IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 1 June 2026, accepted 13 July 2026, date of publication 17 July 2026, date of current version 23 July 2026.

Digital Object Identifier 10.1109/ACCESS.2026.3714030

RESEARCH ARTICLE

An Efficient SPICE Framework for Simulation of BTI Effects at the Circuit Level

JOSEF SALZMANN[1,2], ALEXANDER KARL[2], DOMINIC WALDHÖR[2], TIBOR GRASSER[2], (Fellow, IEEE), AND MICHAEL WALTL[1,2], (Senior Member, IEEE)

[1]Christian Doppler Laboratory for Single Defect Spectroscopy, Institute for Microelectronics, TU Wien, 1040 Vienna, Austria
[2]Institute for Microelectronics, 1040 TU Wien, Vienna, Austria

Corresponding authors: Josef Salzmann (salzmann@iue.tuwien.ac.at) and Michael Waltl (waltl@iue.tuwien.ac.at)

This work was supported in part by Austrian Federal Ministry for Digital and Economic Affairs; in part by the National Foundation for Research, Technology and Development; in part by the Christian Doppler Research Association; in part by Austrian Research Promotion Agency Österreichische Forschungsförderungsgesellschaft (FFG) and their Competence Centers for Excellent Technologies (COMET) Program through the Project Reliability Evaluation and Analysis of Circuits and Emerging Transistor Technologies (REACT) under Grant FO999923935; and in part by Technische Universität Wien (TU Wien) Bibliothek through the Open Access Funding Program.

ABSTRACT Bias temperature instability (BTI) due to charge trapping at defect sites and defect generation is a major reliability concern in robust analog and digital circuits. Incorporating accurate BTI degradation into circuit-level simulations that captures recovery history and workload dependence remains challenging, because many existing modeling approaches either lack sufficient physical fidelity or suffer from poor computational efficiency. Extensive transistor reliability studies have established that BTI is due to a combination of charge trapping/detrapping and the creation of quasi-permanent interface defects. Regarding charge trapping, it has been shown that the two-state defect model based on the non-radiative multiphonon (NMP) model provides one of the most accurate descriptions under arbitrary mission profiles. However, the comparatively high computational cost of recent NMP implementations limits their applicability in circuit simulations. In this work, we employ the open-source SPICE simulator ngspice and implement an NMP-based defect model that enables computationally efficient emulation of BTI effects at circuit level. To demonstrate its efficiency, we simulate a 501-stage ring oscillator and an operational amplifier with devices containing up to several thousand individual traps. While recent implementations have resulted in an increase of two orders of magnitude in computational cost for only a few hundred traps, our implementation introduces an overhead of less than one order of magnitude even when several thousand defects are included. This modest computational cost, combined with the high physical accuracy of our approach, paves the way to verification across diverse operating conditions as part of the design process, supporting high yield.

INDEX TERMS Bias temperature instability (BTI), electronic design and automation (EDA), SPICE, circuit-aging simulation, defect-aware circuit simulation, charge trapping.

I. INTRODUCTION

For many scaled CMOS technologies, bias temperature instability (BTI) is a key reliability concern that is typically expressed as a drift in the threshold voltage of a device, i.e., ΔV_{th} [1]. In CMOS technologies the drift of V_{th} is typically caused by the capture and release of charge carriers at defects located inside the insulator (slow oxide defects) or at the insulator/channel interface (fast interface defects), as well as defect generation from precursor sites. Although power-law expressions have been successfully used to fit experimentally observed ΔV_{th} shifts, such approaches generally exhibit limited predictive capabilities, particularly when relaxation effects become important, as typically observed in dynamic workloads, owing to their empirical nature. To overcome this limitation and provide a physics-based description of defect dynamics and their impact on device characteristics, a defect-centric model based on the non-radiative multiphonon (NMP) model has been proposed [2]. This framework has been successfully applied to explain various aspects of BTI phenomena across a wide range of technologies, such as Si/SiO_2 systems [3], high-k technologies [4], power devices

The associate editor coordinating the review of this manuscript and approving it for publication was Ludovico Minati.

using SiC [5] and the most recent novel transistors based on 2D materials [6]. In these studies, the observed BTI-induced shift was successfully explained within the NMP model by charge trapping at interface defects and oxide defects, as illustrated in Fig. 1. The NMP model describes charge capture and emission at defects as transitions between two charge states mediated by lattice vibrations. The defect is represented by two harmonic potential energy surfaces along a configuration coordinate Fig. 2, and transitions between them occur via thermally activated multiphonon processes, which determine the capture and emission rates. In extensive experiments, it was observed that the capture and emission times of defects can span several orders of magnitude, ranging from picoseconds to years [7], which can be explained well by the NMP model. Note that the model parameters for a given defect type can also be extracted from ab initio calculations, ensuring a physically sound parameter set [8]. In addition to charge trapping into preexisting defects, it has been shown that additional defects can be created via the activation of precursor sites. Although not strictly rigorous, this activation can also be described using the equations of the NMP model when empirical parameters are used [9].

An implementation of the NMP model is available in the open-source compact physics framework Comphy [10] and to efficiently calibrate the defect model on an extensive experimental data set, the effective single defect decomposition algorithm has been proposed [9]. Incorporating the NMP model into a circuit simulation environment provides handling of arbitrary workloads, extrapolation beyond calibration and a physics based unified description of BTI, RTN and $1/f$ noise. This enables detailed simulation of how BTI affects the characteristics of non-trivial circuits. However, due to the computational complexity of the NMP model, this integration poses a significant challenge. The only implementations of the NMP-based defect model for analog simulation we are aware of are [11], [12]. Despite the efforts to speed-up the Verilog-A implementation in [12] this approach still results in a tremendous computational overhead and slows down circuit simulations by several orders of magnitude, as the NMP model is evaluated at each simulation time step for every single defect. For example, the simulation of devices with 500 defects can increase the simulation time by a factor of more than 100 [12].

A variety of frameworks to assess BTI aging in both digital and analog circuits based on different approaches that do not suffer from such computational overheads exist. Commercially available tools such as RelXpert [13] and MOSRA [14], [15] offer foundry PDK support, are computationally efficient and used in practice, however, these models have various drawbacks, as they are empirical. For example, the poor prediction performance of recovery under arbitrary workloads [16] or the inability of capturing the stochastic nature of BTI degradation in scaled devices [17]. Reaction-diffusion (RD) based frameworks have been presented [18], [19], however it has been questioned whether they are based on physically meaningful assumptions [1].

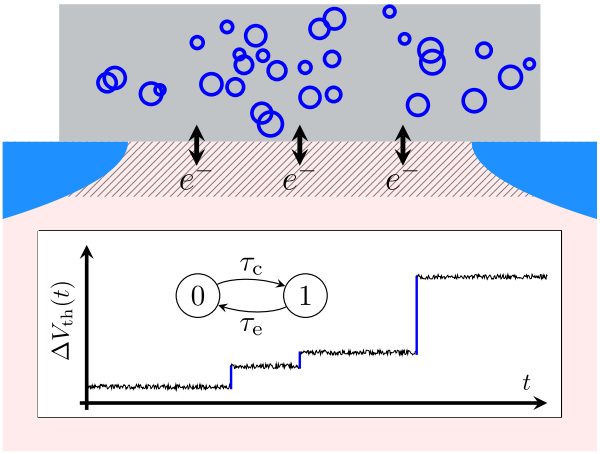

FIGURE 1. Defects in the oxide, near-interface and interface defects can exchange charge carriers with the channel. These charge transitions change the electrostatic potential along the conducting channel and can emerge as discrete changes in the threshold voltage of a device, i.e. ΔV_{th}. This can cause a severe degradation of the signal-to-noise ratio in analog amplifier circuits, but can also lead to jitter in digital circuits.

For example, they require parameters inconsistent with the known behavior of hydrogen, also it has been shown that RD models do not capture the observed stochastic behavior in scaled devices [20]. Although computationally efficient, probabilistic defect occupation frameworks [21], [22] cannot handle arbitrary workloads in a physically sound manner. Machine learning approaches that enable fast simulation have been presented [16], however, any machine learning model inherits the drawbacks of the underlying model used for training. Overall, the drawbacks of any model can be compensated by excessive guard bands and conservative sign-off criteria, at the expense of area and performance.

