An Analytical Gate Delay Model in Near/Subthreshold Domain Considering Process Variation

PENG CAO[ID], **(Member, IEEE), ZHIYUAN LIU**[ID]**, JINGJING GUO**[ID]**, AND JIANGPING WU**[ID]

National ASIC System Engineering Research Center, Southeast University, Nanjing 210096, China

Corresponding author: Peng Cao (caopeng@seu.edu.cn)

ABSTRACT Voltage scaling technique is widely employed in state-of-the-art low power circuits with excellent power reduction. However, voltage scaling to sub-threshold (STV) and near-threshold (NTV) domain introduces performance degradation and high process variation sensitivity. Accurate modeling of the statistical characteristics especially the probability distribution function (PDF) and the cumulative distribution function (CDF) is urgently required with process variation consideration. In this paper, a novel analytical model is derived based on log-skew-normal (LSN) distribution to precisely evaluate the gate delay variation. The multi-variate threshold variation in stacked gates are modeled with a linear approximation method in delay distribution derivation. By applying the CDF of the proposed model, the maximum and minimum delay indicated by $\pm 3\sigma$ percentile point can be calculated essentially different from the common method with much higher accuracy. Experimental results show the proposed model is highly fitted with Monte Carlo (MC) results for stochastic delay modeling of generic logic gates in near/subthreshold regime with less than 8% and 6% error in delay variability and $\pm 3\sigma$ delay prediction, showing maximum accuracy improvement about 40 times compared to preproposal models.

INDEX TERMS Process variation, analytical distribution model, log-skew-normal, near-threshold regime.

I. INTRODUCTION

Low voltage design [1]–[4], including near/subthreshold design, has become an attractive solution for applications where performance is not the primary concern to save power. However, due to the small gate voltage drive of the transistors operating in the near/subthreshold voltage regime, the logic gates suffer from high sensitivity to process variation, thus leading to a wider spread in the statistical distribution of performance compared with the designs at super-threshold voltage [5]. Moreover, the gate delay distribution becomes non-Gaussian at low voltages and poses the requirement for the advanced statistical delay variation model [6].

The statistical delay modeling solutions for low voltage circuit can be classified into two categories: the fitting-based approach and the analytical one. The fitting-based models were built by approximating the delay variation as non-Gaussian distributions based on numerous Monte

Carlo (MC) simulations, such as the log-skew-normal (LSN) distribution [7], the inverse-Gaussian (IG) distribution [8], and the Weibull distribution [9]. They suffer from the inevitable long simulation runtimes for each specific circuit to fit the distribution parameters and could not obtain physical insights for circuit design parameters. The analytical models for the effect of random process variations were presented in [10]–[13]. In [10], the gate delay in subthreshold regime was considered to follow log-normal (LN) distribution with threshold voltage fluctuation, which was extended for NAND gate by approximating the threshold voltages of the stacked NMOS transistors follow the identified Gaussian distribution. The delay variability (σ/μ) was derived analytically in [11] for subthreshold circuit where correlation between transistors in a stack is extracted for generic logic gates and verified under specific corner. A transregional current and delay model was developed in [12] in closed-form for near-threshold and utilized to generate a LN-distributed delay model under threshold variation, which is further developed in [13] to model statistical timing behavior at

microarchitectural level. Although most analytical models approximate the stochastic circuit delay follows the LN distribution at low voltage, it does not hold for near-threshold region as indicated in [7].

In this paper, an analytical delay variation model is developed based on LSN distribution for generic logic gates in near/subthreshold regime considering threshold voltage as the dominated process variation. The main contributions of this work are summarized as follows.

- The LSN-based delay model is derived analytically via moment matching technique where the distribution parameters are derived as the function of circuit parameters including supply voltage, threshold voltage, the number of stacked transistors, which has been validated in near/subthreshold domain with inverter and NAND2/NOR2 gates. Different from other delay variation models based on LN distribution [10]–[13], our model is verified to be more compatible with the current/delay model in near-threshold regime. Compared with the fitting-based approach with LSN approximation [7], this work is more effective by avoiding time-consuming MC simulations for each circuit with accuracy.
- In order to model multi-variate threshold variations in stacked gates, a linear approximation method is proposed to model the statistical impact of stacked transistors, which shows better accuracy when applied in near/subthreshold voltage domain compared with the relative approach [12].

The rest of paper is organized as follows. Following the introduction, the critical factor of process variation, the current and delay model, and the evaluation of the delay distribution in near/subthreshold regime are presented in Section 2. Subsequently, in Section 3, the statistical delay models for several logic gates are derived based on LSN distribution in near-threshold region via moment matching technique. Experiments and comparisons are given in Section 4. Finally, conclusions are drawn in Section 5.

II. PRELIMINARIES

At the beginning of this section, the threshold voltage is validated to be the critical factor of process variation by calculating the impact of process parameters on the skewness of current distribution in different voltage nodes. Thereafter, the drain current equation is formulated in near/subthreshold regime as well as the delay equation.

A. CRITICAL FACTOR OF PROCESS VARIATION

In general, the process variation can be classified into global and local ones. Although channel length variations and random dopant fluctuations(RDF) are equally important for super-threshold operation, threshold voltage variability owing to RDF is considered as the main source of current and delay variations in low voltage regime so that the impact of channel length variations can be neglected [14], [15]. As many prior publications have re-iterated, the main cause of

local variation in modern technology is the threshold voltage (V_{th}) variation [10]–[12], which leads to asymmetric current and delay distribution at low voltage.

The skewness is a measure of the asymmetry of the probability distribution of a random variable, which is usually defined as the third standardized moment. At low voltage, the statistical distributions of drain current and gate delay become no longer symmetric under process variation, whose shapes could not be determined precisely by the mean and the standard deviation, namely the first and second moment, without higher moments such as skewness. Fig. 1 demonstrate the skewness of the current distribution for an inverter gate from SMIC40LL library versus the supply voltage, which were obtained using 10K MC simulations under the variations of threshold voltage (V_{th}) only, all process parameters except V_{th} and all process parameters. It can be observed from the figure that the impact of the threshold voltage variation on the skewness of the current distribution is considerably higher than that of the variation from other parameters, especially at lower supply voltages.
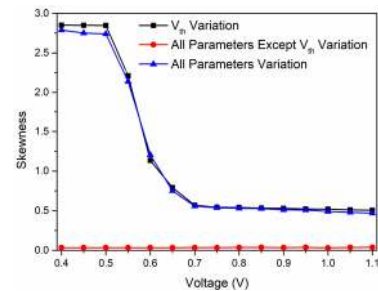

FIGURE 1. Relationship between process parameters and the skewness of drain current at different voltage nodes.

