	80	82
0.18	90	92
0.19	95	97
0.2	100	102
0.21	109	111
0.22	119	121
0.25	139	140
0.26	149	150

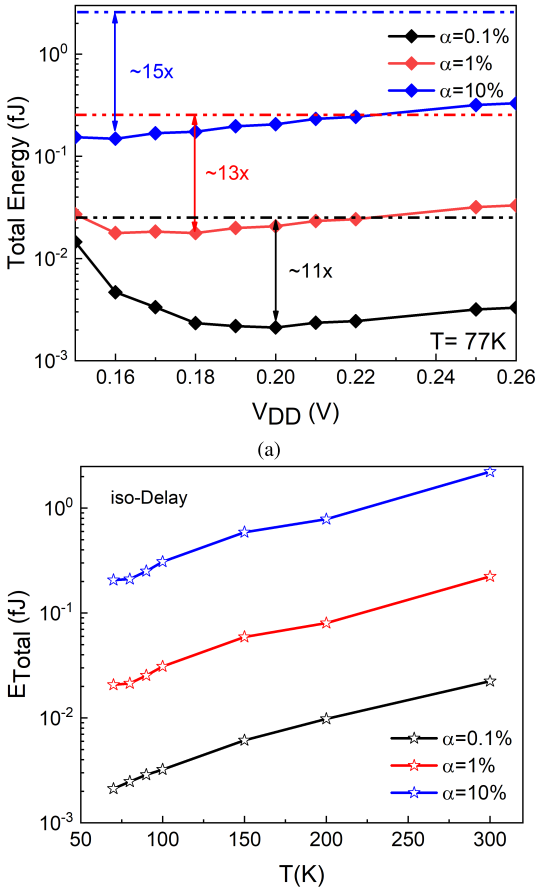
(c)

FIGURE 9. (a) Circuit schematic of a 35-stage NAND-based ring oscillator (RO) with FO = 2. (b) RO delay at 77 K for different V_{TH} **at every** V_{DD} **values. (c) Combinations of** V_{TH} **and** V_{DD} **values at 77 K for iso–delay condition.**

Based on this set of simulations in Fig. 9, we analyze total energy consumption of the RO under different switching activity factors under the iso–Delay condition. The total energy (E_{Total}) of the RO is calculated using

$$E_{\mathrm{Total}} = \alpha E_{\mathrm{Active}} + (1-\alpha)E_{\mathrm{Leak}} \tag{5}$$

where, α is the switching activity factor, E_{Active} is the energy consumption during switching, and E_{Leak} is the energy consumption at the idle state. The total energy for each V_{DD} for different switching activity factors are demonstrated in Fig. 10(a). For low switching activities, e.g., $\alpha = 0.1\%$, the minimum total energy is observed at $V_{\mathrm{DD}} = 0.2$ V, where 11× energy reduction is observed at 77 K compared to the RT case. For an intermediate switching activity, e.g., $\alpha = 1\%$, the minimum total energy is observed at $V_{\mathrm{DD}} = 0.18$ V, where 13× energy reduction is achieved. Furthermore, energy reduction improves to 15× for a higher switching activity, e.g., $\alpha = 10\%$.



(a)

(b)

Temp (K)	$V_{\mathrm{THn}}(mV)$	$\vert V_{\mathrm{THp}}\vert(mV)$	$V_{\mathrm{DD}}(V)$
70	100	102	0.2
80	101	103	0.2
90	113	114	0.22
100	134	133	0.25
150	193	185	0.35
200	209	203	0.4
300	390	380	0.7

(c)

FIGURE 10. (a) Total energy of the RO at 77 K to achieve the minimum energy consumption. The optimal V_{DD} shifts to a lower value for a higher switching activity factor under the iso–Delay condition. The dotted lines are the reference energy values which represent the total energy for different switching factors at RT without scaling V_{DD}, i.e., at V_{DD} = 0.7 V. (b) Total energy at different temperatures under the iso–Delay condition. (c) Optimal V_{TH} (for both n/p-FinFETs) and V_{DD} for different temperatures under the iso–Delay condition, for α = 0.1%.

This trend is consistent with our understanding. When the switching activity is high, the switching energy is more important than the leakage energy. In addition, the active energy has a quadratic dependence on V_{DD} while the leakage energy only has a linear dependence on V_{DD}. Therefore, V_{DD} reduction is more effective to save the total energy at higher switching activities, as observed in Fig. 10(a).

