Estimation of Delay Variations Due to Random-Dopant Fluctuations in Nano-Scaled CMOS Circuits*  

Hamid Mahmoodi-Meimand, Saibal Mukhopadhyay, and Kaushik Roy  
Dept. of ECE, Purdue University, West Lafayette, IN-47907, USA  
<mahmoodi, sm, kaushik>@ecn.purdue.edu

Abstract  
In nano-scaled CMOS circuits the random dopant fluctuations cause significant threshold voltage (Vt) variations in transistors. In this paper, we propose a semi-analytical estimation methodology to predict the delay distribution (mean and standard deviation) of logic circuits considering Vt variation in transistors. The proposed method is fast and can be used to predict delay distribution in nano-scaled CMOS technologies both at the circuit and the device design phase.

1. Introduction  
In nano-scaled CMOS devices, the random variations in number and placement of dopant atoms in the channel region cause random variations in the transistor threshold voltage (Vt) [1-3], known as “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]. Moreover, the delay distribution of a gate 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 gate is necessary both at the circuit and device design phase to enhance the yield of logic circuits in nano-meter regimes. Although the Monte-Carlo simulation of gates is accurate (e.g. using circuit simulator like SPICE during circuit design and device simulator like MEDICI during device design) 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. Particularly, in this work:

- We have developed, a general semi-analytical method to predict the mean, the standard deviation (STD) and the Probability Distribution Function (PDF) of delay in logic circuits considering random 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.

2. Vt Variation Due to Random Dopant Fluctuation (RDF)  
The Vt variations (6Vt) (due to random dopant fluctuation) of different transistors in a circuit are considered as independent Gaussian random variables (mean=O) [1]. The standard deviation (6Vt) depends on the manufacturing process, doping profile, and the geometry and doping profile.

\[
\sigma_{Vt} = \frac{qT_{ox}}{c_{ox}} \left( \frac{N_{a}W_{d}}{3L_{min}W_{min}} \right) \times \frac{L_{min}W_{min}}{LW} = \sigma_{Vt0} \times \frac{L_{min}W_{min}}{LW} \quad (1)
\]

where, \(N_{a}\) is the effective channel doping, \(W_{d}\) is the depletion region thickness, \(T_{ox}\) is the oxide thickness, and \(L_{min}\) and \(W_{min}\) are the minimum channel length and width, respectively. During the estimation of delay distribution at the circuit level, we use \(\sigma_{Vt0}\) as an input parameter.

3. Estimation of Delay Distribution  
Let us consider a general logic gate with n transistors (Fig. 1). In general, the propagation delay from input \(I_{N}\) to output \(O_{P}\) depends on the Vt of all n transistors (i.e. \(V_{t}\)) in the gate. Hence, considering the Vt fluctuation of each transistor \(\delta V_{t}\) from their nominal values \(V_{t0}\), \(\delta t_{g}\) can be written as:

\[
\delta t_{g} = f(V_{t1},...,V_{tn}) = f(V_{t0} + \delta V_{t1},...,V_{t0} + \delta V_{tn}) \quad (2)
\]

Since the Vt fluctuation in different transistors due to RDF is independent of each other, \(\delta V_{t1},...,\delta V_{tn}\) are considered as independent Gaussian random variables with zero mean, and STD \(\sigma_{Vt}\) given by (1). Expanding \(\delta t_{g}\) in multi-variable Taylor series for the variables \(\delta V_{t1},...,\delta V_{tn}\) around their mean (=0), the mean \((\mu_{t})\) and STD \((\sigma_{t})\) of delay can be expressed as:

\[
\mu_{t} = \delta t_{g0} + \frac{1}{2} \sum_{\text{all transitions}} \left[ \frac{\partial^{2} \delta t_{g}}{\partial(V_{t1})^{2}} \right]_{V_{t0}} \sigma_{Vt1}^{2}
\]

\[
\sigma_{t}^{2} = \sum_{\text{all transitions}} \left[ \frac{\partial^{2} \delta t_{g}}{\partial(V_{t1})^{2}} \right]_{V_{t0}} \sigma_{Vt1}^{2} \quad (3)
\]

where \(\delta t_{g0}\) is the nominal delay (\(\delta t_{g0}=f(V_{t0},...,V_{t0})\) i.e. delay when \(\delta V_{t}=0\) 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 \(\delta V_{t1}\) can be estimated by evaluating \(\delta t_{g1}=f(V_{t0},...,V_{t0}+\delta V_{t1},...,V_{t0})\) and \(\delta t_{g}=f(V_{t0},...,V_{t0}+\delta V_{t1},...,V_{t0})\). 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 (Fig. [4]). Evaluating more delay values with respect to \(\delta V_{t1}\) 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 \(\delta t_{g}\). 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 5 to estimate delay distributions in flip-flops. Using the estimated values of the mean and the STD from (3), the PDF of \(\delta t_{g}\) can be approximated as a Gaussian distribution (this approximation is validated in section 4 e.g. 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 (\(t_{g}^{lh}\)) and high-to-low (\(t_{g}^{hl}\)) delays in the nominal case, under random process variations these two delays can be different. Therefore, the overall delay from \(I_{N}\) to output is given by:

