Fundamental Thermal Limits on Data Retention in Low-Voltage CMOS Latches and SRAM

Elahe Rezaei, *Member, IEEE*, Marco Donato, *Member, IEEE*, William R. Patterson, *Life Member, IEEE*, Alexander Zaslavsky, and R. Iris Bahar, *Senior Member, IEEE*

***Abstract*—Ultra-low-power systems with substantial computing capacity require latches and SRAMs to operate at extremely low supply voltages. However, with aggressive technology scaling, reliability becomes a major challenge due to unavoidable process variations and the presence of multiple noise sources, including intrinsic thermal noise. This paper provides a quantitative measure of reliability by calculating the probability distribution function (PDF) of errors induced by thermal noise in latches and SRAMs operating in subthreshold conditions. Implemented in a novel simulation tool for thermal-noise analysis of CMOS circuits (STTACC), our algorithm uses a stochastic differential equation circuit model that preserves the proper Poissonian statistics for thermal-noise-driven current fluctuations in MOSFETs. Our probabilistic error model can handle error rate analysis for arrays of latches or full SRAMs on time scales from seconds to years without excessive computational overhead. We demonstrate that the time-to-error (TTE) statistics of subthreshold SRAMs obey log-normal distributions that depend on parameters such as node and device capacitance, device threshold variations and operating conditions of supply voltage and temperature. This makes it possible to quantitatively evaluate the asymptotic behavior of extremely rare error events that are inaccessible to standard SPICE-based simulations.**

***Index Terms*—Thermal noise-induced errors, subthreshold CMOS SRAM, time-domain simulation.**

I. Introduction

EMBEDDED systems for Internet-of-Things (IoT), wearable electronics and medical devices, are generally designed to operate under strict power constraints. As on-chip SRAM arrays often consume much of the area in a System-on-Chip (SoC), optimizing their power consumption is critical for enabling energy-efficient computing.

Manuscript received January 14, 2020; revised March 31, 2020; accepted May 16, 2020. Date of publication May 22, 2020; date of current version September 3, 2020. This work was supported by the National Science Foundation Grant through the CISE Directorate, under Grant CCF1525486. *(Corresponding author: Elahe Rezaei.)*

Elahe Rezaei was with the School of Engineering, Brown University, Providence, RI 02912 USA. She is now with the Department of Machine Learning-HW, Qualcomm Technologies Inc., San Diego, CA 92121 USA (e-mail: elahe.rezaei@ieee.org).

Marco Donato is with the School of Engineering and Applied Sciences, Harvard University, Cambridge, MA 02138 USA (e-mail: mdonato@seas.harvard.edu).

William R. Patterson, Alexander Zaslavsky, and R. Iris Bahar are with the School of Engineering, Brown University, Providence, RI 02912 USA (e-mail: william_patterson_iii@brown.edu; alexander_zaslavsky@brown.edu; iris_bahar@brown.edu).

Color versions of one or more of the figures in this article are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TDMR.2020.2996627

In the domain of ultra-low power (ULP) applications, lowering the V_{DD} below device threshold voltage V_{TH} both reduces memory power and may also optimize total power in the circuits in which the memory is embedded. However, ULP-optimized designs come at the cost of reduced reliability. In particular, due to the exponential relationship of subthreshold currents on V_{TH}, process variations are a major concern as they can compromise SRAM data retention [1].

Furthermore, even in non-ULP systems, idle sections are commonly operated at reduced supply voltage to save power. Obviously, such a strategy must be limited to a minimum V_{DD} that guarantees data retention, a limit that is likely reached when the system is in subthreshold operation [2]. In this scenario, SRAM bitcells are even more susceptible to process variations [3]–[5] and intrinsic noise sources, due to aggressive technology scaling, and thereby suffer increasing bit error rates [6], [7].

In 1969, Keyes pointed out that data storage requires nonlinear devices [8]. Since nonlinearity only occurs on the scale of the thermal voltage, kT/q, he speculated that data retention would set a fundamental limit on V_{DD} of some multiple of kT/q due to the unavoidable presence of thermal noise. The present article derives such a limit for CMOS SRAM that is based on a probabilistic measure of minimum retention time and is only weakly dependent on process technology.

Accurate thermal noise modeling in ULP subthreshold CMOS circuitry requires a precise understanding of the probabilistic behavior of current fluctuations. Traditionally, thermal noise has been modeled as stationary additive white Gaussian noise with zero mean and constant variance that is inserted at the input of each logic gate [9], [10]. However, in aggressively scaled circuits, the number of electrons stored on a nodal capacitance is very small and even a fluctuation of two or three electrons can have a significant effect. This is aggravated by the highly nonlinear voltage dependence of drain currents in subthreshold operation. In subthreshold operation, MOSFET drain-source current arises from electrons transiting the channel in forward and reverse directions with probabilities described by a two-sided Poisson process, as originally pointed out by Sarpeshkar *et al.* [11]. Each electron changes the gate-source and drain-source voltages. This affects the probability of subsequent electron transfers, so the statistics are non-stationary. To simulate such a process requires very short simulation times over which the rates are approximately constant and the number of events is small enough to require attention to Poissonian statistics [11], [12]. Previously, we

proposed a fast time-domain subthreshold noise simulation engine [13] based on the solution of stochastic differential equations (SDE) that achieved orders of magnitude speed-up compared to SPICE-based simulations. Recently, we extended this SDE-based model to subthreshold circuits based on a central pair of cross-coupled inverters (e.g., latches and SRAM cells) by including all of the channel and displacement currents coupling the inverters into the formulation of the driving SDE [14].

We are able to detect thermal-noise-driven bit-flip errors using Monte Carlo simulation to solve a probabilistic model of the incremental evolution of transients. Direct SPICE-based time domain simulation cannot capture such bit-flips, as they are too infrequent. Instead, our algorithm performs repeated simulations of nodal charge fluctuations on a sub-picosecond time scale over small voltage ranges. It accumulates the probabilities of moderately improbable events to compute the time-to-error (TTE) distribution for a subthreshold SRAM latch. Based on these results, we also find that the mean time to error [12], [15]–[17] (MTTE) is not an informative measure of thermal-noise-induced error statistics.

In this work, we present STTACC, an extension of our time-domain simulation framework [14] that specifically targets the quantitative statistics and detection of bit-flip errors in subthreshold SRAM cells and latches. Our simulation framework addresses the reliability of future generations of SRAM devices beyond the current state-of-the-art. Indeed, the simulation results presented in Section IV are based on a 7nm predictive technology [18], for which test circuits are not generally available, and can capture bit-flip errors that are sufficiently rare to be undetectable in reasonably-sized test chips. In addition to describing a detailed implementation of our simulation framework, we now present several new results:

- We demonstrate the fundamental limits of reliable operation for subthreshold SRAM cells due to thermal noise fluctuations, and identify the log-normal distribution as the appropriate description for the time-to-error (TTE) statistics;
- We introduce the use of cumulative distribution functions (CDFs) of TTEs as an alternative methodology to the mean time to error (MTTE) for evaluating SRAM reliability;
- We explore the bit-flip resiliency of SRAM cells by determining their sensitivity to process variations, temperature and operating voltage in a model 7nm CMOS technology node.

The remainder of this paper is organized as follows: Section II describes our thermal noise model and the SDE-based formulation for the operation of subthreshold CMOS SRAM cells; Section III provides the implementation details and features of STTACC; Section IV reports our STTACC-based simulation results showing the impact of technology parameter variation on noise immunity using as an example a 7nm predictive technology model from the ASAP7 PDK [18]; and finally, Section V contains concluding remarks and suggestions for future work.