Additionally, as device dimensions continue to shrink, device-to-device variability increases since the impact of individual defects becomes more pronounced [23]. In particular, a single defect can have a severe effect when it is located in the dominant current path of the device, where it strongly perturbs the local conduction. This can result in discrete steps in the transient behavior of ΔV_{th} in scaled devices, as illustrated in Fig. 1. Consequently, assessing the impact of device variability at the circuit level is becoming increasingly important for system design. To consider the impact of charge trapping effects on circuits during the design process, we developed a computationally efficient implementation of the NMP model, augmenting the open source SPICE simulator ngspice via the XSPICE interface, as this offers higher flexibility than directly implementing it in a device model. Using XSPICE the model is independent of any device model, at the cost of not being able to directly modify the internal parameters of the device model. Adding this model to the gate of a device, see Fig. 3, allows defect-aware SPICE simulations. By circumventing the costly Comphy invocations in each iteration through the use of lookup tables, harnessing single instruction, multiple

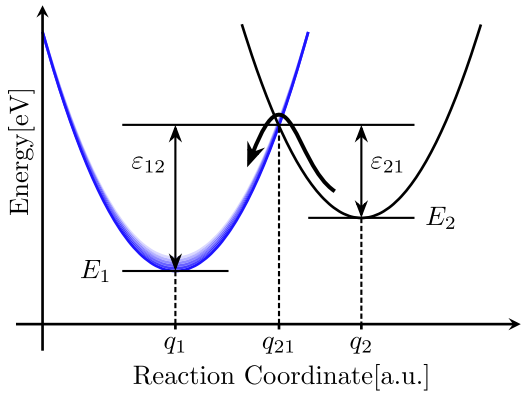

FIGURE 2. Defect states are modeled as harmonic oscillators within the NMP model. In the high temperature limit [25] the reaction barrier for a charge transfer is determined by the bias-dependent intersection point of the two parabolas.

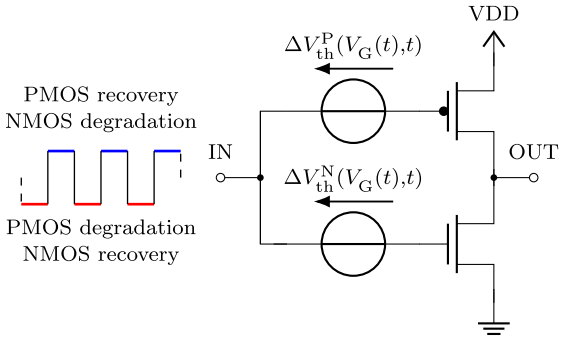

FIGURE 3. Embedding of the XSPICE model into the SPICE simulation. The contributions of defects to $\Delta V_{\mathrm{th}}(V_{\mathrm{G}}(t), t)$ for every device are calculated at every time step and are considered as voltage sources which are connected in series to the gate terminal of each transistor.

data (SIMD) extensions and employing multiprocessing, the computational performance is considerably increased, making dynamic BTI simulation of medium sized circuits feasible. As the model is implemented in C, migration to other SPICE based tools is easily possible, e.g., by directly augmenting existing device models. In addition, the proposed framework is compatible with higher level techniques like workload compression [24], which can be built on top of the proposed framework.

II. PROPOSED AGING SIMULATION WORKFLOW

The aging simulation workflow in state-of-the-art EDA tools commonly follows the Berkeley Reliability Tool (BERT) methodology [26], as illustrated in Fig. 4. Within this framework, accumulated stress is first evaluated through a nominal simulation based on a representative workload profile. Subsequently, an aging model is applied to determine the resulting modification of device parameters, after which the aged circuit behavior is assessed using the updated parameter set. A key advantage of this methodology lies in its straightforward implementation due to its reliance on empirical degradation formulas, which ensures high computational efficiency and practical usability in industrial design environments. Based on these simulation outcomes, design guard bands are extracted in a subsequent step. However, this simplified treatment typically provides only coarse estimates of circuit degradation and often neglects dynamic interactions between device degradation and recovery and circuit operation. Owing to these modeling inaccuracies and simplifications, the extracted guard bands may be overly conservative. While such conservatism can improve technology yield and reliability margins, excessively large guard bands may unnecessarily constrain performance and hinder technology optimization. For instance, supply-voltage scaling can be limited because any overestimation of ΔV_{th} degradation must be compensated by increasing V_{DD} in digital circuits, which in turn increases power consumption and reduces overall efficiency.

In contrast, we propose a three-step workflow that eliminates the conventional separation between nominal and aged simulations, as illustrated in Fig. 5. In the first step, the electrostatics of the BSIM model are calibrated to the target technology using measured C–V and I_{D}–V_{G} characteristics. Because BTI-induced charge trapping is represented as a series voltage source at the gate terminal, as depicted in Fig. 3, this method can also be applied to any foundry-calibrated process design kit (PDK). Although similar concepts have been discussed in earlier literature, existing implementations generally lack a computationally efficient realization of a physics-based charge-trapping model.

In the second step, the NMP defect parameters are extracted from MSM sequences (see Fig. 6) using the effective single-defect decomposition (ESiD) method [9]. Since defect transition rates are highly dependent on temperature, MSM sequences at high temperatures can be used to capture nominally slow defects, while lower temperatures are needed to capture nominally fast defects [27]. For ESiD to be able to extract representative defect ensembles MSM sequences at various temperatures are generally necessary. Defects in a device are distributed both in space and in terms of their defect parameters such as the trap level, as shown by a band diagram Fig. 7. Here, an example of a trap band extracted with ESiD is shown. The occupation of a defect, symbolized by its color in Fig. 7, depends on stress history, bias and temperature. From a specified parameter search region, ESiD extracts the defect distribution that reproduces the measured MSM sequences. The ESiD methodology has been integrated into SPICE simulations as part of this work and enables highly faithful reproduction of device behavior under varying operating conditions. Consequently, the proposed framework captures BTI defect dynamics directly within the SPICE simulation, eliminating the need for artificial separation between stress and aging phases and thereby enabling a streamlined, physically consistent workflow for circuit-level BTI analysis.

III. SPICE-LEVEL DEFECT SIMULATION

With the XSPICE extension, ngspice [28] provides an interface that allows the addition of custom components to the

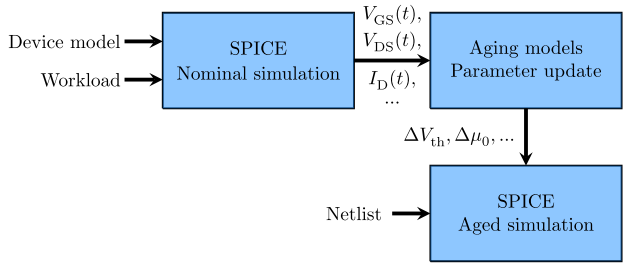

FIGURE 4. Workflow of reliability simulations in state-of-the art EDA tools following the conventional BERT approach. Aging is captured by an aging model outside SPICE that supplies changes in device parameters.

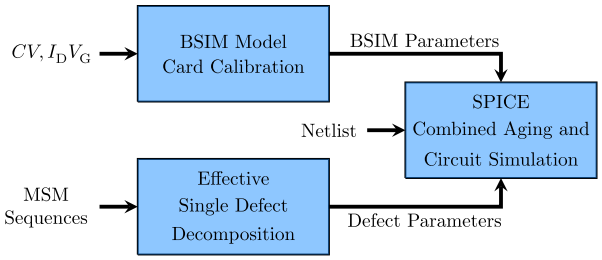

FIGURE 5. In our implementation, stress and aging are treated simultaneously rather than in separate simulation phases. This unified formulation enables fully self-consistent simulations in which the instantaneously aged state of the circuit, arising from BTI, is inherently propagated to subsequent time steps.

simulation environment. Via this interface we added a voltage source that captures the defect behavior in dependence of bias and temperature, according to the NMP model. The defect parameters extracted by ESiD can be applied directly to this model, enabling a seamless workflow. This model allows us to capture the defect behavior of each transistor in the circuit.

