Statistical Delay Calculation with Multiple Input Simultaneous Switching

Qin Tang, Amir Zjajo, Michel Berkelaar and Nick van der Meijs
Circuits and Systems Group, Delft University of Technology
Q.tang@tudelft.nl

***Abstract*—The increasing process variations which goes along with the continuing CMOS technology shrinking necessitate accurate statistical timing analysis. Multiple Input Simultaneous Switching (MISS) is simplified to Single Input Switching (SIS) in most of the recent approaches, which introduces significant errors in Statistical Static Timing Analysis (SSTA). Hence, we propose a new modeling and statistical analysis method to capture statistical gate delay variations, able to accurately handle MISS. Experiment results obtained with a 45nm technology show that our approach accurately obtains not only mean and standard deviation, but also the third moment, skewness.**

I. Introduction

Static Timing Analysis (STA) tools are widely used for performance verification due to their ability to perform efficient timing checks on large chips. However, STA faces a number of accuracy issues related to false paths and MISS. Since process variations do not shrink at the same ratio as the process geometries, they have an increasing impact with every new technology generation. Small changes in transistor channel length and doping density cause more dramatic changes in transistor behavior compared to older technologies. The traditional corner-based STA is not able to accurately and efficiently model delay variability. A better way to estimate the variability of timing is to perform Statistical Static Timing Analysis (SSTA), which models path or circuit delay variations as a function of all process variations of interest using statistical techniques. SSTA comes in two flavors: path-based SSTA and block-based SSTA. Although block-based SSTA does not have the tough task to select critical paths considering process variations like path-based SSTA, it requires a solution to the basic statistical sum and maximum operations for the propagation of arrival times from the source node to the sink node. Since the maximum is a nonlinear function, the maximum of two normal (Gaussian distributed) arrival times at the inputs of a gate will result in a non-normal arrival time at its output, typically with positive skewness. The error caused by ignoring skewness in the maximum operation is larger if the input arrival times have similar means but dissimilar variances [1]. Approximations for the statistical maximum operator have been proposed for both Gaussian [2] and non-Gaussian random variables [3], but they use an assumption of statistical independence between the input signals. The computation of the exact statistical maximum also needs exact correlation information among the input arrival time variables, which requires extensive computation and storage of dependence. In [6] and [7] discrete probability density functions (pdf) or cumulative density functions (cdf) are used to propagate statistical information, avoiding the maximum operation, however they also assume independent arrival times and circuit delay.

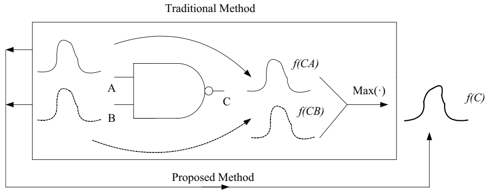

Fig. 1. Exact statistical simulation for MISS

New methods for handling reconvergence and spatial correlations have been proposed in block-based SSTA, for example [4], [5]. In most of these approaches, there is a focus on the correct computation of the statistical maximum, taking proper correlation into consideration. However, an essential source of calculation error is due to MISS, which can not be handled accurately by the above methods, as they all use black-box gate delay models, in which the electrical effect of multiple input signals switching (near-)simultaneously is not modeled. These black-box models only model SIS. The approach presented in this paper uses gate models built out of statistical transistor models, and is hence fundamentally capable of modeling the effects of MISS.

MISS arises when multiple inputs of a gate switch in close proximity in time. Fig. 1 explains the SIS assumption for statistical delay analysis of a NAND2. In a SIS approach, it is always assumed that only one input is switching while the others are deterministic stable (V_{dd} for NAND). The output arrival time distributions $f(CA)$ and $f(CB)$ are calculated by propagating input distributions through the gate separately based on SIS. The final distribution $f(C)$ is found after statistical maximum operation of $f(CA)$ and $f(CB)$. In traditional timing analysis, MISS is a significant problem in both STA and SSTA. It has been reported that not modeling MISS can result in as much as 100% error in STA [8]. [9] shows SIS underestimates the mean delay of a stage by up to 20% and overestimates the standard deviation up to 26%. The existing SSTA approaches considering MISS mainly model delay as a function (linear function, orthogonal polynomial or high dimensional function) of the absolute input arrival times [9]–[13]. Fixed distributions are required in [9]–[11] without

This research is sponsored by the European Union and the Dutch government as part of the ENIAC MODERN project

considering varying input slope in SSTA, which can also cause significant errors [12], [13].