The variability is defined as the standard deviation (σ) of a random variable normalized by its mean value (μ), which indicates the global deviation of the statistical distribution. Further investigation is made in Table 1 which illustrates the variability (σ/μ) of the statistical drain current (I_{on}) and gate delay (T_d) for inverter gate and NAND gate at the supply voltage of 0.35V through 10K MC SPICE simulations with the output load capacitance of 1fF. By comparing the variabilities of I_{on} and T_d with threshold voltage variation only and with all sources of variation including channel length variation, it can be seen that the impact of other process

TABLE 1. Variabilities of the statistical drain current and gate delay for INV and NAND gate.

	Threshold voltage variation only		All sources of variation		Error	
	I_{on}	T_d	I_{on}	T_d	I_{on}	T_d
INV	0.625	0.724	0.616	0.742	1.4 %	2.4 %
NAND	0.493	0.520	0.503	0.534	2.1 %	2.5%

variation excluding threshold voltage is negligible with the error of less than 3% for both current and gate delay.

Hence, in this paper, V_{th} is specified as the critical process variation, which is considered to follow Gaussian distribution with the mean (μ_0) and the variance (σ_0^2). It should be mentioned that although changing the layout would pose a clear impact on threshold voltage, the impact of layout dependent effects (LDEs) have not been considered in this work since the proposed statistical delay model is mainly targeted for the characterization of logical cells in standard cell library before placement.

B. CURRENT AND DELAY EQUATION IN NEAR/SUBTHRESHOLD REGIME

In subthreshold voltage region, the drain current (I_{ds}) is determined by

$$I_{on} = I_0 \frac{W}{L} e^{\frac{V_{gs}-V_{th}}{n\phi_t}} \left(1 - e^{-\frac{V_{ds}}{\phi_t}}\right) \tag{1}$$

where V_{gs} and V_{ds} are respectively the gate-source voltage and drain-source voltage, V_{th} denotes the threshold voltage, ϕ_t is the thermal voltage, n is the sub-threshold slope factor. In addition, I_0 is a process-dependent parameter determined by

$$I_0 = \mu C_{ox} (n-1) \phi_t^2 \tag{2}$$

where μ is the effective electron mobility, C_{ox} is the oxide capacitance per unit, W is the channel width, L is the channel length.

As shown in [12], [16], (1) is imprecise in near-threshold regime, thus the drain current is modified for low voltage regime, whose form is presented by

$$I_{ds} = I_0 \frac{W}{L} K_0 e^{K_1 \frac{V_{gs}-V_{th}}{n\phi_t} + K_2 \left(\frac{V_{gs}-V_{th}}{n\phi_t}\right)^2} \left(1 - e^{-\frac{V_{ds}}{\phi_t}}\right) \tag{3}$$

where the parameters $K_0/K_1/K_2$ are process-independent fitting constants due to the definition of pinch-off voltage and the utilization of normalized variables during the deviation of the current model [16].

For on state current I_{on}, $V_{ds} = V_{gs} = V_{DD}$, since V_{DD} is several times of ϕ_t even at near/subthreshold voltage regime, yields

$$I_{on} = I_0 \frac{W}{L} K_0 e^X \tag{4}$$

where X is a quartic polynomial of V_{th} as

$$X = \frac{K_1}{n\phi_t} (V_{DD} - V_{th}) + \frac{K_2}{(n\phi_t)^2} (V_{DD} - V_{th})^2 \tag{5}$$

Once I_{on} is determined, the gate delay (T_d) can be approximated with a linear RC-delay model, which can be expressed as

$$T_d = K_f \frac{V_{DD} C_L}{I_{on}} = K_f \frac{V_{DD} C_L}{I_0 \frac{W}{L} K_0} e^{-X} \tag{6}$$

where K_f is a process-dependent fitting parameter, C_L is the load capacitance at the output node of a gate. The fitting parameter K_f serves to normalize the RC time constant so that the delay model could track the drain current more closely, which was obtained by measuring the gate delay and on state current with DC and transient SPICE simulation separately when sweeping load capacitance. Although the value of K_f remains nearly constant for different load capacitances, the mean value of fitted K_f is used for the later statistical derivation to reduce the fitting error.

III. STATISTICAL DELAY MODEL FOR CMOS GATES BASED ON LOG-SKEW-NORMAL DISTRIBUTION

In this section, by assuming the gate delay in NTV regime follows LSN distribution, the statistical delay model for generic logic gate is derived using moment matching technique, which presents the impact of threshold voltage variation on the gate delay distribution. The derivation starts from the inverter gate, then extends to the stacked gates.

A. ANALYTICAL DISTRIBUTION MODEL FOR THE INVERTER GATE

The gate delay in (6) for inverter can be expressed with the logarithmic form as follows

$$Y = \ln (T_d) = -X + C_{Td} \tag{7}$$

where

$$C_{T_d} = \ln \left(K_f \frac{V_{DD} C_L}{I_0 \frac{W}{L} K_0}\right) \tag{8}$$

The gate delay (T_d) can be considered to follow LSN distribution as validated in [7], then its logarithmic form (Y) shown in (7) follows SN distribution [17], whose PDF ($f_{SN}(Y)$) is represented as

$$f_{SN} (Y) = \frac{2}{\omega} \phi \left(\frac{Y-\varepsilon}{\omega}\right) \Phi \left(\lambda \frac{Y-\varepsilon}{\omega}\right) \tag{9}$$

where $\phi (\cdot)$ and $\Phi (\cdot)$ is the PDF and CDF of the standard normal distribution, ε, ω, and λ are the location, scale, and shape parameter of SN distribution, respectively. According to the property of SN distribution, the distribution parameters ε, ω, and λ can be represented by the function of its first, second and third moments, which are its mean (μ), variance (σ^2), and skewness (γ_1). The abovementioned relation can be represented as

$$\begin{cases} \varepsilon = \mu - \omega\beta\sqrt{\frac{2}{\pi}} \\ \omega = \sqrt{\frac{\sigma^2}{1 - \frac{2}{\pi}\beta^2}} \\ \beta = \frac{\pi}{2} \frac{\left(\frac{2\gamma_1}{4-\pi}\right)^{\frac{2}{3}}}{1 + \left(\frac{2\gamma_1}{4-\pi}\right)^{\frac{2}{3}}} \end{cases} \tag{10}$$

where β is calculated to obtain λ as

$$\lambda = \frac{\beta}{\sqrt{1 - \beta^2}} \tag{11}$$

On the other hand, since Y are the function of threshold voltage as shown in (5) and (7), the first, second, and third moments of Y can be also calculated as the function of mean (μ_0) and standard deviation (σ_0) of the Gaussian distributed threshold voltage. Concrete expressions are given as follows

$$
\begin{cases}
\mu = -\dfrac{K_1}{n\phi_t}E\left[V_{DT}\right] - \dfrac{K_2}{(n\phi_t)^2}E\left[V_{DT}^2\right] + C_{T_d} \\
\sigma^2 = \dfrac{K_2^2}{(n\phi_t)^4}\left(E\left[V_{DT'}^4\right] - E\left[V_{DT'}^2\right]^2\right) \\
\gamma_1 = \dfrac{\dfrac{-K_2^3}{(n\phi_t)^6}E\left[V_{DT'}^6\right] + \dfrac{3}{4}\dfrac{K_1^2 K_2}{(n\phi_t)^4}E\left[V_{DT'}^4\right]}{\sigma^3} \\
\quad + \dfrac{\dfrac{3}{16}\dfrac{K_1^4}{(n\phi_t)^2 K_2}E\left[V_{DT'}^2\right]}{\sigma^3} \\
\quad + \dfrac{\dfrac{K_1^6}{64 K_2^3} - 3\mu\sigma^2 - \mu^3}{\sigma^3}
\end{cases} \tag{12}
$$

