A Timing Yield Model for SRAM Cells at Sub/Near-Threshold Voltages Based on a Compact Drain Current Model

Shan Shen, Peng Cao, Member, IEEE, Ming Ling, Member, IEEE, and Longxing Shi, Senior Member, IEEE

Abstract—Sub/near-threshold static random-access memory (SRAM) design is crucial for addressing the memory bottleneck in power-constrained applications. However, the high integration density and reliability under process variations demand an accurate estimation of extremely small failure probabilities. To capture such a “rare event” in memory circuits, the time and storage overhead of conventional simulations based on the Monte Carlo (MC) analysis cannot be tolerated. On the other hand, classic analytical methods predicting failure probabilities from a physical expression become inaccurate in the sub/near-threshold voltage domain due to the hypothetical distribution or the oversimplified drain current ($I_{ds}$) model for nanoscale devices. This work first proposes a simple but efficient empirical $I_{ds}$ model to describe the drain-induced barrier lowering (DIBL) effect. Based on that, the probability density functions of the interest metrics in SRAM are derived. Two analytical models are then put forward to evaluate SRAM dynamic stabilities, including the access time failure and the write failure. The proposed models can be extended easily to different types of SRAM with different read/write-assist circuits. The models are validated against MC simulations across different operating voltages and temperatures. The average relative errors at 0.5-V $V_{DD}$ are only 8.8% for the access-time failure model and 10.4% for the write failure model. The size of the required sample data set is 43.6× smaller than that of the state-of-the-art method.

Index Terms—Cell failure analysis, compact model, near-threshold voltage, static random-access memory (SRAM).

I. INTRODUCTION

MEMORY cell circuits need to be replicated millions of times to achieve extremely high integration density in a small footprint, where cutting-edge process technology is demanded. In this case, a stringent yield requirement of memory chips can be translated into an extremely small failure probability of each component cell, thereby making the circuit failure a “rare event.” In general, the failure probability estimation is usually analytically intractable due to the high complexity of memory circuits. The most straightforward approach is the Monte Carlo (MC) method, which repeatedly draws samples and evaluates circuit performance by transistor-level SPICE simulations. However, MC is extremely time consuming for the rare-event estimation, because millions or even billions of samples are needed to capture one failure.

Several analytical models [1]–[9] were proposed to solve this problem. For example, work [4] uses Gaussian-tail-fitting to model the distribution of one-sided static noise margin (SNM). Although they achieve great accuracy at nominal voltages, predictions deviate from simulation results at low voltages or low failure probabilities [5]. This is because: 1) the cell read current ($I_{READ}$) or differential voltage ($\Delta V$) on bit-lines (BLs) becomes non-Gaussian at low supply voltages and 2) the drain current ($I_{ds}$) model from which analytical expressions are derived is either unsuitable for the transregional near-threshold domain or ignores the short-channel effect. Analytical methods based on SNM [6], [7] and dynamic noise margin (DNM) [8] were proposed to help the bitcell design, but these metrics only account for read/write margins to balance the sizes of transistors, rather than timing violations. The voltage acceleration method (VAM) [9] measures $I_{READ}$ at the high-$\sigma$ region from actual silicon chips by aggregating the $I_{READ}$ distributions at lower voltages. Besides $I_{READ}$, the functional failures of static random-access memory (SRAM) are also affected by variations of peripheral circuits and the length of the wordline (WL) pulse [3].

Furthermore, the SRAM frequency depends on the longest path of the access and write operations. As a result, we need to consider both of them in failure analyses. To some extent, other types of failure models, such as those for static timing analysis (STA) [10], [11], can be also used to analyze SRAM write failures (an SRAM cell uses 2 cross-coupled inverters to store data). But the yield constraints of a logic path are much looser than that of SRAM.

Other researchers use importance sampling to largely reduce the number of samples required by MC simulations. The
adaptive importance sampling (AIS) method [12] develops an iterative resampling approach to search for failure regions, which can further reduce the sample number. High-dimension importance sampling (HDIS) [13] was proposed to solve the high-dimension problem of process parameters when the scale of circuits is large. Unfortunately, the importance-sampling-based methods have two main problems. SRAM comprised of most analog/mixed-signal circuits must be simulated at the transistor level. The time of a single simulation is still large. Although these types of analysis methods can decrease the number of simulations, they still require tens of thousands of training samples to converge compared to analytical models. On the other hand, the predicted results of these methods may have large fluctuations in some cases [5], which is because the likelihood ratios between the original sampling distribution and the distorted one have huge numerical instability [34].

Cell yield analyses using low-rank tensor approximation (LRTA) [14] and spline-high-dimensional model representation (SP-HDMR) [15] substitute metamodels for the expensive simulations in failure probability estimation and integrate their models into AIS and HDIS, respectively. These methods are equivalent to trading the estimation precision for the simulation speedup because the accuracy of metamodels greatly affects the final predictions. Besides, constructing and solving the metamodels are still complicated.

In a word, accuracy and speedup are a tradeoff in failure analyses, where analytical models are efficient enough but suffer from poor accuracy, particularly at low supply voltages. Sampling-based methods improve the precision but cost lots of samplings or computation resources. Thus, in this article, our goal is to probe the boundary of the tradeoff in the fail-

1) We improve the accuracy of the compact $I_{ds}$ model by considering the drain-induced barrier lowering (DIBL) effect, which is the most dominant characteristic of nanoscale devices in the near/subthreshold voltage domain. Moreover, our empirical $I_{ds}$ model achieves a great balance between precision and simplicity. It can be used for hand-calculation and early-stage design analysis.

2) The non-Gaussian-distributed circuit metrics in SRAM, such as $\Delta V$ between the BL and write delay, are derived analytically and expressed as a simple nonlinear function of a Gaussian random variable based on the proposed $I_{ds}$ model. The corresponding distribution parameters in their probability density functions (PDF) can be characterized by trivial simulation efforts.

3) Finally, the analytical models are constructed to evaluate the probabilities of both the access-time failure and the write failure for SRAM cells with a remarkable speedup. Due to their good scalabilities, the proposed failure models can be directly used by SRAMS comprised of different read/write assist circuits. Moreover, they can also be extended easily for different types of cell topologies. By taking the impact of the supply voltage, transistor size, and load capacitance into account, the results provide physical insights and guidelines for both cell and peripheral circuit designs.

The remainder of this article is arranged as follows: Section II introduces the preliminaries, Section III develops the $I_{ds}$ model, access-time failure, and write failure models sequentially, Section IV shows evaluation results, and Section V concludes this article.

II. PRELIMINARIES

A. Inversion-Charge-Based $I_{ds}$ Models

The classical $I_{ds}$ model [16] for MOS transistors is widely used by other analytical failure models for logic paths and SRAMs [7]

$$I_{ds} = I_0 e^{\frac{V_{gs}-V_{th}}{n} - \frac{V_{ds}}{V_{th}}} \cdot (1 - e^{-\frac{V_{ds}}{V_{th}}})$$

where $V_{th}$ is the threshold voltage, $n$ is the subthreshold swing parameter, $\lambda$ is the DIBL coefficient, and $V_t$ is the thermal voltage. Although (1) is simple and suitable for hand-calculation, it introduces strike errors in the near-threshold voltage domain due to the oversimplified relationship between $I_{ds}$ and $V_{gs}$.