In this paper, we propose a stochastic waveform model (SWM) and delay calculation method that include MISS effects and process variations. As illustrated in Fig. 1, we propose to consider all the inputs together and directly calculate output statistical information avoiding the maximum operation. The SWM has both varying crossing times and input slopes. In our delay calculation algorithm, the discrete *pdf* of crossing time was used. We tested our approach in circuits with MISS up to four inputs. The first three moments—mean (μ), standard deviation (σ) and skewness (γ) values were compared to transistor-level Monte Carlo simulation. Experimental results show that our approach gives accurate results for these three moments even under MISS conditions.

II. Modeling and Statistical Analysis Considering MISS

A. Stochastic Waveform Model (SWM)

In traditional SSTA, the arrival time and gate delay are represented as statistical variables while the slope variability is neglected and treated as a deterministic value in the simplest case. As only the arrival time distribution and deterministic slope are known, the statistical waveform is represented by a set of ramp signals shown in Fig. 2. However, the slope is also an important factor which has direct effect on gate delay [14]. [12] shows that large errors occur without considering the varying slope of input signals. Fig. 2 also shows the realistic stochastic waveforms of a buffer (left) and the output waveforms of a 3-stage inverter chain (right) from Spectre MC simulations. Clearly the stochastic waveforms are not exactly symmetric with respect to the time axis. Similar to [15] and [16], the variational voltage waveform is represented by a time domain stochastic variable:

$$v(t) = v_0(t) + \sum_{k=1}^{M} \alpha_k(t) \cdot \xi_k \tag{1}$$

where $\alpha_k(t)$ is the sensitivity of voltage $v(t)$ to the corresponding process variation ξ_k. Therefore in our SWM, the voltage, rather than crossing time, is modeled as a stochastic variable. The sensitivity matrix $\alpha(t)$ is the parameter which must be calculated during delay calculation. Given the statistical information of process variations, the moments of and covariance between voltages are easily obtained.

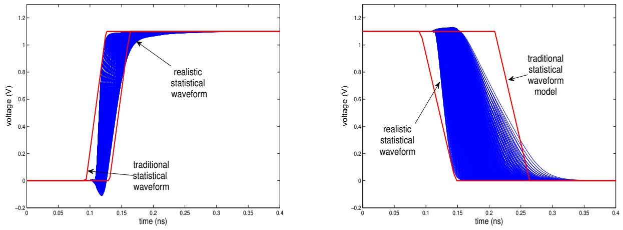

Fig. 2. Stochastic waveform modeling

B. RDE-based Statistical Simulator

As technology scales to 45nm and below, the process variations have greater impact on transistor behavior. To make our gate model as physical (and hence accurate) as possible, we create a gate model at the transistor level. For this paper, a table-based statistical transistor model (STM) is used for gate modeling [17]. Every transistor is modeled as a current source I_{ds} and five capacitors (C_{gs}, C_{gb}, C_{gd}, C_{sb} and C_{db}) as shown in Fig. 3. All elements in the STM are represented as a linear function of process variations of interest.

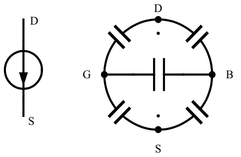

Fig. 3. Transistor model

Monte Carlo is too CPU time intensive for statistical timing analysis since normally at least 1000 runs are required and the number grows when more random variables are considered. As stated in Section II-A, the nominal voltages $v_0(t)$ and sensitivity matrix $\alpha(t)$ must be calculated for SWM. In our RDE-based statistical simulation algorithm [18], the random circuit equation is processed directly to compute $v_0(t)$ and $\alpha(t)$ in (1).