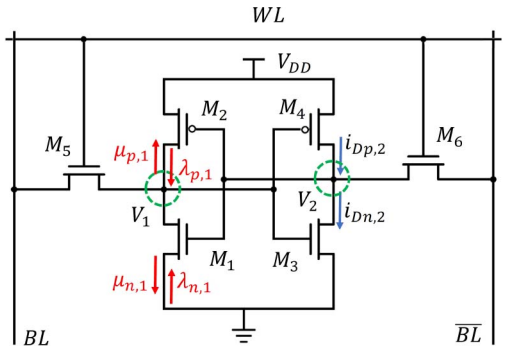

Fig. 1. The structure of a conventional 6T (six-transistor) CMOS SRAM bitcell. The Poisson charging rates (shown in red) can be derived from the forward and reverse components of the transistors' drain currents (shown in blue).

II. Thermal Noise Modeling

In this section, we provide an overview of a statistically accurate Poisson model of thermal fluctuations in the transistor currents. Based on this model, we develop a system of SDEs that describe the transient behavior of a 6T SRAM cell.

A. Stochastic Thermal Noise Fluctuations in Subthreshold CMOS

The drain current of a transistor in subthreshold is given by [19], [20]:

$$I_D = I_0 \exp\left(\frac{qV_{gs}}{mkT}\right) \exp\left(\frac{qBV_{ds}}{kT}\right)\left[1 - \exp\left(-\frac{qV_{ds}}{kT}\right)\right], \quad (1)$$

where B is the DIBL parameter, and I_0 as well as the body factor m are technology-dependent factors. This drain current can be decomposed into opposing forward and reverse electron flows, characterized by average rates $\mu_{n,p}$ and $\lambda_{n,p}$ for NMOS and PMOS transistors respectively (in electrons per second, after division by electron charge q) [13]:

$$\mu_n = \frac{I_0}{q} \exp\left(\frac{qBV_{ds}}{kT}\right) \exp\left(\frac{qV_{gs}}{mkT}\right) \quad (2)$$

$$\lambda_n = \mu_n \exp\left(\frac{-qV_{ds}}{kT}\right) \quad (3)$$

$$\mu_p = \frac{I_0}{q} \exp\left(\frac{qBV_{sd}}{kT}\right) \exp\left(\frac{qV_{sg}}{mkT}\right) \quad (4)$$

$$\lambda_p = \mu_p \exp\left(\frac{-qV_{sd}}{kT}\right), \quad (5)$$

as illustrated in Fig. 1. The intrinsically stochastic nature of these two flows allows us to model the components as independent Poisson processes [11]:

$$X_\mu \sim Poisson(\mu)$$
$$X_\lambda \sim Poisson(\lambda),$$

where X_μ and X_λ are time series of Poisson-distributed random variables with means μ and λ.

B. SDE-Based Formulation for CMOS SRAM Operation

The fundamental storage element for this study is the basic six-transistor (6T) CMOS SRAM shown in Fig. 1, where transistors M_1, M_2, M_3, M_4 implement the cross-coupled inverters

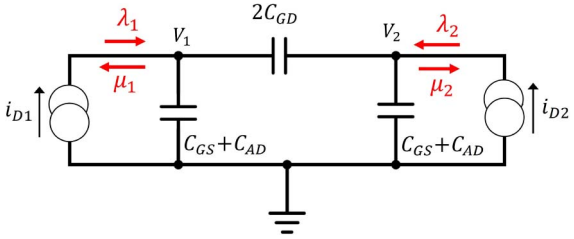

Fig. 2. Equivalent circuit for 6T CMOS SRAM bitcell of Fig. 1 where the Miller and gate-source capacitances of the inverters are C_{GD} and C_{GS}. The C_{AD} term models the additional capacitance due to access transistors and wiring.

and M_5 and M_6 are the access transistors. We can apply the same methodology introduced in [13] to develop the SDEs for this cell.

Fig. 2 shows the equivalent circuit of the SRAM cell, where the current sources i_{D1} and i_{D2} represent the net drain currents at each node. Capacitances C_{GD} and C_{GS} denote the gate-drain (Miller) and gate-source capacitances of the inverters, with an additional capacitance C_{AD} modeling the effect of access transistors and other parasitics. For circuits operated in sub-threshold with a total voltage swing of ~200 mV, all transistor and parasitic capacitances are roughly constant, whereas the charging and discharging rates vary significantly as a function of node voltages. The Kirchhoff current law (KCL) equations at nodes 1 and 2 in Fig. 2 can be written in matrix form as:

$$\begin{bmatrix} i_{D_1} \\ i_{D_2} \end{bmatrix} dt = \begin{bmatrix} C_{GS}+C_{AD}+2C_{GD} & -2C_{GD} \\ -2C_{GD} & C_{GS}+C_{AD}+2C_{GD} \end{bmatrix} \begin{bmatrix} dV_1 \\ dV_2 \end{bmatrix}. \tag{6}$$

Equation (6) becomes a set of SDEs when the drain currents are treated explicitly as random processes with $i_{D_1} = q(X_{\lambda_1} - X_{\mu_1})$ and $i_{D_2} = q(X_{\lambda_2} - X_{\mu_2})$. The Poisson rates of an inverter, e.g., $\lambda_1 = \lambda_{p,1} + \lambda_{n,1}$ and $\mu_1 = \mu_{p,1} + \mu_{n,1}$ for the first inverter, are highly voltage-dependent, so these SDEs are nonlinear.

To solve (6) for possible trajectories of the voltage state of the system, we use Euler's method with a time step $\delta t = t_n - t_{n-1}$ small enough to ensure a negligible probability of a significant change in λ's or μ's within a time step.[1] These rates are updated at the end of each time step. With these assumptions, the solution can be written as:

$$\boldsymbol{V[t_n] = V[t_{n-1}] + q\ C^{-1}\ dX[t_{n-1}]}, \tag{7}$$

where $\boldsymbol{C}$ is the capacitance matrix in (6). The components of $\boldsymbol{dX}$ are changes in the number of electrons on each node $dX_i = X_i[t_n] - X_i[t_{n-1}]$ during a single time step.

III. STTACC Implementation

Figure 3 shows the transient response of voltages $V_1(t)$ and $V_2(t)$ calculated by our simulator using (7), with the time axis chosen to enclose an event or transition of interest and a 0.1 ns scale bar added for clarity. In equilibrium, the two logic states

[1]Typically δt is sub-picosecond and its product with λ or μ is less than one transferred electron per time step.

