A Statistical Gate-Delay Model Considering Intra-Gate Variability

Kenichi Okada
Dept. Communications & Computer Engineering
Kyoto University
kokada@vlsi.kuee.kyoto-u.ac.jp

Kento Yamaoka
Dept. Communications & Computer Engineering
Kyoto University
kento@vlsi.kuee.kyoto-u.ac.jp

Hidetoshi Onodera
Dept. Communications & Computer Engineering
Kyoto University
onodera@i.kyoto-u.ac.jp

ABSTRACT

This paper proposes a model for calculating statistical gate-delay variation caused by intra-chip and inter-chip variability. As the variation of individual gate delays directly influences the circuit-delay variation, it is important to characterize each gate-delay variation accurately. Furthermore, as every transistor in a gate affects the transient characteristics of the gate, it is also necessary to consider the *intra-gate* variability in the model of gate-delay variation. This effect is not captured in existing statistical delay analyses. The proposed model considers the intra-gate variability through the introduction of sensitivity constants. The accuracy of the model is evaluated, and some simulation results for circuit delay variation are presented.

1. INTRODUCTION

The consideration of device fluctuations is an important theme in designing CMOS integrated circuits, particularly on the deep submicron technology where fluctuation of device characteristics is much more pronounced. In a conventional worst-case timing analysis, the device characteristics are usually assumed to be uniform within a chip, whereas in fact considerable intra-chip variation may be present [1, 2], potentially degrading product yield[2, 3, 4]. In this paper, we suppose an inter-chip variability and an intra-chip variability of transistor characteristics. The intra-chip variability represents the fluctuation of characteristics between individual transistors on a single chip, and inter-chip variability represents the fluctuation in characteristics between chips assuming uniform characteristics within each chip.

It is necessary to consider the variability of all transistors within a chip to calculate the intra- and inter-chip variations of gate delay. A number of methods for statistical timing analysis based on the intra-chip delay variability have been proposed [13, 14, 5, 6, 7]. However, these previous methods were gate-level analyses, approximating intra-chip variability of gate delay by a constant ratio, for example, 15% or 20% of the average gate-delay. In order to accurately simulate the circuit-delay variation by statistical timing analysis, the delay variations of each gate should be calculated accurately, not approximated to a single value.

Intra-chip variability of gate delay is caused by intra-chip variation in transistor characteristics in the gate. As such, every transistor on a charging/discharging path influences the gate-delay variability. Therefore, the ratio of the intra-chip delay variation is not constant, instead depending on the number of transistors on the path. The model should therefore consider the relative variability among intra-gate transistors, in this paper referred to as *the intra-gate variability*. The intra-gate variability influences each gate-delay variation much. As a result, each gate-delay variation influences the total circuit-delay variation directly, so we need to consider the intra-gate variability in a statistical timing analysis. It is important to estimate each gate-delay variation accurately. This paper proposes a model for a statistical gate-delay calculation with consideration of the intra-gate variability.

Although intra-gate variability may be calculated simply by assigning variables to every transistor in the gate, such a model would quickly become infeasible when considering large numbers of transistors. The proposed model was therefore developed to calculate the intra-chip variability of gate delay using fewer variables.

In this paper, the gate-delay model based on a response surface method (RSM)[8] and incorporating intra-gate variability is first presented, with mention of the modeling of transistor characteristics. The method for calculating the sensitivity constant of the gate-delay model is then introduced, and the accuracy of calculation is evaluated. Some simulation results are then presented, demonstrating the importance of considering intra-gate variability.

2. STATISTICAL GATE-DELAY MODELING

The proposed gate-delay model for characterizing inter-chip and intra-chip delay variation is based on a response surface method (RSM)[8]. When multiple transistors exist on a charging/discharging path, each of the transistors affects the transient behavior of the gate. For example, Fig. 1 shows fall-delay distributions calculated with and without consideration of intra-gate variability for a 2-input NAND gate. In this example, only the intra-chip variability of transistors is considered, and the influences of each transistor on gate delay are assumed to be independent. Two nMOS transistors lie on the discharging path, giving rise to delay variance of $2\sigma^2$ ($=\sigma^2 + \sigma^2$) when calculated with intra-gate variability. Without considering intra-gate variability, characteristics of each transistor on the charging/discharging path are always equal, and the delay variance should be roughly $4\sigma^2$ ($=(\sigma + \sigma)^2$). This demonstrates that in calculating both inter-chip and intra-chip gate-delay variation, it is necessary to consider the intra-gate variability.

The simplest method for calculating intra-gate variability would be to assign RSM variables to each transistor. However, such an approach would require many variables. For example, assuming 4 variables assigned to each transistor, a 4-input 4-folded NAND gate with 16 nMOS and 16 pMOS transistors would required 128 variables {(16nmos + 16pmos) $\times$ 4variables}. Such a large number of variables will lengthen the time require to generate RSMs and calculate delay. In the proposed delay model, calculation of intra-chip and inter-chip variations is carried out using a single statistic to represent the intra-gate variability.

The variables of the proposed model are combined based on the statistical theorem that the sum of multiple normal distributions obeys a normal distribution. The concept of *a sensitivity constant* is introduced to allow the intra-chip delay variation to be calculated from the inter-chip delay variation. The coefficients of the RSM can then be calculated without considering intra-gate variability explic-

ICCAD'03, November 11-13, 2003, San Jose, California, USA.

