Gate Delay Modeling for Multiple To-non-controlling Stimuli

Liang Chi Chen
Professor Melvin Breuer

Technical Report Number: CENG 02-16

Department of Electrical Engineering – Systems
University of Southern California
Los Angeles, California 90089-2562
213-740-4465

Gate Delay Modeling for Multiple To-non-controlling Stimuli

Liang-Chi Chen, Sandeep K. Gupta, Melvin A. Breuer
Department of EE - Systems, University of Southern California, Los Angeles, CA 90089-2562
+1 (213) 740-4460, +1 (213) 740-2251, +1 (213) 740-4469
{lichen, sandeep, mb}@poisson.usc.edu

ABSTRACT

In sub-100nm technology, gate delay is affected by many delay phenomena. Our objective is to develop an accurate empirical gate delay model that captures these delay phenomena. This model is more accurate than analytical models since close form analytical gate delay models cannot achieve high accuracy due to modeling complexity. This model is also more useful than table lookup methods, since it make it possible to identify at low complexity the worst case combinations of arrival and required times for timing analysis. We identify several phenomena which impact gate delay including two new ones, quantify their importances, identify input waveforms that capture these phenomena, and develop a delay model for accurately predicting to-non-controlling response. For simplicity, we show the version of the model that captures the effect of a pair of input transitions and demonstrate that it significantly improves accuracy over previous models.

1. Introduction

Static timing analysis (STA) [1] is widely used for validating circuit performance. It computes min-max timing ranges (also called timing windows) for rising and falling transitions at each line in a circuit without explicitly considering any vectors. The min-max delay at the primary outputs of a combinational block, hold/setup times of latches/flip-flops, and the clock period are used to determine if a circuit has hold time/setup time violations.

The accuracy of STA depends heavily on the delay model used for each gate. Although SPICE-like models [2][3][4] provide good timing accuracy, they can not be used in STA because they require fully-specified input waveforms. Due to the high complexity of using these models, it is impractical to obtain the timing ranges for STA by using these models and simulating all possible fully specified vectors.

Pin-to-pin delay models [5] are hence used for STA. One main deficiency of pin-to-pin delay models is that near-simultaneous switching delay [6][7] is not captured. *Near-simultaneous switching* means that the transitions at a few inputs of a gate occur at the same time or with skews close to 0. It is simplified as simultaneous switching in this paper. *Simultaneous to-controlling transitions* at inputs of a primitive gate decrease gate delay due to activation of multiple charge/discharge paths [8]. *Simultaneous to-non-controlling transitions* at inputs of a primitive gate increase gate delay due to short circuit current, Miller effect, and body effect.

A delay model capturing the simultaneous to-controlling switching and also applicable to STA has been developed in [8]. In this paper, we explore delay phenomena that impact gate delay due to simultaneous to-non-controlling transitions and develop a delay model for this situation.

2. Previous Delay Models

Simulators have been developed for digital circuits with different accuracy/computation cost trade-offs. *Timing simulators* [2][3] generate voltage waveforms more efficiently (lower computation costs) than SPICE-like *circuit simulators* [4], but are less accurate. *Delay calculators* are very efficient in determining circuit delay and therefore become the dominating tools for timing estimation and timing analysis.

Several approaches for modeling inverters exists, including *resistance-capacitance (RC) based systems* [9], *analytical delay function systems* [10][11][12], and *empirical delay based systems* that use *lookup tables* [13][14][15].

More complicated gates such as NAND/NOR structures are difficult to model because of their multinodal circuitry and multiple inputs. Two types of models for NAND/NOR gates are based on analytical inverter models. The first approach generalizes the inverter model to NAND/NOR gates. Since the inverter model can not be directly extended to consider simultaneous switching, these generalizations are based on simple RC delay models [13][15][16], or capturing only pin-to-pin delays [17][18].

In a second approach, a NAND/NOR gate is collapsed into an equivalent inverter. Many approaches that handle simultaneous switching [6][19][20][21] are based on this inverter-collapsing method. These approaches focus on finding the equivalent transistor widths and an equivalent input waveform. Early approaches replaced n parallel and n serial transistors of the same size by single transistors with widths n and $1/n$ times of the original, respectively. Multiple transistors in series are modeled as an inverter with a delay degradation factor [21]. In [6][19][20], the authors provided enhanced models for collapsing serial transistors into a single conducting transistor of appropriate width and an equivalent input waveform.

Two empirical approaches for characterizing NAND/NOR gates have been developed [8][22]. To make empirical char-

acterization less complex, Chandramouli et al. [22] use dimensional analysis to reduce the number of variables that impact gate delay. Chen et al. [8] develop formulae for capturing simultaneous to-controlling transitions, but not simultaneous to-non-controlling transitions.

Among all the above approaches for NAND/NOR gates, only [8] can be used in STA for large circuits. In the other approaches it is too difficult to identify the combinations of transition and arrival times at gate inputs that lead to extreme values of timing ranges at gate outputs, unless all possible pairs of vectors are simulated.

In this paper we develop a new empirical model that captures simultaneous to-non-controlling switching effects and applicable to STA. Together with our work reported earlier [8], this model can significantly improve the accuracy of STA.

3. Notation

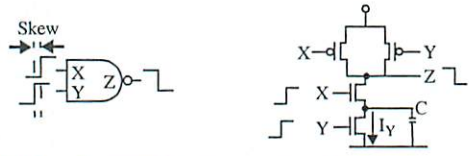

FIGURE 1. An example of two-input NAND gate.

A NAND gate, with output Z and two inputs X and Y (Figure 1), is used as an example to illustrate the definitions. Here Z represents the gate output and also the gate. **C** is the parasitic internal capacitance. $\mathbf{N_X}$ and $\mathbf{P_X}$ represent the N transistor and the P transistor driven by input X. The **controlling value** of a multi-input gate Z, CV^Z, is the value when applied to any of the gate's inputs, completely determines the value at its output, independent of the values at its other inputs. The **non-controlling value** of a gate Z, $\overline{CV}^Z$, is the complement of its controlling value. The **to-non-controlling transition** at an input of gate Z is denoted as a sequence of values $<CV^Z, \overline{CV}^Z>$. If to-non-controlling transitions occur at one or more inputs of a gate, and the gate's non-controlling value is applied to its remaining inputs, then the transition at the gate output is called a **to-non-controlling response**.

The **transition time** (T^X_{tr}) of a transition *tr*, where $tr \in \{R, F\}$, at line X is the time required for a rising transition (**R**) to go from $0.1V_{dd}$ to $0.9V_{dd}$, or from $0.9V_{dd}$ to $0.1V_{dd}$ for a falling transition (**F**). The **arrival time** (A^X_{tr}) of a transition *tr* at line X is the time when the voltage at the line X reaches $0.5\ V_{dd}$. The **skew** ($\delta^{X,Y}$) between transitions at lines X and Y is $A^Y_{tr} - A^X_{tr}$. $\delta^{X,Y} > 0$ means the transition at Y arrives later than that at X ($A^Y_{tr} > A^X_{tr}$). The **to-non-controlling gate delay** d^Z_{tr}, defined as $A^Z_{tr} - \max(A^X_{\overline{tr}}, A^Y_{\overline{tr}})$, is the gate delay of Z, where the output transition $tr \in \{R, F\}$ is a to-non-controlling response and input transitions $\overline{tr}$ are to-non-controlling transitions where $R = \overline{F}$ and $F = \overline{R}$. The **pin-to-pin delay** from X to Z is the difference between arrival times at Z and X when Y is steady at the non-controlling value and a transition is applied at X.

4. Delay Phenomena

During to-non-controlling response, the output is discharged/charged through the serial transistors. We first identify important delay phenomena and then study the magnitude of their effect for characterization in our empirical model. There are at least five phenomena that influence gate delay when simultaneous to-non-controlling transitions occur, namely short circuit current, initial states of internal capacitances, Miller effect, body effect, and charge sharing. In addition, we have identified two new phenomena, called stopping early discharge and impedance matching.

4.1 Short Circuit Current

Effects of the short circuit current on an inverter were reported in [12][18]. Below we reveal how short circuit current impacts simultaneous switching delay in a manner not reported before.

