INTEGRATION, the VLSI journal 46 (2013) 345–358

Contents lists available at SciVerse ScienceDirect

INTEGRATION, the VLSI journal

journal homepage: www.elsevier.com/locate/vlsi

A gate-delay model focusing on current fluctuation over wide range of process–voltage–temperature variations

Ken-ichi Shinkai, Masanori Hashimoto *, Takao Onoye

Department of Information Systems Engineering, Graduate School of Information Science and Technology, Osaka University, 1-5 Yamadaoka, Suita, Osaka, Japan

ARTICLE INFO

Article history:
Received 20 February 2011
Received in revised form
14 September 2012
Accepted 19 January 2013
Available online 5 February 2013

Keywords:
Gate-delay model
Process–voltage–temperature (PVT) variations
Static timing analysis
Statistical timing analysis

ABSTRACT

This paper proposes a gate-delay model suitable for timing analysis that takes into consideration wide-ranging process–voltage–temperature (PVT) variations. The proposed model translates an output-current fluctuation due to PVT variations into modifications of the output load and input waveform. After translation, any conventional model can compute delay taking into account PVT variations by using the modified output load and reshaped input waveform. Experimental results with 90- and 45-nm technologies demonstrate that the average error of the fall and rise delay estimation in single- and multi-stage gates was approximately 5% on average over a wide range of input slews, output loads, and PVT variations. The proposed model can be used in Monte Carlo STA (static timing analysis) in addition to corner-based timing analysis. It can be also used in statistical STA to calculate the sensitivities of delays to variation parameters on-the-fly even when the nominal operating condition changes as well.

1. Introduction

Process variations have been predicted to become severer [1], while environmental variations such as voltage noise and temperature fluctuations are also becoming aggravated [2–5], as technology continues to advance. These sources are often referred to as PVT (Process–Voltage–Temperature) variations. Due to the effect of PVT variations, fabricated circuits perform differently from targeted one in the design phase, which prompted designers to develop a timing-analysis method that could be used to estimate variations in delay before fabrication. Conventional STA (static timing analysis) assumes worst case scenarios of PVT variations, and its estimates are often excessively pessimistic. Presently, SSTA (statistical STA) has been intensively studied [6–10]. It treats delay stochastically and is expected to provide realistic estimates.

In the meantime, recent chips have been designed with a lower supply voltage, V_{dd}, while power dissipation has been rather increasing [11]. This has forced designers to endure delay variations due to V_{dd} fluctuations [12,13]. As the demand for low-power devices has been increasing, post-fabrication techniques of reducing leakage and switching-power dissipation have often been used. For example, designs using DVS (dynamic voltage scaling) and variable V_{th} (threshold voltage) by body biasing are being incorporated into the SoC (system-on-a-chip) design flow. Intentional variations in V_{dd} and V_{th} associated with operation modes are generally larger in these designs than typical unintentional variations. Verification of timing in numerous operation modes and corners is urgently demanded, although the design time is tightly constrained. STA and SSTA are often used for such verifications, but it requires data of delay-characteristics in numerous corners, referred to as library data. Conventional gate-delay models that only cover a single mode/corner are totally insufficient for creating the whole library that covers a number of modes and corners. Statistical gate-delay and waveform-based models [14–17] have been proposed, and they characterize stochastic times when the output voltage of a gate reaches at certain levels. Thus, these models aim to give a gate-delay distribution around a given center corner for given distributions of process variations, and do not output a determinate delay value for a set of concrete process variations. Besides, Refs. [18,19] give a deterministic delay value for a set of concrete process variations. In particular, a model proposed in [19] copes with voltage variations as well as process variations, though most of the above models focus on process variations only. Such models that give a deterministic delay value are useful for implementing Monte Carlo STA (MC-STA) [20,21], since the gate delay and slew values must be computed for each set of concrete process variations. However, they are based on CSMs (current source models), and hence they are less compatible with frameworks of timing analysis that use conventional gate-delay models, such as sensitivity-based model and table look-up model.

As the discussion above, accurate and reasonable modeling of gate-delay variations is required to consider PVT variations in

* Corresponding author. Tel.: +81 6 6879 4526; fax: +81 6 6879 4529.
E-mail address: hasimoto@ist.osaka-u.ac.jp (M. Hashimoto).

0167-9260/$ - see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.vlsi.2013.01.003

STA, MC-STA and/or SSTA. This paper presents a gate-delay model that copes with wide-ranging PVT variations and has compatibility with conventional models. The key point is that the proposed model directly focuses on an output-current fluctuation and not on a delay variation. Once a set of determinate PVT variation values are given, the proposed model translates the current fluctuation caused by PVT variations into the output load on the basis of the fact that gate delay is the time required to charge/discharge output load. In addition, the input waveform is shifted to compensate for the timing difference that is irrelevant to the current fluctuation. The proposed model can be used as a wrapper to any conventional models with a small amount of additional characterization cost.

The proposed model preliminarily reported in Ref. [22] works well in most conditions. However, we found that the accuracy degraded in some cases where the input transition was very slow. We improved the accuracy in such cases by making two improvements: (1) More accurately computing the input voltage at timing when the output voltage goes across 50% of V_{dd}. (2) Calculating the delay offset not dependent on current variations more properly. In this work, we verified the proposed model with 45-nm process technology in addition to 90-nm one. Characterization costs of the proposed and conventional models were evaluated. Moreover, we explain how we used the proposed model to calculate the sensitivity for SSTA, and verify that on-demand sensitivity updating is helpful for enhancing accuracy.

The rest of the paper is organized as follows. Conventional gate-delay models are outlined and their inefficiency in variation-aware analysis is described in Section 2. Section 3 first overviews the proposed model and then explains details on its calculation steps. To corroborate accuracy, Section 4 presents the experimental results in 90- and 45-nm process technologies. Section 5 discusses the demonstration in applying the proposed model to on-demand sensitivity recalculations in SSTA. Finally, the discussion is concluded in Section 6.

2. Problems with conventional gate-delay models

To construct variation-aware timing analysis, four features are usually required for a gate-delay model:

1. Small migration cost from conventional model.
2. Small characterization cost.
3. Capability of handling the increase in the number of variation parameters.
4. Capability of being utilized in SSTA.

Here, we explain conventional gate-delay models and discuss their advantages and disadvantages with respect to the above requirements.

For handling PVT variations, a naive approach is to increase the number of parameters in the table look-up model. Fig. 1 shows an example of 2D (two-dimensional) table look-up model with regard to input slew and output load of a gate. If input slew is 10 ps and output load is 4 fF, for example, 28 ps of gate-delay is calculated. However, the table look-up model is obviously difficult to employ, because its characterization cost explodes exponentially as the number of variation parameters increases.

The sensitivity-based model using Taylor expansion shown in Eq. (1) is a standard approach to variational analysis

$$d = d_0 + \sum_{\forall p_i} \left.\frac{\partial d}{\partial p_i}\right|_{p_i = p_{i\_0}} \Delta p_i, \tag{1}$$

where d is the delay, d_0 is the nominal value of d, and p_i is the i-th variation parameter. Here, $(\partial d/\partial p_i)|_{p_i = p_{i\_0}}$ is the delay sensitivity

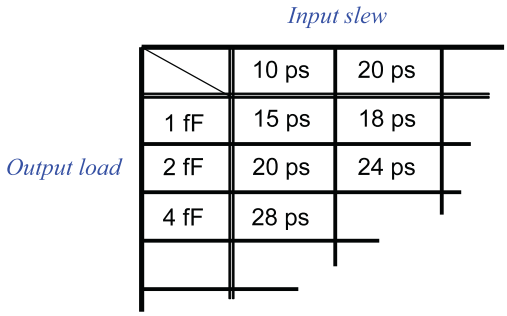

Fig. 1. Table look-up model.

where parameter p_i equals a nominal value, i.e., p_{i_0}, and Δp_i is the difference of p_i from p_{i_0}. The sensitivity-based model is widely used in SSTA [6,7,23] and parametric STA [24,25], because it has advantageous statistical characteristics [7] and can consistently cope with various parameters. However, sensitivities depend on the output load and input-signal waveform, and hence a number of additional transient analyses are necessary for characterization, which is computationally very expensive. Another problem is that the sensitivity-based model is accurate as long as parameter variations are small, but might not be tolerant to large variations due to the linear approximation. Moreover, the sensitivity computed under the nominal condition by using Eq. (1) might differ greatly from sensitivity under actual operating conditions, such as where the average V_{dd} declines.

3. Proposed gate-delay model

A feature of the proposed model is that it focuses on current-variation modeling instead of direct-delay modeling. The delay in a CMOS (complementary metal oxide semiconductor) logic gate is the time required to charge/discharge its output load, and hence an output-current variation caused by PVT variations changes the gate delay. In this section, the relations between gate delay, output current, and variation sources are reviewed to explain why this work focuses on the output-current variation. After that, this section outlines the proposed model and describes details on the computational procedures.

3.1. Relations between gate delay, output current, and parameter variations

The saturation current of a MOSFET (MOS field-effect transistor), I_{dsat}, is usually expressed in the α-power-law MOSFET model [26]

$$I_{dsat} = k\frac{\mu\epsilon_{ox}W}{t_{ox}L}(V_{gs} - V_{th})^{\alpha}, \tag{2}$$

where V_{gs} means a gate-source voltage, μ denotes the effective mobility, ϵ_{ox} is the dielectric constant of the gate oxide, and t_{ox} is the gate-oxide thickness. Here, W is the channel width and L is the channel length. α is a coefficient to express the carrier-velocity saturation effect, and this reaches close to one in advanced technologies. k is also a coefficient. Here is an example of saturation current variations in 90-nm CMOS technology in Fig. 2. It is obvious that the output current has an almost linear relation to variation sources and is tractable.