\(t_{g}=\text{Max}(t_{g}^{lh},t_{g}^{hl})\). The distributions of \(t_{g}^{lh}\) and \(t_{g}^{hl}\) (approximated as Gaussian) can be individually estimated using (3). Now the goal is to estimate the distribution of \(t_{g}\) from those of \(t_{g}^{lh}\) and \(t_{g}^{hl}\). Assume

*This work is supported in part by Semiconductor Research Corp. (1078.001), Gigascale System Research Center (MARCO), Intel and IBM Corp.

17

Authorized licensed use limited to: San Francisco State Univ. Downloaded on December 10, 2008 at 20:54 from IEEE Xplore. Restrictions apply.
where \( p(a) = \frac{1}{\sqrt{2\pi}} \exp(-\frac{a^2}{2}) \) (PDF of a standard normal distribution).

Using the distributions of \( t_{dLH} \) and \( t_{dHL} \), the moments of the distribution of \( t_d \) can be calculated as [7]:

\[
m_0 = 1
\]

\[
m_1 = \mu_1 \Phi(a) - \mu_2 \Phi(-a) + \alpha \varphi(a)
\]

\[
m_2 = (\mu_1 + 2\sigma_1)^2 \varphi(a) + (\mu_2 + 2\sigma_2)^2 \varphi(-a) + (\mu_1 + \mu_2 + \alpha) \varphi(a)
\]

\[
\alpha = (\mu_1 - \mu_2) / \sigma_1 - 2 \sigma_2 \sigma_1 / \sigma_1 \rho
\]

where \( \varphi(a) = (2\pi)^{-1/2} \exp(-a^2/2) \) (PDF of a standard normal distribution) and \( m_k \) is the moment of order \( k \). Since both \( t_{dLH} \) and \( t_{dHL} \) depend on the \( \text{Vt} \) 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:

\[
\rho = \frac{E(t_{dLH} t_{dHL}) - E(t_{dLH}) E(t_{dHL})}{\sigma(t_{dLH}) \sigma(t_{dHL})}
\]

Function of a standard normal distribution, \( \rho \) is the correlation coefficient and \( m_k \) is the moment of order \( k \). Since both \( t_{dLH} \) and \( t_{dHL} \) depend on the \( \text{Vt} \) 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:

\[
\rho = \frac{E(t_{dLH} t_{dHL}) - E(t_{dLH}) E(t_{dHL})}{\sigma(t_{dLH}) \sigma(t_{dHL})}
\]

\[
E(t_{dLH} t_{dHL}) = T_{dLH} T_{dHL} + \frac{1}{2} \sum_{n=0}^{\infty} \frac{(-1)^n}{n!} T_{dLH} T_{dHL} \left( \frac{\partial^2}{\partial \text{Vt}_1 \partial \text{Vt}_2} \right)^n
\]

\[
\sigma(t_{dLH} t_{dHL}) = T_{dLH} T_{dHL} + \frac{1}{2} \sum_{n=0}^{\infty} \frac{(-1)^n}{n!} T_{dLH} T_{dHL} \left( \frac{\partial^2}{\partial \text{Vt}_1 \partial \text{Vt}_2} \right)^n
\]

Hence, using (4) and (5) the mean \( \mu_d \) and the STD \( \sigma_d \) of the overall delay \( t_d \) can be calculated as [5]:

\[
\mu_d = m_0
\]

\[
\sigma_d^2 = m_2 - m_1^2
\]

4. Statistical Gate Delay Model

Delay distributions of the gates in the standard cell library can be obtained using the proposed models. In this section we present the results for two basic gates, namely, inverter and 2-input NAND, designed using the 70nm Berkeley 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 \( t_{dHL} \), two basic gates, namely, inverter and 2-input NAND, designed using the 70nm Berkley Predictive Technology Models (BPTM) [8].

The inverter is designed for same LH delay \( t_{dLH} \) and HL delay \( t_{dHL} \), two basic gates, namely, inverter and 2-input NAND, designed using the 70nm Berkley Predictive Technology Models (BPTM) [8].