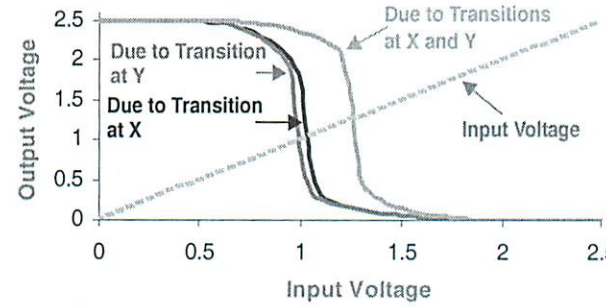

FIGURE 2. DC characteristics of a two-input NAND gate.

In Figure 2 we show the DC characteristics for the two-input NAND gate shown in Figure 1, where we vary (1) only the input voltage at X (Y is kept at logic 1), (2) only the input voltage at Y (X is kept at logic 1), and (3) the voltages at X and Y together.

The DC characteristics capture the voltage at gate output when the parasitics and load achieve steady state. Thus, when applying a rising transition at gate inputs, we observe that in case (3), the output reaches $0.5V_{dd}$ for higher input voltages (i.e., later in time, in dynamic characteristics) than that for cases (1) and (2) (Vin=1.25 in case (3) vs. Vin=1.05 in case (1)). The main reason is that to pull down the output voltage, the pull-down transistor(s) discharging the output must also sink the current from the pull-up transistors that are switching off. In case (3), the transistor P_X as well as P_Y contribute to short circuit current, while in the two other cases, only one of these two transistors does.

From static characteristics in Figure 2, we observe that the voltage at the gate output starts to decrease when the transistor(s) that charge the gate output are weaker than the transistors that discharge. During the period when both P and N transistors are on, there is a short circuit current. In case (3), the peak short circuit current will be larger than that in cases (1) and (2). Short circuit current is of concern in low power designs. Here we use the same phenomena to explain the increase in delay during simultaneous switching.

4.2 Initial States of Internal Capacitances

The impact of the initial state of internal capacitances on delay was illustrated in [7]. When we apply a rising step waveform to both inputs of the gate in Figure 1, different initial states of the internal capacitances cause gate delay to vary by 7% (Table 1). The reason is that to discharge the output, internal capacitances between the simultaneous transitions also need to be discharged.

TABLE 1. Delay of multiple-to-non-controlling transitions without/with internal capacitance pre-charged.

Internal capacitance C	X	Y	Z	Delay	Ratio
If pre-discharged	0→1	0→1	1→0	84.3	1.00
If pre-charged	0→1	0→1	1→0	90.3	1.07

4.3 Stopping Early Discharge

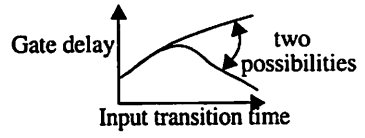

FIGURE 3. The general shapes of pin-to-pin gate delay as a function of input transition time.

It has been shown in [8] that pin-to-pin gate delay, as a function of input transition time (see Figure 3), may be either (1) monotonically increasing or (2) bi-tonic (monotonically increasing and then monotonically decreasing). In case (2), this bi-tonicity is due to the fact that a transition at an input starts to pull up (down) the output voltage before the modeled arrival time of the input transition, i.e., the time when the corresponding voltage reaches $0.5V_{dd}$. The effective $\beta n/\beta p$ ratio determines which shape this delay function takes. Case (1) can be treated as a special case of (2) where the curve's peak occurs when the input transition time is very large. Case (2) occurs for the output falling delay of the two-input NAND gate shown in Figure 1.

In Figure 4 we show the skew vs. delay relation for a fixed value of T_Y (400ps) and two different T_X (280 and 600ps), and where the internal capacitance is pre-discharged. (These two T_X values are selected because this cause the same pin-to-pin delay from X to Z, as shown in Figure 5.) The two curves in Figure 4 show identical delays when the skew is large (i.e. have the same pin-to-pin delay), but the delay difference is as high as 38.9% (skew = 60ps) when the skew is small, and 22.8% when the skew is 0.

The main reason for this difference in delay is due to a new phenomena that we call *stopping of early discharge*. In the case where only input X has a transition and $T_X = 600$, the output starts discharging before input X reaches $0.5V_{dd}$. Interestingly, the output starts to discharge earlier for case $T_X = 600$ compared to case with $T_X = 280$. (By our definition, for T_X values of 600 and 280, the transition at input X starts at 300 and 140 before A_X, respectively.) But when both inputs have transitions, the output will not be discharged until both N_X and N_Y are on. This phenomena delays the time when discharging starts, and hence causes more delay increase for the $T_X = 600$ case since it relies more on early discharge than the $T_X = 280$ case.

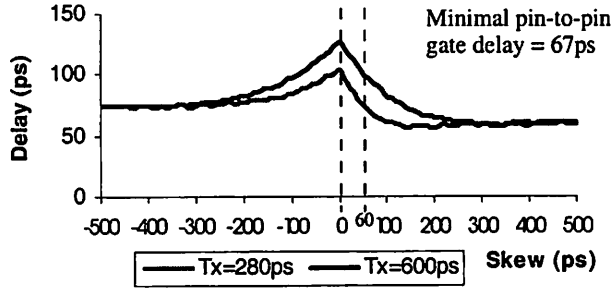

FIGURE 4. Two simultaneous switching with a fixed T_Y.

In general, stopping early discharge occurs when both of the following two conditions hold: (1) pin-to-pin delay is bi-tonic with respect to input transition time; (2) simultaneous to-non-controlling transitions occur at the inputs of a gate.

4.4 Impedance Matching

The pin-to-pin delay for output falling of the two-input NAND gate in Figure 1 is shown in Figure 5. Because of the initial state of the internal capacitance, when the input transition time is small, the pin-to-pin delay from X is smaller than that from Y. If the input transition time is larger, we have the opposite result. For an input transition time of 700ps, in Figure 6 we show the voltages at the gate input, the gate output, and the internal node C for the two cases where the input transition is at X and Y, respectively. The most significant difference between these two plots is the initial voltage at node C. In Figure 6 (a) where the input transition is at X, V_C is initially 0 and remains close to 0 and therefore limits the current flowing through N_Y. In Figure 6 (b), V_C is initially $V_{dd} - V_{th} \approx 2.1$Volt. The average value of V_C is higher than before, hence N_Y sinks more current. More precisely, during the switching of X, N_Y is always in the linear region, where I_Y, the current flows though N_Y is

proportional to $V_Y*V_C-V_C^2$. In Figure 6 (a) as V_C approaches 0, so does I_Y. In contrast, in Figure 6 (b), V_C as well as I_Y are large. Thus the delay in Figure 6 (a) is longer, even if less charge needs to be discharged.

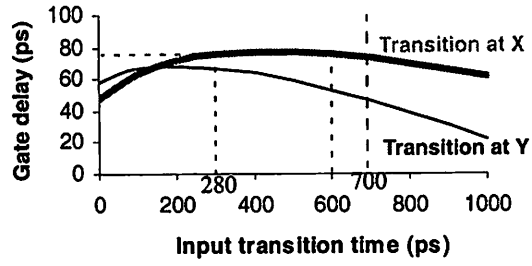

FIGURE 5. Pin-to-pin delay of a two-input NAND.

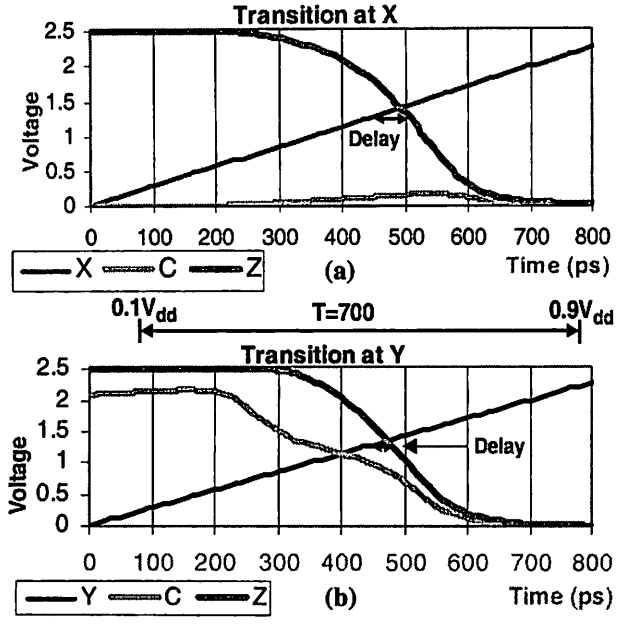

FIGURE 6. Single input transition response of a two-input NAND gate for the transition (a) at X and (b) at Y, where input transition time is 700ps.