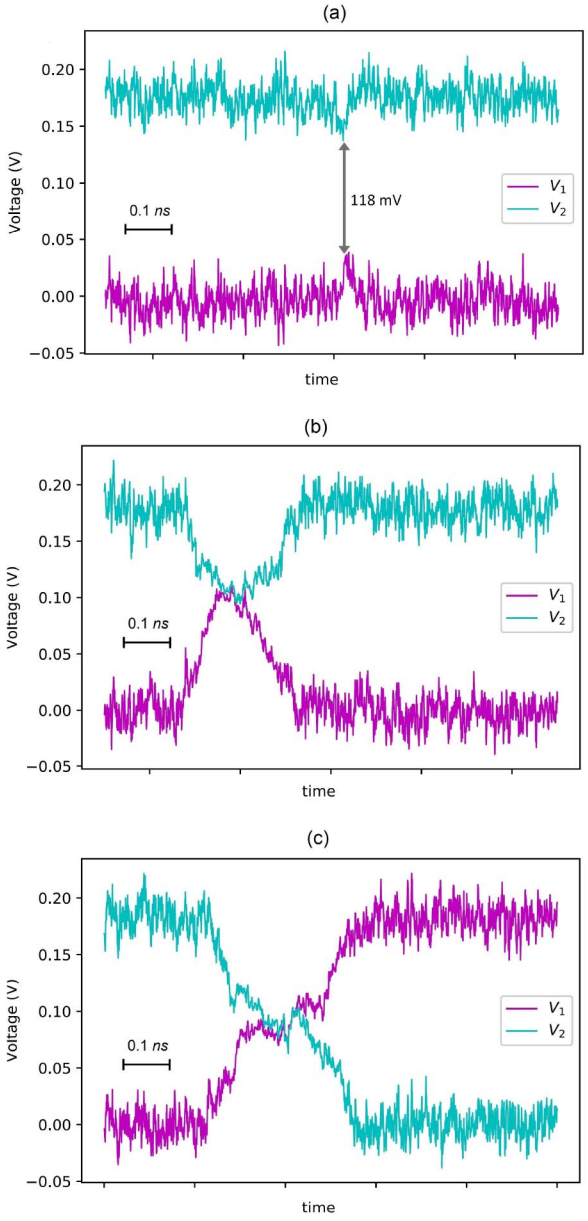

Fig. 3. (a) An event of $\Delta v(t) = 118$ mV extracted by STTACC using continuous simulation. (b) An event of $\Delta v(t) \leq 0$ mV with metastable resolution returning to equilibrium. (c) A bit-flip error event with $\Delta v(t) \ll 0$ mV. The shallow slope and extended middle region of this transition are due to metastability in the system as it passes through the switching threshold. The events in both (b) and (c) were extracted by STTACC using the incremental algorithm.

"1" and "0" ideally correspond to full charge or zero charge on the output node capacitances, and an inter-nodal voltage difference $\Delta v(t) \equiv V_2(t) - V_1(t) \simeq V_{DD}$. Thermal noise fluctuations on both storage nodes can decrease $\Delta v(t)$. Ultimately, if the voltage difference $\Delta v(t)$ approaches 0, positive feedback in the system can switch the logic state, resulting in a bit-flip error.

If we start with $\Delta v(t) = V_{DD} = 180$ mV, we can try to capture a bit-flip error by simulating until $\Delta v(t) < 0$. However, the minimum $\Delta v(t)$ captured by continuously computing the numerical solution of (7), after a month of continuous simulation, was 118 mV, as shown in Fig. 3(a). Thus, reaching

$\Delta v(t) \leq 0$ is computationally unfeasible as the simulation time grows exponentially for lower Δv values [14]. This issue is overcome by an incremental algorithm described below, where we take advantage of the fact that electron and hole charging/discharging rates depend only on the present instantaneous values of the node voltages $V_1(t)$ and $V_2(t)$ and not on history.

Our algorithm for calculating TTEs has two phases. The first phase starts from equilibrium and simulates continuously, looking for times when Δv goes below a preset threshold, L_1. For each such crossing event, we record the time $t[1]$ and values of $V_1[1]$ and $V_2[1]$. A large number of such runs is assembled to create a representative set of initial conditions. The values $t[1]$, $V_1[1]$, $V_2[1]$ are used to initialize the system for the second phase in which we calculate additional reductions of Δv by user-defined incremental amounts. At each stage n, the values $t[n]$, $V_1[n]$, $V_2[n]$, which correspond to the crossing of a threshold L_n, are used as the starting point to compute the decrement in the next stage, $n+1$, and the process repeats until a bit-flip occurs, see Fig. 3(b) and (c).

Algorithm 1: Incremental Algorithm Pseudo-Code for TTE Simulation

```
Result: TTE for bit-flip occurrence
Initialization;
while Δv(t) > 0 do
    Update λ(t), μ(t)
    Randomly generate X(t)
    Solve (7) and calculate Δv(t);
    if Δv(t) < L_n then
        Record V1[n], V2[n], t[n];
        Update L_n;
    else
        if Δv(t) > L_{n-2} then
            Exit
            Restart from V1[n − 1], V2[n − 1];
        else
            Continue;
        end
    end
end
```

A. Incremental Algorithm for Estimating TTEs

To accelerate the simulation of thermally-induced bit-flips, we combine the statistical accuracy of our SDE numerical model with the computational efficiency of memoryless stochastic processes. The Poisson processes governing the drain current imply that the evolution of $\Delta v(t)$ is independent of history. We exploit this in an incremental procedure for extracting the TTE as follows:

- We start the transient simulation by looking for a deviation from equilibrium, defined by $\Delta v(t) \leq L_1 \equiv V_{DD} - \rho_{init}$, where ρ_{init} is a user-defined initial decrement step. Once the desired event is captured, the event time is recorded, and the corresponding nodal voltage values, $V_1[1]$ and $V_2[1]$, are saved as a checkpoint.
- The simulator progresses incrementally to the next stages by updating the voltage deviation as $L_n \equiv L_{n-1} - \rho$, where ρ is another user-defined step size. While tracking $\Delta v(t)$ at stage n, if $\Delta v(t) > L_{n-2}$, indicating a return towards equilibrium, a loop exit condition is triggered. Then, the simulation is reset to the last checkpoint, $V_1 = V_1[n-1]$ and $V_2 = V_2[n-1]$.
- This procedure repeats until $\Delta v(t) \leq 0$. The simulator then completes the run allowing the voltages to reach their opposite equilibrium point, indicating a bit-flip.

At any given stage, as long as $\Delta v(t)$ is much larger than 0, the internal feedback of the latch tends to return it to equilibrium, which is why a bit-flip error is exponentially unlikely. However, by counting loop exit conditions and discarding their data at each stage, the simulation run-time can be reduced greatly. The time to reach stage n can be found via the recursive formula:

$$t_{\Delta v}[n] \approx (M_n + 1) \times t_{\Delta v}[n-1] + b_n, \tag{8}$$

where $t_{\Delta v}[n-1]$ is the time required to reach stage $(n-1)$, b_n is the time spent in stage n, and M_n counts the loop exits at stage n. In addition to the key improvement in simulation run-time provided by this algorithm, multi-threading is used to run multiple simulations in parallel and the computation time for bit-flip detection in a latch can be reduced to minutes.

The components of the vector $\boldsymbol{dX}$ in (7) come from a Poisson random number generator (RNG) that uses the average number of electrons per time step δt from the product of δt with the rates λ and μ for MOSFETs biased at the most recent values of V_1 and V_2. To account for the circuit's nonlinear response to noise fluctuations, these rates are updated at each simulation time step based on the instantaneous values of V_1 and V_2. For efficiency, all rates for all transistors are pre-computed and saved in a look-up table (LUT). STTACC is entirely built in C++, with the computational flow described in Fig. 4.

B. Simulator Setup

The simulation accuracy relies on characterizing the particular CMOS technology to extract the Poisson rates and the device capacitances. To optimize the computation, this characterization step is performed off-line. In this study, we base our characterization on a SPICE BSIM-CMG predictive model [21] for a 7nm FinFET technology [18]. The rate LUTs are extracted and stored for further use according to (1-4) by DC analysis sweeping the values of V_{GS} and V_{DS}. The minimum granularity of LUTs is 1 mV; i.e., smaller than the voltage changes of V_1 and V_2 due to the addition of a single electron to one of the nodes as estimated from (7). The transistor capacitances calculated in SPICE are used to define the lumped and coupling capacitances in appropriate gate models. The STTACC implementation uses a C++ gate class to enhance flexibility to handle more complex gates than just the inverters needed for latch data retention analysis. The gate class also includes methods to compute the net charging and discharging rates depending on the gate topology and to supply data to the computational engine efficiently.