By introducing process variations, the random circuit equation can be expressed in compact form:

$$F(x', x, t, \xi) = C(t, \xi)x'(t, \xi) + G(t, \xi)x(t, \xi) = J(t, \xi) \tag{2}$$

where x is the nodal voltage vector including all input and output voltages and x' is its time derivative. G and C are the conductance and capacitance matrices and J is a current source value vector. The equation can be further simplified by eliminating and moving V_{dd} and input voltages to the right side since they are known. If there are no process variations (ξ=0), the equation is a typical transient analysis equation $F(x', x, t) = 0$ and the solution is denoted as $x_s(t)$ in this paper. In order to manage the random nonlinear equation (2), Taylor expansion is used to linearize (2) at the nominal value $x'_s(t)$, $x_s(t)$ and the mean values of the process variables P_0.

After linearization, the random equation is converted to a linear random differential equation (RDE). Denoting the voltage variation as $y(t) = x(t) - x_s(t)$, the RDE equation can be written as:

$$\begin{aligned} y'(t) &= R(t)y(t) + Q(t)\xi && (3) \\ R(t) &= -C^{-1}(G - \partial J/\partial x) && (4) \\ Q(t) &= C^{-1}\partial J/\partial p && (5) \end{aligned}$$

If the $\partial C/\partial p$ and $\partial G/\partial p$ are comparable to $\partial J/\partial p$, they must be included in (5). If C is singular, it stays in the left of (3).

According to the mean square integral theorem, the solution of (3) is proportional to ξ assuming the initial value is not random [19]. Using $\alpha(t)$ as the coefficient of proportionality and substituting $y = \alpha\xi$ in (3), the equation for $\alpha(t)$ turns out to be an ordinary differential equation:

$$\alpha'(t) = R(t)\alpha(t) + Q(t) \tag{6}$$

which can be solved by fast numerical methods. After solving $x_s(t)$ and (6), the stochastic output waveform model is obtained in (1). Based on the moments and correlations of process variations, the moments of voltage can be calculated by using common statistical operations.

C. Delay Moments Calculation

For timing analysis, the problem of interest is to compute the moments of arrival time, gate delay or in general the crossing time. The crossing time t_η is defined as the first time for voltages to cross the threshold voltage $V_\eta = \eta\% \cdot V_{dd}$. By using a numerical integral method, e.g, backward Euler or the trapezoidal rule, the solution of x_s and α at a specific time point are calculated from that at the previous time point, making the output $x(t)$ a Markovian process. During the period when the nominal voltage is in transition, the calculation of crossing time *cdf* (F_n in (7)) starts and for a rising transition this is expressed as:

$$
\begin{aligned}
F_n &= P(t_\eta \le t_n) = 1 - P(t_\eta > t_n) = 1 - G_n && (7)\\
G_n &= P(v_1 \le V_\eta \cap v_2 \le V_\eta \cap \ldots \cap v_n \le V_\eta) && (8)\\
&= P(v_n \le V_\eta | v_{n-1} \le V_\eta, \ldots, v_1 \le V_\eta) \cdot G_{n-1} && (9)\\
&= P(v_n \le V_\eta | v_{n-1} \le V_\eta) \cdot G_{n-1} (n = 2 : N) && (10)\\
&= \frac{P(v_n \le V_\eta \cap v_{n-1} \le V_\eta)}{P(v_{n-1} \le V_\eta))} \cdot G_{n-1} && (11)
\end{aligned}
$$

where v_i is the voltage of interest at time t_i. According to the properties of a Markovian process $v(t_n)$, (9) is rewritten in (10). Based on (7) to (11) an iteration method is used to calculate the *cdf* of the corresponding crossing time with initial condition G_1=1. Given the moments and covariances calculated in the RDE-based statistical simulator, the joint probability and single probability in (11) are easy to obtain.