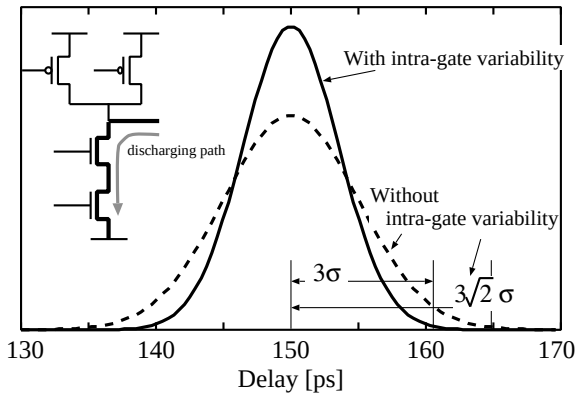

Figure 1: Consideration of intra-gate variability (fall-delay distribution of 2-input NAND gate)

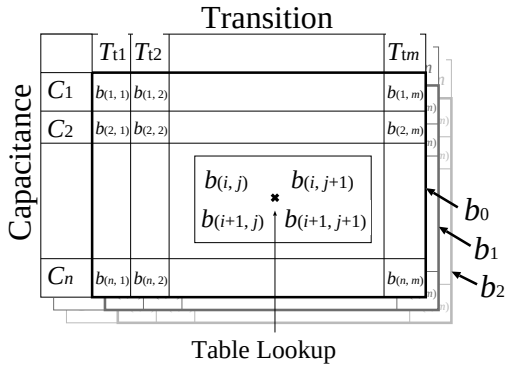

Figure 2: Concept of table-lookup RSM

itly. The intra-gate variation of gate delay is then determined from the sensitivity constant and the RSM calculated without intra-gate variability.

2.1 Statistical Modeling of Transistor Characteristics

The transistor characteristics are mapped to physical parameters[9]. In this paper, V_{TH0}, T_{OX}, W_{int} and L_{int} are used as the physical parameters. For convenience, the transistor characteristics are expressed by a vector $\boldsymbol{p} = (p_1, p_2, \ldots, p_n)^T$, the elements of which are the physical parameters for nMOS and pMOS transistors. The transistor characteristics $\boldsymbol{p}$ are also expressed as the sum of the average $\boldsymbol{\mu}$, the inter-chip variation $\boldsymbol{p}_g$, and the intra-chip variation $\boldsymbol{p}_r$, as follows.

$$\boldsymbol{p} = \boldsymbol{\mu} + \boldsymbol{p}_g + \boldsymbol{p}_r \tag{1}$$

The fluctuation of transistor characteristics consists of both correlative components and independent components. On a chip, the inter-chip variation represents the correlative component, and the intra-chip variation represents the independent component. The average $\boldsymbol{\mu}$ is common to both. The inter-chip variation $\boldsymbol{p}_g$ is different for every chip, but is uniform within each chip, whereas the intra-chip variation $\boldsymbol{p}_r$ is different for every transistor. The intra-chip variation fluctuates independently and represents the stochastic "white noise" of the fabrication process. The inter-chip and intra-chip variations are expressed as normal distributions, and the variations are statistically independent. Both variations are characterized by variances of physical parameters and correlations among parameters. The intra-chip variance is modeled so as to be inverse proportional to the gate area [10, 9], and not to be dependent on the geometrical location and distance between transistors [1, 11, 9]. In this scheme, the smaller the gate-size, the larger the standard deviation of intra-chip variation.

2.2 RSM Using Lookup Tables

The RSM of gate delay without intra-gate variability is generated as follows. The delay time t_d is expressed in terms of physical parameters $\boldsymbol{p}$ using a linear RSM $\boldsymbol{b}$, as given by

$$\begin{aligned} t_d &= \mathrm{rsm}(\boldsymbol{p}) \\ &= b_0 + (b_1\ b_2\ \cdots\ b_n)\begin{pmatrix} p_1 \\ \vdots \\ p_n \end{pmatrix} \qquad (2) \\ &= b_0 + \boldsymbol{b}^T\boldsymbol{p} \qquad (3) \end{aligned}$$

The delay time depends on not only the transistor characteristics, but also the transition time and load capacitance. Therefore, the coefficients of the RSM also depend on the transition time and the load capacitance. As it is difficult to build an accurate RSM that includes transition time and capacitance as parameters over a wide parameter range, a table-lookup RSM is employed, the coefficients $(b_0, \ldots, b_n)$ of which are calculated from a table of transition time and capacitance. While tables are only used for the coefficient b_0 in traditional static timing analysis without consideration of delay variations, the proposed model uses tables to determine the coefficients $(b_0, \ldots, b_n)$, as shown in Fig. 2.

In each table, the coefficients of the RSM are derived from several SPICE simulations without considering intra-gate variability. A set of coefficients is simulated for each transition and each capacitance forming the two-dimensional table. The table-lookup RSM is built for each gate and each rise and fall time.