A recent work [17] presents a new approximation of this relation via a Taylor series expansion, resulting in $I_{ds} \propto e^{K_1((V_{gs}-V_{th})/nvt)+K_2((V_{gs}-V_{th})/nvt)^2}$, where $K_1$ and $K_2$ are fitting constants. This approximation yields higher accuracy and continuity in the near-threshold region than the model (1) and is adopted by other works [10], [11]. However, the short-channel effect, such as DIBL, was not accurately incorporated in [17].

B. Failure Mechanisms of SRAM

Variations of process parameters during chip fabrication include interdie (global) and intradie (local) variations. The latter can be further divided into systematic variations and random variations. Systematic variations exhibit spatial correlations, where devices that are close to each other have a higher probability of being alike than devices that are placed far apart [18]. An example of systematic variations is the change of the channel length ($L$) of different transistors in a die that are spatially correlated. Oppositely, random variations refer to the shifts in parameters of neighboring transistors that are completely independent and uncorrelated to their spatial distances. The random dopant fluctuation (RDF) induced $V_{th}$ variation is a classic example of random variations. Fortunately, the systematic variations do not result in large differences between two transistors that are in close spatial proximity. The random intradie variations can result in a significant mismatch of $V_{th}$ between the two neighboring transistors in a die [1]. It will further cause functional failures of SRAM cells. Here, we briefly discuss the mechanisms of these failures.

1) Access Time Failure: Before accessing a cell, a pair of BLs is first precharged to $V_{DD}$ by the precharge pMOS. Then, the access transistors M2 and M3 are turned on during cell reading, which is shown in Fig. 1 ($V_Q = “1”$ and $V_QB = “0”$). Transistor M3 will try to pull one of the BLs, BLB, down to the ground, while the voltage at BL remains high. Reading
Fig. 1. Schematic of SRAM in a read operation. Assume the cell stores “1” and BLB is discharging.

a 0 is a symmetrical situation, where M2 and M0 will discharge the BL to the ground and BLB is high. After some time, a sense amplifier (SA) will detect the differential voltage between BL and BLB ($\Delta V = V_{BL} - V_{BLB} > 0$). However, SAs also suffer from $V_{th}$ mismatches and have “bias” (offset) voltages. The offset voltage ($V_{OS}$) can be a positive value or a negative value. As long as the input voltage difference is less than $V_{OS}$ (whether it is positive or negative), the read result will be a 0; otherwise, it will be a 1. Within a tolerable WL assertion time ($T_{WL}$), as illustrated in Fig. 1, the access-time failure occurs when $\Delta V < V_{OS}$ of the activated SA. Oppositely, the access-time violation happens when $\Delta V > V_{OS}$ for reading a 0 (notice that $\Delta V < 0$). This failure is caused by the strength reduction of the access and the pull-down nMOS (M3 and M1, or M2 and M0).

2) Read Upset: In Fig. 1, due to the voltage divided by pull-down transistor M1 and the access transistor M3, $V_{QB}$ increases to a positive value $V_{READ}$. If $V_{READ}$ is higher than the trip-point of the inverter M4-M0 ($V_{TRIP_L}$), then the very act of reading unintentionally overwrites the cell. This represents a read-upset event (also called read destruction or read failure). A relatively stronger access nMOS (M3 or M2) and a weaker pull-down nMOS (M1 or M0) increase $V_{READ}$, thereby leading to a probability increase of read upsets.

3) Write Failure: While writing a 0 to a cell storing 1 ($V_{BL} = 0$, $V_{BLB} = V_{DD}$), node Q discharges through BL to a low value determined by the voltage division between the pull-up pMOS M4 and the access nMOS M2. At the same time, the node QB also is charged by BLB. If $V_Q$ cannot be reduced below the trip point of the inverter M5-M1 ($V_{TRIP_R}$) as well as $V_{QB}$ cannot be pumped above $V_{TRIP_L}$ within the time when word-line is high ($T_{WL}$), then a write failure occurs (Fig. 2). At the node QB, $V_{QB} > 0$ due to the voltage division, therefore, M3 always has a forward body bias, making its $V_{th}$ increase during the entire write operation (the body connects to the ground in an SRAM cell). This results in a smaller write current through M3 than that through M2. A faster discharge of node Q means $V_Q$ is more likely to reach $V_{TRIP_R}$ triggering the positive feedback of the SRAM cell. As a consequence, the discharged side decides a successful write in an SRAM cell. It is a symmetrical case when overwriting a 0 cell, that is, the QB starts discharging through M3 to a low voltage determined by M5 and M3. A write failure occurs when $V_{QB}$ cannot be reduced below $V_{TRIP_L}$ within $T_{WL}$.

Hence, a stronger pMOS and a weaker access nMOS can significantly slow down the discharging process, thereby causing a write failure.

4) Hold Failure: In the standby mode, the supply voltage of SRAM is reduced to minimize the leakage power consumption. However, if the lowering of $V_{DD}$ causes the data stored in the cell to be destroyed, then it is said to have failed in the hold mode. As the supply voltage of the cell is lowered, the voltage at the node storing 1 (node Q in Fig. 1) also gets reduced. Moreover, leakage of the pull-down nMOS M0 can also reduce $V_Q$. If the voltage is reduced below $V_{TRIP_R}$, then flipping occurs and the data are lost in the hold mode. Although all these failures may occur due to process parameter variations, our work focuses on the most dominant failure that relates to the $V_{DD}$ and temperature variations in the near-threshold domain. Therefore, we run $1.3 \times 10^7$ MC simulations to compare the probabilities and the sensitivities to the PVT conditions of all those four types of cell failures. TSMC 28 nm technology is used and the SRAM demo is foundry-provided in the process design kit (PDK). The access-time constraint is set to 4 ns and the write-time constraint is set to 0.2 ns. Table I shows the probability comparison of different

| $V_{DD}$ (V) @ 25°C | 0.2 | 0.3 | 0.4 | 0.5 | 0.6 |
|----------------------|-----|-----|-----|-----|-----|
| Hold failures        | 0   | 0   | 0   | 0   | 0   |
| Access-time failures | 1   | 1   | 0.154 | 1.37E-5 | 0   |
| Read upsets          | 0.028 | 2.78E-4 | 6.60E-6 | 0   |
| Write failures       | 1   | 1   | 0.716 | 5.25E-4 | 0   |

| Temp. (°C) @ 0.5V    | -75 | -25 | 25  | 75  | 125 |
|-----------------------|-----|-----|-----|-----|-----|
| Hold failures         | 0   | 0   | 0   | 0   |
| Access-time failures  | 0.116 | 3.90E-3 | 1.37E-5 | 0   |
| Read upsets           | 1.86E-4 | 4.01E-5 | 0   |
| Write failures        | 1   | 0.84 | 5.25E-4 | 0   |

| Table I              | (a) Probability Comparison of Different Types of Failures Across Different Supply Voltages | (b) Probability Comparison of Different Types of Failures Across Different Temperatures |
|----------------------|---------------------------------------------------------------|-----------------------------------------------|
| (a)                  | (b)                                                          |                                               |
| $V_{DD}$ (V) @ 25°C  | Hold failures | Access-time failures | Read upsets | Write failures |
|----------------------|----------------|----------------------|--------------|---------------|
| 0.2                  | 0              | 1                    | 0.028        | 1             |
| 0.3                  | 0              | 1                    | 2.78E-4      | 1             |
| 0.4                  | 0              | 1.154                | 6.60E-6      | 1             |
| 0.5                  | 0              | 1.37E-5              | 0            | 1             |
| 0.6                  | 0              | 0                    | 5.25E-4      | 0             |
| Temp. (°C) @ 0.5V    | Hold failures | Access-time failures | Read upsets  | Write failures |
|----------------------|----------------|----------------------|--------------|---------------|
| -75                  | 0              | 0.116                | 1.86E-4      | 1             |
| -25                  | 0              | 3.90E-3              | 4.01E-5      | 1             |
| 25                   | 0              | 1.37E-5              | 0            | 1             |
| 75                   | 0              | 0                    | 5.25E-4      | 0             |
| 125                  | 0              | 0                    | 0            | 0             |