The $E[\cdot]$ in equation (12) calculates the mean value, V_{DT} and $V_{DT'}$ represent the following two expressions and both of them follow the Gaussian distribution.

$$
\begin{cases}
V_{DT} = V_{DD} - V_{th} \\
\quad \sim N\left(V_{DD} - \mu_0, \sigma_0\right) \\
V_{DT'} = V_{DD} - V_{th} + \dfrac{n\phi_t K_1}{2K_2} \\
\quad \sim N\left(V_{DD} - \mu_0 + \dfrac{n\phi_t K_1}{2K_2}, \sigma_0\right)
\end{cases} \tag{13}
$$

With integral deduction, several terms in (12) can be expressed by

$$
\begin{cases}
E\left[V_{DT'}^2\right] = \sigma_0^2 + \left(E\left[V_{DT'}\right]\right)^2 \\
E\left[V_{DT'}^4\right] = 3\sigma_0^4 + 6\sigma_0^2\left(E\left[V_{DT''}\right]\right)^2 + \left(E\left[V_{DT''}\right]\right)^4 \\
E\left[V_{DT'}^6\right] = 15\sigma_0^6 + 45\sigma_0^4\left(E\left[V_{DT''}\right]\right)^2 \\
\quad + 15\sigma_0^2\left(E\left[V_{DT''}\right]\right)^4 + \left(E\left[V_{DT''}\right]\right)^6
\end{cases} \tag{14}
$$

where

$$
\begin{cases}
E\left[V_{DT}\right] = V_{DD} - \mu_0 \\
E\left[V_{DT'}\right] = V_{DD} - \mu_0 + \dfrac{n\phi_t K_1}{2K_2} \\
E\left[V_{DT''}\right] = V_{DD} - \mu_0 + \dfrac{K_1/n\phi_t}{K_2/(n\phi_t)^2}
\end{cases} \tag{15}
$$

It is worth noting that except the mean (μ_0) and standard deviation (σ_0) of threshold voltage, most parameters in (12) are environment parameters, process-independent parameters, or process-dependent parameters obtained by fitting. Although the mean of Gaussian-distributed threshold voltage could be extracted as the nominal value by SPICE simulations for transistors with different sizes, MC SPICE simulations are required for the minimum-sized transistor at each specific process node to extract its standard deviation, which could be further utilized to deduce σ_0 for transistors with larger sizes by Pelgrom's law [18]. According Pelgrom's law, the standard deviation is inversely related to the square root of the transistor size. By taking a four-time minimum-width

transistor as an example, whose channel length is commonly equal to the minimum for most standard cell libraries, its standard deviation could be calculated as $\sigma_0/2$. Compared with the traditional MC-based methods to acquire the statistical characteristics of gate delay, the effort of MC simulations is remarkably reduced to be only once for various gate sizes and output load capacitances owing to the model expressed in (12).

By employing moment matching technique, the distribution parameters ε, ω, and λ can be solved by joining (10) and (12), which are expressed as the function of the mean and standard deviation of threshold voltage, the supply voltage and other related constants.

According to [16], T_d follows LSN distribution with the same distribution parameters ε, ω, and λ as the SN-distributed Y.

$$
T_d \sim LSN\left(\varepsilon, \omega^2, \lambda\right) \tag{16}
$$

Its PDF ($f_{LSN}(T_d)$) and CDF ($F_{LSN}(T_d)$) with the location (ε), scale (ω), and shape (λ) parameters are shown as follows

$$
f_{LSN}(T_d) = \frac{2}{\omega y}\phi\left(\frac{\ln(T_d) - \varepsilon}{\omega}\right)\Phi\left(\lambda\frac{\ln(T_d) - \varepsilon}{\omega}\right) \tag{17}
$$

$$
F_{LSN}(T_d) = \Phi\left(\frac{\ln(T_d) - \varepsilon}{\omega}\right) - 2T\left(\frac{\ln(T_d) - \varepsilon}{\omega}, \lambda\right) \tag{18}
$$

in which $T(H, A)$ is the Owen's T function and can be given by

$$
T\left(\frac{\ln(T_d) - \varepsilon}{\omega}, \lambda\right) = \frac{1}{2\pi}\int_0^\lambda \frac{e^{-\frac{\ln(T_d) - \varepsilon}{\omega}^2}{2}(1+t^2)}}{1 + t^2}dt \tag{19}
$$

After determining the PDF and CDF, the mean (E_{T_d}) and variance (D_{T_d}) of gate delay can be formulated accordingly

$$
\begin{cases}
E_{T_d} = 2e^\varepsilon e^{\frac{\omega^2}{2}}\phi(\beta\omega) \\
D_{T_d} = 2e^{2\varepsilon}e^{\omega^2}\left(e^{\omega^2}\phi(2\beta\omega) - 2\phi^2(\beta\omega)\right)
\end{cases} \tag{20}
$$

B. ANALYTICAL DISTRIBUTION MODEL FOR STACKED GATE

The previous derivation for statistical inverter delay distribution can be expanded to stack topology including NMOS stack (NAND gate) and PMOS stack (NOR gate), whose delay distribution modeling is similar. Without loss of generality, the derivation of statistical delay model for a NAND2 gate shown in Fig. 2 is investigated, where the intermediate node between the two NMOS, T_U and T_L, is annotated with the voltage value V_x. The discharge current through upper (I_U) and lower (I_L) NMOS transistors can be

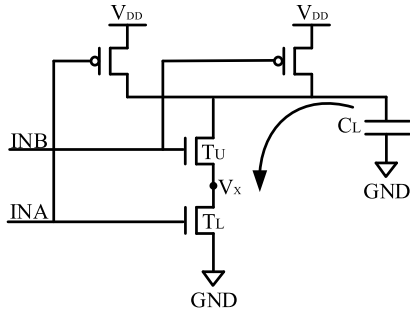**FIGURE 2.** Schematic of NAND2 gate.

determined by

$$
\begin{cases}
I_U = I_0 \dfrac{W}{L} K_0 e^{K_1 \frac{V_{DD}-V_X-V_{thU}}{n\phi_t} + K_2 \left(\frac{V_{DD}-V_X-V_{thU}}{n\phi_t}\right)^2} \\
\quad \left(1 - e^{-\frac{V_{DD}-V_X}{\phi_t}}\right) \\
I_L = I_0 \dfrac{W}{L} K_0 e^{K_1 \frac{V_{DD}-V_{thL}}{n\phi_t} + K_2 \left(\frac{V_{DD}-V_{thL}}{n\phi_t}\right)^2} \\
\quad \left(1 - e^{-\frac{V_X}{\phi_t}}\right)
\end{cases}
\tag{21}
$$

where V_{thU} and V_{thL} is the threshold voltage for transistor T_U and T_L, respectively.

