Cryo-DTCO: A Disorder-Aware Compact Model for Deep-Cryogenic FETs Enabling Multi-bit Computing-in-Memory with 2T-eDRAM

Xufan Li[1,2], Zeyu Guo[1], Shengzhe Yan[1], Shuaidi Zhang[1], Jiawei Wang[1], Xiaoxin Xu[1], Feng Zhang[1], Jinshan Yue[1*], Zhicheng Wu[1*], Lingfei Wang[1,2*] and Ling Li[1,2*]

[1]State Key Laboratory of Fabrication Technologies for Integrated Circuits, Institute of Microelectronics, Chinese Academy of Sciences, Beijing, 100029, China; [2]University of Chinese Academy of Sciences, Beijing, 100049, China

*Email: yuejinshan@ime.ac.cn; wuzhicheng@ime.ac.cn; wanglingfei@ime.ac.cn; lingli@ime.ac.cn

***Abstract*—Cryogenic embedded-DRAM (eDRAM) is a promising candidate for energy-efficient high-performance computing (HPC), offering long retention time, high density, and low power consumption. However, disorder effects at deep-cryogenic temperatures pose significant challenges to classical compact modeling, particularly in the deep-subthreshold regime. This work presents a disorder-aware compact model for cryogenic FETs incorporating band-tail states and quantum confinement effects. A self-consistent Poisson-Schrödinger solver provides the physical foundation for quantum correction, and the model is validated against 14-nm FinFET experiments. The leakage characteristics are further calibrated to gate-induced drain leakage and retention time measurements from 2T-eDRAM cells at 10 K, extending the model capability beyond the instrument detection limit. With variability-aware statistical analysis, the model enables 4-bit analog computing-in-memory (CIM) in a cryogenic 2T-eDRAM array, achieving an area efficiency of 14.9 TOPS/mm² with 3 orders of magnitude longer retention time than at room temperature. A negative-feedback current write-verify (NF-CWV) scheme is further proposed to mitigate inter-level overlap, offering a potential accuracy improvement from 12.5% to 70.4% with 1.4% overhead.**

Keywords—DTCO, 2T-eDRAM, cryogenic, compact model

I. Introduction

Cryogenic embedded-DRAM (eDRAM) is a promising building block for future high-performance computing (HPC) architectures, offering long retention time, high density, and low power consumption [1, 2]. However, device variability and reliability challenges at advanced technology nodes are significantly amplified at cryogenic temperatures due to disorder effects and device miniaturization [3, 4], resulting in variable retention time and hindering variation-tolerant circuit design. Given the strong correlation between device leakage and cell retention, a design-technology co-optimization (DTCO) flow [Fig. 1], targeting the deep-cryogenic and deep-subthreshold regime, beyond the scope of classical compact models, is highly demanded for reliable circuit optimization. In this work, a self-consistent Poisson-Schrödinger solver is employed to provide the physical foundation for cryogenic compact modeling with quantum correction. To characterize the leakage behavior beyond the instrument detection limit, gate-induced drain leakage and retention time are experimentally extracted from 2T-eDRAM cells. The resulting compact model is applicable across a wide range of temperatures and operating conditions, and incorporates multi-device variability to enable 4-bit cryogenic computing-in-memory (cryo-CIM) in an eDRAM architecture.

II. Device Modeling

A. Potential Calculation

A self-consistent Poisson-Schrödinger solver incorporating the material-aware sub-band density-of-states (DOS) [5] is first developed to obtain the quantum-mechanical carrier density profile across the silicon body at cryogenic

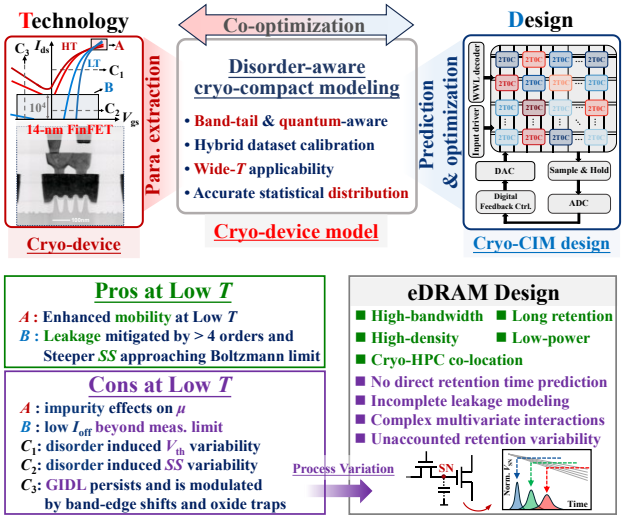

Fig. 1. Pros and cons of cryogenic device operation and the proposed DTCO framework based on 14-nm FinFET technology. While low-temperature operation benefits eDRAM design, disorder-induced variability, incomplete leakage modeling, and retention uncertainty require a disorder-aware compact model for multi-bit CIM optimization.

temperatures [Fig. 2(a)], serving as the calibration reference for the quantum correction detailed later in this section.