For the case shown in Figure 6 (b), after N_Y starts to conduct, V_C tends to stay at a value that is close to balance the current flow through N_X and N_Y, i.e. the impedances of N_X and N_Y are somewhat matched, such that more current flows through the series chain, and therefore decreases the gate delay.

Hence, depending on the input transition time, a pre-charged internal capacitance may lead to higher or lower delay due to impedance matching.

4.5 Miller Effect

Parasitic capacitances exist between different terminals of a transistor [19]. Parasitic capacitances between gate and drain, and between gate and source are shown for N_X and N_Y in Figure 7. When gate inputs switch to 1, charge is transferred from the corresponding gate inputs to the gate output and the internal capacitance. Such charge transfer slows down the output transition, since the additional charge must also be discharged.

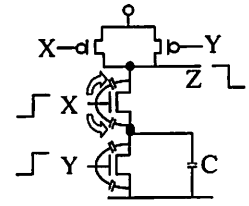

FIGURE 7. Charge transferred through parasitic capacitances due to Miller effect.

4.6 Body Effect

Body effect increases the gate delay when the voltage between source and bulk of a transistor (V_{SB}) is greater than 0 [23]. In Figure 7, inputs X and Y have simultaneous rising transitions. If the internal capacitance is pre-charged to 1, then V_C, the voltage of the source of transistor N_X, is greater than 0. The threshold voltage of N_X increases when $V_{SB} > 0$. This phenomenon adds to the delay of the gate with respect to the time when the transition at X begins.

4.7 Charge Sharing

Consider the four-input NAND gate shown in Figure 8 where all internal capacitance C_0, C_1, and C_2, are of equal size. When the gate output stays at 1 (controlled response), the internal capacitances may be charged to some intermediate voltages between 0 and $V_{dd}-V_{th}$ (represented simply as 1 in Figure 8), depending on the sequence of values at the inputs.

However, the following rules govern the voltages of internal nodes: (1) The capacitances above the topmost off transistor are pre-charged. The capacitances below the lowest off transistor are pre-discharged. (2) All internal capacitances can only have a voltage between 0 and $V_{dd}-V_{th}$. (3) An internal capacitance will have a voltage that is higher or equal to that of a capacitance at a lower position.

Hence right before all inputs becomes 1 (and therefore the output switches to 0), the states of internal capacitances may have many possible values, and therefore impact the gate delay.

4.8 An Example

An example showing the impact of these effects on the simultaneous to-non-controlling transitions for gate delay of the two-input NAND gate in Figure 1 is illustrated in Figure 9. Here input skew and the initial state of the internal capacitance are varied, while fixed and identical transition times used at both inputs.

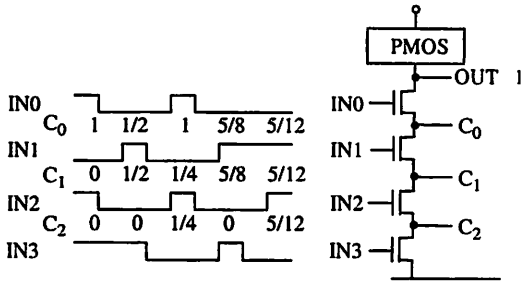

FIGURE 8. Charge sharing at a four-input NAND gate.

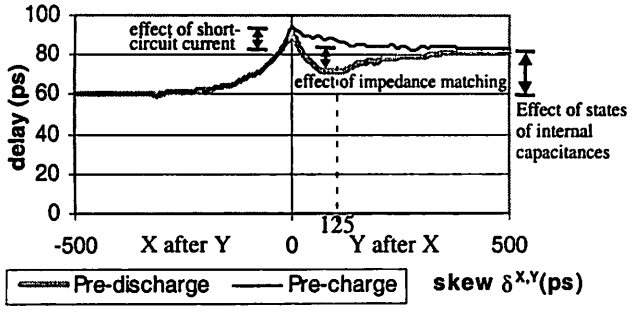

FIGURE 9. Skew-delay relation for to-non-controlling response at a two-input NAND gate where both inputs have equal transition time.

When the absolute value of skew is large, one transition occurs much later than the other. Then the delay is the same as the pin-to-pin delay, since the early transition does not affect the gate delay. So the delay for skew = -500 and 500ps are identical to pin-to-pin delays from X and Y, respectively. The two curves merge when skew = 500ps, since there is sufficient time to charge the internal capacitance to V_{dd}-V_{th} even if it was pre-discharged. In other words, the initial state of the internal capacitance does not matter. The difference of the pin-to-pin delays to Z from X and Y is caused mainly by the states of the internal capacitance (discharged for pin-to-pin delay from X and charged for that from Y). It is also slightly impacted by Miller effect and body effect.

When the skew is 0, the delay is greater than each of the two pin-to-pin delays, primary due to the higher short circuit current. In Figure 9, the minimum delay for positive skew occurs when the internal capacitance is pre-discharged and the skew is 125ps. The internal capacitance is partially charged after N_X turns on. Hence, when N_Y turns on, the two transistors have better matched impedances and the delay is reduced. Just to the left of the minimum delay point, the delay increases due to higher short circuit current. Just to the right of this point, the delay increases due to increased internal capacitance charge. For the pre-charged case, the internal capacitance is already completely charged and so cannot accept any more charge. So the curve is smooth and no impedance matching effect occurs.

5. Impacts of Delay Phenomena

We explore the delay effect of each phenomenon based on a minimum-size two-input NAND gate with a minimum load. For each experiment, we eliminate the impact of one phenomenon by modifying the circuit, when necessary. We then apply various input transition times (slews) and skews over realistic ranges to cover all possible cases, and observe the delay difference caused by each phenomenon.

To remove the effect of the short circuit current, we disconnected the P transistors from the output, and pre-charge the internal capacitances, so the N transistors only discharge the load and the internal capacitances. To eliminate Miller effect, we implemented a version of SPICE level 1 model with no coupling capacitances between terminals of a transistor. To eliminate the body effect, we connected the substrate terminal of the top N transistor (N_X) to the source of the same transistor, making $V_{SB} = 0$.

TABLE 2. Magnitude variation of delay phenomena at a two-input NAND gate.

Delay phenomenon	Delay difference %		
	Ave	Min	Max
Short circuit current	83.2	27	230
Initial states of internal capacitances (pin-to-pin delay)	12.4	-29	24
Impedance matching (in simultaneous switching with skew)	9	3.5	14
Miller effect	73.8	39.4	181
Body effect	7	4	9.5

The results are shown in Table 2. Stopping early discharge is not listed because we were not able to eliminate only this effect. This gate has only one internal capacitance hence there is no charge sharing. When we pre-charge the internal capacitance, the delay may increase (+24%) or decrease (-29%) compared to the pre-discharged case. (Any decrease is caused due to impedance matching.) The average delay differences are computed based on absolute values of delay differences.

From Table 2, we conclude that the effect of short circuit current, Miller effect, and the initial state of internal capacitances are very significant. This shows that inverter-collapsing based models [6][19][20] that ignore the internal capacitances will result in significant errors for some combinations of input transition times and skews.

6. Identifying Input Waveforms That Capture the Delay Phenomena

The advantage of empirical models is that the delay effect of each phenomenon does not need to be considered separately and then combined together. For any given input waveform, the simulator takes care of all relevant effects.

However, developers need to identify and simulate the cases where each targeted delay phenomenon is excited, to ensure that its effects are captured. A simulation scenario is characterized by the vectors, transition times, skews, initial states of internal capacitances, and so on.

Short circuit current occurs when both P and N transistors are on during input switching. Its delay effect for a single input transition can be easily captured, since only one input variable (input slew) needs to be varied. In contrast, it is difficult to accurately capture the short circuit current effect due to simultaneous switching of multiple inputs. Stopping early discharge also occurs during simultaneous switching. To capture the effects of initial states of internal capacitances, we pre-charge/discharge internal capacitances to different possible values during characterization. Impedance matching is also captured in the same way. The Miller effect is partially excited when simultaneous transitions are applied to capture effects of short circuit current. Transitions that occur much earlier than the final simultaneous transitions also excite the Miller effect. Body effect is positively correlated to "states of internal capacitances". Hence, no extra requirement is enforced on gate inputs to excite this effect. The internal capacitances need to be characterized for a larger number of distinct initial voltages for considering charge sharing.

