# Estimation of Delay Variations due to Random-Dopant Fluctuations in Nanoscale CMOS Circuits

Hamid Mahmoodi, *Student Member, IEEE*, Saibal Mukhopadhyay, *Student Member, IEEE*, and Kaushik Roy, *Fellow, IEEE*

***Abstract*—In nanoscale CMOS circuits the random dopant fluctuations (RDF) cause significant threshold voltage ($\mathrm{Vt}$) variations in transistors. In this paper, we propose a semi-analytical estimation methodology to predict the delay distribution [Mean and Standard Deviation (STD)] of logic circuits considering $\mathrm{Vt}$ variation in transistors. The proposed method is fast and can be used to predict delay distribution in nanoscale CMOS technologies both at the circuit and the device design phase. The method is applied to predict the delay distributions in different logic gates and flip-flops and is verified with detail Monte Carlo simulations. It is observed that a 30% spread (STD/Mean) in $\mathrm{Vt}$ variation results in 5% spread in the delay of logic gates (inverter, NAND, etc.). The effect of $\mathrm{Vt}$ variation due to RDF is more significant in the setup time (STD/Mean $= 11\%$) and clock-to-output delay (STD/Mean $= 5\%$ to $25\%$) of flip-flops.**

***Index Terms*—CMOS circuits, delay variations, logic gates, process variations, random dopant fluctuations, statistical modeling, threshold voltage.**

## I. Introduction

In NANOSCALE CMOS devices, the random variations in number and placement of dopant atoms in the channel region cause random variations in the transistor threshold voltage ($\mathrm{Vt}$) [1]–[3], known as the "random (or discrete) dopant effect." This can result in threshold voltage mismatch between transistors on die (intra-die variations) resulting in significant delay variation of logic gates and circuits [3]. The effect of random dopant fluctuations (RDF) on $\mathrm{Vt}$ increases with technology scaling. This is due to the fact that the average number of dopant atoms in the channel of a transistor reduces with technology scaling. For example, assuming a doping density of $10^{18}/\mathrm{cm}^3$, the average number of dopant atoms in the channel of a minimum size (width $= 2\times$length) 70-nm device (effective channel length of 40 nm) is approximately 100. The random variation in this small number of dopant atoms can result in significant variation in the $\mathrm{Vt}$ of the transistors. Hence, the $\mathrm{Vt}$ variation due to RDF can result in significant variation in the delay of a logic circuit. Moreover, the effect of $\mathrm{Vt}$ variation on the delay distribution of a circuit strongly depends on the device geometry (channel length, width, oxide thickness, etc.) and doping profile. Hence, a statistical modeling and analysis of the delay of logic gates (considering $\mathrm{Vt}$ variation due to RDF) is necessary both at the circuit and device design phase to enhance the yield of logic circuits in nanometer regimes. Although the Monte Carlo simulation (e.g., using a circuit simulator like SPICE during circuit design and a device simulator like MEDICI during device design) of gates is accurate in estimating the delay distributions, it considerably increases the design time. The Response Surface generation based Methods (RSM) [4] for statistical delay models considering intra-gate variability also require large number of simulations to generate the response surface. This is also computationally expensive particularly if the estimation is required at the device design phase. In this paper, we propose a semi-analytical method to estimate the delay distributions of logic gates. Particularly, in this work:

- We have developed a general semi-analytical method to predict the mean, standard deviation (STD), and probability distribution function (PDF) of delay in logic circuits considering random $\mathrm{Vt}$ variation in transistors.
- We have applied the proposed method to estimate:
  - distribution of propagation delay in logic gates;
  - distribution of the clock-to-output delay and the setup time in flip-flops;
  - the sensitivity of the delay distribution to the device geometry and doping profile.

Using the proposed models, we have estimated the delay distribution of logic gates (in particular, inverter and NAND gate) considering $\mathrm{Vt}$ variation due to RDF. The models are verified with detailed Monte Carlo simulations to ensure the accuracy. It is observed that application of 30% of $\mathrm{Vt}$ spread (standard deviation of $\mathrm{Vt}$ variation as a percentage of the mean of $\mathrm{Vt}$, i.e., STD/Mean) results in 5% spread (STD/Mean) in the delay of the logic gates. It is further observed that as the $\mathrm{Vt}$'s of transistors in a logic gate become correlated, the delay variations tend to increase. Application of the models to estimate the distribution of the rise and fall times of a gate shows that the rise and fall times can have a significant variation. The models are also applied to estimate the distribution of the setup time and the clock-to-output delay of a flip-flop. The results show that a 30% spread in $\mathrm{Vt}$ variation results in 5% to 25% spread in the clock-to-output delay of a flip-flop depending on data arrival time. The setup time of the flip-flop also shows significant variation (STD/Mean $= 11\%$). Using the proposed models, we have analyzed the impact of device design parameters, in particular, the doping profile and the oxide thickness, on the delay distribution of an inverter. It is observed that increasing the doping and the oxide thickness increases the delay variation.

Manuscript received December 16, 2004; revised March 17, 2005.

The authors are with the Department of Electrical and Computer Engineering, Purdue University, West Lafayette, IN 47907 USA (e-mail: mahmoodi@ecn.purdue.edu).

Digital Object Identifier 10.1109/JSSC.2005.852164

The rest of this paper is organized as follows. In Section II, the mathematical formulation for estimation of gate delays is described. Section III describes the statistical modeling of the delay of logic gates. Section IV describes the statistical model for the delay of a flip-flop. In Section V, the effect of correlation among threshold voltages is analyzed. Section VI analyzes the effect of device parameters on the delay distribution of logic gates. Finally, Section VII concludes the paper.

## II. Methodology of Statistical Gate Delay Estimntion

In this section, we describe the semi-analytical models for estimation of delay distributions in logic gates. We first describe the models used to represent Vt variation due to the RDF. We next describe the mathematical formulation of the proposed semi-analytical models for the estimation of delay variations.

### A. Vt Variation due to Random Dopant Fluctuation (RDF)

The Vt variations ($\delta Vt$) (due to RDF) of different transistors in a circuit are considered as independent Gaussian random variables (mean $= 0$) [1]. The standard deviation ($\sigma_{Vt}$) depends on the manufacturing process, doping profile, and the transistor size, and is given by [2]

$$\sigma_{Vt} = \left[\frac{qT_{\rm ox}}{\varepsilon_{\rm ox}}\sqrt{\frac{(N_aW_d)}{3L_{\min}W_{\min}}}\right] \times \sqrt{\frac{L_{\min}W_{\min}}{LW}} = \sigma_{Vt0} \times \sqrt{\frac{L_{\min}W_{\min}}{LW}} \quad (1)$$

where $N_a$ is the effective channel doping, $W_d$ is the depletion region width, $T_{\rm ox}$ is the oxide thickness, and $L_{\min}$ and $W_{\min}$ are the minimum channel length and width, repectively. During the estimation of delay distribution at the circuit level, we use $\sigma_{Vt0}$ as an input parameter. In Section VI, we describe the impact of variation in $N_a$ and $T_{\rm ox}$ (hence $\sigma_{Vt0}$) on the delay distribution.