Fig. 2. Schematic of SRAM in a write operation. Assume we write a “0” into a “1” cell.
types of failures. First, we find no hold failure occurs during our simulations with $V_{DD}$ rising from 0.2 to 0.6 V and the temperature increasing from $-75 \degree C$ to $125 \degree C$. The hold failure is well suppressed due to the low leakage current provided by the advanced nanometer technology. Second, the probability of the read upset is several orders of magnitude lower than that of the access-time failure. It increases with the lowered operating voltage and temperature. Third, the write failure and the access-time failure are very sensitive to the changes of the operating voltage and the temperature. Their probabilities increase abruptly when $V_{DD}$ scales down from 0.5 to 0.4 V. For the write failure, the probability dramatically rises as the temperature drops down below 25 $\degree C$. Finally, compared to the access time, the write time is much shorter with similar yield constraints.

To analyze these two cell failures, designers cannot only rely on the read/write SNMs because they are used together with the assumption of infinite time duration ($T_{WL} = \infty$) [8]. Dynamic metrics, such as access-time and write failures, represent the read and write behaviors in state-of-the-art SRAM designs with read/write-assist circuitry [19] and/or shrinking access cycle time [20], [21]. Moreover, accurate estimation of SRAM dynamic stability enables the evaluation and comparison of different dynamic circuit techniques, achieving agreement between estimates and measurements, fewer design iterations, and less time-to-market [22]. Thus, we construct analytical models to describe and abstract mechanisms of the access-time failure and the write failure in this work.

C. Modeling Access-Time Failure

Due to the high speed and storage density, 6T SRAM is widely used as the high-capacity on-chip storage [19] and can also be operated at near-threshold voltages [20], [21], [23] with simple assist circuits. For more complicated cell structures with separated read ports, such as single-port 8T [24] and 10T [25] illustrated in Figs. 1 and 2, their read paths comprised of an access transistor and a pull-down transistor are identical to that of 6T. The only difference is that the single-read-port 8T cell [24] employs an inverter or domino logic to detect the voltage drop on the read bitline. For simplicity, we use a standard 6T SRAM cell as an example in the rest of this article. However, a similar modeling process can be reproduced in other complicated scenarios and left to our future work.

Complete SRAM access can be divided into several cycles to satisfy the logic core frequency. However, the critical path always contains the following steps as illustrated in Fig. 1: ❶ the WL is first driven to high, ❷ the BLB with a large load starts to discharge from $V_{DD}$ to the ground through M3 and M1, and ❸ SAE signal activates the SAs [3]. At near-threshold voltages, discharging the BLs with large capacitances through those minimum-sized transistors becomes slower due to the reduced $I_{READ}$. The status of M2 is unchanged so that $\Delta V$ is mainly contributed by discharging BLB. As we discussed in Section II-B1), by setting $T_{WL}$ to a specified time duration ($T_{READ}$), the probability of an access-time failure emergence can be calculated as [1]

$$
P_f = P(\Delta v < \Delta V | T_{READ}) = \Phi_{\Delta}(v_{OS}, T_{READ}) = \left\{ \begin{array}{ll} \int_{-\infty}^{v_{OS}} f_{\Delta}(\Delta v, T_{READ})d\Delta v, & v_{OS} > 0 \\ 0, & v_{OS} \leq 0 \end{array} \right. \tag{2}$$

where $\Delta v$ is a variable of differential voltage, $v_{OS}$ is a possible offset voltage of the SA, $f_{\Delta}$, and $\Phi_{\Delta}$ are the PDF and the cumulated density function (CDF) of $\Delta v$. When reading a 0 cell, $\Delta v < 0$, the probability of an access-time failure changes to

$$
P_f = P(\Delta v > v_{OS} | T_{READ}) = \Phi_{\Delta}(v_{OS}, T_{READ}) = \left\{ \begin{array}{ll} \int_{v_{OS}}^{+\infty} f_{\Delta}(\Delta v, T_{READ})d\Delta v, & v_{OS} < 0 \\ 0, & v_{OS} \geq 0 \end{array} \right. \tag{3}$$

Though $\Delta V$ is positive during reading a 1 and negative during reading a 0, modeling $\Delta V$ is identical in both cases. This can be explained by the randomness of intradie variations and the symmetry layout of the 6T cell structure. In other words, the process parameters, such as $V_{th}$, from the symmetrical transistor pairs (M0-M1, M2-M3, M4-M5) have the same distribution and are totally independent. As a result, the analysis of reading a 1 cell where BLB discharges through M3 and M1 is identical to that of reading a 0 cell where BL discharges through M2 and M0. Prior works [1]–[3] treat $\Delta V$ (or $I_{READ}$) as a Gaussian random variable according to simulation results at the nominal voltage. However, this assumption no longer holds in the sub/near-threshold voltage domain, leading to increased errors as supply voltage scales down [5].

Each cell failure event is related to those of other cells that share the same SA. This relationship among cell failures can be described by a joint PDF and the independence of $V_{OS}$ and $\Delta V$. The offset voltage is mainly due to the $V_{th}$ mismatch from the cross-coupled inverters in the SA, and it also follows a Gaussian distribution even under low voltages [3]. Some classical models [1], [2] assume $V_{OS}$ to be the worst-case value, leading to a pessimistic yield estimation. The final failure probability for a random cell in an SRAM array, called the bit error rate (BER), can be expressed by

$$
P_f(T_{READ}) = \int_{-\infty}^{+\infty} f_{v_{OS}}(v) \cdot \Phi_{\Delta}(v_{OS}, T_{READ})dv_{OS}$$

$$
f_{v_{OS}}(v) = \frac{1}{\sqrt{2\pi\sigma_{v_{OS}}}} e^{-\frac{(v-\mu_{v_{OS}})^2}{2\sigma_{v_{OS}}^2}} \tag{4}$$

where $f_{v_{OS}}$ is the PDF, and $\mu_{v_{OS}}$ and $\sigma_{v_{OS}}$ are the mean and standard deviation of $V_{OS}$.

D. Modeling Write-Time Failure

Since analyses of writing a 0 and a 1 are totally symmetric as we discussed above, failure probabilities of writing a 0 and a 1 are equal. Here, we use Fig. 2 as an example to model the operation. The operation contains only two steps: ❶ the WL is asserted, and ❷ node $Q$ is discharged while node QB is charged through access transistors and BLs (10T [25] has two access transistors
on the write path. As we discussed in Section II-B3), if \( V_Q \) cannot be below \( V_{\text{TRIP}_R} \) within a given \( T_{\text{WL}} \), a write failure occurs definitely. Thus, the minimum write time required by a write operation is [1], [2]

\[
t = \left[ C_Q \int_{V_{\text{DD}}}^{V_{\text{TRIP}_R}} \frac{dV_Q}{IM_2 - IM_0 + IM_4} \right]
\]

where \( IM_2 \), \( IM_4 \), and \( IM_0 \) are the channel current for M4, M2, and M0, respectively. \( C_Q \) is the load capacitance of node \( Q \). The trip-point voltage \( V_{\text{TRIP}_R} \) can be extracted from

\[
IM_5(V_{gs} = V_{ds} = V_{\text{TRIP}_R} - V_{\text{DD}}) + IM_5(V_{gs} = V_{\text{WL}} - V_{\text{TRIP}_R}, V_{ds} = V_{\text{DD}} - V_{\text{TRIP}_R}) = IM_1(V_{gs} = V_{ds} = V_{\text{TRIP}_R}).
\]

Given a specific time constraint \( T_{\text{WL}} = T_{\text{WRITE}} \), the probability of write failures can be calculated as [1]

\[
P_f(T_{\text{WRITE}}) = \int_{T_{\text{WRITE}}}^{\infty} f_W(t)dt
\]

where \( f_W \) is the PDF of the write time.

