IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 13 July 2023, accepted 21 August 2023, date of publication 29 August 2023, date of current version 7 September 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3309649

RESEARCH ARTICLE

Simulation Acceleration of Bit Error Rate Prediction and Yield Optimization of 3D V-NAND Flash Memory

YOHAN KIM[1,2], (Member, IEEE), AND SOYOUNG KIM[3], (Senior Member, IEEE)

[1]Department of Semiconductor and Display Engineering, College of Information and Communication Engineering, Sungkyunkwan University, Suwon 16419, South Korea
[2]Computational Science and Engineering Team, Innovation Center, Samsung Electronics, Suwon 16677, South Korea
[3]College of Information and Communication Engineering, Sungkyunkwan University, Suwon 16419, South Korea

Corresponding author: Soyoung Kim (ksyoung@skku.edu)

This work was supported by the National Research Foundation of Korea (NRF) grants funded by the Korea Government Ministry of Science and Information and Communication Technology (MSIT) (No. 2020R1A5A1019649 and No. 2020M3H2A1076786).
The EDA tool was supported by the IC Design Education Center (IDEC), Korea.

ABSTRACT When designing 3D V-NAND technologies with a gate induced drain leakage (GIDL) assisted erase scheme, many experiments must be conducted to determine the optimal GIDL design targets to achieve fast erase performance and secure yield characteristics. However, only a limited amount of data can be used since V-NAND processes are time-consuming and expensive in the early stage of development. TCAD and numerical methods also require a considerable amount of time and effort to calculate bit error rate (BER), and it is impossible to explore the entire design spaces in time. In this paper, we propose a novel simulation acceleration technique for bit error rate prediction and yield optimization in 3D V-NAND technology. This acceleration framework includes a machine learning (ML)-based compact model for the lognormal variability of GIDL currents and a physics-inspired slow cell model for the read margin reduction. Using a combination of these models with efficient Monte Carlo (MC) circuit simulations, we can accurately estimate threshold voltage (V_{th}) distributions to explore the entire design spaces using a limited amount of data. Based on the proposed technique, the predictive model achieves high accuracy in the current 176-layer V-NAND technology, and it also provides high scalability with respect to GIDL transistor geometries, temperatures, supply voltages, variabilities, and the number of stacking layers. Moreover, a contour map of bit error rate is newly introduced for the efficient design space exploration and read margin prediction. Therefore, the results indicate that the proposed framework can be further extended to large-scale experimental data and new architectures to accelerate the yield optimization in next-generation 3D V-NAND flash memory development.

INDEX TERMS Acceleration, artificial neural network, bit error rate, circuit simulation, compact model, GIDL-assisted erase, machine learning, pathfinding, read margin, V-NAND flash memory.

I. INTRODUCTION

Three-dimensional vertical-NAND (3D V-NAND) devices for high speed and capacity products have been extensively utilized in data-driven computing environments, over conventional computation-driven computing, and has been further strengthened by the critical involvement of big data. This has also enabled the modern everywhere-always-connected life through smart handheld devices, along with telecommunication technologies [1]. Based on vertical stacking technologies of many memory layers, cell arrays of the 3D V-NAND induced the tipping point into more aggressive scaling of the bit storage density without relying on the reduction in cell dimensions. Fig. 1(a) shows the structure of 3D V-NAND devices with a single word line (WL) scheme which was the leading candidate for the early 3D V-NAND technologies.

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

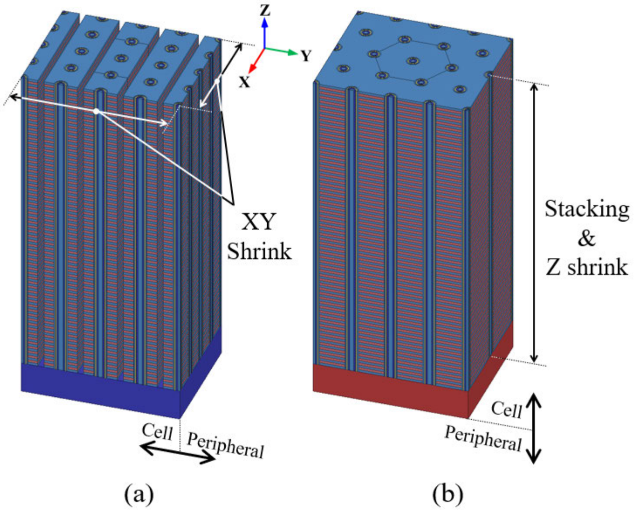


FIGURE 1. The 3D V-NAND structures of (a) single word line and (b) word plane scheme with the cell over peripheral (COP).

In this era of V-NAND development, the horizontal placement of the cell arrays with peripheral circuits was common and reducing the cell pitch (XY shrink) was a key driver. However, it was extremely vulnerable to deep channel holes because of leaning problems during the high aspect ratio (HAR) plasma etch process [2]. Therefore, the word plane scheme was utilized to overcome these HAR issues [3], and it was specialized in the 3D vertical stacking process, which has over 100 layers with the multiple-stack technologies, as shown in Fig. 1(b). However, it is estimated that the allowed maximum number of stacking layers would be limited to 400 stages without any cell dimension reductions (Z shrink) [4]. To prolong the historical scaling trends of the storage density, the cell over peripheral (COP) structure is newly introduced [5], which offers the chance to reduce the chip area by arranging the peripheral circuits under the memory cell array rather than alongside the array. However, the COP structure cannot operate the bulk erase scheme, which increases the channel string potential (V_{ch}) using direct contact with the memory cells, as shown in Fig. 2(a). The bulk erase has excellent erase performances and large read voltage windows. To simplify the technological process and increase the integration density of the COP structures, the GIDL-assisted erase scheme (GIDL erase) has become the gold standard for modern state-of-the-art 3D V-NAND products [6]. The problem is that its erase efficiency is very poor and GIDL-related erase failures are frequent due to the inherent GIDL generation process of hole carriers. The GIDL erase is also sensitive to the supplying electric fields for band-to-band-tunneling [7], and it causes the inevitable channel potential delays, as shown in Fig. 2(b).

In the flash memory industry, technology computer aided design (TCAD) and numerical simulations are actively utilized to resolve these GIDL-related failures and yield problems, because many experiments must be conducted to obtain the optimal design parameters in the early stages of development. However, the numerical methods require a considerable amount of time and effort to predict the bit error rate (BER), and it is nearly impossible to explore the whole design space for optimizing product yield in time. In addition, we cannot perform as many experiments as desired in a process development step because only a limited amount of experimental data can be obtained since the V-NAND processes are time-consuming and expensive.

Therefore, an efficient and accurate BER estimation technique is essential to evaluate the various design options and optimize the yield in a timely manner. However, prior studies investigating BER estimation in the GIDL-assisted erase scheme have been very limited. In [8] and [9], an analytical compact model is presented to describe the time dynamics of the GIDL-assisted erase operation in the 3D V-NAND structures. This analytical model is suitable for reproducing the GIDL-assisted transient analysis results from TCAD simulations. However, this approach does not support a scalable model for widely different channel sizes, bias voltages, and temperatures. This lack of scalability makes it difficult to perform an evaluation with various technological options accurately in circuit simulators. In addition, considering process variation is essential for the bit error rate prediction, and this approach does not provide the variability models for GIDL characteristics. In [10] and [11], the authors proposed a machine learning approach that reproduces the variations of threshold voltage (V_{th}) and current (I_{on}) in the 3D V-NAND cells. The model is based on an artificial neural network (ANN) whose inputs are variability sources and electrical parameters. However, this method focuses on predicting the variations of V_{th} and I_{on} for the wear-out in pre-production steps that can achieve the same accuracy of TCAD simulations. However, this model cannot reproduce I-V characteristics that can be implemented in SPICE simulators for accurate bit error prediction. In [12], [13], [14], and [15], an analytical fitting method, ANN model, and support vector regression are proposed to predict the bit error rate as a function of program/erase (P/E) cycle, read cycle, retention time, and the total ionizing dose effects. However, these predictive models are entirely based on the data measured from an experiment with various P/E cycle and retention time. They aim to improve the error correction codes and the wear equalization algorithms regardless of time and cost. Therefore, these approaches cannot be applied to accelerate and evaluate the read margin loss of the GIDL-assisted erase operations. In [16], the authors present a parameter estimation algorithm to find the means and variances of the threshold voltage distribution that is modeled as a Gaussian mixture. However, this approach has limits on predicting the read margin of GIDL-assisted erase, because the distributions are approximated to the Gaussian mixture and the errors from non-Gaussian characteristics are relatively large. To overcome this limitation, the work proposed in [17] to choose from the various distributions as well as Gaussian mixture are evaluated to find the accurate predictive model