2.3 Gate Delay Model

Here, we propose a model to calculate a delay variation, which is derived through the following two steps. We combine the variables of our model based on the statistical theorem that a sum of multiple normal distributions obeys a normal distribution.

First, we discuss the intra- and inter-chip delay variations in Sect. *2.3.1*. We show that the intra-chip delay variation can be expressed by one variable per transistor. Second, we explain that the intra-chip delay variation can be expressed by one variable per gate in Sect. *2.3.2*. We use a sensitivity constant of each transistor to the gate delay, and we assign a variable to each transistor for the intra-gate variability. We explain that all variables, for the intra-gate variability, can be combined into one variable.

2.3.1 Normalization and Combination of RSM Variables

The calculation of gate-delay is discussed here using an INV gate as an example, which consists of one nMOS transistor and one pMOS transistor. For this part of the model, it is not necessary to consider the relative variability of transistors (intra-gate variability) to calculate the gate-delay variation. The physical parameters are represented by the sum of inter-chip variation $\boldsymbol{p}_g$ and intra-chip variation $\boldsymbol{p}_r$. The vectors $\boldsymbol{p}_g$ are common to all transistors within each chip, whereas the vectors $\boldsymbol{p}_r$ are different for every transistor. The delay time t_d is calculated from the inter-chip and intra-chip variation as follows.

$$\begin{aligned} t_d &= b_0 + \boldsymbol{b}^T\boldsymbol{p} \qquad (4) \\ &= b_0 + \boldsymbol{b}^T(\boldsymbol{\mu} + \boldsymbol{p}_g + \boldsymbol{p}_r) \qquad (5) \end{aligned}$$

As the physical parameters are generally correlated in some way, the physical parameters are normalized. The correlative pa-

rameters $\boldsymbol{p}$ are derived from non-correlated variables $\boldsymbol{x}$ using principal component analysis (PCA) as follows [12].

$$\boldsymbol{p} = \boldsymbol{D}\boldsymbol{U}\boldsymbol{\Lambda}^{1/2}\boldsymbol{x} \tag{6}$$

where $\boldsymbol{U}$ is a matrix with eigenvectors of a correlation matrix of $\boldsymbol{p}$ as column vectors, and $\boldsymbol{\Lambda}$ is a matrix with eigenvalues of the correlation matrix on the diagonal. The diagonal elements of matrix $\boldsymbol{D}$ represent the standard deviations of the physical parameters. Each element of vector $\boldsymbol{x}$ is a random number with a normal distribution N(0, 1) having a mean value of 0 and variance of 1.

Considering the inter-chip variation and the intra-chip variation, the delay t_d is modeled as follows.

$$\begin{aligned} t_\mathrm{d} &= b_0 + \boldsymbol{b}^T\boldsymbol{\mu} + \boldsymbol{b}^T\boldsymbol{D}_\mathrm{g}\boldsymbol{U}_\mathrm{g}{\boldsymbol{\Lambda}_\mathbf{g}}^{1/2}\boldsymbol{x}_\mathrm{g} + \boldsymbol{b}^T\boldsymbol{D}_\mathrm{r}\boldsymbol{U}_\mathrm{r}{\boldsymbol{\Lambda}_\mathbf{r}}^{1/2}\boldsymbol{x}_\mathrm{r} && (7)\\ &= t_0 + \boldsymbol{\tau}_\mathrm{g}^T\boldsymbol{x}_\mathrm{g} + \boldsymbol{\tau}_\mathrm{r}^T\boldsymbol{x}_\mathrm{r} && (8)\end{aligned}$$

Each element of vector $\boldsymbol{x}_\mathrm{g}$ and $\boldsymbol{x}_\mathrm{r}$ is a random number with a normal distribution having a mean value of 0 and variance of 1. The terms $\boldsymbol{\tau}_\mathrm{g}^T\boldsymbol{x}_\mathrm{g}$ and $\boldsymbol{\tau}_\mathrm{r}^T\boldsymbol{x}_\mathrm{r}$ represent the inter-chip and intra-chip delay variation, respectively.

The intra-chip delay variation $\boldsymbol{\tau}_\mathrm{r}^T\boldsymbol{x}_\mathrm{r}$ is the sum of the normal distributions, and as such can be expressed as a combined normal distribution. The delay variation t_d can then be formulated as the following equation.

$$t_\mathrm{d} = t_0 + \boldsymbol{\tau}_\mathrm{g}^T\boldsymbol{x}_\mathrm{g} + \tau_\mathrm{r}x_\mathrm{r} \tag{9}$$

The constant t_0 is the mean value of gate delay t_d, and the scalar x_r is a random number N(0, 1). The elements of $\boldsymbol{x}_\mathrm{g}$ cannot be combined, being common to all transistors within each chip. The subscripts g and r represent the inter-chip and intra-chip variability, respectively. The matrices $\boldsymbol{D}_\mathrm{g}$, $\boldsymbol{U}_\mathrm{g}$ and $\boldsymbol{\Lambda}_\mathrm{g}$ are calculated from the variances and covariances of the physical parameters. The intra-gate delay variation τ_r is the root-mean-square of the variances of each intra-gate variation $\boldsymbol{\tau}_\mathrm{r}^T\boldsymbol{x}_\mathrm{r}$. The parameters t_0, $\boldsymbol{\tau}_\mathrm{g}$ and τ_r are calculated as follows.