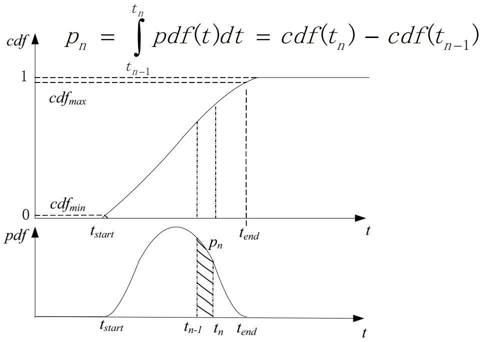

Fig. 4. Cumulative density function and discrete probability density function

To simplify the calculations, the cdfs and pdfs have these properties: $i)$ $cdf = 1$ if $cdf \ge cdf_{max}$ and $cdf = 0$ if $cdf \le cdf_{min}$. The time t_{start} and t_{end} are the time corresponding to cdf_{min} and cdf_{max} respectively shown in Fig. 4; $ii)$ the pdf is calculated during the period $[t_{start}\ t_{end}]$. At any time t_n, the discrete pdf is approximated by $p_n = \int_{t_{n-1}}^{t_n} pdf(t)dt$ with $p(t_{start}) = 0$.

The next step is to restrict the *cdf* within t_{start} and t_{end} (denoted $cdfr$ in this paper), shown in Fig. 4. Since the simulation uses a dynamic time step algorithm for efficiency, the $cdfr$ needs to be uniformly sampled for *pdf* computation. After uniformly sampling $cdfr$ with N_s samples and interpolating $cdfr$, the resulting time vector T_1 and $cdfu$ vectors are used to calculate the *pdf* vector Λ with element $\Lambda_k = cdfu_k - cdfu_{k-1}$ ($\Lambda_1 = 0$).

The last step is to calculate the moments of crossing time. Denoting Λ^T as the transposition of the row vector Λ, the calculation method can be explained as following:

$$
\begin{aligned}
\mu &= T_1 \Lambda^T && (12)\\
\sigma &= T_2 \Lambda^T - \mu^2 \quad (T_{Nk} = T_{1k}^N \quad k = 1 : N_s) && (13)\\
\gamma &= (\Gamma - 3\mu\sigma^2 - \mu^3)/(\sigma^3) \quad (\Gamma = T_3 \Lambda^T) && (14)
\end{aligned}
$$

The calculation method for a falling transition is similar to the above methods with the only difference in (8). By replacing v_i by $V_{dd} - v_i$, (8) to (11) are still used in the same way. If the waveform is non-monotonic and crosses V_η multiple times, the method above can be used to iteratively find all crossing times.

III. Experimental results

The proposed model and algorithm were implemented and tested on a set of gates and circuits in the 45nm PTMVTG technology [20]. Transistor-level Monte Carlo(MC) simulation results (2500 samples) are regarded as the golden reference. Fig. 5 shows the discrete *pdf* with 50 samples and the histogram of MC simulation in Spectre of a NAND2 with falling output, AOI21 and AOI22 with rising output. All inputs of each gate have the exact same mean value of arrival times. The discrete *pdf* was scaled to provide a straightforward shape comparison. In the beginning, the statistical input signal at every input of a multi-input gate was modeled as ramp signals of 100ps transition time with voltage variations. The σ of voltages and arrival time differences among input signals are varied to obtain results at diverse scenarios.

Table I[1] lists the average error of mean (μ), standard deviation (σ) and skewness (γ) of delay in gates with different levels of complexity. It shows that the worst average σ and γ error occur in NAND4 and NAND3 respectively. These gates have the most transistors stacked among the gates with corresponding same number of inputs. Fig. 6 illustrates the errors of all the experiments. The μ errors are within 1%. All the σ errors are within 5% except two biggest σ cases (6.02% and 6.42%) coming from NAND4 with rising output and falling output respectively. All of the skewness errors are within 7%. We also simulated two combinational circuits (COM1 and COM2 in Table I) with identical paths to the output gate. The process variables are taken to be length and width with 3σ of 20% and 43% of the mean value. By using identical input signals switching at the same time, the inputs of the output gate has MISS and realistic waveforms produced by process variations like the curves in Fig. 2. The results of COM1 and COM2 are compared to 10000× MC results in Spectre. It is observed that the smaller the arrival time difference, the larger skewness, so the skewness should not