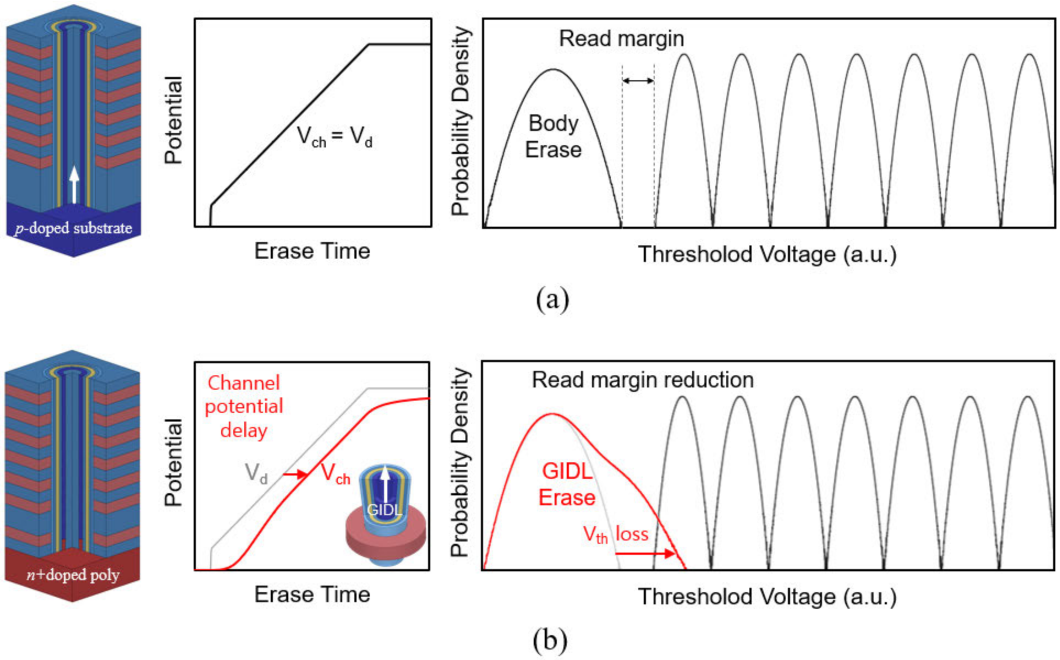


FIGURE 2. Threshold voltage distributions of the body erase and GIDL-assisted erase scheme. (a) The body erase scheme operates bulk erase, which increases the channel potential using the direct contact with p-doped substrates and has excellent erase performance. (b) The GIDL erase scheme of COP is built on n+doped poly-silicon and increases the channel potential using GIDL current, which has an inherent channel potential delay and slow cells (V_{th} loss and small read margin).

of the threshold voltage distribution. However, the evaluation results show that the average accuracy is 95%, and it also causes a large error in the bit error rate prediction at the very low probability density. Because of the lack of effective yield estimation methods in the GIDL-assisted erase scheme, we propose a novel simulation acceleration technique for bit error rate prediction and optimization using an ANN-based compact model and Monte Carlo circuit simulations.

The rest of this paper is organized as follows. Section II describes the mechanisms of channel potential delay and read margin reduction in the GIDL-assisted erase scheme, section III discusses the methodology of the acceleration models with their scalability and accuracy, and also describes a Monte Carlo (MC) simulation technique, and section IV highlights the bit error rate prediction and yield optimization for next-generation candidate structures. The last section concludes by outlining the efficient yield prediction framework to deal with the challenges of the extreme high stack flash memories.

II. INVESTIGATION ON READ MARGIN REDUCTION

Fig. 3(a) shows the structure of the 3D V-NAND flash memory addressed in this work. It includes repeated vertical cell arrays and a transistor to generate hole carriers by band-to-band-tunneling (BTBT). The cell array has a cylindrical poly-silicon channel with an oxide layer filling the channel cavity. The charge trap layer (CTL) plays the role of storage node for the memory cells and the $SiO_2/SiN_x/SiO_2$ (O/N/O) gate stack runs all along the channel string.

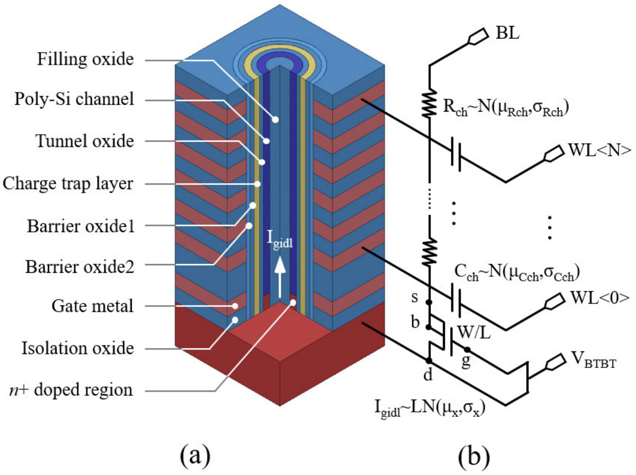


FIGURE 3. (a) The structure of the 3D V-NAND for GIDL-assisted erase in this work. (b) The schematic figure of an equivalent circuit for GIDL-assisted erase operations. The parameters are summarized in Table 1.

A. MECHANISMS OF CHANNEL POTENTIAL DELAY

To understand the mechanism of the hole carrier accumulation process, we performed a TCAD simulation by solving the Poisson and drift-diffusion equations with the dynamic nonlocal BTBT model in a channel string structure. Fig. 4 shows the simulated channel potential (V_{ch}), external biases (V_d and V_g), internal potential differences (V_{gs}, V_{ds}) in the floating body, and hole current injected into the channel

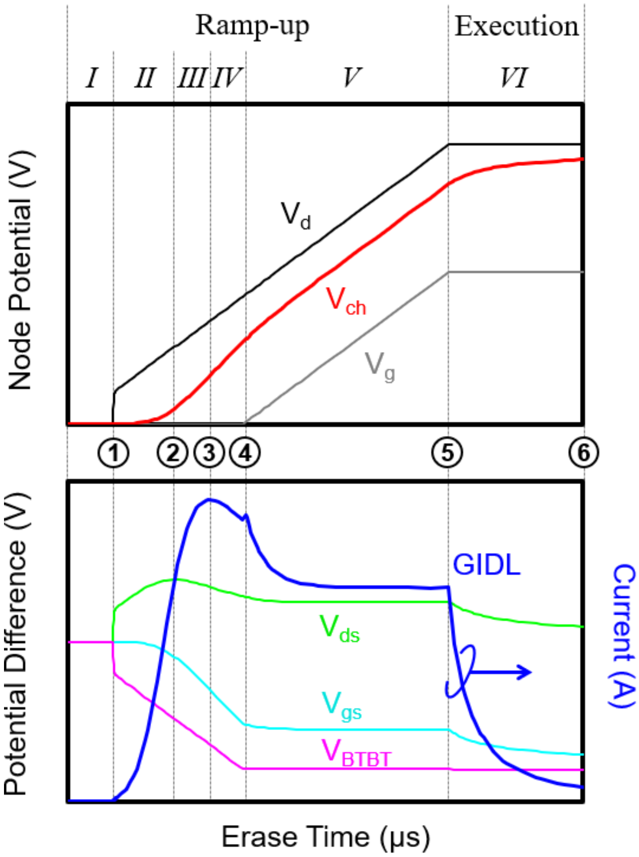


FIGURE 4. The simulation results for the transient behaviors of the GIDL-assisted erase operation.

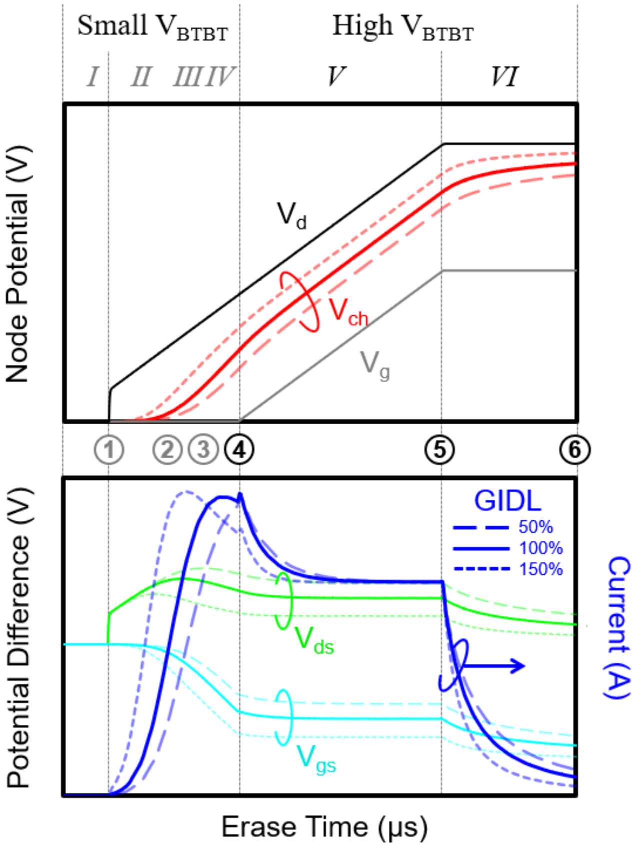


FIGURE 5. The simulation results for the transient behaviors of different GIDL transistor performances.

string. V_{BTBT} is the supply voltage (V_g-V_d), which determines the BTBT electron-hole pair generation rate, and V_{ds} is the internal bias, which affects the hole injection rate into the channel.