$$\begin{aligned} t_0 &= b_0 + \boldsymbol{b}^T\boldsymbol{\mu} && (10)\\ \boldsymbol{\tau}_\mathrm{g}^T &= \boldsymbol{b}^T\boldsymbol{D}_\mathrm{g}\boldsymbol{U}_\mathrm{g}\boldsymbol{\Lambda}_\mathrm{g}^{1/2} && (11)\\ \tau_\mathrm{r} &= \sigma(\tau_\mathrm{r}x_\mathrm{r}) = \sigma(\boldsymbol{b}^T\boldsymbol{p}_\mathrm{r}) && (12)\end{aligned}$$

2.3.2 *Intra-Gate Variability*

The model as described in the previous subsection does not consider intra-gate variability, assuming one transistor on the charging/discharging path. The model can now be extended to the case of multiple transistors on the charging/discharging paths by introducing a sensitivity constant s_k. The intra-chip delay variation can then be calculated from the RSM $\boldsymbol{b}$ and the sensitivity constant.

In order to generalize the model, a term $\boldsymbol{b}_k^T\boldsymbol{p}_k$ is assigned to each transistor on the charging/discharging paths. Equation (5) is then rewritten as follows.

$$\begin{aligned} t_\mathrm{d} &= b_0 + \boldsymbol{b}_1^T\boldsymbol{p}_1 + \boldsymbol{b}_2^T\boldsymbol{p}_2 + \ldots + \boldsymbol{b}_k^T\boldsymbol{p}_k + \ldots + \boldsymbol{b}_m^T\boldsymbol{p}_m && (13)\\ &= b_0 + \sum_k^m \boldsymbol{b}_k^T\boldsymbol{\mu} + \sum_k^m \boldsymbol{b}_k^T\boldsymbol{p}_\mathrm{g} + \sum_k^m \boldsymbol{b}_k^T\boldsymbol{p}_{\mathrm{r}k} && (14)\end{aligned}$$

where m is the number of transistors on the charging/discharging paths. The terms $\boldsymbol{b}_k^T\boldsymbol{p}_k$ are allocated to each transistor. The RSM $\boldsymbol{b}$ is similarly expressed as

$$\boldsymbol{b}^T = \sum_k^m \boldsymbol{b}_k^T. \tag{15}$$

The delay t_d is then calculated as follows.

$$\begin{aligned} t_\mathrm{d} &= b_0 + \boldsymbol{b}^T(\boldsymbol{\mu} + \boldsymbol{p}_\mathrm{g}) + \sum_k^m \boldsymbol{b}_k^T\boldsymbol{p}_{\mathrm{r}k} && (16)\\ &= t_0 + \boldsymbol{\tau}_\mathrm{g}^T\boldsymbol{x}_\mathrm{g} + \sum_k^m \tau_{\mathrm{r}k}x_{\mathrm{r}k} && (17)\end{aligned}$$

$$\boldsymbol{\tau}_\mathrm{g}^T = \boldsymbol{b}^T\boldsymbol{D}_\mathrm{g}\boldsymbol{U}_\mathrm{g}{\boldsymbol{\Lambda}_\mathbf{g}}^{1/2} \tag{18}$$

The vectors $\boldsymbol{x}_\mathrm{g}$ are common to all transistors within each chip, whereas the scalars $x_{\mathrm{r}k}$ are different for every transistor. The variables $\boldsymbol{x}_\mathrm{g}$ and $x_{\mathrm{r}k}$ are assumed to be independent.

It is then necessary to estimate the intra-chip delay variations $\sum_k^m \tau_{\mathrm{r}k}x_{\mathrm{r}k}$ of Eq. (17). However, although the coefficient $\boldsymbol{\tau}_\mathrm{g}$ can be calculated from the RSM $\boldsymbol{b}$, the coefficients $\tau_{\mathrm{r}k}$ cannot be calculated from the RSM $\boldsymbol{b}$ directly. Therefore, a sensitivity constant s_k is introduced to represent the sensitivity of transistor variation to gate delay, defined as being proportional to $\tau_{\mathrm{r}k}$. If the sensitivity constants can be calculated for each transistor, the coefficients $\tau_{\mathrm{r}k}$ can be calculated from the RSM $\boldsymbol{b}$ and the sensitivity constants. The RSM $\boldsymbol{b}$ is first calculated without intra-gate variability, and then the effect of intra-gate variability can be expressed by sensitivity constants. The coefficient $\tau_{\mathrm{r}k}$ can then be expressed by an arbitrary value $\tau_{\mathrm{r}0}$ and a sensitivity constant s_k as a proportionality constant.

$$\tau_{\mathrm{r}k} = s_k\tau_{\mathrm{r}0} \tag{19}$$

The sensitivity constant s_k can be readily simulated through sensitivity analysis, as described later.