[1] importantance sampling-based MC was used for all senarios of NAND2 and NOR2 as comparison references

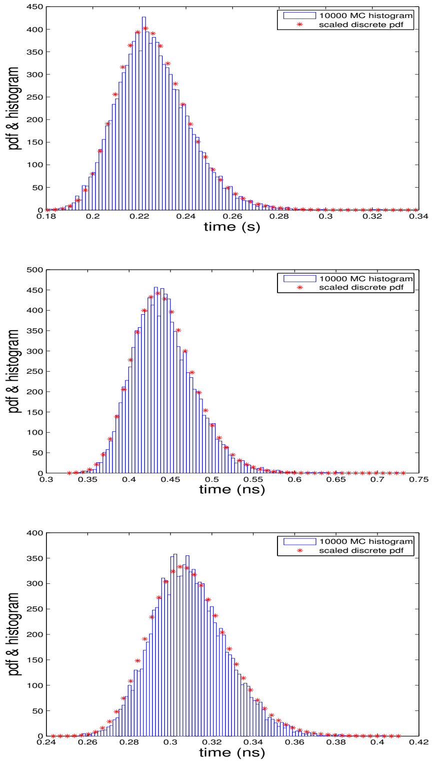

Fig. 5. pdf and histogram comparison of NAND2, AOI21 and AOI22

be ignored. The experiment results indicate the necessity of skewness estimation and the ability of the proposed method to accurately calculate three moments.

Compared to 1000 MC runs, our method achieves 62×speedup on average. Although our method only needs to simulate once for statistical output voltages, the equations of the RDE-based statistical simulation have an extra set of equations for α computation in (1) and need sensitivity calculation for (3), which slows down the simulation. We are working on using a faster differential equation solver, even higher speedup is expected.

TABLE I
ACCURACY COMPARISON OF THREE MOMENTS FOR MULTI-INPUT GATES

	errors of rising output			errors of falling output		
cell	μ	σ	γ	μ	σ	γ
NAND2	0.52%	2.00%	2.52%	0.20%	1.14%	2.20%
NOR2	0.38%	1.03%	3.07%	0.15%	1.73%	2.91%
NOR3	0.55%	0.82%	3.26%	0.15%	1.90%	2.82%
NAND3	0.70%	2.38%	5.52%	0.41%	2.04%	3.05%
AOI21	0.04%	1.56%	2.00%	0.09%	1.34%	1.71%
AOI211	0.13%	3.06%	2.62%	0.04%	1.84%	4.33%
AOI22	0.04%	0.65%	1.60%	0.19%	1.68%	3.06%
NAND4	0.75%	4.48%	2.70%	0.33%	4.54%	3.27%
COM1	0.93%	1.99%	4.00%	0.37%	3.89%	4.26%
COM2	0.64%	4.39%	5.10%	0.41%	5.58%	5.75%

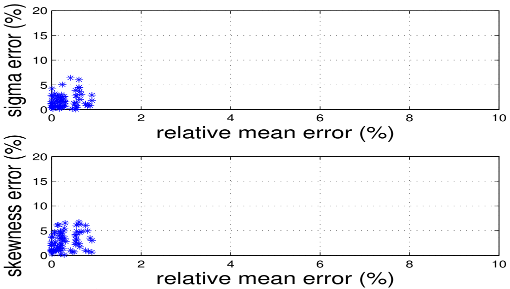

Fig. 6. All moment percentage errors comparison

IV. CONCLUSION

The errors introduced by the SIS assumption and statistical maximum operation motivate us to propose a novel modeling and simulation method to capture process variations and MISS, avoiding the maximum operation. We represent variational waveforms of any shape in time domain statistical variables including the influence of slope. The transistor-level gate modeling and RDE-based statistical simulation method provides variational output waveforms, from which the moments of crossing times are computed. Due to the increasing process variations and continuing shrinking technology, the μ and σ are not always enough to represent variational gate delay. Additionally in this paper we computed the skewness of gate delay. Experimental results indicated the high accuracy of our approach compared to Monte Carlo simulations.