Regarding simulation run time, computing the capture and emission times, denoted as $\tau_c(V_G)$ and $\tau_e(V_G)$ respectively, for each defect at every time step represents a major computational bottleneck. To accelerate the simulation, these transition times are precomputed and stored in lookup tables at the beginning of the simulation. Although this introduces a non-negligible initialization overhead, the overall run time is significantly reduced – because only simple interpolations need to be performed during the subsequent simulation steps. Log-space interpolation is used, as capture and emission times are distributed over several orders of magnitude. Our model offers two different simulation modes: *i)* a probabilistic mode in which each defect contributes to the total threshold voltage shift based on its occupational probability, and *ii)* a discrete mode where the occupancy of each defect is determined using a kinetic Monte Carlo algorithm. An example use case for the probabilistic mode is the simulation of large devices in which the mean degradation is of interest, rather than the individual defect contributions. Example usages of the discrete mode are the simulation of random telegraph noise (RTN) and the assessment of device-to-device variability by sampling individual defects from a given defect distribution.

A. PROBABILISTIC SIMULATION MODE

The objective of the probabilistic simulation mode is to simulate the average impact of a defect on ΔV_{th}. Therefore, each defect state is represented by an occupational probability $p(t)$, which corresponds to the charge state of a defect. While $p(t) = 0$ represents a neutral trap (in the electron picture) the defect is considered negatively charged when $p(t) = 1$. The master equation of the underlying 2-state Markov chain, see Fig. 1, for the occupational probability can be solved analytically [3]. Given the occupational probability at time t, the capture time $\tau_c(V_G)$ and emission time $\tau_e(V_G)$ of one defect, and the simulation time step Δt, using Eq. (1) allows the computation of the occupational probability of the next time step $p(t + \Delta t)$.

$$p(t+\Delta t) = \frac{\tau_e}{\tau_c+\tau_e} + \left(p(t) - \frac{\tau_e}{\tau_c+\tau_e}\right)e^{-\Delta t\left(\frac{1}{\tau_c}+\frac{1}{\tau_e}\right)} \quad (1)$$

In each time step the occupational probability of each defect is updated using Eq. (1), and from that the threshold voltage shift of all defects can be calculated as sum of all occupational probability multiplied by $\Delta V_{th,i}$, which is the impact of each defect on the overall ΔV_{th}, see Eq. (2).

$$\Delta V_{th}(t) = \sum_{i=1}^{N} p_i(t)\Delta V_{th,i} \quad (2)$$

This formula is applicable under the assumption that the defect concentration is comparatively low. For high defect concentrations a fully self-consistent calculation, which captures the interplay of defect charge and electrostatics, is needed, however, this is not feasible at circuit level.

Note that $\Delta V_{th,i}$ is often modeled using the charge sheet approximation (CSA), where the value of $\Delta V_{th,i}$ is considered to be proportional to the trap depth. However, single-defect studies revealed that the charge sheet approximation underestimates the impact of each defect to the overall ΔV_{th} [29]. This further leads to an underestimation of the device variability effects. As our model does not impose any restriction on the supplied defect parameters this deviation from the CSA can easily be taken into account by modifying the respective ΔV_{th} values.

B. DISCRETE SIMULATION MODE

In contrast to the probabilistic simulation mode, where each defect state is characterized by a probability, the discrete simulation mode assigns a Boolean variable to each defect, indicating whether the defect is occupied (occ = True) or unoccupied (occ = False). In order to compute the binary defect state, the behavior of the underlying Markov chain has to be simulated. Given a defect with capture time τ_c, emission time τ_e and occupation occ at time t, Algorithm 1 computes the occupation at time $t + \Delta t$. This simplified version of the Gillespie algorithm [30] samples the capture and emission time until Δt has elapsed. If the condition $(\tau_c + \tau_e) < \Delta t$ is true, the defect is so fast that sampling would require a potentially huge number of iterations of the while loop, but

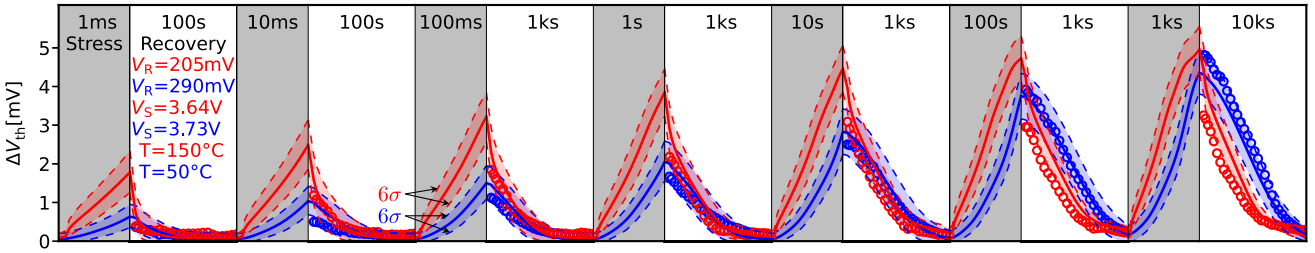

FIGURE 6. The SPICE model is directly calibrated to ΔV_{th} measurements of MSM sequences at different temperatures and stress voltages using the effective single defect decomposition (ESiD) method [9]. Each stress and recovery segment is depicted using a relative logarithmic time scale. This enables highly accurate and efficient calibration of the circuit model. Data is depicted as circles, taken from [9]. The solid line depicts the mean, with dashed lines enclosing the 6σ region, which is important for the consideration of variability effects.

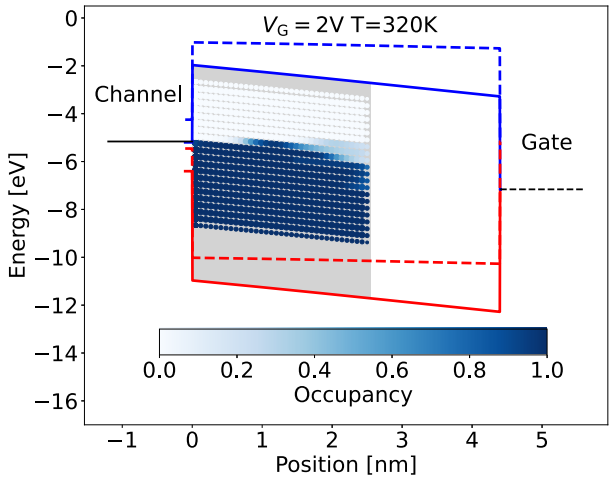

FIGURE 7. Band diagram ($V_G = V_{DD}$ (solid) flat band (dashed)) showing the occupation of uniformly sampled defects in the first 2.5nm. The trap band has been extracted using the ESiD method applied to the MSM sequence from Figure 6.

Algorithm 1 Discrete Simulation of Single Defect

Input: *occ* // Occupation at time t
Output: *occ* // Occupation at time $t + \Delta t$
if $(\tau_c + \tau_e) < \Delta t$ **then**
  $rand \leftarrow uniform(0, 1)$;
  $occ\_probability \leftarrow \tau_e/(\tau_c + \tau_e)$;
  return $rand < occ\_probability$;
else
  $t_{acc} \leftarrow 0$;
  while $t_{acc} < \Delta t$ **do**
    $rand \leftarrow -log(uniform(0, 1))$;
    if *occ* **then**
      $t_{acc} \leftarrow t_{acc} + \tau_e \cdot rand$;
    else
      $t_{acc} \leftarrow t_{acc} + \tau_c \cdot rand$;
    $occ \leftarrow$ not *occ*;
  return not *occ*;

in this case a simple coin flip weighted by the transition times suffices. Similar to the probabilistic simulation mode the resulting threshold voltage shift of a device is the sum of the threshold voltage shifts of its occupied defects.

Whereas the probabilistic simulation mode only allows simulating the average case, relevant for large-area devices, using the discrete simulation mode enables us to also consider device-to-device variability in transient simulations. Fig. 8 shows several simulated traces using the discrete mode compared to the probabilistic simulation. Performing many number of discrete mode simulations allows extracting the variability at each time step.