As each individual intra-gate variation $s_k\tau_{\mathrm{r}0}x_{\mathrm{r}k}$ is independent, the intra-gate variation can be expressed by a single scalar number x_r. From Eqs. (17) and (19), we obtain

$$t_\mathrm{d} = t_0 + \boldsymbol{\tau}_\mathrm{g}^T\boldsymbol{x}_\mathrm{g} + \sqrt{\sum_k^m s_k^2}\ \tau_{\mathrm{r}0}x_\mathrm{r}. \tag{20}$$

The variable τ_{r0} depends on the absolute values of the sensitivity constants s_k, and cannot be calculated from the RSM $\boldsymbol{b}$. Hence, the intra-gate delay variation τ_r simulated without consideration of the relative variation of each transistor is employed. When p_r is assigned to all transistor on the charging/discharging paths, the delay variation τ_r can then be calculated from the RSM $\boldsymbol{b}$ as follows.

$$\begin{aligned} \tau_\mathrm{r} &= \sigma(\boldsymbol{b}^T\boldsymbol{p}_\mathrm{r}) && (21)\\ &= \sigma\left(\sum_k^m \tau_{\mathrm{r}k}x_\mathrm{r}\right) && (22)\\ &= \sum_k^m s_k\tau_{\mathrm{r}0} && (23)\\ \tau_{\mathrm{r}0} &= \frac{1}{\sum_k^m s_k}\tau_\mathrm{r} && (24)\end{aligned}$$

Therefore, p_r can be used by calculating τ_r without consideration of the relative variations of each transistor.

From Eqs. (20) and (24), the proposed gate-delay model becomes

$$t_\mathrm{d} = t_0 + \boldsymbol{\tau}_\mathrm{g}^T\boldsymbol{x}_\mathrm{g} + \frac{\sqrt{\sum_k^m s_k^2}}{\sum_k^m s_k}\tau_\mathrm{r}x_\mathrm{r}. \tag{25}$$

The second term $\boldsymbol{\tau}_\mathrm{g}^T\boldsymbol{x}_\mathrm{g}$ represents the inter-chip delay variation, the third term represents the intra-chip delay variation with consideration of the intra-gate variability, and the term $\tau_\mathrm{r}x_\mathrm{r}$ represents the intra-chip delay variation calculated *without intra-gate variability*.

The coefficient $\frac{\sqrt{\sum_k^m s_k^2}}{\sum_k^m s_k}$ is referred to in this model as the intra-gate-variability factor. The difference in delay variation due to intra-gate variability can then be expressed solely in terms of the intra-gate-variability factor, eliminating the need to build RSMs $\boldsymbol{b}_k$ for each transistor. The variations $\boldsymbol{\tau}_g$ and τ_r are calculated from the RSM $\boldsymbol{b}$ using Eqs. (18) (21), respectively.

The proposed model can express the intra-gate variability in terms of a single variable, in contrast to the $n \cdot m$ variables required for a fully allocated model Eq. (14), where n is the number of physical parameters, and m is the number of transistors on the charging/discharging paths.

This methodology is expected to be readily applicable to the statistical circuit-delay analyses reported so far [13, 14, 5, 6], because the gate-delay variation with intra-gate variability can be calculated using the intra-gate-variability factor. Worst-case analysis can be employed to simulate the values t_0, $\boldsymbol{\tau}_g^T \boldsymbol{x}_g$ and $\tau_r x_r$ because the values can be simulated without considering intra-gate variability. The intra-gate-variability factor can be obtained by simply simulating the sensitivity constants.

3. SENSITIVITY CONSTANT

3.1 Calculation Methods of the Sensitivity Constant

The sensitivity constants can be simulated through sensitivity analysis for each transistor. Although sensitivity constants depend on the transition time and load capacitance, it is computationally expensive to calculate sensitivity constants for every transition time and every load capacitance. Four candidate methods for calculating the sensitivity constant are evaluated below, two considering intra-gate variation, and two examples without.

(A) *Calculation of s_k for each transition time and each load capacitance*

In this simple method, the sensitivity constant s_k is simulated for every transition time and every load capacitance. In order to calculate the sensitivity constant, SPICE parameters for a slow case and a typical case are employed, as determined for the intra-chip variability [9]. The sensitivity constant is calculated as the difference between the delays simulated by the slow and typical parameters, as follows.

$$s_k = (\text{slow delay}) - (\text{typical delay}) \tag{26}$$

where the absolute scale of s_k is not significant, but the relative values are significant, as can be seen in Eq. (25). When simulating the slow delay, the slow case parameters are assigned to only one transistor: the other transistors are simulated using the typical parameters. The slow case parameters are determined by the gate area [10, 9]. Although incurring higher computational cost, this method calculates the gate-delay variation with higher accurately than the other methods.

(B) *Substitution by typical s_k*

This method uses sensitivity constants calculated for typical slew and load conditions rather than all slew and load conditions. The sensitivity constant is calculated once for each gate and each input pin and pull-up/down state. Under the typical conditions, the gate-delay variation determined by this method is the same as that determined by the calculation of s_k in method (A) above.

