Pushing the Boundaries of AI Chips: From Monolithic 3D CMOS to Cryogenic Computing

Mahdi Benkhelifa[1,◇], Shivendra S. Parihar[2,◇], Anirban Kar[1], Girish Pahwa[3], Yogesh S. Chauhan[4], Hussam Amrouch[1,*]
[1]Technical University of Munich; TUM School of Computation, Information and Technology;
Chair of AI Processor Design; Munich Institute of Robotics and Machine Intelligence, Munich, Germany
[2]Department of Electrical Engineering and Computer Sciences, University of California, Berkeley, USA
[3]National Yang Ming Chiao Tung University; International College of Semiconductor Technology, Taiwan
[4]Indian Institute of Technology Kanpur, India
[◇]Equal contribution; [*]Corresponding author: amrouch@tum.de

***Abstract*—As CMOS scaling approaches its fundamental limits, the explosive rise of AI and LLMs has unveiled profound bottlenecks in computing architectures. This paper presents two groundbreaking paradigms poised to reshape the landscape of high-performance computing and meet the surging demands of AI-driven workloads. The first paradigm is 3D monolithic integration, a revolutionary approach that achieves unprecedented logic density through Complementary FETs (CFETs), where pMOS and nMOS transistors are vertically stacked, and a dramatic expansion of on-chip memory capacity by integrating memory layers atop logic transistors. The second paradigm leverages the transformative potential of operating chips at cryogenic temperatures where transistors exhibit enhanced performance, and parasitic resistances are substantially minimized. These advancements hold the promise of redefining computing efficiency and performance for the AI era.**

***Index Terms*—Cryogenic, AI acceleration, 3D integration**

I. Introduction

The rapid advancements in artificial intelligence and the recent proliferation of large language models have underscored the critical limitations of conventional computing architectures. As transistor scaling approaches its fundamental physical boundaries, there is an imperative to pioneer transformative technologies that can sustain the exponential growth in computational demands. This paper introduces two new paradigms poised to redefine high-performance computing in the AI era: 3D monolithic integration and cryogenic chip operation. 3D monolithic integration represents a disruptive leap in the pursuit of higher performance and energy-efficient computing [1]. By vertically stacking pMOS and nMOS transistors in the form of Complementary FETs (CFETs), this approach achieves unprecedented logic density [2]. Critically, it revives the stalled progress of SRAM scaling, which has decelerated markedly in recent years and stagnated in the latest technology nodes. Reducing SRAM cell area directly increases on-chip memory capacity, which is pivotal for enhancing energy efficiency, as off-chip memory access accounts for the majority of energy dissipation in AI workloads. Furthermore, the integration of memory layers directly atop logic transistors enables a dramatic expansion of on-chip memory capacity, substantially improving computational throughput while simultaneously reducing latency and energy consumption. However, achieving such a monolithic 3D integration is fundamentally constrained by the inability to stack silicon layers due to thermal budget limitations. In response, oxide semiconductors, such as indium gallium zinc oxide (IGZO) and indium tungsten oxide (IWO), offer a compelling solution. These materials can be processed within the strict thermal budgets of silicon (i.e., $\leq 400°C$) [3], unlocking new possibilities for advanced device integration. In addition, operating chips at cryogenic temperatures, specifically around 77 K, presents another transformative pathway for enhancing computational efficiency [4]. At these temperatures, transistors exhibit improved performance metrics, including higher carrier mobility and minimized parasitic resistances [5]. These enhancements translate into higher performance and/or reduced power consumption, addressing the stringent demands of modern AI workloads. In addition, cryogenic operation is particularly suited for emerging applications such as quantum computing and aerospace systems, where CMOS circuits must operate at extremely low temperatures.

Together, monolithic 3D integration and cryogenic computation paradigms can tackle the dual challenges of logic density and computational efficiency and, hence, might lay the groundwork for next-generation chips tailored for the escalating demands of AI-driven workloads.

II. Cryogenic Computing using 5 nm FinFETs

For circuit-level analysis, transistors fabricated using commercial 5 nm FinFET technology are experimentally measured at temperatures down to 77 K. Then, a SPICE model is generated by calibrating a cryogenic-aware BISM-CMG compact model against the measured characteristics. Fig. 1 shows the setup for DC measurements at cryogenic temperatures, including a Lakeshore “CRX-VF” cryogenic probe station, a closed-cycle refrigerator, a switch matrix, and a Keysight B1500A semiconductor parameter analyzer. The FinFET, fabricated on a bulk silicon substrate in a ground-signal-ground configu-