It can be seen from (21) that compared with inverter, the discharge current is related to the intermediate node voltage V_x for stacked gate, which raises the difficulty of analytical derivation for current model in NTV regime. Firstly, the assumption of $I_U = I_L$ may not be valid during the discharge procedure due to the intermediate node capacitance. Secondly, even when letting $I_U = I_L$, the expression of V_x turns out to be the solution issue of transcendental equation from (21), which means it cannot be solved analytically.

To solve the abovementioned problem, a linear approximation method is introduced to express V_X with V_{thL} and V_{thU}. The relationship among V_X, V_{thL}, and V_{thU} are demonstrated in Fig. 3 via MC simulation for a 40nm process NAND2 cell at 0.45V supply voltage. It should be noticed that since the values of V_X, V_{thL}, and V_{thU} are extracted from simulation results, the impact of the intermediate node capacitance as well as the difference between I_U and I_L is not ignored. It can
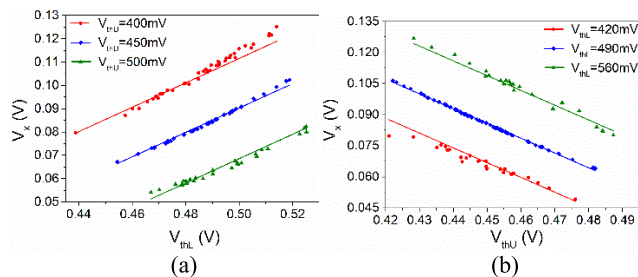
be observed from Fig. 3(a) that the simulated V_X changes linearly with V_{thL}. Same tendency can be observed from Fig. 3(b), which demonstrates that V_X and V_{thU} also show linear relationship.

Therefore, V_X can be expressed in an empirical manner by a bivariate linear model, which is shown as follows

$$
V_X = k_U V_{thU} + k_L V_{thL} + k_C
\tag{22}
$$

where k_U, k_L, and k_C are the fitting parameters obtained by MC SPICE simulation. Since the fitting coefficients of the empirical model vary little for stacked transistors with different sizes, MC SPICE simulation could be performed only once to fit coefficients for the stacked transistors with some certain size, which could be used to estimate V_x for transistors of other sizes.

The effect of the proposed empirical bivariate linear model can be validated by comparing the mean and standard deviation of V_X obtained from (22) and MC simulation. Related fitting errors of V_X for NAND2 and NOR2 gates under 40nm process are listed in Table 2, which shows the average fitting error for mean ($\mu(V_X)$) and standard deviation ($\sigma(V_X)$) is less than 1% when supply voltage ranges from 0.35V to 0.55V.

TABLE 2. Related fitting errors of gates with stack topology for different supply voltages.

Voltage		0.35V	0.45V	0.55V
NAND2	$\mu(V_X)$	<0.1%	<0.1%	<0.1%
	$\sigma(V_X)$	0.53%	0.42%	0.37%
NOR2	$\mu(V_X)$	<0.1%	<0.1%	<0.1%
	$\sigma(V_X)$	0.32%	0.29%	0.33%

Although threshold voltage changes linearly with temperature, the linear relation among threshold voltages of stacked transistors and V_x holds under different temperatures. Further investigation for the accuracy of the bivariate linear model is shown in Table 3 for NAND2 and NOR2 gate working at the voltage of 0.45V with temperature ranging from $-25°$ to $100°$. It can be seen that the error of mean ($\mu(V_X)$) and standard deviation ($\sigma(V_X)$) is still negligible considering temperature effect.

TABLE 3. Related fitting errors of gates with stack topology for different temperatures.

Temperature		-25℃	25℃	75℃	100℃
NAND2	$\mu(V_X)$	<0.1%	<0.1%	<0.1%	<0.1%
	$\sigma(V_X)$	0.48%	0.42%	0.37%	0.33%
NOR2	$\mu(V_X)$	<0.1%	<0.1%	<0.1%	<0.1%
	$\sigma(V_X)$	0.36%	0.29%	0.34%	0.31%

By substituting V_X into the expression of I_U in (21), the discharge current for the NAND2 gate (I_{st}) can be deduced with the consideration of the impact of intermediate node capacitance and written as

$$
I_{st} = I_0 \dfrac{W}{L} K_0 e^{K_1 \frac{V_{DD}-V_{th\_st}}{n\phi_t} + K_2 \left(\frac{V_{DD}-V_{th\_st}}{n\phi_t}\right)^2}
\tag{23}
$$

FIGURE 3. Relationship between V_X and: (a) V_{thL}; (b) V_{thU}.

where V_{th_st} is the equivalent random variable which presents the joint effect of V_{thU} and V_{thL}. The form of V_{th_st} is given by

$$V_{th\_st} = (k_U + 1) V_{thU} + k_L V_{thL} + k_c \tag{24}$$

Due to the fact that V_{thU} and V_{thL} both follows Gaussian distribution, thus by symbolizing $V_{thU} \sim N(\mu_U, \sigma_U)$ and $V_{thL} \sim N(\mu_L, \sigma_L)$, the equivalent variable V_{th_st} with the linear combination of V_{thU} and V_{thL} also follows Gaussian distribution, whose mean (μ_{st}) and standard deviation (σ_{st}) can be determined by

$$\begin{cases} \mu_{th\_st} = (k_U + 1) \mu_U + k_L \mu_L + k_C \\ \sigma_{th\_st} = \sqrt{(k_U + 1)^2 \sigma_U^2 + k_L^2 \sigma_L^2} \end{cases} \tag{25}$$

After deriving V_X, I_{st}, and V_{th_st}, the delay for stacked transistors (T_{d_st}) can be expressed by taking V_{th_st} as the equivalent threshold voltage in (6) as

$$T_{d\_st} = K_f \frac{V_{DD} C_L}{I_0 \frac{W}{L} K_0} e^{-K_1 \frac{V_{DD} - V_{th\_st}}{n\phi_t} - K_2 \left(\frac{V_{DD} - V_{th\_st}}{n\phi_t}\right)^2} \tag{26}$$

Similarly as the derivation flow for inverter in the last section, the first, second, and third moments of the logarithmic form of T_{d_st} can be calculated to obtain the corresponding mean, standard deviation, and skewness according to (26). Thereafter, using moment matching technique, the LSN distribution related parameters including shape, scale, and location can be obtained through substituting above mentioned statistical parameters of the logarithm form of T_{d_st} into (10) and (11).

