IEEE Open Journal of
Circuits and Systems

Received 8 June 2023; revised 10 August 2023; accepted 23 August 2023. Date of publication 29 August 2023; date of current version 19 September 2023.

Digital Object Identifier 10.1109/OJCAS.2023.3309478

Cryogenic In-Memory Computing for Quantum Processors Using Commercial 5-nm FinFETs

SHIVENDRA SINGH PARIHAR[1,2] (Member, IEEE), SIMON THOMANN[1] (Member, IEEE), GIRISH PAHWA[3] (Member, IEEE), YOGESH SINGH CHAUHAN[2] (Fellow, IEEE), AND HUSSAM AMROUCH[1,4,5] (Member, IEEE)

[1]Semiconductor Test and Reliability, University of Stuttgart, 70174 Stuttgart, Germany
[2]Department of Electrical Engineering, Indian Institute of Technology Kanpur, Kanpur 208016, India
[3]Department of Electrical Engineering and Computer Sciences, University of California at Berkeley, Berkeley, CA 94720, USA
[4]Munich Institute of Robotics and Machine Intelligence, Technical University of Munich, 80333 Munich, Germany
[5]Chair of AI Processor Design, Technical University of Munich, 80333 Munich, Germany

This article was recommended by Associate Editor A. James.

CORRESPONDING AUTHOR: S. S. PARIHAR (e-mail: parihasa@iti.uni-stuttgart.de)

This work was supported by the German Research Foundation (DFG) "Open Access Publication Funding/2023-2024/ University of Stuttgart" under Grant 512689491.

ABSTRACT Cryogenic CMOS circuits that efficiently connect the classical domain with the quantum world are the cornerstone in bringing large-scale quantum processors to reality. The major challenges are, however, the tight power budget (in the order of milliwatts) and small latency (in the order of microseconds) requirements that such circuits inevitably must fulfill when operating at cryogenic temperatures. In-memory computing (IMC) is rapidly emerging as an attractive paradigm that holds the promise of performing computations efficiently where the data does not need to move back and forth between the CPU and the memory. Hence, it overcomes the fundamental bottleneck in classical von Neumann architectures, which provides considerable savings in power and latency. In this work, for the first time, we propose and implement an end-to-end approach that investigates SRAM-based IMC for cryogenic CMOS. To achieve that, we first characterize commercial 5 nm FinFETs from 300 K down to 10 K. Then, we employ the first cryogenic-aware industry-standard compact model for the FinFET technology (BSIM-CMG) to empower SPICE to accurately capture how cryogenic temperatures alter the electrical characteristics of transistors (e.g., threshold voltage, carrier mobility, sub-threshold slope, etc.). Our key contributions span from (1) carefully calibrating the cryogenic-aware BSIM-CMG against commercial 5 nm FinFET measurements in which SPICE simulations come with an excellent agreement with the experimental data, (2) exploring how the robustness of SRAM cells against noise (during the hold, read, and write operations) changes at extremely low temperatures, (3) investigating how the behavior of SRAM-based IMC circuits changes at 10 K compared to 300 K, and (4) modeling the error probabilities of IMC circuits that are used to calculate the Hamming distance, which is one of the essential similarity calculations to perform classifications.

INDEX TERMS SRAM, cryogenic, quantum computing, transistor modeling, in-memory computing.

I. INTRODUCTION

QUANTUM computing promises to resolve a wide range of computational problems that are fundamentally challenging, if not impossible, to be resolved by classical computing. Synthesizing new materials, optimizing drugs [1], and notably, simulating quantum systems [2] are examples of the types of problems that could be superbly tackled by quantum computing, potentially reshaping the future of mankind. Nevertheless, for this to happen, a large number of high-fidelity qubits are indispensable, and hence, quantum computer up-scaling becomes a necessity. However, this demands CMOS-based compute circuits that effectively connect the classical domain (where information is processed) and the quantum domain (where qubits are present).

These circuits are prerequisites in such scaled-up quantum computers as they are responsible for: (1) processing the measurements received from the qubits, (2) performing the required classification to translate the readout data to the digital world as well as (3) performing the necessary error corrections [3], [4].

The inevitable need for cryogenic circuits: Today, qubits operate at near absolute zero (e.g., 10 mK) to ensure they stay in the required superimposed state for as long as possible. Such a coherence time is often short (ranging from nanoseconds to milliseconds) and, more importantly, extremely sensitive to noise and heat due to the fragility of qubits. On the other hand, control circuits currently operate at room temperature (i.e., 300 K), which causes a profound input-output bottleneck for existing quantum computers. This is further exacerbated by the fact that every qubit might even need to be individually controlled [5]. This overwhelming problem has been recently exemplified by state-of-the-art experiment that demonstrated the need for approximately 200 wide-band coaxial cables along with 45 bulky microwave circulators and a rack of electronic circuits to control merely 53 qubits [5], [6]. Despite the isolation, a significant temperature gradient (300 K $\leftrightarrow$ 0.1 K) induced at the two ends of every wire creates a heat flux that might still leak from the control circuits (outside the fridge) towards the qubits (inside the fridge) jeopardizing the entire quantum system. Further, timing constraints are already tightened by the short coherence time of qubits. As long cables introduce large latencies, satisfying the timing constraints can be a challenge or an entirely infeasible task in the worst-case scenario. Hence, it becomes inevitable to move the CMOS circuits from room temperature down to cryogenic temperatures to locate them as close as possible to qubits. Otherwise, scaled-up quantum computers, where a large number of qubits are being coherently and reliably operated, are not possible.

The key challenges behind cryogenic circuits: Operating CMOS circuits at cryogenic temperatures impose tough power constraints on the circuits. This is because of the highly limited power dissipation capability of these circuits that becomes consequential at extremely low temperatures. For instance, at 10 K and 0.1 K, the control circuits must operate within a power budget that does not exceed merely 100 mW and 10 mW, respectively [7]. Otherwise, the generated heat can rapidly disturb the state of qubits and even destroy them. Therefore, when operating CMOS circuits at cryogenic temperature, power optimization becomes the primary goal. In addition, the latency of the performed computing must be short enough to meet the tight timing constraints imposed by the short coherence time of qubits.

In a nutshell, cryogenic circuits required for quantum processors must not only be ultra-low power but also fast.

The need for cryogenic-aware compact model: Current commercial SPICE tools and compact models are not yet aware of the fundamental changes that cryogenic temperatures cause in the underlying semiconductor physics governing CMOS transistors, and the research is still in its infancy. Some key changes caused at cryogenic temperatures are as follows: leakage current decreases, transistor sub-threshold swing decreases, and carrier mobility improves while transistor threshold voltage increases. This makes the existing compact models lack the necessary information on how the key electrical characteristics of p- and n-FinFET are impacted at extremely scaled-down operating temperatures.

The promise of in-memory computing: Classical computing using the existing von-Neumann architecture inherently suffers from significant power and latency overheads due to the physical separation between the computing units and memory units. On the contrary, In-memory Computing (IMC) is rapidly emerging as an attractive alternative in which the memory is augmented by a certain capability to perform some types of computation. This eliminates the fundamental bottleneck and considerably accelerates computation while reducing the consumed power. IMC can be realized using either classical CMOS-based SRAM memories [8], [9], [10], [11], [12], [13], [14], [15], [16] or emerging beyond-CMOS nonvolatile memories such as ferroelectric FET [17], [18], [19], [20], [21], [22]. Unlike beyond-CMOS technologies, which are still in their infancy, classical CMOS-based SRAMs are much more mature and suffer significantly less from process variation and variability effects. Therefore, in this work, we focus on implementing IMC circuits using conventional SRAMs.

Our main contributions within this paper are:

(1) We characterize commercial 5 nm FinFETs from 300 K all the way down to 10 K. We then employ the obtained measurements to carefully calibrate the first cryogenic-aware industry-standard compact model for FinFET technologies for accurate SPICE simulations.

(2) Our calibrated FinFET compact model (which empowers SPICE to come up with an excellent agreement with the experimental data) is then used to analyze how the resiliency of 6-T SRAMs against noise (during the hold, read, and write operations) from 300 K down to 10 K.

(3) We investigate how the behavior of SRAM-based IMC circuits changes at extremely low temperatures and how errors in computations can emerge.

(4) We model the induced error probabilities of IMC circuits (when used to calculate the Hamming distance) at 10 K compared to 300 K, revealing how cryogenic temperature can affect the reliability.

II. RELATED WORK

In-memory computation increases energy efficiency by performing computations within the memory itself instead of transferring enormous amounts of data back and forth to the processing units. Recently, there has been significant research interest in using IMC to perform data-intensive computations, e.g., in deep neural networks and machine learning [8], [9], [10], [11]. It has been studied for both the conventional CMOS [8], [9], [10], [11], [12], [13], [14], [15], [16] and beyond CMOS devices [17], [18], [19], [20], [21], [22]. However, only a few works have studied SRAM

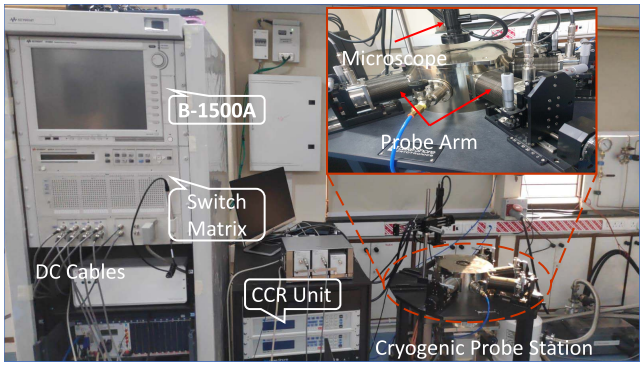


FIGURE 1. On-wafer Lakeshore's Cryogenic Probe Station. During measurements, B-1500A and CCR units are used to precisely control the voltages and temperature, respectively.

in the context of IMC at cryogenic temperatures. Authors in [8], [9] analyzed the IMC-based deep neural network and convolutional neural network performance at cryogenic temperatures using the 28 nm and 55 nm CMOS technology, respectively. In [10], the authors propose to implement the surface code for Quantum Error Correction (QEC) using the SRAM-based IMC. However, these studies were limited to the older generation of CMOS technologies. Although authors in [11] reported the SRAM IMC macro on the 7 nm technology platform, they performed the evaluation only at 300 K. In another recent publication, X-SRAM, a modified variant of conventional SRAM, was introduced by [16] for performing IMC. However, the study primarily relied on a predictive technology modelcard for bulk MOSFET and restricted the analysis to temperatures of 300 K and above. Previous studies were carried out either with emerging memory technologies or older-generation CMOS technologies. SRAM is a more mature memory technology and well-suited for quantum processors due to its compatibility with the cryogenic CMOS circuitry [4], [23].