TABLE 3. Capturing delay phenomena in empirical models.

Delay phenomenon	Methods to capture its effects
Short circuit current	Simultaneous switching
Initial states of internal capacitances	Pre-charge/discharge internal capacitances
Stopping early discharge	Simultaneous switching
Impedance matching	Pre-charge/discharge internal capacitances to different possible values
Miller effect	Simultaneous switching and effects of early transitions
Body effect	State of internal capacitances
Charge sharing	Pre-charge/discharge internal capacitances to different possible values

The requirements of exciting the seven proposed delay phenomena by input waveform discussed above are summarized in Table 3. Note that the waveforms used to capture all phenomena must only cover a range of *simultaneous switching* (slews and skews), *state of internal capacitances*, as well as *early transitions* that are not a part of simultaneous switching. By capturing the first two of these three situations, almost all effects of the seven delay phenomena are captured. What is missed is the Miller effect related to charge injected by early transitions.

7. Modeling Two-Input Gates

The accuracy of simulation results can be greatly improved by enhancing the most frequently encountered cases. The most frequently used case, pin-to-pin propagation, is already well taken care in the traditional pin-to-pin model. Taking advantage of the observation that two simultaneous transitions occur much more often than three or four, we propose to enhance modeling accuracy by improving how two simultaneous transitions are processed. Thus our new model will capture all delay effects due to states of internal capacitances and the last two simultaneous transitions.

We first consider only two-input NAND gates and then generalize it. We will only characterize the cases where the initial state of each internal capacitance is either fully charged or fully discharged. An initial state at a different intermediate voltages caused by charge sharing will be covered in Section 8. All discussion below is based on NAND gates. NOR gates can be studied as duals of NAND gates.

7.1 Timing Functions (for a Two-input NAND)

During test generation, all circuit parameters (e.g., device sizes and loads) remain fixed. In contrast, timing parameters (e.g., arrival times, transition times) may change from vector to vector. Hence, the delay and transition times for a two-input NAND gate need to be represented as functions of timing variables.

Given the *arrival times* and *transition times* of transitions at a gate's inputs and the *initial states of internal capacitances*, we compute *gate delays* and *output transition times*. The *output arrival time* of a gate is computed using the input arrival times and appropriate gate delays.

Consider only the cases where all inputs of a gate have either non-controlling values or transitions to the same value as in [8] and each internal capacitance is fully charged or discharged. We use **C** to represent the number of internal capacitances which are pre-charged to V_{dd} - V_{th}. The gate delay and output transition time of a two-input NAND gate is represented by the following timing functions (Figure 10): (a) a rise delay function for two simultaneous input transitions, $\mathbf{d^Z_R}(T^X_F, T^Y_F, \delta^{X,Y})$; and (b) a rise transition time function for two simultaneous input transitions, $\mathbf{t^Z_R}(T^X_F, T^Y_F, \delta^{X,Y})$; (c) a fall delay function for two simultaneous input transitions, $\mathbf{d^Z_F}(T^X_R, T^Y_R, \delta^{X,Y}, C)$; and (d) a fall transition time function for two simultaneous input transitions, $\mathbf{t^Z_F}(T^X_R, T^Y_R, \delta^{X,Y}, C)$.

7.2 Trends with Respect to Single Variables

We will adopt the rise delay and transition functions from [8] and develop new functions for fall delay and transition time. In the following, unless otherwise specified the direction of input transitions will be to-non-controlling, *i.e. rising* for NAND gates and falling for NOR gates. The relations

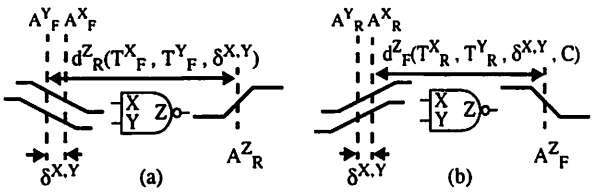

FIGURE 10. (a) Rise delay function and (b) fall delay function.

between output variables and each input variable for the new functions at a two-input NAND gate (Figure 1) are further detailed in Figure 11.

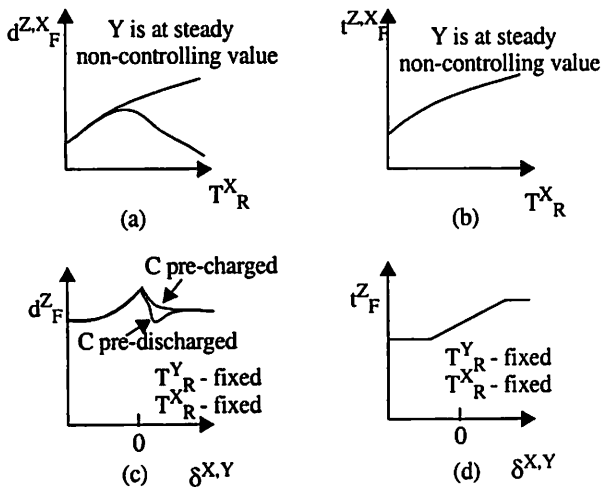

FIGURE 11. Fall timing functions vs. input variables.

Although pin-to-pin delay is either monotonic or bitonic, the output transition time always increases as input transition time increases (Figure 11.(b)). An example of Figure 11.(c) has been discussed in Section 4.8, the maximal delay always occurs at $\delta^{X,Y} = 0$. The minimal delay always occurs at the side (left for negative skew and right for positive skew) where the transition at an input closer to the output leads the other transition. In Figure 11.(d), for both pre-charged and pre-discharged cases, the output transition time is not impacted significantly by simultaneous switching and can be easily approximated by three piecewise linear segments.

7.3 Finding Empirical Formulae

We capture the skew-delay relation in Figure 11.(c) using two separate three/four point approximations for pre-charged/pre-discharged cases, respectively. The formulae for four point approximation (Figure 12) are listed below.

$S0 = 0$

$D0(T^X_R, T^Y_R) = K_{10}*(T^X_R)^2 + K_{11}*(T^Y_R)^2 + K_{12}*T^X_R + K_{13}*T^Y_R + K_{14}$

$D1(T^X_R) = K_{20}*(T^X_R)^2 + K_{21}*T^X_R + K_{22}$

$S1(T^X_R, T^Y_R) = K_{30}*(T^Y_R)^2 + K_{31}*T^X_R + K_{32}*T^Y_R + K_{33}$

Here K_{ab} is a constant obtained from empirical results. Formulae for S2 and S3 are similar to that for S1. D2 is similar to D1. D3 is similar to D0. The set of scenarios identified in Section 6 are simulated using SPICE. Then classical curve fitting approaches are used to compute all coefficients in these formulae. The three point approximation for skew-transition time relation in Figure 11 is similar.

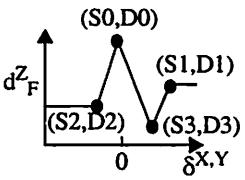

FIGURE 12. Four point piecewise linear approximation for skew-delay relation when the internal capacitance is pre-discharged.

8. Extension to Obtain a General Model

We have developed a delay model considering simultaneous switching for a two-input gate with fixed load. Next we extend this model to cover general cases using the following steps: (1) Capture two simultaneous transitions in simple gates with more than two inputs. (2) Capture the delay for larger numbers of simultaneous transitions using upper and lower bounds. (3) Capture charge sharing by upper and lower bounds. (4) Capture the impact of Miller effects caused by earlier transitions. (5) Capture the effect of load. (6) Finally, we automate the process of generating this delay model.

8.1 Simple Gates with More than Two Inputs

When characterizing simple gates with more than two inputs, we assume that simultaneous transitions may occur at any two gate inputs and perform the same characterization as Section 7 for every pair of inputs. For example, for a four-input NAND gate, the number of input pairs to be characterized is C(4,2)=6.