### B. Mathematical Formulation for Estimation of Delay Distribution

Let us consider a general logic gate with $n$ transistors (Fig. 1). In general, the propagation delay from input $\mathrm{IN}_j$ to output $t_{dj}$ depends on the Vt of all $n$ transistors (i.e., $\mathrm{Vt}_i$) in the gate. Hence, considering the Vt fluctuation of each transistor ($\delta Vt_i$) from their nominal values ($Vt_{i0}$), $t_{dj}$ can be written as

$$t_{dj} = f(Vt_1,\ldots,Vt_n) = f(Vt_{10}+\delta Vt_1,\ldots,Vt_{n0}+\delta Vt_n). \quad (2)$$

Since the Vt fluctuations in different transistors due to RDF are independent of each other, $\delta Vt_1,\ldots,\delta Vt_n$ are considered as independent Gaussian random variables with zero mean, and STD ($\sigma_{Vti}$) is given by (1). Expanding $t_{dj}$ in multi-variable Taylor series for the variables $\delta Vt_1,\ldots,\ \delta Vt_n$, around their mean ($= 0$), the mean ($\mu_{tdj}$) and STD ($\sigma_{tdj}$) of delay can be expressed as [5]

$$\mu_{tdj} = T_{dj0} + \frac{1}{2}\sum_{\text{all transistors}}\left[\frac{\partial^2 t_{dj}}{\partial(\delta Vt_i)^2}\right]_{\delta Vt_i=0}\sigma^2_{Vt_i}$$
$$\sigma^2_{tdj} = \sum_{\text{all transistors}}\left[\left(\frac{\partial t_{dj}}{\partial(\delta Vt_i)}\right)^2\right]_{\delta Vt_i=0}\sigma^2_{Vt_i} \quad (3)$$

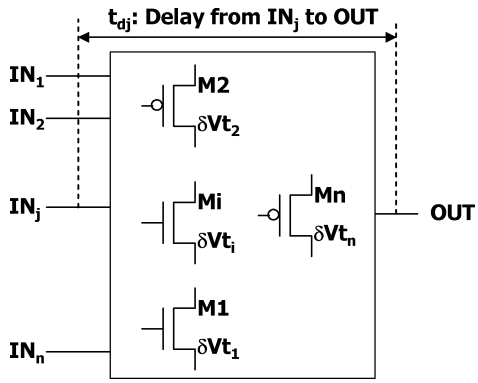


Fig. 1. General circuit of $n$ transistors.

where $T_{dj0}$ is the nominal delay ($T_{dj0} = f(Vt_{10},\ldots,Vt_{n0})$, i.e., delay when $\delta Vt = 0$ for all transistors). These partial derivatives represent the sensitivity of delay to threshold voltage of individual transistors. The analytical evaluation of the nominal delay or the partial derivatives can be obtained using simplified delay models (e.g., Sakurai's model [6]). However, in this work, we have evaluated them numerically using circuit simulator SPICE or device simulator MEDICI, to ensure better accuracy. The partials with respect to $\delta Vt_i$ can be estimated by evaluating $t_{dj1} = f(Vt_{10},\ldots,Vt_{i0}+\Delta,\ldots,Vt_{n0})$ and $t_{dj2} = f(Vt_{10},\ldots,Vt_{i0}-\Delta,\ldots,Vt_{n0})$. Hence, the total number of simulations required is $(1+2n)$ (i.e., a linear complexity). This is considerably less compared to the number of simulations required (i.e., complexity) in a Monte Carlo simulation or response surface based method (e.g., [4]). Evaluating more delay values with respect to $\delta Vt_i$ and use of polynomial curve fitting can further reduce the error in the estimation of the partials. The complexity can be further reduced by analyzing the circuit and eliminating the transistors that do not have a strong impact on $t_{dj}$. This will be helpful to reduce the number of required simulations for complex gates with large number of transistors. We will use this reduction strategy in Section IV to estimate delay distributions in flip-flops. Using the estimated values of the mean and the STD from (3), the PDF of $t_{dj}$ can be approximated as a Gaussian distribution (this approximation is validated in Section III, e.g., see Fig. 4).

There are two possible transitions at the output: Low-to-High (*LH*) and High-to-Low (*HL*). Although the gates may be designed for same low-to-high ($t_{djLH}$) and high-to-low ($t_{djHL}$) delays in the nominal case, under random process variations these two delays can be different. Therefore, the overall delay from $\mathrm{IN}_j$ to output is given by $\boldsymbol{t_{dj}} = \mathbf{Max}(\boldsymbol{t_{djLH}}, \boldsymbol{t_{djHL}})$. The distributions of $t_{djLH}$ and $t_{djHL}$ (approximated as Gaussian) can be individually estimated using (3). Now the goal is to estimate the distribution of $t_{dj}$ from those of $t_{djLH}$ and $t_{djHL}$. Assume (mean, STD) of $t_{djLH}$ and $t_{djHL}$ are $(\mu_1,\sigma_1)$ and $(\mu_2,\sigma_2)$, respectively. Using the distributions of $t_{djLH}$ and $t_{djHL}$, the moments of the distribution of $t_{dj}$ can be calculated as [7]

$$m_0 = 1$$
$$m_1 = \mu_1\Phi(\alpha) + \mu_2\Phi(-\alpha) + a\varphi(\alpha)$$
$$m_2 = (\mu_1^2+\sigma_1^2)\Phi(\alpha) + (\mu_2^2+\sigma_2^2)\Phi(-\alpha) + (\mu_1+\mu_2)a\varphi(\alpha)$$
$$\alpha = \frac{(\mu 1-\mu 2)}{a}; a^2 = \sigma_1^2+\sigma_2^2-2\sigma_1\sigma_2\rho \quad (4)$$

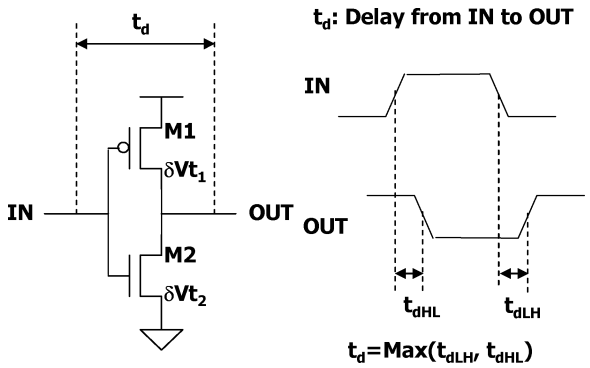


Fig. 2. Inverter and delay definitions.

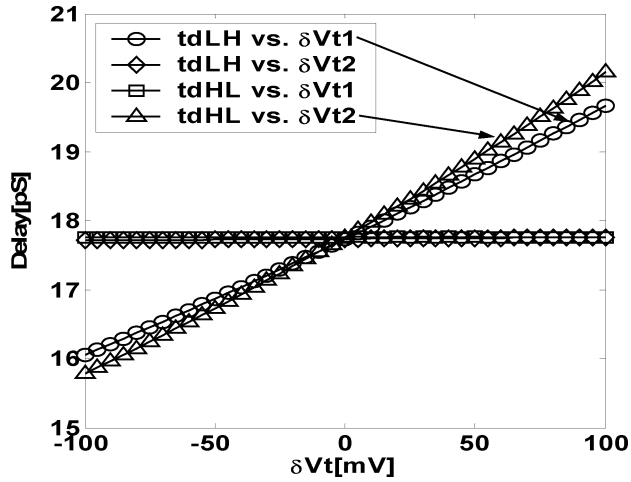


Fig. 3. Delay versus $\delta Vt_i$ for inverter.