All in all, in this work, we present the 5 nm FinFET SRAM-based IMC using the Ternary Content Addressable Memory (TCAM) and X-SRAM arrays. We also investigate the impact of process variations for both cryogenic and room temperatures. Our study spans from the transistor level to the circuit level all the way to the error probabilities that can be induced at the system level.

III. CRYOGENIC MEASUREMENT SETUP

This work presents the measurement of the minimum channel length, multi-fin, multi-finger FinFETs of a commercial 5 nm FinFET technology. We use a cryogenic probe station called "Lakeshore CRX-VF" to perform the on-wafer DC measurements in the temperature range of 10 K to 300 K Fig. 1). The primary components of the probe station are a two-stage closed cycle refrigerator (CCR), tungsten probes, probe positioners, the vacuum pump, sample stage, and microscope. The first stage of CCR unit cools the probes. The second stage cools the sample stage. With the help of temperature sensor and CCR unit, the probe station automatically cools the ambient temperature down to desired cryogenic temperatures.

In order to minimize the chance of condensation, we keep the sample at a higher temperature during the cooling process. As the probe arm temperature can go as low as 15 K, the thermal load from the probe arms limits the sample stage temperature up to 3.5 K to 8.5 K. To minimize temperature fluctuations during the measurements, the probe station operates till the lower limit of 10 K. Once we set the cryogenic temperature for measurement, the CCR unit starts the cooldown process. The measurement requires voltage/current sources and meters with the smallest noise floor possible. We use Keysights' B1500A parameter analyzer with high-resolution and medium-power source measurement units (SMUs). To avoid human interference in the measurement process, we control all the measuring instruments with a computer using the GPIB interface.

IV. 5 NM FINFET COMPACT MODEL CALIBRATION FROM ROOM TEMPERATURE DOWN TO CRYOGENIC TEMPERATURE

The electrical characteristics of semiconductor devices, such as transistors, diodes, MOS varactors, etc., are highly temperature-dependent due to the temperature-dependent charge carrier concentration and mobility. Hence, setting the appropriate simulation environment becomes the most important step for extracting compact model parameters. To begin the model parameter extraction, we first set the appropriate temperature environment in every simulation setup using the SPICE keyword TEMP (the ambient temperature for each simulation) and the model parameter TNOM (transistor's nominal temperature). We have utilized a modified cryogenic-aware FinFET compact model to capture the transistor electrical characteristics from 300 K to 10 K [24], [25]. The following subsection describes the model extraction procedure for room and cryogenic temperatures.

A. TRANSISTOR MODEL CALIBRATION

After setting the correct simulation environment, we extract the process-dependent model parameters, e.g., doping, oxide thickness, and gate material work function. Interface trap charges and source-drain coupling substantially influence the sub-threshold behavior of the transistors. Using the BSIM-CMG model parameters PHIG, CIT, and CDSC, the simulations accurately imitate room temperature sub-threshold characteristics (300 K) of the measured FinFETs [25]. We extract the low field mobility and field-dependent mobility degradation parameters (i.e., U0, UA, UD, EU, and ETAMOB) from the transfer characteristics ($I_{DS} - V_{GS}$) when the transistor operates at low drain voltage (V_{DS}) and moderate inversion Fig. 2(a)). On the other hand, we extract the series resistances model parameters (RSW, RDW, RSWMIN, and RDWMIN) from the strong-inversion regime (higher gate voltage). The model parameters ETA0, PDIBL2, and CDSCD, capture the impact of Drain-Induced Barrier Lowering (DIBL). To extract the influence of DIBL, we focus on the sub-threshold region of operation and optimize the above-mentioned model parameters by simultaneously

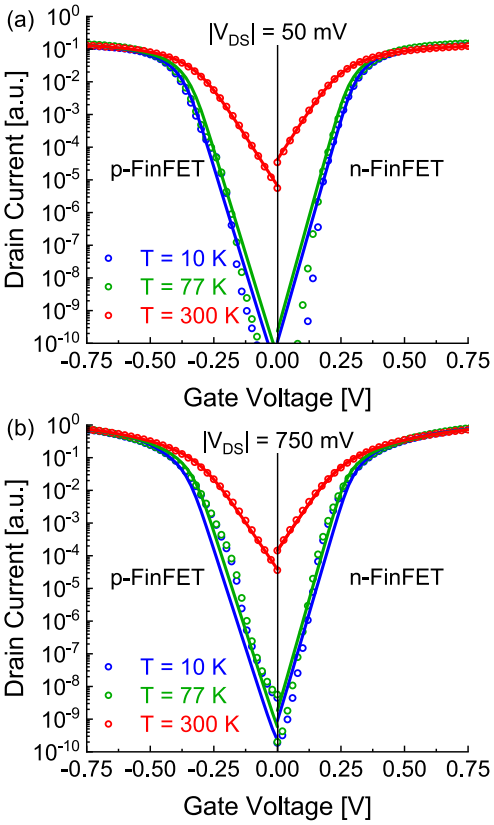


FIGURE 2. Transfer characteristics of p- and n-FinFET for multiple temperatures ranging from 10K to 300K in (a) Linear (V_{DS} = 50mV) and (b) Saturation (V_{DS} = 750mV). Symbols and lines show the data from measurements and calibrated model simulations.

observing the $I_{DS} - V_{GS}$ Fig. 2(b)) characteristics at lower and higher V_{DS}. With V_{DS} increase, carrier velocity begins to saturate, and transfer ($I_{DS} - V_{GS}$) and output characteristic ($I_{DS} - V_{DS}$) show a very small increase in drain current with a further increase in V_{DS}. The velocity saturation model parameters VSAT, VSAT1, MEXP, and KSATIV accurately capture this effect. At higher V_{DS} and gate voltage (V_{GS}), we extract the impact of velocity saturation and channel length modulation by minimizing the error between measurement and simulation data of $I_{DS} - V_{GS}$ and $I_{DS} - V_{DS}$.

The cryogenic operation of MOS transistors improves the transistor performance due to a reduction in carrier scattering [24]. Because of the decrease in carriers' thermal energy at cryogenic temperatures, a lower over-the-barrier transport reduces Sub-threshold Swing (SS) and OFF-state current (I_{OFF}). Consequently, silicon-based transistor characteristics differ considerably at cryogenic temperatures from 300 K. Some dominant effects at cryogenic temperatures are as follows: nonlinear temperature-dependence in SS characteristics, increase in threshold voltage (V_{TH}), surface roughness scattering, coulomb scattering, and nonlinear velocity saturation effect [24], [26], [27]. To account for these effects in SPICE simulations of FinFETs, we use the model equations presented in [24] along with the industry-standard BSIM-CMG compact model [25]. As the existing BSIM-CMG model is based on Maxwell-Boltzmann (MB) statistics, we use the MB statistics, along with the modifications presented in [24]. This allows us to capture the impact of Fermi-Dirac (FD) statistics from 300 K to cryogenic temperatures for electron density calculation. Since the effective density of states, surface potential, and charges are highly temperature dependent, we first extract an effective temperature (T_{eff}) at cryogenic temperatures from the SS behavior with respect to temperature [24]. The SS deviates from the Boltzmann limit of kT/q in the cryogenic range, which is caused by the presence of band-tail states [26], [27]. To model the SS saturation, the ambient temperature (T_{amb}) is smoothly clamped to the temperature below which SS starts saturating using the Equation (1) [24], which is subsequently employed to determine the total charge density.

$$T_{eff} = \frac{T_{amb} + T_0 + \left[(T_{amb} - T_0)^2 + 0.25 \cdot D_0^2\right]^{0.5}}{2} \quad (1)$$

Here, T_0 is the T_{amb} at which SS saturates, and D_0 is a smoothing parameter. To model the impact of an increase in V_{TH} due to band-tail states and an increase in trap states below Fermi energy level, a V_{TH} correction factor is introduced into the existing BSIM-CMG with KT11, KT12, and TVTH as the model parameters [24]. The temperature-dependent lattice vibrations decrease at cryogenic temperatures and thereby improve peak mobility. Nevertheless, as the temperature drops, the thermal velocity of the charge carriers also decreases. At cryogenic temperatures, these low thermal velocity carriers experience increased Coulomb and surface roughness scattering, which decreases their effective mobility at very low and high vertical electric fields, respectively. To account for the temperature dependence of these mobility components, the current mobility models, which rely on simple power-law relationships, are substituted with a linear temperature-dependent power law formulation, as demonstrated in Equation (2) [24].

$$\mu_P(T) = \mu_P(T_{nom}) \cdot \left[\frac{T_{amb}}{T_{nom}}\right]^{\mu_P 1 + \mu_P 2(T_{amb}/T_{nom})} \quad (2)$$

Here, μ_P refers to different mobility model parameters, namely U0, UA, and UD. μ_{P1} and μ_{P2} represent temperature-independent parameters of U0, UA, and UD. The influence of surface roughness on mobility at higher electric fields is taken into account by EU, which is modeled as a linear temperature-dependent parameter. The impact of Coulomb and surface roughness scattering at cryogenic temperatures has been incorporated into the extracted mobility model by optimizing the following temperature coefficient of the mobility model: UD1, UD2, UA1, UA2, EU1, etc. To account for the non-linear temperature dependency of the saturation velocity, the parameters related to velocity saturation effects, such as VSAT and MEXP, are modeled using Equation (3). This formulation enables a more accurate representation of the impact of temperature on the saturation velocity [24].