The failure model [1] for SRAM yield analysis assumes \( T_{\text{WRITE}} \) follows a noncentral \( F \) distribution, while other models for STA find the delay time of an inverter chain follows Log-Normal [10], or the log-skew-normal (LSN) distribution [11]. Rather than relying on the specific shape of PDFs of access or write time, the importance-sampling-based methods have been proposed to construct a distorted sampling distribution by shifting the mean-vector (a vector containing process parameters of the six transistors in a cell) of the original distribution to the boundary or the center of the failure region. The metamodel-based methods treat the relation between performance metrics and process parameters as a black box, fitting it through nonlinear functions.

In general, the key to yield modeling is to find the exact PDFs of the circuit metrics of interest, which is the main purpose of this work.

III. PROPOSED ANALYTICAL TIMING YIELD MODEL

A. \( I_{ds} \) Model for the Sub/Near-Threshold Voltage Domain

In an nMOS transistor, the drain and source are connected by back-to-back \( pn \)-junctions (substrate-source and substrate-drain). In the vicinity of the \( pn \)-boundary, the electrons have been diffused away, or have been forced away by an electric field. The only elements left in this region are ionized donors or acceptor impurities. This region is called the depletion region.

Depletion regions of the source and reverse-biased drain junction become relatively more important with shrinking channel lengths (e.g., SRAM cell circuits). Since a part of the region below the gate is already depleted (by the source and drain fields), a smaller threshold voltage suffices to turn on the transistor. In other words, \( V_{th0} \) decreases with channel length for short-channel devices. A similar effect can be obtained by raising the drain-source (bulk) voltage, as this increases the width of the drain-junction depletion region. Consequently, the threshold decreases with increasing \( V_{ds} \), known as the effect of drain-induced barrier lowering (DIBL). It is the major mechanism in the near-threshold region that affects the output resistance in the saturation range [26], [27].

Since the DIBL effect varies with the operating voltage, it is a problem in memories, where the read current of a cell (i.e., the \( I_{ds} \) of the access nMOS M3 in Fig. 1) becomes a function of the voltage on the BL, which complicates the differential voltage (\( \Delta V \)) modeling. Moreover, the leakage current from stand-by cells also depends upon the BL voltage and data patterns due to the DIBL. From the cell perspective, DIBL manifests itself as a data-dependent noise source [26].

Fortunately, we find that this effect can be modeled very well by introducing an extra exponential \( V_{ds} \) term with the DIBL coefficient (\( \lambda \)). Another \( V_{ds} \) term in our model that accounts for the transition between the linear and saturation regions is simplified and regulated by a constant \( K_1 \). Moreover, based on the work [17], the quadratic polynomial (involving \( K_1 \) and \( K_2 \)) describing \( I_{ds} \) changed with \( V_{ds} \) is also adopted by our model. The final expression of the drain current is

\[
I_{ds} = I_0 e^{K_1(V_{gs} - V_{th}^n) + K_2(V_{gs} - V_{th}^p)^2} \cdot e^{\Delta V(V_{dd} - V_{DS})} \left( 1 - e^{-K_1(V_{ds} - V_{th})} \right)
\]

\[
I_0 = \frac{V_d}{L} \frac{1}{v_t} K_0
\]

where \( K_0, K_1, K_2, and \lambda \) are independent fitting constants, \( \mu \) is the effective electron mobility, \( n \) is the subthreshold swing parameter, and \( V_t \) is the thermal voltage. The parameters in (8) including \( V_{th}, \mu, n, \) and \( K_1 \) are all positive for nMOS \( (V_d \geq V_t) \) and negative for pMOS \( (V_d \leq V_t) \). Although this empirical model introduces artificial constants besides the physical parameters, it shows a great approximation to the simulation results that will be discussed in Section IV-A.

B. Access-Time Failure Model

As shown in Fig. 1, the BL discharging process in SRAM access can be described as

\[
-C_{\text{BLB}} \frac{dV_{\text{BLB}}}{dt} = IM_3(V_{gs} = V_{WL} - V_{QB}, V_{ds} = V_{\text{BLB}} - V_{QB}, V_{bs} = -V_{QB})
\]

where \( IM_3 \) is the channel current of M3, \( V_{WL} \) is the WL-driven voltage, and \( V_{QB} \) is the voltage of node QB in Fig. 1. In this equation, the voltage drop on BL caused by the cell leakage current is ignored in the target near-threshold region (thanks to the leakage control by the advanced technology). Moreover, we ignore the substrate bias of M3 due to a stronger M1 and weaker M3 in a balanced design leading to a small \( V_{th} \). Then, (9) can be rewritten by substituting (8) for \( IM_3 \)

\[
-C_{\text{BLB}} \frac{d(V_{DD} - \Delta V)}{dt} = C_{\text{BLB}} \frac{d\Delta V}{dt} = IM_3(V_{gs} = V_{WL} - V_{QB}, V_{ds} = V_{\text{BLB}} - V_{QB}, V_{bs} = -V_{QB})
\]

\[
= I_0 e^{K_1(V_{gs} - V_{th}^n) + K_2(V_{gs} - V_{th}^p)^2} \cdot e^{\frac{V_{dd} - \Delta V}{V_t}} \cdot (1 - e^{-K_1(V_{ds} - V_{th})})
\]

where \( \Delta V \) is a possible differential voltage, and \( C_{\text{BLB}} \) is the equivalent load capacitance of BLB. The last factor \((1 - e^{-K_1(V_{dd} - \Delta V)/V_t})\) is very close to 1 because of the small \( \Delta V \) (usually tens of millivolts). Let \( p(V_{gs}, V_{th}) = \)
where \( T_{\text{READ}} \) is the preset time constraint. After calculating the logarithm of (11) and bringing \( \Delta V \) to the left side, we get

\[
\Delta V = \frac{m_f}{n} \ln \left( e^{\alpha T_{\text{READ}} + e^{-\frac{\lambda_{DD}}{m_f} - p}} \right) + V_{\text{DD}}
\]

where \( \alpha = (I_0/m_f)C_{\text{BLB}} \). Compared to the variation of \( p \) in (12), \(-\lambda_{DD}/m_f\) dominates the exponential term while \( \alpha T_{\text{READ}} \) dominates the total value in the parenthesis. Therefore, we can approximate the second \( p \) in (12) to its mean value \( p_0 \) yielding

\[
\Delta V \approx \frac{m_f}{n} \ln \left( e^{\alpha T_{\text{READ}} + e^{-\frac{\lambda_{DD}}{m_f} - p_0}} \right) + V_{\text{DD}}
\]

where \( g(t) \) is a function of the specified access time and other process parameters. From (13), it is obvious that

\[
\Delta V \propto p(V_{\text{WL}}, V_{\text{thn}}) \Rightarrow \Delta V \propto K_2 \left( \frac{V_{\text{WL}} - V_{\text{thn}}}{m_f} + \frac{K_1}{2K_2} \right)^2.
\]