where $\varphi(\alpha) = (2\pi)^{-1/2}\exp(-\alpha^2/2)$ [PDF of a standard normal distribution (mean $=$ 0, STD $=$ 1)], $\Phi(\alpha) = \int_{-\infty}^{\alpha}\varphi(t)dt$ (cumulative distribution function of a standard normal distribution), $\rho$ is the correlation coefficient, and $m_k$ is the moment of order $k$. Since both $t_{djLH}$ and $t_{djHL}$ depend on the Vt of the same transistors, they are correlated and cannot be considered as independent random variables. Hence, $\rho$ needs to be considered, and it is estimated as in (5), shown at the bottom of the page. Hence, using (4) and (5) the mean ($\mu_{dj}$) and the STD ($\sigma_{dj}$) of the overall delay $t_{dj}$ can be calculated as [5]

$$\mu_{dj} = m_1 \text{ and } \sigma_{dj}^2 = m_2 - m_1^2. \tag{6}$$

## III. Statistical Delay Models for Logic Gates

Delay distributions of logic gates in a standard cell library can be obtained using the semi-analytical models proposed in Section II. In this section, we present the results for two basic gates, namely, inverter and 2-input NAND, designed using the 70-nm Berkley Predictive Technology Models (BPTM) [8]. Fig. 2 shows an inverter gate and the delay definitions. The inverter is designed for same LH delay ($t_{dLH}$) and HL delay ($t_{dHL}$) in the nominal case ($\delta Vt_1 = \delta Vt_2 = 0$). It can be observed that $\delta Vt$ of the pMOS ($\delta Vt_1$) has a strong impact on $t_{dLH}$ (Fig. 3). On the other hand, $t_{dHL}$ is mainly sensitive to $\delta Vt$ of the nMOS ($\delta Vt_2$) (Fig. 3). The distributions of $t_{dLH}$, $t_{dHL}$, and $t_d$ ($= \text{Max}(t_{dLH}, t_{dHL})$) estimated using the proposed model closely match the distributions obtained by Monte Carlo simulations in SPICE (Fig. 4). It is observed that application of 30% Vt spread ($\sigma_{Vt} = 30\%$ of $\mu_{Vt}$) results in 5% spread (STD/Mean) in the overall delay of an inverter. Increasing the Vt variation results in a larger delay spread.

The proposed model enables us to study the impact of different circuit parameters on delay statistics. The delay distribution is impacted by sizing, output load, input transition (rise/fall) time, supply voltage, and temperature (Figs. 5 and 6). As observed from Fig. 5, the increase in sizing (width) decreases not only the mean and STD of delay but also the relative spread (STD/Mean) of the delay. This is because: 1) the nominal delay decreases (assuming a constant load) and 2) larger transistor size reduces Vt variation [see (1)]. The delay linearly depends on the output load and the input transition time [6]. Therefore, the mean and the STD of delay linearly change with the output load and the input transition time such that the delay spread does not change with these parameters. The delay spread reduces at higher supply voltages and lower temperatures (Fig. 6). To understand this effect, let us consider a simple delay model (assuming short-channel velocity-saturated transistor), given by [2]

$$t_d = \frac{CV_{\text{DD}}}{I_D} = \frac{C}{WC_{\text{ox}}\upsilon_{\text{SAT}}\left(1-\frac{Vt}{V_{\text{DD}}}\right)} \Rightarrow$$
$$\frac{\partial t_d}{\partial V_t} = \frac{C}{WC_{\text{ox}}\upsilon_{\text{SAT}}V_{\text{DD}}\left(1-\frac{Vt}{V_{\text{DD}}}\right)^2}. \tag{7}$$

At higher $V_{\text{DD}}$, the delay sensitivity to Vt ($\partial t_d/\partial Vt$) decreases and therefore the delay spread reduces [see (3)]. Similarly, at a lower temperature, the delay sensitivity to Vt reduces (due to increase in saturation velocity $\upsilon_{\text{SAT}}$ [2]), resulting in reduced delay spread.

The proposed model can also be used to estimate the distributions of output rise/fall time of a logic gate. Fig. 7 shows

$$\rho = \frac{E(t_{djLH}t_{djHL}) - E(t_{djLH})E(t_{djHL})}{\sigma(t_{djLH})\sigma(t_{djHL})} = \frac{E(t_{djLH}t_{djHL}) - \mu_1\mu_2}{\sigma_1\sigma_2}$$
$$E\left(t_{djLH}t_{djHL}\right) = T_{djLH0}T_{djHL0} + \frac{1}{2}\sum_{\text{all transistors}}\frac{\partial^2\left(t_{djLH}t_{djHL}\right)}{\partial(\delta Vt_i)^2}\sigma_{Vt_i}^2 = T_{djLH0}T_{djHL0}$$
$$+\frac{1}{2}\sum_{\text{all transistors}}\left(T_{djLH0}\frac{\partial^2\left(t_{djHL}\right)}{\partial(\delta Vt_i)^2} + 2\frac{\partial t_{djHL}}{\partial(\delta Vt_i)}\frac{\partial t_{djLH}}{\partial(\delta Vt_i)} + T_{djHL0}\frac{\partial^2\left(t_{djLH}\right)}{\partial(\delta Vt_i)^2}\right)\sigma_{Vt_i}^2 \tag{5}$$

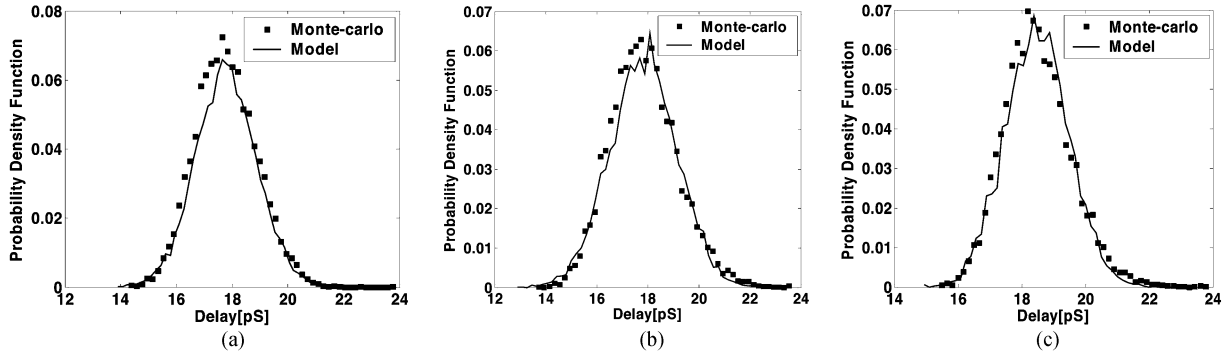


Fig. 4. Model verification: PDF of (a) $t_{dLH}$, (b) $t_{dHL}$, and (c) $t_d = \mathrm{Max}(t_{dLH}, t_{dHL})$ for inverter. ($\sigma Vt0 = 60$ mV is chosen to get a considerable spread in delay distributions; SPICE Monte Carlo simulations are done for 10 000 points).