IV. COMPUTATIONALLY EFFICIENT SIMULATION

Regardless of whether the probabilistic or discrete simulation mode is used, for each time step and for each defect the respective transition times have to be determined. To achieve high performance our algorithm generates a lookup table for each defect, holding the transition times for the initial gate bias conditions at the start of the simulation. In order for this approach to be both correct and computationally efficient several subtleties need to be considered. With this approach the trace of each defect through the (τ_c, τ_e)-plane, i.e., capture and emission time (CET) map, is stored and referenced throughout the simulation. Fig. 9 depicts a few possible traces through the CET map, each trace is one defect starting at $V_G = 2$ V and ending at $V_G = 4$ V.

A. LOOKUP TABLE - GENERAL CONSIDERATIONS

One important aspect of the lookup table (LUT) is the range of the V_G-values for which the transition times should be computed. Our XSPICE model offers the parameters V_{Gmin}, V_{Gmax} and the LUT-size. The resulting interval is divided in a number of equidistant steps. Considering Fig. 9, V_{Gmin} and V_{Gmax} determine the endpoints of the trace, while the LUT-size determines the resolution of the trace.

Choosing the size of the lookup table allows a trade-off between accuracy and computational effort. While a small lookup table is computationally efficient, a table with too few entries will impact the accuracy, vice versa for a too large table. Fig. 10 depicts the $\Delta V_{th}(t)$ traces of the same device

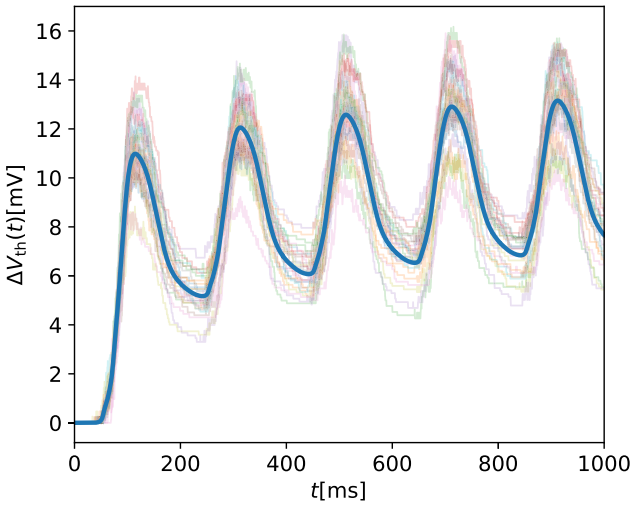

FIGURE 8. Different $\Delta V_{th}(t)$ traces of discrete simulations (pale) compared to the probabilistic trace (thick blue). While the probabilistic simulation allows to investigate the average behavior, variability can be investigated by performing numerous discrete simulations. As can be seen, the variability introduced by charge trapping can lead to a considerable deviation from the mean.

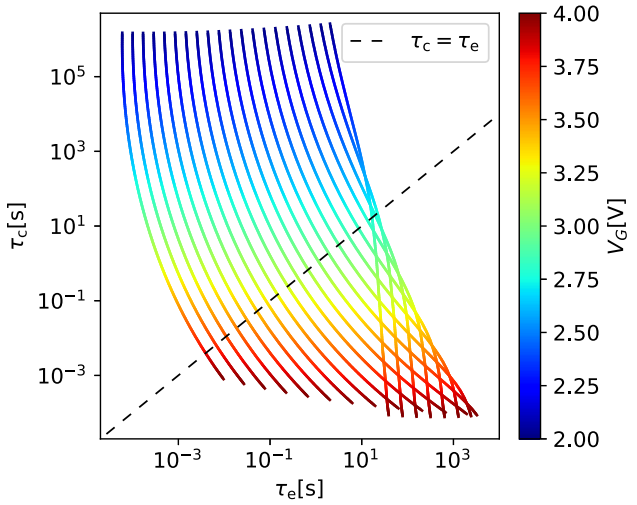

FIGURE 9. Example of defects changing their capture and emission times (CET) depending on the gate voltage, represented as a trace through the CET map. Each trace represents a single defect.

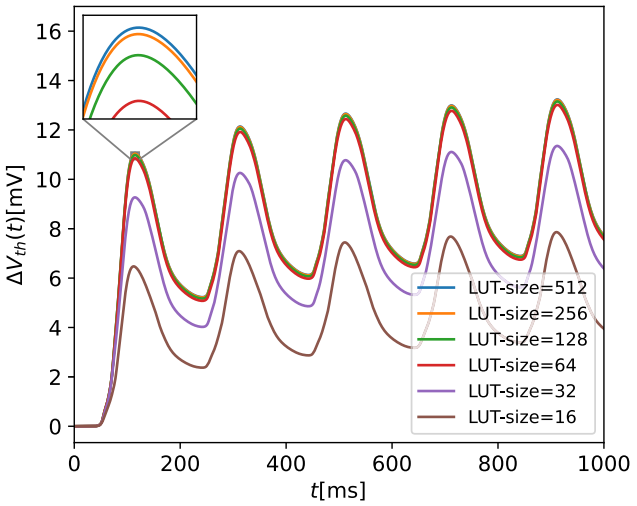

FIGURE 10. $\Delta V_{th}(t)$ traces of the same model with different lookup table sizes. For small lookup table sizes the accuracy severely degrades. Above 100 entries no significant improvement in accuracy is achieved.

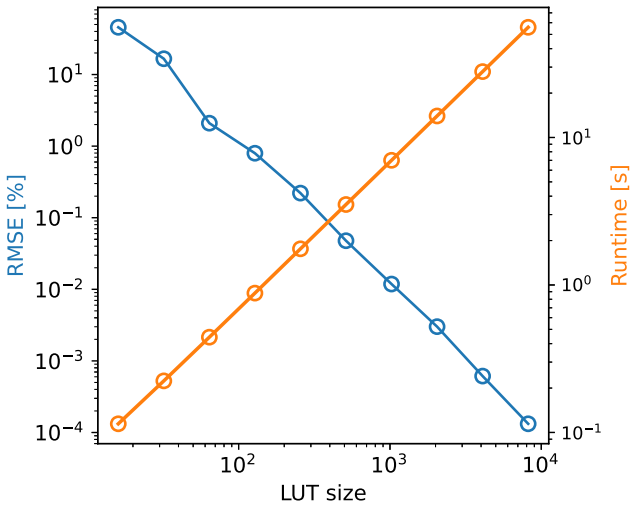

FIGURE 11. Relative RMS error (blue) of $\Delta V_{th}(t)$-traces of Fig. 10 (only traces up to a LUT size of 512 shown in Fig. 10) compared to a simulation with a LUT of size 16384. Already with 256 entries the relative RMS error is in the range of 0.1%. Time needed to compute the LUT is shown in orange.

with different lookup table sizes. As expected, accuracy severely degrades for too small tables and saturates for larger tables. Fig. 11 depicts the relative RMS error of simulations with different LUT sizes compared to simulation with LUT size 16384 (blue) and the time needed to build the LUT (orange), showing a sharp decrease in RMS error with increasing LUT size and the increase in time needed to build larger the LUTs.

B. LOOKUP TABLE CACHE EFFICIENCY

In each time step of the simulation the total threshold voltage shift of a device is determined by iterating through all defects of the device and determining their contribution. Regardless of the computation mode the transition times of each defect for the respective gate voltage need to be gathered from the lookup table. As these lookup tables can become quite large (several megabytes for thousands of defects) a cache-friendly data layout becomes important.

While from an object-oriented perspective it makes sense to store the transition times of each defect from the voltage $V_1 = V_{\mathrm{Gmin}}$ to $V_N = V_{\mathrm{Gmax}}$ contiguously, i.e., the transition times of defect 1, then the transition times of defect 2 until defect D: $\tau_1(V_1), \tau_1(V_2), \ldots, \tau_1(V_N)$, $\tau_2(V_1), \tau_2(V_2), \ldots, \tau_D(V_N)$, this does not harness the prefetching mechanism offered by the cache system. Since in a given time step the transition times of all defects of

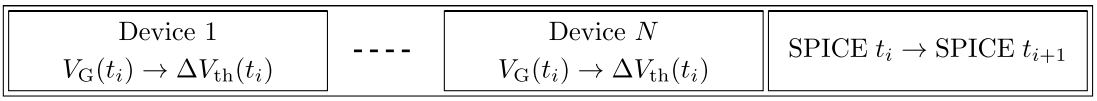

FIGURE 12. One time step performed by the sequential implementation. The ΔV_{th} calculations of each device are invoked sequentially and the SPICE time step update is performed only after all of them are finished.