(C) *Substitution by 0 or 1*

The delay variation depends on the transistor characteristics on the charging/discharging paths, and the sensitivity constants of the transistors are usually different between transistors. This method assumes that the sensitivity constants of transistors on the charging/discharging paths are approximately equal. The sensitivity constant s_k is defined as 1 for transistors on the charging/discharging paths, and 0 for transistors on other paths. The intra-gate-variability factor in Eq. (25) is then simplified as follows.

$$\frac{\sqrt{\sum_k^m s_k^2}}{\sum_k^m s_k} = \frac{\sqrt{ms^2}}{ms} = \frac{1}{\sqrt{m}} \tag{27}$$

where m is the number of transistors on the paths. In the method, the following model can be used to calculate the gate-delay variation rather than Eq. (25).

$$t_d = t_0 + \boldsymbol{\tau}_g^T \boldsymbol{x}_g + \frac{1}{\sqrt{m}} \tau_r x_r \tag{28}$$

This method does not require sensitivity analysis, and will incur larger error than method (B).

(D) *Calculation of s_k without intra-gate variability*

This method does not consider the intra-gate variability. In this case, the intra-gate-variability factor is 1, and the intra-chip delay variation is $\sum_k^m s_k / \sqrt{\sum_k^m s_k^2}$ times larger than that of method (A). The method also does not require sensitivity analysis.

3.2 Classification of Cell Structures

Three types of cell structures are employed for error evaluation; serial, parallel, and multi-stage, referring to the configuration of multiple transistors on the charging/discharging paths. An example of a serial structure is the pull-down path of a NAND gate, as exemplified in Fig. 3 for a 4-input NAND gate. NOR, AND-OR-INV, and OR-AND-INV gate, etc., also have serial structures.

As the cell height in a generic cell-library is uniform, larger transistors cannot be inserted into the cell. Instead, large transistors are folded and connected in parallel. An example of such as parallel structure is shown in Fig. 4 for a 4-folded INV gate (INVP040). Large NAND gates are often folded.

Similarly, large transistors for strong drivers have large input capacitance. To reduce the input capacitance, a multi-stage gate is often employed, which consists of multiple internal gates. Figure 5 shows a multi-stage INV gate with 3 internal INV-gates as an example of a multi-stage structure. A buffer gate is a kind of this structure.

3.3 Evaluation of Methods for Calculating the Sensitivity Constant

The estimated error of sensitivity calculation by methods (B), (C) and (D) are compared with method (A). A generic cell-library designed for a 0.13 μm CMOS process is considered. Figure 6 shows the sensitivity constants of a 4-input NAND gate (NAND4) at a fall time, as simulated by SPICE. The sensitivity constants s_1, s_2, s_3 and s_4 represent the constants of the nMOS transistors in the 4-input NAND gate in the order of connection from the output. As the input signal of input A rises, inputs B, C and D are fixed high. The x-axis represents the input transition time (10ps to 100ps), the y-axis represents the load capacitance (10fF to 100fF), and the z-axis represents sensitivity, the reference of which is s_3 of input C in Fig. 3. The sensitivity constant s_1 of input A, which rises, becomes large compared to the other inputs when the load capacitance is small and the input transition is large.

Figure 7 shows the error in the calculated delay variation for methods (B), (C) and (D) compared to method (A). In this evaluation, only the intra-chip and intra-gate variations are considered.

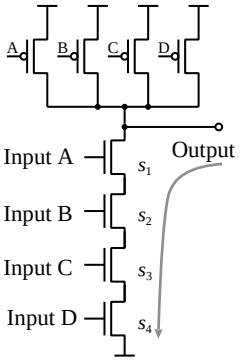

Figure 3: An example of a serial structure (NAND4)

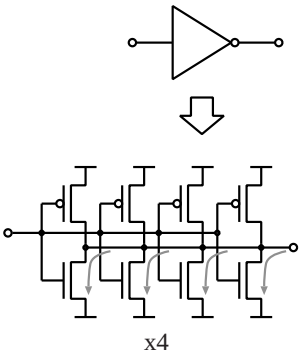

Figure 4: An example of a parallel structure (INVP040)

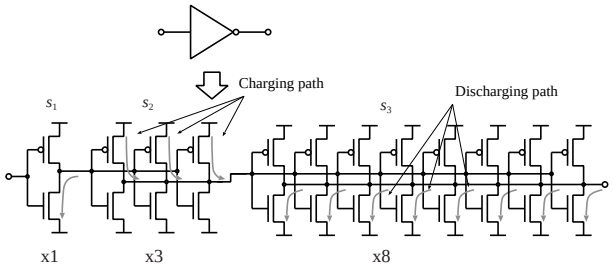

Figure 5: An example of a multi-stage structure (INVP080)

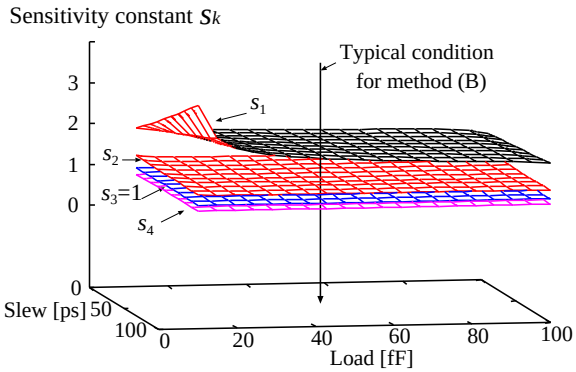

Figure 6: Sensitivity constants for a 4-input NAND gate (NAND4)

