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 (δVt) (due to random dopant fluctuation) of different transistors in a circuit are considered as independent Gaussian random variables (mean=0) [1]. The standard deviation (σ_{Vt}) depends on the manufacturing process, doping profile, and the transistor size, and is given by:

$$\sigma_{Vt} = \left[\frac{qT_{ox}}{\varepsilon_{ox}} \sqrt[4]{\frac{(N_a W_d)}{3L_{min}W_{min}}} \right] \times \sqrt{\frac{L_{min}W_{min}}{LW}} = \sigma_{Vt0} \times \sqrt{\frac{L_{min}W_{min}}{LW}} \qquad (1)$$

where, N_a is the effective channel doping, W_d is the depletion region width, T_{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 σ_{Vt0} as an input parameter. In section 6, we have described the impact of variation in N_a and T_{ox} (hence σ_{Vt0}) on the delay distribution.

3. Estimation of Delay Distribution
Let us consider a general logic gate with n transistors (Fig. 1). In general, the propagation delay from input IN_j to output (t_{dj}) depends on the Vt of all n transistors (i.e. Vt_i) in the gate. Hence, considering the Vt fluctuation of each transistor (δVt_i) from their nominal values (Vt_{i0}), t_{dj} can be written as:

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

Since the Vt fluctuation in different transistors due to RDF is independent of each other, $\delta Vt_1,\ldots, \delta Vt_n$ are considered as independent Gaussian random variables with zero mean, and STD (σ_{Vti}) 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 (μ_{tdj}) and STD (σ_{tdj}) of delay can be expressed as [5]:

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

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 (eg. 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 δVt_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 δVt_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 5 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 4 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 INj to output is given by: $t_{dj}=\mathrm{Max}(t_{djLH}, t_{djHL})$. 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

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

Authorized licensed use limited to: San Francisco State Univ. Downloaded on December 10, 2008 at 20:54 from IEEE Xplore. Restrictions apply.

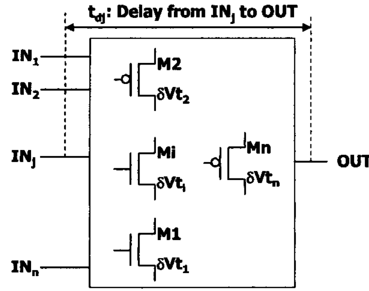

Fig. 1. General circuit of *n* transistors

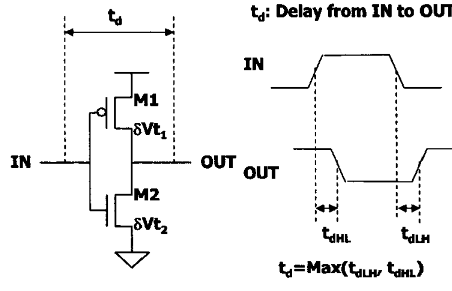

Fig. 2. Inverter and delay definitions

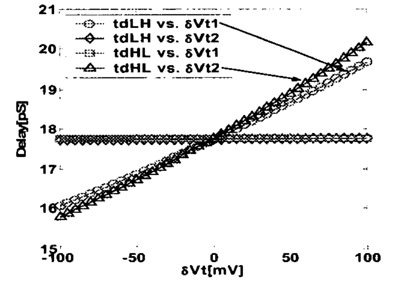

Fig. 3. Delay vs. δVt_i for inverter

(mean, STD) of t_{djLH} and t_{djHL} are (μ_1,σ_1) and (μ_2,σ_2), respectively. Using the distributions of t_{djLH} and t_{djHL}, the moments of the distribution of t_{dj} can be calculated as [7]:

$$\begin{aligned} m_0 &= 1 \\ m_1 &= \mu_1\Phi(\alpha)+\mu_2\Phi(-\alpha)+a\varphi(\alpha) \\ m_2 &= \left(\mu_1^2+\sigma_1^2\right)\Phi(\alpha)+\left(\mu_2^2+\sigma_2^2\right)\Phi(-\alpha)+\left(\mu_1+\mu_2\right)a\varphi(\alpha) \\ \alpha &= (\mu 1-\mu 2)/a \;;\; a^2=\sigma_1^2+\sigma_2^2-2\sigma_1\sigma_2\rho \end{aligned} \tag{4}$$