978-3-9826741-0-0/DATE25/© 2025 EDAA

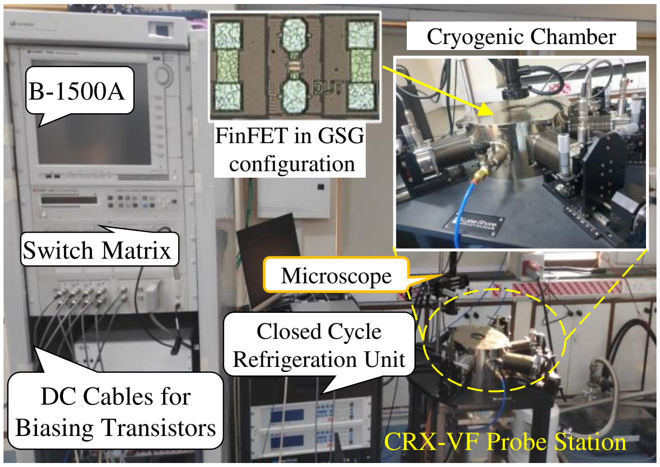
B-1500A
FinFET in GSG configuration
Cryogenic Chamber
Switch Matrix
Microscope
Closed Cycle Refrigeration Unit
DC Cables for Biasing Transistors
CRX-VF Probe Station

Fig. 1: Cryogenic characterization setup used to measure commercial 5 nm FinFETs.

ration, had their source and body terminals grounded, with voltages applied to the gate and drain terminals to measure drain current. After achieving the desired ambient pressure, the temperature is reduced to 77 K. Once stabilized, the DC probe tips are contacted with the wafer and allowed to reach thermal equilibrium. Electrical connections are made to the desired pads, and DC voltages are applied using the source measurement units to measure voltages and currents from the FinFET terminals. Further details on the measurements setup are available in [4].

In advanced CMOS technologies, the impact of Self-Heating (SH) extends up to several GHz due to their 3D architectures [4]. To accurately model the influence of SH and cryogenic temperatures on transistor characteristics, we have modified the industry-standard BSIM-CMG compact model for FinFETs [6]. Specifically, we replaced the default 1st-order thermal network with a 4th-order thermal network to capture SH up to the iso-thermal frequency. Cryogenic temperature effects were incorporated using the modeling methodology outlined in [7]. For model calibration, we first extracted thermal network parameters as described in [8]. To account for SH at cryogenic temperatures, the thermal resistance was scaled by a factor of six, following the cryogenic thermal resistance scaling presented in [5]. The DC transfer characteristics were calibrated across the measured temperature range, starting with parameter extraction at nominal temperature (300 K). Once the 300 K model was established, temperature-dependent parameters of the modified cryogenic-aware BSIM-CMG model were extracted to capture temperature scaling effects. Detailed model extraction procedures, including physical effect parameters, can be found in our prior work [4], [6], [9]. Fig. 2a shows the validation of the calibrated model against 5 nm FinFET measurement data at drain-source voltage = 750 mV.

Operation at cryogenic temperatures increases the V_{TH} of transistors. To maximize performance and minimize power consumption in cryogenic circuits, CMOS transistors can be optimized specifically for cryogenic conditions [5], [10], [11]. Fig. 2b illustrates two scenarios using simulation-based transfer characteristics obtained from the calibrated FinFET model. In this work, we optimize FinFETs for iso-I_{OFF} and iso-V_{TH} conditions. The iso-I_{OFF} optimization (solid green lines in Fig. 2b) ensures that I_{OFF} at cryogenic temperatures matches that of transistors optimized for 300 K. This approach significantly enhances ON-state current (I_{ON}) at cryogenic temperatures due to the higher carrier mobility. On the other hand, iso-V_{TH} optimization (dashed green lines in Fig. 2b) aims to maintain a consistent V_{TH} at cryogenic temperatures compared to transistors optimized for 300 K. This optimization improves both I_{ON} and I_{OFF}, resulting in enhanced performance and reduced leakage power at the circuit levels.