Fig. 10(b) further presents the energy analysis at different temperatures under the iso–Delay condition. When we reduce the operating temperature, the same delay can be achieved with underlinelower V_{DD} and thus, the lower energy consumption. Under iso–Delay, E_{Total} almost linearly decreases with the operating temperature. In the case of $\alpha = 0.1\%$, Fig. 10(c) lists the combination of V_{DD} and V_{TH} to achieve the minimum total energy.

B. ENERGY ANALYSIS UNDER VARIATIONS

To examine the elevated impact of process variations on cryogenic design, we perform Monte Carlo simulations on 1000 samples of the 35-stage ring-oscillator at each temperature. We keep the nominal values of V_{DD} and V_{TH} as mentioned in Fig. 10(c), and inject variations in metal gate WF, which induces random V_{TH} variations of $3\sigma = 10$, 20 or 30 mV for each transistor in the RO. As a result, we observe a significant amount of variations in the leakage energy at cryogenic temperatures as shown in Fig. 11(a). Furthermore, V_{TH} variations at cryogenic temperatures has a pronounced impact on I_{ON}, especially for low V_{DD} operations. Therefore, E_{Active} and RO delay also experience an increasing amount of variations at lower temperatures, as shown in Fig. 11(b) and (c).

Next, we analyze the variations at 77 K under voltage scaling. Three different supply voltages are sampled, i.e., 0.16 V (best case for $\alpha = 10\%$), 0.2 V (best case for $\alpha = 0.1\%$), and 0.25 V ($\alpha = 0.1\%$) as demonstrated in Fig. 12. In Fig. 12(a), we illustrate the variations in E_{Leak} across different V_{DD}, taking into account various levels of V_{TH} variations. It is important to note that we have fine-tuned the mean value of V_{TH} for each V_{DD} setting to achieve an iso–delay condition, as detailed in the table presented in Fig. 8(c). In an ideal scenario, we would expect the variation in E_{Leak} to decrease as V_{DD} decreases because it is directly proportional to the variation in I_{OFF}, which in turn decreases with lower V_{DD}. This is primarily due to the increase in the mean value of I_{OFF} resulting from the reduction in V_{TH} as V_{DD} decreases. Therefore, it is reasonable to observe lower variation in E_{Leak} at $V_{DD} = 0.16$ V compared to $V_{DD} = 0.2$ V. Similarly, we would expect lower energy variation at $V_{DD} = 0.2$ V compared to $V_{DD} = 0.25$ V. However, it is essential to consider that the variation in I_{OFF} also depends on the subthreshold swing, which is influenced by the drain bias. As a result, at higher V_{DD} levels (e.g., $V_{DD} = 0.25$ V), the subthreshold swing increases, leading to a reduction in the variation in I_{OFF}. Consequently, we observe a lower variation in E_{Leak} than at $V_{DD} = 0.2$ V. Furthermore, at lower V_{DD} values, variations in V_{TH} have a more pronounced impact on I_{ON}. Therefore, we observe higher variation in E_{Active} at lower V_{DD} as shown in Fig. 12(b). It is worth noting that even when significant variations exist in E_{Leak}, they do not significantly affect the variations in E_{Total}, as demonstrated in Fig. 12(c). This is because the primary contribution to E_{Total} comes from E_{Active}. Consequently, the variations in E_{Active} and E_{Total} are essentially identical, as evident from Fig. 12(b) and (c).

We can increase V_{DD}, e.g., from 0.2 to 0.25 V, to reduce the variations, as shown in Fig. 12(b) and (c). However, this will degrade the energy saving at 77 K to 7×, compared to the case with less variations at 11×, as shown in Fig. 10(a).