The GIDL-assisted erase scheme has two operation regimes, (ramp-up and execution), and they have distinct phases (*I-VI*) and singular points (1)-(6). *Phase I* is the quasi-static region before the start of pulse-type ramp-up at point 1 to prevent the slow BTBT response. *Phase II* is the acceleration region, which has the biggest increase rate of V_{ch} before the V_{ds} reaches a peak at point 2. During *phase III*, the BTBT rate faces its maximum limit with the deceasing V_{ds} due to the stored hole carriers in the floating channel, and the hole current has its peak at point 3. During *phase IV*, these exponential behaviors of V_{ch} increase ($\Delta V_{ch} > \Delta V_d$) come to the finish with the start of the constant V_{BTBT} at point 4, and the linear behavior of V_{ch} increase ($\Delta V_{ch} = \Delta V_d$) starts, and the hole current is saturated during *phase V* in the ramp-up regime. During *phase VI* (also in the execution regime), it works in a negative feedback configuration due to the accumulated carriers in the floating channel region. Therefore, the erase execution always finishes without reaching the body erase potential (V_d), and this channel potential delay makes the GIDL-assisted scheme vulnerable to erase performance variations.

Fig. 5 shows the simulation results with three different GIDL transistors. In this simulation, we used the same external voltages and cell array structure, except for the BTBT rates of the GIDL transistors. The transient results show that the injection levels of the GIDL currents are significantly different in only small V_{BTBT} phases (II, III, and IV) due to the increasing V_{ds} and V_{BTBT}. However, the small V_{BTBT} phases are relatively short, and they are less than 20 percent of the total erase operation. In contrast, the injection GIDL currents start to saturate after their peak levels, and they finally converge to similar levels in the high V_{BTBT} phases (V and VI) due to decreasing V_{ds} and the negative feedback configuration. These high V_{BTBT} phases take most of the erase operation time and V_{ds} is bound to stay low for the entire period. This shows that high V_{BTBT} and low V_{ds} are the general operating conditions in the GIDL current trajectories, and the industry normally extract the characteristic current (I_{gidl}) from these conditions to represent the delay of channel potential instead of using the entire GIDL current values. Therefore, we define the characteristic current (I_{gidl}) as a condition of $V_{gs} = -8$V and V_{ds}=3V in this work.

B. MECHANISMS OF READ MARGIN REDUCTION

In the GIDL-assisted erase, the V_{th} distribution can be separated into two independent components, as shown in Fig. 6(a).

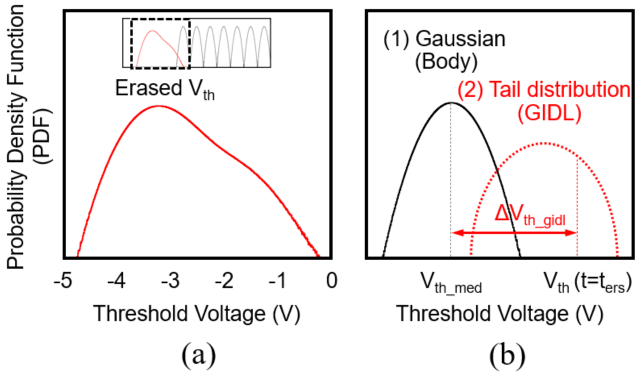


FIGURE 6. (a) The V_{th} distribution of the GIDL-assisted erase, (b) The two independent variation components in the V_{th} distribution, and the V_{th} shift of slow cells (variance of V_{th_gidl}).

One is (1) Gaussian distribution (V_{th_med}, V_{th_std}), which comes from the random variations of tunneling oxide thickness, gate critical dimension, trap site, and Fowler Nordheim (FN) tunneling sensitivities, and this component is identical to the distribution of a body erase scheme. The other is (2) tail distribution, which originates from the inherent channel potential delay of GIDL generation process and its impact on the slow cells. Therefore, the read margin reduction and threshold voltage loss of slow cells (ΔV_{th_gidl} in Fig. 6(b)) caused by the potential delay can be defined as

$$\Delta V_{th\_gidl} = V_{th}(t = t_{ers}) - V_{th\_med} \tag{1}$$

To investigate the relationship between the channel potential delay and the V_{th} loss of slow cells, we performed TCAD simulations in a 176-layer V-NAND structure. In this simulation step, we ignored the random variations of Gaussian distribution (V_{th_std}=0) to reproduce the tail distribution only, and the programmed cells are erased with different channel potential delays using random variables of the BTBT rate, channel string resistance, and capacitance which are significantly related to the delays. Fig. 7(a) shows the simulation results based on the random variables, and we can define the channel potential delays (ΔV_{FN}) on the reference of the body erase scheme, which is primarily responsible for the tail shift of slow cells, as follows,

$$\Delta V_{FN}(t) = V_d(t) - V_{ch}(t) \tag{2}$$

During the erase operation, the carriers are gradually removed from the floating gates due to FN tunneling, and the tunneling effects finish at the erase execution time (t_{ers}). In addition, the amount of V_{th} shift can be determined by the number of carriers remaining in the floating gates at t_{ers}, and it is significantly related to the shortages of applied electric field (ΔE_{FN}) for the FN tunneling as

$$\Delta E_{FN}(t) = \frac{\Delta V_{FN}(t)}{t_{ONO}} \tag{3}$$

These field quantities (ΔE_{FN}(t)) during the erase operation are shown in Fig. 7(b). Therefore, the cumulative factor of channel potential delay and read margin loss from the slow cells is the areas under the curves, and it is the integral ΔE_{FN}(t) from t_{FN} to t_{ers} as follows,

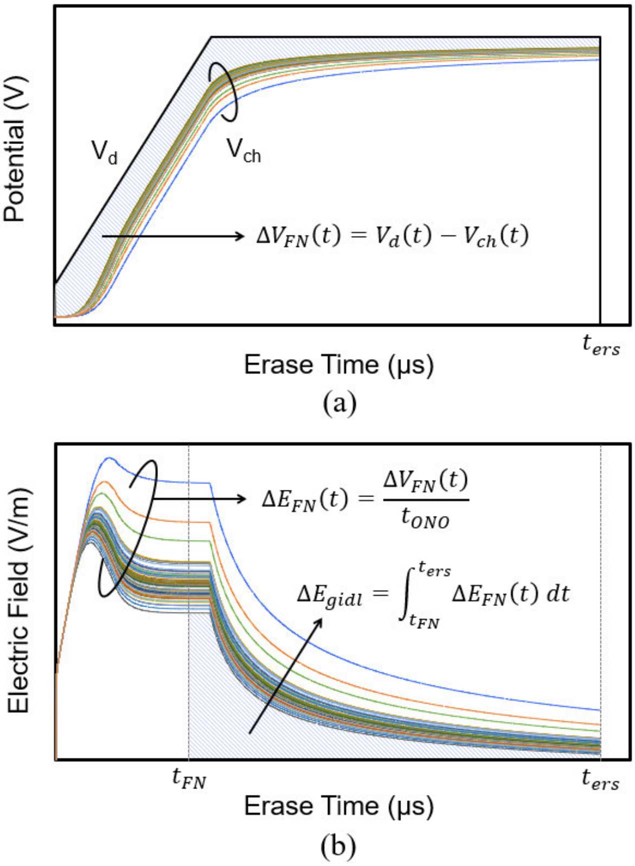


FIGURE 7. (a) The variation of channel potential delay on the reference potential (V_d) of the body erase scheme, (b) The cumulative field factor.

$$\Delta E_{gidl} = \int_{t_{FN}}^{t_{ers}} \Delta E_{FN}(t)\mathrm{d}t \tag{4}$$

where t_{FN} is a model parameter depending on the ramp-up slops and FN tunneling parameters which vary throughout the technologies and processes.

III. FAST SIMULATION MODELS FOR BIT ERROR RATE PREDICTION AND YIELD OPTIMIZATION

A. SLOW CELL MODEL

The bit error rate prediction of slow cells and yield optimization can be accelerated based on the computational efficient and accurate circuit simulation based on the physics-inspired slow cell model and GIDL variability compact model. Using large-scale Monte-Carlo (MC) circuit simulations with these computationally efficient models, the analysis on the V_{th} shift of slow cells and yield loss in the GIDL-assisted erase can be easily performed by the accelerated data.

To obtain the cumulative field factor (ΔE_{gidl} in Eq. (4)) in the current 176-layer V-NAND technology, we used t_{FN}=200μs and t_{ers}=1400μs, and the result of the calculation