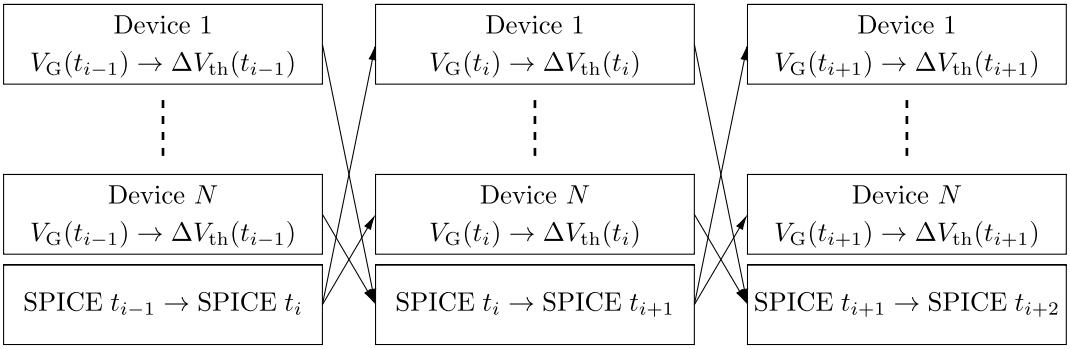

FIGURE 13. Visualization of parallel implementation. The ΔV_{th} calculations of each device happen in a separate process. Although this results in the values of ΔV_{th} lagging behind the rest of the simulation by one time step, we found the loss of accuracy to be well below 0.05%, see Section V-B.

a device for the same gate voltage need to be accessed, it is more efficient to reorganize the transition times such that the transition times of the same voltage of different defects are stored contiguously in memory. This is achieved by the following layout: $\tau_1(V_1), \tau_2(V_1), \ldots, \tau_D(V_1), \tau_1(V_2), \tau_2(V_2), \ldots, \tau_D(V_N)$.

Iterating through the defects of a device with this data layout is much faster, since the access of $\tau_i(V_j)$ also loads $\tau_{i+1}(V_j), \tau_{i+2}(V_j), \ldots$ into the cache, resulting in faster access times overall.

C. EMPLOYING SIMD

Virtually every modern processor implements one or several single instruction, multiple data (SIMD) extensions, which enable a single instruction to operate on multiple data elements at the same time. For example, the widely available AVX2 extension on x86 offers instructions that operate on 256-bit SIMD registers, which enables the processing of 8 single-precision floating-point numbers in parallel. Utilizing the instructions offered by AVX2 the interpolation of τ_c and τ_e and the evaluation of Eq. (1) can be performed on 8 defects in parallel, resulting in a considerable speed up. However, Algorithm 1 does not benefit from SIMD, as generally a different number of iterations in the while loop have to be performed for each defect, i.e., different control flows for each defect.

D. UTILIZING MULTIPROCESSING FOR DEFECT SIMULATION

For large circuits containing devices with a high number of defects, simulation times can become prohibitively long. In such defect-based simulations, the computational effort is dominated by defect-related calculations rather than the matrix operations performed by SPICE. Consequently, parallelizing the defect computations can significantly accelerate the overall simulation. In addition to the conventional XSPICE model, we therefore implemented a second model that performs all defect calculations in separate processes. In contrast to Fig. 12, which depicts the sequence of computations done conventionally, Fig. 13 presents a high-level overview of this parallelization approach, where the defect calculations are distributed across N processors. The main idea is to compute $\Delta V_{th}(t_i)$ for each device while SPICE executes its matrix operations for the same time step t_i. The resulting ΔV_{th} values then become available at time t_{i+1}. Although this introduces a one time step delay in the ΔV_{th} information, we found that the resulting loss in accuracy is negligible due to the inertia of the defect behavior. A thorough investigation of the error imposed by the parallel implementation is performed in Section V-B.

E. DISCARDING INACTIVE DEFECTS

Defect ensembles extracted using ESiD are assumed to be distributed uniformly in the bulk oxide. However, this means that traps deep in the oxide (implying large time constants) are also present due to this enforced uniformity. As such slow defects may never become active throughout a simulation, calculating their contributions is a waste of computational resources, and it is desirable to discard these defects from the simulation. A defect is considered to be "active" if there is a reasonable chance that the defect changes its charge state throughout the simulation. To this end we define three "inactive-windows" and one window that allows simplified computation, see Fig. 14. If a defect's trace through the CET map never leaves one of these regions of inactivity it can be discarded automatically due to inactivity, otherwise it is

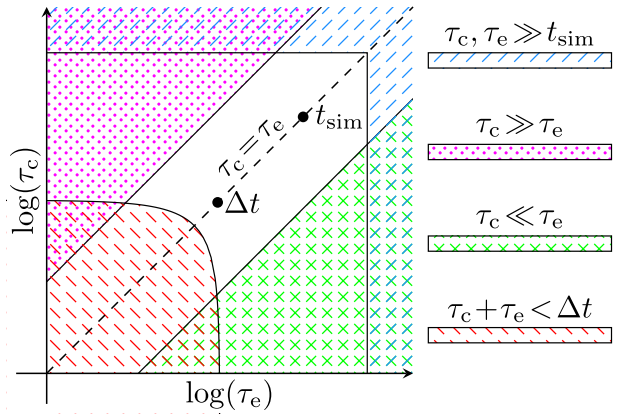

FIGURE 14. CET map divided into different regions. Defects with too large time constants (blue), or defects with one time constant strictly dominating (magenta and green) for all possible gate voltages can safely be discarded at the start of the simulation. Simplified computation can be employed if defects are too fast (red).

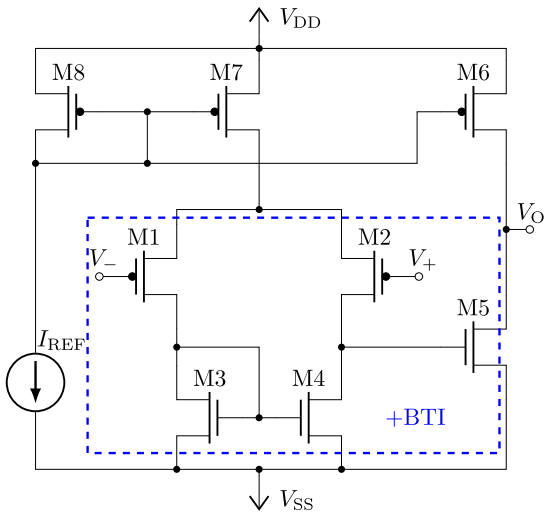

FIGURE 15. Schematic of two-stage CMOS operational amplifier as described in [31]. For the run time and accuracy simulation an inverting amplifier, built from this operational amplifier was simulated, with M1 to M5 (enclosed by dashed blue line) enhanced with our model. BTI dynamics in M6, M7 and M8 are neglected, as they are only used for biasing.

considered active. Fig. 14 highlights four different regions of the CET map. The blue region corresponds to $\tau_c \gg t_{sim}$ and $\tau_e \gg t_{sim}$. A defect that is always (i.e., for all values in $[V_{Gmin}, V_{Gmax}]$) in the blue region is so slow that it will practically never switch state during t_{sim}, and can consequently be discarded. The magenta region corresponds to $\tau_c \gg \tau_e$, meaning that the capture time is several orders of magnitude larger than the emission rate. A defect that is always in the magenta region is essentially always off. The opposite is true for the green region, which represents the region where $\tau_c \ll \tau_e$. A defect that is discarded because its trace never leaves the green or blue region is captured as a constant ΔV_{th} contribution, which just needs to be determined at the start of the simulation. For traces that always remain in the green region the contribution is its full ΔV_{th}, while for the blue region it is between 0 and ΔV_{th}, depending on the initial state.

Since the notions of "larger" and "faster" depend on the context, the model offers two parameters with which the user can adjust the borders of the aforementioned regions. The first parameter determines how many orders of magnitude τ_c needs to be larger (smaller) than τ_e to be in the magenta (green) region. The second parameter directly specifies t_{sim}.