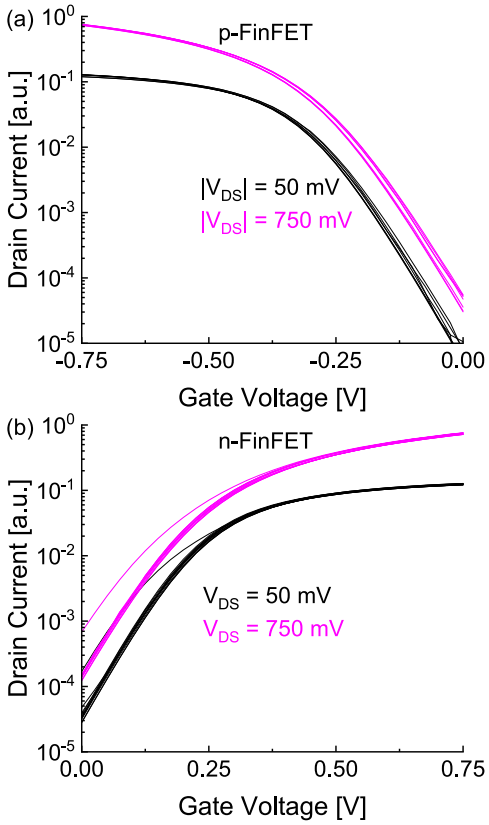


FIGURE 3. (a and b) present the measurement results of p-FinFET and n-FinFET, respectively, and show the impact of process variations on $I_{DS} - V_{GS}$ at 300K. The measurements are done for both linear (V_{DS} = 50mV) and saturation (V_{DS} = 750mV).

$$P(T) = P(T_{nom}) + P_T(T_{amb} - T_{nom}) + P_{T1}(T_{amb} - T_{nom})^2 + P_{T2}(T_{amb} - T_{nom})^3 \quad (3)$$

In Equation (3), the parameters P_T, P_{T1}, and P_{T2} correspond to AT, AT1, and AT2, respectively, for the temperature dependence of both VSAT and VSAT1. Similarly, for the temperature dependency of MEXP, these parameters represent TMEXP, TMEXP1, and TMEXP2, respectively. Lastly, the parameters P_T, P_{T1}, and P_{T2} represent KSATIVT, KSATIVT1, and KSATIVT2 for the temperature dependence of KSATIV. Fig. 2 presents the model validation with experimental data from 10 K to 300 K. At lower gate voltage, the measurement's intrinsic randomness causes noisy data. Noise in the measurements at low currents is the likely cause of discrepancies between the simulated and measured results.

B. IMPACT OF VARIATION ON TRANSISTOR PERFORMANCE

The characteristics of bulk FinFETs are significantly affected by different sources of fluctuation, including interface traps, work function variation, process variations, and random dopants [28]. The impact of device variability is becoming more significant in advanced FinFET technologies and circuits as a result of the stringent circuit design margin driven by the constant down-scaling of transistor dimensions and supply voltage. To characterize the effects of random variations, we have measured the DC transfer characteristics of FinFETs fabricated at different dies on a single wafer. Fig. 3(a) and 3(b) present the $I_{DS} - V_{GS}$ measurement results of p-FinFET and n-FinFET, respectively, operating in linear and saturation regions. We observe that variations impact the $I_{DS} - V_{GS}$ of p-FinFET and n-FinFET to different extents at different V_{DS}. For example, the p-FinFET shows V_{TH} fluctuation (σV_{TH}) of 5 mV and 26 mV for linear and saturation while n-FinFET exhibits σV_{TH} of 9 mV and 18 mV.

V. SRAM RELIABILITY ANALYSIS

The nanoscopic dimensions and complex fabrication process of state-of-the-art FinFET technologies have significantly increased the impact of process variations and inhibited performance improvement. The quantized transistor dimensions in the High-Density Cell (HDC) exacerbate variability-induced yield reduction. The following operational issues in SRAM govern the yield: (1) Read failure, defined as switching the cell content during reading (2) Write failure, the inability to write into a cell (3) Hold failure is flipping the cell state while holding the data and (4) Access failure defined as the unacceptable increase in cell access time. In this work, we focus on the first three issues, i.e., hold, read, and write noise margins. For this purpose, the mismatch or statistical variability of the transistors is included in the SRAM cell analysis by incorporating the measured σV_{TH} in the BSIM-CMG compact model. Authors in [29] reported that silicon lattice constant experiences a mere 0.022 % change at cryogenic temperatures. Therefore, in this work, we assume the lattice constant of silicon to be constant across all temperatures. As the process variations arise during the device fabrication, with the assumption mentioned above, geometry variation will also not change much at cryogenic temperatures. Previous works also report negligible temperature dependence on V_{TH} variability [30], [31], [32]. Hence, we apply the σV_{TH} of p-FinFET, and n-FinFET extracted at 300 K in cryogenic SRAM simulations. We use the Monte-Carlo analysis in our SRAM framework to characterize the variability impact on the Static Noise Margin (SNM) of the HDC. The Monte-Carlo simulations were performed for varying V_{DD} and at multiple temperatures ranging from 10 K to 300 K to assess the impact of V_{DD} and temperature variations.

The schematic view of the conventional six-transistor SRAM (6T-SRAM) cell comprised of two access transistors (PG1 and PG2) and one cross-coupled inverter pair is shown in Fig. 4. The inverter pair in the SRAM keeps the cell in a bi-stable state while holding the data stored inside it. Voltage fluctuation due to the noise at the input node of inverters degrades the ability of the cell to store the data. The SRAM cell can withstand a certain voltage noise during the hold operation before switching internal states, defined as Hold Noise Margin (HNM). To read the data stored in the cell, we pre-charge the bit line (BL) up

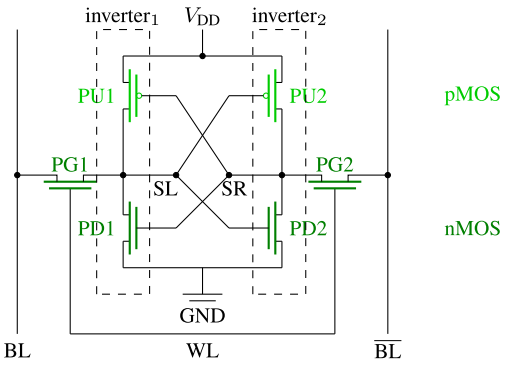


FIGURE 4. Schematic view of the 6T-SRAM cell used later in the TCAM array.

to V_{DD}. The pre-charged BL starts discharging after Word Line (WL) activation if the data stored in the cell (voltage at node SL) is 0. The maximum amount of static noise at which a cell can retain the stored data during a read operation is known as the Read Noise Margin (RNM). The lowest BL voltage needed to modify the internal cell state during the write operation is called Write Noise Margin (WNM). During the write operation, initially, the data that needs to be written is transferred on the BL and complement bit line (BLB). Subsequently, access transistors are turned ON by WL activation to access the internal storage node of the cell.

Fig. 5 shows the statistical spread in noise margins for the HDC from 10 K to 300 K, at nominal supply voltage ($V_{DD,nom}$ = 0.75 V). Cryogenic temperatures result in a lower I_{OFF} and improve the data retention capacity of SRAM cells. Fig. 5(a) shows that the HNM at 10 K has a higher mean (μ) value than at 300 K, which reflects the increase in HNM at cryogenic temperatures.

In a read operation, n-FinFET of the pull-down-network (PD2) and access transistor (PG2) form a voltage divider, and the voltage of node SR (V_{SR}) starts increasing. This leads to an increase in the sub-threshold leakage current of the n-FinFET (PD1) and lowers the voltage at node SL (V_{SL}). It further results in an increase in V_{SR}, such that if V_{SR} exceeds the switching threshold of the left side inverter, then the SRAM cells' state flips. This results in a destructive or unstable read operation. Therefore, the worst-case scenario due to process variations for read stability arises when the access transistor (PG2) and pull-up transistor (PU2) become strong and PD2 becomes weak. Fig. 5(b) presents variations' impact on RNM at three different temperatures. Since at cryogenic temperatures, the leakage current is considerably small compared to 300 K, the read stability of the cell improves, and we observe a relative increase in the RNM at cryogenic temperatures Fig. 6(a)). However, as SS starts saturating below 77 K Fig. 2), further lowering the temperature does not show the same amount of improvement in I_{OFF} and subsequently in noise margins, as we observe from 300 K to 77 K. Cryogenic temperature also results in higher V_{TH}, which increases the on-resistance (R_{ON}) of the PD2 and PG2 transistors. The rise in R_{ON} of PD2 causes an increase in