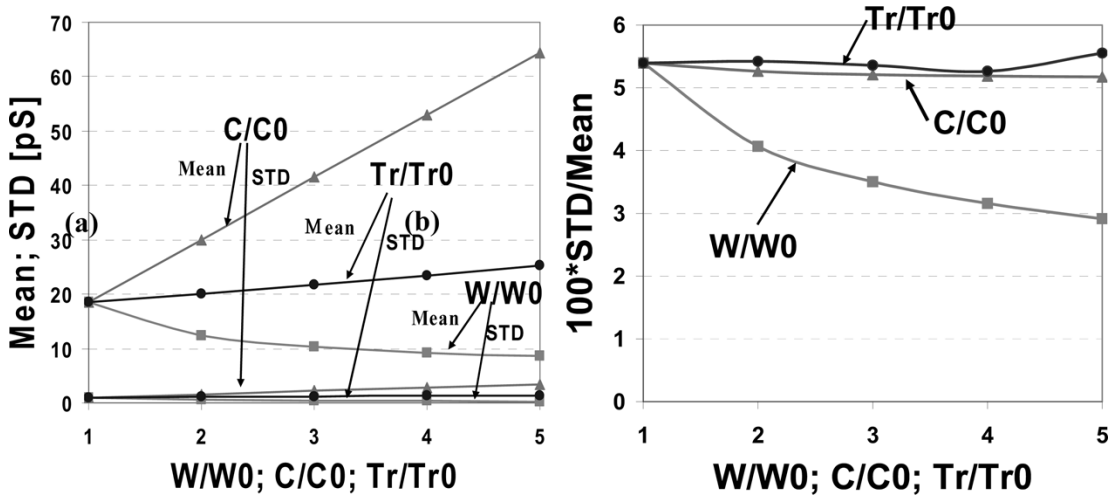


Fig. 5. Impact of sizing (W), output load (C), and input rise/fall time (Tr) on delay PDF of inverter.

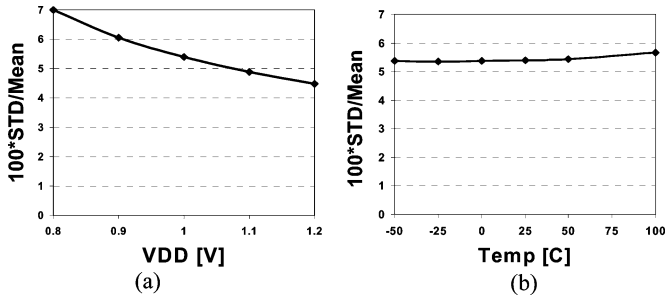


Fig. 6. Impact of (a) supply voltage (VDD) and (b) temperature on delay PDF of inverter.

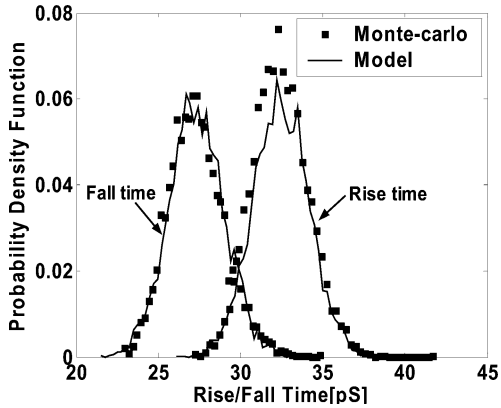


Fig. 7. PDF of output rise/fall time of inverter gate.

the rise/fall time distributions of an inverter estimated using the proposed model. It can be observed that the estimated PDF closely follows the SPICE Monte Carlo simulations. It is observed that 30% Vt spread ($\sigma_{Vt} = 30\%$ of $\mu_{Vt}$) results in 5% spread (STD/Mean) in the rise and 6% spread in the fall time. From Fig. 7, it is observed that the intra-gate Vt variation changes the output transition slope (i.e., rise/fall time) of a gate. On the other hand, the delay distribution of a gate depends on its input transition slope (i.e., rise/fall time). Hence, when a logic gate is driving another logic gate, their delay distributions are not completely independent. They are correlated through the slope of the transition at the intermediate node.

Now, let us consider a NAND gate as shown in Fig. 8. In this case, there are two paths from inputs to output, and therefore two possible delays ($t_{d1}$ and $t_{d2}$). Under nominal conditions, the delay from IN2 to OUT ($t_{d2}$) is expected to be larger [9]. However, under process variations, this may not be true. Therefore, distributions of both $t_{d1}$ and $t_{d2}$ need to be estimated using

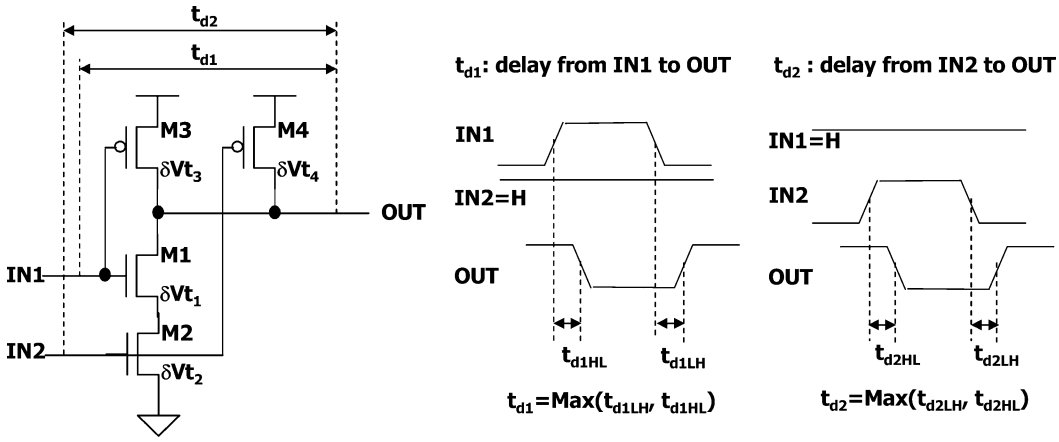


Fig. 8. NAND gate and delay definitions.

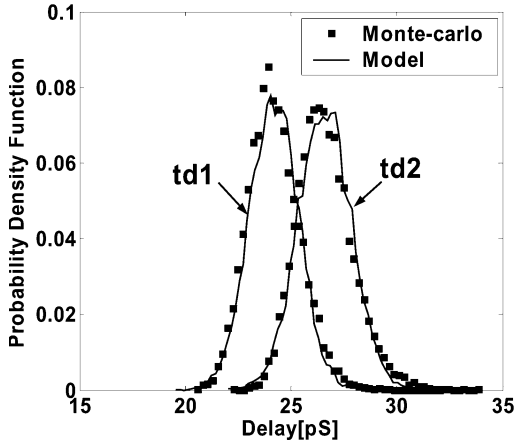


Fig. 9. PDF of $t_{d1}$ and $t_{d2}$ of NAND gate.

the proposed methodology. Fig. 9 shows that the estimated distributions of the delays closely match their PDF obtained from the SPICE Monte Carlo simulations. It is observed that for 30% Vt spread, the delay spread of the NAND gate is 5%.

In the estimation of $t_{d1}$ ($t_{d2}$), we assumed that IN2 (IN1) was stable at high level long before the switching of IN1 (IN2) (Fig. 8). However, in a real circuit where the inputs are provided through other gates and from different paths, there might be a small time difference between the transition events at the two inputs. Let us consider an LH transition at IN2 (IN1) followed by a transition (LH or HL) at IN1 (IN2). Let us assume that the arrival time difference between IN2 and IN1 is $\Delta t$. Thus, $\Delta t > 0$ implies that LH transition at IN2 arrived earlier than the transition at IN1 and delay of interest is from IN1 to output (i.e., $t_{d1}$). Similarly, $\Delta t < 0$ implies that LH transition at IN1 arrived earlier than the transition at IN2 and delay of interest is from IN2 to output (i.e., $t_{d2}$). Using our proposed model in Section II, the impact of $\Delta t$ on delay distributions of the NAND gate is studied (Fig. 10). As $\Delta t$ gets closer to zero, the mean and STD of delay increases because more transistors (both pMOS transistors) can influence the delay. For example, if we assume that LH transition at IN2 arrives long before the arrival of IN1 (i.e., large $\Delta t$), the pMOS M4 (see Fig. 8) is already "off". Hence, it does not influence $t_{d1}$ (i.e., $(\partial t_{d1}/\partial Vt_{\mathrm{M4}}) = 0$). However, if $\Delta t$ is close to zero, then M4 is not completely turned off when IN1 arrives. Thus, the variation in the current through M4 (due to Vt fluctuation) will also impact $t_{d1}$ (i.e., $(\partial t_{d1}/\partial Vt_{\mathrm{M4}}) \neq 0$). The influence of $\Delta t$ on the delay distribution points to the fact that the delay distribution of a gate not only depends on the output transition slopes of the previous gates (as explained earlier) but also the delay of the previous gates (as it changes the arrival time).