With a sensible choice of these two parameters a large number of inactive defects can be automatically discarded at the start of the simulation. This generally results in a considerable reduction of computational effort with practically no impact on simulation accuracy. Note that disregarding the defects outside the active time window for dynamic charge trapping has no impact on the time-zero state of the device.

In contrast, the red region in Fig. 14, corresponding to $\tau_c + \tau_e < \Delta t$ and handled by the if-branch in Algorithm 1, represents defects that are so fast that employing the else-branch of Algorithm 1 does not make sense due to the high amount of iterations needed to arrive at a result. Since such defects switch faster than can be sampled with Δt, a simple weighted coin flip with weight $\tau_e/(\tau_c + \tau_e)$ suffices to determine their occupation. In contrast to the blue, magenta and green regions this criterion needs to be checked in each iteration throughout the simulation. In the probabilistic mode this criterion does not need to be checked, as the computation of the occupational probability is just an evaluation of Eq. (1).

V. SIMULATION PERFORMANCE

To evaluate the simulation performance of our implementations, we simulate a 501-stage ring oscillator and an inverting amplifier, built using an operational amplifier as depicted in Fig. 15, with and without defects and investigate the slowdown of the simulation with defects depending on the number of defects per device. While in the case of the ring oscillator there are 1002 instances (PMOS and NMOS of each inverter) of our model, in the case of the inverting amplifier only M1 to M5 are enhanced with our model, resulting in 5 instances of our model.

The resulting slowdown factors are shown in Fig. 16 for the sequential implementation and Fig. 17 for the parallel implementation, where solid lines denote the probabilistic mode and dashed lines the discrete mode. Even when each device in the circuit contains 1000 defects, the slowdown relative to the defect-free simulation remains well below a factor of 10 for both approaches.

In the case of the inverting amplifier the slowdown of the parallel implementation starts to follow a linear trend in the case of more than 1000 defects. Since the inverting amplifier circuit only contains five instances of our model, which results in 5 additional processes, the speed-up of the parallel implementation compared to the sequential implementation is strictly bounded to a factor of $5 + 1$. For 10^4 defects the observed speed-up is approximately a factor of 3.

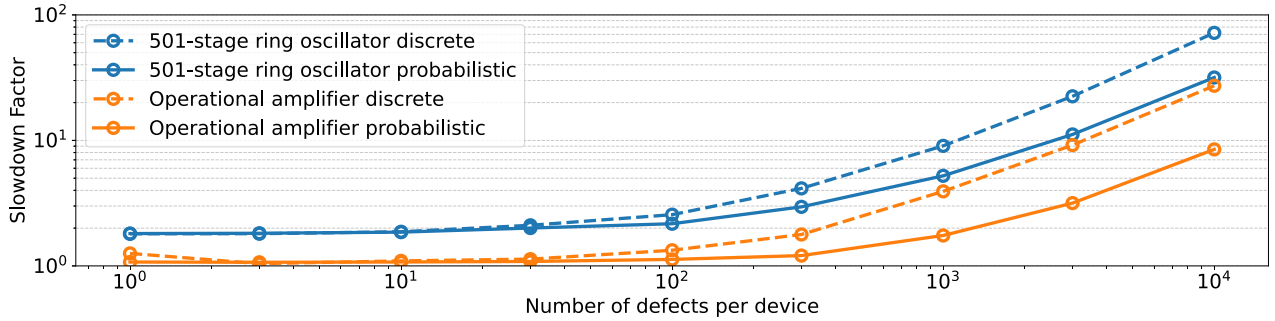

FIGURE 16. Comparison of the run time of SPICE simulations using the sequential implementation of a 501-stage ring oscillator (blue) and operational amplifier (orange) with varying numbers of defects per device against the respective defect-free reference simulation.

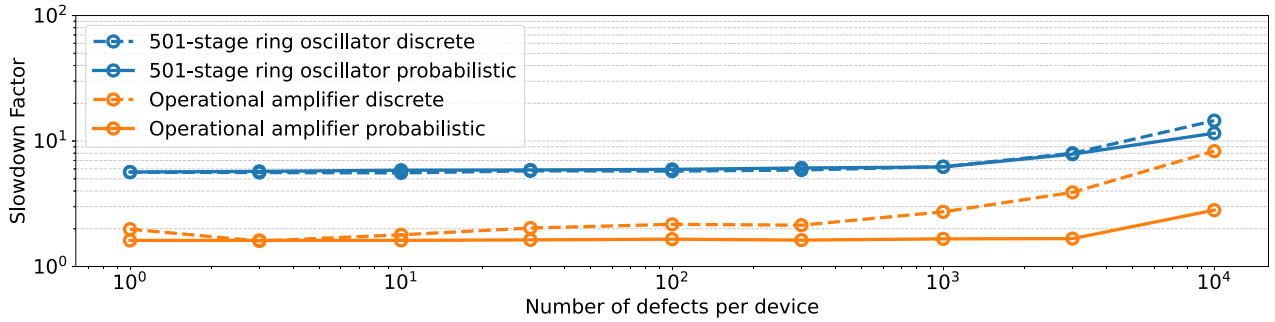

FIGURE 17. Same as Fig. 16, but using the parallel implementation. In simulations with large numbers of defects this results in a considerable speed-up, although for small numbers of defects the parallel implementation actually results in slow-down compared to the sequential implementation. Especially in the case of the 501-stage ring oscillator, where 1002 additional processes are spawned, inter-process synchronization adds considerable overhead.

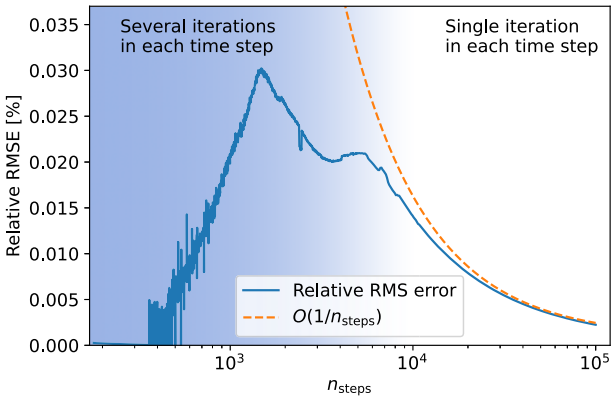

FIGURE 18. Relative RMS of the ΔV_{th} signal of the parallel implementation compared to the sequential implementation of M5 when using Fig. 15 in an inverting amplifier. For large n_{steps} the error starts to follow the expected $1/n_{steps}$ trend, while for small n_{steps} the error between parallel implementation and sequential implementation becomes small and erratic as several iterations are performed in each time step by SPICE.

In the case of the ring oscillator the slowdown of the parallel implementation remains nearly constant with increasing defect count, the slowdown of the sequential implementation scales approximately linearly for larger numbers of defects. Consequently, for high defect densities the parallel approach clearly achieves shorter simulation times, as it can distribute the computational workload across multiple processes. In contrast, for a few hundred defects the sequential implementation performs better because the inter-process communication overhead of the parallel scheme dominates when only few defects are present. All simulations were carried out on a dual-socket AMD EPYC 9355 system operating at 4.4 GHz and providing a total of 128 threads. Overall this represents a substantial improvement over a previously reported implementation, which reports slowdowns of more than a factor of 100 for 500 defects [12].

A. RUN TIME OF PROBABILISTIC AND DISCRETE MODE

In terms of run time it is observable in Fig. 16 and Fig. 17 that for high defect counts the discrete simulation mode needs considerably more run time than the probabilistic mode, which can be explained by examining Algorithm 1. Whereas in probabilistic mode only Eq. (1) has to be evaluated in each iteration, which benefits from SIMD and is constant effort regardless of τ_c and τ_e, in discrete mode Algorithm 1 has to be evaluated, which may have to perform several iterations depending on the values of τ_c and τ_e, with no theoretical upper bound on the number of iterations needed until termination. Especially when $\tau_c + \tau_e \geq \Delta t$ and $\tau_c \approx \tau_e \approx \Delta t/2$ the while-loop of Algorithm 1 will likely be evaluated several times, which further increases run time in discrete mode.

Apart from these differences both probabilistic and discrete mode first need to compute (i.e., look up and interpolate) the values of τ_c and τ_e in each iteration, which is roughly as computationally intensive as evaluating Eq. (1).