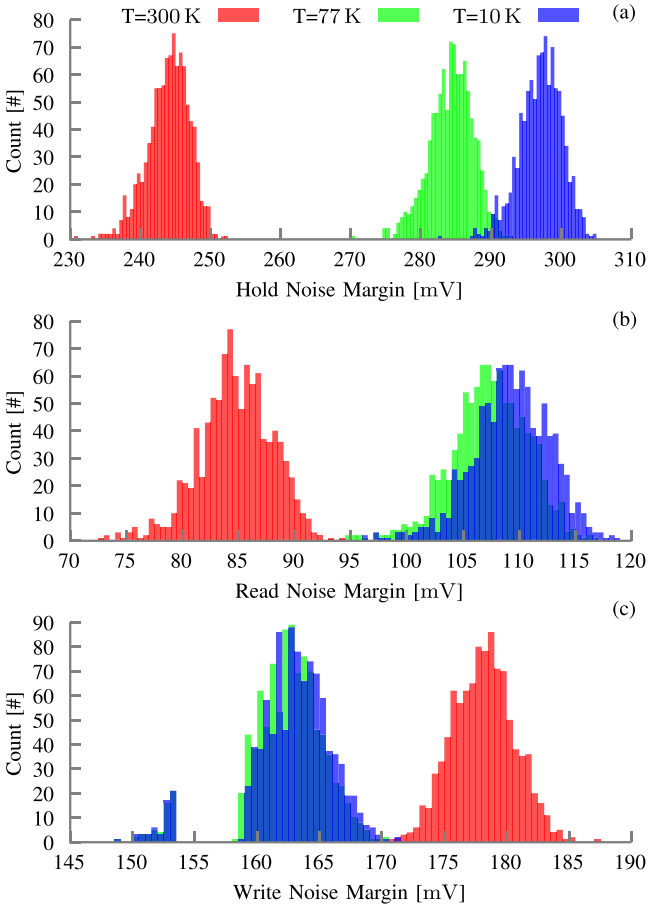


FIGURE 5. Impact of process variations on High-Density Cell SRAM (a) hold (b) read and (c) write noise margin for temperatures ranging from 10K to 300K at V_{DD} = 0.75V.

V_{SR}, and higher V_{TH} of PG2 decreases the current flowing into the node V_{SR}. The hindrance imposed by PG2 in the current flow is more crucial and further leads to an increase in read stability.

During the write operation (e.g., writing '0' in the cell), PU1 and PG1 conduct. In this case, if V_{SL} falls below the right side inverter's switching threshold, a successful write takes place. At room temperature, a higher drain current helps in the discharging of V_{SL}. Hence, the WNM of the SRAM is larger at 300 K compared to cryogenic temperatures (Fig. 5(c)). With an increase of V_{TH} at cryogenic temperatures, the switching threshold of the right side inverter of SRAM increases (improves the writing ability). However, higher V_{TH} of PG1 increases the V_{SL} and leads to degradation of write stability. As V_{TH} increase in the n-FinFET is higher compared to p-FinFET, we observe a stronger impact of PG1 on the WNM at cryogenic temperatures. Fig. 6(a) shows that the μ of WNM reduces from 178 mV to 162 mV for a temperature change from 300 K to 10 K. The HNM and RNM μ increases from 244 mV to 297 mV and 84.7 mV to 109.2 mV, respectively, with temperature decrease. Fig. 6(b) and 6(c) show that the variation in the temperature has a negligible impact on the normalized standard deviation (σ/μ) for $V_{DD,nom}$. However, at

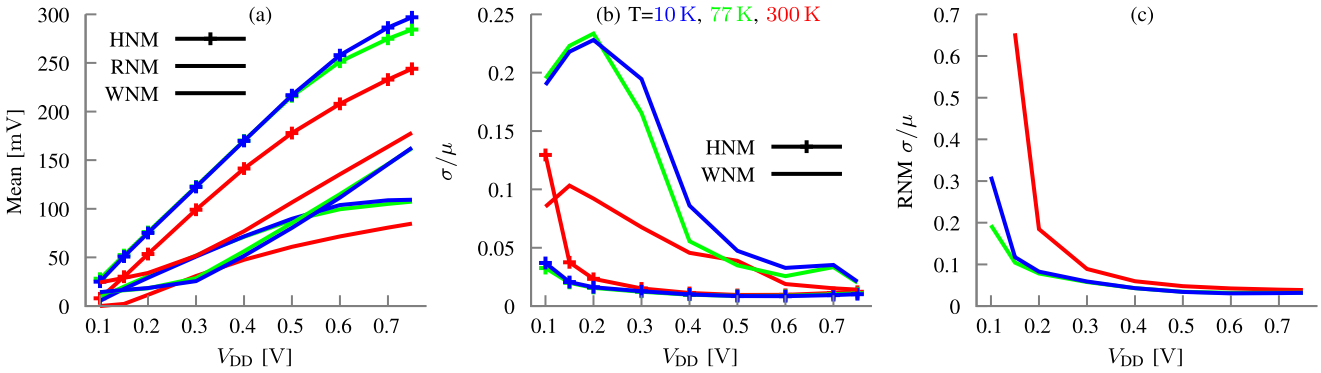


FIGURE 6. (a) Impact of temperature and supply voltage on Mean value of the SNM. The normalized standard deviation for (b) hold and write noise margin and (c) read noise margin.

TABLE 1. SNM of SRAM for different temperatures at $V_{DD} = 0.75$ V

Noise Margin	T = 10K			T = 77K			T = 300K		
	Mean μ (mV)	SD σ (mV)	σ/μ (%)	Mean μ (mV)	SD σ (mV)	σ/μ (%)	Mean μ (mV)	SD σ (mV)	σ/μ (%)
HNM	297.00	3.02	1.01	284.00	3.11	1.09	244.00	2.98	1.22
RNM	109.20	3.45	3.16	107.30	3.42	3.18	84.70	3.26	3.84
WNM	162.00	3.41	2.09	162.00	3.25	2.00	178.00	2.50	1.40

lower supply voltage (V_{DD}), we observe a minor temperature dependency on σ/μ. The percentage variation in σ/μ shows that the characterized SRAM cell at $V_{DD,nom}$ is highly resilient to temperature variations. Table 1 summarises the impact of process variations and temperature on the SRAM SNM.

VI. SRAM IN-MEMORY COMPUTING

A. TCAM ARRAYS FOR HAMMING DISTANCE COMPUTATION

1) SINGLE TCAM CELL

To implement a single TCAM cell, 4 CMOS-transistors along with two Static Random Access Memory (SRAM) cells (S1 and S2) are used as shown in Fig. 7(a) [33]. The data of a TCAM cell (C) is stored by the two SRAMs in a complementary fashion. For example, at $C = 1$, S1 and S2 are in the logical 1 and 0 states, which are read out at the labeled nodes ('L' and 'R'), respectively. Although S1 holds 1, the read-out nodes are placed on the negated side to ensure correct functionality. Therefore, in this example, 'L' holds the inverted value of S1, i.e., 0. Before the lookup, the Match Line (ML) is pre-charged to V_{DD}. Then the query data Q is applied to Query Line (QL) (corresponding to left/S1) and inverted data to Query Line Bar (QLB) (corresponding to right/S2). When $C = Q$ *match*, the inverted 'L' and QL are complementary ($\overline{C} \neq Q$); therefore, no conductive path forms from ML to GND. Similarly, the non-inverted 'R' and inverted QLB ($C \neq \overline{Q}$) are complementary on the right-hand side. In this case, the TCAM cell is OFF as both discharge paths are blocked, and the voltage stays high (output 1). In the case of a *miss*, either the left or the right discharge path is active as their associated transistors are 1 at the same time, forming a conductive path to discharge ML. The TCAM cell is ON, and the output is 0.

2) TCAM ARRAY FOR HAMMING DISTANCE CALCULATION

Using the TCAM cells and combining them via a shared ML forms a *block*. Inside such a block, all TCAM cells share the same periphery and access logic as shown in Fig. 7(c). This includes the Bit Line (BL) and Word Line (WL) to write the data into the cells. Typically, the write operation is a one-time initialization phase for associative memories, whereas this work focuses on the read-out. Hence, we exclude the write operation from the evaluation, and the SPICE circuit implementation simplifies by using individual voltage sources for BL and WL.

The shared ML of all TCAM cells within a single block is an integral part of the block design. After the ML is pre-charged, a query bit string is applied through the respective lines to the block. Subsequently, each individual TCAM cell compares stored and query data as described in Section I. A miss leads to a conductive path and discharges ML. Due to the parallel circuit configuration of the TCAM block, more misses form more parallel conductive paths, leading to a faster ML discharge as the total resistance decreases. Consequently, the discharge rate is proportional to the number of cells reporting a miss. Computing the sum of bit-wise similarity checks is more widely known as the Hamming distance and can be performed with such a TCAM block. To derive a sharp, distinct output signal depending on the discharge rate, the ML is connected to

TABLE 2. Performance comparison of TCAM circuits designed using different technologies.

Technology	Array Size	Temperature (K)	Read Latency (ps)	Maximum Error Probability
5 nm Bulk-FinFET	1 ×10	300	56	35 %
(This Work)	1 ×10	10	62	59 %
	1 ×10	10 (iso-I_{OFF})	47	12 %
14 nm Bulk-FinFET [12]	1 ×15	300	99	39 %
14 nm Fe-FinFET [12]	1 ×15	300	305	64 %
14 nm SOI-FinFET [13]	128 ×32	300	150	–
28 nm Bulk-CMOS [14]	1024 ×80	300	1700	–
65 nm Bulk-CMOS [15]	2048 ×72	300	1900	–

In [12] maximum detectable Hamming Distance (HD) is 7 bits. However, in our work, we present that 5 nm FinFET-based TCAM can successfully detect HD of up to 10 bits.