For compact modeling, the band-tail DOS is assumed to follow an exponential distribution characterized by T_0. These disorder effects, together with dopant freeze-out and bandgap variation, collectively contribute to a V_{th} shift:

$$\Delta V_{th} = \pm \eta_{ion} \cdot \frac{qN_{dop} \cdot t_{si}}{2\,C_{ox}} + \frac{Q_{it}}{C_{ox}} + \frac{\alpha T^2}{2q(T+\beta)}, \quad (1)$$

where the sign is chosen according to the carrier type, and α and β are the Varshni coefficients for bandgap variation. Here, $\eta_{ion} = \left(-1+\sqrt{1+2\xi}\right)/2\xi$ is the incomplete ionization coefficient with $\xi = 2g_{d,a}\left(N_{dop}/N_{c,v}\right)\exp\left(\left|E_{c,v}\text{-}E_{d,a}\right|/qV_T\right)$. To obtain an analytical solution to the Poisson equation, the doping-related term coupled with incomplete ionization is approximated by a V_{th}-shift due to light doping [6]. Then, the total charge density consists of two terms, i.e., free carrier density (N_F) governed by thermal voltage $V_T = k_B T/q$, and band-tail state carrier density (N_T) characterized by $V_{TA} = k_B T_0/q$ related to the band-tail width. This dual-thermal-voltage formulation enables a unified room-to-cryogenic description: N_F dominates at high temperatures while N_T dominates at cryogenic temperatures. For simplicity, the density is approximated in separated temperature regions, yielding a unified Poisson equation:

$$\frac{\partial^2 \varphi(x)}{\partial x^2} = A_i \cdot \exp\left(\frac{\varphi - V_{ch}}{V_i}\right), \quad (2)$$

where A_i is the density-related prefactor, and the index i denotes either the free-carrier ($V_i = V_T$) or trapped-carrier ($V_i = V_{TA}$) branch. Based on Y. Taur's method [7], Eq. (2) is

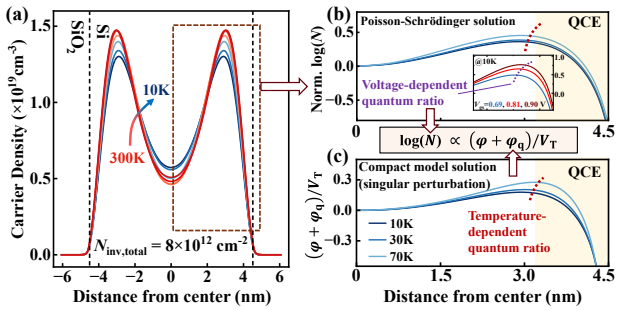

Fig. 2. (a) Carrier density profile across the silicon body solved by the self-consistent Poisson-Schrödinger solver. (b) Normalized carrier density near the surface; inset shows the voltage-dependent quantum ratio. (c) Quantum-corrected potential derived by singular perturbation method, calibrated to (b), for compact model implementation.

transformed into a transcendental equation for the parameter B_p, which is related to the center potential:

$$F = U_g - \left[\ln B_p - \ln(\cos B_p) + \eta_c B_p \tan B_p\right] = 0, \quad (3)$$

where $U_g = (V_{gs}\text{-}V_{th}\text{-}\Delta V_{th}\text{-}V_{ch})/2V_i$ is the normalized gate voltage, $\eta_c = 2C_{Si}/C_{ox}$, and $B_p = \frac{1}{2}\eta_c t_{Si}\exp[(\varphi_0\text{-}V_{ch})/2V_i]$ is the normalized surface potential parameter. Considering the dominant factors in the above-V_{th} and below-V_{th} regions, the intermediate variables B_a and B_b are expressed as:

$$B_a = \ln\left[1+\exp\left(R_0 \cdot \arctan\left(\frac{U_g}{\eta_c}\right)\right)\right]/R_0, \quad (4)$$

$$B_b = \left[\frac{1}{2\eta_c} W\left(2\eta_c \cdot e^{2U_g}\right)\right]^{1/2}, \quad (5)$$

where B_a and B_b denote the corresponding approximate solutions, respectively, and $W(\cdot)$ denotes the Lambert-W function. A Schröder correction is applied to improve the accuracy of B_a near the threshold transition:

$$B_{ac} = B_a - \frac{F(B_a)\cdot F'(B_a)}{[F'(B_a)]^2 - F(B_a)\cdot F''(B_a)}, \quad (6)$$

where all derivatives are evaluated at B_a. The corrected branch B_{ac} and the below-V_{th} branch B_b are then connected using a smooth stitching function:

$$B_{p,j} = \frac{1}{2}\left(B_{ac,j}+B_{b,j} - \sqrt{(B_{ac,j} - B_{b,j})^2+\delta_B^2}\right), \delta_B \ll 1 \quad (7)$$

where $j \in \{s, d\}$ denotes the source or drain terminal. By substituting density parameters of N_F ($V_i = V_T$) and N_T ($V_i = V_{TA}$), the potential can be expressed at both low-T and high-T. Fig. 3(a, b) shows good agreement between analytical and numerical results under different voltages and temperatures.

B. Current Expression

The normalized terminal charge is defined as $Q_{i,j} = B_{p,j}^{(i)}\tan(B_{p,j}^{(i)})$. The free-carrier and band-tail charge components are then synthesized into an effective charge:

$$Q_{eff,j} = \frac{\sqrt{(V_T \cdot Q_{free,j})^2 + (V_{TA}\cdot Q_{trap,j})^2}}{V_T + V_{TA}}, \quad (8)$$

which provides a continuous transition from free-carrier-dominated transport at room temperature to band-tail-state-assisted transport at cryo-T. The carrier transport is further described by a semi-empirical T-dependent mobility model incorporating phonon and surface-roughness scattering:

$$\mu_{eff} = \frac{U_0 \cdot \mu_0}{1 + V_T^{m1}/U_1 + V_T^{m2}\cdot V_{ov,eff}^{m3}/U_2}, \quad (9)$$

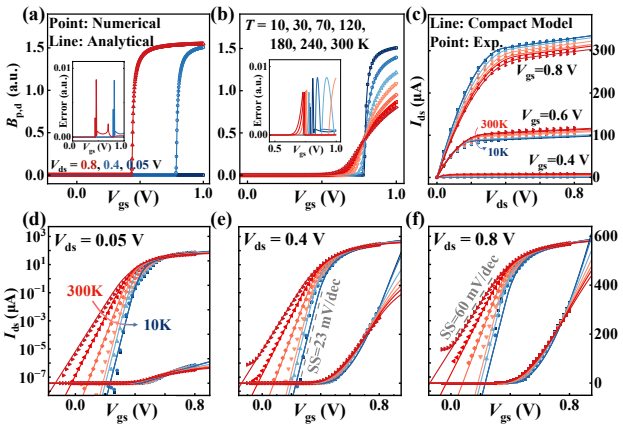

Fig. 3. Compact model calibrated to experimental data for parameter extraction. Agreement between numerical and analytical results for B_p at different (a) V_{ds} and (b) T with errors in the insets validating the accuracy of compact model. (c) shows the fitting results for output curves, and (d-f) show the corresponding transfer characteristics at V_{ds} of 0.05V, 0.4V and 0.8V.