B. ACCURACY OF PARALLEL IMPLEMENTATION

As utilizing multiprocessing on the basis of single SPICE models as mentioned above incurs a one time step delay for the SPICE model, an investigation of the error introduced by this method is performed. To compare the error introduced by the parallel implementation we performed a simulation that contained the exact same circuit twice, once using the sequential implementation and once using the parallel implementation. The RMS error of the ΔV_{th} curves of the two circuits was computed. For this comparison we used and inverting amplifier built using the operational amplifier depicted in Fig. 15, comparing the ΔV_{th} curves of M5. It turns out that the error of the parallel implementation compared to the sequential implementation is highly dependent on the number of time steps in the simulation n_{steps}, Fig. 18 depicts the relative RMS error between the two modes. Based on the one time step delay of the parallel implementation it is expected that the RMS error follows $1/n_{steps}$. However, Fig. 18 shows that the RMS error starts at basically 0, follows a somewhat erratic pattern, peaks at about 0.03% for roughly 10^3 steps and gradually transitions into the expected $1/n_{steps}$ shape. This behavior can be explained by the algorithm SPICE uses for transient analysis. Depending on the gradients present during the simulation and options like `RELTOL`, `ABSTOL`, `TRTOL`, `CHGTOL`, etc., SPICE has to perform several Newton–Raphson iterations for a single time step to ensure convergence. Especially for small n_{steps} values several iterations are performed in each time step. Therefore, also the parallel and sequential implementations are invoked several times in each time step, which allows the parallel implementation to ''catch up'' and compute the same value as the sequential implementation, leading to an RMS error close to zero. For large values of n_{steps} only a single iteration is performed leading to the expected behavior. In between the two extremes of small and large n_{steps} SPICE performs several iterations for some time steps, while only performing one iteration in other time steps, leading to the erratic curve in Fig. 18.

VI. CONCLUSION

We have presented an XSPICE extension for ngspice that enables computationally efficient physics-based bias temperature instability (BTI) simulations using defect-centric models. Both NBTI and PBTI are captured by a threshold voltage shift ΔV_{th}. By combining considerable improvements, such as lookup tables, cache efficient data layout, single instruction, multiple data extensions (SIMD) and multiprocessing, our C implementation substantially reduces the amount of time needed to simulate NMP-based BTI SPICE simulations. Whereas previous Verilog-A implementations report slow-downs of around two orders of magnitude for a few hundred defects [12], our implementation offers a slow-down of around a factor 10 for several thousand defects. This opens new possibilities of assessing the impact of BTI at the circuit level. Based on the feasibility of such simulations, we advocate a direct defect-centric simulation approach in which BTI effects are directly captured by the NMP model, eliminating the traditional separation between nominal and aged simulations. Future work will focus on integrating our implementation in existing device models, which also allows to capture BTI-induced mobility degradation [32].

REFERENCES

[1] J. H. Stathis, S. Mahapatra, and T. Grasser, ''Controversial issues in negative bias temperature instability,'' *Microelectron. Rel.*, vol. 81, pp. 244–251, Feb. 2018.
[2] T. Grasser, B. Kaczer, W. Goes, H. Reisinger, T. Aichinger, P. Hehenberger, P.-J. Wagner, F. Schanovsky, J. Franco, M. T. Toledano Luque, and M. Nelhiebel, ''The paradigm shift in understanding the bias temperature instability: From reaction–diffusion to switching oxide traps,'' *IEEE Trans. Electron Devices*, vol. 58, no. 11, pp. 3652–3666, Nov. 2011.
[3] T. Grasser, ''Stochastic charge trapping in oxides: From random telegraph noise to bias temperature instabilities,'' *Microelectron. Rel.*, vol. 52, no. 1, pp. 39–70, Jan. 2012. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0026271411004203
[4] G. Rzepa, M. Waltl, W. Goes, B. Kaczer, and T. Grasser, ''Microscopic oxide defects causing BTI, RTN, and SILC on high-k FinFETs,'' in *Proc. Int. Conf. Simulation Semiconductor Processes Devices (SISPAD)*, Sep. 2015, pp. 144–147.
[5] J. Berens, M. Weger, G. Pobegen, T. Aichinger, G. Rescher, C. Schleich, and T. Grasser, ''Similarities and differences of BTI in SiC and si power MOSFETs,'' in *Proc. IEEE Int. Rel. Phys. Symp. (IRPS)*, Apr. 2020, pp. 1–7.
[6] T. Knobloch, D. Waldhoer, and T. Grasser, ''Modeling 2D material-based nanoelectronic devices in the presence of defects,'' *IEEE Nanotechnol. Mag.*, vol. 17, no. 4, pp. 15–25, Aug. 2023.
[7] T. Grasser, W. Goes, Y. Wimmer, F. Schanovsky, G. Rzepa, M. Waltl, K. Rott, H. Reisinger, V. V. Afanas'ev, A. Stesmans, A.-M. El-Sayed, and A. L. Shluger, ''On the microscopic structure of hole traps in pMOSFETs,'' in *IEDM Tech. Dig.*, Dec. 2014, pp. 21.1.1–21.1.4.
[8] C. Wilhelmer, D. Waldhoer, M. Jech, A.-M.-B. El-Sayed, L. Cvitkovich, M. Waltl, and T. Grasser, ''Ab initio investigations in amorphous silicon dioxide: Proposing a multi-state defect model for electron and hole capture,'' *Microelectron. Rel.*, vol. 139, Dec. 2022, Art. no. 114801. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0026271422003250
[9] D. Waldhoer, C. Schleich, J. Michl, B. Stampfer, K. Tselios, E. G. Ioannidis, H. Enichlmair, M. Waltl, and T. Grasser, ''Toward automated defect extraction from bias temperature instability measurements,'' *IEEE Trans. Electron Devices*, vol. 68, no. 8, pp. 4057–4063, Aug. 2021.
[10] G. Rzepa, J. Franco, B. O'Sullivan, A. Subirats, M. Simicic, G. Hellings, P. Weckx, M. Jech, T. Knobloch, M. Waltl, P. J. Roussel, D. Linten, B. Kaczer, and T. Grasser, ''Comphy—A compact-physics framework for unified modeling of BTI,'' *Microelectron. Rel.*, vol. 85, pp. 49–65, Jun. 2018. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0026271418301641
[11] P. Weckx, M. Simicic, K. Nomoto, M. Ono, B. Parvais, B. Kaczer, P. Raghavan, D. Linten, K. Sawada, H. Ammo, S. Yamakawa, A. Spessot, D. Verkest, and A. Mocuta, ''Defect-based compact modeling for RTN and BTI variability,'' in *Proc. IEEE Int. Rel. Phys. Symp. (IRPS)*, Apr. 2017, pp. CR-7.1–CR-7.6.
[12] D. Sangani, D. Claes, P. Weckx, B. Kaczer, and G. Gielen, ''Towards a computationally efficient verilog—A defect-centric BTI compact model for circuit aging simulations,'' in *Proc. IEEE Int. Rel. Phys. Symp. (IRPS)*, Mar. 2025, pp. 1–6.
[13] (2025). *Spectre RelXpert Reliability Simulator User Guide*. Accessed: May 11, 2026. [Online]. Available: https://support.cadence.com/
[14] B. Tudor, J. Wang, C. Sun, Z. Chen, Z. Liao, R. Tan, W. Liu, and F. Lee, ''MOSRA: An efficient and versatile MOS aging modeling and reliability analysis solution for 45nm and below,'' in *Proc. 10th IEEE Int. Conf. Solid-State Integr. Circuit Technol.*, Nov. 2010, pp. 1645–1647.