However, gate delay is basically inversely proportional to current [27]. Intuitively, gate delay is also inversely proportional to $(V_{dd} - V_{th})$ considering Eq. (2). This means that the sensitivity-based delay model is not suitable for large variations of V_{dd} and V_{th}, whereas it really works well for small variations. To develop a gate-delay model that can be applied to a wide range of variations

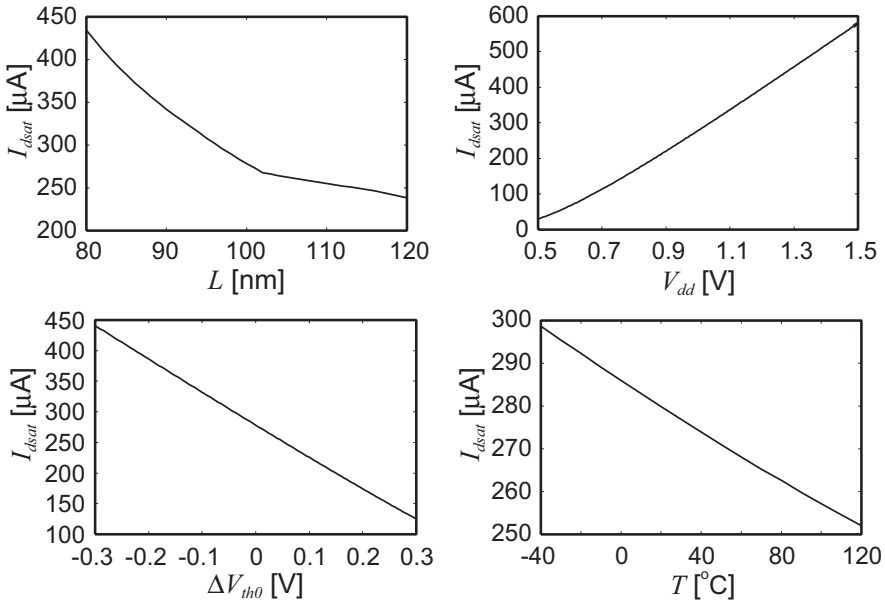

Fig. 2. I_{dsat} at $V_{gs} = V_{ds} = V_{dd}$ vs. PVT variations in 90-nm process technology (L: channel length, V_{dd}: supply voltage, ΔV_{th0}: variation in V_{th0} from its nominal value, and T: temperature).

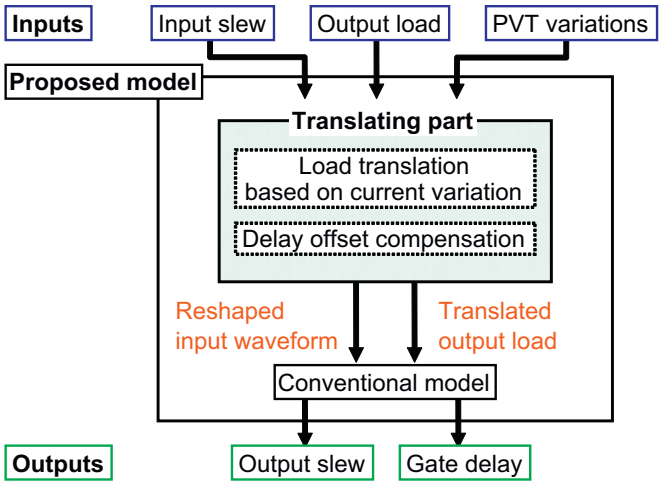

Fig. 3. Overview of flow for proposed model.

of V_{dd} and V_{th}, which is eagerly required in DVS and variable V_{th} design, focusing on and modeling a current variation represents a reasonable approach.

3.2. Overview of proposed gate-delay model

Fig. 3 overviews the flow for the proposed model, which translates three inputs (input slew, output load, and a set of PVT variations) into two outputs (output slew and gate delay). In other words, PVT variation information is embedded in the translated load and reshaped input waveform. Using them, any conventional model can compute delay taking PVT variations into consideration. Note that PVT variations in this work are not random variables but determinate parameter-differences from their nominal values. Input slew and output load are also determinate values. The proposed gate delay model provides a gate delay for a set of determinate device parameters given to each transistor, which means that the proposed model is independent from the device-parameters correlation. In fact, if there is a correlation, a model user generates samples (sets of device-parameters) according to the device-parameter distributions with

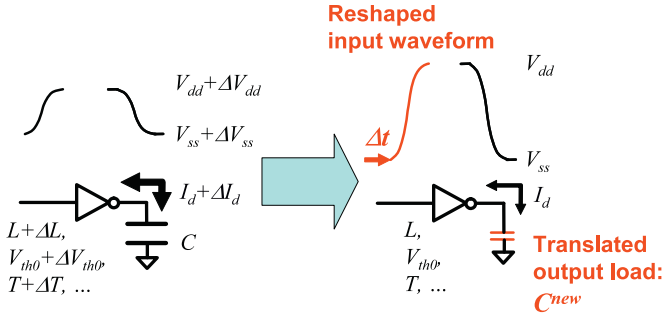

Fig. 4. Operation of translating part.

Fig. 5. Conceptual diagram that shows how PVT variations affect the current waveform. A solid line denotes a waveform without PVT variations. Dashed and dotted lines are waveforms suffering from PVT variations.

correlations and gives the samples to the gate delay model. This property is exploited in MC-STA. The proposed model gives determinate delay values for each sample. Thus, the proposed model is independent with how to handle correlations between process variables for path or circuit delay analyses. .

Fig. 4 illustrates the operation of the translating part in Fig. 3. All variations are eliminated by reshaping the input waveform and replacing output load as shown in the right side of Fig. 4. The modification of the input waveform corresponds to scaling it in terms of voltage and shifting it in time axis.

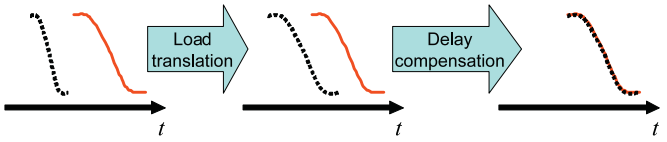

Fig. 6. Detail of translation on output waveform. Solid lines denote actual waveform with PVT variations. Dashed lines are waveforms in case that all PVT variations are removed. By applying load translation and delay compensation, the dashed line approaches to the actual waveform. Note that the actual waveforms are normalized to nominal V_{dd}.

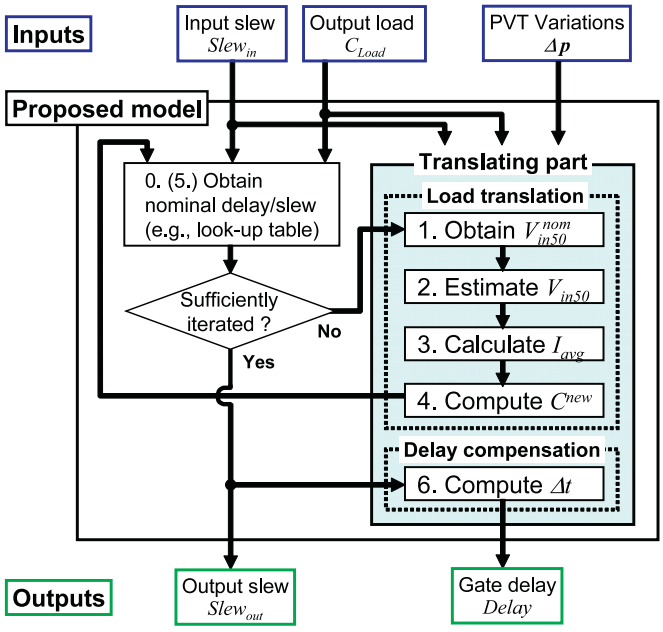

Fig. 7. Entire proposed procedure to calculate delay.

Here, it is important that this work focuses on the fluctuation in output current I_d, and not directly on delay variations. As shown in Fig. 5, PVT variations scale the current waveform up or down, and shift the starting time for transistor switching. This fact prompts the proposed model to translate output load on the basis of current variations, and to compensate for such time-shifting effects. Fig. 6 shows how the output waveform is altered at the two steps of the translating part in Fig. 3. The solid lines represent the actual waveform with PVT variations, and the dashed lines correspond to the waveform in the case that all the PVT variations are removed. In the left of Fig. 6, all the variations are simply removed, and then two waveforms are different. After the first step of load translation, the output slew is adjusted, but the timing is still different. The second step of delay compensation shifts the waveform, and finally the delay is well estimated. In the proposed model, the input waveform is not scaled with regard to time axis.

Fig. 7 illustrates a complete flowchart of the entire procedure. Load translation is carried out in Steps 1–4 and delay compensation in Step 6. The information required for this procedure, such as I_{avg}, will be described in the following sections. The proposed model needs two prior preparations. The first involves constructing a conventional delay model that takes the input transition time and output load as indices. This is used for calculating the nominal delay in Steps 0 and 5. This work assumes that a conventional 2D look-up table would be used to calculate nominal delay, even though other models can similarly be incorporated. The second involves characterization of the output current used in Step 3. This characterization only requires a limited number of DC (direct current) analyses independent of the input slew and output load, and thus its characterization cost is quite small.

The advantages of the proposed model associated with the demands discussed in Section 2 are four-fold: (1) The gate delay can be computed in any conventional models by using the translated output load and reshaped input waveform, and hence all timing analyzers can carry out variation-aware timing analysis by pre-processing the proposed model. (2) The additional characterization cost is quite limited. No time-consuming transient analysis is required for additional characterization. The modeling of current variations only needs a relatively small number of DC analyses. (3) Variation parameters are treated similarly as a factor that varies the output current. Therefore, when the number of parameters increases, the proposed model can handle them consistently. (4) The proposed model provides accurate results over a wide range of parameter variations, because current variations are easily and precisely modeled. Therefore, it is suitable not only for variation-aware STA but also for DVS and V_{th} control design. The large range of variations to which it can be applied may eliminate delay characterizations in numerous corners. By using the proposed model, libraries at PVT corners can be generated. In addition, when an actual operating condition varies from the nominal condition, the proposed model can update delay sensitivities.

The following sections describe the details on each step in Fig. 7. It begins with Step 4 and goes through Step 3 to Step 1. After that, this section explains why Steps 0–5 need to be iterated, and ends with Step 6. This order is irregular but easier to follow, because the description of a step explains why a parameter, such as I_{avg}, calculated in the previous step is required. To simplify explanations in these sections, the following (1)–(4) are assumed. (1) The fall delay of a single-stage inverter will be estimated. (2) The V_{dd} variations on the driver and receiver sides are the same. (3) Ground voltages V_{ss} have a constant value of 0 V. Note that the proposed model can also be applied to other gates and the rise delay. The mismatches of V_{dd} and V_{ss} at each pin between the driver and receiver sides can be considered in the proposed model, and the experiments in Section 4 take these into account. (4) An input waveform is monotonic. The shape of input waveform is dependent on the gate-delay model used for nominal delay computation.

When the given input waveform is totally different from that assumed in the gate-delay model, for example a non-monotonic waveform due to coupling noise, waveform approximation to a monotonic waveform by a method, such as in Refs. [28] and [29] might be necessary, since the proposed model could not directly deal with noises.