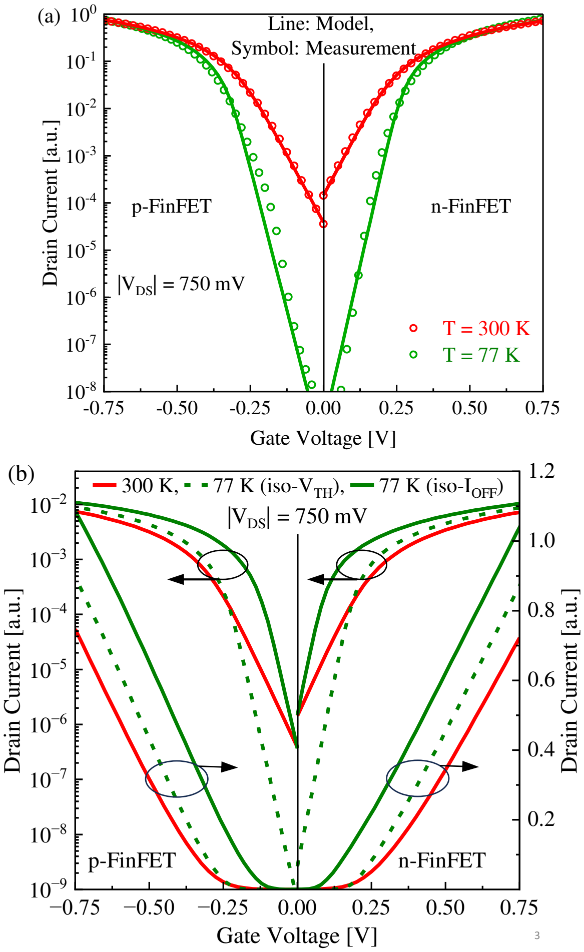

Fig. 2: (a) Compact model validation against experimentally measured 5 nm p-FinFET and n-FinFET at 300 K and 77 K. (b) Engineered FinFET transfer characteristics for iso-threshold voltage (V_{TH}) and iso-OFF-state current (I_{OFF}) operations. Green dashed and solid lines represent iso-V_{TH} and iso-I_{OFF} optimizations, respectively, at 77 K.

A. Cryogenic-Aware SRAM Evaluation at 5 nm FinFET

In AI chips, where computing cores require continuous access to on-chip SRAM-based cache memories, SRAM performance is critical for overall performance and efficiency.

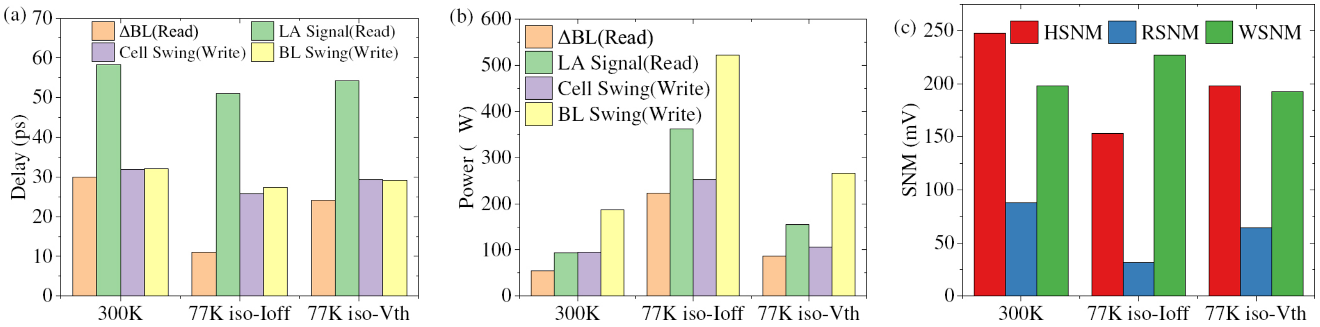

Fig. 3: (a) SRAM performance evaluation at 300 K and 77 K including iso-I_{OFF} and iso-V_{TH} conditions in terms of (a) read and write operation delays. (b) Power consumption during read and write operations. (c) Change in SNM of SRAM for hold, read, and write operations at room and cryogenic temperatures.

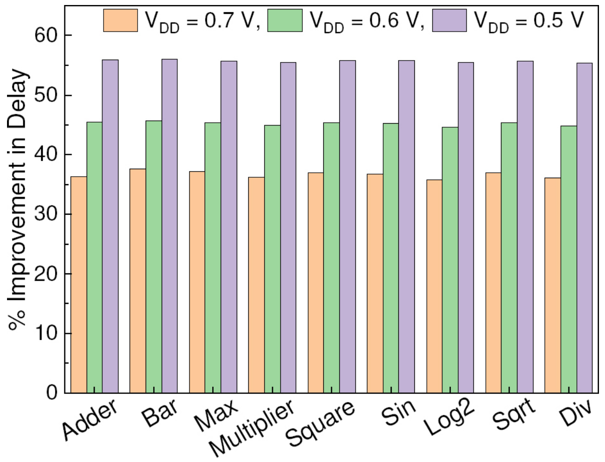

Fig. 4: Performance improvement in various benchmark circuits' speed when transistors are designed to operate in iso-I_{OFF} conditions at 77 K compared to 300 K operation.