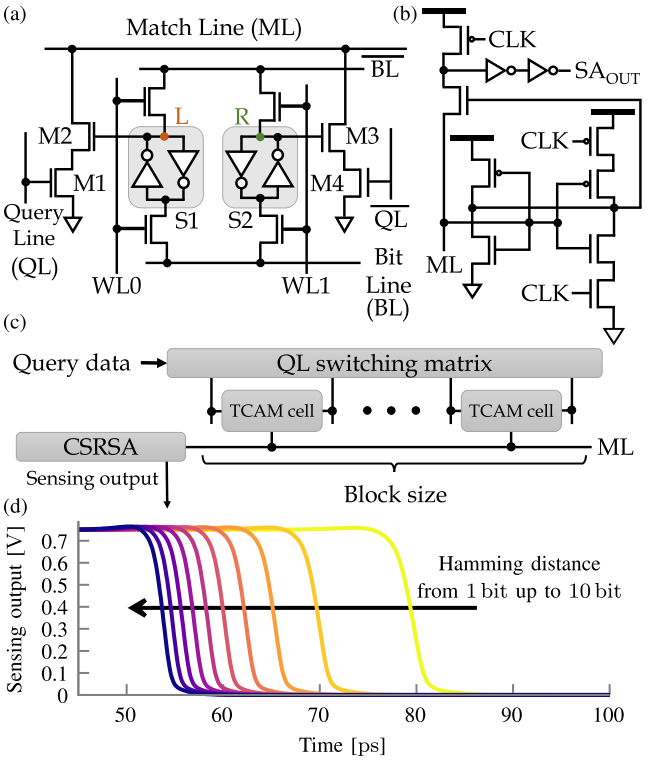


FIGURE 7. (a) Standard 16T TCAM cell schematic. (b) Clocked Self-Referenced Sense Amplifier (CSRSA) schematic, with CLK being the enable signal. (c) TCAM block with variable cell count. Only the query line (QL, input) and match line (ML, output) are drawn. (d) Output voltage waveforms of the CSRSA for a block size of 10 bit at room temperature.

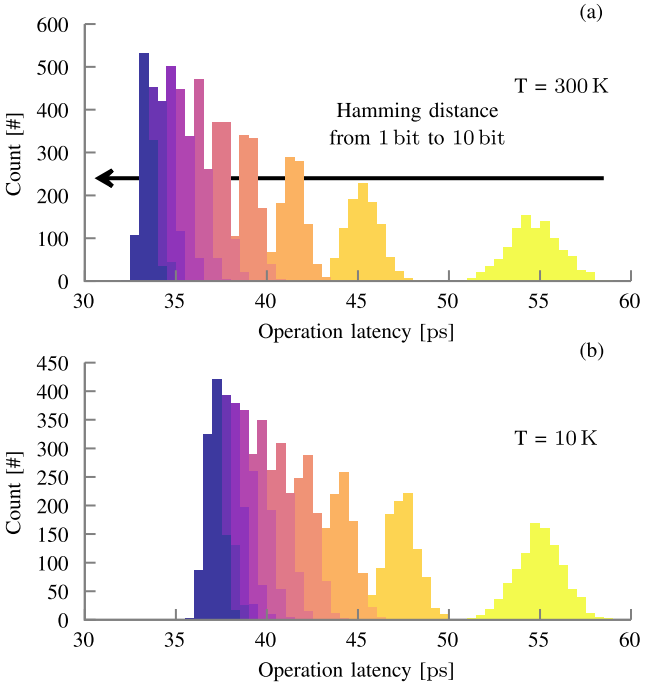


FIGURE 8. Impact of process variations on operation latency at (a) 300K, and (b) 10K. Simulation results are from a block size of 10 bit with the input signal (QL/QLB) rise time (T_{rise}) of 50 ps. We perform 1000 Monte-Carlo SPICE simulations for each Hamming distance.

a Clocked Self-Referenced Sense Amplifier (CSRSA) with its schematic shown in Fig. 7(b) [33]. It converts the discharge rate from the voltage domain to the temporal domain (i.e., from how fast to when the voltage drops). Thus, the operation latency determines the Hamming distance of the block, and an example of a block with 10 cells and all possible Hamming distances is shown in Fig.7(d). It can be observed that the margins between the misses decrease, which makes differentiating the individual misses increasingly harder. With the decreasing margins, the variability tolerance decreases, e.g., coming from process variations, which we will evaluate next.

B. SRAM-BASED IN-MEMORY BOOLEAN COMPUTATION

X-SRAM is an enhanced version of 8$^+$T SRAM that incorporates a modified peripheral circuit to perform IMC [16]. Fig. 9(a) and 9(b) illustrate the modified 8$^+$T SRAM and Sense Amplifier (SA), respectively. Unlike the conventional 6T SRAM cell, X-SRAM employs decoupled read-write ports, allowing simultaneous activation of two read word lines (RWLs) without any read disturb issues. During the read operation of X-SRAM, read bit-lines (RBL and RBLB) are first pre-charged to V_{DD}. Subsequently, the RWLs corresponding to the desired rows are activated. RBL or RBLB

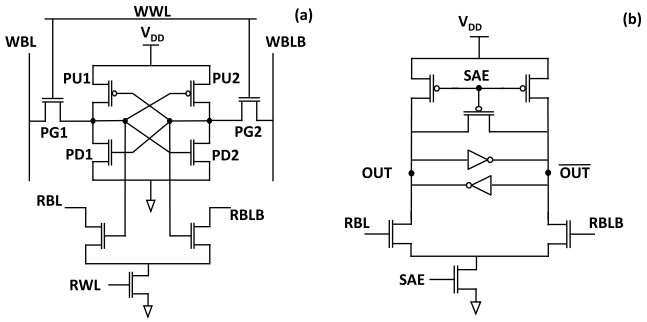


FIGURE 9. Circuit schematic of (a) the 8$^+$T SRAM and (b) asymmetric differential sense amplifier.

discharges depending on the stored value of the bit-cell (either "1" or "0"). An asymmetric differential SA is utilized to detect the voltage difference between RBL and RBLB, and further generate stable voltages at its outputs. To make the differential SA asymmetric, transistors connected at RBL and RBLB are made of different widths. This asymmetry leads to varying discharge rates of the SA output nodes (OUT and $\overline{OUT}$) based on the current-carrying capabilities of the transistors. This modification enables the implementation of bit-wise NAND and NOR operations using the stored data.

The Boolean operations within X-SRAM can be demonstrated through the following examples, where we discuss the OR/NOR operation. The read operation of X-SRAM is similar to a conventional 6T SRAM cell. To perform the read and IMC operation, first, we precharge both of the RBLs to V_{DD}, and next, we activate the RWLs of the respective cells, followed by enabling the SA. For IMC, two RWLs corresponding to two different bit cells are activated at the same time. In the case when both cells store logic "0" or "1", RBLB (and subsequently "OUT") or RBL (and subsequently "$\overline{OUT}$") discharges from V_{DD} (logic "1") to 0 V (logic "0"). However, when the stored data in the two rows differ (one storing "0" and the other storing "1," or vice versa), both RBL and RBLB discharge simultaneously. If we increase the width of the transistor in SA that is connected at RBLB, then the complementary output node of SA ($\overline{OUT}$) discharges faster, causing the SA output node (OUT) to stabilize at logic "1." Based on the aforementioned discussion, it can be concluded that when the transistor connected to RBLB in the SA is wider than the transistor connected to RBL, the SA generates the OR gate output at the node OUT and the NOR gate output at the node $\overline{OUT}$. Similarly, by making the transistor connected to RBL wider than the one connected to RBLB, the AND/NAND output of the stored values can be obtained. Furthermore, by incorporating an additional NOR gate at the AND/NAND and OR/NOR outputs of SA, the XOR operation can be performed on the stored values within the memory.

C. EVALUATION OF IMC TEMPERATURE DEPENDENCE

To evaluate the temperature impact on the described in-memory compute scheme, we apply process variations to the

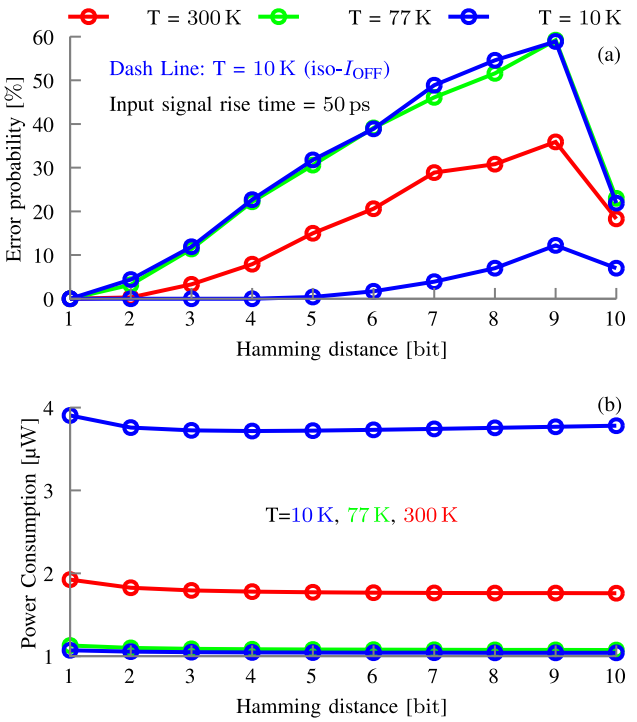


FIGURE 10. Impact of cryogenic temperatures on (a) error probabilities and (b) power consumption with T_{rise} of 50 ps for the input signals (QL/QLB). Dashed line with symbol shows the simulation result at 10K for iso-I_{OFF} condition. Here, iso-I_{OFF} refers to the condition when I_{OFF} of the transistor at 10K is equal to its I_{OFF} at 300K.

TCAM cell impacting their drive strength in the mismatch case. The variation in discharge rate due to process variations will consequently affect the operation latency of the CSRSA. Here, operation latency is the time interval within which the CSRSA is enabled, and its output reaches 10 % V_{DD}. Fig. 8(a) and 8(b) show the operation latency distribution per Hamming distance at 300 K and 10 K, respectively. While the heaps for the Hamming distance of 1 bit and 2 bit are clearly separated in Fig. 8(a), with further increase of the Hamming distance, the margins reduce, and the heaps start to overlap more.