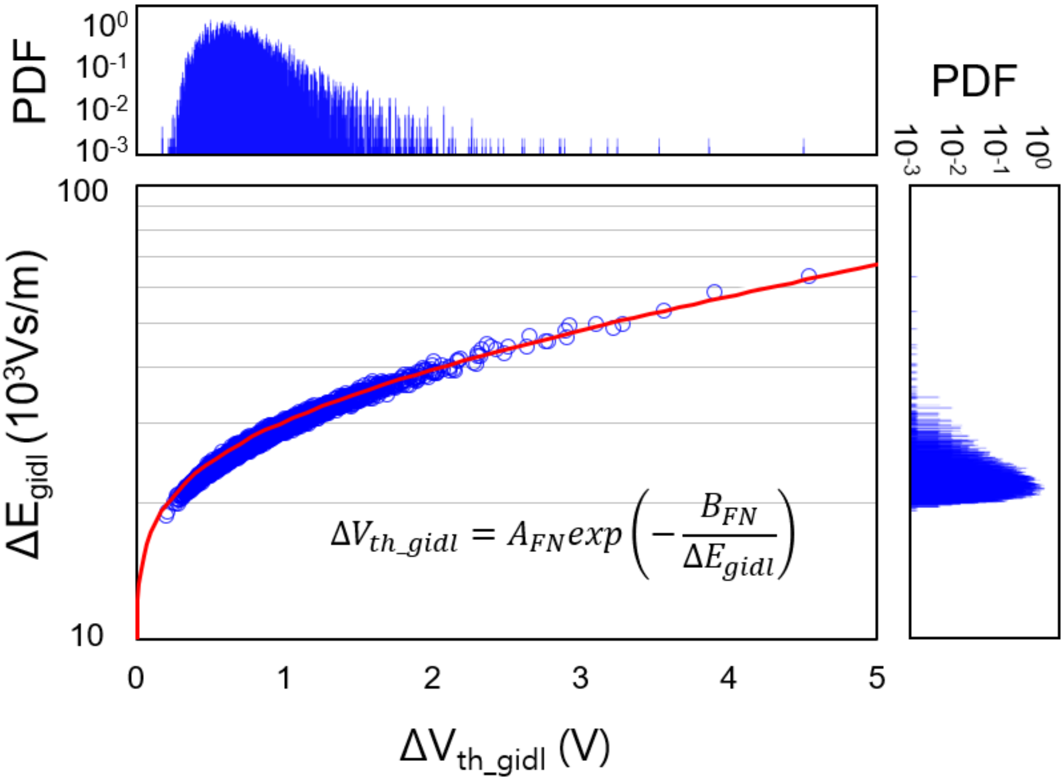


FIGURE 8. The probability density functions (PDF) of the two random variables of the V_{th} shift of slow cells and the cumulative field factor of the channel potential delay in the GIDL-assisted erase scheme. The distributions that trap high correlation have a relationship represented by a slow cell model.

is plotted in Fig. 8 with the V_{th} shift of slow cells (ΔV_{th_gidl} in Eq. (1)). It shows that the two random variables are highly correlated, therefore, the slow cell model can be described as follows,

$$\Delta V_{th\_gidl} = A_{FN} \exp\left(-\frac{B_{FN}}{\Delta E_{gidl}}\right) \tag{5}$$

where A_{FN} is a threshold voltage coefficient parameter and B_{FN} is a FN tunneling exponential coefficient of the slow cell which depend on the processes. In this work, these fitting parameters are extracted with 1.85×10^1 and 8.9×10^4 for A_{FN} and B_{FN}, respectively, and the slow cell model has good agreement with the numerical simulation data, as shown in Fig. 8.

B. GIDL VARIABILITY MODEL

Firstly, TCAD model and parameters were calibrated to 176 layer V-NAND technology based on the assumed values such as geometries of a GIDL transistor (W, L), temperature (T), bias conditions for GIDL generation (Vg, Vd), GIDL current characteristics (I_{gidl}, δI_{gidl}), and RC delay network and its variations (R_{ch}, C_{ch}, δR_{ch}, and δC_{ch}). For the GIDL-assisted erase operations in this paper, GIDL transistors with channel width (W) and length (L) of 201nm and 24nm, respectively, are used under external bias conditions with gate voltage (V_g) of 0-10V and drain voltage (V_d) of 0-18V. In this condition, the GIDL current is 0.9nA, and it is set up to drive a 176 layer delay network of poly-silicon with a resistance of $2.4 \times 10^{-6}\Omega$ and a capacitance of 3.2×10^{-17}F per layer. Table 1 summarizes these calibration parameters of TCAD and circuit simulations.

The accelerated circuit simulation requires an accurate compact GIDL transistor model, which can reproduce the logarithmic characteristics of the hole injection into a macaroni-shaped floating channel and its variability of log-normal distribution. Based on the measured I-V data, we take the natural logarithm of the corresponding current values to generate a normal distribution, which has a mean (μ_Y), standard deviation (σ_Y), and its coefficient of variation (CV). Then, we arrange the log-transformed data into the ANN-based compact model [18], [19], [20] for feeding, as shown in Fig. 9(a). This ANN-based compact model has two hidden layers with 20 and 15 neurons in each layer (N_1=20, N_2=15). The input features are channel width (W), length (L), temperature (T), and bias voltages (V_g,V_d,V_s, and V_b), and the targets are the log-transformed current values ($\ln(I_d)$, $\ln(I_s)$, and $\ln(I_b)$), as shown in fig. 9(b). In the Verilog-A implementation step for the SPICE circuit simulations, we multiply the inference results of the ANN-based compact model by a multiplication factor, which follows a normal distribution of $N(1,\sigma_Y)$, and then take the exponential of them to change the distribution back into a lognormal, as shown in

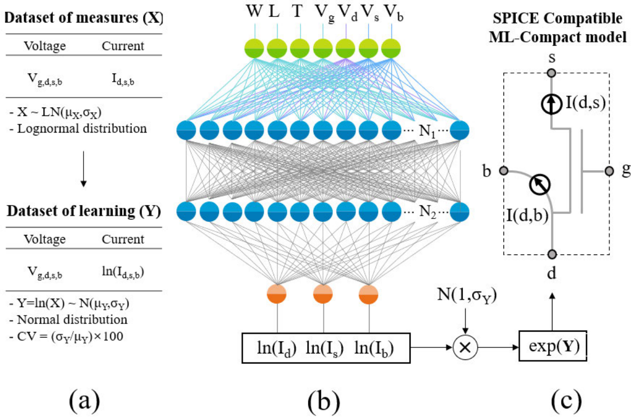


FIGURE 9. The flowchart for lognormal variability modeling of the GIDL transistor using machine learning. (a) Generation of the dataset for machine learning model feeding, (b) a fully connected artificial neural network-based compact model, and (c) the Verilog-A implementation for the computationally efficient circuit simulation.

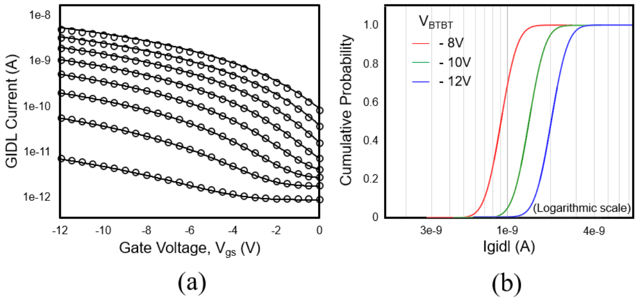


FIGURE 10. The inference results of the artificial neural network-based compact model. (a) I-V curve inferences (lines) versus TCAD targets (symbols), and (b) Inferences of the GIDL characteristic currents following the lognormal distributions in the cumulative probability plot at different V_{BTBT} values.

Fig. 9(c). After 8,000 epoch training cycles, high accuracy of 99.3% (1-mean absolute percentage error, 1-MAPE) can be achieved, as shown in Fig. 10(a). In an MC simulation, the inference results of the characteristic current, I_{gidl} follow the lognormal distributions in the cumulative probability plot for three different V_{BTBT} values, as shown in Fig. 10(b).

C. SCALABILITY OF THE PROPOSED MODELS

The high generality of an ANN-based compact model enables high scalability in modeling complex characteristics of tunneling physics. Therefore, we perform additional simulations at different transistor geometries and temperatures to evaluate the model scalability and accuracy. In the TCAD simulations, the GIDL current changes nonlinearly with the channel width because a stronger electric field is applied and a higher BTBT current density is generated at the channel edge region as the channel width decreases. The ANN-based compact model is trained and verified with the nonlinear simulation data and shows high average accuracy of 99.3% as shown in Fig. 11. The GIDL current also has a positive temperature dependence because the bandgap narrows as temperature increases, resulting in an increasing BTBT rate. In this process, the BTBT rate is varied by a factor of 10.8 from 248K (cool temperature, CT) to 358K (hot temperature, HT). The ANN-based compact model is also jointly trained with the temperature dependent data and shows high average accuracy of 99.2% as shown in Fig. 12.

In addition, the slow cell model (read margin reduction model) is implemented in the SPICE simulator. The model verification results of the different GIDL transistor widths show good agreement (average R2 score is 0.991) with TCAD simulation data as shown in Fig. 13. For the temperature variation, the effects of an increase in GIDL current (I_{gidl}) and a decrease in channel potential delay (V_{FN}) according to increasing temperature needs to be considered in a GIDL-assisted erase scheme. The operating temperature can be different from the nominal temperature (TNOM) at which the slow cell model parameters are extracted. In this work, the slow cell model parameters (t_{FN} A_{FN}, and B_{FN}) are extracted and verified at the nominal room temperature (RT), TNOM=298K. Therefore, the slow cell model can account for the effects of temperature by making parameter B_{FN} temperature dependent, as follows:

$$\Delta V_{th\_gidl} = A_{FN} \exp\left(-\frac{B_{FN}(T)}{\Delta E_{gidl}}\right) \tag{6}$$

where

$$B_{FN}(T) = B_{FN0}\left(\frac{T}{TNOM}\right)^{FNT} \tag{7}$$

FNT is introduced as a FN tunneling temperature exponent for the exponential coefficient (B_{FN}) which is an image of the potential barrier that corresponds to the effective band gap of the barrier materials. The temperature dependence model shows good agreement (average R2 score is 0.992) with TCAD simulation data from 248K (CT) to 358K (HT) as shown in Fig. 14. Table 2 summarizes these fitting parameters of the slow cell model for bit error rate prediction and yield optimization.