8.2 More Simultaneous Transitions

Considering three simultaneous to-non-controlling transitions at the inputs of a simple gate, the effect of latest two transitions can be well captured since they are characterized directly. The third latest transition will not occur earlier than the second latest transition. So the latest time that it may occur is the same as that of second latest transition (skew between these two transitions is 0). Adding the third latest transition, the gate delay can be upper bounded by either (1) the third latest transition occurs at the same time with the second latest transition which excites stronger short circuit current effect and Miller effect, or (2) the third transition occurs much earlier (this case may be the same as one where only the last two transitions occur) that helps charge up the internal capacitances and causes longer delay.

Similarly, the minimal gate delay may occur when (1) only the last two transitions impact the gate delay, or (2) the

third latest transition partially charges up the internal capacitances, and so forms a good impedance match and therefore reduces the gate delay.

In case (1) where the third last transition does not induce impedance match, since the third transition will not decrease the gate delay, so the gate delay is lower bounded by the delay caused by only the last two transitions

For case (2), as mentioned in Section 4.8, impedance match does not occur when all internal capacitances are pre-charged. The most significant impedance match occurs when the internal capacitances are fully pre-discharged and the third latest transition is at the top of the serial transistors. In such cases, the internal capacitances will be partially charged up before the last two transitions occur. For a four-input NAND gate with all three internal capacitances pre-discharged, the impedance match may cause up to 40% delay reduction over the delay caused by the last two simultaneous transitions. When the number of pre-discharged internal capacitances is reduced to two, this reduction caused by impedance match becomes no more than 5%. When the third latest transition reduces the gate delay, we bound the worst minimal delay by *n*% reduction of the delay of last two simultaneous transitions, where *n* may be different for each case in the classification.

Consider a three-input NAND gate with output Z and inputs X, Y, W. Assume that all three inputs have simultaneous to-non-controlling transitions and W has the earliest transition and has the least impact on the gate delay. Using the above approximation, we reduce the variables of the characterization from six variables (T^X, T^Y, T^W, $\delta^{X,Y}$, $\delta^{X,W}$, C) to five variables (T^X, T^Y, T^W, $\delta^{X,Y}$, C) since variable $\delta^{X,W}$ is not required for our computation of upper and lower bounds. (Recall that lower bound is computed by assigning $\delta^{X,W}$ as the best impedance match and -∞. Upper bound is computed by assigning $\delta^{X,W}$ as min($\delta^{X,Y}$, $\delta^{Y,X}$) and -∞). We further replace the five-variable characterization to two four-variable characterizations - (T^X, T^Y, T^Y, $\delta^{X,Y}$, C) and (T^X, T^Z, T^Z, $\delta^{X,Y}$, C) by using either T^Y or T^Z to replace the other, and use the larger and smaller of the two to compute upper and lower bounds, respectively. Similarly, this method can be used to reduce the characterization to four variables for more simultaneous transitions.

8.3 Charge Sharing

An example illustrating the delay effect of charge sharing at a four-input NAND gate (Figure 8) is shown in Figure 13. Here two transitions with the same transition time are applied to the top and bottom transistors of a four-input NAND gate with a fixed skew. We vary the input transition time and measure the delay for the four cases - the voltages of all internal capacitances are (V_{dd}-V_{th}) multiplied by 0/3, 1/3, 2/3, and 3/3 respectively.

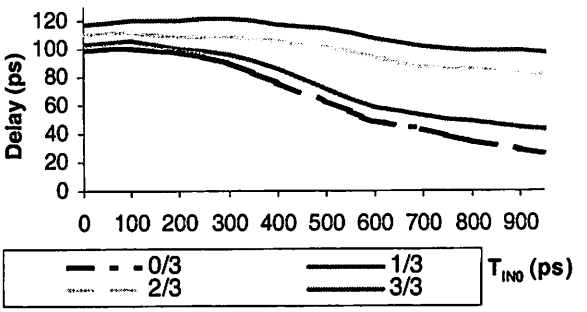

FIGURE 13. Effect of charge-sharing on gate delay (T_{IN0}=T_{IN3}, skew = 300ps).

Charge sharing causes the voltages of internal capacitances to some intermediate voltages (1/3 and 2/3 in Figure 13), the gate delay can be bounded by the two cases where the internal capacitances involved in charge sharing are fully charged (3/3) and fully discharged (0/3). The observation is true in general and has been validated for different input transition times, skews, and number of internal capacitances involved. This is true because impedance matching, which causes the cases with more charge to have shorter delay, never occurs here. So the delay increases monotonically as the charge at internal capacitances increases.

For each charge sharing case in Figure 13, we compute the V_{DS} of each transistor in Table 4. If the same transition is applied at all inputs of a gate, effect of impedance match is expected to be strongest if each V_{DS} is around 0.25*V_{dd}. In contrast, in all these four cases, the V_{DS} of two middle transistors are 0. This limits the current flowing through these two transistors and therefore leads to a poor impedance match.

TABLE 4. V_{DS} value of each serial transistor in each charge sharing case.

Transistor input name	Voltages at the three internal capacitances			
	0	1/3	2/3	3/3
IN0	V_{dd}	2/3V_{dd}	1/3V_{dd}	0
IN1	0	0	0	0
IN2	0	0	0	0
IN3	0	1/3V_{dd}	2/3V_{dd}	V_{dd}

Our characterization will cover all the possible cases where each internal capacitance is either fully charged or fully discharged. So the cases that provide upper/lower bounds for charge sharing are covered.

8.4 Capture Early Transitions

As mentioned in Table 3, to capture Miller effect, we need to capture not only simultaneous switching but also the early transitions that do not participate in simultaneous

switching. Below we call the position of the transistor where a transition is applied as the position of the transition.

Early transitions which do not participate in the final simultaneous switching may inject extra charge to internal capacitances. However in most cases, its impact on the gate delay is insignificant. (1) When the early transition is above all final transitions or below all transitions, the injected charges do not change the voltages of internal capacitances. (2) When the internal capacitances are fully pre-charged, the amount of injection is very small because it is difficult to inject charge to a fully charged capacitances (only 1.5% difference according to our experiments). In contrast, when internal capacitances are pre-discharged and the position of the early transition is in between two final transitions, the effect of early injection is most significant. (3) However, even in this case, if the final transition closest to ground arrives earlier than the other final transitions, the charge injected by early transitions may be removed before all the transistors in series conduct. So the early charge does not impact the gate delay.

We measure the delay effect of early transition at various input corners of a 3-input NAND gate. The most significant extra delay is 16ps. We add this extra delay to upper bound the delay only for the cases when the early transitions can impact the gate delay, as identified above.

8.5 Capture the Effect of Load

By our characterization, the magnitude of increased delay caused by extra load is quite regular, and can be simply captured by adding a degree 2 polynomial function of load into the model proposed in Section 7.

8.6 Characterization Effort

We automate the process for cell characterization as shown in Figure 14. The characterization has four portions: characterizing pin-to-pin delay, delay for two simultaneous transitions, delay for more than two simultaneous transition, and load respectively. We will use a 4-input NAND gate to illustrate the characterization effort. The time for characterizing pin-to-pin delay and load is ignorable compared to the time for characterizing simultaneous switching.

For characterizing two simultaneous transitions proposed in Section 7 for a two-input NAND gate, we perform the characterization for 6 pairs of inputs (pick any two out of four inputs) with 4 different transition times for each input and 200 skew values. The internal capacitance between these two inputs may be initialized as pre-charged or pre-discharged. In total 38400 (6*4*4*200*2) simulations are performed which take 7 hours on a SUN Ultra SPARC III 750MHz machine.

For characterizing three simultaneous transitions proposed in Section 8, with the same numbers of input transition time and skew as above, there are 12 input

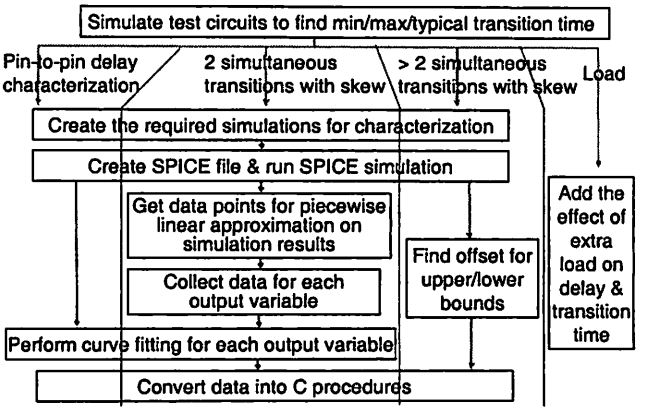

FIGURE 14. Flow chart of cell characterization automation.

combinations to be characterized (4 possible positions for latest transition and 3 possibilities to pick two from the rest three inputs). There are three different initial states of the internal capacitances. (All capacitances above and below transistor X, Y, and W are pre-charged and pre-discharged respectively.) The number of simulation will be 115200 (12*4*4*200*3). Similarly characterizing four simultaneous transitions takes 51200 (4*4*4*200*4) simulations.