IV. WORST/BEST CASE GATE DELAY EVALUATION

With the location, scale and shape parameters of the LSN distributed T_d shown in (10) and (11), their CDF can be deduced as in (18) so that any percentile points can be determined accordingly.

As for T_d, $\pm 3\sigma$ percentile points are commonly used to indicate the maximum and minimum delay with Gaussian distribution assumption, which are $\Phi(3) \approx 99.87\%$ percentile point for maximum delay and $\Phi(-3) \approx 0.13\%$ percentile point for minimum delay. However, as for non-Gaussian, the $\mu \pm 3\sigma$ delay may be far away from the 99.87% and 0.13% percentile points due to asymmetry.

Based on the LSN-based model, the so-called $\pm 3\sigma$ percentile points can be derived by letting the CDF of LSN distribution in (19) to be $\Phi(\pm 3)$, which is

$$F_{LSN}(Y) = \Phi\left(\frac{\ln(Y) - \varepsilon_Y}{\omega_Y}\right) - 2T\left(\frac{\ln(Y) - \varepsilon_Y}{\omega_Y}, \lambda_Y\right) = \Phi(\pm 3) \tag{27}$$

To solve this equation, an important finding is that in practice, experiment results show that for T_d, the value of shape parameter (λ_{Td}) in LSN distribution is close to 1.

By applying the following property of Owen' function,

$$T(h, 1) = \frac{1}{2} \Phi(h)(1 - \Phi(h)) \tag{28}$$

Eq. (28) can be derived by

$$\Phi^2\left(\frac{\ln(Y) - \varepsilon_Y}{\omega_Y}\right) = \Phi(\pm 3) \tag{29}$$

It can be clearly shown from (29) that the maximum/minimum delay at $3\sigma/-3\sigma$ percentile point in LSN-distributed T_d locates at $T_{d\max}/T_{d\min}$, which is

$$\begin{cases} T_{d\max} = e^{\varepsilon_Y + \Phi^{-1}(\sqrt{\Phi(3)})\omega_Y} \approx e^{\varepsilon_Y + 3.21\omega_Y} \\ T_{d\min} = e^{\varepsilon_Y + \Phi^{-1}(\sqrt{\Phi(3)})\omega_Y} \approx e^{\varepsilon_Y - 1.79\omega_Y} \end{cases} \tag{30}$$

Substituting ε_{Td}, ω_{Td} and ε_{Td_st}, ω_{Td_st} derived by (10) and (11) into (30), the maximum/minimum delay for inverter and NAND2 can be obtained, respectively.

V. EXPERIMENTAL RESULTS AND COMPARISON

The goal of this section is to verify the proposed statistical model for gate delay at near/subthreshold region and to compare it with MC simulation results as golden reference and other competitive methods.

A. EXPERIMENTAL SETUP

The proposed statistical delay model was validated under the process of SMIC40LL technology in low voltage regime, by which the estimated statistical characteristics were compared to MC SPICE simulations using foundry-provided BSIM4 models. The inverter gate and stacked gates including NAND2 and NOR2 were selected for comparison with FO4 load at one subthreshold voltage (0.35V) and two near-threshold voltage nodes (0.45V/0.55V). The number of MC simulations were 10K for each gate at each operating voltage.

The validation process of the proposed model is given by the flow diagram shown in Fig. 4, which are mainly composed of two steps.

Step 1: Parameter acquisition. When the process technology and standard cell library have been determined, the parameters utilized in the proposed model should be obtained via SPICE simulations. To fit the process-dependent parameters for the drain current model in (3) and delay model in (6) for low voltage domain, several SPICE simulations are required by sweeping the working conditions such as supply voltages, gate sizes and output load capacitances. The mean of Gaussian-distributed threshold voltage is extracted as nominal value under the SPICE simulation for transistors with different sizes. As for the standard deviation of threshold voltage of the minimum-sized transistor and the coefficients of the linear approximation model in (22) for stacked gates, MC simulations are needed to be performed only once separately to extract parameters for all working conditions.

Step 2: Model Deviation. The proposed models are derived with the acquired parameters in Step 1 as well as the specified working conditions including the size, stack topology, output load capacitance of gate and supply voltage. With the
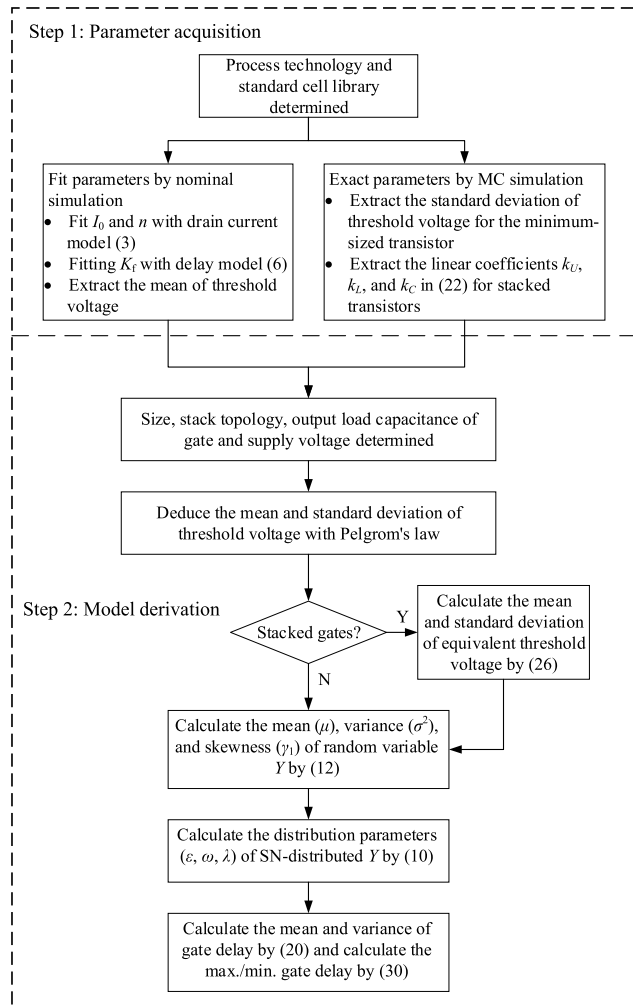
FIGURE 4. The flow diagram of statistical gate delay modeling.

extracted mean and standard deviation of threshold voltage of the minimum-sized transistor by MC simulations, those for the gates with larger sizes could be deduced with Pelgrom's law. As for the staked gates, the mean and standard deviation of the equivalent threshold voltage can be calculated by (26). Subsequently, the mean, variance and skewness of the random

variable Y, the SN distributed logarithm of gate delay, can be calculated by (12) before calculating its distribution parameters by employing moment matching technique in (10), which are also the distribution parameters of LSN distributed gate delay. Finally, based on the properties of LSN distribution, the mean and variance of statistical gate delay could be calculated by (20) and the maximum/minimum gate delay could be deduced by (30).