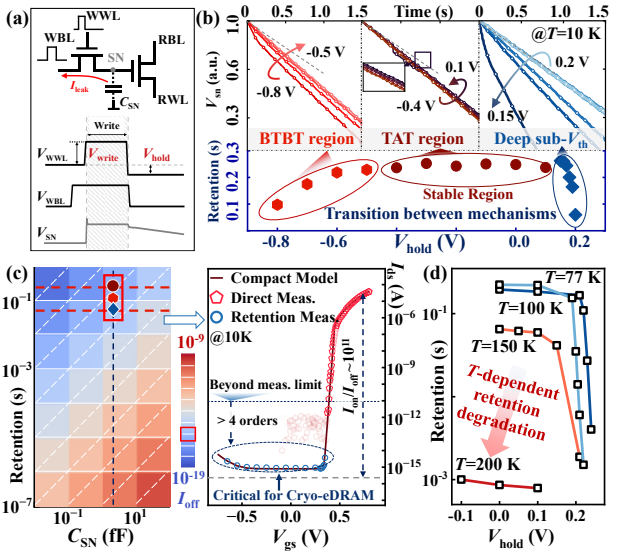

Fig. 4. (a) 2T-eDRAM cell schematic and operation waveforms. (b) SN discharge curves at various V_{hold} at 10 K, exhibiting retention time governed by band-to-band tunneling (BTBT), trap-assisted tunneling (TAT), and deep sub-V_{th} leakage regimes. (c) Retention-to-I_{off} mapping validated against measurements. (d) T and V_{hold} dependence of measured retention time.

where μ_0 is the low-field mobility, U_0-U_2 are fitting coefficients, and m_1-m_3 are empirical mobility-degradation exponents. The effective smooth overdrive voltage is defined as $V_{ov,eff} = V_{ref} + \ln\left[1+\exp\left(R_1\cdot(V_{gs}\text{-}V_{th} - \Delta V_{th})\right)\right]/R_1$, where V_{ref} and R_1 control the low-bias offset and threshold smoothness, respectively. The current is finally expressed as:

$$I_{ds} = I_0 \cdot (G_s - G_d) + I_{GIDL}, \quad (10)$$

where $G_j = Q_{eff,j} + \eta_c \cdot Q_{eff,j}^2$, and the prefactor I_0 is as follows:

$$I_0 = \mu_{eff}\frac{W}{L} 16 C_{Si}\cdot\left[(V_T^2 + V_{TA}^2)/(V_T + V_{TA})\right]^2. \quad (11)$$

Besides, the quantum confinement effect (QCE) is critical in the degradation of on-state current (I_{on}) in ultra-thin-body devices [8]. Due to QCE, the charge centroid shifts away from the Si/SiO$_2$ interface, weakening the effective gate control. By combining the density-gradient method with singular perturbation analysis [9], the peak position of the carrier distribution can be predicted and used to describe the effective charge-centroid shift. A quantum-confinement correction factor is then defined as

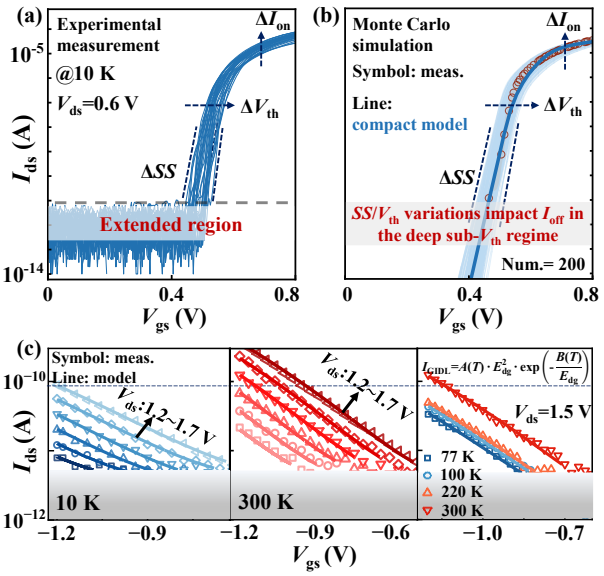

Fig. 5. Measured I_{ds}-V_{gs} of multiple devices at 10 K with off-current floor at 1 pA. (b) Model fitting with 200 Monte Carlo curves, enabling statistical prediction beyond the measurement limit. (c) GIDL characteristics under varying V_{ds} and temperatures (77-300 K), dominated by BTBT.

$$\gamma_{QCE} = v \cdot \frac{x_p}{t_{Si}/2}, \tag{12}$$

where x_p is obtained via a Schröder iteration similar to Eq. (6), and v is a fitting parameter. The quantum-corrected drain current is then expressed as $I_{ds,q} = \gamma_{QCE} I_0 \cdot (G_s - G_d) + I_{GIDL}$. As shown in Fig. 2(b, c), QCE is T-dependent and becomes more pronounced at low-T. Based on this model, the simulated T-dependent transfer and output curves show excellent agreement with the experimental data in Fig. 3(c-f), validating the compact modeling framework for cryogenic circuit design.