Fig. 1. General circuit of \( n \) transistors

Fig. 2. Inverter and delay definitions

Fig. 3. Delay vs. \( \text{Vt} \) for inverter

in the nominal case \( \delta \text{Vt} = \delta \text{Vt}_f = 0 \). It can be observed that, \( \delta \text{Vt} \) of the PMOS \( \delta \text{Vt}_f \) has a strong impact on \( t_{dLH} \) (Fig. 3). On the other hand, \( t_{dHL} \) is mainly sensitive to \( \delta \text{Vt} \) of the NMOS \( \delta \text{Vt}_n \). The distributions of \( t_{dLH}, t_{dHL} \) and \( t_f = \text{Max}(t_{dLH}, t_{dHL}) \) estimated using the proposed model closely match with the distributions obtained by Monte-Carlo simulations in SPICE (Fig. 4).

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 (Fig. 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 (a) the nominal delay decreases (assuming a constant load) and (b) larger transistor size reduces \( \text{Vt} \) 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]:

\[
t_f = \frac{C}{\text{I}_o} = \frac{C}{W \text{C}_o \text{V}_{II} \text{II}} \frac{\partial \text{I}_o}{\partial \text{Vt}} = \frac{C}{W \text{C}_o \text{V}_{II} \text{II}} \frac{\partial \text{I}_o}{\partial \text{Vt}}
\]

At higher \( \text{VDD} \) the delay sensitivity to \( \text{Vt} \) (\( \partial \text{I}_o / \partial \text{Vt} \)) decreases and therefore the delay spread reduces (see (3)). Similarly at a lower temperature, the delay sensitivity to \( \text{Vt} \) reduces (due to increase in saturation velocity, \( \text{V}_{II} \)) [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 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. From Fig. 7 it is observed that the intragate \( \text{Vt} \) 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

Fig. 4. Model verification: PDF of (a) \( t_{dLH} \), (b) \( t_{dHL} \), and (c) \( t_f = \text{Max}(t_{dLH}, t_{dHL}) \) for inverter. (\( \sigma \text{VDD} = 60\text{mV} \) is chosen to get a considerable spread in delay distributions; Spice monte-carlo simulations are done for 10000 points).
Fig. 5. Impact of sizing (W), output load (C), and input rise/fall time (Tr) on delay PDF of inverter

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 the proposed methodology. Fig. 9 shows that the estimated distributions of the delays closely match their PDF obtained from the SPICE Monte-Carlo simulations.

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 just a small time difference between the transition events at the two inputs. Let us consider a 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 3, 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. $\partial t_{d1}/\partial V_{tM4}$ = 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 $V_t$ fluctuation) will also impact $t_{d1}$ (i.e. $\partial t_{d1}/\partial V_{tM4}$ ≠ 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).

5. Statistical Flip-Flop Delay Model

As mentioned in Section 2, the proposed methodology can be used for estimating delay distribution of any circuit of $n$ transistors. In this case, the PDF of output rise/fall time of inverter gate is estimated using Monte-Carlo simulations. Fig. 7 shows the estimated PDF of rise and fall times, and also the PDF obtained using our proposed model. Similarly, the PDF of $t_{d1}$ and $t_{d2}$ for the NAND gate is shown in Fig. 9. As $\Delta t$ gets closer to zero, the mean and STD of delay increases because more transistors (both PMOS transistors) can influence the delay.
have considerable impact on the path from the clock (CLK) to the output (Q) need to be considered (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 be critical transistors for setup time (\(t_{su}\)) that are in the path from inputs (CLK and D) to the master latch (only 10 transistors as shown in Fig. 11). For estimation of 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 that they can be neglected in our estimation model (Section 3).

In flip-flops the \(t_{q}\) 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, the \(t_{q}\) is initially constant, then it increases, and finally when \(\Delta t\) reaches the setup time (\(t_{su}\)), the flip-flop fails to sample that data correctly. Using the proposed modeling, the impact of \(\Delta t\) on distribution of \(t_{q}\) is studied and plotted in Fig. 13. In addition to increase in both mean and STD of \(t_{q}\) (Fig. 13(a)), the delay spread also increase (Fig. 13(b)), as \(\Delta t\) approaches the setup time.

6. 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 on (a) mean and STD of the delay of TGFF (Fig. 14(a)), the effective channel doping thereby increasing delay. The STD of the delay increases in thick oxide devices (Fig. 14(b)). Hence, the STD of delay increases in thick oxide devices (Fig. 14(b)).

7. Conclusion

In this paper we have developed a semi-analytical model to predict the delay distributions in circuits considering Vt variation in the transistors due to random dopant fluctuations. The proposed models can be effectively used to estimate delay distributions both at the circuit and the device design phase.

References

[5]. A. Papoulis, Probability, Random Variables and Stochastic Process
[8]. Berkeley Predictive Technology Model, www.device.eecs.berkeley.edu/