## IV. Statistical Delay Models for Flip-Flop

As mentioned in Section II, the proposed methodology can be used for estimating delay distribution of any circuit of $n$ transistors. In this section, we study the impact of process variations on flip-flop delay measures including clock-to-output delay ($t_{cq}$) and setup time ($t_{su}$). Setup time is defined as the minimum time required for the data input (D) to be stable before clock rising edge, so that the data can be correctly captured to the output [9]. Fig. 11 shows the transmission-gate flip-flop (TGFF), which is a static master–slave flip-flop [9], [10]. There are 20 transistors in this flip-flop; however, not all of them can have considerable impact on $t_{cq}$ or $t_{su}$. In order to estimate the distribution of $t_{cq}$, Vt variations of only the transistors that are in the path from the clock (CLK) to the output (Q) need to be considered (only eight transistors as shown in Fig. 11). For estimation of the distribution of $t_{su}$, the critical transistors are those in the path from inputs (CLK and D) to the master latch (only 10 transistors as shown in Fig. 11). The variations of other transistors have much less and almost negligible impact, so they can be neglected in our estimation model (Section II). This is an example of the transistor set reduction strategy mentioned in Section II. Fig. 12 shows that the estimated distributions closely match the distributions extracted by SPICE Monte Carlo simulations. It can be observed that a 30% Vt spread results in 5% spread in the clock-to-output delay. On the other hand, the spread in the setup time is considerably large. A 30% spread in the Vt results in 11% spread in the setup time.

In flip-flops the $t_{cq}$ typically depends on the input data arrival time with respect to the clock rising edge ($\Delta t$) [10]. As the data transition gets closer to the clock rising edge, $t_{cq}$ is initially constant, then it increases, and finally when $\Delta t$ reaches the setup time ($t_{su}$), the flip-flop fails to sample the data correctly.

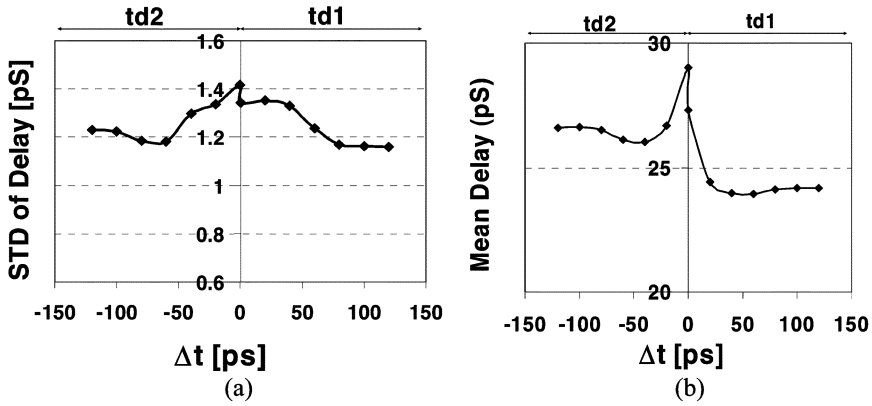


Fig. 10. Impact of input arrival time difference ($\Delta t$) on (a) mean and (b) STD of NAND gate delay.

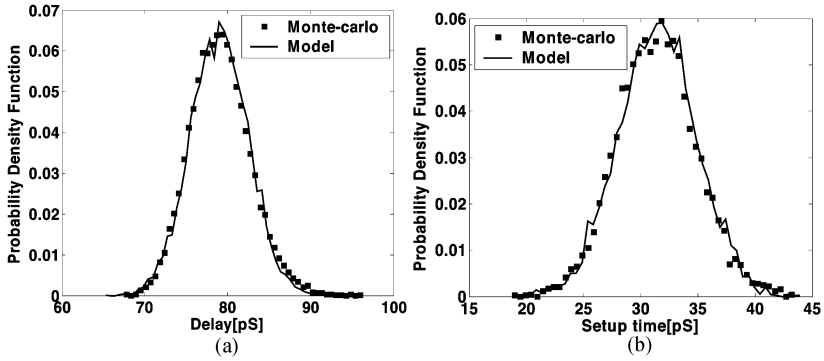


Fig. 12. PDF of (a) $t_{cq}$ and (b) setup time ($t_{su}$) of TGFF.

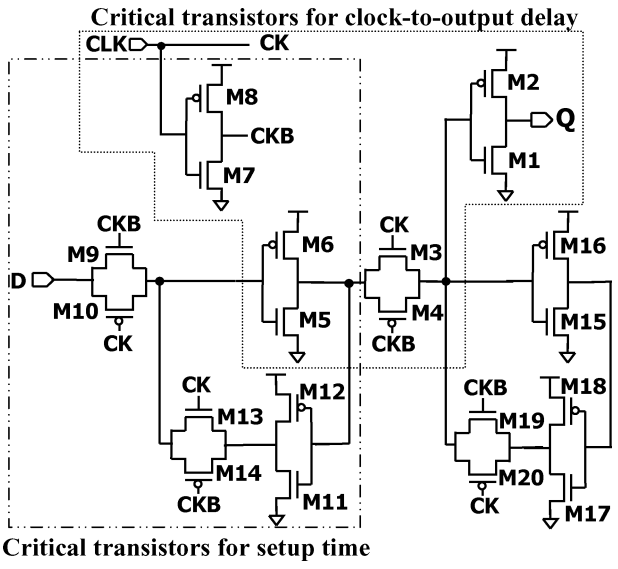


Fig. 11. Transmission gate flip-flop (TGFF).

Using the proposed modeling, the impact of $\Delta t$ on distribution of $t_{cq}$ is studied and plotted in Fig. 13. In addition to increase in both mean and STD of $t_{cq}$ [Fig. 13(a)], the delay spread also increases [Fig. 13(b)] as $\Delta t$ approaches the setup time. It is observed that, depending on the data arrival time, the spread of $t_{cq}$ varies from 5% to 25%.

## V. Effect of Correlation of Threshold Voltages of Different Transistors

In the previous discussions, we have assumed the Vt of different transistors in a logic gate are independent random variables. This assumption is valid if we are considering the Vt variation due only to RDF [3], [13]. However, in general, due to the presence of systematic intra-die variations (e.g., Vt variation due to channel length variation), Vt's of different transistors can be correlated. In this section, the impact of such correlation on delay variations is investigated. The proposed models in Section II [see (3)] can be easily extended to account for the effect of correlation, as shown in (8) at the bottom of the following page [5], where $r_{(i,k)}$ is the correlation coefficient between $\delta Vt_i$ and $\delta Vt_k$. Since all the transistors in a logic gate are in a very close spatial proximity, we can assume the correlation coefficients among different transistors are same, i.e., $r_{(i,k)} = r$ for all $i$ and $k$. Although the proposed model can handle different values of correlation coefficients for different transistors, we have used the above assumption to simplify the calculation. Fig. 14 shows that the mean and the standard deviations estimated using the extended analytical model shown in (8) closely follow the Monte Carlo distributions for $t_{d1HL}$ and $t_{d1LH}$ of a NAND gate (Fig. 8) in the presence of correlation. It is also observed that the effect of the correlation on the mean value of delays is not very significant, whereas it has a stronger impact on the standard deviations of the delay distributions. This is because as observed from (8), correlations are multiplied by second-order derivatives of delay with respect to threshold voltage variables to modify