It can be clearly shown in Fig. 4 that the most time consuming part of the total procedure lies in the MC simulation in the first step while the calculations in the model derivation step are all in closed form which take almost no time. Different from the fitting based methods [7]–[9] where MC simulations are requited for each gates at each specific working conditions, only two sets of MC simulations are required under each specific process technology to extract necessary parameters for the following gate delay modelling. On a modern server with Intel Xeon processor, less than one minute CPU time would be consumed for each set of 10K MC BSIM4 transient SPICE simulations. Compared with other analytical models as [10], [12], although additional MC simulations are performed for the extraction of liner coefficients for stacked gates, significant accuracy improvement is achieved as discussed later at the cost of computation effort in tens of seconds.

B. MODEL EVALUATION AND COMPARISON

The accuracy of the proposed statistical gate delay models is demonstrated by means of PDF and compared with MC simulation results as well as the models from [10], [12] in Fig. 5, where the PDFs for the inverter gate, NAND2 and NOR2 are illustrated in Fig. 5(a), Fig. 5(b) and Fig. 5(c), respectively. It can be demonstrated that our model fits well with MC simulation results for all gates due to the analytical derivation based on LSN distribution. For the competitive models based on LN distribution, the stacking effect exaggerates the incorrectness of PDFs for NAND2 and NOR2 gate as shown in Fig. 5(b) and Fig. 5(c).

To assess the accuracy of different delay variation models comprehensively, the delay variability and the predicted maximum/minimum delay are used as the evaluation metrics.
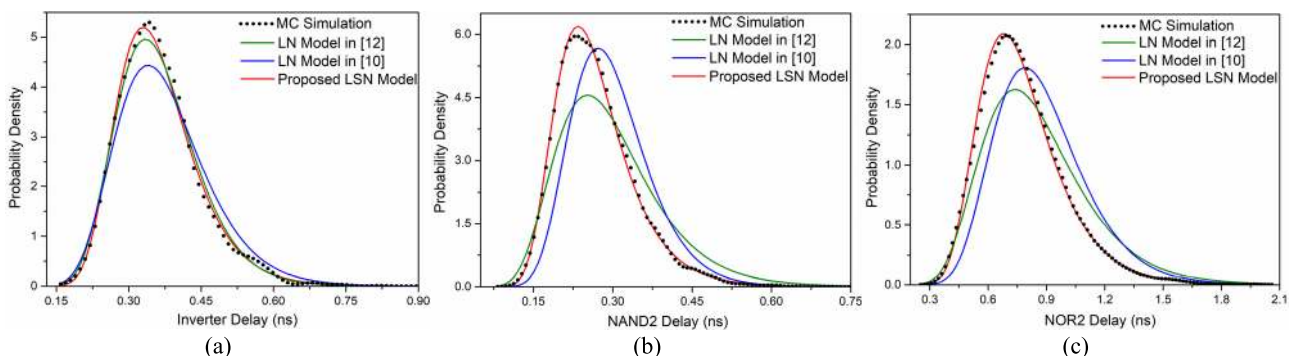

FIGURE 5. The probability density of gates with FO4 load in supply voltage of 0.45V: (a) Inverter; (b) NAND2; (c) NOR2.
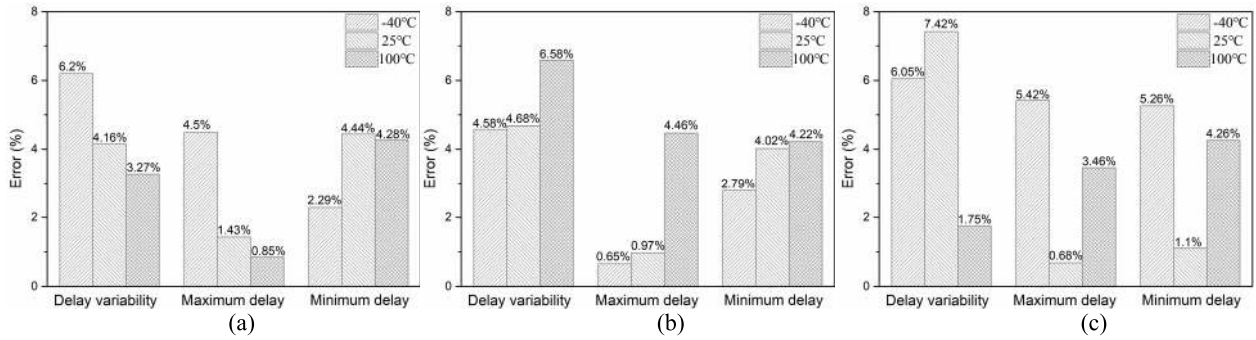
FIGURE 6. Error in terms of the delay variability and the maximum/minimum delay in near/subthreshold regime under different temperatures for inverter: (a) 0.35V; (b) 0.45V; (c) 0.55V.
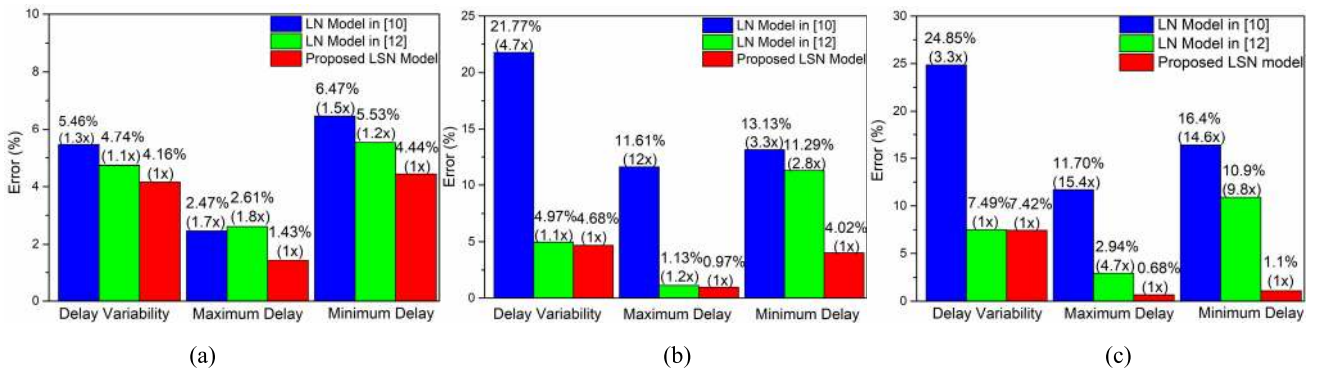

FIGURE 7. Error comparison in terms of the delay variability and the maximum/minimum delay in near/subthreshold regime for inverter: (a) 0.35V; (b) 0.45V; (c) 0.55V.