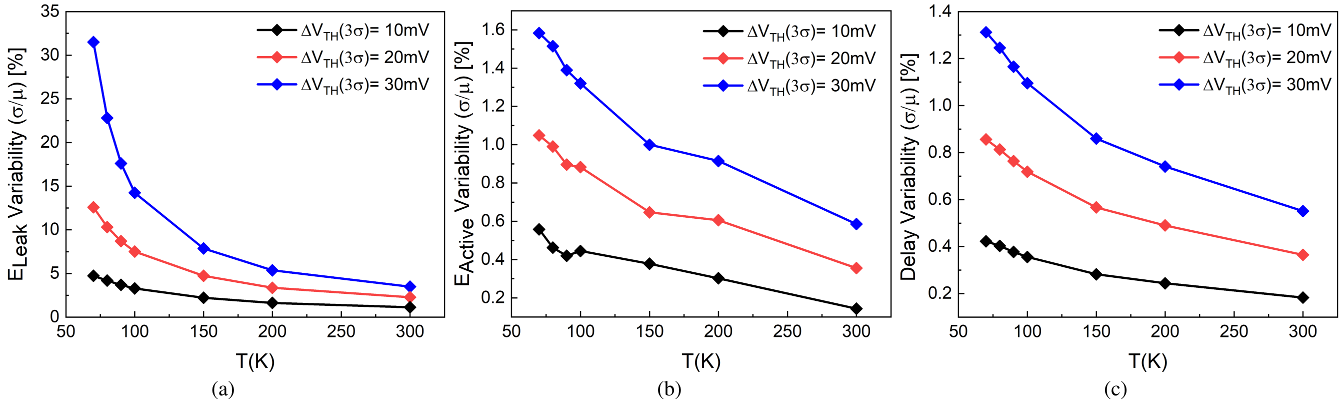


FIGURE 11. Monte Carlo simulations on 1000 sample of the 35-stage NAND-based RO under different values of V_{TH} variations and different operating temperatures for $\alpha = 0.1\%$. (a) Leakage energy (E_{Leak}) variability, (b) active energy (E_{Active}) variability, and (c) delay variability. The variations in both energy and delay increases at lower operating temperatures.

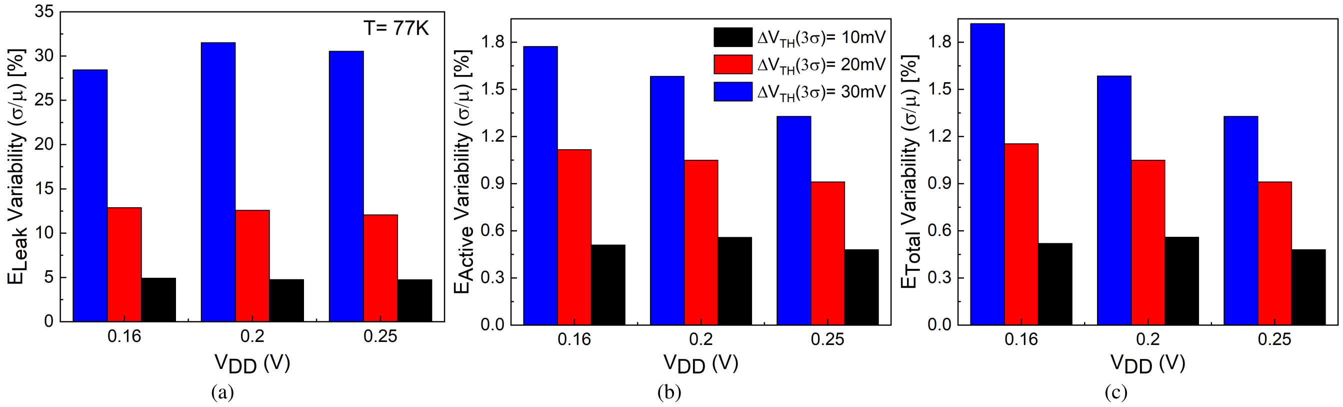


FIGURE 12. (a) Leakage energy (E_{Leak}) variability, (b) active energy (E_{Active}) variability, and (c) total energy (E_{Total}) variability of the RO under V_{TH} variations. To maintain the iso–delay condition, it is necessary to increase both V_{DD} and V_{TH}, which can effectively reduce the variation in energy.