IV. ACCELERATION SIMULATION FOR BIT ERROR RATE PREDICTION AND YIELD OPTIMIZATION

Based on the ANN-based compact model and the slow cell model, the V_{th} shift evaluation and BER estimation can be accelerated in the efficient circuit simulation domain. The flow of this acceleration work is demonstrated in Fig. 15(a): (1) the Gaussian distribution of V_{th} is generated using the $agauss(V_{th\_med}, V_{th\_std})$ function of the SPICE simulators, (2) the random variables of the V_{th} shift (ΔV_{th_gidl}) are obtained based on the large-scale analysis data from the acceleration simulation using the instantaneous calculation of Eq. (5). Finally, we can obtain the entire V_{th} distribution of the GIDL-assisted erase operation by calculating the sum of the two independent random variables and also evaluate the tail bit of the slow cells using a metric, V_{th} loss in Fig. 15(b).

We performed the large-scale MC circuit simulations based on the present generation of a 176-layer V-NAND technology

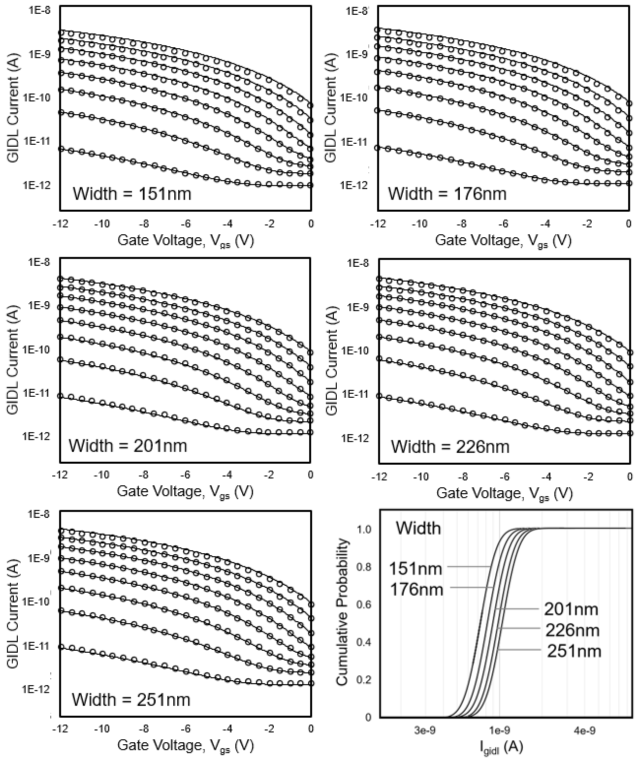


FIGURE 11. The inference results (lines) of the artificial neural network-based compact model compared to TCAD targets (symbols) and GIDL characteristic currents following the lognormal distributions in the cumulative probability plot at different transistor widths.

as shown in Table 1. The circuit simulation includes the fast models which is trained and calibrated with the same TCAD simulations of the 176-layer V-NAND supplying -8V BTBT voltage ($V_{BTBT} = -8$V). Fig. 16 shows the comparison results of the calibrated TCAD simulation (ground truth) and the acceleration simulation. The fast models show good agreement with the ground truth depending on the various medians of the GIDL currents. Fig. 17 represents the comparison results of the ground truth and the proposed models depending on the variation levels of the GIDL current, which have good agreement. The temperature dependent models in Eq. (6) and (7) are also implemented and performed in the accelerated circuit simulation. Fig. 18 shows good agreement between the ground truth and the temperature dependent models.

To explore the design space and evaluate the impact of the GIDL transistor performances on the V_{th} loss and yield, we can perform a large-scale simulation based on the accelerating models. Fig. 19(a) shows the acceleration simulation results of V_{th} loss according to the GIDL transistor performances of $V_{BTBT} = -8$V in the 176-layer structure. The optimal design target of the GIDL transistor performances (I_{gidl}, CV of I_{gidl}, and V_{BTBT}) is the value set when the V_{th} loss becomes small enough to distinguish the erase from the programmed states. It generally depends on the valid window and bit error rate criteria of various technologies. In this work, we define the V_{th} loss of 0.5V or less ($V_{th\_loss} \leq 0.5$) at the probability density of 1×10^{-3} as an optimal boundary to achieve the successful read operation in this demonstration. A small V_{th} loss indicates larger read voltage window, higher error correction probability, and a smaller number of read

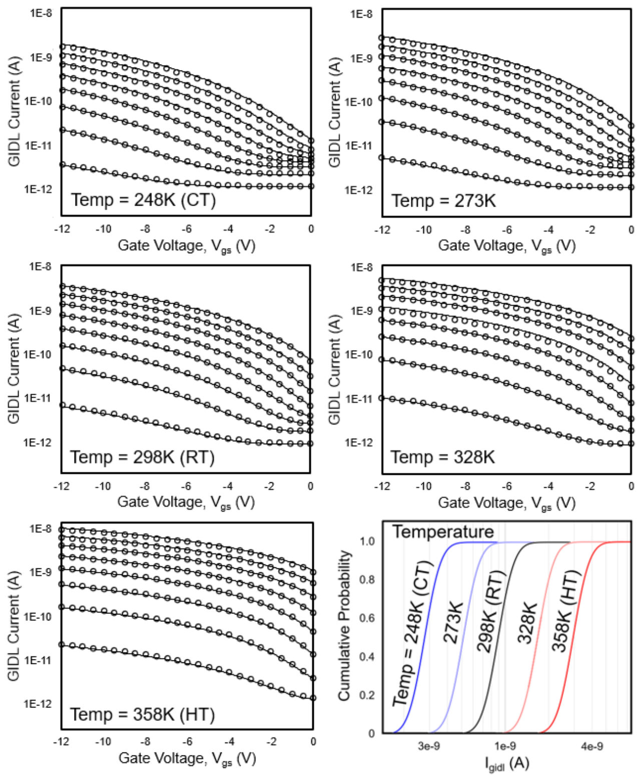


FIGURE 12. The inference results (lines) of the artificial neural network-based compact model compared to TCAD targets (symbols) and GIDL characteristic currents following the lognormal distributions in the cumulative probability plot at different temperatures.

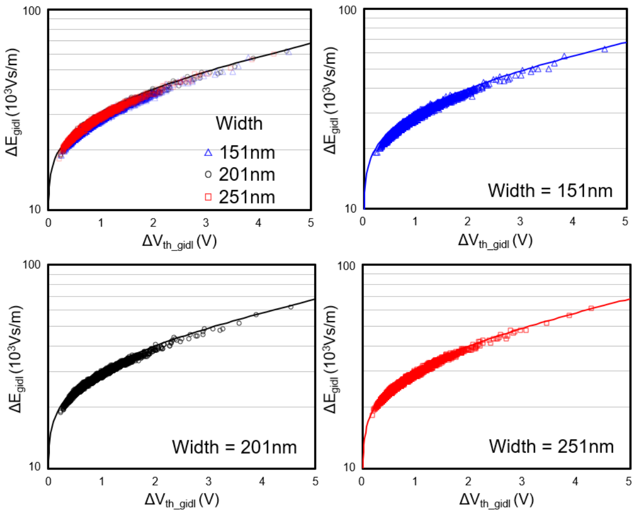


FIGURE 13. The slow cell model (line) and TCAD targets (symbols) for the two probability densities of V_{th} shift and the cumulative field factor of the channel potential delay in the GIDL-assisted erase scheme at different transistor widths.

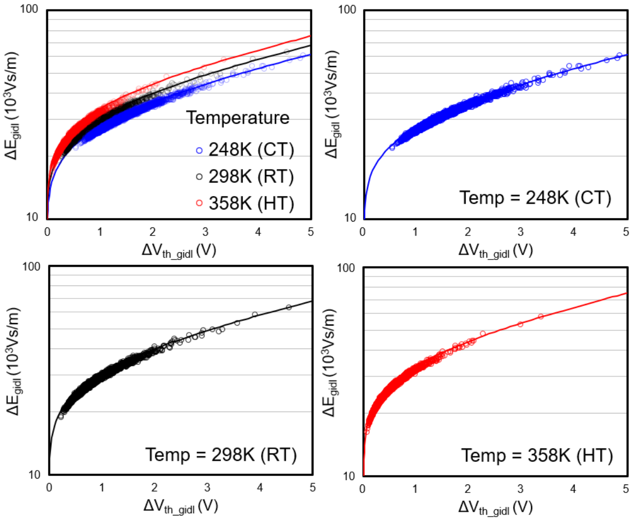


FIGURE 14. The slow cell models (lines) and TCAD targets (symbols) for the two probability densities of V_{th} shift and the cumulative field factor of the channel potential delay in the GIDL-assisted erase scheme at different temperatures.

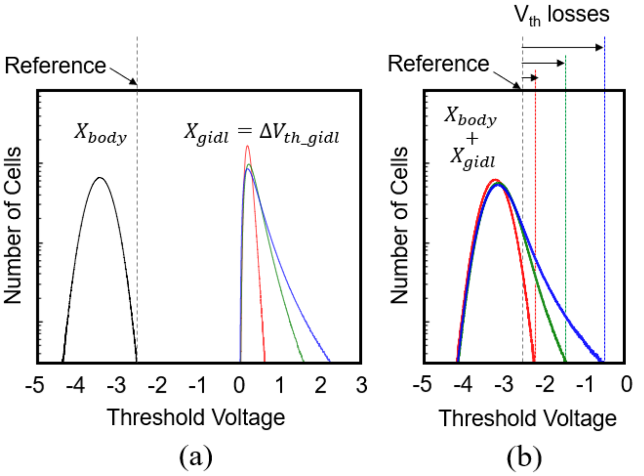


FIGURE 15. The flow of acceleration simulation for generating the V_{th} distribution of the GIDL-assisted erase. (a) Independent generations of two independent V_{th} distributions, (b) the sum of two distributions and a new metric, V_{th} loss for the tail bit evaluation.