The former is defined as the standard deviation (σ) of gate delay normalized by its mean value (μ), which indicates the global deviation of the statistical models, whereas the latter is denoted by the $\pm 3\sigma$ quantile points of delay, embodying the accuracy in extreme conditions.

The accuracy of the proposed statistical delay model was validated in near/subthreshold domain under different temperatures. As shown in Fig. 6, the estimated delay variability and the maximum/minimum delay of inverter gate are limited to be less than 8% and 6% respectively in low voltage domain with no significant increase considering temperature effect.

Fig. 7 illustrates the error of inverter delay models at three operating voltages compared with MC simulation results, where the error of the competitive models is normalized by that of ours. It can be found that although the error of delay variability in of our model is similar as that of [12], the maximum/minimum delay prediction achieves up to $4.7 \times /9.8 \times$ precision improvement in near-threshold domain, which can be taken as the evidence of the contribution from LSN distribution although modeled with the same statistical characteristics. The error of the delay variability and the extreme delay of the model in [10] is at up to $4.7 \times$ and $15.4 \times$ larger than ours in near-threshold domain due to the fact that the relation with the threshold voltage is not considered appropriately at near-threshold voltage.

Fig. 8 and Fig. 9 compare the accuracy of delay variation models when applied for stacked gates including NAND2 and NOR2. The accuracy of the delay variability and the extreme value of model [12] degrades distinctly to be respectively over $9.5 \times$ and $6.4 \times$ worse than our model at subthreshold voltage for stacked gates although it is comparative with this work when modeling the inverter delay variation. The reason is that the proposed linear approximation method for stacking effect matches the multivariate threshold variation much better than the fitting approach used in [12] for stacked gates under nominal condition. The loss of accuracy is exaggerated for near-threshold stacked gates where up to $40.8 \times$ and $6.5 \times$ error surge can be found for delay variability and the maximum/minimum delay respectively. When considering the accuracy of the model from [10] in subthreshold regime, the error of the delay variability and maximum/minimum delay are at least $1.7 \times$ and $2.1 \times$ worse than that of ours, indicating the advantage of our model for stacked gates over the identically threshold variation assumption used in [10].

In addition, since the current model utilized in this work [16] for low supply voltage is applicable for FinFET devices with considerable precision [19], the proposed statistical delay model can also be applied for FinFET with threshold variation besides bulk CMOS technology.
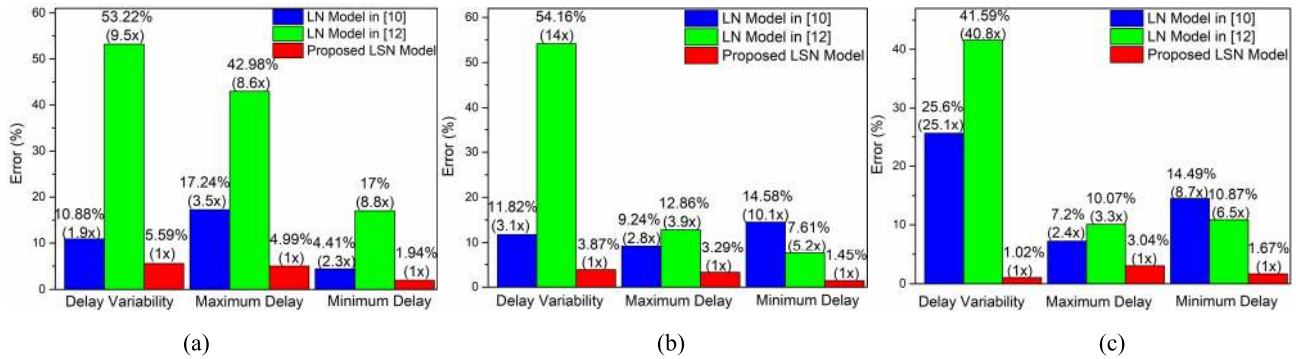
FIGURE 8. Error comparison in terms of the delay variability and the maximum/minimum delay in near/subthreshold regime for NAND2: (a) 0.35V; (b) 0.45V; (c) 0.55V.
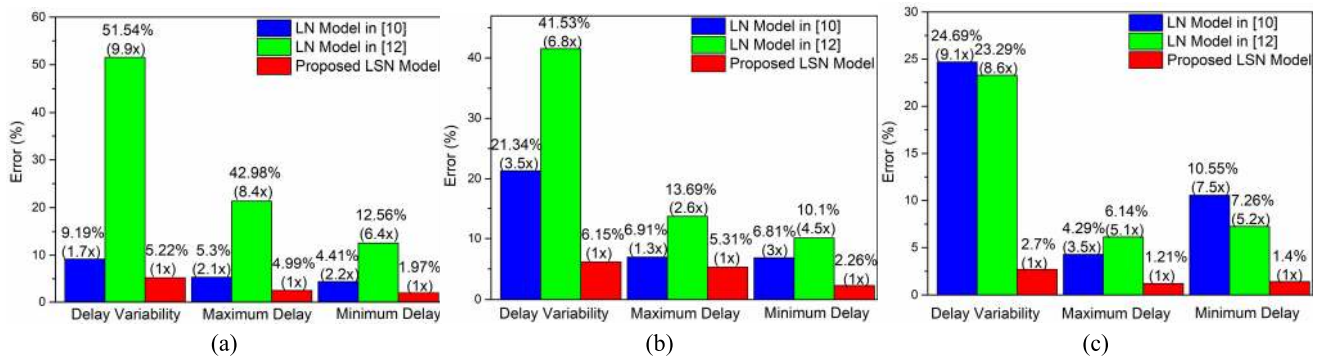

FIGURE 9. Error comparison in terms of the delay variability and the maximum/minimum delay in near/subthreshold regime for NOR2: (a) 0.35V; (b) 0.45V; (c) 0.55V.