Here, we only consider the \( V_{\text{th}} \) variations in our model because, first, the threshold variation is the most dominant factor to cause a cell hard failure (the same conclusion drawn from [1]–[3], [10], [11], [13], [14]). Second, introducing several random variables in our model will complicate the expression of the metrics of interest, and result in difficulty to find their PDFs. In MOSFET models [26], [27], the threshold voltage \( V_{\text{thn}} \) is a linear function of the threshold voltage at a zero substrate bias \( V_{\text{thn0}} \) which is one of the Gaussian-distributed process parameters defined in the technology library. Consequently, \( V_{\text{thn}} \) is a Gaussian random variable. By treating \( V_{\text{WL}} \) as a fixed value \( (= V_{\text{DD}}) \), \( \Delta V \) is a square function of a Gaussian \( V_{\text{thn}} \), which becomes a non-central Chi-square random variable with 1 degree of freedom. This conclusion also matches the observation reported by Shen et al. [5]. Through a simple manipulation, the PDF of \( \Delta V \) is

\[
f_{\Delta V} (\Delta v) = \frac{1}{2\sigma_{\Delta} \sqrt{2\pi} \Delta v} e^{-\frac{(\sqrt{\Delta v - \mu_{\Delta}})^2}{2\sigma_{\Delta}^2}}
\]

where parameters \( \mu_{\Delta} \) and \( \sigma_{\Delta} \) are the mean and standard deviation of \( \Delta V \). They can be directly collected in MC simulations at different transient times in SPICE. By taking (15) into (2)–(4), the final yield at a given \( T_{\text{READ}} \) can be obtained.

Fig. 3 shows the Q-Q plot of \( \Delta V \) at TTG. It shows the tail of quantiles, where a small \( \Delta V \) accounts for an access failure. We use \( 1.3 \times 10^7 \Delta V \) samples from MC simulations at 0.5 V and \( T_{\text{WL}} = 5.1 \text{ns} \) to quantify \( \mu, \sigma \), and the quantiles at the \( y \)-axis. The converging line of quantiles indicates the distribution of data samples is linearly related to the modeled distribution, which means (15) can depict the distribution of \( \Delta V \) accurately.

C. Write Failure Model

To evaluate write-time failures, the current in (5) needs to be replaced by our \( I_{\text{th}} \) model (8). First, we find \( I_{\text{th0}} \) can be ignored compared to \( I_{\text{M4}} \) and \( I_{\text{M2}} \). Second, the transregional term \((1 - e^{-K_1(V_{\text{ds}}/V_t)})\) is close to 1 as long as \( V_{\text{ds}} > 3V_t \) (thermal voltage \( \approx 25 \text{ mV} \)). \( V_{\text{TRIP,R}} \) varies from 40% \( V_{\text{DD}} \) to 62% \( V_{\text{DD}} \) after our evaluation (the target \( V_{\text{DD}} \) is from 0.5 to 0.7 V). As a result, the transregional terms in \( I_{\text{M4}} \) and \( I_{\text{M2}} \) can be removed during write-failure modeling. By using subscripts to distinguish the process parameters of pMOS M4 and nMOS M2, the current difference at node Q in (5) can be expressed as

\[
-I_{\text{M2}} - I_{\text{M0}} + I_{\text{M4}} \approx -I_{\text{M2}} + I_{\text{M4}}
\]

\[
\approx -I_{\text{th0}} \cdot e^{\beta_0} \cdot e^{\frac{\lambda_{VQ}V_{\text{thn}}}{\sqrt{n}} - \beta_0} + I_{\text{th0}} \cdot e^{\beta_0} \cdot e^{\frac{\lambda_{VQ}V_{\text{thn}}}{\sqrt{n}V_{\text{DD}}}}
\]

\[
= -e^{\beta_0} \left( I_{\text{th0}} \cdot e^{\frac{\lambda_{VQ}V_{\text{thn}}}{\sqrt{n}V_{\text{DD}}}} - \beta_0 \right)
\]

where \( p_n(V_{\text{WL}}, V_{\text{thn}}) = K_1p_n([V_{\text{WL}} - V_{\text{thn}}]/nV_t) + K_2p_n([V_{\text{WL}} - V_{\text{thn}}]/nV_t)^2 \), \( p_p(V_{\text{Q}} - V_{\text{DD}}, V_{\text{thn}}) = K_{1p}([V_{\text{Q}} - V_{\text{DD}} - V_{\text{thp}}]/nV_p) + K_{2p}([V_{\text{Q}} - V_{\text{DD}} - V_{\text{thp}}]/nV_p)^2 \), and \( \beta = e^{\beta_0} - p_{n} \). Because the first term in (16) is several times larger than the second in the parenthesis, same as the approximation done in (13), we replace \( \beta \) with its mean value \( \beta_0 \). Then, take (16) into (5) and separate \( e^{\beta_0} \) from the denominator, the minimum write-time becomes

\[
t \approx -CQe^{-p_n(V_{\text{WL}}, V_{\text{thn}})} \cdot w(V_{\text{TRIP,R}})
\]

\[
w(v) = \int_{V_{\text{DD}}}^{v} \frac{dV_Q}{I_{\text{th0}} \cdot e^{\frac{\lambda_{VQ}V_{\text{thn}}}{\sqrt{n}V_{\text{DD}}}} - \beta_0} \cdot e^{\frac{\lambda_{VQ}V_{\text{thn}}}{\sqrt{n}V_{\text{DD}}}}
\]

where \( w(v) \) is the function of \( V_{\text{TRIP,R}} \) and the process parameters of M4 and M2. \( V_{\text{TRIP,R}} \) can be obtained from (6), which relates to the parameter variations of M5, M3, and M1. However, it does not have an analytical solution using the
proposed \( I_{ds} \) model, which further makes (17) an implicit function and the PDF of write time cannot be derived. To obtain an analytical expression, similar to the prior works [1], [2], we neglect the variation of \( V_{\text{TRIP, R}} \) here and set it to the nominal value. Furthermore, modeling variations of \( M2 \) and \( M4 \) are more important than those of \( M3 \) and \( M1 \) because discharging node \( Q \) is faster than charging the node \( QB \) and determines the minimum write time (discussed in Section II-B). Fortunately, the simplification only results in a little underestimation (Fig. 10).

Now, the logarithm of \( t \) is proportional to \( p_n(V_{\text{WL}}, V_{\text{thn}}) \)

\[
\ln(t) \propto -K_{2n} \left( \frac{V_{\text{WL}} - V_{\text{thn}}}{n_q v_t} + \frac{K_{1n}}{2K_{2n}} \right)^2. \tag{18}
\]

Since \( V_{\text{WL}} \) is a fixed value and \( V_{\text{thn}} \) is an independent Gaussian random variable, the logarithm of write time follows a non-central Chi-square distribution with 1 degree of freedom. The non-Gaussian \( t \) is now a simple nonlinear function of a Gaussian random variable, \( \sqrt{\ln(t)} \sim N(\mu_w, \sigma_w^2) \). The final PDF of \( T_{\text{WRITE}} \) is

\[
f_w(t) = \frac{1}{2 \sigma_w \cdot t \sqrt{2 \pi} \cdot \ln(t)} e^{-\left(\frac{\ln(t) - \mu_w}{2\sigma_w} \right)^2}. \tag{19}
\]

where \( \mu_w \) and \( \sigma_w \) are the mean and standard deviation of \( \sqrt{\ln(t)} \). Finally, the write yield can be calculated through (19) and (7).