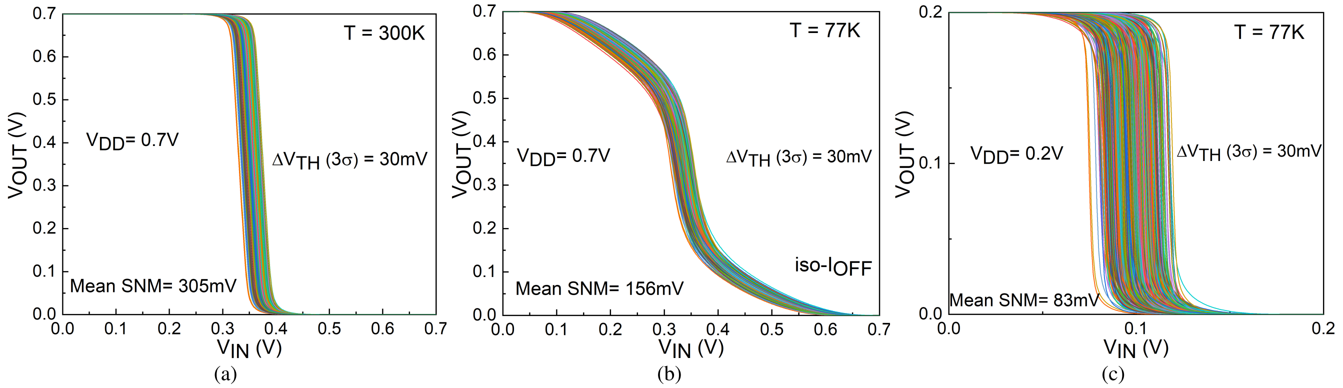


FIGURE 13. Statistical simulation of the VTC of a CMOS inverter. (a) Inverter VTC at 300 K under V_{TH} variations at $V_{DD} = 0.7$ V. (b) Inverter VTC at 77 K under V_{TH} variations at $V_{DD} = 0.7$ V under the iso–I_{OFF} condition (i.e., lower V_{TH}). (c) Inverter VTC at 77 K under V_{TH} variations at $V_{DD} = 0.2$ V.

C. SNM UNDER VARIATIONS

To further understand the impact of variations on circuit stability at 77 K, we analyze the SNM of a CMOS inverter under metal gate WF variations (i.e., V_{TH} variations) using Monte Carlo simulations over 1000 samples. Fig. 13(a) shows an inverter voltage transfer curve (VTC) for 1000 samples at the RT for $V_{DD} = 0.7$ V, where 3σ of V_{TH} is at 30 mV. The mean value of SNM at the RT is 305 mV. Fig. 13(b) shows the inverter VTC at 77 K under the iso–I_{OFF} condition (i.e., V_{TH} is reduced at 77 K from the value at the RT). Due to lower V_{TH}, the n-FinFET is turned on at a lower input voltage, but the p-FinFET is turned off at a higher input voltage.

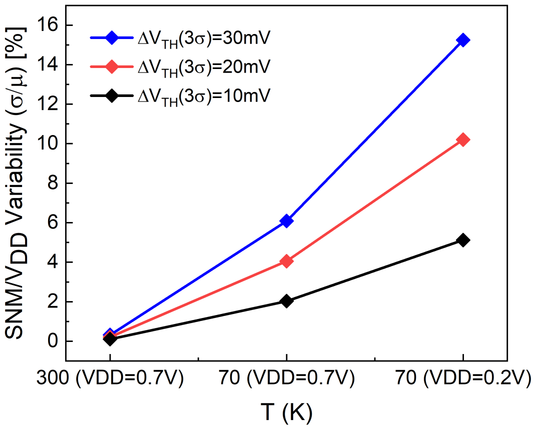


FIGURE 14. **SNM/V_{DD} variability for three cases in Fig. 13 under different amount of variations in V_{TH}.**

Thus, we observe a significant degradation in SNM down to 156 mV at 77 K compared to the RT case under the iso–I_{OFF} condition. This degradation can be fixed by appropriate tuning of V_{TH} and V_{DD}. For example, Fig. 13(c) demonstrates a steep inverter VTC with the mean value of SNM as 83 mV at $V_{DD} = 0.2$ V, which has a similar ratio of SNM/V_{DD} as Fig. 13(a).

On the other hand, similar to the logic circuit, the inverter VTC is more sensitive to process variations at cryogenic temperatures. We compare the variability of SNM$/V_{DD}$ at three different amount of variations in V_{TH} for all aforementioned three cases in Fig. 14. The variation in SNM significantly increases at lower operating temperature, for the same V_{DD}, and becomes even worse with lower V_{DD} at cryogenic temperatures. Indeed, compared to logic circuits, data stability is a primary concern in cryogenic design as observed in this study.