Another approach could be an extension to handle superposition-based model for capacitive coupling [30]. With this model, the noise waveform at the victim is derived for each aggressor. These noise waveforms are superposed to the victim transition waveform without aggressor transitions, and the victim transition waveform with aggressor transition is obtained. In this case, the victim transition waveform without aggressor transitions can be computed by the proposed model. The remaining issue is the noise waveform at the victim. A possible computation might be: (1) the equivalent circuit of the aggressor load, which includes coupling capacitance, is scaled by the similar idea in Eq. (5) and then (2) conventional noise computation, such as one with the Thevenin resistance expression, is performed assuming that PVT variation is removed by the load translation. The extension for coping with coupling noise is one of our future works.

3.3. Step 4: translation of current variations into output load

This section explains the most important part of the proposed model, i.e., how to translate current fluctuations into output load.

Generally speaking, the delay varies due to a current fluctuation caused by PVT variations, while output-load variations also change the delay. When the output current becomes large due to PVT variations, the delay decreases. On the other hand, when the output load shrinks, the delay also decreases. Hence, these two cases are highly related in terms of delay. This means current variations caused by PVT variations can be translated into output load, while gate delays, strictly speaking the output slew, before and after translation are equal. This load translation is outlined in Fig. 4. Here, I_d denotes the output current which mostly characterizes the performance of a gate, as will be discussed in the next section. Current variation ΔI_d caused by PVT variations is translated into a new output load, C^{new}. After translation, the gate is supposed to be under a nominal condition, and therefore its output slew can be calculated with any conventional delay model. The delay is computed considering the delay offset in Step 6.

First, the case where output load is capacitive is explained. Before translation, the amount of the whole charge, Q, that must be discharged is equal to CV_{dd}, where C is the total capacitance to discharge, i.e., the sum of the load capacitance, C_{Load}, and parasitic capacitances[1] inside the cell. Considering that Q varies with V_{dd} fluctuations, C^{new} is expressed by Eq. (4) from the relation of Eq. (3)

$$(I_d+\Delta I_d) : I_d = Q : Q^{new}, \\ = C(V_{dd}+\Delta V_{dd}) : C^{new}V_{dd}, \tag{3}$$

$$\therefore\ C^{new} = \frac{I_d}{I_d+\Delta I_d}\frac{V_{dd}+\Delta V_{dd}}{V_{dd}}C, \tag{4}$$

where Δ denotes a parameter variation, and superscript^{new} corresponds to a translated value under a nominal condition. As a simple example, it is supposed that parameter variations double the current ($\Delta I_d = I_d$) and the supply voltage is constant ($\Delta V_{dd} = 0$). Eq. (4) means this translation reduces the output load to half instead of doubling the current so that the estimated output slew approaches to the actual one. The delay and output slew can be calculated with any conventional delay models using C^{new}, where the delay estimation will be further improved in Step 6.

The idea behind this load translation can be consistently applied to the generic output load. For example, when the output load is resistive interconnects, the capacitance in Eq. (4) should be replaced with admittance at the driving point. This is expressed by Eq. (5)

$$Y^{new} = \frac{I_d}{I_d+\Delta I_d}\frac{V_{dd}+\Delta V_{dd}}{V_{dd}}Y, \tag{5}$$

where Y is the driving-point admittance of the actual circuit, and Y^{new} is the translated admittance. On-chip interconnects are often compacted to CRC π model in Fig. 8 for gate-delay computation [31]. C_1^{new}, C_2^{new} and R^{new} can simply be expressed as follows[2]:

$$C_1^{new} = \frac{I_d}{I_d+\Delta I_d}\frac{V_{dd}+\Delta V_{dd}}{V_{dd}}C_1, \tag{6}$$

[1] This work considers parasitic capacitances, such as the overlap and fringing capacitances between gate and drain/source terminals, and diffusion capacitances of drain/source terminals. Each drain/source terminal associated with the output transition is taken into account. These capacitances are incorporated into output capacitance considering the Miller effect.

[2] We considered that the admittance of a π circuit is $(j\omega C_2+1/(R_1+(1/j\omega C_1)))$. Thus

$$Y^{new} = \left(j\omega C_2^{new} + \frac{1}{R_1^{new}+\frac{1}{j\omega C_1^{new}}}\right).$$

Substituting Eq. (5) into this, we obtain Eqs. (6)–(8).

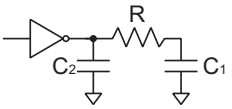

Fig. 8. CRC π model.

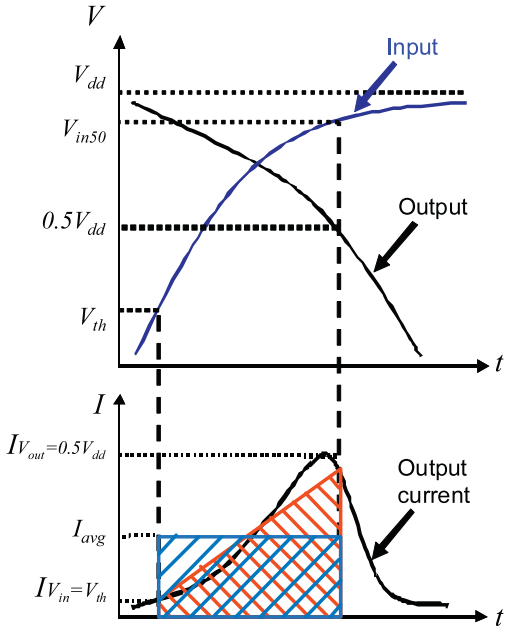

Fig. 9. Typical current waveform.

$$C_2^{new} = \frac{I_d}{I_d+\Delta I_d}\frac{V_{dd}+\Delta V_{dd}}{V_{dd}}C_2, \tag{7}$$

$$R^{new} = \frac{I_d+\Delta I_d}{I_d}\frac{V_{dd}}{V_{dd}+\Delta V_{dd}}R. \tag{8}$$

Note that Eq. (4) is one of the special cases of Eq. (5) with $R=0$.

3.4. Step 3: estimating current variations

Next, the estimate of I_d used for load translation in Eqs. (4) and (5) is explained.

3.4.1. Current used for load translation

The delay is generally determined by the shape of the current in the interval between the time when an input voltage, V_{in}, equals V_{th} and the time when the output voltage, V_{out}, goes across 50% of V_{dd}. Therefore, the average current in this interval determines the delay and is appropriate for I_d in Eqs. (4) and (5). To explain how to compute it, let us take an example of a typical current waveform in Fig. 9. Strictly speaking, the average current mentioned above should be numerically calculated by the integral of the current shape, but its computation is difficult because it requires a huge amount of variational current characterizations and pseudo-transient analyses. Instead, it can be estimated by a first-order approximation. The integral of the current shape is approximated as a trapezoid, and then the integrated area is equal to the area of a rectangle whose height is I_{avg} in Eq. (9)

$$I_{avg} = \frac{1}{2}(I_{V_{in}=V_{th}} + I_{V_{out}=0.5V_{dd}}), \tag{9}$$

where $I_{V_{in}=V_{th}}$ represents the output current when V_{in} equals V_{th}, and $I_{V_{out}=0.5V_{dd}}$ is that when V_{out} is 50% of V_{dd}.

It is experimentally verified that this approximation is reasonable in most combinations of C_{Load}, $Slew_{in}$, and parameter variations. Therefore I_{avg} is used as I_d. Instead, $I_{V_{out}=0.5V_{dd}}$ can be used as I_d, because in most situations $I_{V_{in}=V_{th}}$ is negligibly small. $I_{V_{in}=V_{th}}$ should be included to cover the situation that V_{dd} is close to V_{th}.

3.4.2. Estimation of current

Designers need to prepare the flexible and robust function of $I_d = f(\mathbf{\Delta p})$ against the increase in variation parameters. In this work, the response-surface method [32] is employed as a candidate to estimate I_{avg}, because the order of polynomials could be chosen to satisfy the required accuracy. However, other methods can also be used as long as their accuracy and cost are reasonable. The polynomials for $I_{V_{out}=0.5V_{dd}}$ is expressed in Eq. (10)

$$I_{V_{out}=0.5V_{dd}} = f(\mathbf{\Delta p}, V_{in50}). \tag{10}$$

Constructing Eq. (10) does not need any transient analyses for the combinations of various input waveforms and output loads, and only needs a limited number of DC analyses. The parameter left unknown is V_{in50}, i.e., V_{in} at $V_{out}=0.5V_{dd}$. The next section will explain how V_{in50} is computed.

3.5. Step 2: estimate of V_{in50}

It is generally difficult to accurately determine V_{in50}, because V_{in50} depends on the input slew, output load, and PVT variations. For easy modeling, this section demonstrates how to approximate V_{in50} using V_{in50}^{nom}, where V_{in50}^{nom} is V_{in50} after the output load C_{Load} has been replaced with C^{new}. After the output load replacement, the supply voltage differs as seen in Fig. 4, and hence there must be an offset, ΔV_{in50}, as plotted in Fig. 10

$$V_{in50} = V_{in50}^{nom} + \Delta V_{in50}. \tag{11}$$

V_{in50}^{nom} should be calculated with C^{new} and $Slew_{in}$, although C^{new} is still unknown. To solve this problem, an iterative approach is taken, which will be explained in Section 3.7. The following discusses how ΔV_{in50} is estimated.

This section derives a single expression of ΔV_{in50} that covers a wide range of input and output slews. It was experimentally observed that ΔV_{in50} strongly depends on the *Slew* ratio, which is defined as $Slew_{out}/Slew_{in}$. Therefore, ΔV_{in50} is expected to be between two extreme cases of ΔV_{in50} in terms of the *Slew* ratio, and is computed by averaging them with a weight related to the *Slew* ratio. One of them is $\Delta V_{in50,s}$ where input translation is much slower than output translation, and another is $\Delta V_{in50,f}$ in the opposite case. However, the estimate of V_{in50} cannot directly use the *Slew* ratio as a weighting factor, because $Slew_{out}$ is unknown in Step 2. It was also experimentally found that V_{in50}^{nom} has a strong correlation with the *Slew* ratio, as shown in Fig. 11. This is because

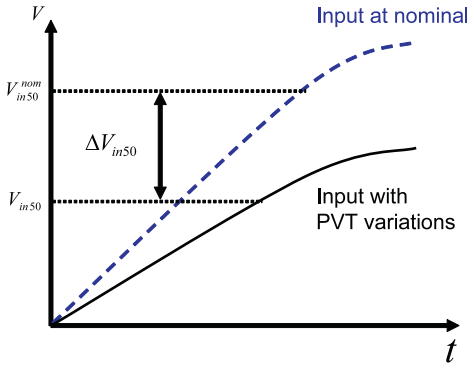

Fig. 10. Relation between V_{in50} and V_{in50}^{nom}.