Despite advancements in CMOS technology leading to faster and smaller transistors, increased I_{OFF} and wire parasitics hinder the performance and power density of SRAM and processors. These effects result in higher static power dissipation and latency in SRAM arrays, impacting processor performance. To increase memory capacity, SRAM arrays are typically subdivided into smaller banks, incurring area and power overheads due to additional circuitry such as write drivers and sense amplifiers. Operating CMOS transistors at 77 K offers several advantages, including steeper sub-threshold swing, higher I_{ON}, and lower wire resistance, compared to 300 K. For example, operating SRAM arrays at 77 K enables a ~1.35 × increase in size for high-density cell-based arrays without requiring assist schemes [12]. This work investigates the impact of cryogenic temperatures on SRAM reliability and performance. We utilize a cryogenic-aware model within our SRAM evaluation framework [4], which takes inputs like cell type, transistor type, supply voltage (V_{DD}), frequency, and temperature, and provides analysis of several figure of merits such as noise margins, performance, and power.

Additionally, transistors can be experimentally optimized for iso-I_{OFF} or iso-V_{TH} operation at 77 K, matching their characteristics at 300 K [5]. We compare the SNM of SRAM cells designed with iso-I_{OFF} and iso-V_{TH} transistors. Fig. 3 shows delay and power results under these conditions. Read measurements focus on two metrics: the time to produce a 100 mV difference on complementary bit lines (ΔBL), and the time to generate a latch output signal (LA). Write measurements include the time to drive the SRAM cell output node "Q" and the bit lines to 90 % of target potentials (i.e., cell swing and BL swing, respectively). As expected, iso-I_{OFF} achieves the lowest delay, with a ΔBL delay of 11.05 ps, followed by iso-V_{TH} at 24.18 ps, and the room temperature operation at 29.95 ps. However, iso-I_{OFF} results in higher power dissipation (223.7 μW) due to increased I_{ON} and I_{OFF}. Regarding reliability, SNM values at 77 K are slightly lower than at 300 K, but iso-V_{TH} achieves better Read Noise Margin (RNM) (64.5 mV) compared to iso-I_{OFF} (31.45 mV).

B. *Cryogenic-Aware Circuit Evaluation at 5 nm FinFET*

After evaluating the SRAM array, we investigated performance improvements in various benchmark circuits. Standard cell libraries for 300 K and 77 K were generated and characterized based on an FinFET PDK [13]. Circuit analysis was performed using commercial EDA tools. For arithmetic circuits, we used RTL designs from the EPFL benchmark suite [14]. Evaluations were conducted at different V_{DD} levels for both room and cryogenic temperatures. Due to the higher I_{ON} achieved with iso-I_{OFF} operation, cryogenic circuit simulations were performed exclusively with iso-I_{OFF} calibrated transistors at 77 K. Fig. 4 illustrates the improvements in critical path delays for nine benchmark circuits at different V_{DD} levels compared to 300 K. Notably, performance gains were more pronounced at lower V_{DD}, e.g., 0.5 V, compared to 0.6 V or 0.7 V. This significant improvement at reduced voltages arises from the relative increase in I_{ON} for each FinFET at 77 K compared to 300 K. These results suggest that the benefits of cryogenic operation are maximized by designing and operating circuits at lower V_{DD}, which also reduces overall power consumption. By optimizing V_{DD}, designers can achieve an optimal balance between performance gains and power efficiency. Apart from iso-I_{OFF} transistors, additional benefits can be leveraged by designing circuits with transistors

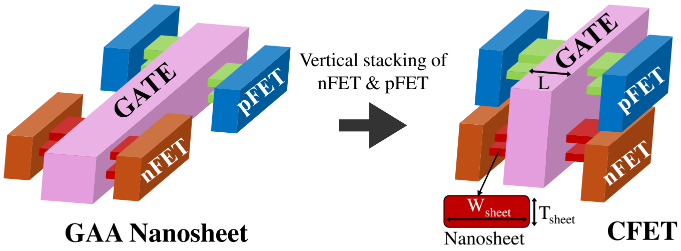

Fig. 5: Vertical stacking of pFET over nFET nanosheets results in the silicon footprint reduction [2]. CFET leads to a reduction of tracks in standard cell design.

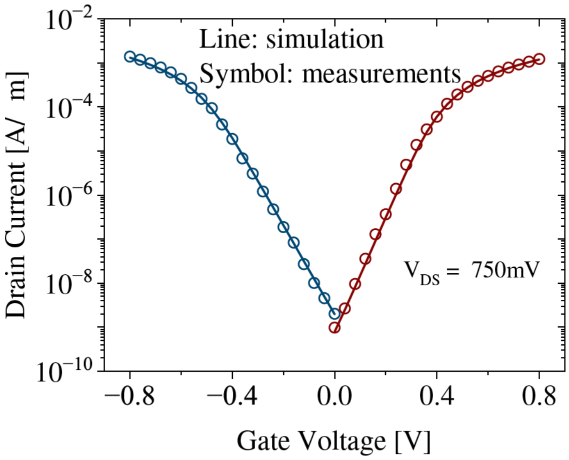

Fig. 6: SPICE simulations using our calibrated compact model of CFET accurately capture the I_{DS}-V_{GS} behavior of both nFET and pFET.

optimized for iso-V_{TH} conditions at cryogenic temperatures. In summary, cryogenic applications offer performance enhancement opportunities through the optimization of transistor and operating voltage, paving the way for more efficient chips.