TABLE 1. The calibration parameters for TCAD and Monte-Carlo circuit simulations.

Name	Description	Value
W	Channel width of GIDL transistor	201nm
L	Channel length of GIDL transistor	24nm
T	Temperature	298K
V_g	External gate bias for GIDL generation	0-10V
V_d	External drain bias for GIDL generation	0-18V
V_s, V_d	Floating channel potentials	(calculated)
I_{gidl}	Median of GIDL current, μ_Y	0.9nA
R_{ch}	Median of channel resistance per WL, μ_R	$2.4 \times 10^{-6}\Omega$
C_{ch}	Median of channel capacitance per WL, μ_C	3.2×10^{-17}F
δI_{gidl}	CV of GIDL current, σ_Y/μ_Y	23%
δR_{ch}	CV of channel resistance per WL, σ_R/μ_R	5%
δC_{ch}	CV of channel capacitance per WL, σ_C/μ_C	5%

I_{gidl} is the characteristic current of GIDL transistor when V_{gs}=-8V and V_{ds}=3V.
CV is coefficient of variation percentage when the random variable follows normal distribution.

TABLE 2. The fitting parameters of the slow cell model for bit error rate prediction and yield optimization.

Name	Description	Value
t_{FN}	Ramp-up slop dependent parameter	200μ s
A_{FN}	Threshold voltage coefficient parameter	1.85×10^1
B_{FN0}	Exponential coefficient at T=TNOM	8.9×10^4
TNOM	Nominal temperature	298K
FNT	FN tunneling temperature exponent	0.55

TNOM is the temperature at which parameters are extracted. In this work, the model parameters are extracted at room temperature (T=298K).

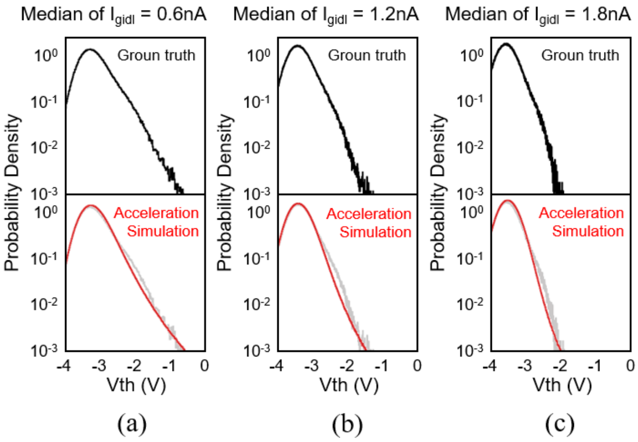


FIGURE 16. The comparison of TCAD simulation (top) and accelerating circuit simulation results (bottom) as a function of GIDL transistor performances. (a) I_{gidl}=0.6nA with CV=26%, T=298K, (b) I_{gidl}=1.2nA with CV=26%, T=298K, and (c) I_{gidl}=1.8nA with CV=26%, T=298K.

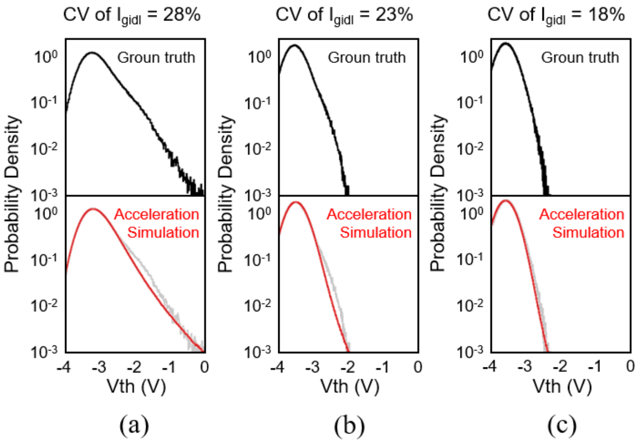


FIGURE 17. The comparison of TCAD simulation (top) and accelerating circuit simulation results (bottom) as a function of GIDL current variations. (a) CV of I_{gidl}=28% with I_{gidl}=0.9nA, T=298K, (b) CV of I_{gidl}=23% with I_{gidl}=0.9nA, T=298K, and (c) CV of I_{gidl}=18% with I_{gidl}=0.9nA, T=298K.

retry. In addition, the yield characteristics is determined by bit error rate, which strongly depends on the error correction abilities and read retry techniques of each industry. To demonstrate the flow of yield prediction and optimization in this 176-layer V-NAND technology, we assume that the optimal boundary criterion of valid window ($V_{th\_loss} \leq 0.5$) is equivalent to a bit error rate of 0.1%, and it increases by a factor of 3 for every 0.5V increase in V_{th} loss. Therefore, a contour map of bit error rate for yield analysis can be derived from the V_{th} loss as shown in Fig. 19(b).

Based on the acceleration simulations combined with the bit error occurrences, we can obtain a metric for predictive

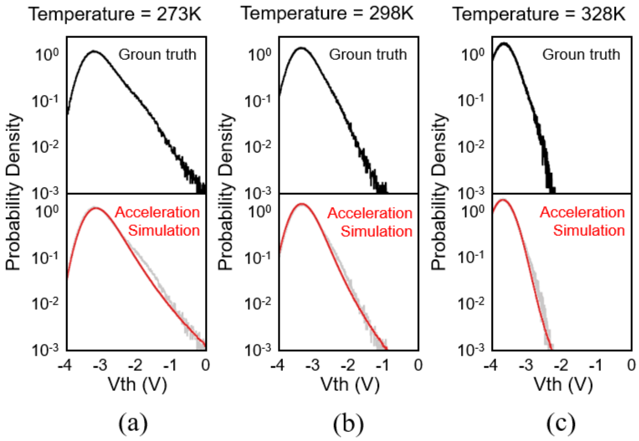


FIGURE 18. The comparison of TCAD simulation (top) and accelerating circuit simulation results (bottom) at different temperatures. (a) T=273K with CV=26%, (b) T=298K with CV=26%, and (c) T=298K with CV=26%.

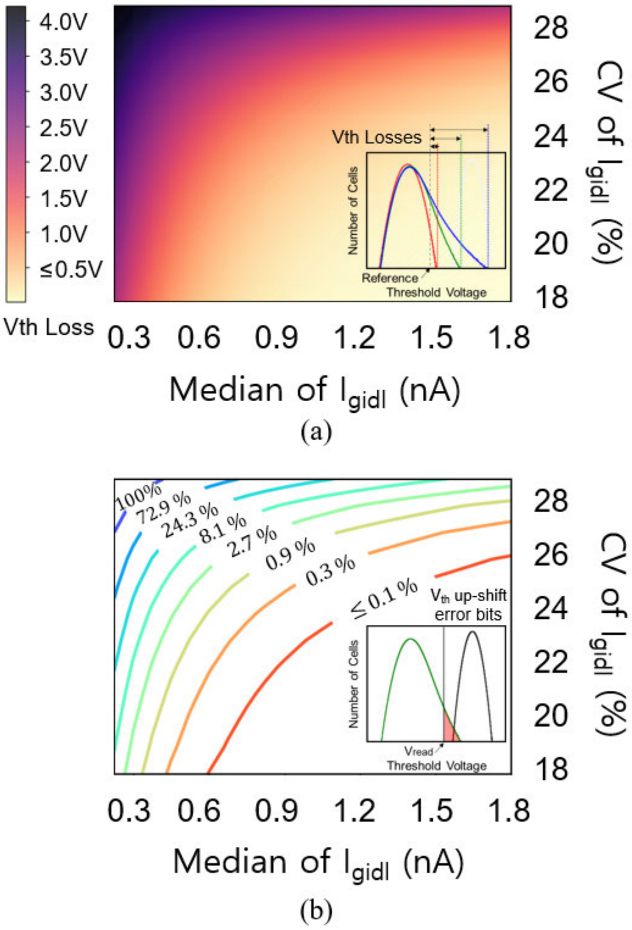


FIGURE 19. (a) The acceleration simulation results of V_{th} loss in the 176-layer V-NAND technology. (b) The contour map of yield based on the bit error rate in the 176-layer V-NAND technology. To describe the flow of yield optimization in this technology, we assume the V_{th} loss of 0.5V is equivalent to 0.1% bit error rate and the rate increases exponentially for every 0.5V increase in the V_{th} loss.