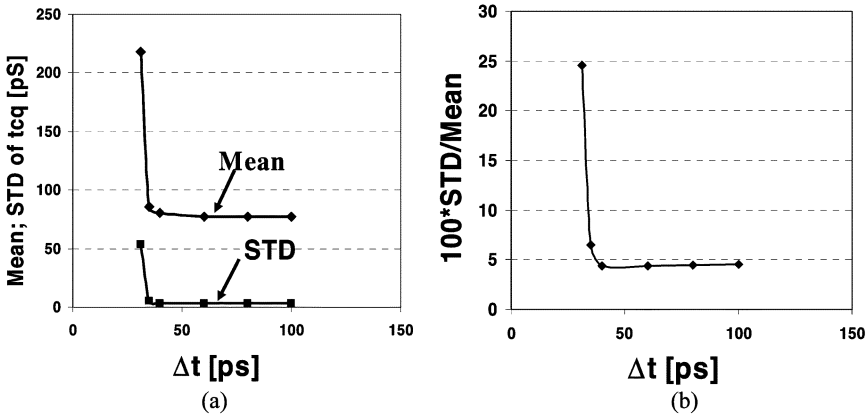


Fig. 13. Impact of input data arrival time ($\Delta t$) on (a) mean and STD and (b) spread of $t_{cq}$ delay of TGFF.

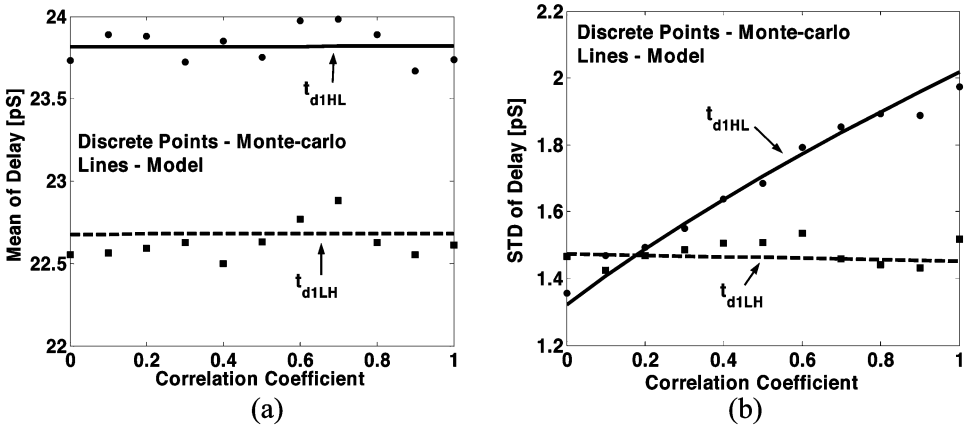


Fig. 14. Impact of Vt correlation on (a) mean and (b) standard deviation of delay distribution of a NAND gate ($t_{d1}$).

the mean of the delay. However, the second-order derivatives of delay with respect to Vt's are very small as the delay has an approximately linear dependence on Vt's (Fig. 3). Another observation from Fig. 14 is that the distribution of the low-to-high delay ($t_{d1LH}$) is not impacted by correlations. The reason is that the low-to-high delay in NAND gates is mostly impacted by only the threshold voltage of a single pMOS transistor. Since $t_{d1LH}$ is sensitive to a single threshold voltage variable, therefore correlations do not impact its distribution. On the other hand, the distribution of the high-to-low delay ($t_{d1HL}$) is sensitive to threshold voltages of both nMOS pull-down transistors (two threshold voltage variables). Therefore, the delay distribution (STD) of $t_{d1HL}$ is impacted by correlations as shown in Fig. 14(b). The STD of $t_{d1HL}$ increases as a result of positive correlation among Vt's. That is because a positive correlation coefficient among Vt's indicate that the threshold voltage of the transistors are more likely to shift in same directions, rather than opposite directions. Therefore, the delay cancellation effect due to Vt shift in opposite directions for the nMOS pull down transistors of the NAND gate reduces, resulting in more delay variations. Similarly, it is observed that in an inverter gate, the distribution of neither the low-to-high ($t_{dHL}$) nor the high-to-low ($t_{dHL}$) delay is affected by Vt correlations. That is because in an inverter gate, $t_{dLH}$ and $t_{dHL}$ are sensitive to a single Vt variable (Fig. 3). The above-mentioned modeling is applicable for considering the impact of correlations on other circuit responses such as rise/fall times and delay of other circuits such as setup and clock-to-output delay of flip-flops.

$$\mu_{tdj} = T_{dj0} + \frac{1}{2}\sum_{i=1}^{n} \frac{\partial^2 t_{dj}}{\partial(\delta Vt_i)^2}\bigg|_{\delta Vt_i=0} \sigma_{Vti}^2 + \sum_{k=1}^{n}\sum_{i=1;i\neq k}^{n} \frac{\partial^2 t_{dj}}{\partial(\delta Vt_i)\partial(\delta Vt_k)}\bigg|_{\delta Vt_i=\delta Vt_k=0} r_{(i,k)}\sigma_{Vti}\sigma_{Vtk}$$

$$\sigma_{tdj}^2 = \sum_{i=1}^{n}\left(\frac{\partial t_{dj}}{\partial(\delta Vt_i)}\bigg|_{\delta Vt_i=0}\right)^2 \sigma_{Vti}^2 + 2\sum_{k=1}^{n}\sum_{i=1;i\neq k}^{n}\left(\frac{\partial t_{dj}}{\partial(\delta Vt_i)}\bigg|_{\delta Vt_i=0}\right)\left(\frac{\partial t_{dj}}{\partial(\delta Vt_k)}\bigg|_{\delta Vt_k=0}\right) r_{(i,k)}\sigma_{Vti}\sigma_{Vtk} \quad (8)$$

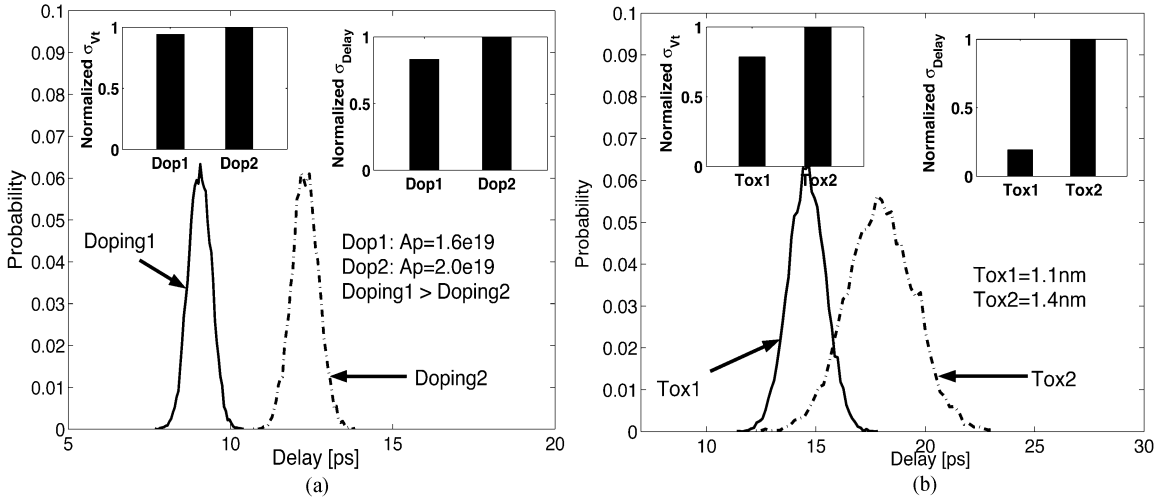


Fig. 15. Impact of (a) doping and (b) oxide thickness on delay distribution of an inverter (designed with 25-nm devices) obtained using the proposed model.