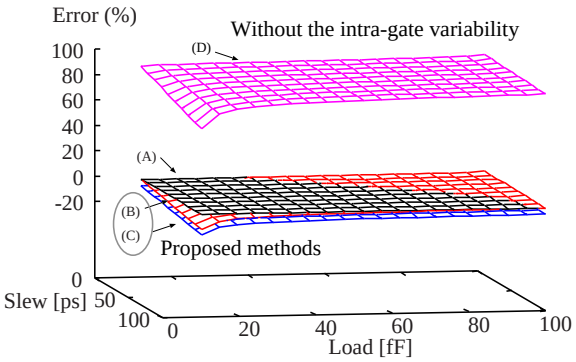

Figure 7: Error in gate-delay variation (NAND4)

The z-axis represents the standard deviation of delay variation, the reference of which is the standard deviation of method (A). The error in the delay variation becomes large when the load capacitance is small and the input transition is large because the sensitivity constants determined by methods (B), (C) and (D) are constant. Table 1 shows the maximum, average and minimum errors for Fig. 7 compared to method (A). For method (B), the set of sensitivity constants was determined for a typical load capacitance of 50fF and an input transition time of 50ps in Fig. 6. At a rise time, there is only one transistor on the charging path, and as such it is not necessary to consider the intra-gate variability because the intra-gate-variability factor in Eq. (25) becomes 1, as follows.

$$\frac{\sqrt{\sum_k^m s_k^2}}{\sum_k^m s_k} = \frac{\sqrt{s_k^2}}{s_k} = 1 \qquad (29)$$

However, when more than one input rises simultaneously, the characteristics of the multiple transistors on the charging path will influence the delay variation.

The sensitivity constants of parallel structures are equal because all transistors in a parallel structure are designed identically and connected to the same nets. The delay variations simulated by methods (A), (B) and (C) are therefore equal. The delay variation determined by method (D) is $\sqrt{m}$ times larger than that of method (A) according to Eq. (28).

Figure 8 shows the sensitivity constants for the multi-stage INV gate in Fig. 5 at a fall time. The sensitivity constants s_1, s_2 and s_3 are simulated for the 1st, 2nd and 3rd stages, respectively. Figure 9 shows the standard deviation, the reference of which is that of method (A). For method (B), the set of sensitivity constants was determined for a typical load capacitance of 400fF and an input transition time of 50ps in Fig. 8. Table 1 also shows the maximum,

Table 1: Error in gate-delay variation

	NAND4			INVP080		
method	min.	ave.	max.	min.	ave.	max.
(B)	−11.9%	**−0.76%**	+1.54%	−26.0%	**−1.12%**	+13.8%
(C)	−16.0%	**−5.42%**	−3.24%	−35.7%	**−12.1%**	−1.05%
(D)	+68%	**+89%**	+94%	+122%	**+204%**	+243%

average and minimum errors for Fig. 9 compared to method (A). The average error of method (B) is −1.12%, whereas the average error of method (C) is −12.1%. The error of method (B) also depends on the choice of typical conditions of load and slew.

As the sensitivity ratio of the multi-stage structure is larger than that of the serial structure, the difference in the delay variation between methods (B) and (C) is larger than for the serial structure. The choice between method (B) or (C) should therefore be made in consideration of accuracy and computational cost.

4. EXPERIMENTAL RESULT OF DELAY ANALYSIS

In this section, we show an experimental result for the evaluation of the intra-gate variability using our method. We present the importance for consideration of the intra-gate variability.

For comparison, circuit delays were simulated with and without consideration of the intra-gate variability. For these two simulations, the standard deviations of transistor characteristics were equal, which were iderived from measurements [9]. Sensitivity constant s_k was assumed to be 1.

Figure 10 shows the distribution of circuit delay simulated with and without consideration of intra-gate variability. The simulated

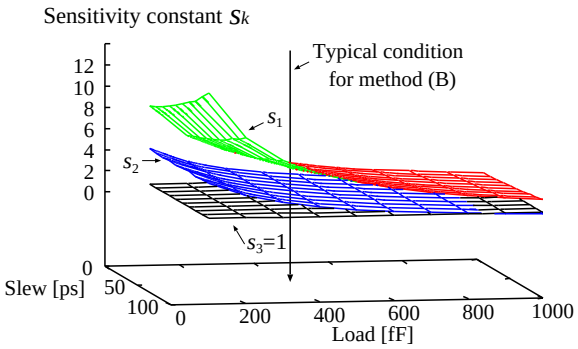

Figure 8: Sensitivity constants for a multi-stage INV gate (INVP080)