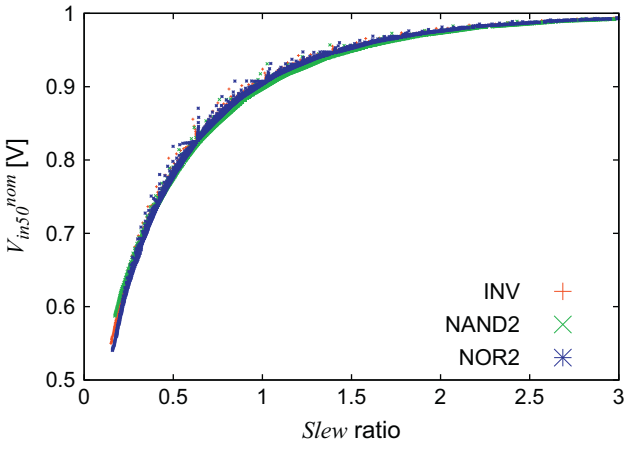

Fig. 11. Relation between *Slew* ratio and V_{in50}^{nom} with INV (INVerter), NAND2 (two-input NAND), and NOR2 (two-input NOR) when a rise-input waveform is injected. This evaluation was carried out with 9800 combinations of C_{Load} and $Slew_{in}$.

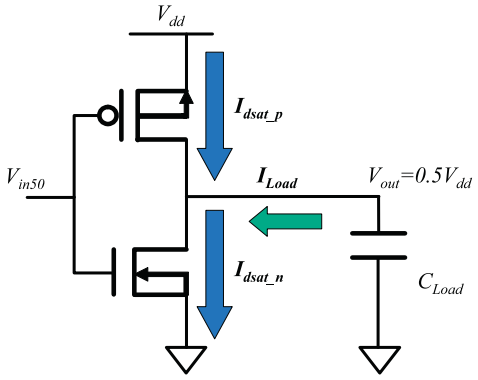

Fig. 12. Currents in inverter when output reaches $0.5V_{dd}$.

when V_{in50}^{nom} is large, the input transition finishes earlier than the output transition, and vice versa.

Considering the above and Eq. (11), a generalized form of V_{in50} is constructed using V_{in50}^{nom} as a weighting factor instead of the *Slew* ratio

$$V_{in50} \sim V_{in50}^{nom} + (r\Delta V_{in50,f} + (1-r)\Delta V_{in50,s}), \tag{12}$$

$$r = \frac{V_{in50}^{nom} - V_{in50,s}^{nom}}{V_{in50,f}^{nom} - V_{in50,s}^{nom}}, \tag{13}$$

where $V_{in50,s}^{nom}$ ($V_{in50,f}^{nom}$) means V_{in50}^{nom} when the input transition is much slower (faster) than the output transition. r is the empirical coefficient that is an alternative to the *Slew* ratio.

3.5.1. Estimate of $V_{in50,s}$ (V_{in50} when $Slew_{in} \gg Slew_{out}$)

Where the output load is very small and the input slew is much larger than the output slew, inverter operation is determined by DC rather than AC (alternating current) characteristics. When V_{out} reaches $0.5V_{dd}$ as seen in Fig. 12, the current I_{Load} from C_{Load} would be negligibly small. It is assumed that both of NMOS and PMOS turn on and $|I_{dsat\_n}| = |I_{dsat\_p}|$. Considering the channel-length modulation, these saturation currents are expressed in Eqs. (14) and (15) on the basis of α-power-law MOSFET model [33]

$$I_{dsat\_n} = \frac{\beta_n}{2}(V_{in} - V_{th\_n})^{\alpha_n}(1 + \lambda_n V_{out}), \tag{14}$$

$$I_{dsat\_p} = -\frac{\beta_p}{2}(V_{dd} - V_{in} - |V_{th\_p}|)^{\alpha_p}(1 + \lambda_p(V_{dd} - V_{out})), \tag{15}$$

where β is a transconductance coefficient and λ denotes a channel-length modulation parameter. For simplicity, it is assumed that the current characteristics of NMOS and PMOS are the same, i.e., $\beta_n=\beta_p$, $\alpha_n=\alpha_p=\alpha$ and $\lambda_n=\lambda_p=\lambda$. Here, $V_{in50,s}$ and $\Delta V_{in50,s}$ can be expressed by Eqs. (16) and (17)

$$V_{in50,s}=\frac{V_{dd}+V_{th\_n}-|V_{th\_p}|}{2}. \tag{16}$$

$$\Delta V_{in50,s}=\frac{\Delta V_{dd}+\Delta V_{th\_n}-|\Delta V_{th\_p}|}{2}. \tag{17}$$

3.5.2. Estimate of $V_{in50,f}$ (V_{in50} when $Slew_{in} \ll Slew_{out}$)

Where the input slew is much smaller than the output slew, V_{in} has reached V_{dd} by the time when V_{out} reaches $0.5V_{dd}$. Thus, $V_{in50,f}$ and $\Delta V_{in50,f}$ only depend on V_{dd} variations and are expressed in Eqs. (18) and (19)

$$V_{in50,f}=V_{dd}+\Delta V_{dd}, \tag{18}$$

$$\Delta V_{in50,f}=\Delta V_{dd}. \tag{19}$$

3.6. Step 1: obtaining V_{in50}^{nom}

V_{in50}^{nom} can easily be obtained without any transient simulations, because the information required to compute it is available before Step 1. This information, i.e., $Slew_{in}$, $Delay$, and $Slew_{out}$, can be calculated using the input and output waveforms estimated in Step 0.

3.7. Steps 0–5: iteration

Iterative calculations are necessary in the proposed procedure for load translation. To accurately estimate C^{new}, precise V_{in50}^{nom} is required. However, to precisely estimate V_{in50}^{nom}, accurate C^{new} is essential. This is a kind of chicken-and-egg problem. Hence, V_{in50}^{nom} is computed with C_{Load} in the first iteration and is recalculated with C^{new} in subsequent iterations. Two or three iterations were empirically sufficient to accurately estimate C^{new}.

3.8. Step 6: delay offset compensation

The delay and output slew are obtained after exiting iteration. Unfortunately, the delay obtained differs from what are desired to know. This section first explains why such differences arise, and next presents ways of compensating for them.

The proposed method aims to be applied to a wide range of V_{dd} and V_{th} variations, while the time when the output transition starts could be considerably varied in such a situation. Here is a case that the supply voltage has dropped to $V_{dd}+\Delta V_{dd}$, as shown in Fig. 13. It is clear that "input at nominal" (input after the proposed translation) goes across a certain voltage, V_x (e.g., V_{th}), earlier than "input with ΔV_{dd}" (input before the translation), because the proposed model stretches an input waveform in terms of the voltage, keeping the input transition times unchanged. In contrast, when the supply voltage is larger than its nominal value, an input waveform is shrunk and "input at nominal" comes later than "input with ΔV_{dd}". Such a difference in time before and after the proposed load translation causes the crossing timing offset, Δt, which is irrelevant to current variations. Hence, Δt should be separately computed. A similar problem also occurs with V_{th} variations.

From the above, the correct delay, $Delay$, should be compensated. As shown in Fig. 4, $Delay$ is expressed by Eq. (20)

$$Delay=Delay_{pro}+\Delta t, \tag{20}$$

where $Delay_{pro}$ is the delay estimated with C^{new} without compensation. In principle, Δt is simply regarded as the difference between two times. First one is $t_{50,b}$ that means the time when V_{out} reaches $0.5V_{dd}$ before load translation, and the other is $t_{50,a}$ that is the time when V_{out} does so after load translation. Therefore Δt is defined as Eq. (21)

$$\Delta t=t_{50,b}-t_{50,a}. \tag{21}$$

When $Slew_{in} > Slew_{out}$, $t_{50,b}$ can be computed using V_{in50} in Step 2, and then Δt can be obtained, which will be explained in the following section. However, when $Slew_{in} < Slew_{out}$, $t_{50,b}$ cannot be computed using V_{in50}, because V_{in50} may be V_{dd}. Thus, the calculation of Δt should be changed according to the $Slew$ ratio. The proposed model takes an approach similar to that employed in Section 3.5, i.e., it estimates Δt in two cases and empirically generalizes Δt using V_{in50}^{nom} similarly to ΔV_{in50} in Eq. (12). This is expressed as Eq. (22)

$$\Delta t \sim r\Delta t_f+(1-r)\Delta t_s. \tag{22}$$

Here, Δt_s and Δt_f are Δt where the input transition is much slower and faster than the output transition, respectively. r is the same as that in Eq. (13). Fortunately, the information required to compute Δt is given by STA or calculated in the previous steps with the proposed model, which will be explained in what follows.

3.8.1. Estimation of Δt_s (Δt when $Slew_{in} > Slew_{out}$)

Fig. 14 shows a case where $Slew_{in}$ is larger than $Slew_{out}$. As mentioned above, Δt_s can be expressed by

$$\Delta t_s=t_{50,b}-t_{50,a}. \tag{23}$$

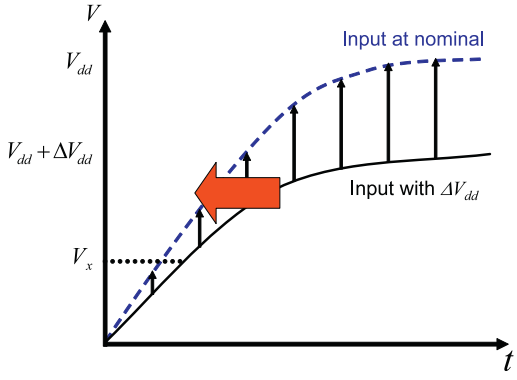

Fig. 13. PVT variations shift timing when MOS turns on/off, and beginning of output transition varies. *Note*: In Figs. 13–15, the solid and dashed lines denote respective waveforms before and after load translation. These figures show a case that the supply voltage has dropped to $V_{dd}+\Delta V_{dd}$.

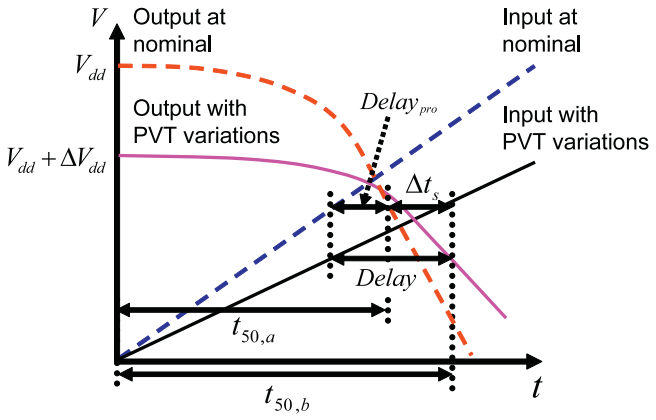

Fig. 14. Δt_s calculation.