1) IMPACT OF CRYOGENIC TEMPERATURES ON ERROR PROBABILITY IN TCAM ARRAY

To quantify the overlap between heaps of two Hamming distances, we calculate error probabilities in the Hamming distance. First, we use the nominal cases to place boundaries halfway between neighboring operation latencies. The boundaries serve as ranges for the latency intervals associated with the respective Hamming distances. We then sort the Monte–Carlo samples of each Hamming distance into the ranges and count how many samples are outside the correct range. The distributions of the Hamming distances are closer to each other at 10 K, as shown in Fig. 8(b), compared to the 300 K case shown in Fig. 8(a). Fig. 10(a) presents that the error probability increases with increasing Hamming distances. The error probabilities reach the maximum at the second to last Hamming distance. As the

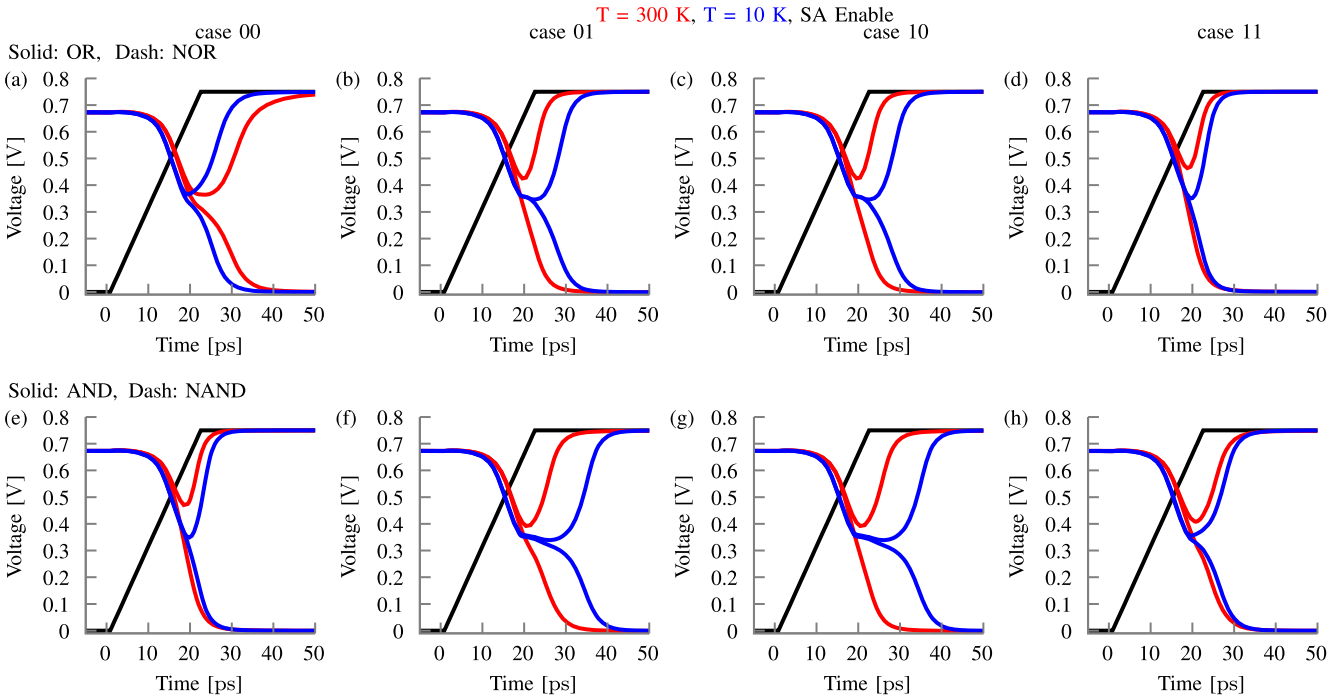


FIGURE 11. Impact of temperature on X-SRAM performance: (a)-(d) OR/NOR operation, (e)-(h) AND/NAND operation.

last Hamming distance is only bounded on one side, it can only overlap to one side, drastically reducing the error probability.

In the preceding paragraph, we provided an explanation of our methodology for calculating the error probability in the TCAM array. In the subsequent text, we will delve into the underlying factors that contribute to the increased error probability observed at cryogenic temperatures. The input signals are connected at QL and QLB, which act as the gate voltage of M1 and M4, respectively, as shown in Fig. 7. The rise time (T_{rise}) of the input signals (QL/QLB) decides the rate of increase in V_{GS} for M1 and M4. M1 (or M4) will turn on as soon as the input voltage reaches the V_{TH} of these transistors. After M1 (or M4) is turned on, there will be a significant amount of current flow in the discharge path, and ML will start discharging. As the discharge process is very fast, some of the discharge work is done while the input signals are still pulled up. Therefore the discharge process heavily depends on the rise time of the input signals and clock signal (CLK). Due to an increased V_{TH} at cryogenic temperatures, the I_{DS} at 10 K for $V_{GS} \leq 0.55$ V in saturation regime is smaller than the I_{DS} at 300 K Fig. 2(b), resulting in higher operation latency. This operation latency increases at lower temperatures, especially for higher Hamming distances Fig. 8(a) and 8(b). These higher latencies increase the overlaps between distributions of consecutive Hamming distances Fig. 8(b). This leads to a maximum error probability increase of 1.65 × at 10 K compared to 300 K Fig. 10(a).

2) IMPACT OF CRYOGENIC TEMPERATURES ON POWER DISSIPATION OF TCAM ARRAY

At cryogenic temperature, due to the Fermi-Dirac Statistics, the probability of finding an electron in the conduction band reduces significantly. Consequently, there are not enough high-energy electrons to climb the source-to-channel barrier, and hence, at a constant V_{GS}, there is a very small electron concentration in the conduction band. This reduction results in a substantial improvement in the OFF state leakage current of the transistors. The higher static power consumption is a major drawback in SRAM-based circuits. The significant reduction in the I_{OFF} at cryogenic temperatures helps in mitigating the static power consumption of SRAM cells and, subsequently, the total power consumption of SRAM-based TCAM array. Fig. 10(b) presents the total power consumed during the search/read operation of a 1×10 TCAM array. An approximately five-order of reduction in I_{OFF} at cryogenic temperatures results in ∼50 % improvement in power consumption.

3) IMPACT OF TEMPERATURE ON 8^{+}T SRAM ARRAY PERFORMANCE

In Section VI-B, we have discussed the architecture and working principle of X-SRAM to perform the basic Boolean operations. Here, we present the performance evaluation of 5 nm technology-based 1×32 X-SRAM array at both room and cryogenic temperatures. Fig. 11 illustrates the results of the SA output for different Boolean operations, including OR, NOR, AND, and NAND. It is observed that operating

TABLE 3. Comparison of the X-SRAM performance between room and cryogenic temperatures.

Operation	Stored Bits	Latency (ps)		
		T = 300 K	T = 10 K	T = 300 K [16]
IMC-based OR	00	39.75	26.75	1000.0
	01	24.05	30.15	1000.0
	10	24.05	30.15	1000.0
	11	22.14	24.43	1000.0
IMC-based NOR	00	42.66	28.53	1000.0
	01	23.67	29.29	1000.0
	10	23.67	29.29	1000.0
	11	21.93	23.03	1000.0
IMC-based AND	00	21.89	20.30	1000.0
	01	27.28	35.69	1000.0
	10	27.28	35.69	1000.0
	11	27.10	29.12	1000.0
IMC-based NAND	00	20.55	24.44	1000.0
	01	27.40	36.39	1000.0
	10	27.40	36.39	1000.0
	11	26.45	27.96	1000.0

at cryogenic temperatures results in increased latency for most of the IMC operations in X-SRAM. Initially, both outputs of SA (OUT and $\overline{\text{OUT}}$) decrease simultaneously until they surpass the threshold of the p-FinFET in the cross-coupled inverter pair. Beyond this point, the signals diverge, and the positive feedback loop in cross-coupled inverters guides them into their respective stable states. At cryogenic temperatures, a greater reduction is necessary (due to the increased threshold voltage), leading to increased delay. To compare ourselves with the existing literature, we present a comparison of operating latencies at 300 K in Table 3. The 5 nm technology-based X-SRAM is faster than the previously reported X-SRAM and is indeed suitable for the fast interface necessary for classification in quantum computers.

VII. DESIGN GUIDELINES FOR CRYOGENIC TEMPERATURES

In the previous section, we observed that lowering the temperature increases the operating latencies and error probability for the higher number of mismatch cases. In this Section, we present two methods to improve the reliability of the designed TCAM array.

A. TRANSISTOR OPTIMIZATION

At cryogenic temperatures, an increase in V_{TH} results in a higher overlap between two consecutive heaps of Hamming distances, as depicted in Fig. 8(b). However, this increase in V_{TH} can be mitigated through work function engineering. By doing so, one can achieve the iso-I_{OFF} operation at cryogenic temperature (i.e., I_{OFF} at a cryogenic temperature is similar to I_{OFF} at 300 K). Additionally, work function engineering results in an increase in the transistor's current at all V_{GS}, as illustrated in Fig. 12. Fig. 10(a) shows that when transistors operate at the iso-I_{OFF} condition, the error probability at 10 K significantly reduces. The higher current levels achieved through work function engineering contribute to lower operation latency and an error probability approximately 2.94 × lower than that observed at 300 K.