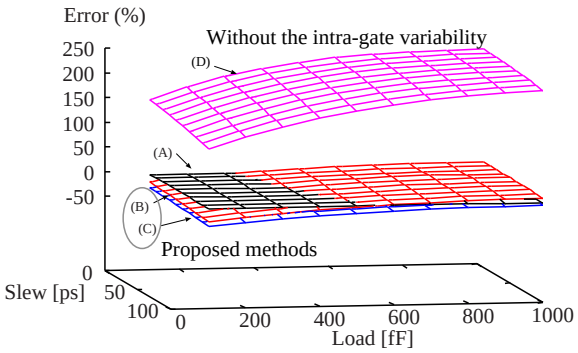

Figure 9: Error in gate-delay variation (INVP080)

circuit (des) consists of 3759 logic gates, and is a combinational circuit included in the LGSynth93 benchmark set. The broken line represents the distribution of circuit delay simulated without the relative, intra-gate, variability according to method (D), and the solid line represents the distribution of circuit delay calculated by the proposed model according to method (C). Both results are simulated by Monte Carlo analysis using a STA. The execution time for 1000-step Monte Carlo analysis was 14.2 seconds on a Pentium4 1.7GHz running Linux. The proposed model can also be applied to the other simulation methods [13, 14, 5, 6]. The error of mean value for the conventional method is 0.042 [ns], and for the worst-case value ($\mu + 3\sigma$) is 0.18 [ns]. When intra-gate variability is not considered, the delay variation is 31.8% larger. This demonstrates that it is necessary to consider the intra-gate variability in order to analyze gate-delay variation accurately.

5. CONCLUSION

A model for statistical calculation of gate-delay considering both intra-chip and inter-chip variability was proposed.

Every transistor on charging/discharging paths influences the gate delay. At the statistical modeling of gate delay, it is necessary to consider the intra-gate variability of transistor characteristics. The proposed model introduces sensitivity constants to facilitate the calculation of intra-gate variability without assigning variables to every individual transistor. The gate delay variation is calculated from the sensitivity constants and the delay variation simulated without consideration of the intra-gate variability. The influence of the intra-gate variability is expressed in terms of an

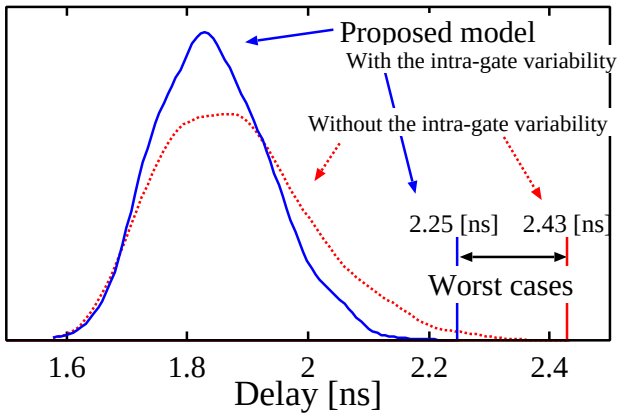

Figure 10: Comparison between delay distribution of proposed and conventional models (des circuit)

intra-gate-variability factor, which is calculated from the sensitivity constants.

Through circuit simulation, the proposed method was demonstrated to give a delay variation 31.8% smaller than when intra-gate variability is not considered, which is demonstrating that it is important to consider intra-gate variability in statistical timing analysis.

6. REFERENCES

[1] S. Nassif, "Within-chip variability analysis," *IEDM Tech. Digest*, pp. 283–286, Dec. 1998.

[2] S. Nassif, "Modeling and analysis of manufacturing variations," *Proc. CICC*, pp. 223–228, 2001.

[3] K. A. Bowman and J. D. Meindl, "Impact of within-die parameter fluctuations on future maximum clock," *Proc. CICC*, pp. 229–232, 2001.

[4] M. Orshansky, C. Spanos and C. Hu, "Circuit performance variability decomposition," *Proc. IWSM*, pp. 10–13, 1999.

[5] M. Berkelaar, "Statistical delay calculation, a linear time method," *Proc. TAU*, pp. 15–24, 1997.

[6] M. Hashimoto and H. Onodera, "A performance optimization method by gate sizing using statistical static timing analysis," *Proc. ISPD*, pp. 111–116, 2000.

[7] S. Tsukiyama, M. Tanaka and M. Fukui, "A statistical static timing analysis considering correlations between delays," *Proc. ASPDAC*, pp. 353–358, 2001.

[8] G. E. P. Box and N. R. Draper, "Empirical Model-Building and Response Surfaces," John Wiley & Sons, 1987.

[9] K. Okada, K. Yamaoka and H. Onodera, "A statistical gate delay model for intra-chip and inter-chip variabilities," *Proc. ASPDAC*, pp. 31–36, Jan. 2003.

[10] M. Pelgrom, A. Duinmaijer and A. Welbers, "Matching properties of MOS transistors," *IEEE J. Solid-State Circuits*, Vol. 24, No. 5, pp. 1433–1439, Oct. 1989.

[11] J. Bastos, M. Steyaert, R. Roovers, P. Kinget, W. Sansen, B. Graindourze, A. Pergoot and E. Janssens, "Mismatch characterization of small size MOS transistors," *Proc. ICMTS*, Vol. 8, pp. 271–276, 1995.

[12] M. Turk and A. Pentland, "Eigenfaces for recognition," *Journal of Cognitive Neuroscience*, Vol. 3, No. 1, pp. 71–86, 1991.

[13] R. Hitchcock, "Timing verification and the timing analysis program," *Proc. DAC*, pp. 594–604, 1982.

[14] H.-F. Jyu, S. Malik, S. Devadas and K. Keutzer, "Statistical timing analysis of combinational logic circuits," *IEEE Trans. VLSI Systems*, Vol. 1, No. 2, pp. 126–137, June 1993.