C. Runtime Attributes

Our simulation approach requires several user-defined runtime parameters to balance accuracy and computational efficiency. The simulator time step δt determines both the

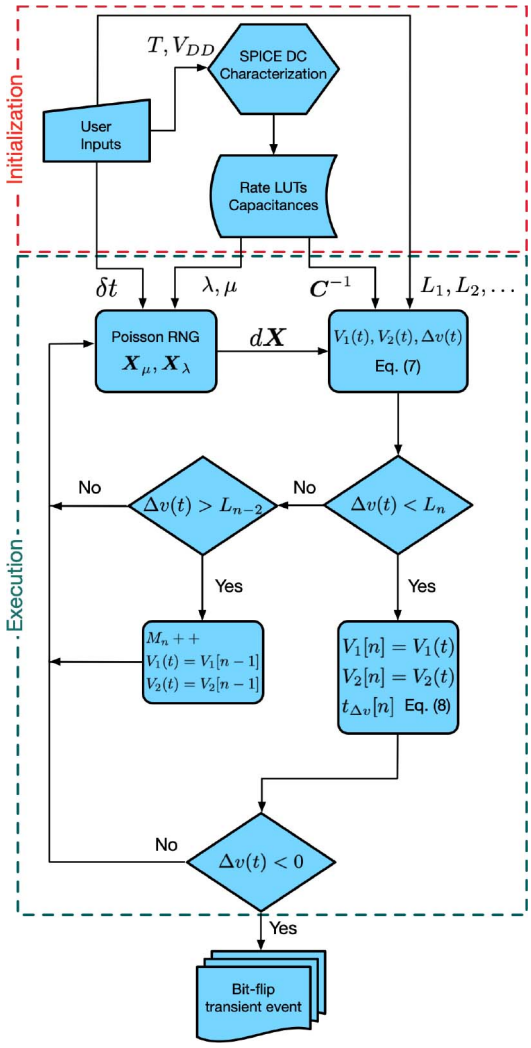

Fig. 4. Flowchart of the incremental phase of TTE simulation. Charge and discharge rates and capacitance values are extracted from DC analysis of the transistors. The circuit topology is derived from a netlist. Execution begins by setting $M_1 = 0$ and $t_1, V_1[1]$, and $V_2[1]$ to values computed by a prior continuous simulation (not shown).

average voltage step sizes and the rate at which the charging/discharging rates are recalculated. Long δt speeds up the calculation but degrades accuracy. The upper bound on δt is set by the requirement that the $\lambda\delta t$ and $\mu\delta t$ products not generate voltage jumps that appreciably alter the rates within a single time step.

The other required runtime attributes determine the threshold levels L_i, $i \in [1, N_{lvls}]$, used as checkpoints as the simulation progresses. A simple choice is to divide the voltage range $V_{DD} \geq L_i \geq -V_{DD}$ into an initial voltage decrement, ρ_{init}, and $(N_{lvls} - 1)$ additional equal decrements, such that

$$L_i = V_{DD} - \rho_{init} - (i-1) \cdot \rho,$$

where ρ determines the level of voltage difference between the stages L_{i-1} and L_i for $i \geq 2$. A lower bound on ρ is set by the requirement that ρ be several times larger than the change in $\Delta v(t)$ due to the addition or removal of a single electron on one node. From the inverse capacitance matrix in (7), changing the node charge by a single electron changes the node voltage by ± 1.3 mV, in turn changing $\Delta v(t)$ by ± 0.86 mV. A suitable minimum ρ for this system is about 10 mV. Increasing the value of ρ would result in greater accuracy at the cost of longer computation time.

We reserve a dedicated parameter for the initial voltage decrement, ρ_{init}. The choice of ρ_{init} determines the initial state of the system for the incremental phase. A critical requirement for such state is to reach the maximum deviation from equilibrium in a reasonable compute time. For our system we determined a sensible value of $\rho_{init} = 30$ mV, which on average can be reached in a simulation time on the order of seconds.

IV. Results and Discussion

A. Simulator Setup

As a test platform for examining thermally induced errors in advanced subthreshold SRAMs, we employ the 7nm predictive technology model from the ASAP7 PDK [18]. The SRAM circuit is designed according to Fig. 1 using minimum-sized low-V_{TH} (LVT) transistors, for which $V_{TH} \simeq 250$ mV. We used single-fin transistors for the NMOS and PMOS, minimizing capacitance and taking advantage of their well-matched currents in subthreshold, as shown in the $I_D(V_G)$ characteristics of Fig. 5(a).

First, we extracted the capacitances and rates using a DC analysis of the transistors (initialization phase in Fig. 4) for $V_{DD} = 180$ mV, $T = 100°C$, which is in line with other research on noise-immune design for subthreshold circuits [22]. The capacitance values extracted from the transistor model in [18] were $C_{GS} = C_{GD} = 30$ aF, and were largely independent of gate bias, as expected in subthreshold operation.

The rates λ and μ for NMOS and PMOS devices as a function of output voltage V_{out} were calculated according to (2)–(5) and are shown in Fig. 5(b)–(c) for $V_{in} = V_{DD}$ and $V_{in} = 0$ respectively, along with the net current I_D flowing into or out of the V_{out} node, $I_D = (\lambda_n + \lambda_p) - (\mu_n + \mu_p)$ (see inset of Fig. 5(b)). As expected, at the equilibrium points, the charging and discharging rates match: $\lambda_n + \lambda_p = \mu_n + \mu_p$ when $V_{in} = 180$ mV or $V_{in} = 0$. The maximum charging and discharging rate λ and μ are less than 3 electrons/ps for $0 \leq V_{in} \leq 180$ mV, so we set $\delta t = 0.2$ ps to ensure that the $(\lambda + \mu)\delta t$ product did not change the nodal voltages by more than 3.2 mV, according to (7). We found that the switching time of an inverter loaded by a single second inverter in this technology operated at $V_{DD} = 180$ mV is ~50 ps. The longer, ~200 ps duration of the spontaneous transition shown in Figure 3(c) is due to the slower evolution of an undriven transition and an extended period (~100 ps) of metastability.

We chose $\rho_{init} = 30$ mV as a first decrement step size easily reachable by continuous simulation and set the second decrement step size $\rho = 10$ mV, a value larger than the expected voltage change within a time step δt to filter out small fluctuations.

B. Evaluating TTE Distribution

To date, published analytical approaches to quantifying thermal noise immunity in flip-flops [12], [15], [16] and