VI. CONCLUSION

This paper proposes an accurate LSN-distributed model for the gate delay in low voltage region considering process variation via moment matching technique. For deducing the delay distribution of the stacked gates, the multi-variate threshold variation is modeled with a linear approximation method in delay distribution derivation. Moreover, in order to evaluate the worst/best case delay of gates, the derived location, scale and shape parameters and cumulative distribution function (CDF) are employed to deduce the delay at $\pm 3\sigma$ percentile points. The proposed distribution model shows an excellent fit with Monte Carlo (MC) results for stochastic delay modeling of generic logic gates in near/subthreshold regime. In addition, the error in delay variability and 3σ delay prediction show obvious enhancement compared to preproposal models.

REFERENCES

[1] B.-H. Chen, P.-Y. Chou, Y.-B. Fang, L.-K. Yong, T.-J. Lin, and J.-S. Wang, "Design of ultra-low-leakage near-threshold dynamic circuits in nano CMOS for IoT applications," in *Proc. IEEE 16th Int. Conf. Nanotechnol.*, Aug. 2016, pp. 537–540.

[2] H. Kaul, M. Anders, S. Hsu, A. Agarwal, R. Krishnamurthy, and S. Borkar, "Near-threshold voltage (NTV) design: Opportunities and challenges," in *Proc. 49th Annu. Design Automat. Conf.*, 2012, pp. 1153–1158.

[3] M. Alioto, E. Consoli, and G. Palumbo, "Analysis and comparison in the energy-delay-area domain of nanometer CMOS flip-flops: Part I—Methodology and design strategies," *IEEE Trans. Very Large Scale Integr. (VLSI) Syst.*, vol. 19, no. 5, pp. 725–736, May 2011.

[4] D. Bol, R. Ambroise, D. Flandre, and J. D. Legat, "Interests and limitations of technology scaling for subthreshold logic," *IEEE Trans. Very Large Scale Integr. (VLSI) Syst.*, vol. 17, no. 10, pp. 1508–1519, Oct. 2009.

[5] R. G. Dreslinski, M. Wieckowski, D. Blaauw, D. Sylvester, and T. Mudge, "Near-threshold computing: Reclaiming Moore's law through energy efficient integrated circuits," *Proc. IEEE*, vol. 98, no. 2, pp. 253–266, Feb. 2010.

[6] C. Hou, "A smart design paradigm for smart chips," in *IEEE ISSCC Dig. Tech. Papers*, Feb. 2017, pp. 8–13.

[7] H. A. Balef, M. Kamal, A. Afzali-Kusha, and M. Pedram, "All-region statistical model for delay variation based on log-skew-normal distribution," *IEEE Trans. Comput.-Aided Des. Integr. Circuits Syst.*, vol. 35, no. 9, pp. 1503–1508, Sep. 2016.

[8] J. Chen, S. Cotofana, S. Grandhi, C. Spagnol, and E. Popovici, "Inverse Gaussian distribution based timing analysis of Sub-threshold CMOS circuits," *Microelectron. Rel.*, vol. 55, no. 12, pp. 2754–2761, Oct. 2015.

[9] L. Zhang, J. Shao, and C. C.-P. Chen, "Non-Gaussian statistical parameter modeling for ssta with confidence interval analysis," in *Proc. Int. Symp. Phys. Design*, Apr. 2016, pp. 33–38.

[10] F. Frustaci, P. Corsonello, and S. Perri, "Analytical delay model considering variability effects in subthreshold domain," *IEEE Trans. Circuits Syst. II Express Briefs*, vol. 59, no. 3, pp. 168–172, Mar. 2012.

[11] Y. Zhang and B. H. Calhoun, "Fast, accurate variation-aware path timing computation for sub-threshold circuits," in *Proc. 15th Int. Symp. Qual. Electron. Design*, Mar. 2014, pp. 243–248.

[12] S. Keller, D. M. Harris, and A. J. Martin, "A compact transregional model for digital CMOS circuits operating near threshold," *IEEE Trans. Very Large Scale Integr. (VLSI) Syst.*, vol. 22, no. 10, pp. 2041–2053, Oct. 2014.

[13] J. Shiomi, T. Ishihara, and H. Onodera, "Microarchitectural-level statistical timing models for near-threshold circuit design," in *Proc. 20th Asia South Pacific Design Automat. Conf.*, Jan. 2015, pp. 87–93.

[14] N. Drego, A. Chandrakasan, and D. Boning, "Lack of spatial correlation in MOSFET threshold voltage variation and implications for voltage scaling," *IEEE Trans. Semicond. Manuf.*, vol. 22, no. 2, pp. 245–255, May 2009.

[15] B. Zhai, S. Hanson, D. Blaauw, and D. Sylvester, "Analysis and mitigation of variability in subthreshold design," in *Proc. Int. Symp. Low Power Electron. Design*, Aug. 2005, pp. 20–25.

[16] D. M. Harris, B. Keller, J. Karl, and S. Keller, "A transregional model for near-threshold circuits with application to minimum-energy operation," in *Proc. Int. Conf. Microelectron.*, Dec. 2010, pp. 64–67.

[17] X. Li, Z. Wu, V. D. Chakravarthy, and Z. Wu, "A low-complexity approximation to lognormal sum distributions via transformed log skew normal distribution," *IEEE Trans. Veh. Technol.*, vol. 60, no. 8, pp. 4040–4045, Oct. 2011.

[18] M. J. M. Pelgrom, A. C. J. Duinmaijer, and A. P. G. Welbers, "Matching properties of MOS transistors," *IEEE J. Solid-State Circuits*, vol. 24, no. 5, pp. 1433–1439, Oct. 1989.

[19] X. Lin, Y. Wang, and M. Pedram, "Joint sizing and adaptive independent gate control for FinFET circuits operating in multiple voltage regimes using the logical effort method," in *Proc. IEEE/ACM Int. Conf. Comput.-Aided Design (ICCAD)*, Nov. 2013, pp. 444–449.

PENG CAO (M'14) received the B.S. and Ph.D. degrees in microelectronics and solid state electronics from Southeast University, Nanjing, China, in 2002 and 2010, respectively.

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

ZHIYUAN LIU received the B.S. degree in electronic science and technology from the University of Shanghai for Science and Technology, Shanghai, China, in 2017. He is currently pursuing the master's degree in electronic engineering with Southeast University, Nanjing, China. His research interests include timing analysis and optimization for near-threshold design.

JINGJING GUO received the B.S. degree in microelectronics and the M.S. degree in integrated circuit engineering from Xidian University, China, in 2011 and 2014, respectively. She is currently pursuing the Ph.D. degree in microelectronics and solid state electronics with Southeast University, Nanjing, China. Her research interest includes low power and low voltage digital circuit design.

JIANGPING WU received the B.S. degree in electronic engineering from Southeast University, Nanjing, China, in 2016, where she is currently pursuing the master's degree in electronic engineering. Her research interest includes delay deviation analysis under near-threshold process variation.

• • •