## VI. Estimation of Delay Distribution at Device Design Phase

The low complexity of the proposed model makes it very effective in estimating the impact of device design parameters on the statistical delay of different circuits. In this paper, we have studied the effect of the doping profile and the oxide thickness ($T_{\rm ox}$) on the delay distribution of an inverter designed with predictive 25-nm devices [11]. Device simulator MEDICI was used to estimate the partial derivatives in (3). The designed devices have 2-D nonuniform super halo ("Halo" and "Retrograde") channel doping profile (say, $N_a(x,y)$, approximated as Gaussian function) as given below [11], [12]:

$$\begin{aligned} &x>0,\\ &N_a(x,y)=A_p\Gamma_{xa}(x)\mathrm{K}_{ya}(y)+N_{\rm SUB}\\ &\text{where } \mathrm{K}_{ya}(\mathrm{y})=\exp\left(\frac{-(y-\alpha_a)^2}{{\sigma_{ya}}^2}\right)\\ &\text{and } \left[\begin{array}{l}\Gamma_{xa}(x)=\exp\left(\frac{-(x-\beta_a)^2}{{\sigma_{xa}}^2}\right), \quad 0\le x\le\beta_a\\ \qquad\quad =1, \quad x>\beta_a\end{array}\right] \end{aligned} \tag{9}$$

where $A_p$ represents the peak "halo" doping. $N_{\rm SUB}$ is the constant uniform doping in the bulk and is much less compared to contributions from Gaussian profiles at and near the channel and source/drain regions. Parameters $\alpha_a$ and $\beta_a$ control the positions and $\sigma_{ya}$ and $\sigma_{xa}$ control the variances of the Gaussian profiles in channel regions [11]. The peak halo doping value ($A_p$) was used to modify the doping profile. The effective channel doping ($N_{\rm cheff}$) is calculated using the following method as described in [12]:

$$\begin{aligned} N_{\rm cheff}&=\frac{1}{\Delta_{\rm ch}}\int\int_{\Delta_{\rm ch}}N_a(x,y)dxdy+N_{\rm sub}\\ &=\frac{A_p}{\alpha_a L_{\rm eff}}\int_{x=-L_{\rm eff}/2}^{x=+L_{\rm eff}/2}\Gamma_{xa}(x)dx\\ &\quad\times\int_{y=0}^{y=\alpha_a}\mathrm{K}_{ya}(y)dy+N_{\rm sub} \end{aligned} \tag{10}$$

where $\Delta_{\rm ch}=\alpha_a L_{\rm eff}$ is the area of the channel region which is under the influence of gate. To calculate the effective doping, we have assumed that most of the depletion charge is confined in the region $y=0$ to $y=\alpha_a$.. Using the estimated value of the effective doping density from (10), the width of the depletion region ($W_d$) can be calculated as [2]

$$W_{dm}=\sqrt{\frac{2\varepsilon_{si}\left(2\psi_{\rm B}-V_{\rm BS}\right)}{qN_{\rm cheff}}} \tag{11}$$

where $\psi_{\rm B}$ is the Fermi potential in bulk ($=2k_{\rm B}Tln(N_{\rm cheff}/n_i)$) and $V_{\rm BS}$ is the substrate bias. Using the expression of $W_{dm}$ (with $V_{\rm BS}=0$) given in (11) into (1), the standard deviation of the Vt variation due to RDF is given by

$$\sigma_{Vt0}=\frac{qT_{\rm ox}}{\varepsilon_{\rm ox}}\frac{\left(\left(2\varepsilon_{si}/q\right)\left(4k_{\rm B}Tln\left(N_{\rm cheff}/n_i\right)\right)N_{\rm cheff}\right)^{1/4}}{\sqrt{3L_{\rm MIN}W_{\rm MIN}}}. \tag{12}$$

From (12), it can be observed that $\sigma_{Vt0}$ increases with an increase in the doping density and the oxide thickness. Let us now analyze the effect of doping density and the oxide thickness on the delay distribution. For the analysis, we can re-use the simple delay equation given in (7) to estimate the high-to-low delay of an inverter. Using (7), the high-to-low delay of the inverter is given by

$$t_{dHL}=\frac{C_L V_{\rm DD}}{I_D}=\frac{C_L T_{\rm ox}}{W\varepsilon_{\rm ox}\upsilon_{\rm SAT}\left(1-\frac{Vt_N}{V_{\rm DD}}\right)} \tag{13a}$$

$$\Rightarrow\frac{\partial t_{dHL}}{\partial Vt_N}=\frac{C_L T_{\rm ox}}{W\varepsilon_{\rm ox}\upsilon_{\rm SAT}V_{\rm DD}\left(1-\frac{Vt_N}{V_{\rm DD}}\right)^2}. \tag{13b}$$

Increasing $A_p$ increases the effective doping density, thereby increasing the Vt variation of due to RDF (i.e., $\sigma_{Vt0}$) [see (10) and (12), and Fig. 15(a)]. Increasing the effective density has three impacts on the delay distribution due to RDF for an inverter. First, a higher doping increases the Vt of the transistor, thereby increasing the nominal and mean delay [see Fig. 13(a)] [2]. Second, increasing the Vt of the transistor increases the sensitivity of the delay to Vt [see Fig. 13(b)] which increases the standard deviation of the delay variation [see (3)]. Finally, a higher doping results in a higher value of $\sigma_{Vt}$ (as $\sigma_{Vt0}$ increases), which also increases the standard deviation of the delay variation [see (3)]. Hence, increasing doping increases both the mean and the standard deviation of the delay variation as shown in Fig. 15(a). However, STD of delay does not

strongly depend on the doping, as $\sigma_{Vt0}$ is a weak function of doping.

Increasing oxide thickness ($T_{\rm ox}$) increases the Vt variation due to RDF (i.e., $\sigma_{Vt0}$) [see (10) and (12), and Fig. 15(b)]. Increasing the oxide thickness has several impacts on the delay distribution of a transistor. First, increasing the oxide thickness reduces the current through a transistor, thereby increasing the nominal and the mean value of the delay [see Fig. 13(a) assuming a constant load] [2]. Second, increasing the oxide thickness also increases the sensitivity of the delay to Vt [see Fig. 13(b)] which increases the standard deviation of the delay variation [see (3), assuming a constant load]. Third, a higher $T_{\rm ox}$ results in a higher value of $\sigma_{Vt}$ (as $\sigma_{Vt0}$ increases), which also increases the standard deviation of the delay variation [see (3)]. Finally, increasing $T_{\rm ox}$ also impacts the Vt of a transistor [2]. The simplified expression for the threshold voltage of a short-channel device can be given by [2], [12], [14], [15]

$$V_t = \underbrace{V_{fb} + \psi_{\rm B} + \frac{T_{\rm ox}\sqrt{2q\varepsilon_{si}N_{\rm cheff}\psi_{\rm B}}}{\varepsilon_{\rm ox}}}_{Vt(\text{long-channel})} - \underbrace{\left[[2(V_{bi}-\psi_{\rm B})+V_{ds}]\left[e^{-L_{\rm eff}/2l_c}+2e^{-L_{\rm eff}/l_c}\right]\right]}_{\Delta Vt(\text{short-channel})}$$

$$\text{where } l_C = \sqrt{\frac{\varepsilon_{si}W_dT_{\rm ox}}{\varepsilon_{\rm ox}\varsigma}}. \tag{14}$$