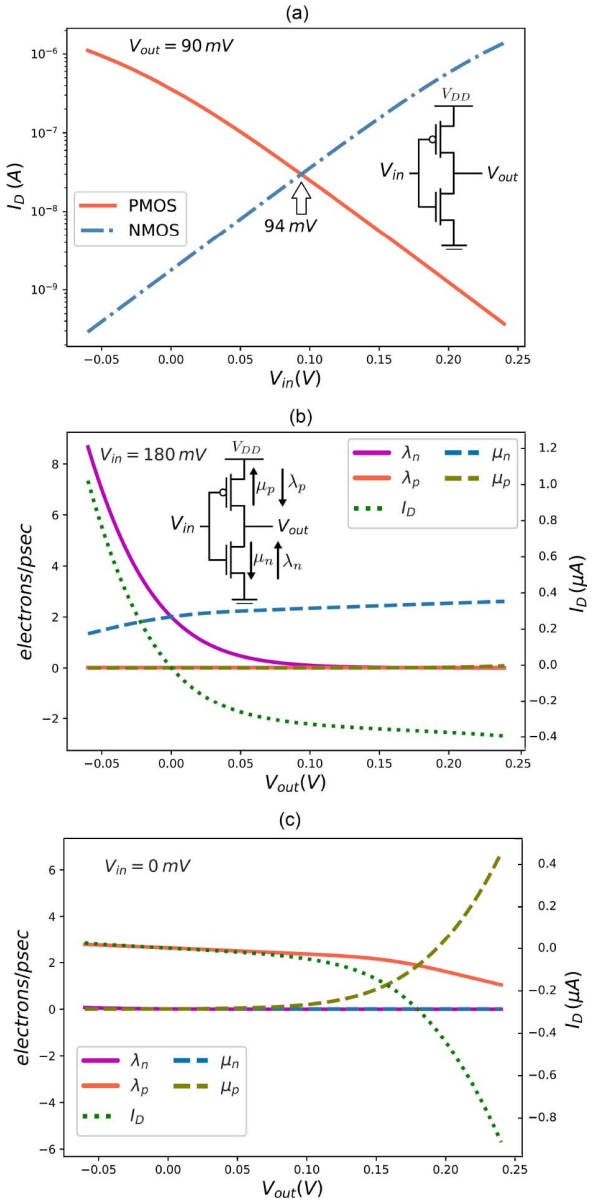

Fig. 5. (a) Subthreshold currents in the NMOS and PMOS transistors for $V_{in} = 0 - 180$ mV, showing close match at $V_{DD}/2$. (b, c) NMOS and PMOS charging and discharging rates as a function of the output voltage V_{out} and the net current $I_D = I_{DN} - I_{DP}$ into or out of the V_{out} node for $V_{in} = 180$ mV (b) and $V_{in} = 0$ (c). Note that the rates λ_p and μ_p in (b) and λ_n and μ_n in (c) are overlapped since I_{off} is over two orders of magnitude smaller than I_{on}, see (a).

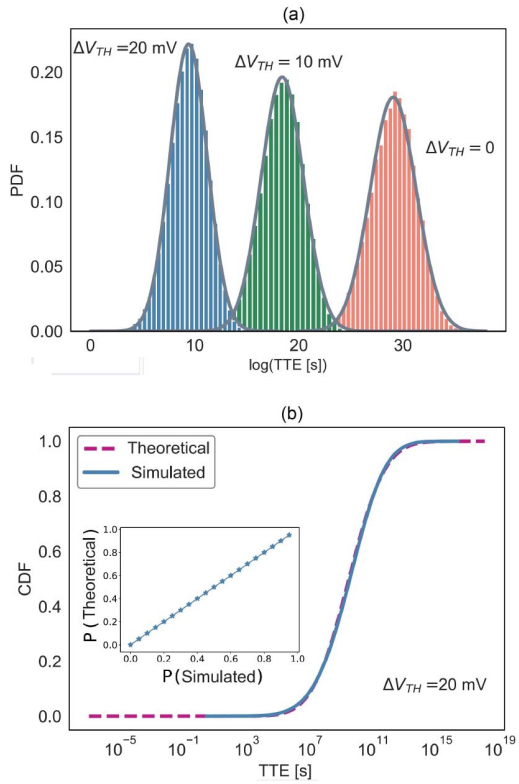

Fig. 6. (a) Statistical TTE histograms of subthreshold CMOS latches extracted by STTACC for $V_{DD} = 180$ mV, $T = 100°C$, and threshold mismatch $\Delta V_{TH} = 0$, 10, and 20 mV, together with the corresponding log-normal fits using the histogram mean and variance. (b) TTE CDF of a subthreshold CMOS latch extracted by STTACC vs. theoretical log-normal distribution from the mean and variance of the TTE histogram ($V_{DD} = 180$ mV, $T = 100°C$, $\Delta V_{TH} = 20$ mV). Inset shows the P-P plot for comparing the simulation results and the theoretical fit.

combinational circuits [17] have focused on mathematically modeling the mean time-to-error (MTTE). Conversely, recent experimental noise measurements in advanced short-channel MOSFETs have focused on other types of noise, such as random telegraph noise (RTN) and flicker ($1/f$) noise [23], [24], since the fundamental thermal noise is not yet the major driver of errors [25]. In our preliminary work, we obtained the simulated cumulative distribution functions (CDFs) of TTE [14] and showed that the CDFs vary over many orders of magnitude in time. This renders MTTE a poor indicator of circuit reliability, since MTTE is dominated by the largest terms in the distribution, whereas it is the shortest TTE values that are of practical interest.

Now, using a large set of accelerated Monte-Carlo simulations facilitated by our runtime-efficient incremental algorithm, we identify the full statistics of the TTE distribution and find that it is well-matched by a log-normal distribution. Fig. 6(a) shows the log-scale histograms of TTE values generated by STTACC for $V_{DD} = 180$ mV and $T = 100°C$ for three different cases of transistor voltage mismatch: nominal V_{TH} values of PMOS and NMOS transistors (i.e., $\Delta V_{TH} = 0$) as in Fig. 5(a), and asymmetric $\Delta V_{TH} = 10$ mV and $\Delta V_{TH} = 20$ mV threshold mismatch that makes the latch far less stable.

These histograms are constructed from at least 60,000 Monte Carlo simulation runs each, using the incremental STTACC simulator. Superimposed on each histogram is a log-normal distribution obtained from the mean and variance of the TTE data with no fitting parameters, demonstrating a very good fit. While not entirely surprising, since the TTE distribution is a multiplicative product of positive independent random variables $M_n + 1$ according to (8), this result is extremely useful. Using the mean and variance of the log-normal distribution from the histogram in Fig. 6(a), we can construct a theoretical CDF curve and compare it to the simulation results to find an excellent match, see Fig. 6(b).

In the inset of Fig. 6(b) we also construct the probability-probability (P-P) plot as a quantitative measure of the agreement between the histogram of simulation results and the corresponding log-normal distribution from the $\Delta V_{TH} = 20$ mV

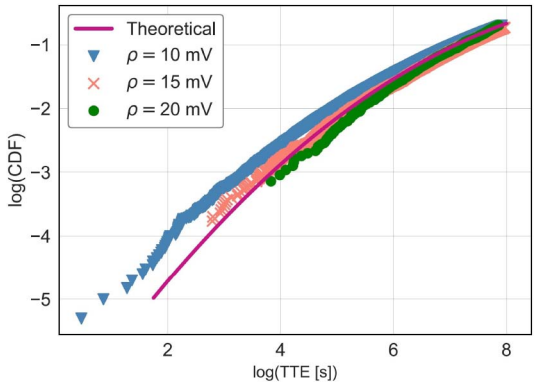

Fig. 7. Comparison of the left tail of CDF curves extracted by STTACC for different values of decrement size ρ with the theoretical log-normal function constructed from the mean and variance of the $V_{DD} = 180$ mV and $T = 100°C$ histogram for an asymmetric worst-case threshold mismatch of $\Delta V_{TH} = 20$ mV.

case. The straight line observed in the P-P plot confirms the log-normal distribution describes the simulated TTE histograms [26]. The log-normal nature of the TTE distribution allows for a quantitative estimation of the frequency of exponentially rare events in the left tail of the TTE distribution, corresponding to short failure times that cannot be addressed by direct simulation.

As mentioned earlier, the decrement step ρ should be maximized for improved accuracy. In fact, continuous simulation with $\rho = V_{DD}$ would lead to an exact numerical solution. However, as shown in [14], extracting a bit-flip by continuous simulation is computationally unfeasible. As a result, the runtime choice of ρ involves a trade-off between robustness and computation time. Fig. 7 plots the left tail of the CDF curves, corresponding to shortest TTEs, for different values of ρ and compares them to the theoretical prediction obtained from the log-normal histogram in Fig. 6(a). We find that the simulation results for $\rho \geq 10$ mV are in acceptable agreement with the theoretical prediction. While $\rho = 20$ mV comes closer to the theoretical curve, in our subsequent results we opt for $\rho = 10$ mV since it provides a good trade-off between accuracy and computation time, saving three orders of magnitude in runtime compared to $\rho = 20$ mV.