Fig. 4 shows the tail part of the whole Q-Q plot of \( T_{\text{WRITE}} \). The parameters \( \mu_w \) and \( \sigma_w \) in (19) and the \( y \)-axis quantiles are calculated from \( 1.3 \times 10^5 \) samples of \( T_{\text{WRITE}} \) using MC simulations at 0.5-V 25 \(^\circ\)C TTG. Quantile samples approximately lie on a line demonstrating that (19) models the distribution of \( T_{\text{WRITE}} \) accurately.

\section*{D. Failure Models With Read/Write Assist Circuits}

Many SRAM designs use read-assist circuits to avoid read upsets during access and enhance read stability. For instance, as Fig. 1 shows, WL under-drive [28] weakens the access transistor to keep \( V_{QB} \) low. Cell \( V_{\text{DD boost}} \) boost [29] increases the operating voltage of cells with a separated power supply to strengthen the pull-down nMOS and increase \( V_{\text{TRIP}} \) during reading. Our models can also be applied in SRAM with read-assist techniques. When WL under-drive is employed, it is equivalent to reducing \( V_{\text{WL}} \) by a fixed value. The resistance of \( M3 \) increases because of its lowered \( V_{gs} \). Our assumption of the zero substrate bias of \( M3 \) still holds. Then, \( V_{\text{WL}} \) in (10)–(13) is replaced with the reduced WL-driven voltage \( V_{\text{WLUD}} \). Equation (14) becomes:

\[
\Delta V \propto K_2 \left( \frac{V_{\text{WLUD}} - V_{\text{thn}}}{n_q v_t} + \frac{K_{1n}}{2K_{2n}} \right)^2. \tag{20}
\]

The shape of the PDF in (15) does not change. Similarly, when using an increased cell supply voltage (\( V_{\text{DDC}} \)), the rise of \( V_{\text{DDC}} \) only affects the value of \( \mu_\Delta \) and \( \sigma_\Delta \), while keeping our assumption and the form of PDF unchanged.

For write-assist techniques, as shown in Fig. 2, boosting the WL-driven voltage [30] helps the write margin by enhancing the contention ability of the access nMOS \( M2 \) and speeding up the completion of writing a 1 to the other side. It can be considered as increasing \( V_{\text{WL}} \) by a fixed value (\( V_{\text{WLB}} \)). The stronger \( M2 \) with a larger write current further reduces the error of our approximation in (17). By replacing \( V_{\text{WL}} \) with \( V_{\text{WLB}} \) in (16)–(18), the logarithm of write time is proportional to

\[
\ln(t) \propto -K_{2n} \left( \frac{V_{\text{WLB}} - V_{\text{thn}}}{n_q v_t} + \frac{K_{1n}}{2K_{2n}} \right)^2. \tag{21}
\]

This does not change the form of PDF in (19). In another scenario, cell supply voltage collapse [31] suppresses the latching capability of an SRAM cell to make a fast and easy write operation. This method reduces the \( V_{\text{ds}} \) and \( V_{gs} \) of \( M4 \) to \( V_Q - V_{\text{DDC}} \) and \( V_{QB} - V_{\text{DDC}} \), respectively. The trip point of the right-side inverter is also reduced to \( V_{\text{TRIP, RC}} \). Then, \( w \) in (17) is rewritten as

\[
w(V_{\text{TRIP, RC}}) = \int_{V_{\text{DD}}}^{\lambda w V_{Q}} \frac{dV_{Q}}{\beta_0 I_{0} e^{\frac{-\lambda w V_{Q}}{n_v V_{Q}}} - \beta_0 I_{0} e^{\frac{-\lambda w V_{Q} - V_{\text{DDC}}}{n_v V_{Q}}}}. \tag{22}
\]

However, all the small differences between the derived functions with or without write assist have no impact on our final conclusion but only the exact values of \( \mu_w \) and \( \sigma_w \) in (19).

In the next section, we will validate our analytical yield models in the scenarios where the WL under-drive and WL boost techniques are applied in SRAM.

\section*{IV. Evaluation}

This section first shows the validation of our compact \( I_{ds} \) model, then compares the proposed failure models from both accuracy and efficiency perspectives.

All experiments are performed with HSPICE on a server with Intel Xeon Gold 5118 CPU @ 2.30 GHz. We use the TSMC 28-nm library based on the BSIM 4.5.0 model [27] to set up our experiments. There are 402 process parameters for each transistor totally, including 78 global variables and 14 local variables, such as \( V_{\text{tho}} \), coefficient of the drain voltage reduction (\( \eta \)), low field mobility at nominal temperature (\( u_0 \)), electrical gate oxide thickness in meters (\( t_{\text{oxe}} \),
and channel doping concentration at the depletion edge for the zero body bias ($n_{d_{depl}}$). All process parameters’ variabilities are turned on in our simulations. The BSIM model [27] is an accurate device model and integrated into the SPICE tool. Both global and local process parameters are Gaussian random variables whose mean values and standard deviations are defined in the technology library. The variations of device geometry, spatial variations, and other correlated variations are also set in the technology library. The standard 6T SRAM demo is designed by TSMC existing in their PDK. The nominal supply voltage of the SRAM transistors is 0.8 V.

In the experiment part, we use different global process corners to prove the performance of our cell failure models. TTG, FFG, and FSG represent different types of global process corners, where TTG stands for typical nMOS typical pMOS, FFG for fast nMOS fast pMOS, and FSG for fast nMOS slow pMOS.

### A. Validation of the Compact $I_{ds}$ Model

We first collect $I_{ds}$ versus $V_{gs}$ and $I_{ds}$ versus $V_{ds}$ data through SPICE simulations with the DC analysis. The data set is used to train our $I_{ds}$ model for each type of device with a minimum size. The range of $V_{gs}$ and $V_{ds}$ is from 0 to 0.7 V with a step of 0.01 V. Then, the *nlinfit* procedure in MATLAB is used for model constants fitting in (8). Note that fitting constants (listed in Table II) in our compact model are unrelated to variations of process parameters defined in the technology library, which means the proposed $I_{ds}$ model can also be utilized in other CMOS technology.

Fig. 5 depicts different compact drain-current models against simulation results. Due to the oversimplified relation between inversion charge and the terminal voltage, the classic model [16] introduces strike errors in the near-threshold region. The improved work [17] performs well in modeling $I_{ds}$ versus $V_{gs}$, but not well in modeling $I_{ds}$ versus $V_{ds}$. This is because the DIBL effect makes $I_{ds}$ increase with $V_{ds}$ in the velocity saturated region. Our model shows a good match to simulation results across different $V_{gs}$ and $V_{ds}$.

The proposed model can also predict the drain current of other types of transistors with high, standard, and low $V_{th}$. Table II lists evaluation errors at TTG 25 °C. The max/average relative errors are calculated in the saturated region (where $V_{ds} > V_{dsat}$). The maximum relative error is only 12% for the high-threshold pMOS (denoted as pch_hvt). The average relative error in the saturated region is always below 4.2% across different types of transistors. The decrease of average errors for low-$V_{th}$ devices at lower $V_{gs}$ is because the model depicts the tendency of $I_{ds}$ more precisely at sub/near-threshold voltages (already shown in Fig. 5(a)).