From (14) it can be observed that the long-channel threshold voltage of a device increases with an increase in $T_{\rm ox}$ [2], [14], [15]. However, the negative Vt shift due to the short-channel effect increases with an increase in $T_{\rm ox}$ [2], [14], [15]. The net effect of $T_{\rm ox}$ depends on the strength of the short-channel effect. For the 25-nm device with the "super halo" doping used in this analysis, we observed that Vt increases by a small amount with the increase in the oxide thickness from 1.1 to 1.4 nm. The increase in Vt further increases the STD of the delay distribution at a higher oxide thickness. Hence, increasing the oxide thickness increases both the mean and the standard deviation of the delay variation as shown in Fig. 15(b). Due to the linear dependence of $\sigma_{Vt0}$ on the oxide thickness, increasing the $T_{\rm ox}$ has a stronger impact on the STD of the delay compared to increasing the doping.

## VII. Conclusion

With technology scaling, the effect of the random dopant fluctuation (RDF) on the threshold voltage of a transistor increases, resulting in variation in the delay of different logic gates. In this paper, we have modeled and analyzed the effect of the RDF-induced Vt variation on the delay of different logic gates. We have proposed semi-analytical models to predict the delay distributions in combinational and sequential logic circuits. It is observed that the RDF-induced Vt variation results in significant variation in the delay of logic gates. However, the effect is much stronger in the setup time of a flip-flop. Moreover, the proposed models can be effectively used to estimate delay distributions at both the circuit and the device design phase. It is observed that device design parameters such as doping profile and oxide thickness have a strong impact on the delay variation in logic gates due to RDF. The analysis presented in this paper shows that the Vt variation due to RDF has a strong effect in different delay parameters of logic gates and flip-flops. Hence, the effect of RDF needs to be considered in the design phase of both the devices and the logic circuits (gates and flip-flops) in nanoscale CMOS. The proposed models are very helpful for estimating the effect of Vt variation due to RDF on the delay of logic circuits.

## References


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

[2] Y. Taur and T. H. Ning, *Fundamentals of Modern VLSI Devices*. New York: Cambridge Univ. Press, 1998.

[3] S. Borkar, T. Karnik, S. Narendra, J. Tschanz, A. Keshavarzi, and V. De, "Parameter variation and impact on circuits and microarchitecture," in *Proc. Design Automation Conf.*, 2003, pp. 338–342.

[4] K. Okada, K. Yamaoka, and H. Onodera, "A statistical gate-delay model considering intra-gate variability," in *Proc. Int. Conf. Computer Aided Design*, Nov. 2003, pp. 908–913.

[5] A. Papoulis, *Probability, Random Variables and Stochastic Process*. New York: MacGraw-Hill, 2002.

[6] T. Sakurai and A. R. Newton, "A simple short-channel MOSFET model and its application to delay analysis of inverters and series-connected MOSFETs," in *Proc. IEEE Int. Symp. Circuits and Systems*, May 1990, pp. 105–108.

[7] C. E. Clark, *Oper. Res.*, vol. 9, no. 2, pp. 145–162, 1961.

[8] Berkeley Predictive Technology Model [Online]. Available: http://www-device.eecs.berkeley.edu/~ptm/

[9] J. M. Rabaey, *Digital Integrated Circuits*. Englewood Cliffs, NJ: Prentice Hall, 1996.

[10] B. Nikolic *et al.*, "Improved sense-amplifier-based flip-flop: Design and measurements," *IEEE J. Solid-State Circuits*, vol. 35, no. 6, pp. 876–884, Jun. 2000.

[11] "Well-Tempered" Bulk-Si NMOSFET Device Home Page (2001). [Online]. Available: http://www-mtl.mit.edu:/researchgroups/Well/

[12] S. Mukhopadhyay, A. Raychowdhury, and K. Roy, "Accurate estimation of total leakage current in scaled CMOS logic circuits based on compact current modeling," in *Proc. Design Automation Conf.*, Jun. 2003, pp. 169–174.

[13] S. R. Nassif, "Modeling and analysis of manufacturing variations," in *Proc. Custom Integrated Circuit Conf.*, 2001, pp. 223–228.

[14] K. Roy and S. C. Prasad, *Low-Power CMOS VLSI Circuit Design*. New York: Wiley, 2000.

[15] D. Fotty, *MOSFET Modeling With SPICE*. Englewood Cliffs, NJ: Prentice-Hall, 1997.



**Hamid Mahmoodi** (S'00) received the B.S. degree in electrical engineering from Iran University of Science and Technology, Tehran, Iran, in 1998, and the M.S. degree in electrical and computer engineering from the University of Tehran, Iran, in 2000. He is currently pursuing the Ph.D. degree in electrical and computer engineering at Purdue University, West Lafayette, IN.

His research interests include low-power, robust, and high-performance circuit design for nanoscale bulk CMOS and SOI technologies. He has more than 35 refereed publications in journals and conferences.

Mr. Mahmoodi was a recipient of the Best Paper Award of the 2004 International Conference on Computer Design.

**Saibal Mukhopadhyay** (S'99) received the B.E. degree in electronics and telecommunication engineering from Jadavpur University, Calcutta, India, in 2000. He is currently pursuing the Ph.D. degree in electrical and computer engineering at Purdue University, West Lafayette, IN.

He was an intern with the IBM T. J. Watson Research Lab, Yorktown Heights, NY, in summer of 2003 and 2004, in the High Performance Circuit Design Department. His research interests include analysis and design of low-power and robust circuits using nanoscaled CMOS and circuit design using double gate transistors.

Mr. Mukhopadhyay received the IBM Fellowship Award for 2004–2005.

**Kaushik Roy** (SM'95–F'01) received the B.Tech. degree in electronics and electrical communications engineering from the Indian Institute of Technology, Kharagpur, India, and the Ph.D. degree in electrical and computer engineering from the University of Illinois at Urbana-Champaign in 1990.

He was with the Semiconductor Process and Design Center of Texas Instruments Inc., Dallas, TX, where he worked on FPGA architecture development and low-power circuit design. He joined the electrical and computer engineering faculty at Purdue University, West Lafayette, IN, in 1993, where he is currently a Professor and University Faculty Scholar. His research interests include VLSI design/CAD for nanoscale silicon and non-silicon technologies, low-power electronics for portable computing and wireless communications, VLSI testing and verification, and reconfigurable computing. He has published more than 300 papers in refereed journals and conferences, holds eight patents, and is a coauthor of two books on low power CMOS VLSI design. He is the Chief Technical Advisor of Zenasis Inc. and Research Visionary Board Member of Motorola Labs (2002).

Dr. Roy received the National Science Foundation Career Development Award in 1995, an IBM faculty partnership award, a ATT/Lucent Foundation award, and best paper awards at 1997 International Test Conference, IEEE 2000 International Symposium on Quality of IC Design, 2003 IEEE Latin American Test Workshop, and 2003 IEEE Nano. He has been on the editorial boards of *IEEE Design and Test*, IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS, and IEEE TRANSACTIONS ON VLSI SYSTEMS. He was Guest Editor for the Special Issue on Low-Power VLSI, *IEEE Design and Test* (1994), IEEE TRANSACTIONS ON VLSI SYSTEMS (June 2000), and *IEE Proceedings—Computers and Digital Techniques* (July 2002).