REFERENCES

[1] C. Clark, “The greatest of a finite set of random variables,” *J. Oper. Res.*, vol. 9, no. 2, pp. 145–162, Mar./Apr. 1961.

[2] C. Visweswariah, K. Ravindran, S. W. K. Kalafala, and S. Narayan, “First-order incremental block-based statistical timing analysis,” in *DAC*, 2004, pp. 331–336.

[3] J. Singh and S. Sapatnekar, “Statistical timing analysis with correlated non-Gaussian parameters using independent component analysis,” in *DAC*, 2006, pp. 155–160.

[4] L. Zhang, W. Chen, Y. Hu, J. A. Gubner, and C. C. Chen, “Correlation-preserved statistical timing with a quadratic form of Gaussian variables,” *Trans. Comput.-Aided Des. of Integr. Circuits Syst.*, vol. 25, no. 11, pp. 2437–2449, 2006.

[5] V. Khandelwal and A. Srivastava, “A quadratic modeling-based framework for accurate statistical timing analysis considering correlations,” *Trans. on VLSI Syst.*, vol. 15, no. 2, p. 206-215, 2007.

[6] J.-J. Liou, K.-T. Cheng, S. Kundu, and A. Krstic, “Fast statistical timing analysis by probabilistic event propagation,” in *DAC*, 2001, pp. 661–666.

[7] A. Devgan and C. Kashyap, “Block-based static timing analysis with uncertainty,” in *ICCAD*, 2003, pp. 607–614.

[8] C. Amin, C. Kashyap, N. Menezes, k. Killpack, and E. Chiprout, “A multi-port current source model for multiple-input switching effects in CMOS library cells,” in *In DAC 2006*, 2006, pp. 247–252.

[9] A. Agarwal, F. Dartu, and D. Blaauw, “Statistical gate delay model considering multiple input switching,” in *DAC*, 2004, pp. 658–663.

[10] K. Y. Satish, J. Li, C. Talarico, and J. Wang, “A probabilistic collocation method based statistical gate delay model considering process variations and multiple input switching,” in *DATE*, 2005, p. 770-773.

[11] S. Yanamanamanda, J. Li, and J. Wang, “Uncertainty modeling of gate delay considering multiple input switching,” in *ISCAS*, 2005, pp. 2457–2460.

[12] J. Sridharan and T. Chen, “Modeling multiple input switching of CMOS gates in DSM technology using HDMR,” in *DATE*, vol. 1, 2006, pp. 1–6.

[13] J. Sridharan and T. Chen, “Gate delay modeling with multiple input switching for static (statistical) timing analysis,” in *VLSID*, 2006, pp. 323–328.

[14] T. Kouno and H. Onodera, “Consideration of transition-time variability in statistical timing analysis,” in *SOCC*, 2006, pp. 207–210.

[15] H. Fatemi, S. Nazarian, and M. Pedram, “Statistical logic cell delay analysis using a current-based model,” in *DAC*, 2006, pp. 253–256.

[16] B. Liu and A. B. Kahng, “Statistical gate level simulation via voltage controlled current source models,” in *BMAS Workshop*, 2006, p. 23.

[17] Q. Tang, A. Zjajo, M. Berkelaar, and N. van der Meijs, “Transistor level waveform evaluation for timing analysis,” in *VARI*, 2010, pp. 1–6.

[18] Q. Tang, A. Zjajo, M. Berkelaar, and N. van der Meijs, “RDE-based transistor-level gate simulation for statistical static timing analysis,” in *DAC*, 2010, pp. 787–792.

[19] T.T. Soong, “Random differential equations in science and engineering,” New York: Academic Press, 1973.

[20] Arizona State University, “Predictive Technology Model (PTM)”, http://ptm.asu.edu.