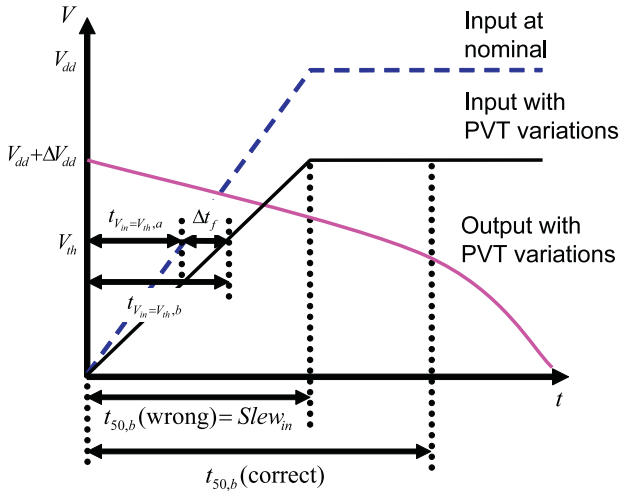

Fig. 15. Δt_f calculation.

$t_{50,b}$ is easily obtained with V_{in50} and $Slew_{in}$. For instance, if the input is a ramp waveform, $t_{50,b}$ is computed by

$$t_{50,b} = Slew_{in}\frac{V_{in50}}{V_{dd}+\Delta V_{dd}}. \quad (24)$$

$t_{50,a}$ is also readily computed by

$$t_{50,a} = 0.5Slew_{in} + Delay_{pro}. \quad (25)$$

3.8.2. Estimation of Δt_f (Δt when $Slew_{in} < Slew_{out}$)

The $t_{50,b}$ computation using V_{in50} in Eq. (24) can be used as long as $V_{in50} < V_{dd}$. If Eq. (24) is applied to the case in Fig. 15, the estimated $t_{50,b}$ ($=t_{50,b}$(wrong)) is different from the correct $t_{50,b}$ ($=t_{50,b}$(correct)). Thus, Eq. (23) cannot be used in this case.

Here, another approach is adopted to estimate Δt_f. Recall that output transition is supposed to start when V_{in} exceeds V_{th}. Therefore, Δt_f can be estimated as the difference in timing when V_{in} traverses the threshold voltage before and after load translation, i.e., $t_{V_{in}=V_{th},b}$ and $t_{V_{in}=V_{th},a}$.

$$\Delta t_f = t_{V_{in}=V_{th},b} - t_{V_{in}=V_{th},a}. \quad (26)$$

For example, when the input is a ramp waveform, these are calculated by

$$t_{V_{in}=V_{th},b} = Slew_{in}\frac{V_{th}+\Delta V_{th}}{V_{dd}+\Delta V_{dd}}, \quad (27)$$

$$t_{V_{in}=V_{th},a} = Slew_{in}\frac{V_{th}}{V_{dd}}. \quad (28)$$

4. Experimental results

This section discusses the accuracy and characterization cost of the proposed gate-delay model with 90- and 45-nm process technologies. Here, a 90-nm industrial standard cell library and a 45-nm open-source standard cell library [34] are used for evaluation.

4.1. Accuracy of delay and output slew estimates

4.1.1. Experimental conditions

Five gates are used for the evaluation, which are three simple single-stage gates (INV, NAND2, and NOR2) and two complex single-stage/multi-stage gates (AND2 (two-input AND) and AOI21 (two-input AND-OR-INV)). It is assumed that L, ΔV_{th0}, V_{dd}, V_{ss}, and T vary simultaneously under the condition in Table 1. All parameters are normally distributed and independent. Note that this assumption only defines the experimental conditions and limits neither the range nor conditions to which the proposed model can be applied. L and ΔV_{th0} differ transistor by transistor, and ΔT varies gate by gate. V_{dd} and V_{ss} of driver and receiver sides fluctuate separately. In particular, V_{dd}, V_{ss}, and V_{th} have widely varied, because the proposed model also aims to cover the DVS and variable V_{th} design.

Table 1
Experimental conditions for PVT variations.

Variation parameter	90-nm technology		45-nm technology	
	μ	3σ	μ	3σ
Channel length L	100 nm	20 nm	50 nm	10 nm
V_{th0} offset ΔV_{th0}	0 V	0.3 V	0 V	0.3 V
Supply voltage V_{dd}	1.0 V	0.25 V	1.1 V	0.25 V
Ground voltage V_{ss}	0 V	0.25 V	0 V	0.25 V
Temperature T	37.5 °C	82.5 °C	37.5 °C	82.5 °C

Table 2
Experimental conditions for C_{Load} and $Slew_{in}$.

Parameter	Min	Max
C_{Load}	fan-out of 1	fan-out of 50
$Slew_{in}$	20 ps	500 ps
$\frac{Slew_{out}}{Slew_{in}}$	$\frac{1}{3}$	3

Note: here $Slew_{out}$ is under the nominal condition.

Given this parameter set, there are some combinations in which MOSFETs work in the sub-threshold region. Here, such combinations of V_{dd}, V_{ss}, and V_{th} are eliminated, because the gate delay becomes unacceptably large in these combinations. The sub-threshold current changes exponentially as a function of them, and hence it is difficult to appropriately express the sub-threshold current by the regression to polynomials given in Section 3.4.2.

To evaluate accuracy under practical conditions often found in actual designs, C_{Load} and $Slew_{in}$ used in the simulation were randomly generated so that the conditions listed in Table 2 were satisfied. Note that in the 90-(45-)nm technology used for the experiments, 1.5 fF(0.5 fF) corresponds to the input capacitance of a 1 × INV, and 20 ps is the output slew when a step input is given to a 1 × INV with a fan-out of one. Here, the input waveform is expressed by a linear segment followed by an exponential tail [35]. Moreover, *Slew*s are defined as five times 40–60% swing of the waveform.

The number of the iterative calculations shown in Fig. 7 is set to two. Third-order polynomials are used in the current estimation model shown in Section 3.4.2. The stepwise regression procedure is adopted to derive them [36,37] using a numerical computing software, MATLAB [38]. The adjusted coefficient of determination, adjusted R^2, is larger than 0.99, and this is sufficient to accurately estimate the current. The characterization cost will be shown in Section 4.3.

In the following evaluations in this section, nominal delay calculation used in Steps 0 and 5 in Fig. 7 is performed by HSPICE [39], because this section aims to evaluate the error caused by the proposed model itself and isolates unnecessary errors from other error sources, such as table interpolation.

4.1.2. Accuracy of estimates for 90-nm technology

Simple single-stage gate with capacitive load. This section compares the results of actual simulation with PVT variations to

Table 3
Average error [ps (%)] in estimation of delay and output slew for simple gates (90-nm process technology).

Gate	Improved model				Preliminary model in [22]			
	Delay		$Slew_{out}$		*Delay*		$Slew_{out}$	
	Rise	Fall	Rise	Fall	Rise	Fall	Rise	Fall
INV	6.5 (4.5)	4.7 (3.9)	6.6 (2.7)	4.4 (2.4)	12.7 (8.8)	8.0 (6.7)	13.2 (5.4)	5.3 (2.9)
NAND2	6.1 (4.3)	7.7 (4.5)	5.3 (2.2)	10.2 (3.7)	11.6 (8.1)	11.4 (6.7)	10.5 (4.3)	12.6 (4.6)
NOR2	10.0 (4.4)	4.6 (3.9)	17.2 (4.3)	4.2 (2.3)	16.5 (7.3)	7.5 (6.2)	24.9 (6.3)	4.7 (2.5)
Avg.	– (4.4)	– (4.1)	– (3.1)	– (2.8)	– (8.1)	– (6.5)	– (5.3)	– (3.3)
Avg. between rise and fall	– (4.3)		– (2.9)		– (7.3)		– (4.3)	

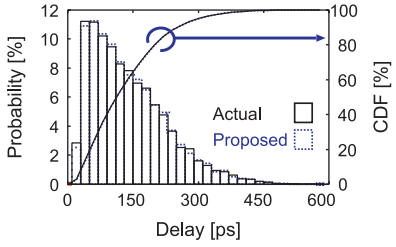

Fig. 16. Probability distribution of INV fall delay (#10,000).

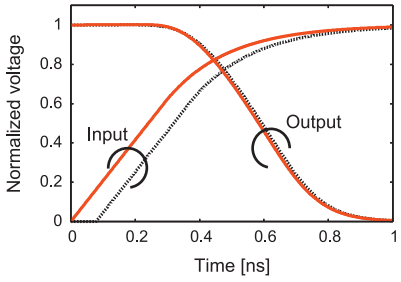

Fig. 17. Estimated transition waveform of INV. Solid lines denote actual waveforms with PVT variations and dashed lines denote waveforms estimated by the proposed model. Actual waveforms are normalized to nominal V_{dd}. Input after load translation is shifted by Δt.

Table 4
Contribution to accuracy (for fall-delay of 90-nm INV).

Load translation	Delay compensation	Average err. [ps (%)]	
		Delay	$Slew_{out}$
√	√	4.7 (3.9)	4.4 (2.4)
√		14.1 (11.9)	4.4 (2.4)
		29.1 (24.4)	25.9 (14.1)

that of the proposed calculation for INV, NAND2, and NOR2 by using a 10,000 set of generated C_{Load}, $Slew_{in}$, and variation parameters. Table 3 lists the average errors in *Delay* and $Slew_{out}$. Average errors with the proposed model are 4.3% for *Delay* and 2.9% for $Slew_{out}$. Fig. 16 shows a fall-delay histogram and CDF (cumulative distribution function) of INV. The shape of distribution obtained with the proposed model is consistent with the actual distribution. Fig. 17 illustrates an example of the transition waveform. The shape of the waveform as well as the propagation delay is accurately estimated.