IV. CONCLUSION

In this work, we develop a set of predictive model cards for cryogenic operations. Based on the BSIM-CMG template, we calibrate these model cards with TCAD simulations and silicon data at 14 nm. Leveraging these new model cards, we benchmark energy reduction at the cryogenic temperatures, at the same switching speed. The reduction in energy consumption is up to 15× at 77 K in the absence of variations, or 7× under severe process variations. The noise margin of CMOS circuits, which is an index of data stability, is even more sensitive to process variations at 77 K, especially under low V_{DD}. It is essential to conduct joint device and design research to maximize the benefit of cryogenic design in the presence of variations.

REFERENCES

[1] A. Huang, "Moore's law is dying (and that could be good)," *IEEE Spectr.*, vol. 52, no. 4, pp. 43–47, Apr. 2015.

[2] W. F. Clark, B. El-Kareh, R. G. Pires, S. L. Titcomb, and R. L. Anderson, "Low temperature CMOS—A brief review," *IEEE Trans. Compon., Hybrids, Manuf. Technol.*, vol. 15, no. 3, pp. 397–404, Jun. 1992.

[3] F. H. Gaensslen, V. L. Rideout, E. J. Walker, and J. J. Walker, "Very small MOSFET's for low-temperature operation," *IEEE Trans. Electron Devices*, vol. ED-24, no. 3, pp. 218–229, Mar. 1977.

[4] H. L. Chiang et al., "Cold CMOS as a power-performance-reliability booster for advanced FinFETs," in *Proc. IEEE Symp. VLSI Technol.*, Jun. 2020, pp. 1–2.

[5] J. B. Burr, "Cryogenic ultra low power CMOS," in *IEEE Symp. Low Power Electron. Dig. Tech. Papers*, Oct. 1995, pp. 82–83.

[6] S. S. Parihar, V. M. van Santen, S. Thomann, G. Pahwa, Y. S. Chauhan, and H. Amrouch, "Cryogenic CMOS for quantum processing: 5-nm FinFET-based SRAM arrays at 10 K," *IEEE Trans. Circuits Syst. I, Reg. Papers*, vol. 70, no. 8, pp. 3089–3102, Aug. 2023.

[7] 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.

[8] D. Prasad et al., "Cryo-computing for infrastructure applications: A technology-to-microarchitecture co-optimization study," in *IEDM Tech. Dig.*, Dec. 2022, p. 23.

[9] C. J. Glassbrenner and G. A. Slack, "Thermal conductivity of silicon and germanium from 3 °K to the melting point," *Phys. Rev.*, vol. 134, pp. 1058–1069, May 1964.

[10] A. Grill et al., "Reliability and variability of advanced CMOS devices at cryogenic temperatures," in *Proc. IEEE Int. Rel. Phys. Symp. (IRPS)*, Apr. 2020, pp. 1–6.

[11] V. Moroz et al., "Challenges in design and modeling of cold CMOS HPC technology," in *Proc. Int. Conf. Simul. Semiconductor Processes Devices (SISPAD)*, Sep. 2021, pp. 107–110.

[12] A. Chabane et al., "Cryogenic characterization and modeling of 14 nm bulk FinFET technology," in *Proc. IEEE 47th Eur. Solid State Circuits Conf. (ESSCIRC)*, Sep. 2021, pp. 67–70.

[13] *Sentaurus Device User Guide, Version T-2022.03*, Synopsys, Mountain View, CA, USA, 2022.

[14] A. Beckers, F. Jazaeri, and C. Enz, "Characterization and modeling of 28-nm bulk CMOS technology down to 4.2 K," *IEEE J. Electron Devices Soc.*, vol. 6, pp. 1007–1018, 2018.

[15] Y. S. Chauhan, *FinFET Modeling for IC Simulation and Design: Using the BSIM-CMG Standard*. Amsterdam, The Netherlands: Academic Press, 2015.

[16] A. Schenk, "Unified bulk mobility model for low- and high-field transport in silicon," *J. Appl. Phys.*, vol. 79, no. 2, pp. 814–831, Jan. 1996.

[17] R. Quay, C. Moglestue, V. Palankovski, and S. Selberherr, "A temperature dependent model for the saturation velocity in semiconductor materials," *Mater. Sci. Semicond. Process.*, vol. 3, nos. 1–2, pp. 149–155, Mar. 2000.

[18] G. Pahwa, P. Kushwaha, A. Dasgupta, S. Salahuddin, and C. Hu, "Compact modeling of temperature effects in FDSOI and FinFET devices down to cryogenic temperatures," *IEEE Trans. Electron Devices*, vol. 68, no. 9, pp. 4223–4230, Sep. 2021.