[15] *PrimeSim HSPICE User Guide: Basic Simulation and Analysis*, Synopsys, Inc., Sunnyvale, CA, USA, 2024.
[16] P. R. Genssler, H. E. Barkam, K. Pandaram, M. Imani, and H. Amrouch, "Modeling and predicting transistor aging under workload dependency using machine learning," *IEEE Trans. Circuits Syst. I, Reg. Papers*, vol. 70, no. 9, pp. 3699–3711, Sep. 2023.
[17] A. Toro-Frías, P. Martín-Lloret, J. Martin-Martinez, R. Castro-López, E. Roca, R. Rodriguez, M. Nafria, and F. V. Fernández, "Reliability simulation for analog ICs: Goals, solutions, and challenges," *Integration*, vol. 55, pp. 341–348, Sep. 2016. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0167926016300153
[18] P. Gholve, P. Chatterjee, C. Pasupuleti, H. Amrouch, N. Gangwar, S. Das, U. Sharma, V. M. van Santen, and S. Mahapatra, "CARAT—A reliability analysis framework for BTI-HCD aging in circuits," *Solid-State Electron.*, vol. 201, Mar. 2023, Art. no. 108586. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0038110122003586
[19] N. Parihar, N. Goel, S. Mukhopadhyay, and S. Mahapatra, "BTI analysis tool—Modeling of NBTI DC, AC stress and recovery time kinetics, nitrogen impact, and EOL estimation," *IEEE Trans. Electron Devices*, vol. 65, no. 2, pp. 392–403, Feb. 2018.
[20] T. Grasser, K. Rott, H. Reisinger, M. Waltl, F. Schanovsky, and B. Kaczer, "NBTI in nanoscale MOSFETs—The ultimate modeling benchmark," *IEEE Trans. Electron Devices*, vol. 61, no. 11, pp. 3586–3593, Nov. 2014.
[21] J. Martin-Martinez, B. Kaczer, M. Toledano-Luque, R. Rodriguez, M. Nafria, X. Aymerich, and G. Groeseneken, "Probabilistic defect occupancy model for NBTI," in *Proc. Int. Rel. Phys. Symp.*, Apr. 2011, pp. XT.4.1–XT.4.6.
[22] V. V. A. Camargo, B. Kaczer, T. Grasser, and G. Wirth, "Circuit simulation of workload-dependent RTN and BTI based on trap kinetics," *Microelectron. Rel.*, vol. 54, no. 11, pp. 2364–2370, Nov. 2014. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0026271414001826
[23] K. Tselios, T. Knobloch, D. Waldhoer, B. Stampfer, E. Ioannidis, H. Enichlmair, R. Minixhofer, T. Grasser, and M. Waltl, "Revealing the impact of gate area scaling on charge trapping employing SiO_2Transistors," *IEEE Trans. Device Mater. Rel.*, vol. 23, no. 3, pp. 355–362, Sep. 2023.
[24] A. S. Andreo, V. M. van Santen, R. Castro-López, E. Roca, H. Amrouch, and F. V. Fernández, "Workload compression techniques to scale defect-centric BTI models to the circuit level," *IEEE Trans. Circuits Syst. I, Reg. Papers*, vol. 72, no. 11, pp. 7044–7057, Nov. 2025.
[25] J. Michl, A. Grill, D. Claes, G. Rzepa, B. Kaczer, D. Linten, I. Radu, T. Grasser, and M. Waltl, "Quantum mechanical charge trap modeling to explain BTI at cryogenic temperatures," in *Proc. IEEE Int. Rel. Phys. Symp. (IRPS)*, Apr. 2020, pp. 1–6.
[26] R. Tu, E. Rosenbaum, W. Chan, C. Li, E. Minami, K. Quader, P. Ko, and C. Hu, "Berkeley reliability tools-BERT," *IEEE Trans. Comput.-Aided Design Integr. Circuits Syst.*, vol. 12, no. 10, pp. 1524–1534, Oct. 1993.
[27] G. Pobegen and T. Grasser, "On the distribution of NBTI time constants on a long, temperature-accelerated time scale," *IEEE Trans. Electron Devices*, vol. 60, no. 7, pp. 2148–2155, Jul. 2013.
[28] H. Vogt, G. Atkinson, and P. Nenzi. (2025). *Ngspice Users Manual Version 45*. [Online]. Available: http://www.ngspice.org
[29] K. Tselios, B. Stampfer, J. Michl, E. Ioannidis, H. Enichlmair, and M. Waltl, "Distribution of step heights of electron and hole traps in SiON nMOS transistors," in *Proc. IEEE Int. Integr. Rel. Workshop (IIRW)*, Oct. 2020, pp. 1–6.
[30] D. T. Gillespie, "A general method for numerically simulating the stochastic time evolution of coupled chemical reactions," *J. Comput. Phys.*, vol. 22, no. 4, pp. 403–434, Dec. 1976.
[31] S. Franco, *Analog Circuit Design*. New York, NY, USA: McGraw-Hill, 2015.
[32] C. Bogner, H. Reisinger, I. Lahbib, A. Martin, W. Gustin, T. Grasser, and C. Schlünder, "On the role of NBTI and PBTI induced mobility degradation for compact modeling in metal-gate/high-k FETs," in *Proc. IEEE Int. Integr. Rel. Workshop (IIRW)*, Oct. 2023, pp. 1–7.

JOSEF SALZMANN received the bachelor's and master's degrees in computer engineering from TU Wien, in 2021 and 2024, respectively. He is currently pursuing the Ph.D. degree with the Institute for Microelectronics. He works on the development of computer models to simulate the impact of charge trapping effects in integrated circuits.

ALEXANDER KARL received the bachelor's and master's degrees in technical physics from TU Wien, in 2020 and 2022, respectively. He is currently pursuing the Ph.D. degree with the Institute for Microelectronics. His research focuses on the modeling of field-effect transistors based on two-dimensional materials. His research interests include reliability and stability issues in these novel devices, such as bias temperature instability, random telegraph noise, and gate leakage, with a particular emphasis on understanding their microscopic origins.

DOMINIC WALDHÖR received the joint bachelor's degree in electrical engineering and the master's degree in microelectronics and photonics from TU Wien, in 2018, and the Ph.D. degree from the Institute for Microelectronics, in 2024, with a focus on multiscale modeling of charge trapping processes in transistors. In his doctoral work, he combined density functional theory with physics-based device simulations to advance the understanding of defect-related reliability phenomena, such as bias temperature instability and random telegraph noise, and contributed to the development of the Comphy device simulator, including models based on non-radiative multiphonon theory and methods for defect parameter extraction. His research interests include theoretical semiconductor device physics, ab initio modeling of crystalline and disordered materials, defect physics, and reliability modeling in emerging device technologies, including transistors based on 2-D materials.

TIBOR GRASSER (Fellow, IEEE) is currently a Professor with TU Wien. He has edited various books, e.g., on the bias temperature instability, hot carrier degradation, and low-frequency noise (all with Springer). He is a Distinguished Lecturer of IEEE EDS. He was a recipient of the Best and Outstanding Paper Awards at IRPS, in 2008, 2010, 2012, and 2014; IPFA, in 2013 and 2014; ESREF in 2008; and the IEEE EDS Paul Rappaport Award, in 2011. He has been involved in outstanding conferences, such as IEDM (the General Chair, in 2021), IRPS, ESSDERC (the TPC Co-Chair, in 2015), IIRW (the General Chair, in 2014), and SISPAD (the General Chair, in 2007). He served as an Associate Editor for IEEE TRANSACTIONS ON ELECTRON DEVICES and *Microelectronics Reliability* (Elsevier).

MICHAEL WALTL (Senior Member, IEEE) is currently a University Professor of robust microelectronics and the Director of the Device Characterization Laboratory, Institute for Microelectronics. He also leads the Christian Doppler Laboratory for Single-Defect Spectroscopy, as well as several other funded research projects. In addition, he heads multiple collaborative projects with leading international industrial partners, including imec, ams-OSRAM AG, Infineon, and GlobalFoundries. In addition, he has authored and co-authored more than 200 publications. He is a member of the technical program committees of several other conferences in Europe (ESSERC) and USA (IRPS). He received the Best Paper Award at the 2014 IRPS and the Best Poster Award at the 2019 ICPTDC. He was a co-recipient of the Best Student Paper Award at the 2019 IIRW, the 2019 IEDM, and the 2020 DRC. He serves on the management committees of the IEEE International Electron Devices Meeting (IEDM), the IEEE Integrated Reliability Workshop (IIRW), and European Symposium on the Reliability of Electron Devices, Failure Physics and Analysis (ESREF). He serves as an Associate Editor for *Microelectronics Engineering* and regularly reviews for leading journals, including *Nature Electronics*, IEEE TRANSACTIONS ON ELECTRON DEVICES, *Microelectronics Reliability*, and *Journal of Applied Physics*.