Table 4 demonstrates which part, i.e., the load translation or the delay compensation part, contributes to the reduction in error. The case that neither technique is used is consistent with nominal delay calculations ignoring all variations of parameters. Load translation reduces the average error of *Delay* from 24.4% to 11.9% and that of $Slew_{out}$ from 14.1% to 2.4%. In addition, the error

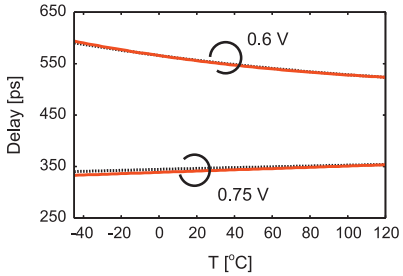

Fig. 18. Demonstration of inverted temperature dependence with 90-nm INV (rise-input with $Slew_{in}$ = 100 ps, and C_{Load} = fan-out of 40). Solid lines denote delays obtained by SPICE, and dashed lines denote delays estimated by the proposed model.

Table 5
Average error [ps (%)] in estimation of delay and output slew for complex gates (90-nm technology).

Gate	Improved model			
	Delay		$Slew_{out}$	
	Rise	Fall	Rise	Fall
AND2	7.3 (5.3)	6.4 (3.9)	4.4 (2.6)	4.6 (2.0)
AOI21	10.3 (4.7)	6.9 (4.0)	21.3 (5.5)	10.6 (3.8)

in delay estimation is decreased to 3.9% by delay compensation. It is important to simultaneously employ both techniques. A dominant error source remaining is thought to be the current waveform approximation shown in Section 3.4.1. The average error between the true average current and I_{avg} was 3.25% for INV. The error due to the current waveform approximation is very close to the gate delay error in Table 3.

At low supply voltages, a cell may suffer ITD (inverted temperature dependence) [40]. With ITD, the cell delay increases as temperature decreases. This phenomenon is caused by a competition between mobility and threshold voltage to dominate the cell delay variation. With the proposed model, as long as the output current is characterized at a range of low-voltage region, it can handle ITD effect. Fig. 18 shows an example that the proposed model reproduces ITD of an inverter at 0.6 V. The temperature dependence of the cell delay is well reproduced both at 0.75 and 0.6 V.

Complex single-stage/multi-stage gate with capacitive load. The same evaluation in the above paragraph is performed for complex gates. In this evaluation, a multi-stage gate is divided into single-stages and each stage is analyzed by using the proposed model. For example, AND2 is divided into two single-stage gates, i.e., NAND2 and INV. Table 5 demonstrates that results in complex gates exhibit the same degree of accuracy with those in simple gates.

When the proposed model is implemented in a timing analyzer, a wrapper that conceals the division of a multi-stage gate into

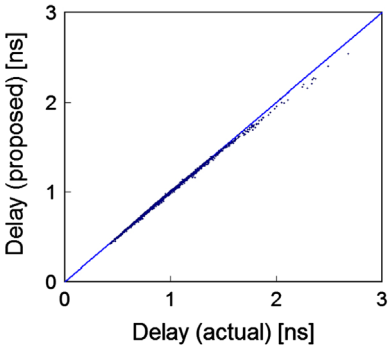

Fig. 19. Accuracy of delay estimates for 90-nm INV with RC output load (2.0-mm-long wire).

single-stage gates might be necessary. The characterization should be simply carried out for each single-stage in the following step: (1) Divide a multi-stage gate into single-stages. (2) Characterize I_{avg} and Δt for each stage. (3) Characterize *Delay* and $Slew_{out}$ for each stage.

Simple single-stage gate with RC load. This section next evaluates the accuracy when RC output load is connected with INV. A 2.0 mm-long wire was translated into a CRC π model shown in Fig. 8 by O'Brien and Savarino [31], and was used as RC output load. Fig. 19 shows the accuracy of the proposed model for a 3000 set of generated parameters and the average error is 10.6 ps (1.1%). The proposed method works well for RC output load as well as for capacitive load.

Multiple-stage path. The next evaluation is to apply the proposed model to path-delay estimation. It is assumed a 50-stage path consisting of five gates (INV, NAND2, NOR2, AND2, and AOI21) randomly. A load capacitance of each gate was randomly attached according to Table 2. $Slew_{in}$ given to a first gate is also randomly generated with Table 2. One thousand sets of 50-stage paths were generated. Note that, in this evaluation, gate-delay computation in Steps 0 and 5 in Fig. 7 was performed by not HSPICE but conventional 2D look-up tables on the assumption of a practical implementation.

Fig. 20 shows the error of the path-delay from the input of the first gate to the output of each stage. The standard deviation of the error becomes small as the number of stage increases, which means the error is not accumulated with waveform propagation. In fact, the error of each single stage is almost the same. In this evaluation, the error is slightly larger than those shown above, since the error involved by table interpolation is included.

4.1.3. Accuracy of estimates for 45-nm technology

The accuracy of delay and output slew estimates is verified using 45-nm technology library in a way similar to Section 4.1.2. Table 6 shows the results, and the estimation errors are between 2.3% and 6.7%. Summarizing the results for 90- and 45-nm technologies, the average error of the proposed model in estimation of both delay and output slew is approximately 5% on average.

4.2. Comparison with conventional methods

This section demonstrates that the proposed model has advantages over conventional ones in terms of accuracy and its applicable range of variations.

4.2.1. Comparison with sensitivity-based model

Fig. 21 compares the errors in delay estimation when ΔV_{th0} is varied. The proposed model provides accurate estimates even

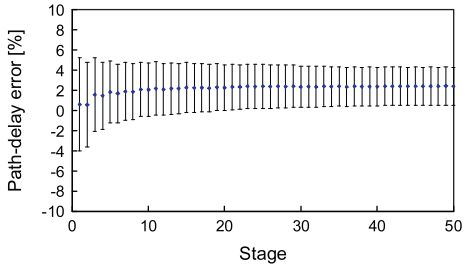

Fig. 20. Error of path-delay from primary input to each stage. Circles denote average errors and bars indicate standard deviations of errors.

Table 6
Average error [ps (%)] in estimation of delay and output slew for simple/complex gates (45-nm process technology).

Gate	*Delay*		$Slew_{out}$	
	Rise	Fall	Rise	Fall
INV	11.1 (5.4)	4.9 (4.2)	18.8 (5.3)	5.4 (2.8)
NAND2	12.1 (5.7)	8.5 (6.6)	21.4 (5.8)	7.0 (3.2)
NOR2	15.4 (6.1)	4.5 (3.8)	29.7 (6.7)	4.5 (2.3)
AND2	9.8 (5.1)	9.7 (4.6)	4.4 (2.3)	14.3 (4.1)
AOI21	13.6 (5.2)	5.6 (4.2)	26.8 (5.8)	6.7 (3.0)
Avg.	– (5.5)	– (4.7)	– (5.2)	– (3.1)
Avg. between rise and fall	– (5.1)		– (4.1)	

though ΔV_{th0} varies by ± 0.35 V. When delay estimates within 10% error are required, the V_{dd} variation of ± 0.5 V and ΔV_{th0} variation of ± 0.35 V are acceptable in the proposed model as listed in Table 7, which reveals that the proposed model is suitable for DVS and variable V_{th} design.

4.2.2. Comparison with direct delay modeling

The proposed model is compared with one that treats delay directly. A direct-delay model is constructed with second-order polynomials using the response surface method. Similarly, second-order polynomials are used as the current estimates for the proposed model. The data-sampling points for both models were the same in terms of PVT variations. The typical, minimum ($=\mu-3\sigma$), and maximum ($=\mu+3\sigma$) for each variation parameter were used for modeling. The input transition time and output load were fixed here to 100 ps and 100 fF, for simplicity.

Figs. 22 and 23 show the accuracy of estimating delay with these models. The average error in the direct-delay model is 39.9 ps (14.2%), whereas that of the proposed model is 4.5 ps (1.6%). As mentioned in Section 3.1, the current modeling is more reasonable than direct-delay modeling.

4.3. Characterization cost

The proposed model can reduce the characterization cost of the conventional method. Here, it is assumed that a 2D look-up table model which takes the output load and input slew as indices is used for nominal-delay calculation. This comparison adopts a following policy of computing delay variations using the 2D look-up table. A delay variation due to PVT variations is calculated using sensitivities. Sensitivities of process variations are assumed to be constant, and hence one table is prepared for each process variation, while those due to VT variations are interpolated with two tables of sensitivity so as to cover environmental corners.

Table 8 summarizes the characterization costs for delay of one cell, where T_{table} is the time required to generate a 2D look-up table, T_{DC} is the time required to perform one DC analysis, N is the number of variation parameters, R is that of reference points for

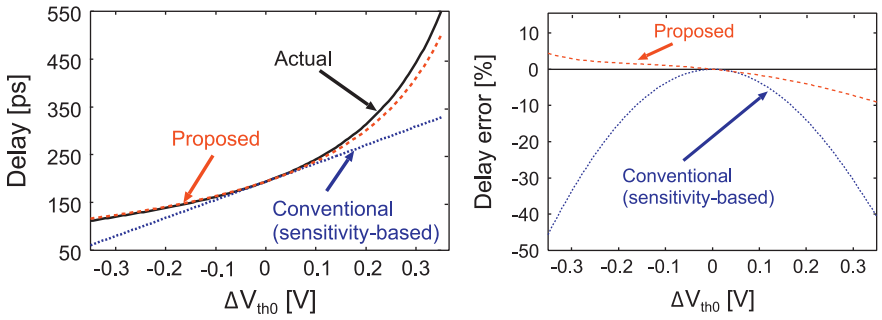

Fig. 21. Comparison in applicable range of ΔV_{th0}.

Table 7
Applicable range within delay error of 10%.

Variation parameters	Proposed	Sensitivity-based
V_{dd}	± 0.5 V	± 0.2 V
ΔV_{th0}	± 0.35 V	± 0.16 V

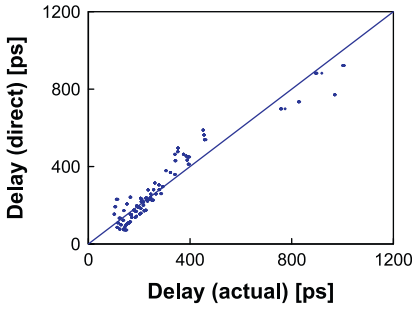

Fig. 22. Accuracy of delay estimates with direct-delay model.

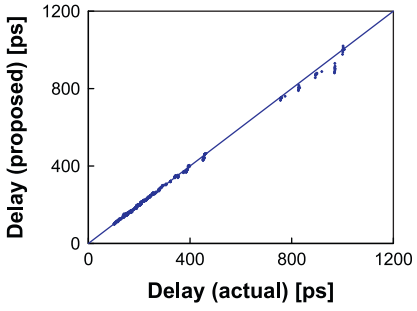

Fig. 23. Accuracy of delay estimates with proposed model.