### B. Model Predictions Versus Monte Carlo Results

Simulation results are collected from a 256-row by 64-column sized array with a word-line driver. Criteria of all types of failures in our experiments have been discussed in Section II-B. All initial data of cells is set to 1 to match the assumption during modeling. For the access-time failure simulation, we first run MC simulations with SA circuits to

| Param. | $I_0$ | $K_f$ | $K_s$ | $\lambda$ | $V_{gs}=0.6V$ Max. error | Avg. error | $V_{gs}=0.4V$ Max. error | Avg. error |
|--------|-------|-------|-------|--------|----------------------|-----------|----------------------|-----------|
| nch_hvt | 5.3796E-6 | 0.3981 | -0.0296 | 0.0201 | 7.7% | 3.7% | 6.6% | 3.3% |
| pch_hvt | -1.3225E-6 | -0.5068 | -0.0360 | 0.0315 | 12.0% | 3.4% | 10.8% | 3.2% |
| nch_vtt | 2.2900E-5 | 0.1411 | -0.0028 | 0.0012 | 5.8% | 3.9% | 9.1% | 4.0% |
| pch_vtt | -7.9742E-6 | -0.3102 | -0.0093 | 0.0014 | 5.3% | 3.6% | 7.9% | 3.6% |
| nch_vlt | 1.4854E-5 | 0.3157 | -0.0118 | 0.0161 | 7.4% | 2.6% | 2.9% | 2.0% |
| pch_vlt | -1.5647E-5 | -0.2547 | -0.0082 | 0.0135 | 10.5% | 4.2% | 3.9% | 1.8% |
obtain a data set of $V_{OS}$. Then, the voltage differences $\Delta V$ are collected from an SRAM array and compared to the $V_{OS}$ in new MC simulations. For the write-failure simulation, we first conduct DC analysis for the right side of a cell, collecting samples of $V_{TRIP_R}$ through MC sweeps. Then, new MC simulations with transient analysis are run, and the voltage of the node Q of the written cell is compared to the corresponding $V_{TRIP_R}$ to find the minimum write time.

Considering acceptable time/storage overhead, we run $1.3 \times 10^7$ MC simulations that account for $4\sigma$ $(3.17 \times 10^{-5})$ of failure estimation with a 95% confidence interval as the gold standard for each PVT. This magnitude of yield is also adopted by other yield analysis methods [5], [7], [12]–[14]. Note that our model has the capability to predict the failure probabilities at a higher-$\sigma$ region, however, the validation at $6\sigma$ $(9.87 \times 10^{-10})$ requires more than $10^{13}$ samples to simulate which may take several years to run. The relative error is defined as

$$\text{error} = \left| \frac{P_f - \hat{P}_f}{P_f} \right| \times 100\% \quad (23)$$

where $P_f$ is the failure probability estimated by MC simulations and $\hat{P}_f$ is the prediction from our access-time or write-failure probability models.

Fig. 6 compares the relative errors at $4\sigma$ between other access-time failure models and ours. The failure model [1] has larger deviations across all operating voltages because it only considers the variations of $I_{READ}$ and assumes access time as a Gaussian random variable. VAM [9] that only models $I_{READ}$ distribution also has large errors because $I_{READ}$ is a function of the two factors, $\Delta V$ and $T_{WL}$. In work [3], the differential voltage $\Delta V$ is supposed to follow the Gaussian distribution, which results in the error growth at a lower $V_{DD}$. Our model improves the evaluation accuracy of [1], [3], and [9] by $31.6 \times$, $17.6 \times$, and $25.9 \times$ at 0.5 V, respectively. Studies [12] and [13] mentioned in Section I are the importance-sampling-based methods and achieve similar accuracy but with more data samples compared to our work (Table III).

Fig. 7(a) shows failure probabilities down to the order of $10^{-6}$ at 25 °C TTG. Our work achieves great accuracy with the 15.7% maximum error and 8.8% average error in the wide range of $T_{READ}$ across different voltages. As $V_{DD}$ reduces, the minimum access time at 0.5 V satisfying the $4\sigma$ yield constraint is $7.85 \times$ longer than that at 0.7 V. Such a performance degradation makes $V_{MIN}$ of the SRAM module 0.1 to 0.2 V higher than $V_{MIN}$ of the logic module. Fig. 7(b) further shows the yield changes with temperatures at 0.5-V FFG. The model prediction errors are 4.9%, 6.4%, 10.5%, and 12.7% at $4\sigma$ for $-25$ °C, 25 °C, 75 °C, and
125 °C, respectively. At 75 °C, the $T_{READ}$ is only 50% of that at 25 °C, while increases to 1.41 × at −25 °C. This character is fully analyzed and leveraged by several low-power SRAM designs [20], [23].

Fig. 8 shows $T_{READ}$ at different WL voltages using WL under-drive assist circuits. The maximum relative error of our model is 9.7% with 0.525-V $V_{WL}$. The read assist circuits sacrifice the access delay for reducing the read upset probability. A 0.1-V $V_{WL}$ reduction worsens $T_{READ}$ nearly 2× at 0.6-V $V_{DD}$, by contrast, a 0.25-V reduction has no significant impact on the read speed. Such a tradeoff between the WL voltage and the access delay is the motivation of the design [32].

Regarding write failure estimations, Fig. 9 compares different write failure models at the 25 °C TTG corner. The failure model [1] uses a noncentral $F$ distribution to describe write time. Statistical timing models [10] and [11] for STA using the log-normal (LN) and LSN distributions yield similar precisions. Errors of their predictions grow slowly as $V_{DD}$ scales down, caused by the assumption of unchanged $V_{ds}$ and regular input pules. The accuracy improvements of our model are 15×, 7.5×, and 6.3× at 0.5 V compared to models [1], [10], and [11], respectively. AIS [12] shows its instability with 34.1% relative error at 0.7 V in the write-failures prediction. The maximum error of our model is 17.3% at 0.55-V FFG. Fig. 10 shows the model predictions against the MC simulation results at different Fig. 10(a) voltages and Fig. 10(b) temperatures. For a write operation, driving the internal storage node with small capacitance is much faster than the read operation. Thus, $T_{WRITE}$ has a concentrated distribution with a very small mean and variance but a large skewness. Such a distribution increases the average errors of our model predictions and makes itself very sensitive to PVT. For example, in the 4σ region, the write delay is shortened to 52% when $V_{DD}$ scales from 0.5 to 0.55 V and 42% when the temperature rises from 25 °C to 75 °C. However, the average error of our model across all PVT is always below 10.4% in a wide range of failure probabilities. The underestimation of our write failure model is mainly caused by ignoring the variations of $V_{TRIP_R}$ in (17), equivalent to not taking into account the process variations of M5, M3, and M1. Besides, the underestimated error grows as the voltage or temperature goes down. This demonstrates that a lower voltage
or temperature enhances the impact of process variations on memory transistors.

Fig. 11 depicts the normalized $T_{\text{WRITE}}$ and relative errors of the write failure model versus different WL voltages using WL-boost assist circuits with the $4\sigma$ yield constraint. The maximum relative error of our model is 16.1% with 0.6-V $V_{\text{WL}}$. The assist circuits enhance both the write margin and speed at low supply voltages. A 25-mV $V_{\text{WL}}$ boost reduces more than half of $T_{\text{WRITE}}$ at 0.5 V 25 °C. However, the write assist technique introduces additional power and area overhead, especially when a large boost capacitor is used [30].

Different from other studies [9]–[11], [14], our simulations run with full process parameter variations of all transistors in the netlist. As a consequence, the evaluation results are closer to the real situation in the design flow and more valuable for designers.