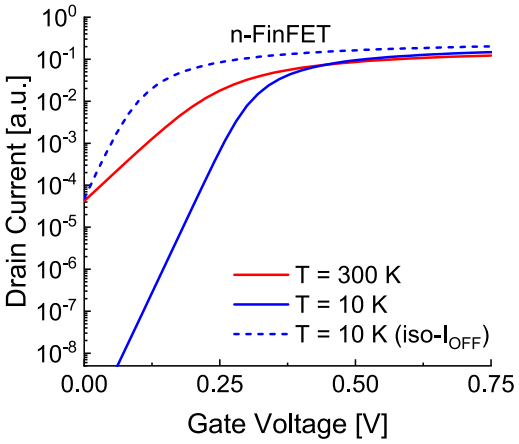


FIGURE 12. Transfer characteristics of n-FinFET for iso-I_{OFF} at 10K in Linear (V_{DS} = 50mV) regime.

B. SEARCH PULSE RISE TIME OPTIMIZATION

As the transistor I_{DS} has an exponential dependence on V_{GS} in the sub-threshold region and quadratic dependence in the strong inversion region, one can expect a higher impact of process variations in the sub-threshold region. In a TCAM cell, the ML starts discharging even before the voltage at QL or QLB reaches V_{TH} of M1 or M4, respectively. For a faster QL signal (smaller T_{rise}), ML discharging takes place when the transistor operates only around $V_{\text{GS}} = V_{\text{DD}}$. On the other hand, in the case of the slower QL signal (higher T_{rise}), the ML discharging process involves transistor operation from sub-V_{TH} to above-V_{TH} to strong inversion. Hence, we observe the higher impact of process variation in the case of a slower QL signal. The reduction in T_{rise} of QL/QLB results in a higher voltage at the gate terminal of M1/M4 at a particular time instant. This leads to operating the transistor more in the inversion than in the sub-threshold region. Hence, lower impact of variations, higher I_{DS}, and a faster discharge of the ML. The ML discharge with a smaller T_{rise} results in a smaller overlap between the distributions of Hamming distances. Fig. 13(a) demonstrates that a smaller T_{rise} gives lower error probabilities for TCAM cell operating at all temperatures. The faster discharge of ML also helps to lower the operating latency and results in reduced overall power consumption, as shown in Fig. 13(b).

VIII. CONCLUSION

In this work, we have analyzed the 5 nm FinFETs-based IMC circuits at cryogenic temperatures for the first time. To do so, we have experimentally characterized the 5 nm technology FinFETs from 10 K to 300 K and the impact of process variations at 300 K. We have shown the impact of process variations and cryogenic temperatures on SRAM noise margins and TCAM error probabilities. We have presented that TCAM cells exhibit higher error probabilities at cryogenic temperatures than 300 K. The impact of the process

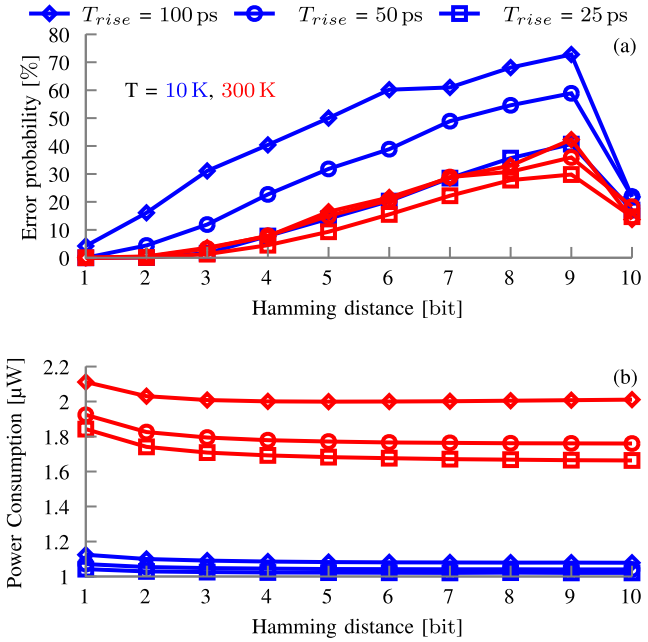


FIGURE 13. Impact of input signal rise time on (a) error probabilities and (b) power consumption at 300K and 10K.

variations on TCAM cells can be minimized by reducing the T_{rise} of the QL/QLB signal and operating the transistors at iso-I_{OFF} conditions. Transistors operating in iso-I_{OFF} condition resulted in 2.94× lower error probability at 10 K compared to 300 K. Only a few ps of delay in the X-SRAM array constructed using 5 nm technology highlights the suitability of SRAM-based IMC circuits for Cryogenic CMOS circuitry in the interfacing layer of quantum computers

ACKNOWLEDGMENT

The authors would like to thank Central Research Facility IIT Delhi, India, for facilitating the cryogenic characterization of FinFETs.

The authors would also like to thank Paul R. Genssler, Victor van Santen, and Munazza Sayed from the Chair of Semiconductor Test and Reliability (STAR) for their valuable support and helpful discussions.

REFERENCES

[1] M. Reiher, N. Wiebe, K. M. Svore, D. Wecker, and M. Troyer, "Elucidating reaction mechanisms on quantum computers," *Proc. Natl. Acad. Sci.*, vol. 114, no. 29, pp. 7555–7560, 2017. [Online]. Available: https://www.pnas.org/doi/abs/10.1073/pnas.1619152114

[2] A. G. Fowler, M. Mariantoni, J. M. Martinis, and A. N. Cleland, "Surface codes: Towards practical large-scale quantum computation," *Phys. Rev. A*, vol. 86, Sep. 2012, Art. no. 032324. [Online]. Available: https://link.aps.org/doi/10.1103/PhysRevA.86.032324

[3] N. C. Jones et al., "Layered architecture for quantum computing," *Phys. Rev. X*, vol. 2, Jul. 2012, Art. no. 031007. [Online]. Available: https://link.aps.org/doi/10.1103/PhysRevX.2.031007

[4] J. P. G. Van Dijk et al., "A scalable cryo-CMOS controller for the wideband frequency-multiplexed control of spin qubits and transmons," *IEEE Sensors J. Solid-State Circuits*, vol. 55, no. 11, pp. 2930–2946, Nov. 2020.

[5] S. J. Pauka et al., "A cryogenic CMOS chip for generating control signals for multiple qubits," *Nat. Electron.*, vol. 4, no. 1, pp. 64–70, Jan. 2021. [Online]. Available: https://doi.org/10.1038/s41928-020-00528-y

[6] F. Arute et al., "Quantum supremacy using a programmable superconducting processor," *Nature*, vol. 574, no. 7779, pp. 505–510, Oct. 2019. [Online]. Available: https://doi.org/10.1038/s41586-019-1666-5

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

[8] X. Si et al., "24.5 A twin-8T SRAM Computation-in-memory macro for multiple-bit CNN-based machine learning," in *Proc. 2019 IEEE Int. Solid-State Circuits Conf.–(ISSCC)*, 2019, pp. 396–398.

[9] P. Wang, X. Peng, W. Chakraborty, A. Khan, S. Datta, and S. Yu, "Cryogenic performance for compute-in-memory based deep neural network accelerator," in *Proc. 2021 IEEE Int. Symp. Circuits Syst. (ISCAS)*, 2021, pp. 1–4.

[10] P. Wang, X. Peng, W. Chakraborty, A. I. Khan, S. Datta, and S. Yu, "Cryogenic benchmarks of embedded memory technologies for recurrent neural network based quantum error correction," in *Proc. 2020 IEEE Int. Electron Devices Meeting (IEDM)*, 2020, pp. 38.5.1–38.5.4.

[11] Q. Dong et al., "15.3 A 351TOPS/W and 372.4GOPS compute-in-memory SRAM macro in 7nm FinFET CMOS for machine-learning applications," in *Proc. 2020 IEEE Int. Solid–State Circuits Conf.–(ISSCC)*, 2020, pp. 242–244.

[12] S. Thomann, P. R. Genssler, and H. Amrouch, "HW/SW co-design for reliable TCAM-based in-memory brain-inspired hyperdimensional computing," *IEEE Trans. Comput.*, vol. 72, no. 8, pp. 2404–2417, Aug. 2023.

[13] A. Fritsch et al., "A 4GHz, low latency TCAM in 14nm SOI FinFET technology using a high performance current sense amplifier for AC current surge reduction," in *Proc. ESSCIRC Conf. 2015–41st Eur. Solid-State Circuits Conf. (ESSCIRC)*, 2015, pp. 343–346.

[14] K. Nii et al., "13.6 A 28nm 400MHz 4-parallel 1.6Gsearch/s 80Mb ternary CAM," in *2014 IEEE Int. Solid-State Circuits Conf. Dig. Tech. Papers (ISSCC)*, 2014, pp. 240–241.

[15] I. Hayashi et al., "A 250-MHz 18-Mb full ternary CAM with low-voltage matchline sensing scheme in 65-nm CMOS," *IEEE J. Solid-State Circuits*, vol. 48, no. 11, pp. 2671–2680, Nov. 2013.

[16] A. Agrawal, A. Jaiswal, C. Lee, and K. Roy, "X-SRAM: Enabling in-memory Boolean computations in CMOS static random access memories," *IEEE Trans. Circuits Syst. I, Reg. Papers*, vol. 65, no. 12, pp. 4219–4232, Dec. 2018.

[17] Y. Hou et al., "Cryogenic in-MRAM computing," in *Proc. 2021 IEEE/ACM Int. Symp. Nanoscale Archit. (NANOARCH)*, 2021, pp. 1–6.

[18] W. Li, J. Read, H. Jiang, and S. Yu, "MAC-ECC: In-situ error correction and its design methodology for reliable NVM-based compute-in-memory inference engine," *IEEE Trans. Emerg. Sel. Topics Circuits Syst.*, vol. 12, no. 4, pp. 835–845, Dec. 2022.