Table 8
Delay-characterization cost for one cell.

Model	Proposed	Conventional
Look-up tables	$T_{table} \times 1$	$T_{table} \times (1+N_p+2N_eR_e)$
Current-estimation model	$T_{DC} \times (N_{V_{in50}^{nom}} R_p^{N_p} R_e^{N_e})$	–

each variation parameter, and subscripts p and e, respectively, mean the parameters for process and environmental variations. $N_{V_{in50}^{nom}}$ means the number of DC analyses for creating Eq. (10). With the above policy, the number of tables that the conventional characterization requires is $(1+N_p+2N_eR_e)$, where '1' denotes the number of a nominal-delay table and the others represent that of sensitivity tables. By contrast, the proposed model only needs one table for a nominal calculation, because it can easily calculate sensitivities and hence no sensitivity tables are required. As a result, the number of transient analyses is greatly reduced. The proposed model requires the current estimation model and the number of necessary DC analyses is $R_p^{N_p} R_e^{N_e}$, but T_{DC} is significantly smaller than T_{table}. In addition, the number of DC analyses could be reduced by adopting a smart design of experiments and/or deriving the orthogonal polynomials, such as those in Ref. [17].

Here is a numerical example. T_{table} for a 7×7 look-up table is 0.10 s and T_{DC} is below 0.0003 s with a 3.2-GHz CPU on a Linux server, not including other overheads, such as reading a file. It is assumed that $N_{V_{in50}^{nom}} = 10$ in order to deal with both of rise- and fall-delay. Using these values and assuming that $N_p = N_e = 2$, $R_p = 3$, and $R_e = 4$, the computational time to construct a cell library is less than 0.532 s for the proposed model and is 1.90 s for the conventional method. Therefore, the characterization process becomes approximately 3.6 times faster by using the proposed model as well as the estimation accuracy improves.

5. Application of proposed model to SSTA

The sensitivity-based model is widely adopted in SSTA. However, efficient methods of characterizing sensitivities are still studied. Using the sensitivity-based delay model with SSTA, when the average of a variation parameter itself varies, the actual sensitivity could be different with one calculated under the nominal condition. This causes error in estimating delay. Particular for SSTA, which copes with VT variations, such as voltage noise [41], as well as process variations, the inaccuracy of sensitivity can be a serious problem since the average supply voltage given to gates differs. For example, the average voltage of gates near V_{dd}-pads must be different from that of gates far from V_{dd}-pads. The proposed model can be used for on-demand sensitivity updates without employing any additional transitional analyses.

5.1. Sensitivity calculation in SSTA

We will first explain why on-demand sensitivity updates are necessary. Suppose that the average of a variation parameter p_i varies from p_{i_0} to $p_{i'\_0}$, as shown in Fig. 24. Here, the nominal sensitivity, $(\partial d/\partial p_i)|_{p_i = p_{i\_0}}$, is inappropriate, and the actual sensitivity, $(\partial d/\partial p_i)|_{p_i = p_{i'\_0}}$, should be used in calculating delay. To eliminate this sensitivity mismatch, on-demand sensitivity updates are necessary. However, the conventional characterization of sensitivity using a number of transient simulations with regard to output load, input slew and variation parameters, suffers from a large increase in the computational time.

Here, the details on sensitivity updates with the proposed model are explained. For simplicity, it is assumed that variation parameter is only the channel length, L. Gate-delay d is

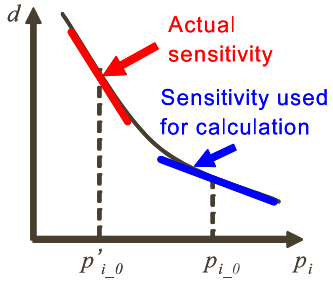

Fig. 24. Sensitivity mismatch in the case that nominal value of parameter p_i shifts from p_{i_0} to $p_{i'\_0}$.

expressed by

$$d = d_0 + k_{PDN}\Delta L_{PDN} + k_{PUN}\Delta L_{PUN} + \sum_i k_{ri}\Delta L_{ri}, \tag{29}$$

where *PDN* stands for a pull-down network, *PUN* stands for a pull-up network, $\Delta L_{PDN/PUN}$ is uniform variation on the *PDN*/*PUN* side, ΔL_{ri} represents an uncorrelated random variation of the i-th transistor, $k_{PDN/PUN}$ is the sensitivity to $\Delta L_{PDN/PUN}$, and k_{ri} is the sensitivity to ΔL_{ri}. The last term on the right-hand side in Eq. (29) is the sum of normal distributions where ΔL_{ri} is normally distributed. In this case, one can incorporate them into the one normal distribution, $k_r N(0,1)$, in Eq. (30), where k_r is calculated in Eq. (31)

$$d = d_0 + k_{PDN}\Delta L_{PDN} + k_{PUN}\Delta L_{PUN} + k_r N(0,1^2), \tag{30}$$

$$k_r = \sqrt{\sum_i \sigma^2(k_{ri}\Delta L_{ri})}. \tag{31}$$

Each sensitivity is expressed as a partial differential of d with regard to each parameter at their nominal value

$$k_{PDN} = \left.\frac{\partial d}{\partial L_{PDN}}\right|_{\Delta L_{PUN}=0, {}^\forall j \Delta L_{rj}=0}, \tag{32}$$

$$k_{PUN} = \left.\frac{\partial d}{\partial L_{PUN}}\right|_{\Delta L_{PDN}=0, {}^\forall j \Delta L_{rj}=0}, \tag{33}$$

$$k_{ri} = \left.\frac{\partial d}{\partial L_{ri}}\right|_{\Delta L_{PDN}=0, \Delta L_{PUN}=0, {}^\forall j (j\neq i)\Delta L_{rj}=0}. \tag{34}$$

The number of gate-delay calculations required to recalculate the set of sensitivities in Eq. (30) is $mn+2m+1$, where m is the number of variation parameters and n is the number of transistors in the gate. mn corresponds to the number of k_{ri}, $2m$ to k_{PDN} and k_{PUN} of each p, and one to the nominal condition. When k_{PDN} and k_{PUN} are approximated as the sum of k_{ri} with $i \in PDN$ and $i \in PDN$, the above number could be reduced to $mn+1$. Thanks to the proposed gate-delay model, sensitivities can be updated on-the-fly without performing transient simulations.

5.2. Experiments

5.2.1. Accuracy enhancement with updated sensitivity

This section tests on-demand sensitivity updates in order to verify how effective they were. Here, a 90-nm industrial standard cell library is used for evaluation. As an example, this evaluation examines the path delay of a three-stage INV in the circuit diagram of Fig. 25. Two results of estimating the delay with updating and without updating sensitivity are compared. The average supply voltages of each inverter, $\mu(V_{dd})$s, are indicated in the figure, and all the $\mu(V_{ss})$s are equal to 0 V. It is assumed that $3\sigma(V_{dd})$s and $3\sigma(V_{ss})$s were 0.05 V. Moreover, every transistor has ΔV_{th0} as a process variation parameter and $3\sigma(\Delta V_{th0})$ is also

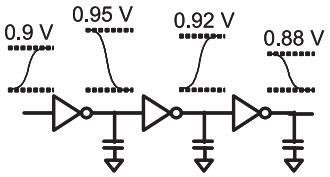

Fig. 25. Three-stage INV used for evaluation.

0.05 V. The wire capacitances between inverters are 10 fF and the transition time of input given to the first inverter is 100 ps.

Delays of each inverter were computed similarly to that Eqs. (30) and (31). Additionally, the variations of input slews, i.e., the term of $k_{Slew_{in}}\Delta Slew_{in}$, were took into account, because the output slew in the previous stage corresponds to the input slew at the next gates, and hence the variation of input slew should be considered [7]. The slews of each inverter were computed similarly to that used in calculating the delay.

Fig. 26 shows the estimated delay distribution. The distribution with sensitivity updating is almost identical to the actual results summarized in Table 9. The error in estimating the standard deviation is reduced from −22.7% to 1.6% thanks to the on-demand sensitivity updates.

5.2.2. Computation time of updating sensitivity

As mentioned above, the usage of updated sensitivities improves the accuracy of SSTA. This section shows its additional calculation cost in SSTA. The same policy described in Section 4.3 is adopted and each table is created with a third-order polynomial. Next, the proposed model is implemented to calculate sensitivities for a SSTA reported in [41]. The specification of implemented proposed model is same with that in Section 4.1.1. A run-time of the SSTA was evaluated using an industrial embedded processor, MeP (media embedded processor) [42]. The processor was synthesized and layed out by commercial tools [43,44] using a 90-nm standard cell library. The layout size was 2×2 mm^2, and the processor was composed of 83,552 cells and 10 SRAM (static random access memory) macros. A C program of image processing was implemented and compiled using GCC (GNU compiler collection) for MeP. In this image processing, MeP reads 256×256-pixel data from SRAM, executes arithmetic operation, and stores the result to SRAM. Using the executable binary, a gate-level simulation was performed and noised-voltage data for 8920 cycles were obtained. Finally, SSTA was executed by the same server with that used in Section 4.3.

The computation times of SSTA are shown in Table 10. The additional cost of updating sensitivity is 5.96 s (9.89%). In this evaluation, the proposed model recalculated all sensitivities, but actually some of them are expected to be close to the nominal sensitivities in regions where voltage drops are small. Moreover, on-the-fly sensitivity analysis is not always necessary and it could be replaced with off-line sensitivity analysis. For example, if sensitivities at a dropped supply voltage were calculated for every cell beforehand in the process of cell library preparation, we do not need to calculate them on-the-fly and we can conceal the computational overhead. Thus, by using the proposed model, designers can receive the benefits of updating sensitivity with acceptable additional cost.

6. Conclusion

This paper proposed a gate-delay model coping with wide-ranging PVT variations. Additional characterization needed for the proposed model is only output-current modeling using a limited number of DC analyses. The proposed model translates the

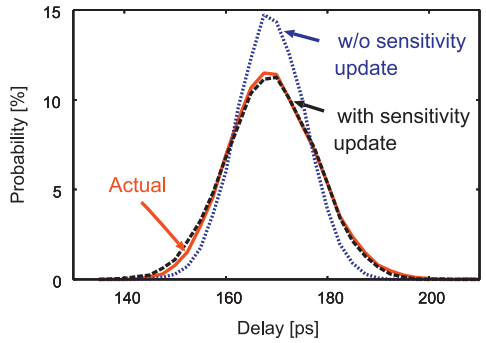

Fig. 26. Comparison of delay distributions (#25,000). Solid line denotes actual distribution. Dashed and dotted lines are with and without on-the-fly sensitivity updating.