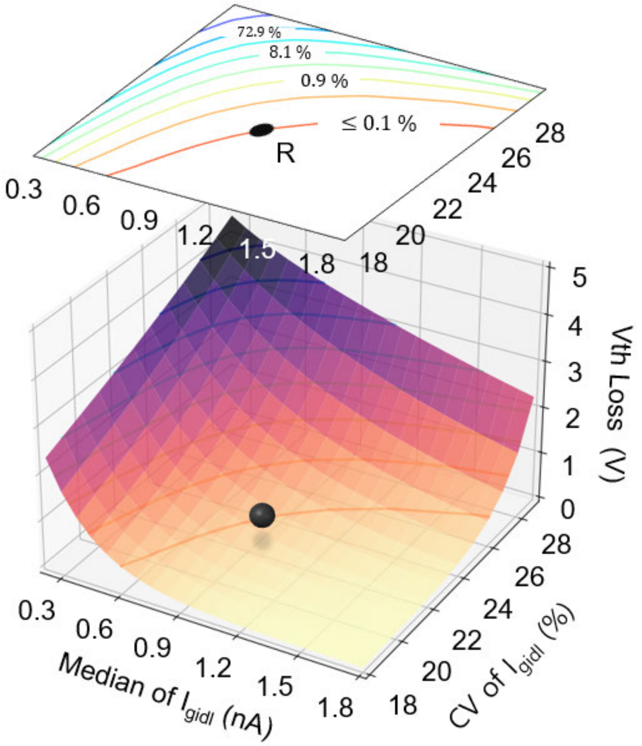


FIGURE 20. The contour map of yield based on the bit error rate and its projection to the 3D plot of V_{th} loss in $V_{BTBT} = -8$V according to the characteristic current I_{gidl} and the CV of I_{gidl} in the 176-layer V-NAND technology.

yield modeling in the 176-layer V-NAND structure, as shown in Fig. 20. Starting with the reference point of this current GIDL transistor (R in Fig. 20), any performance shifts of the median I_{gidl} or its CV immediately will reduce the read margin (V_{th_loss}>0.5V) and degrade the bit error rate (BER>0.1%). However, we can boost the read margins by supplying the higher BTBT voltage from −8V to −10V to drive the same GIDL transistor, as shown in Fig. 21, which must be carefully examined from various side effects on the degradation of reliability and program-erase cycles.

V. PATHFINDING FOR NEXT GENERATION 3D V-NAND CANDIDATE STRUCTURES

Based on the calibrations in the current 176-layer V-NAND technology, the bit error rate prediction and yield optimization of the next-generation candidate technologies can be performed for the pathfinding activity. We extend the acceleration simulations to the next-generation 256 and 352-layer V-NAND technologies to determine the impact of the high aspect ratio etch processes of vertical stack-up on the GIDL-assisted erase characteristics.

A. 256-LAYER 3D V-NAND STRUCTURE

Fig. 22(a) shows the contour map of yield and its projection ot the 3D plot of V_{th} loss in the 256-layer V-NAND structure with the same −8V BTBT voltage (V_{BTBT}= −8V) as the previous 176-layer V-NAND. The prediction results show that the V_{th} loss will be degraded by up to 1.2V due to the extra RC delay (R in Fig. 22(a)), and three paths of yield improvement are possible. Firstly, nearly 1.3nA GIDL

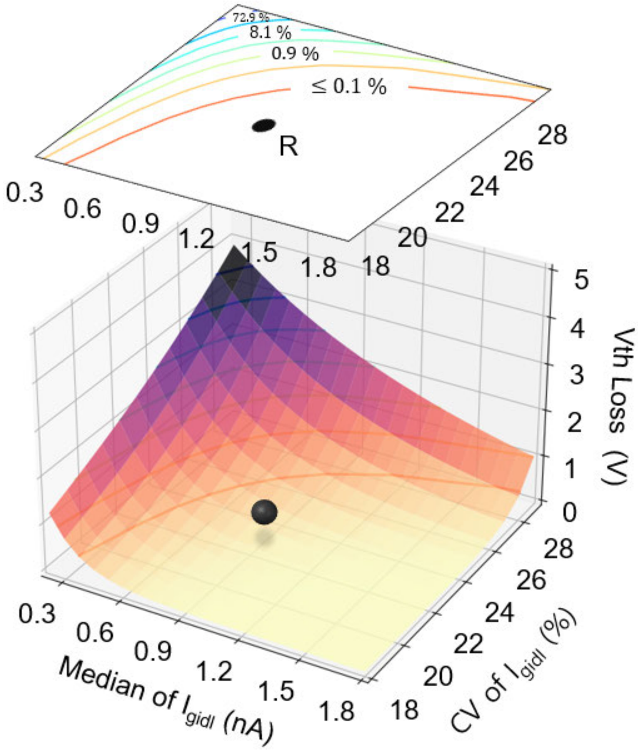
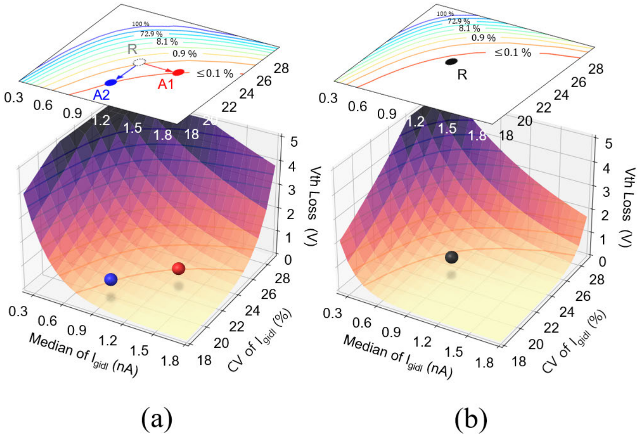


FIGURE 21. The contour map of yield based on the bit error rate and its projection to the 3D plot of V_{th} loss in $V_{BTBT} = -10$V according to the characteristic current I_{gidl} and the CV of I_{gidl} in the 176-layer V-NAND technology.

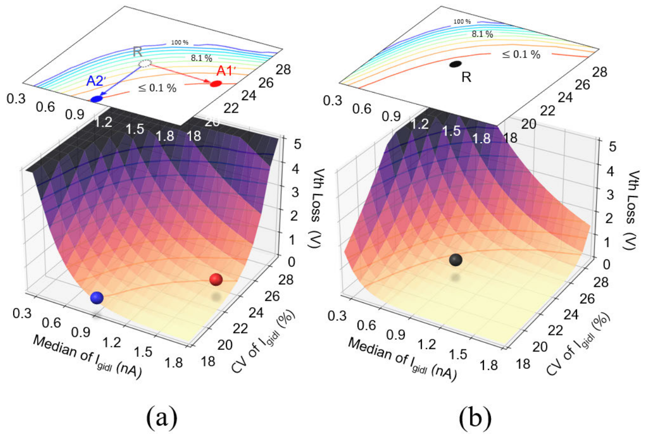


(a)

(b)

FIGURE 22. The contour map of yield based on the bit error rate and its projection to the 3D plot of V_{th} loss in (a) $V_{BTBT} = -8$V, (b) $V_{BTBT} = -10$V according to the characteristic current I_{gidl} and the CV of I_{gidl} in the 256-layer V-NAND technology.

injection current is required without any variation reduction techniques (A1 in Fig. 22(a)). Secondly, the CV of the GIDL currents needs to be reduced up to 20% while the median remains the same (A2 in Fig. 22(a)). Lastly, only 1.1nA injection current will be sufficient if the CV can be reduced to 22% as a compromise. In addition, we can supply the higher BTBT voltages in the identical GIDL transistor to avoid any process challenges. In Fig. 22(b), the same GIDL transistor of



(a)

(b)

FIGURE 23. The contour map of yield based on the bit error rate and its projection to the 3D plot of V_{th} loss in (a) $V_{BTBT} = -8$V, (b) $V_{BTBT} = -12$V according to the characteristic current I_{gidl} and the CV of I_{gidl} in the 352-layer V-NAND technology.

the previous technology is successfully reused to operate the erase in the 256-layer V-NAND structure, and it represents the minimum BTBT voltage of -10V required for ensuring the yield criteria (V_{th_loss}≤0.5 and BER≤0.1%).

B. 352-LAYER 3D V-NAND STRUCTURE

Fig.23(a) shows the prediction results for the 352-layer V-NAND structure with the same -8V BTBT voltage ($V_{BTBT} = -8$V) as the current 176-layer simulation. The results indicate that V_{th} loss is increased to 2.2V (R in Fig. 23(a)) due to the extra delays of additional layers. In the 352-layer V-NAND, there are also three possible paths for yield improvement. Firstly, the GIDL injection current is required to increase to 1.7nA without any variation reductions (A1′ in Fig. 23(a)). Secondly, the CV needs to be improved up to 18% (A2′ in Fig. 23(a)). Lastly, both the 1.3nA GIDL injection current and the 21% CV can be one of the optimal design targets as a compromise solution. To retain the same GIDL transistor process of the current 172-layer V-NAND technology, a minimum of -12V BTBT voltage is required to obtain the same read margin and bit error rate in the next-generation 352-layer V-NAND, as shown in Fig. 23(b).

VI. CONCLUSION

In this paper, we proposed a variability-aware artificial neural network compact model and a fast Monte Carlo circuit simulation technique that accelerate the bit error rate estimation and yield optimization of the GIDL-assisted erase scheme in state-of-the-art flash memories. The GIDL-induced channel potential delay and time dynamics are thoroughly investigated, highlighting the read margin reduction mechanism due to the V_{th} loss of slow cells in the GIDL erase operation. The ANN-based compact model accurately reproduces the GIDL current and its lognormal statistical variations, and the physics-inspired slow cell model is efficiently implemented in circuit simulation to regenerate the underlying relationships between the two probability distributions of electric field and read margin loss. This acceleration simulation technique is a valuable tool for exploring the GIDL design space, yield optimization, and pathfinding activities in next-generation 3D V-NAND flash memory.