To further compare the efficiency of our model, Table III lists the number of samples required by several importance-sampling methods to evaluate the $4\sigma$ failure probability. The number of samples of the proposed models is significantly smaller compared with other sampling-based methods. For each PVT condition, we collect 200 MC samples of $\Delta V$ to calculate $\mu_\Delta$ and $\sigma_\Delta$ in (15) and 200 samples of offset voltages to obtain $\mu_{V_{\text{os}}}$ and $\sigma_{V_{\text{os}}}$ in (4), where the size of the data set is $43.6 \times$ and $127.5 \times$ smaller than that of AIS [12] and HDIS [13], respectively. For the write failure prediction, 1600 $T_{\text{WRITE}}$ samples are enough to quantify $\mu_W$ and $\sigma_W$ in (19) and achieve the best accuracy. Furthermore, we find that the number of samples will affect the final evaluation results where larger samples may result in overfitting the given PDF. The importance-sampling-based methods need a new presampling procedure as PVT changes. Moreover, the failure prediction of AIS is limited to a fixed access time $T_{\text{READ}}$, and it needs to be reconstructed when the timing constraint has changed. Our models can evaluate a wide range of failure probabilities since the PDFs of interest metrics have been constructed.

V. Conclusion

Yield estimation for SRAM cells requires a significantly high $\sigma$ range. It is a hard but critical task because the yield results is an important guideline for choosing error-tolerant strategies or read/write-assist circuit designs for low-power SRAMs. Unfortunately, the precision and the speed of yield analysis are a tradeoff. In this article, we proposed an empirical compact $I_{ds}$ model for nanoscale transistors operated at near/subthreshold voltages is constructed. By using a more accurate $I_{ds}$ model as a foundation, we further proposed two analytical failure models to evaluate the probabilities of the access-time failure and the write failure in an SRAM cell. The evaluation results showed that our models can provide a more accurate prediction with a remarkable speedup compared to the state-of-the-art failure models. Moreover, the proposed models have good scalabilities and contain physical insights.

REFERENCES

[1] S. Mukhopadhyay, H. Mahmoodi, and K. Roy, “Modeling of failure probability and statistical design of SRAM array for yield enhancement in nanoscaled CMOS,” IEEE Trans. Comput.-Aided Design Integr. Circuits Syst., vol. 24, no. 12, pp. 1859–1880, Dec. 2005.
[2] J. Das and S. Ghosh, “Energy barrier model of SRAM for improved energy and error rates,” IEEE Trans. Circuits Syst. I, Reg. Papers, vol. 61, no. 8, pp. 2299–2308, Aug. 2014.
[3] H. Kang, J. Kim, H. Jeong, Y. H. Yang, and S. Jung, “Architecture-aware analytical yield model for read access in static random access memory,” IEEE Trans. Very Large Scale Integr. (VLSI) Syst., vol. 23, no. 4, pp. 752–765, Apr. 2015.
[4] H. Jeong, Y. Yang, J. Lee, J. Kim, and S. Jung, “One-sided static noise margin and Gaussian-tail-fitting method for SRAM,” IEEE Trans. Very Large Scale Integr. (VLSI) Syst., vol. 22, no. 6, pp. 1262–1269, Jun. 2014.
[5] S. Shen, L. Pang, T. Shao, M. Ling, X. Shi, and L. Shi, “TYMER: A yield-based performance model for timing-spectrum SRAM,” Presented at the 57th ACM/IEEE Des. Autom. Conf. (DAC), 2020.
[6] K. Agarwal and S. Nassif, “Statistical analysis of SRAM cell stability,” in Proc. 43rd Annu. Des. Autom. Conf., 2006, pp. 57–62.
[7] X. Fan, R. Wang, and T. Gemmeke, “Physical modeling of bitcell stability in subthreshold SRAMs for leakage–area optimization under PVT variations,” in Proc. Int. Conf. Comput.-Aided Des., 2018, pp. 1–8.
[8] W. Dong, P. Li, and G. M. Huang, “SRAM dynamic stability: Theory, variability and analysis,” in Proc. IEEE/ACM Int. Conf. Comput.-Aided Des., 2008, pp. 378–385.
[9] J. Wang et al., “Non-Gaussian distribution of SRAM read current and design impact to low power memory using voltage acceleration method,” in Proc. Symp. VLSI, 2011, pp. 220–221.
[10] J. Shiomi, T. Ishihara, and H. Onodera, “Microarchitectural-level statistical timing models for near-threshold circuit design,” in Proc. 20th Asia South Pac. Des. Autom. Conf., 2015, pp. 87–93.
[11] P. Cao, Z. Liu, J. Guo, and J. Wu, “An analytical gate delay model in near/subthreshold domain considering process variation,” IEEE Access, vol. 7, pp. 171515–171524, 2019.
[12] X. Shi, F. Liu, J. Yang, and L. He, “A fast and robust failure analysis of memory circuits using adaptive importance sampling method,” in Proc. 55th ACM/ESDA/IEEE Des. Autom. Conf. (DAC), 2018, pp. 1–6.
[13] W. Wu, F. Gong, G. Chen, and L. He, “A fast and provably bounded failure analysis of memory circuits in high dimensions,” in Proc. 19th Asia South Pacific Des. Autom. Conf. (ASP-DAC), 2014, pp. 424–429.
[14] X. Shi, H. Yan, Q. Huang, J. Zhang, L. Shi, and L. He, “Meta-model based high-dimensional yield analysis using low-rank tensor approximation,” in Proc. DAC, 2019, pp. 1–6.
[15] L. Pang, S. Shen, and M. Yao, “A spline-dimensional model representation for SRAM yield estimation in high yield and high dimensional scenarios,” IEEE Access, vol. 9, pp. 47320–47329, 2021.
[16] J. Rabaey, Low Power Design Essentials. New York, NY, USA: Springer, 2009.
Interesting article on various research topics in the field of integrated circuits and systems. The authors discuss various aspects such as timing analysis, SRAM operation, and low-voltage VLSI designs. The research interests of the authors span from statistical timing analysis to memory designs.

Shan Shen was born in 1993. He received the B.S. degree from the Microelectronics Department, Jiangnan University, Wuxi, China, in 2016, and the Ph.D. degree from the School of Electronic Science and Engineering, Southeast University, Nanjing, China.

His research interests mainly include EDA algorithms and memory designs.

Peng Cao (Member, IEEE) received the B.S. and Ph.D. degrees in microelectronics and solid state electronics from Southeast University, Nanjing, China, in 2002 and 2010, respectively.

He joined research with the University of Waterloo, Waterloo, ON, Canada, from 2016 to 2017 as a Visiting Scholar. He is currently an Associate Professor with the National ASIC System Engineering Research Center, Southeast University. His research interests are focused on statistical timing analysis and low-voltage VLSI designs.

Ming Ling (Member, IEEE) received the B.S., M.S., and Ph.D. degrees from Southeast University, Nanjing, China, in 1994, 2001, and 2011, respectively.

He is an Associate Professor working with the National ASIC System Engineering Technology Research Center, Southeast University. His main research interests include memory subsystem of SoC, embedded software, and SoC architecture.

Longxing Shi (Senior Member, IEEE) received the B.S., M.S., and Ph.D. degrees from Southeast University, Nanjing, China, in 1984, 1987, and 1992, respectively.

He was an Associate Professor with the School of Electronic Science and Engineering, Southeast University, from 1992 to 2000, where he has been a Professor and the Dean of the National ASIC System Engineering Research Center since 2001. He has authored one book and over 130 articles. His current research interests include ultralow-power IC design.