To reduce the run time, the number of characterized skews can be reduced with the price of accuracy. Since approximation is also used during curve-fitting for data points in Section 7.3. We expect to see very little accuracy loss when reducing the number of characterized skews.

9. Timing Analysis

Static timing analysis provides min-max timing ranges for each line in a circuit for both rising and falling transitions. The ranges represent bounds on minimum and maximum delay values over all possible pairs of vectors. In timing analysis (Figure 15) arrival times (**A**) and transition times (**T**) at a gate's output are calculated based on these values at gate inputs. These values are computed via a forward traversal starting at the primary inputs. Similarly, the required times (**Q**) are computed via a backward traversal starting at primary outputs. If the arrival time range does not overlap with the required time range for the rising/falling transitions at a line, then the given timing requirements cannot be satisfied and a delay error is found. Delay transfer functions for forward and backward calculations in timing analysis are defined for the proposed model. The min-max ranges in the proposed timing analysis are due to the unspecified input values, pulses, as well as approximations that ignore data dependencies caused by fanouts and reconverges. Since current delay model handles only transitions in the same direction at the inputs of each gate, pulses are ignored.

9.1 Timing Information

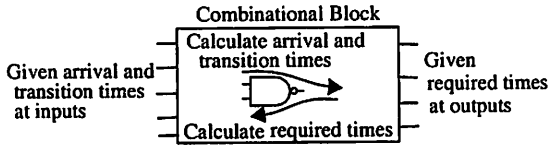

FIGURE 15. Overall structure of timing analysis.

In our min-max range representation, the timing windows in [8] are used (Figure 16). The earliest/latest arrival times and the shortest/longest transition times of rise/fall transitions are recorded for calculating the timing information for the next stage. The smallest (largest) arrival time of falling (rising) transition on line X is represented as $A^X_{F,S}$ ($A^X_{R,L}$). Transition and required times are represented similarly.

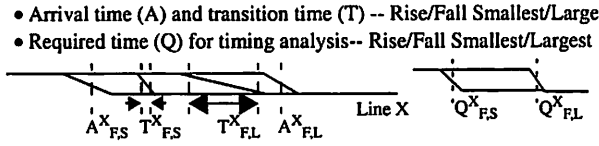

FIGURE 16. Timing information used in our method.

9.2 Static Timing Analysis

9.2.1 Two-Input Gate

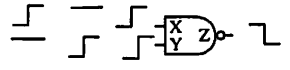

FIGURE 17. Possible input combinations for output rising transition.

Given the arrival and transition times at a gate's inputs, we calculate the corresponding quantities for the gate's outputs. The relations between output variables and input variables in Section 7.2 help identify the input A/T combinations that possibly induce worst case values on the quantities computed for the outputs. Assuming that the initial state of the internal capacitance is not known. A/T calculations for an output falling transition (Figure 17) using our new delay model is shown and explained below.

$$A^Z_{F,S} = \min [A^X_{R,S} + \min \{d^{Z,X}_F(T^X_{R,S}), d^{Z,X}_F(T^X_{R,L})\}, \quad (1)$$
$$A^Y_{R,S} + \min \{d^{Z,Y}_F(T^Y_{R,S}), d^{Z,Y}_F(T^Y_{R,L})\},$$
$$\max \{A^X_{R,S}, A^Y_{R,S}\}$$
$$+ \min_{\beta,\gamma \in \{S,L\}} \{d^Z_F(T^X_{R,\beta}, T^Y_{R,\gamma}, A^Y_{R,S} - A^X_{R,S}, C=0)\}].$$

$$A^Z_{F,L} = \max \{A^X_{R,L}, A^Y_{R,L}\} +$$
$$\max_{\beta,\gamma \in \{S,L,\max\}} \{d^Z_F(T^X_{R,\beta}, T^Y_{R,\gamma}, A^Y_{R,L} - A^X_{R,L}, C=1)\}]. \quad (2)$$

$$\text{where } \overline{T}^X_R = \begin{cases} T^X_{R,\max}, & \text{if } T^X_{R,\max} \in (T^X_{R,S}, T^X_{R,L}); \\ T^X_{R,S}, & \text{else if } d^{Z,X}_F(T^X_{R,S}) > d^{Z,X}_F(T^X_{R,L}); \\ T^X_{R,L}, & \text{otherwise.} \end{cases}$$

Here, $T^X_{R,\max}$ is the value of T^X_R that maximizes $d^{Z,X}_F(T^X_R)$. $T^Y_{R,\max}$ is defined similarly. The transition time can be computed as

$$T^Z_{F,S} = \min \{t^{Z,X}_F(T^X_{R,S}), t^{Z,Y}_F(T^Y_{R,S})\}. \quad (3)$$
$$T^Z_{F,L} = \max \{t^{Z,X}_F(T^X_{R,L}), t^{Z,Y}_F(T^Y_{R,L})\}. \quad (4)$$

For an output falling transition to arrive as early as possible (Equation (1)), we prefer the internal capacitance to be pre-discharged, since the delay in this case will never be larger than that when the internal capacitance is pre-charged (Figure 11.(c)). We also prefer to have a transition only at one input, since multiple transitions usually increase the gate delay. One exception occurs when the internal capacitance is pre-discharged and these two transition have a transition time/skew pairs which decrease the gate delay (e.g. neighborhood of (S3, D3) in Figure 12). Note that we will never prefer to increase input arrival time to decrease skew and therefore decrease gate delay. The reason is that for all skew-delay curves, if we move the skew value 1 unit closer to the minimum delay point (e.g. move from (S1, D1) to (S3, D3) in Figure 12), the delay will never decrease more than 1 (i.e. the slope of each line segment (e.g. (S1, D1) to (S3, D3)) is between -1 and 1). Recall that the output arrival time for simultaneous to-non-controlling transitions is computed by the latest input arrival time plus the gate delay. So if we move the latest input transition one unit later (use $A_{R,S}+1$ instead of $A_{R,S}$) to reduce gate delay, the gate delay will be reduced by less than one unit, then the output arrival time will be increase and therefore violate our purpose to decrease output arrival time. Therefore, we will never try to align two transitions to achieve certain skew for obtaining minimal/maximal output arrival time, since the benefit gain there will always be less than just move all transitions to be as early/late as possible for minimal/maximal output arrival time.

The transition times at X and Y should be either minimal or maximal, depending on which one causes shorter pin-to-pin delay on X and Y, since the shortest delay may be caused by the shortest (Figure 18.(a)) or longest transition time (Figure 18.(b)), but not at any other intermediate time.

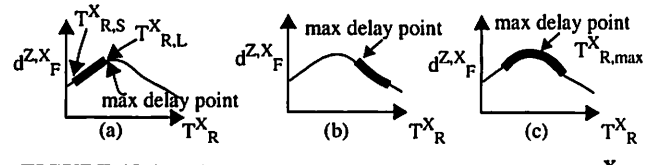

FIGURE 18. Possible transition time min-max range in T^X_R-$d^{Z,X}_F$ curve.

For maximal output arrival time (Equation (2)), we prefer the internal capacitance to be pre-charged. We also prefer to have multiple transitions with zero skew. Even if they are not simultaneous, they produce a output arrival time no

worse than pin-to-pin delay when the internal capacitance is pre-charged.

The maximal gate delay may occur when the input transition times are (a) maximal, (b) minimal, or (c) at some values in between. These three scenarios correspond respectively to the three cases shown in Figure 18, where the min-max range is to the left of the peak, to the right of the peak, or straddles the peak.

Since simultaneous switching does not produce transition time out of the range bounded by the two pin-to-pin output transition times, we can obtain the min/max value of output transition merely by considering pin-to-pin transitions (Equations (3) and (4)).