III. DRAM Cell Retention and Leakage Analysis

A. Cryo-eDRAM Retention Characterization

Fig. 4(a) depicts the schematic of the 2T-eDRAM configuration. The measured retention time at 10 K is shown in Fig. 4(b), where three leakage regimes can be identified depending on the dominant leakage mechanism. Among them, the trap-assisted tunneling (TAT) regime exhibits the best retention performance and is considered to be related to substrate and interfacial defects. To understand the impact of different leakage mechanisms, a mapping relation between leakage and retention is established [Fig. 4(c)], allowing the identification of the optimal operation region for the studied device. The temperature dependence is further examined by measuring the retention time at different temperatures [Fig. 4(d)], which shows a sharp degradation above 150 K.

B. Multivariate Analysis and Statistical Prediction

A multivariate analysis is further performed to capture the variability in cryogenic FETs. Disorder associated with band-tail states induces large carrier-density fluctuations at low temperature and is considered the primary origin of the observed statistical distribution [10]. For accurate circuit prediction, the compact model is calibrated against experimental data including statistical effects, showing good agreement in Fig. 5(a,b). In addition, Fig. 5(c) reveals strong V_{ds}-dependent GIDL at large negative V_{gs}, with only a mild temperature dependence attributed to the variation of the silicon bandgap. Since GIDL strongly affects the leakage current, it is also critical for weak-cell analysis and contributes to the retention-tail distribution [11].

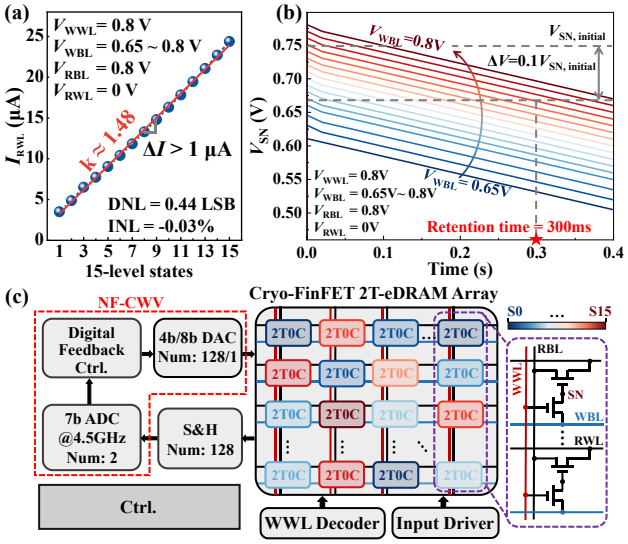

Fig. 6. (a) I_{RWL} linearity across 15 non-zero storage levels (16 levels including zero for 4-bit operation). (b) Retention time for each storage level. (c) Block diagram of the cryo-CIM macro.

IV. Multi-bit Computing-in-2T-eDRAM

A. 4-bit 2T-eDRAM cryo-CIM

To leverage the low leakage current and high mobility of FinFET devices at cryogenic temperatures, the proposed cryo-DTCO flow is used to enable multi-level storage in a single cell and thereby improve storage density. The readout currents, which correspond to multiple storage states of a single 2T0C cell, demonstrate remarkable linearity across 15 distinct non-zero states [Fig. 6(a)]. Together with the off-state, these states enable 4-bit storage and 4-bit analog CIM in FinFET-based 2T-eDRAM arrays at low temperatures. Fig. 6(b) demonstrates the retention time of 15 states. It can be observed that the retention time is approximately 300 ms for all states, sufficient to complete real-time applications. Fig. 6(c) illustrates the cryo-CIM circuit architecture, which includes the FinFET-based 2T-eDRAM array, write wordline (WWL) decoder, input and write drivers, DACs, sample-and-hold (S&H) units, ADCs, negative-feedback current write-verify (NF-CWV) module, and control circuitry. The DACs, feedback controller, and ADCs constitute the NF-CWV module. The main operations performed in the cryo-CIM circuits are as follows:

1) Write operation: The WWL decoder activates a specific row with high voltage on WWL. Once a row of cells is activated, the write driver applies one of the 16 voltage levels to the write bitline (WBL) to program a 4-bit state.

2) Read operation: Similar to the write process, the corresponding row is selected through the read bitline (RBL). The current is sensed from the read wordline (RWL), sampled by the S&H, and then digitized by the ADC.

3) Parallel CIM operation: The input driver applies voltages corresponding to 1-bit inputs '0' and '1' to the RBLs. The MAC result is accumulated as I_{RWL} on each RWL, sampled by the S&H, and digitized by the ADC.