[19] S. Thomann et al., "On the reliability of in-memory computing: Impact of temperature on ferroelectric TCAM," in *Proc. 2021 IEEE 39th VLSI Test Symp. (VTS)*, 2021, pp. 1–6.

[20] O. Prakash, K. Ni, and H. Amrouch, "Ferroelectric FET threshold voltage optimization for reliable in-memory computing," in *Proc. 2022 IEEE Int. Rel. Phys. Symp. (IRPS)*, 2022, pp. 1–10.

[21] C. Marchand, I. O'Connor, M. Cantan, E. T. Breyer, S. Slesazeck, and T. Mikolajick, "A FeFET-based hybrid memory accessible by content and by address," *IEEE J. Explor. Solid-State Computat. Devices Circuits*, vol. 8, no. 1, pp. 19–26, Jun. 2022.

[22] M. Yayla, S. Thomann, S. Buschjäger, K. Morik, J.-J. Chen, and H. Amrouch, "Reliable binarized neural networks on unreliable beyond Von-Neumann architecture," *IEEE Trans. Circuits Syst. I, Reg. Papers*, vol. 69, no. 6, pp. 2516–2528, Jun. 2022.

[23] S. Pellerano et al., "Cryogenic CMOS for qubit control and readout," in *Proc. 2022 IEEE Custom Integr. Circuits Conf. (CICC)*, 2022, pp. 01–08.

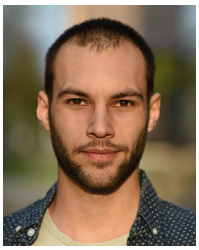
[24] G. Pahwa, P. Kushwaha, A. Dasgupta, S. Salahuddin, and C. Hu, "Compact modeling of temperature effects in FDSOI and FinFET devices down to cryogenic temperatures," *IEEE Trans. Electron Devices*, vol. 68, no. 9, pp. 4223–4230, Sep. 2021.

[25] *BSIM-CMG Technical Manual*. (2022). [Online]. Available: http://bsim.berkeley.edu/models/bsimcmg/

[26] H. Bohuslavskyi et al., "Cryogenic subthreshold swing saturation in FD-SOI MOSFETs described with band broadening," *IEEE Electron Device Lett.*, vol. 40, no. 5, pp. 784–787, May 2019.

[27] A. Beckers, F. Jazaeri, and C. Enz, "Theoretical limit of low temperature subthreshold swing in field-effect transistors," *IEEE Electron Device Lett.*, vol. 41, no. 2, pp. 276–279, Feb. 2020.
[28] Y. Li, H.-W. Cheng, Y.-Y. Chiu, C.-Y. Yiu, and H.-W. Su, "A unified 3D device simulation of random dopant, interface trap and work function fluctuations on high-K/metal gate device," in *Proc. 2011 Int. Electron Devices Meeting*, 2011, pp. 5.5.1–5.5.4.
[29] D. N. Batchelder and R. O. Simmons, "Lattice constants and thermal expansivities of silicon and of calcium fluoride between 6° and 322°k," *J. Chem. Phys.*, vol. 41, no. 8, pp. 2324–2329, 1964. [Online]. Available: https://doi.org/10.1063/1.1726266
[30] A. Grill et al., "Temperature dependent mismatch and variability in a cryo-CMOS array with 30k transistors," in *Proc. 2022 IEEE Int. Rel. Phys. Symp. (IRPS)*, 2022, pp. 10A.1–1–10A.1–6.
[31] M. Cassé, B. C. Paz, G. Ghibaudo, and M. Vinet, "Electrical characterization and modeling of FDSOI MOSFETs for cryo-electronics," in *Proc. 2022 IEEE 15th Workshop Low Temperature Electron. (WOLTE)*, 2022, pp. 1–4.
[32] P. A. 'T Hart, M. Babaie, E. Charbon, A. Vladimirescu, and F. Sebastiano, "Characterization and modeling of mismatch in cryo-CMOS," *IEEE Sensors J. Electron Devices Soc.*, vol. 8, pp. 263–273, Feb. 2020.
[33] K. Ni et al., "Ferroelectric ternary content-addressable memory for one-shot learning," *Nat. Electron.*, vol. 2, no. 11, pp. 521–529, 2019.

SHIVENDRA SINGH PARIHAR (Member, IEEE) is currently pursuing the Doctoral degree with the Indian Institute of Technology Kanpur, Kanpur, India. He is currently associated with the Chair of Semiconductor Test and Reliability, University of Stuttgart, Germany, as a Research Scholar. His primary research interests are characterization and compact modeling of advanced CMOS technologies for circuit design.



SIMON THOMANN (Member, IEEE) received the bachelor's and master's degrees in computer science from the Karlsruhe Institute of Technology, Germany, in 2019 and 2022, respectively. He is currently pursing the Ph.D. degree with the Chair of Semiconductor Test and Reliability, University of Stuttgart. His research interests range from device to system level. Special interest lies on circuit design, emerging technologies, and cross-layer reliability modeling from device to circuit level.

GIRISH PAHWA (Member, IEEE) received the Ph.D. and M.Tech. degrees in electrical engineering from the Indian Institute of Technology (IIT) Kanpur in 2020. He is an Assistant Professional Researcher with the Department of Electrical Engineering and Computer Sciences (EECS), University of California (UC) at Berkeley, where he is also the Executive Director of the Berkeley Device Modeling Center. Prior to this, he worked as a Postdoctoral Researcher with EECS, UC Berkeley from 2020 to 2021. His research primarily focuses on the modeling and simulation of nanoscale devices and device circuit co-design and optimization of emerging transistor technologies with a special emphasis on ferroelectric devices. He is a co-developer of industry-standard BSIM-CMG, BSIM-IMG, BSIM-BULK, BSIM-SOI, and BSIM4 models. He has also developed the first industry-standard models for cryogenic FinFET and FDSOI FETs for quantum computing and cold electronics applications. He has over 50 technical publications in prominent journals and conferences in the field of device modeling and simulation. He is the recipient of the IEEE EDS Early Career Award in 2022, the Outstanding Ph.D. Thesis Award from IIT Kanpur in 2020, and the Best Paper Award at the IEEE International Conference on Emerging Electronics, Mumbai, India, in 2016. He serves as the reviewer of several reputed journals.

YOGESH SINGH CHAUHAN (Fellow, IEEE) is a "Class of 1984" Chair Professor with the Indian Institute of Technology Kanpur, India. He was with Semiconductor Research and Development Center, IBM Bangalore from 2007 to 2010; Tokyo Institute of Technology in 2010; University of California at Berkeley from 2010 to 2012; and ST Microelectronics from 2003 to 2004. He is the developer of several industry standard models, such as ASM-GaN-HEMT model, BSIM-BULK (formerly BSIM6), BSIM-CMG, BSIM-IMG, BSIM4, and BSIM-SOI models. His research group is involved in developing compact models for GaN transistors, FinFET, Nanosheet/Gate-All-Around FETs, FDSOI transistors, Negative Capacitance FETs, and 2-D FETs. He has published more than 300 papers in international journals and conferences. His research interests are characterization, modeling, and simulation of semiconductor devices. He received Ramanujan Fellowship in 2012, IBM Faculty Award in 2013 and P. K. Kelkar Fellowship in 2015, CNR Rao Faculty Award, Humboldt Fellowship, and Swarnajayanti Fellowship in 2018. He is an Editor of IEEE TRANSACTIONS ON ELECTRON DEVICES and a Distinguished Lecturer of the IEEE Electron Devices Society. He is the Chair of IEEE-EDS Compact Modeling Committee. He is the Founding Chairperson of IEEE Electron Devices Society U.P. Chapter and Chairman-Elect of IEEE U.P. Section. He has served in the technical program committees of IEEE International Electron Devices Meeting, IEEE International Conference on Simulation of Semiconductor Processes and Devices, IEEE European Solid-State Device Research Conference, IEEE Electron Devices Technology and Manufacturing, and IEEE International Conference on VLSI Design and International Conference on Embedded Systems. He is a Fellow of Indian National Academy of Engineering.

HUSSAM AMROUCH (Member, IEEE) received the Ph.D. degree with the highest distinction (Summa cum laude) from KIT in 2015. He is a Professor heading the Chair of AI Processor Design, Technical University of Munich. He is, additionally, with the Munich Institute of Robotics and Machine Intelligence, Germany. Further, he is the Head of the Semiconductor Test and Reliability, University of Stuttgart, Germany. Prior to that, he was a Research Group Leader with the Karlsruhe Institute of Technology, where he was leading the research efforts in building dependable embedded systems. He currently serves as an Editor for the *Scientific Reports* (Nature). He has around 230 publications in multidisciplinary research areas (including 90+ journals) across the entire computing stack, starting from semiconductor physics to circuit design all the way up to computer-aided design and computer architecture. His research in HW security and reliability have been funded by the German Research Foundation, Advantest Corporation, and the U.S. Office of Naval Research. His main research interests are design for reliability and testing from device physics to systems, machine learning for CAD, HW security, approximate computing, and emerging technologies with a special focus on ferroelectric devices. He holds eight HiPEAC Paper Awards and the three best paper nominations at top EDA conferences, including DAC'16, DAC'17, and DATE'17, for his work on reliability. He has served in the technical program committees of many major EDA conferences, such as DAC, ASP-DAC, and ICCAD, and as a Reviewer in many top journals like *Electronics* (Nature), IEEE TRANSACTIONS ON ELECTRON DEVICES, IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS—PART I: REGULAR PAPERS, IEEE TRANSACTIONS ON VERY LARGE SCALE INTEGRATION (VLSI) SYSTEMS, IEEE TRANSACTIONS ON COMPUTER-AIDED DESIGN OF INTEGRATED CIRCUITS AND SYSTEMS, and IEEE TRANSACTIONS ON COMPUTERS.