C. Challenges at Cryogenic Temperatures

While cryogenic operation offers significant benefits from the transistor to circuit level, several challenges arise when designing chips for cryogenic applications. Most of these challenges stem from the intrinsic behavior of transistors at low temperatures. In older CMOS technologies, transistors exhibit abnormal behaviors such as carrier freezeout, current kinks, and hysteresis [15]–[17]. A major concern at cryogenic temperatures is SH, which becomes increasingly problematic in advanced CMOS nodes. At cryogenic temperatures, the SH issue is exacerbated as any rise in ambient temperature not only impacts circuit reliability but also leads to significantly higher cooling costs. Our previous studies have shown that while digital circuits are relatively unaffected by SH, analog circuits experience substantial increases in ambient temperature [6], [9]. Additionally, the severity of SH varies across CMOS technologies [5], [6], [9], [18]–[20], requiring circuit designers to carefully select the appropriate CMOS technology for cryogenic applications.

III. Monolithic 3D Integration: From Advanced Stacked Logic to BEOL Memories

FinFET technology, which revolutionized scaling beyond the 10 nm node and overcame the limitations of planar structures [21], is now encountering significant challenges at the sub-5 nm scale. For instance, FinFET structures in the 3 nm node might sometimes be restricted to one fin per transistor to maintain efficient fin heights, which severely impacts device performance and layout-driven scaling [22]. To address these issues, Gate-All-Around (GAA) technology was introduced, offering enhanced channel control through fully surrounding gates. GAA technologies, such as Nanosheet FETs (NSFETs), have demonstrated advancements in scalability and current-carrying capacity. However, further scaling continues to face substantial challenges, highlighting the need for alternative solutions. As device-level miniaturization encounters physical limits, alternative approaches such as exploiting vertical spaces to create 3D computing structures are gaining attention. Monolithic 3D integration (M3D) is emerging as a transformative technology for achieving higher computing densities. In M3D, monolithic interlayer vias, which are significantly smaller than conventional through-silicon vias (TSVs) [23], enable fine-grained vertical integration of transistors on the same substrate. This approach unlocks the Back-End-of-Line (BEOL) space, traditionally reserved for interconnects, for active device fabrication. M3D integration typically employs transistors with minimal thermal budget fabrication processes, such as metal-oxide channel thin-film transistors (TFT), to preserve the integrity of Front-End-of-Line (FEOL) devices. M3D enables two main types of partitioning: block-level and transistor-level integration. Block-level partitioning in M3D focuses on vertically integrating functional blocks, such as memory and logic, to achieve higher memory and computational densities. This integration facilitates near-memory computing by reducing the distance between memory and logic components. Block-level partitioning opens up a wide design space, allowing for increased memory capacity by stacking multiple tiers of memory arrays above FEOL CMOS logic [24]. Furthermore, memory devices with compute-in-memory capabilities, such as Ferroelectric TFTs [25], [26], can be incorporated into the BEOL to develop area-efficient AI accelerators [1].

CFET facilitates scaling into the sub-3 nm regime by vertically stacking pFETs over nFETs. This configuration reduces the number of tracks in standard cell layouts and significantly decreases the silicon footprint compared to traditional NSFET designs, as shown in Fig. 5 [27]. IWO-based TFTs have emerged as a promising option for BEOL compatibility due to their stability, sufficient mobility, and low leakage currents under low-temperature processing conditions (<400 °C). The n-type IWO-TFT can be integrated with various p-type technologies, enabling a flexible and generic mixed-technology approach for transistor-level M3D integration. In this work,