C. Evaluating TTE as a Function of Device Variability and Operating Conditions

Having established the simulation framework, we now turn to the dependence of TTE distributions and error rates on device variability and operating conditions. We will investigate three key parameters: subthreshold supply voltage V_{DD}, temperature, and device variability represented by threshold voltage mismatch. We first focus on ΔV_{TH}, while keeping $V_{DD} = 180$ mV and $T = 100°C$. Random process variations can cause a significant mismatch in neighboring devices particularly for scaled SRAMs [7]. Moreover, optimized SRAM design for stable read/write operation requires precise NMOS and PMOS sizing and threshold control. Even for cells designed with nominally fully-matched devices, short channel effects can impact the effective transistor V_{TH}, particularly in the subthreshold regime [27].

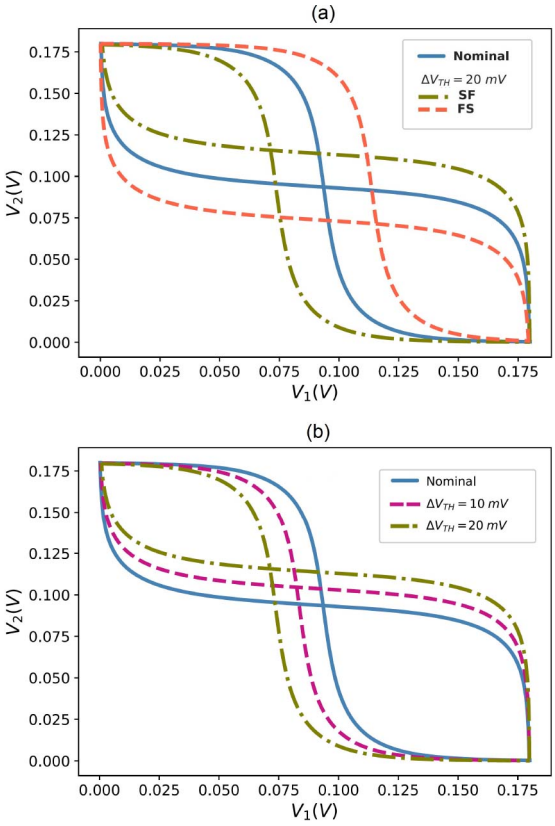

Fig. 8. The effect of threshold voltage variation on the noise margin: (a) when the asymmetric shift $\Delta V_{TH} = 20$ mV is applied on NMOS and PMOS transistors in the latch, i.e., either FS or SF; (b) when the asymmetric shift $\Delta V_{TH} = 10$ and 20 mV is applied on each inverter.

In Fig. 8, we plot the voltage transfer butterfly curves of the two inverters in the latch to describe the effect of threshold voltage variations on the noise margin. We consider three cases for PMOS and NMOS devices: nominal, fast (***F***), and slow (***S***). Nominal refers to optimum-sized NMOS and PMOS devices with symmetric current drive, as in Fig. 5(a); ***F*** represents $-\Delta V_{TH}$ and $+\Delta V_{TH}$ shifts in threshold voltage for NMOS and PMOS devices respectively, and ***S*** represents the opposite shift for each transistor.

For perfectly matched devices, the two curves would intersect at $V_1 = V_2 = V_{DD}/2$ and the two lobes would have the same size, maximizing the static noise margin (SNM) for both logic states [19], [28]. However, when a threshold voltage shift is applied, the voltage transfer curves move away from this symmetric configuration, reducing the SNM for one of the logic states. Fig. 8(a) demonstrates that the worst case in terms of thermal noise immunity is asymmetric mismatch ΔV_{TH} between NMOS and PMOS transistors in the latch (i.e., either ***SF*** or ***FS***), whereas symmetric mismatch ΔV_{TH} between NMOS and PMOS (***SS*** and ***FF***) have similar SNM to the nominal case. In Fig. 8(b), the effect of threshold voltage variations on the noise margin is compared for nominal transistors and those with asymmetric $\Delta V_{TH} = 10$ and 20 mV applied to each inverter.

Figure 9 shows the left tail of the CDF for the theoretical curves extracted from the simulation histograms (as in Fig. 6),

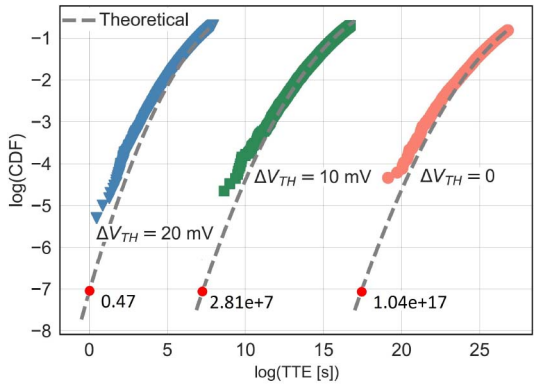

Fig. 9. The effect of threshold voltage variation on the left tail of the TTE distribution at $V_{DD} = 180$ mV and $T = 100°C$. The scatter plots represent simulation results, while dashed lines are the theoretical curves obtained from the mean and variance of the full simulation histograms. The red circular dots indicate the $t_{50\%}$ points representing the time (in seconds) needed to have a single bit-flip in a 1 MB SRAM with 50% probability.

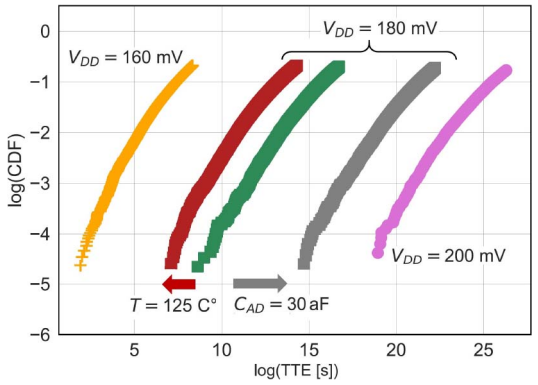

Fig. 10. The effect of varying the operating voltage V_{DD}, temperature, and access transistor capacitance C_{AD} on the left tail of the TTE distribution for $\Delta V_{TH} = 10$ mV, obtained by STTACC simulation.

and for those simulated using STTACC. For all curves we apply worst-case asymmetric mismatch ΔV_{TH} between NMOS and PMOS devices. Threshold mismatch massively degrades reliability, even for modest ΔV_{TH} values of 10 or 20 mV. For example, the average time to have a bit-flip in an SRAM bank with 10k memory cells changes from $\simeq 10^{20}$ s for matched transistors (unconditionally stable) to $\simeq 10$ minutes for asymmetric $\Delta V_{TH} = 20$ mV (which is still less than 10% of the nominal $V_{TH} \simeq 250$ mV [18]).

Figure 10 shows the left tail of the CDF simulated using STTACC for $\Delta V_{TH} = 10$ mV as a function of supply voltage V_{DD} and temperature T. We find that lowering V_{DD} from 180 to 160 mV severely degrades the noise immunity, while raising V_{DD} to 200 mV stabilizes the SRAM. Temperature has less of an impact: raising the temperature to $T = 125°C$ from $100°C$ shifts the left CDF tail by 2 orders of magnitude along the logarithmic time axis.