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), ρ 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, ρ needs to be considered and it is estimated as:

$$\rho=\frac{E\left(t_{djLH}t_{djHL}\right)-E\left(t_{djLH}\right)E\left(t_{djHL}\right)}{\sigma\left(t_{djLH}\right)\sigma\left(t_{djHL}\right)}=\frac{E\left(t_{djLH}t_{djHL}\right)-\mu_1\mu_2}{\sigma_1\sigma_2}$$

$$\begin{aligned} E\left(t_{djLH}t_{djHL}\right) &= T_{djLH0}T_{djHL0}+\frac{1}{2}\sum_{\text{all transistors}}\frac{\partial^2\left(t_{djLH}t_{djHL}\right)}{\partial(\delta Vt_i)^2}\sigma_{Vti}^{\;2}=T_{djLH0}T_{djHL0}+ \\ &\frac{1}{2}\sum_{\text{all transistors}}\left(T_{djLH0}\frac{\partial^2\left(t_{djHL}\right)}{\partial(\delta Vt_i)^2}+2\frac{\partial t_{djHL}}{\partial(\delta Vt_i)}\frac{\partial t_{djLH}}{\partial(\delta Vt_i)}+T_{djHL0}\frac{\partial^2\left(t_{djLH}\right)}{\partial(\delta Vt_i)^2}\right)\sigma_{Vti}^{\;2} \end{aligned} \tag{5}$$

Hence, using (4) and (5) the mean (μ_{dj}) and the STD (σ_{dj}) of the overall delay t_{dj} can be calculated as [5]:

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

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 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 (δVt_1=δVt_2=0). It can be observed that, δVt of the PMOS (δVt_1) has a strong impact on t_{dLH} (Fig. 3). On the other hand, t_{dHL} is mainly sensitive to δVt of the NMOS (δVt_2) (Fig. 3). The distributions of t_{dLH}, t_{dHL}, and t_d (=Max(t_{dLH}, t_{dHL})) estimated using the proposed model closely matche 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 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]:

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

At higher V_{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, υ_{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 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 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

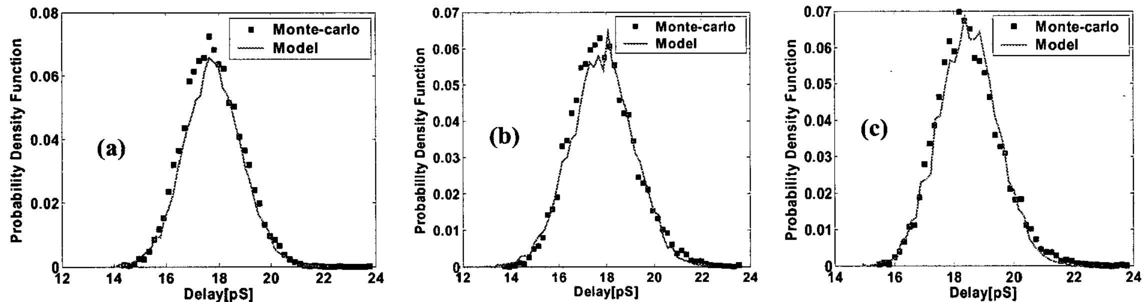

Fig. 4. Model verification: PDF of (a) t_{dLH} (b) t_{dHL} and (c) t_d=Max(t_{dLH}, t_{dHL}) for inverter. ($\sigma Vt0$=60mV is chosen to get a considerable spread in delay distributions; Spice monte-carlo simulations are done for 10000 points).

Authorized licensed use limited to: San Francisco State Univ. Downloaded on December 10, 2008 at 20:54 from IEEE Xplore. Restrictions apply.

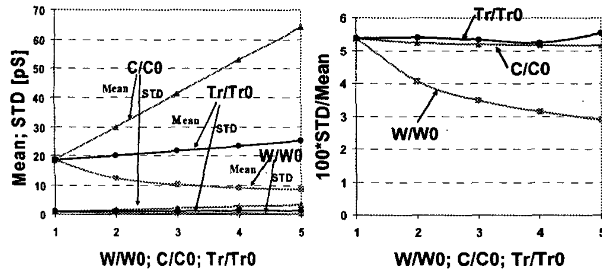

Fig. 5. Impact of sizing (W), output load (C), and input rise/fall time (Tr) on delay PDF of inverter

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 Δt. Thus, $\Delta t > 0$ implies that LH transition at IN2 arrived

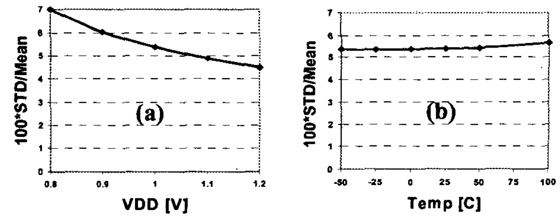

Fig. 6. Impact of (a) supply voltage (VDD), and (b) temperature on delay PDF of inverter

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 Δt on delay distributions of the NAND gate is studied (Fig. 10). As Δt 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 Δt), the PMOS M4 (see Fig. 8) is already "off". Hence, it does not influence t_{d1} (i.e. $(\partial t_{d1}/\partial Vt_{M4}) = 0$). However, if Δt 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. $(\partial t_{d1}/\partial Vt_{M4}) \neq 0$). The influence of Δt 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