During programming, NF-CWV is employed to reduce state overlap caused by device non-uniformity. The programmed current is read back and compared with the target level, and the write condition is iteratively updated until

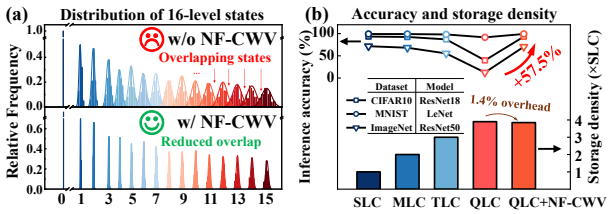

Fig. 7. (a) Distributions of the 16 storage states without and with NF-CWV, illustrating mitigation of inter-level overlap. (b) Inference accuracy and storage density for SLC, MLC, TLC, QLC, and QLC with NF-CWV.

TABLE I. COMPARISON WITH PREVIOUS WORKS

Specification	**This work**	TCAS-I 2023 [13]	IEDM 2020 [14]	ISSCC 2024 [15]	ISSCC 2023 [16]
Technology / Device	14 nm FinFET eDRAM	45 nm a-IGZO eDRAM	22 nm a-IGZO eDRAM	28 nm CMOS eDRAM	28 nm CMOS eDRAM
Array size	512×128	128×128	512×512	256×160	320×320 (8×12 subarrays)
Stored weight	Analog (**4-bit**)	Binary	Binary	Binary	Binary
Array precision (W/I/O)	4, 1, 7	4, Analog, 8	Ternary, 8, 16	8, 1, 22	5/9, 4/8, 13
Energy efficiency (TOPS/W) (8b×8b)	**2281[a], 43.1[b]**	42.88[a] , 8.63[b]	337.5[a], 2.81[b]	18.1[b]	36.2[b]
Area efficiency (TOPS/mm^2) (8b×8b)	**117.2[a], 14.9[b]**	0.61[a], 0.12[b]	6.94[a] , 0.75[b]	16.2[b]	0.63[b]
Retention time (s)	0.3 @ 10 K	>10 @ 300 K	N/A	1.3×10^{-6} @ 300 K	4×10^{-5} @ 300 K
Process maturity	High	Medium	Medium	High	High

[a]CIM array only; [b]CIM array + peripherals

the target window is reached. Its effect on the 16 storage-state distributions is evaluated at the distribution leve [Fig. 7(a)].

B. CIM evaluation on retention/power/area/accuracy

To evaluate the multi-bit storage and CIM capability, this work analyzes the storage density and neural-network (NN) accuracy of the SLC, MLC, TLC, and QLC cases. The NN evaluation follows [12] and considers ResNet18, LeNet, and ResNet50 on CIFAR10, MNIST, and ImageNet, respectively. The I_{RWL} of the multi-bit case is assumed to follow a Gaussian distribution, with the mean and variation extracted from measured devices. The error distribution of CIM results on RWLs is then obtained by Monte Carlo simulation, followed by a PyTorch evaluation of NN accuracy. The NF-CWV effect is evaluated at the distribution level. The storage density increases by 2×, 3×, and 3.9× from SLC to MLC, TLC, and QLC, respectively, while the ResNet50/ImageNet accuracy drops from 71.6% to 12.5%. With NF-CWV, the ResNet50/ ImageNet accuracy is potentially improved by 57.5%, with only 1.4% density overhead [Fig. 7(b)]. This work achieves simultaneous high area/energy efficiency and high robustness/process maturity compared to SOTA [Table I].

V. CONCLUSION

In summary, this work presents a disorder-aware compact model for deep-cryogenic FinFETs, incorporating band-tail states and quantum confinement effects, targeting the deep-subthreshold regime beyond classical modeling. The leakage characteristics are validated against experimental data, and the correlation between device leakage and cell retention is verified through 2T-eDRAM retention measurements. Leveraging variability-aware statistical analysis, the compact model enables 4-bit analog computing-in-memory in cryogenic eDRAM. The inter-level overlap caused by device variability is identified, and an NF-CWV scheme is proposed to alleviate it with minimal density overhead. The cryo-CIM macro achieves an area efficiency of 14.9 TOPS/mm^2 with 3 orders of magnitude longer retention time than at room temperature, demonstrating a practical design-technology co-optimization framework for reliable cryogenic high-performance computing.

ACKNOWLEDGMENT

This work was supported by National Natural Science Foundation of China (Grant 62274178, U25A20481).