If the initial state of internal capacitance, C, is known, equations for $A^Z_{F,S}$ and $A^Z_{F,L}$ may be different. If C = 1, $A^Z_{F,S}$ will be simplified to Equation (5), since simultaneous switch may only increase delay and therefore only pin-to-pin delay is picked to achieve minimal arrival time.

$$A^Z_{F,S} = \min [A^X_{R,S} + \min \{d^{Z,X}_F(T^X_{R,S}), d^{Z,X}_F(T^X_{R,L})\}, \quad A^Y_{R,S} + \min \{d^{Z,Y}_F(T^Y_{R,S}), d^{Z,Y}_F(T^Y_{R,L})\}]. \tag{5}$$

$$A^Z_{F,L} = \max [A^X_{R,L} + d^{Z,X}_F(\bar{T}^X_R), A^Y_{R,L} + d^{Z,Y}_F(\bar{T}^Y_R), \quad \max \{A^X_{R,L}, A^Y_{R,L}\} \quad + \max_{\beta,\gamma \in \{S, L, max\}} \{d^Z_F(T^X_{R,\beta}, T^Y_{R,\gamma}, A^Y_{R,L} - A^X_{R,L}, C=0)\}]. \tag{6}$$

$$\text{where } \bar{T}^X_R = \begin{cases} T^X_{R,max}, & \text{if } T^X_{R,max} \in (T^X_{R,S}, T^X_{R,L}); \\ T^X_{R,S}, & \text{else if } d^{Z,X}_F(T^X_{R,S}) > d^{Z,X}_F(T^X_{R,L}); \\ T^X_{R,L}, & \text{otherwise.} \end{cases}$$

For maximal output arrival time when C = 0 (Equation (6)), we will still prefer to have multiple transitions with zero skew. Even if they are not simultaneous, that will produce an output arrival time no worse than pin-to-pin delay. The exception occurs when the skew range is small and covers only the delay range smaller than pin-to-pin (e.g. range 1 in Figure 19). Then pin-to-pin delay may produce the maximal output arrival time. So we pick up the larger of the two.

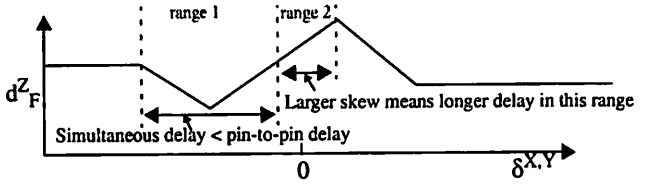

FIGURE 19. Analyze skew-delay relation for achieving maximal delay, given fixed arrival time of last transition where the internal capacitance is pre-discharged.

With respect to the last transition, all other inputs may have only negative skew (arrival earlier). When zero skew is not in the possible range, we will prefer to have a larger skew (a smaller negative value). Although such a skew may lead to a minimal delay (if the left end of the range is at the minimal delay point), but whenever it leads to a delay larger than pin-to-pin (range 2 in Figure 19), larger skew always induces larger delay.

9.2.2 Extension to General NAND Gates

When all internal capacitances are pre-charged, the shape of the skew-delay relation will be like C pre-charged case in Figure 11.(c). So the form of the equations will be similar to Equations (2) and (5). Otherwise, the shape will be like C pre-discharged case in Figure 11.(c). So the form of the equation will be similar to Equations (1) and (6).

For $A^Z_{F,S}$, as Equation (1), we need to consider pin-to-pin delay from each input. In addition, we also need to consider simultaneous switching with minimal number of internal capacitances pre-charged. Since we are computing a minimal arrival time, we will take lower bound of the simultaneous delay. For input corners of the simultaneous delay ($d^Z_F()$), we will take the minimal value of input arrival time and both minimal and maximal transition times at each input as Equation (1). Equations (2), (5), and (6) can be revised in the same way. We will keep Equations (3) and (4) unchanged, since the effect of more simultaneous switching is not significant.

9.3 Timing Analysis for Partially Specified Vectors

After partially specific values are assigned to some circuit lines, worst case corners identified by STA may no longer occur. **Timing analysis for partially specified vectors (TA-PSV)** identifies worst case corners and computes timing information for any given partially specified input values.

TA-PSV on the proposed model is a straight-forward extension of Section 9.2 using the classification given in [24]. The main difference is that we need to handle pre-initialization of internal capacitances. As Section 9.2.2, different pre-initialization values should be classified as two sets: all internal capacitances are pre-charged and otherwise. For both sets, the case by case analysis is similar to that in [24].

10. Experiment Results

10.1 Delay Model

Four previous approaches [6][19][20][22] tried to capture the effect of simultaneous switching for to-non-controlling response. In [8] it is shown that the results in [6][19] can result in significant errors for simultaneous to-controlling transitions, because they do not capture some delay phenomena proposed in [8]. The same is also true for simultaneous to-non-controlling transitions. The approach in [22] is not expected to have high accuracy, since effects of internal capacitances are ignored.

Also in [6][19][20], the authors compare the output voltage vs. time waveforms obtained using their delay models for a few circuit waveforms with SPICE results. These results are misleading for two reasons. (1) The transition time is usually one order of magnitude larger than that of the gate delay, so even a 100% error in delay looks insignificant when compared to SPICE results, since the waveform scale is of the order of transition time. (2) It is more important to demonstrate that the delay model accurately captures the delay phenomena in various corners of input waveforms, rather than demonstrating that waveforms can be matched well in a few cases. The set of scenarios identified in Section 6 insure such coverage.

We implemented a version of our delay model with automatic characterization. In Figure 20 (a), we compare our results with SPICE and the approach in [20]. In this test case, two transitions are applied at IN0 and IN3 of the four-input NAND gate in Figure 8. All internal capacitances are pre-discharged. We vary the skew between these two transitions, where T_{IN0} and T_{IN3} are 300ps and 500ps, respectively. Our model closely correlated to the SPICE results, but the same is not the case for the model proposed in [20], which does not capture the effect of impedance matching. In Figure 20 (b) we show the result with all internal capacitances pre-charged. (This case is not even handled in [20].) Again our model is quite accurate. We believe that error in [20] is mainly due to not considering two dominant effects - short circuit current and initial states of internal capacitances.

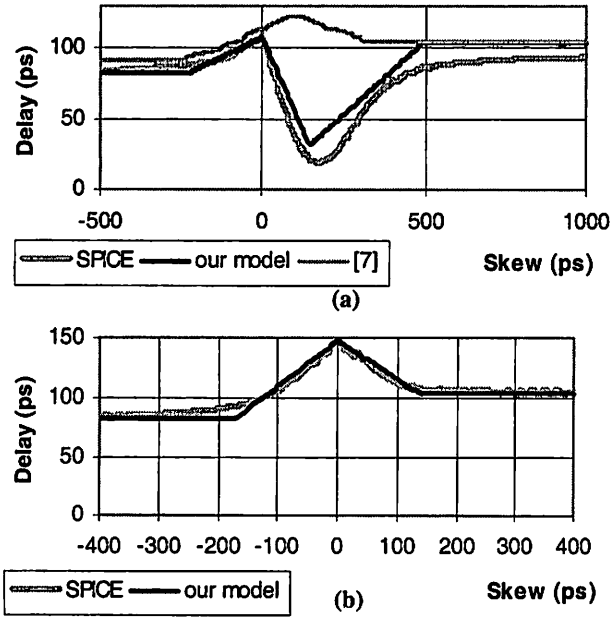

FIGURE 20. Vary $\delta^{IN0, IN3}$ on simultaneous switching at NAND4 when (a) the internal capacitances are all pre-discharged and (b) the internal capacitances are all pre-charged, where T_{IN0}=300ps and T_{IN3}=500ps.

It is interesting to compare the maximal delay effect due to impedance matching (approximately, 100 - 25 = 75 in Figure 20 (a)) and simultaneous switching (approximately, 150 - 100 = 50 in Figure 20 (b)). The maximal effect of three simultaneous switching is 70 (not shown in figures). The relative magnitude of these two phenomena is quite different from that of a two-input NAND gate presented in Section 5 where the delay effects of the phenomena related to simultaneous switching (short circuit current, Miller effect, and stopping early discharge) are much larger than those related to charge/discharge of internal capacitances (initial states of internal capacitances and impedance matching). The reason is that the delay effects of internal capacitances can be accumulated easily as the number of gate inputs increase, but those due to simultaneous switching typically do not.