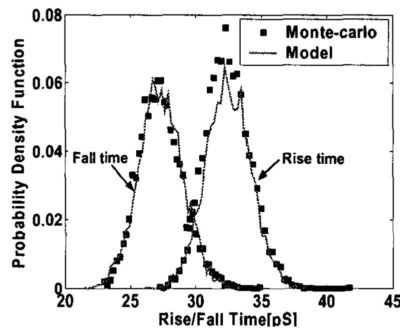

Fig. 7. PDF of output rise/fall time of inverter gate

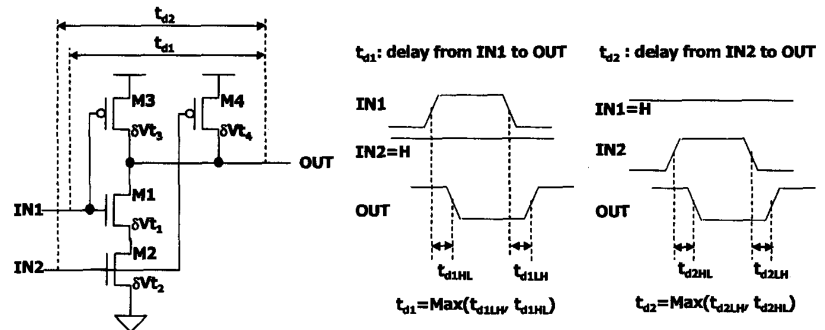

Fig. 8. NAND gate and delay definitions

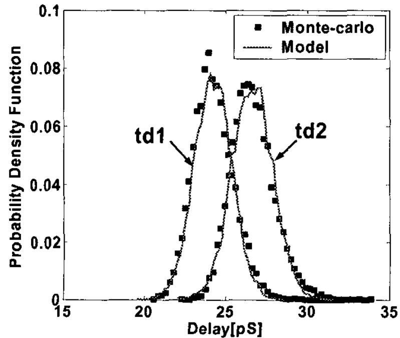

Fig. 9. PDF of t_{d1} and t_{d2} of NAND gate

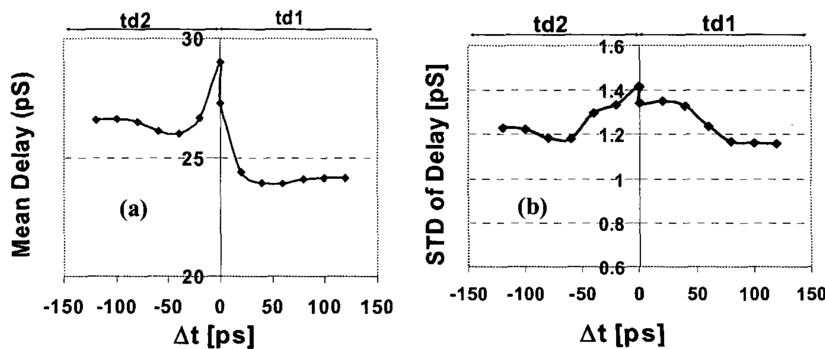

Fig. 10. Impact of input arrival time difference (Δt) on (a) mean, and (b) STD of NAND gate delay

Authorized licensed use limited to: San Francisco State Univ. Downloaded on December 10, 2008 at 20:54 from IEEE Xplore. Restrictions apply.

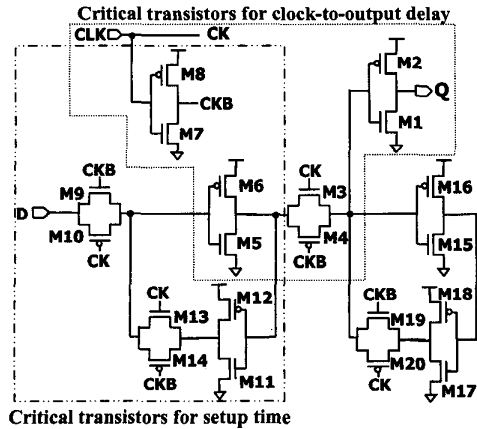

Fig. 11. Transmission Gate Flip-Flop (TGFF)