REFERENCES

[1] Alam S, Hossain M S, Srinivasa S R, et al., "Cryogenic memory technologies," Nature Electronics, vol. 6(3), pp. 185–198, 2023.
[2] Saligram R, Datta S and Raychowdhury A, "CryoMem: A 4K-300K 1.3 GHz eDRAM macro with hybrid 2T-gain-cell in a 28nm logic process for cryogenic applications," 2021 IEEE Custom Integrated Circuits Conference (CICC), pp. 1–2
[3] Hart P T, Babaie M, Charbon E, et al., "Subthreshold mismatch in nanometer CMOS at cryogenic temperatures," IEEE Journal of the Electron Devices Society, vol. 8, pp. 797–806, 2020.
[4] Zhao Y, Wang L, Wu Z, et al., "A Unified Physical BTI Compact Model in Variability-Aware DTCO Flow: Device Characterization and Circuit Evaluation on Reliability of Scaling Technology Nodes," 2021 Symposium on VLSI Technology, pp. 1–2
[5] Zhang X, Wu Z, Bu J, et al., "Modeling of the Subthreshold Swing in Cryogenic MOSFET With the Combination of Gaussian Band Tail and Gaussian Interface State," IEEE Transactions on Electron Devices, vol. 71(2), pp. 1173–1178, 2024.
[6] Lu H, Lu W-Y and Taur Y, "Effect of body doping on double-gate MOSFET characteristics," Semiconductor Science and Technology, vol. 23(1), pp. 015006, 2007.
[7] Lu H and Taur Y, "An analytic potential model for symmetric and asymmetric DG MOSFETs," IEEE Transactions on Electron Devices, vol. 53(5), pp. 1161–1168, 2006.
[8] Aouad M, Martinie S, Triozon F, et al., "Poisson-Schrödinger simulation of inversion charge in FDSOI MOSFET down to 0K-Towards compact modeling for cryo CMOS application," 2020 Joint International EUROSOI Workshop and International Conference on Ultimate Integration on Silicon (EUROSOI-ULIS), pp. 1–4
[9] Huang S, Wu Z, Xu H, et al., "Geometric Variability Aware Quantum Potential based Quasi-ballistic Compact Model for Stacked 6 nm-Thick Silicon Nanosheet GAA-FETs," 2021 IEEE International Electron Devices Meeting (IEDM), pp. 18.5.1–18.5.4
[10] Wang Z, Wang H, Li W F, et al., "Towards Understanding the Dynamic Variation in FinFET at Cryogenic Temperature: New Observations and Physical Modeling," 2024 IEEE International Electron Devices Meeting (IEDM), pp. 1–4
[11] Saino K, Horiba S, Uchiyama S, et al., "Impact of gate-induced drain leakage current on the tail distribution of DRAM data retention time," International Electron Devices Meeting 2000. Technical Digest. IEDM (Cat. No. 00CH37138), pp. 837–840
[12] Wang Z, Luo H, Peng Z, et al., "An 8T SRAM based digital compute-in-memory macro for multiply-and-accumulate accelerating," 2023 IEEE International Symposium on Circuits and Systems (ISCAS), pp. 1–5
[13] Tang W, Liu J, Sun C, et al., "Low-Power and Scalable BEOL-Compatible IGZO TFT eDRAM-Based Charge-Domain Computing," IEEE Transactions on Circuits and Systems I: Regular Papers, vol. 70(12), pp. 5166–5179, 2023.
[14] Houshmand P, Cosemans S, Mei L, et al., "Opportunities and limitations of emerging analog in-memory compute DNN architectures," 2020 IEEE International Electron Devices Meeting (IEDM), pp. 29.1. 1–29.1. 4
[15] He Y, Fan S, Li X, et al., "34.7 A 28nm 2.4 Mb/mm 2 6.9-16.3 TOPS/mm 2 eDRAM-LUT-Based Digital-Computing-in-Memory Macro with In-Memory Encoding and Refreshing," 2024 IEEE International Solid-State Circuits Conference (ISSCC), 67,pp. 578–580
[16] Kim S, Li Z, Um S, et al., "16.5 DynaPlasia: An eDRAM In-Memory-Computing-Based Reconfigurable Spatial Accelerator with Triple-Mode Cell for Dynamic Resource Switching," 2023 IEEE International Solid- State Circuits Conference (ISSCC), pp. 256–258