Our approach also makes it possible to incorporate additional capacitances, either due to other connected devices or due to parasitics, such as wiring. For example, as shown in Fig. 1 for a conventional 6T-SRAM bitcell, there are two access transistors M_5 and M_6 in addition to the cross-coupled

TABLE I
TIME TO A 50% PROBABILITY OF A BIT-FLIP ERROR IN A 1 MB MEMORY UNDER VARIOUS OPERATING CONDITIONS

Conditions		$t_{50\%}$
$T = 100°C$, $V_{DD} = 180$ mV		
$\Delta V_{TH} =$	0 mV	indefinite
	10 mV	7 months
	20 mV	1 second
$T = 100°C$, $\Delta V_{TH} = 10$ mV		
$V_{DD} =$	200 mV	1.5 billion years
	180 mV	7 months
	160 mV	2 seconds
$\Delta V_{TH} = 10$ mV, $V_{DD} = 180$ mV		
$T =$	$100°C$	7 months
	$125°C$	2.6 days
$T = 100°C$, $\Delta V_{TH} = 10$ mV, $V_{DD} = 180$ mV		
$C_{AD} =$	0 aF	7 months
	15 aF	580 years
	30 aF	585,000 years

inverters. Their main effect is to add their gate-drain capacitances (calculated as $C_{GD} = 30$ aF) as capacitance from the output nodes to ground. Those appear in our model as part of the C_{AD} terms of the diagonal elements of the capacitance matrix. Our simulator predicts that the latch will be partially stabilized by the extra capacitance, see Fig. 10. Access transistors might also contribute current fluctuations that could be added as additional λ_n and μ_n terms to those in the inset of Fig. 5(b). However, since these access transistors are always fully off, these terms would be small, like those arising from the corresponding terms in Fig. 5(c) that are indistinguishable from zero.

Finally, by comparing Figs. 9 and 10 it is clear that in order to compensate for the impact of a threshold voltage mismatch of $\Delta V_{TH} = 10$ mV, we need to increase V_{DD} by at least 20 mV (compare $V_{DD} = 180$ mV, $\Delta V_{TH} = 0$ with $V_{DD} = 200$ mV, $\Delta V_{TH} = 10$ mV).

A particularly useful quantitative measure of the failure statistics for an SRAM array is the time for the probability of a failure of at least one stored logic value to reach 50 percent, $t_{50\%}$. The cumulative distribution function describes the probability of at least one bit-flip by time t. For an array of N independent latches, setting the probability of at least one failure to 0.5 gives an equation for $t_{50\%}$ as:

$$0.5 = 1 - \left(1 - CDF\left(t_{50\%}, \overline{m}_{log}, \sigma\right)\right)^N \tag{9}$$

where $\overline{m}_{log}$ and σ are the mean and standard deviation of $log(t)$ as derived by fitting the simulations to a log-normal distribution. Equation (9) is a nonlinear equation for $t_{50\%}$, but is trivial to solve numerically. Fig. 9 shows the $t_{50\%}$ points for a 1 MB SRAM as dots with corresponding time values, providing a quantitative sense of the noise immunity as a function of worst-case threshold voltage mismatch between transistors. Table I summarizes the calculated $t_{50\%}$ values for 1 MB of memory under various operating conditions.

V. CONCLUSION

In this article, we have introduced STTACC, a simulation framework for analyzing thermal noise-driven transients in subthreshold circuits. We employ a parallelizable incremental

algorithm to detect exponentially rare bit-flip errors that would be impossible to estimate with conventional SPICE-based transient simulation methods. We show that the time-to-error (TTE) statisics follow a log-normal distribution and demonstrate that the commonly used mean-time to error (MTTE) metric is not an informative measure of SRAM reliability.

Using parameter values from a 7nm predictive technology model, we ran several experiments with STTACC to evaluate how SRAM reliability is affected by lowered supply voltage, increased process variability (manifested as a shift in threshold voltage mismatch), and temperature increase. Our results offer a way to assess when a circuit with aggressive device scaling coupled with lowered supply voltages will have a given maximum probability of failure. Finally, our simulation framework can be adapted to other technology nodes or other sources of noise, and provides a means of evaluating trade-offs in the robust design of low-power SRAM for ULP memory. Future plans include public release of the tool to the circuit and memory design community.

REFERENCES

[1] H. N. Patel, F. B. Yahya, and B. H. Calhoun, "Subthreshold SRAM: Challenges, design decisions, and solutions," in *Proc. IEEE 60th Int. Midwest Symp. Circuits Syst. (MWSCAS)*, Boston, MA, USA, Aug. 2017, pp. 321–324.

[2] T. Haine, D. Flandre, and D. Bol, "8-T ULV SRAM macro in 28nm FDSOI with 7.4 pW/bit retention power and back-biased-scalable speed/energy trade-off," in *Proc. IEEE SOI 3D Subthreshold Microelectron. Technol. Unified Conf. (S3S)*, Burlingame, CA, USA, 2018, pp. 1–3.

[3] K. Agarwal and S. Nassif, "Statistical analysis of SRAM cell stability," in *Proc. 43rd ACM/IEEE Annu. Design Autom. Conf.*, San Francisco, CA, USA, 2006, pp. 57–62.

[4] A. J. Bhavnagarwala, X. Tang, and J. D. Meindl, "The impact of intrinsic device fluctuations on CMOS SRAM cell stability," *IEEE J. Solid-State Circuits*, vol. 36, no. 4, pp. 658–665, Apr. 2001.

[5] Z. Guo, A. Carlson, L.-T. Pang, K. T. Duong, T.-J. K. Liu, and B. Nikolic, "Large-scale SRAM variability characterization in 45 nm CMOS," *IEEE J. Solid-State Circuits*, vol. 44, no. 11, pp. 3174–3192, Nov. 2009.

[6] J. Wang, S. Yaldiz, X. Li, and L. T. Pileggi, "SRAM parametric failure analysis," in *Proc. 46th ACM/IEEE Annu. Design Autom. Conf.*, San Francisco, CA, USA, 2009, pp. 496–501.

[7] Q. Chen, H. Mahmoodi, S. Bhunia, and K. Roy, "Modeling and testing of SRAM for new failure mechanisms due to process variations in nanoscale CMOS," in *Proc. 23rd IEEE VLSI Test Symp. (VTS'05)*, Palm Springs, CA, USA, 2005, pp. 292–297.

[8] R. W. Keyes, "Physical problems and limits in computer logic," *IEEE Spectr.*, vol. 6, no. 5, pp. 36–45, May 1969.

[9] P. R. Gray, R. G. Meyer, P. J. Hurst, and S. H. Lewis, *Analysis and Design of Analog Integrated Circuits*, 4th ed. New York, NY, USA: Wiley, 2001.

[10] K. Nepal, R. I. Bahar, J. Mundy, W. R. Patterson, and A. Zaslavsky, "Designing nanoscale logic circuits based on principles of Markov random fields," in *Emerging Nanotechnologies* (Frontiers in Electronic Testing), vol. 37, M. Tehranipoor, Eds. Boston, MA, USA: Springer, 2008, pp. 315–338.

[11] R. Sarpeshkar, T. Delbruck, and C. A. Mead, "White noise in MOS transistors and resistors," *IEEE Circuits Devices Mag.*, vol. 9, no. 6, pp. 23–29, Nov. 1993.

[12] H. Li, J. Mundy, W. Patterson, D. Kazazis, A. Zaslavsky, and R. I. Bahar, "A model for soft errors in the subthreshold CMOS inverter," in *Proc. Workshop Syst. Effects Logic Soft Errors*, 2006, pp. 1–4.

[13] M. Donato, R. I. Bahar, W. R. Patterson, and A. Zaslavsky, "A subthreshold noise transient simulator based on integrated random telegraph and thermal noise modeling," *IEEE Trans. Comput.-Aided Design Integr. Circuits Syst.*, vol. 37, no. 3, pp. 643–656, Mar. 2018.