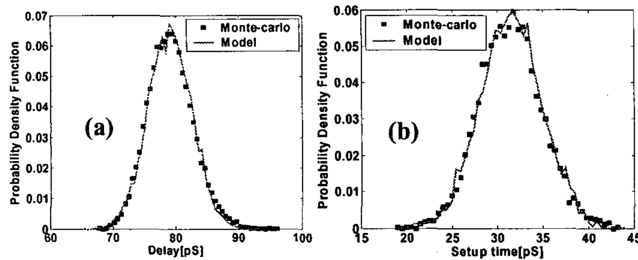

Fig. 12. PDF of (a) t_{cq} and (b) setup time (t_{su}) of TGFF

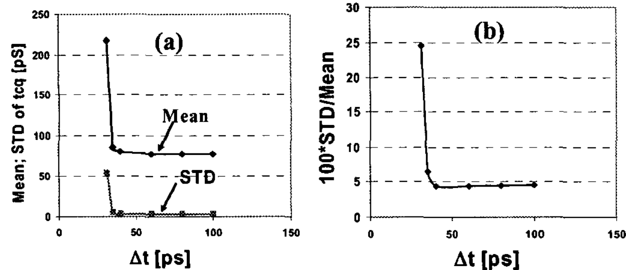

Fig. 13. Impact of input data arrival time (Δt) on (a) mean and STD, and (b) spread of t_{cq} delay of TGFF

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 transistor that are in the path from the clock (CLK) to the output (Q) need to be considered (only 8 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 that they can be neglected in our estimation model (Section 3). This is an example of the transistor set reduction strategy mentioned in section 3. Fig. 12 shows that the

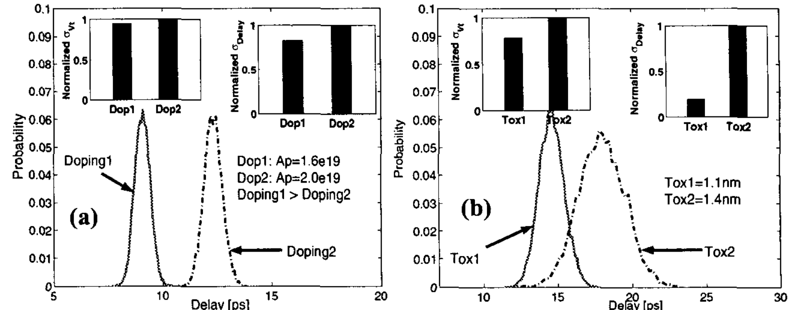

Fig. 14. Impact of (a) doping and (b) oxide thickness on delay distribution of an inverter (designed with 25nm devices) obtained using the proposed model.

estimated distributions closely match the distributions extracted by SPICE Monte-Carlo simulations.

In flip-flops the t_{cq} typically depends on the input data arrival time with respect to the clock rising edge (Δt) [10]. As the data transition gets closer to the clock rising edge, the t_{cq} is initially constant, then it increases, and finally when Δt reaches the setup time (t_{su}), the flip-flop fails to sample that data correctly. Using the proposed modeling, the impact of Δt 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 increase (Fig. 13(b)), as Δt 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 (T_{ox}) on the delay distribution of an inverter designed with predictive 25nm devices [11]. Device simulator MEDICI was used to estimate the partial derivatives in (3). The designed devices have 2-D non-uniform super halo ("Halo" and "Retrograde") doping profile (approximated as Gaussian function) [11]. The peak halo doping value (A_p) was used to modify the doping profile. The effective channel doping (N_a) is calculated using the method described in [12]. Increasing A_p increases the effective doping thereby increasing the STD and the mean of the delay (Fig. 14(a)). It also increases the STD of the Vt fluctuation (by increasing N_a in (1)). Increasing the oxide thickness reduces the current through a transistor thereby increasing delay. The STD of the Vt fluctuation increases with the increase in T_{ox}. Moreover, since a higher T_{ox} increase the short channel effect in a device, the impact of Vt fluctuation on the transistor current becomes more prominent. 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

[1]. A.J. Bhavnagarwala, et. al., IEEE JSSC, vol. 36, pp. 658-665, April 2001.
[2]. Y. Taur and T. H. Ning, *Fundamentals of Modern VLSI Devices*, 1998.
[3]. S. Borkar, et. al., DAC, pp. 338-342, June 2003.
[4]. K. Okada, et. al, ICCAD, pp. 908 – 913, Nov. 2003.
[5]. A. Papoulis, *Probability, Random Variables and Stochastic Process*
[6]. T. Sakurai, et. al, ISCAS, pp. 105 – 108, 1990,
[7]. C. E. Clark, Operations Research 9 (2), pp.145-162, 1961.
[8]. Berkeley Predictive Technology Model, www.device.eecs.berkeley.edu/
[9]. J. Rabaey, *Digital Integrated Circuit*, 2002.
[10]. B. Nikolic, IEEE JSSC, vol. 35, pp. 876-884, June 2000.
[11]. http://www-mtl.mit.edu/Well/
[12] S. Mukhopadhyay, et. al., DAC, pp. 169-174, June 2003

Authorized licensed use limited to: San Francisco State Univ. Downloaded on December 10, 2008 at 20:54 from IEEE Xplore. Restrictions apply.