From our experimental data, we observe a large percentage increase/decrease in delays due to the targeted phenomena compared to pin-to-pin delay. Although it would be rare that the worst delay effects are excited at the same time for many gates along a long path, as combinational logic becomes shallow (say, via pipelining), these delay effects become more significant.

10.2 Static Timing Analysis

Experiment results to be shown after the implementation of our model.

11. New Modeling Advantages

11.1 Compared with Semi-Analytical Models

We have shown the seven main phenomena that impact the delay of simultaneous switching on to-non-controlling transitions and capture them by an empirical model. Capturing all these phenomena by an analytical model is difficult since the time for charging/discharging capacitances is non-linear and the magnitude of one phenomenon may affect another.

Using fully empirical method, we are able to capture all these phenomena together without predicting the effect of each phenomenon and combining them together. By well understanding these phenomena and using proper vectors to excite these phenomena, our delay model is more accurate than previous ones.

Disadvantages of analytical models: (1) A realistic yet simple MOS model is required [17]. formulae are oversimplified for capturing too many kinds of delay effects, i.e. based on SPICE level 1 model, or ignore 2nd/3rd order terms. (2) The effects of each phenomenon are usually not additive, but the analytical functions usually are. (3) Different delay phenomena need not to be captured by different items in formula, but this is what analytical model usually do. (4) Most analytic models are developed based on some

initial conditions, e.g., the initial value of each internal node is zero. In cases where these assumptions are not true, the resulting formulae may lead to significant errors. (5) Automation: We automated the process to characterize our delay model. Once a new process is applied, the new model may be obtained by running the automatic characterization again. Although analytical models is portable as process parameters changes, major modifications are required when the device model is changed (e.g. move from BSIM3 model to UMOS model).

11.2 Compared with Other Empirical Models

Comparing with traditional empirical models, we divide the delay model into many more cases where each case has its own formulae for transition time and delay. We use many coefficients and this improves accuracy. In our classification, each discrete variable (library cell, input position, and number of pre-charged internal capacitances) is enumerated. For each combination in the enumeration, we develop a set of formulae whose input variables are the continuous variables in the original delay functions (input transition times and skew). Although the data size is hundred times larger, the amount of extra data storage is not significant for current program size.

11.3 Able to Model Various Process Corners

By automating the characterization process to make the delay modeling feasible, we also enable the modeling of multiple process corners by running the characterization and obtaining a delay model at each corner. Based on these generated models, we will be able to run timing analysis and test generation at various process corners and validate the timing for each corner. We will also understand how well the test set for nominal process works on exciting worst delay at various process corners, compared to the test set generated at each corner.

12. Conclusions

We identify two new delay phenomena, namely impedance matching and stopping early discharge. We also demonstrate that at least seven delay phenomena must be considered to capture the delays under simultaneous to-non-controlling transitions in sub-100nm VLSI. We quantify the importance of each phenomenon, and identify waveforms and initial state combinations to be used for simulations to excite and capture their effects. By carefully analyzing and capturing the delay effects due to these phenomena, we have developed an accurate model that captures all these effects with manageable complexity.

13. References

[1] R. B. Hitchcock, "Timing verification and timing analysis program", Proc. Design Automation Conf., pp. 594-604, 1982.

[2] C. Visweswariah and R. A. Rohrer, "Piecewise approximate circuit simulation", IEEE Trans. on Computer-Aided Design of Integrated Circuits and Systems, vol. 10, pp. 861-870, July 1991.

[3] Y. H. Shih, Y. Leblebici, and S. M. Kang, "ILLIADS: A fast timing and reliability simulator for digital MOS circuits", IEEE Trans. on Computer-Aided Design of Integrated Circuits and Systems, vol. 12, pp. 1387-1402, Sept. 1993.

[4] L. W. Nagel, "SPICE2, A computer program to simulate semiconductor circuits", Memo UCB / ERL M520, Univ. Cal., Berkeley, May 1975.

[5] IEEE DASC standard delay format (SDF) - web page http://vhdl.org/vi/sdf/.

[6] Y. H. Jun, K. Jun, and S. B. Park, "An accurate and efficient delay time modeling for MOS logic circuits using polynomial approximation", IEEE Trans. on Computer-Aided Design of Integrated Circuits and Systems, vol. 8, pp. 1027-1032, Sept. 1989.

[7] P. Franco and E. J. McCluskey, "Three-pattern tests for delay faults", Proc. VLSI Test Symp., pp. 452-456, 1994.

[8] Liang-Chi Chen, Sandeep Gupta, and Melvin Breuer, "A new gate delay model for simultaneous switching and its applications", Proc. Design Automation Conf., pp. 289-294, 2001.

[9] J. Rubenstein, P. Penfield, Jr., and M. A. Horowitz, "Signal delay in RC networks", IEEE Trans. on Computer-Aided Design of Integrated Circuits and Systems, vol. 2, pp.202-211, July 1983.

[10] T. Sakurai and A. R. Newton, "Alpha-power law MOSFET model and its applications to CMOS inverter delay and other formulas", IEEE J. Solid State Circuits, Vol.25, pp.584-594, Apr. 1990.

[11] L. Bisdounis, S. Nikolaidis, and O. Koufopavlou, "Analytical transient response and propagation delay evaluation of the CMOS inverter for short-channel devices", IEEE J. Solid State Circuits, Vol.33, pp.302-306, Feb. 1998.

[12] K. O. Jeppson, "Modeling the influence of the transistor gain ratio and the input-to-output coupling capacitance on the CMOS inverter delay", IEEE J. Solid State Circuits, vol. 29, no. 6, pp.646-654, June 1994.

[13] V. B. Rao, T. N. Trick, and I. N. Hajj, "A table-driven delay-operator approach to timing simulation of MOS VLSI circuits", Int'l Conf. on Computer Design, pp.445-448, Nov. 1983.

[14] F. C. Chang, C. F. Chen, and P. Subramaniam, "An accurate and efficient gate level delay calculator for MOS circuits", Proc. Design Automation Conf., pp. 282-287, 1988.

[15] D. Overhauser and I. Hajj, "A tabular macromodeling approach to fast timing simulation including parasitics", Proc. Int'l Conf. on Computer Aided Design, pp. 70-73, 1988.

[16] B. S. Cherkauer and E. G. Friedman, "Channel width tapering of serially connected MOSFETs with emphasis on power dissipation". IEEE Trans. on VLSI Systems, vol. 2, no. 1, pp. 100-114. March 1994.

[17] T. Sakurai and A. R. Newton, "Delay analysis of series-connected MOSFET circuits". IEEE J. Solid State Circuits, vol. 26, no. 2, pp. 122-131, February 1991.

[18] J. M. Daga, D. Auvergne, "A comprehensive delay macro modeling for submicrometer CMOS logics", IEEE J. Solid State Circuits, vol. 34, no. 1, pp. 42-55, Jan. 1999.

[19] A. Nabavi-Lishi and N. C. Rumin, "Inverter models of CMOS gates for supply current and delay evaluation",IEEE Trans. on Computer-Aided Design of Integrated Circuits and Systems,

vol. 13, pp. 1271-1279, Oct. 1994.
[20] A. Chatzigeorgiou, S. Nikolaidis, and I. Tsoukalas, "A modeling technique for CMOS gates", IEEE Trans. on Computer-Aided Design of Integrated Circuits and Systems, vol. 18, pp. 557-575, May 1999.
[21] L. Bisdounis and O. Koufopavlou, "Modeling the dynamic behavior of series-connected MOSFETs for delay analysis of multiple-input CMOS gates", Proc. Int'l Symp. on Circuits and System, pp.342 -345, 1998.
[22] V. Chandramouli and K. A. Sakallah, "Modeling the effects of temporal proximity of input transitions on gate propagation delay and transition time", Proc. Design Automation Conf., pp. 617-622, 1996.
[23] N. H. E. Weste and K. Eshraghian, *Principles of CMOS VLSI Design*, 2nd edition, Addison-Wesley Publishing Company, 1993.
[24] Liang-Chi Chen, Sandeep Gupta, and Melvin Breuer, "TA-PSV - Timing Analysis for Partially Specified Vectors", Journal of Electronic Testing: Theory and Applications (JETTA), vol. 18, no. 1, pp. 73-88, Feb. 2002.