[14] E. Rezaei, M. Donato, W. R. Patterson, A. Zaslavsky, and R. I. Bahar, "Thermal noise-induced error simulation framework for subthreshold CMOS SRAM," in *Proc. SOI 3D Subthreshold Microelectron. Technol. Unified Conf. (S3S)*, 2019, pp. 1–3.

[15] H. Li, J. Mundy, W. Patterson, D. Kazazis, A. Zaslavsky, and R. I. Bahar, "Thermally-induced soft errors in nanoscale cmos circuits," in *Proc. IEEE Int. Symp. Nanoscale Architect.*, San Jose, CA, USA, 2007, pp. 62–69.

[16] P. Jannaty *et al.*, "Shot-noise-induced failure in nanoscale flip-flops part II: Failure rates in 10-nm ultimate CMOS," *IEEE Trans. Electron Devices*, vol. 59, no. 3, pp. 807–812, Mar. 2012.

[17] N. Miskov-Zivanov and D. Marculescu, "MARS-C: Modeling and reduction of soft errors in combinational circuits," in *Proc. 43rd ACM/IEEE Annu. Design Autom. Conf.*, San Francisco, CA, USA, 2006, pp. 767–772.

[18] L. T. Clark *et al.*, "ASAP7: A 7-nm finFET predictive process design kit," *Microelectron. J.*, vol. 53, pp. 105–115, Jul. 2016.

[19] Y. Taur and T. H. Ning, *Fundamentals of Modern VLSI Devices*. Cambridge, UK: Cambridge Univ. Press, 2013.

[20] E. A. Gutiérrez-D, *Nano-Scaled Semiconductor Devices: Physics, Modelling, Characterisation, and Societal Impact*. Stevenage, U.K.: Inst. Eng. Technol., 2016.

[21] N. Paydavosi *et al.*, "BSIM—SPICE models enable FinFET and UTB IC designs," *IEEE Access*, vol. 1, pp. 201–215, 2013.

[22] J. P. Kulkarni and K. Roy, "Ultralow-voltage process-variation-tolerant schmitt-trigger-based SRAM design," *IEEE Trans. Very LArge Scale Integr. (VLSI) Syst.*, vol. 20, no. 2, pp. 319–332, Feb. 2012.

[23] L. Van Brandt, B. K. Esfeh, V. Kilchytska, and D. Flandre, "Robust methodology for low-frequency noise power analyses in advanced MOS transistors," in *Proc. 5th Joint Int. EUROSOI Workshop Int. Conf. Ultimate Integr. Silicon (EUROSOI-ULIS)*, 2019, pp. 1–4.

[24] O. Huerta, C. Marquez, A. I. Tec-Chim, F. Guarin, E. A. Gutierrez-D, and F. Gamiz, "Experimental characterization of the random telegraph noise signature in MOSFETs under the influence of magnetic fields," *IEEE Electron Device Lett.*, vol. 39, no. 7, pp. 1054–1057, Jul. 2018.

[25] A. G. Mahmutoglu and A. Demir, "Modeling and analysis of nonstationary low-frequency noise in circuit simulators: Enabling non Monte Carlo techniques," in *Proc. IEEE/ACM Int. Conf. Comput.-Aided Design (ICCAD)*, San Jose, CA, USA, 2014, pp. 309–315.

[26] R. M. Vogel, "The probability plot correlation coefficient test for the normal, lognormal, and Gumbel distributional hypotheses," *Water Resour. Res.*, vol. 22, no. 4, pp. 587–590, 1986.

[27] R. Varma, V. Dokania, A. Sarkar, and A. Islam, "MOSFET aspect ratio optimization for minimized transistor mismatch at UDSM technology nodes," in *Proc. Int. Conf. Comput. Commun. Informat. (ICCCI)*, Coimbatore, India, 2015, pp. 1–4.

[28] M. U. Mohammed, A. Nizam, and M. H. Chowdhury, "Performance stability analysis of SRAM cells based on different finFET devices in 7nm technology," in *Proc. IEEE SOI 3D Subthreshold Microelectron. Technol. Unified Conf. (S3S)*, Burlingame, CA, USA, 2018, pp. 1–3.

Elahe Rezaei (Member, IEEE) received the B.S. and M.S. degrees in electrical engineering from the Sharif University of Technology, Tehran, Iran, in 2012 and 2014, respectively, and the Ph.D. degree in electrical engineering and computer science from the University of Texas at Dallas, Richardson, TX, USA, in 2018.

She joined the School of Engineering, Brown University, Providence, RI, USA, as a Research Associate in 2018. She is currently a Research Scientist with the Department of Machine Learning-HW, Qualcomm Technologies, Inc., San Diego, CA, USA. Her research interests include data-driven approaches for hardware design and computer systems.

Marco Donato (Member, IEEE) received the B.S. and M.S. degrees in electrical engineering from the Sapienza University of Rome, Rome, Italy, in 2008 and 2010, respectively, and the Ph.D. degree in electrical sciences and computer engineering from Brown University, Providence, RI, USA, in 2016.

He is currently a Research Associate with the John A. Paulson School of Engineering and Applied Sciences, Harvard University, Cambridge, MA, USA. His research interests include novel design methodologies targeting energy-efficient and reliable circuits and architectures for emerging computing paradigms.

William R. Patterson (Life Member, IEEE) received the B.Sc. degree in physics and the M.Sc degree in electrical engineering from Brown University, Providence, RI, USA, in 1963 and 1966, respectively.

Since 1977, he has been with the Electrical and Computer Engineering Group, School of Engineering, Brown University, where he is currently a Distinguished Senior Lecturer and a Senior Research Engineer. His current research interests include low-power analog circuit design for biomedical and cryogenic applications, circuits and architectures for probabilistic logic, and instrumentation for geological spectroscopy.

Mr. Patterson was a recipient of the NASA Public Service Medal for his contributions to the Viking Program in 1977.

Alexander Zaslavsky received the B.A. degree in physics from Harvard University, Cambridge, MA, USA, in 1986, and the M.Sc. and Ph.D. degrees in electrical engineering from Princeton University, Princeton, NJ, USA, in 1988 and 1991, respectively.

From 1991 to 1993, he was a Postdoctoral Scientist with the T. J. Watson Research Center, IBM Research, Yorktown Heights, NY, USA. Since 1994, he has been with Brown University, Providence, RI, USA, where he is currently a Professor of engineering and physics. His research interests are in the areas of semiconductor device and nanostructure physics, particularly tunneling and hot-electron devices, as well as reliable computing with nanotransistors.

Dr. Zaslavsky has been an Editor of the *Solid State Electronics* journal since 2003.

R. Iris Bahar (Senior Member, IEEE) received the B.S. and M.S. degrees in computer engineering from the University of Illinois at Urbana–Champaign, Urbana, IL, USA, and the Ph.D. degree in electrical and computer engineering from the University of Colorado Boulder, Boulder, CO, USA, in 1986, 1987, and 1995, respectively.

From 1987 to 1992, she was with Digital Equipment Corporation, Hudson, MA, USA, working on microprocessor hardware design. She has been on the faculty with Brown University, Providence, RI, USA, since 1996, and currently holds a dual appointment as a Professor of engineering and a Professor of computer science. Her research interests have centered on energy-efficient and reliable computer systems, from the system level to device level.

Dr. Bahar is the 2019 recipient of the Marie R. Pistilli Women in Engineering Achievement Award and the Brown University School of Engineering Award for Excellence in Teaching in Engineering. She served as the Program Chair and the General Chair of the International Conference on Computer-Aided Design in 2017 and 2018, and the General Chair of the International Conference on Architectural Support for Programming Languages and Operating Systems in 2019. She is a Distinguished Scientist of ACM.