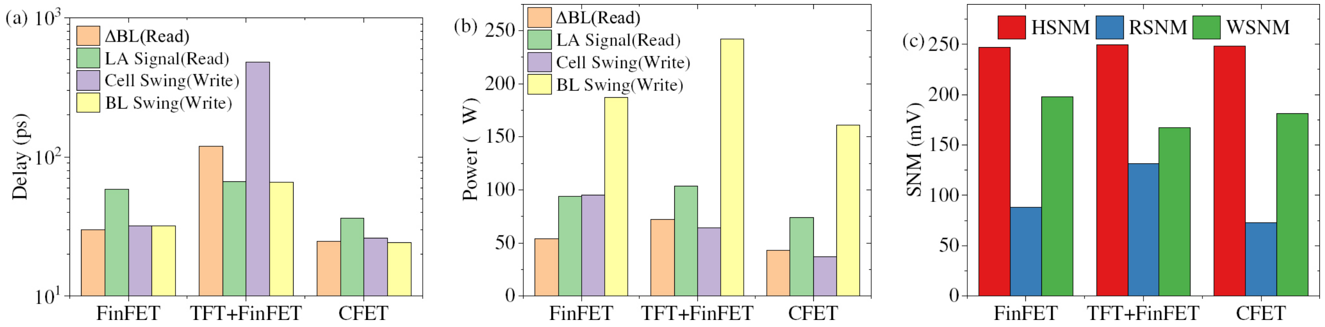

Fig. 7: SRAM performance comparison at 300 K between 3D and 2D SRAM designs under iso-I_{OFF} conditions. (a) Read and write delay results with the CFET array producing smaller delay numbers for all categories due to the improved on current performance. (b) Power consumption during read and write operations. (c) SNM comparison for hold, read, and write states. The higher p-type device performance in CFET compared to the FinFET results in a reduction in Write Noise Margin (WNM). In contrast, the weakened PG device in the case of TFT leads to a higher RNM.

we evaluate the performance of two 3D SRAM designs: one based on CFET technology and the other on BEOL IWO-TFT combined with p-type FEOL FinFET devices. These designs are benchmarked against a classical SRAM based solely on FinFET devices to assess their performance advantages.

A. CFET and IWO-TFT Transistor Compact Modeling

To evaluate the CFET 3D SRAM design, we have employed the industry-standard BSIM-CMG [28] GAA model to implement nanosheet transistors forming the CFET structure. The model was calibrated using transfer characteristics ($I_{DS} - V_{GS}$) of fabricated CFET devices [2]. Calibration involved tuning structural parameters such as equivalent oxide thickness, nanosheet width and thickness, channel length, and doping concentration. Sub-threshold behavior was adjusted through work function, interface traps, and source/drain coupling capacitance. Mobility degradation was captured via low-field mobility, phonon and surface roughness scattering, Coulomb scattering, and effective field parameters for low drain bias conditions. Velocity saturation and channel length modulation effects were captured by calibrating VSAT and PCLM parameters, respectively, with the results shown in Fig. 6. For IWO-TFT simulations, we used a model replicating electrostatic and charge transport behaviors of amorphous oxide semiconductor channels, calibrated with experimental double-gate channel data [29]. Short-channel effects, velocity saturation, and uniform interface trap density were incorporated. Detailed calibration procedures are available in [3]. To benchmark against FinFETs, the work functions of CFET and TFT models were engineered to achieve iso-I_{OFF} conditions with the FinFET reference device.

B. Performance and Reliability Evaluation of 3D SRAMs

Fig. 7 presents the performance of CFET and mixed-technology TFT+FinFET 3D SRAM designs, compared against the 5 nm FinFET 2D SRAM at room temperature. Fig. 7a highlights a speed-up for CFET across all metrics, while the TFT-based design shows increased delays due to performance limitations of PG and PD devices. The CFET design achieved a ΔBL delay of 24.64 ps, outperforming the FinFET (29.95 ps), while the TFT design lagged at 118.8 ps. For the final output LA Signal delay, CFET delivered 37.9 % and 45.7 % improvements over FinFET and TFT, respectively. During write operations, the CFET cell reached a Cell Swing delay of 26.06 ps, 18.4 % faster than FinFET, while the TFT design experienced significant delays at 480.5 ps. The BL Swing delay for CFET was 24.16 ps, attributed to higher I_{ON} compared to FinFET under iso-I_{OFF} conditions. In Fig. 7b, power consumption associated with these delays is shown. The CFET array demonstrated a 21.3 % reduction in dynamic read power, from WL activation to latch output. Write power consumption was notably lower for CFET due to reduced effective BL capacitance and faster discharging through bitlines. Fig. 7c presents the SNM during hold, read, and write operations. Hold SNM was consistent across designs, but CFET's WNM showed an 8 % drop below FinFET's 198.0 mV due to PG:PU strength degradation. Conversely, the TFT design exhibited an increased RNM of 131.77 mV due to reduced PG I_{ON}. Overall, CFET offers SRAM cell area scaling with significant speed-ups in 6T-SRAM operations compared to FinFET arrays. While TFT-based designs expand design flexibility and reduce area, they require precise engineering and matching with FEOL technologies to mitigate performance trade-offs.

IV. Conclusion

This work presents monolithic 3D integration and cryogenic operation as transformative solutions for high-performance computing. CFET-based SRAM designs achieve notable speed-ups, reduced silicon footprint, and improved efficiency, while IWO-TFT integration offers flexibility for BEOL applications with some performance trade-offs. The developed cryogenic-aware models enable optimization of circuits, enhancing their performance. Together, these advancements address scaling challenges and unlock new opportunities for efficient, high-density computing, paving the way for the next generation of chips the AI era.