ACKNOWLEDGMENT

The EDA tool was supported by the IC Design Education Center (IDEC), South Korea.

REFERENCES

[1] S. S. Kim, S. K. Yong, W. Kim, S. Kang, H. W. Park, K. J. Yoon, D. S. Sheen, S. Lee, and C. S. Hwang, "Review of semiconductor flash memory devices for material and process issues," *Adv. Mater.*, 2022. [Online]. Available: https://onlinelibrary.wiley.com/action/showCitFormats?doi=10.1002%2Fadma.202200659, doi: 10.1002/adma.202200659.

[2] H.-W. Chen, S. Verhaverbeke, R. Gouk, K. Leschkies, S. Sun, N. Bekiaris, and R. J. Visser, "Supercritical drying: A sustainable solution to pattern collapse of high-aspect-ratio and low-mechanical-strength device structures," *ECS Trans.*, vol. 69, no. 8, pp. 119–130, Sep. 2015, doi:10.1149/06908.0119ecst.

[3] K.-T. Park, "Three-dimensional 128 Gb MLC vertical NAND flash memory with 24-WL stacked layers and 50 MB/s high-speed programming," *IEEE J. Solid-State Circuits*, vol. 50, no. 1, pp. 204–213, Jan. 2015, doi: 10.1109/JSSC.2014.2352293.

[4] A. Goda, "3-D NAND technology achievements and future scaling perspectives," *IEEE Trans. Electron Devices*, vol. 67, no. 4, pp. 1373–1381, Apr. 2020, doi: 10.1109/TED.2020.2968079.

[5] J. Lee, J. Jang, J. Lim, Y. G. Shin, K. Lee, and E. Jung, "A new ruler on the storage market: 3D-NAND flash for high-density memory and its technology evolutions and challenges on the future," in *IEDM Tech. Dig.*, Dec. 2016, p. 11, doi: 10.1109/IEDM.2016.7838394.

[6] C. Caillat, K. Beaman, A. Bicksler, E. Camozzi, T. Ghilardi, G. Huang, H. Liu, Y. Liu, D. Mao, S. Mujumdar, N. Righetti, M. Ulrich, C. Venkatasubramanian, X. Yang, A. Goda, S. Gowda, H. Mebrahtu, H. Sanda, Y. Yuwen, and R. Koval, "3DNAND GIDL-assisted body biasing for erase enabling CMOS under array (CUA) architecture," in *Proc. IEEE Int. Memory Workshop (IMW)*, May 2017, pp. 1–4, doi: 10.1109/IMW.2017.7939067.

[7] S. Choi, C. Choi, J. K. Jeong, and Y.-H. Song, "Innovative structure to improve erase speed in 3-D NAND flash memory with cell-on-peri (COP) applied," *IEEE Trans. Electron Devices*, vol. 69, no. 9, pp. 4883–4888, Sep. 2022, doi: 10.1109/TED.2022.3188581.

[8] G. Malavena, A. L. Lacaita, A. S. Spinelli, and C. M. Compagnoni, "Investigation and compact modeling of the time dynamics of the GIDL-assisted increase of the string potential in 3-D NAND flash arrays," *IEEE Trans. Electron Devices*, vol. 65, no. 7, pp. 2804–2811, Jul. 2018, doi: 10.1109/TED.2018.2831902.

[9] G. Malavena, A. Mannara, A. L. Lacaita, A. Sottocornola Spinelli, and C. M. Compagnoni, "Compact modeling of GIDL-assisted erase in 3-D NAND flash strings," *J. Comput. Electron.*, vol. 18, no. 2, pp. 561–568, Apr. 2019, doi: 10.1007/s10825-019-01328-0.

[10] K. Ko, J. K. Lee, and H. Shin, "Variability-aware machine learning strategy for 3-D NAND flash memories," *IEEE Trans. Electron Devices*, vol. 67, no. 4, pp. 1575–1580, Apr. 2020, doi: 10.1109/TED.2020.2971784.

[11] D. C. Lee, J. K. Lee, and H. Shin, "Machine learning model for predicting threshold voltage by taper angle variation and word line position in 3D NAND flash memory," *IEICE Electron. Exp.*, vol. 17, no. 22, 2020, Art. no. 20200345, doi: 10.1587/elex.17.20200345.

[12] D. Wei, L. Qiao, X. Chen, H. Feng, and X. Peng, "Prediction models of bit errors for NAND flash memory using 200 days of measured data," *Rev. Scientific Instrum.*, vol. 90, no. 6, Jun. 2019, Art. no. 064702, doi: 10.1063/1.5064655.

[13] P. Kumari, U. Surendranathan, M. Wasiolek, K. Hattar, N. Bhat, and B. Ray, "Analytical bit-error model of NAND flash memories for dosimetry application," *IEEE Trans. Nucl. Sci.*, vol. 69, no. 3, pp. 478–484, Mar. 2022, doi: 10.1109/TNS.2021.3125652.

[14] S. Korkotsides, G. Bikas, E. Eftaxiadis, and T. Antonakopoulos, "BER analysis of MLC NAND flash memories based on an asymmetric PAM model," in *Proc. 6th Int. Symp. Commun., Control Signal Process. (ISCCSP)*, May 2014, pp. 558–561, doi: 10.1109/ISCCSP.2014.6877936.

[15] D. Wei, L. Qiao, W. Shiyuan, F. Ning, and P. Xiyuan, "Research on prediction model for NAND flash bit errors," in *Proc. 12th IEEE Int. Conf. Electron. Meas. Instrum. (ICEMI)*, vol. 1, Jul. 2015, pp. 233–238, doi: 10.1109/ICEMI.2015.7494259.

[16] D.-H. Lee and W. Sung, "Estimation of NAND flash memory threshold voltage distribution for optimum soft-decision error correction," *IEEE Trans. Signal Process.*, vol. 61, no. 2, pp. 440–449, Jan. 2013, doi: 10.1109/TSP.2012.2222399.

[17] Y. Cai, E. F. Haratsch, O. Mutlu, and K. Mai, "Threshold voltage distribution in MLC NAND flash memory: Characterization, analysis, and modeling," in *Proc. Design, Autom. Test Eur. Conf. Exhib. (DATE)*, Mar. 2013, pp. 1285–1290, doi: 10.7873/DATE.2013.266.

[18] Y. Kim, S. Myung, J. Ryu, C. Jeong, and D. S. Kim, "Physics-augmented neural compact model for emerging device technologies," in *Proc. Int. Conf. Simul. Semiconductor Processes Devices (SISPAD)*, Sep. 2020, pp. 257–260, doi: 10.23919/SISPAD49475.2020.9241638.

[19] J. Wang, Y.-H. Kim, J. Ryu, C. Jeong, W. Choi, and D. Kim, "Artificial neural network-based compact modeling methodology for advanced transistors," *IEEE Trans. Electron Devices*, vol. 68, no. 3, pp. 1318–1325, Mar. 2021, doi: 10.1109/TED.2020.3048918.

[20] Y. Kim and S. Kim, "A process-aware compact model for GIDL-assisted erase optimization of 3-D V-NAND flash memory," *IEEE Trans. Electron Devices*, vol. 70, no. 4, pp. 1664–1670, Apr. 2023, doi: 10.1109/TED.2023.3246024.

YOHAN KIM (Member, IEEE) received the B.S. and M.S. degrees in electrical engineering from Korea University, Seoul, Republic of Korea, in 2006 and 2008, respectively.

Since 2008, he has been a Principal Engineer with the Computational Science and Engineering Team, Innovation Center, Samsung Electronics, Suwon, Republic of Korea, where he involved in developing compact models and design technology co-optimization methodologies of advanced semiconductor devices, such as FinFET, MBCFET, STT-MRAM, DRAM, and V-NAND flash memory. He is also with the Department of Semiconductor and Display Engineering, College of Information and Communication Engineering, Sungkyunkwan University, Suwon. He is the Samsung Electronics representative on the Compact Model Coalition (CMC) for the standardization of the advanced devices and holds several U.S. patents. His current research interests include the modeling, characterization, and simulator acceleration of emerging and neuromorphic devices, such as nanowire, tunneling FET, NCFET, ReRAM, PCM, and FeFET.

SOYOUNG KIM (Senior Member, IEEE) received the B.S. degree in electrical engineering from Seoul National University, Seoul, Republic of Korea, in 1997, and the M.S. and Ph.D. degrees in electrical engineering from Stanford University, Stanford, CA, USA, in 1999 and 2004, respectively.

From 2004 to 2008, she was with Intel Corporation, Santa Clara, CA, USA, where she involved in parasitic extraction and simulation of on-chip interconnects. From 2008 to 2009, she was with Cadence Design Systems, San Jose, CA, USA, where she involved in developing IC power analysis tools. She is currently a Professor with the Department of Semiconductor Systems Engineering, College of Information and Communication Engineering, Sungkyunkwan University, Suwon, Republic of Korea. Her research interests include novel semiconductor device modeling, VLSI computer-aided design, signal integrity, power integrity, and electromagnetic interference in electronic systems.