Table 9
Comparison of statistics.

Statistics	Actual (ps)	With sensitivity updating [ps (% err.)]	W/o sensitivity updating [ps (% err.)]
μ	168.2	167.7 (−0.32)	167.7 (−0.32)
σ	8.7	8.8 (1.60)	6.7 (−22.7)
$\mu+3\sigma$	194.2	194.1 (−0.07)	187.8 (−3.31)

Table 10
Computation time of SSTA.

Proposed (s)	Conventional (s)	Difference [s (%)]
66.2	60.2	5.96 (9.89)

output-current variation into output load, and hence the delay can be calculated with any conventional gate-delay model using the translated load and reshaped input waveform. The average errors in estimating the delay and output slew are 5%. The proposed model can be used for implementing MC-STA and SSTA. An application of the proposed model to updating sensitivity in SSTA was also described. It is demonstrated that the error in estimating the standard deviation could be considerably reduced using sensitivities that are updated on-the-fly by the proposed model with acceptable computational overhead.

Acknowledgments

This work was supported in part by NEDO and STARC, as well as by VLSI Design and Education Center (VDEC), the University of Tokyo in collaboration with Synopsys, Inc.

References

[1] H. Masuda, S. Ohkawa, A. Kurokawa, M. Aoki, Challenge: variability characterization and modeling for 65- to 90-nm processes, in: Proceedings of the CICC, 2005, pp. 593–599.
[2] A. Mezhiba, E. Friedman, Scaling trends of on-chip power distribution noise, IEEE Transactions on Very Large Scale Integration (VLSI) Systems 12 (4) (2004) 386–394.
[3] M. Badaroglu, P. Wambacq, G. Van der Plas, S. Donnay, G. Gielen, H. De Man, Evolution of substrate noise generation mechanisms with CMOS technology scaling, IEEE Transactions on Circuits and Systems I: Regular Papers 53 (2) (2006) 296–305.
[4] N. Ahmed, M. Tehranipoor, V. Jayaram, Transition delay fault test pattern generation considering supply voltage noise in a SOC design, in: Proceedings of the Design Automation Conference, ACM, 2007, pp. 533–538.
[5] K. Banerjee, A. Mehrotra, Global (interconnect) warming, IEEE Circuits and Devices Magazine 17 (5) (2001) 16–32.
[6] C. Visweswariah, K. Ravindran, K. Kalafala, S. Walker, S. Narayan, First-order incremental block-based statistical timing analysis, in: Proceedings of the DAC, ACM, 2004, pp. 331–336.
[7] H. Chang, S. Sapatnekar, Statistical timing analysis under spatial correlations, IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems 24 (9) (2005) 1467–1482.
[8] J. Le, L. Pileggi, STAC: statistical timing analysis with correlation, in: Proceedings of the DAC, ACM New York, NY, USA, 2004, pp. 343–348.
[9] W.-S. Wang, V. Kreinovich, M. Orshansky, Statistical timing based on incomplete probabilistic descriptions of parameter uncertainty, in: Proceedings of the ACM/IEEE Design Automation Conference (DAC), ACM, New York, NY, USA, 2006, pp. 161–166. http://dx.doi.acm.org/10.1145/1146909.1146954.
[10] S. Bhardwaj, S. Vrudhula, A. Goel, A unified approach for full chip statistical timing and leakage analysis of nanoscale circuits considering intradie process variations, IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems 27(10) (2008) 1812–1825.
[11] International Technology Roadmap for Semiconductors, ITRS Reports 2009 Edition, 〈http://www.itrs.net/〉.
[12] S. Pant, D. Blaauw, Static timing analysis considering power supply variations, in: Proceedings of the ICCAD, IEEE Computer Society, 2005, pp. 365–371.
[13] M. Hashimoto, J. Yamaguchi, H. Onodera, Timing analysis considering spatial power/ground level variation, in: Proceedings of the ICCAD, 2004, pp. 814–820.
[14] K. Okada, K. Yamaoka, H. Onodera, A statistical gate delay model for intra-chip and inter-chip variabilities, in: Proceedings of the ASP-DAC, ACM, 2003, pp. 31–36.
[15] H. Fatemi, S. Nazarian, M. Pedram, Statistical logic cell delay analysis using a current-based model, in: 2006 43rd ACM/IEEE Design Automation Conference, 2006, pp. 253–256.
[16] M. Schmidt, H. Kinzelbach, U. Schlichtmann, Variational waveform propagation for accurate statistical timing analysis, in: TAU Workshop, 2008, pp. 20–25.
[17] A. Goel, S. Vrudhula, Statistical waveform and current source based standard cell models for accurate timing analysis, in: Proceedings of the 45th Annual Design Automation Conference, ACM New York, NY, USA, 2008, pp. 227–230.
[18] A. Mitev, D. Ganesan, D. Shanmugasundaram, Y. Cao, J. Wang, A robust finite-point based gate model considering process variations, in: Proceedings of the ICCAD, IEEE Press, 2007, pp. 692–697.
[19] V. Zolotov, J. Xiong, S. Abbaspour, D. Hathaway, C. Visweswariah, Compact modeling of variational waveforms, in: Proceedings of the ICCAD, IEEE Press, 2007, pp. 705–712.
[20] V. Veetil, K. Chopra, D. Blaauw, D. Sylvester, Fast statistical static timing analysis using smart Monte Carlo techniques, IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems 30 (6) (2011) 852–865.
[21] J. Jaffari, M. Anis, On efficient Monte Carlo-based statistical static timing analysis of digital circuits, in: Proceedings of the ICCAD, 2008, pp. 196–203.
[22] K. Shinkai, M. Hashimoto, A. Kurokawa, T. Onoye, A gate delay model focusing on current fluctuation over wide-range of process and environmental variability, in: Proceedings of the ICCAD, ACM, 2006, pp. 47–53.
[23] S. Sapatnekar, Timing, Kluwer Academic Publishers, 2004.
[24] S. Onaissi, F. Najm, A linear-time approach for static timing analysis covering all process corners, IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems 27 (7) (2008) 1291–1304.
[25] S. Kumar, C. Kashyap, S. Sapatnekar, A framework for block-based timing sensitivity analysis, in: Proceedings of the DAC, ACM, 2008, pp. 688–693.
[26] T. Sakurai, A. Newton, Alpha-power law MOSFET model and its applications to CMOS inverter delay and other formulas, IEEE Journal of Solid-State Circuits 25 (2) (1990) 584–594.
[27] J. Rabaey, A. Chandrakasan, B. Nikolić, Digital Integrated Circuits: A Design Perspective, Prentice Hall, New Jersey, 1996.
[28] M. Hashimoto, Y. Yamada, K. Onodera, Equivalent waveform propagation for static timing analysis, IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems 23 (4) (2004) 498–508.
[29] S. Abbaspour, M. Pedram, Gate delay calculation considering the crosstalk capacitances, in: Proceedings of the ASPDAC, 2004, pp. 853–858.
[30] V. Zolotov, D. Blaauw, R. Panda, C. Oh, Cross-coupled noise propagation in VLSI designs, International Journal of Analog Integrated Circuits and Signal Processing, vol. 35(2–3), I-2, Kluwer Press, 2003, pp. 133–142.
[31] P.R. O'Brien, T.L. Savarino, Modeling the driving-point characteristic of resistive interconnect for accurate delay estimation, in: Proceedings of the ICCAD, Santa Clara, CA, 1989, pp. 512–515. http:dx.doi.org/10.1109/ICCAD.1989.77002.
[32] R. Myers, D. Montgomery, C. Anderson-Cook, Response Surface Methodology: Process and Product Optimization Using Designed Experiments, Wiley, 2009.
[33] T. Sakurai, A. Newton, A simple MOSFET model for circuit analysis, IEEE Transactions on Electron Devices 38 (4) (1991) 887–894.
[34] NANGATE, Inc., Nangate 45 nm Open Cell Library Datasheet, 2008. URL 〈http://www.nangate.com/〉.
[35] F. Chang, C. Chen, P. Subramaniam, An accurate and efficient gate level delay calculator for MOS circuits, in: Proceedings of the DAC, 1988, pp. 282–287.
[36] M. Efroymson, Multiple regression analysis, Mathematical Methods for Digital Computers 1 (1960) 191–203.

[37] N.R. Draper, H. Smith, Applied Regression Analysis, 3rd Ed., Wiley-Interscience, 1998. URL 〈http://amazon.com/o/ASIN/0471170828/〉.
[38] The MathWorks, Inc., MATLAB.
[39] Synopsys, Inc., HSPICE Documentation.
[40] A. Dasdan, I. Hom, Handling inverted temperature dependence in static timing analysis, ACM Transactions on Design Automation of Electronic Systems (TODAES) 11 (2) (2006) 306–324.
[41] T. Enami, S. Ninomiya, M. Hashimoto, Statistical timing analysis considering spatially and temporally correlated dynamic power supply noise, IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems 28 (4) (2009) 541–553.
[42] Toshiba, Inc., MeP (Media embedded Processor). URL 〈http://www.MePcore.com/〉.
[43] Synopsys, Inc., Design Compiler User Guide, 2007.
[44] Synopsys, Inc., Astro User Guide, 2007.

Ken-ichi Shinkai received the B.E., M.E. and Ph.D. degrees in Information Systems Engineering from Osaka University, Osaka, Japan, in 2006, 2008, and 2011, respectively. He is currently with NEC Corporation. His major interest is variation-aware timing and reliability analysis. He is a member of IEEE.

Masanori Hashimoto received the B.E., M.E. and Ph.D. degrees in Communications and Computer Engineering from Kyoto University, Kyoto, Japan, in 1997, 1999, and 2001, respectively. Since 2004, he has been an Associate Professor in Department of Information Systems Engineering, Graduate School of Information Science and Technology, Osaka University. His research interest includes computer-aided-design for digital integrated circuits and high-speed circuit design. Dr. Hashimoto served on the technical program committees for international conferences including DAC, ICCAD, ASP-DAC, DATE, ISPD, ICCD, and ISQED. He is a member of IEEE, ACM, IEICE, and IPSJ.

Takao Onoye received the B.E. and M.E. degrees in Electronic Engineering, and Dr.Eng. degree in Information Systems Engineering all from Osaka University, Japan, in 1991, 1993, and 1997, respectively. He is currently a Professor in the Department of Information Systems Engineering, Osaka University. His research interests include media-centric low-power architecture and its SoC implementation. He is a member of IEEE, IEICE, IPSJ, and ITE-J.