Acknowledgment

This work was supported by the German Research Foundation (DFG) under grant AM 534/5-1 (NN-Thunder, project: 506419033). We would like to thank Central Research Facility IIT Delhi for facilitating the cryogenic characterization. We would like to thank from the Chair of AI Processor Design at TU Munich Simon Thomann for his development of the SRAM framework as well as Florian Klemme for his cryogenic-aware circuit analysis.

References

[1] S. Kumar, Y. S. Chauhan, and H. Amrouch, "Invited paper: Ultra-efficient edge ai using fefet-based monolithic 3d integration," in *2023 IEEE/ACM International Conference on Computer Aided Design (ICCAD)*, 2023, pp. 1–6. DOI: 10.1109/ICCAD57390.2023.10323845.

[2] S. Liao, L. Yang, T. Chiu, *et al.*, "Complementary Field-Effect Transistor (CFET) Demonstration at 48nm Gate Pitch for Future Logic Technology Scaling," in *2023 International Electron Devices Meeting (IEDM)*, 2023, pp. 1–4. DOI: 10.1109/IEDM45741.2023.10413672.

[3] G. Pahwa, S. Salahuddin, and C. Hu, "An all-region bsim thin-film transistor model for display and beol 3-d integration applications," *IEEE Transactions on Electron Devices*, vol. 71, no. 8, pp. 4701–4709, 2024. DOI: 10.1109/TED.2024.3416083.

[4] 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 Transactions on Circuits and Systems I: Regular Papers*, vol. 70, no. 8, pp. 3089–3102, 2023. DOI: 10.1109/TCSI.2023.3278351.

[5] H. L. Chiang, T. C. Chen, J. F. Wang, *et al.*, "Cold CMOS as a Power-Performance-Reliability Booster for Advanced FinFETs," in *2020 IEEE Symposium on VLSI Technology*, 2020, pp. 1–2. DOI: 10.1109/VLSITechnology18217.2020.9265065.

[6] S. S. Parihar, G. Pahwa, Y. S. Chauhan, and H. Amrouch, "Impact of self-heating in 5nm finfets at cryogenic temperatures for reliable quantum computing: Device-circuit interaction," in *2024 IEEE International Reliability Physics Symposium (IRPS)*, 2024, pp. 1–7. DOI: 10.1109/IRPS48228.2024.10529431.

[7] 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 Transactions on Electron Devices*, vol. 68, no. 9, pp. 4223–4230, 2021. DOI: 10.1109/TED.2021.3097971.

[8] S. S. Parihar, A. Pampori, P. Dwivedi, *et al.*, "A Comprehensive RF Characterization and Modeling Methodology for the 5nm Technology Node FinFETs," *IEEE Journal of the Electron Devices Society*, vol. 11, pp. 444–455, 2023. DOI: 10.1109/JEDS.2023.3298290.

[9] A. Kar, F. Klemme, Y. S. Chauhan, and H. Amrouch, "On the severity of self-heating in fdsoi at cryogenic temperatures: In-depth analysis from transistors to full processor," in *2024 IEEE International Reliability Physics Symposium (IRPS)*, 2024, pp. 1–6. DOI: 10.1109/IRPS48228.2024.10529429.

[10] P. R. Genssler, F. Klemme, S. S. Parihar, *et al.*, "Cryogenic embedded system to support quantum computing: From 5-nm finfet to full processor," *IEEE Transactions on Quantum Engineering*, vol. 4, pp. 1–11, 2023. DOI: 10.1109/TQE.2023.3300833.

[11] B. Hien, M. Walter, V. M. van Santen, *et al.*, "Technology mapping for cryogenic cmos circuits," in *2024 IEEE Computer Society Annual Symposium on VLSI (ISVLSI)*, 2024, pp. 272–277. DOI: 10.1109/ISVLSI61997.2024.00057.

[12] S. S. Parihar, G. Pahwa, B. Mohammad, Y. S. Chauhan, and H. Amrouch, "Novel trade-offs in 5nm finfet sram arrays at extreme low temperatures," *IEEE Transactions on Quantum Engineering*, pp. 1–16, 2024. DOI: 10.1109/TQE.2024.3512367.

[13] L. T. Clark, V. Vashishtha, L. Shifren, *et al.*, "ASAP7: A 7-nm finFET predictive process design kit," *Microelectronics Journal*, vol. 53, pp. 105–115, 2016. DOI: https://doi.org/10.1016/j.mejo.2016.04.006. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S002626921630026X.

[14] L. Amarú, P.-E. Gaillardon, and G. De Micheli, "The EPFL combinational benchmark suite," in *Proceedings of the 24th International Workshop on Logic & Synthesis (IWLS)*, 2015.

[15] F. Balestra, L. Audaire, and C. Lucas, "Influence of substrate freeze-out on the characteristics of mos transistors at very low temperatures," *Solid-State Electronics*, vol. 30, no. 3, pp. 321–327, 1987. DOI: https://doi.org/10.1016/0038-1101(87)90190-0. [Online]. Available: https://www.sciencedirect.com/science/article/pii/0038110187901900.

[16] H. Hanamura, M. Aoki, T. Masuhara, O. Minato, Y. Sakai, and T. Hayashida, "Operation of bulk cmos devices at very low temperatures," *IEEE Journal of Solid-State Circuits*, vol. 21, no. 3, pp. 484–490, 1986. DOI: 10.1109/JSSC.1986.1052555.

[17] B. Dierickx, L. Warmerdam, E. Simoen, J. Vermeiren, and C. Claeys, "Model for hysteresis and kink behavior of mos transistors operating at 4.2 k," *IEEE Transactions on Electron Devices*, vol. 35, no. 7, pp. 1120–1125, 1988. DOI: 10.1109/16.3372.

[18] G. Ghibaudo, M. Cassé, F. S. di Santa Maria, C. Theodorou, and F. Balestra, "Modelling of self-heating effect in FDSOI and bulk MOSFETs operated in deep cryogenic conditions," *Solid-State Electronics*, vol. 192, p. 108 265, 2022. DOI: https://doi.org/10.1016/j.sse.2022.108265. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0038110122000375.

[19] F. Klemme, S. Salamin, and H. Amrouch, "Upheaving self-heating effects from transistor to circuit level using conventional eda tool flows," in *2023 Design, Automation Test in Europe Conference Exhibition (DATE)*, 2023, pp. 1–6. DOI: 10.23919/DATE56975.2023.10137162.

[20] F. Klemme and H. Amrouch, "Transistor self-heating-aware synthesis for reliable digital circuit designs," *IEEE Transactions on Circuits and Systems I: Regular Papers*, vol. 70, no. 12, pp. 5366–5379, 2023. DOI: 10.1109/TCSI.2023.3315293.

[21] C. Auth, C. Allen, A. Blattner, *et al.*, "A 22nm high performance and low-power CMOS technology featuring fully-depleted tri-gate transistors, self-aligned contacts and high density MIM capacitors," in *2012 Symposium on VLSI Technology (VLSIT)*, 2012, pp. 131–132. DOI: 10.1109/VLSIT.2012.6242496.

[22] M. G. Bardon, P. Schuddinck, P. Raghavan, *et al.*, "Dimensioning for power and performance under 10nm: The limits of FinFETs scaling," in *2015 International Conference on IC Design & Technology (ICICDT)*, 2015, pp. 1–4. DOI: 10.1109/ICICDT.2015.7165883.

[23] S. Panth, K. Samadi, Y. Du, and S. K. Lim, "High-density integration of functional modules using monolithic 3d-ic technology," in *2013 18th Asia and South Pacific Design Automation Conference (ASP-DAC)*, 2013, pp. 681–686. DOI: 10.1109/ASPDAC.2013.6509679.

[24] N. Ramaswamy, A. Calderoni, J. Zahurak, *et al.*, "Nvdram: A 32gb dual layer 3d stacked non-volatile ferroelectric memory with near-dram performance for demanding ai workloads," in *2023 International Electron Devices Meeting (IEDM)*, 2023, pp. 1–4. DOI: 10.1109/IEDM45741.2023.10413848.

[25] T. S. Soliman, S. Chatterjee, N. Laleni, *et al.*, "First demonstration of in-memory computing crossbar using multi-level cell fefet," *Nature Communications*, vol. 14, p. 6348, 1 2023. DOI: 10.1038/s41467-023-42110-y.

[26] P. R. Genssler, V. M. van Santen, J. Henkel, and H. Amrouch, "On the reliability of fefet on-chip memory," *IEEE Transactions on Computers*, vol. 71, no. 4, pp. 947–958, 2022. DOI: 10.1109/TC.2021.3066899.

[27] E. Park and T. Song, "Complementary FET (CFET) Standard Cell Design for Low Parasitics and Its Impact on VLSI Prediction at 3-nm Process," *IEEE Transactions on Very Large Scale Integration (VLSI) Systems*, vol. 31, no. 2, pp. 177–187, 2023. DOI: 10.1109/TVLSI.2022.3220339.

[28] *BSIM–CMG Technical Manual*. [Online]. Available: http://bsim.berkeley.edu/models/bsimcmg/.

[29] W. Chakraborty, H. Ye, B. Grisafe, I. Lightcap, and S. Datta, "Low thermal budget (¡250 °c) dual-gate amorphous indium tungsten oxide (iwo) thin-film transistor for monolithic 3-d integration," *IEEE Transactions on Electron Devices*, vol. 67, no. 12, pp. 5336–5342, 2020. DOI: 10.1109/TED.2020.3034063.