Gate Load Delay Computation Using Analytical Models *

Andrew B. Kahng
UCLA Computer Science Department
Los Angeles, CA 90095-1596 USA
abk@cs.ucla.edu

Sudhakar Muddu
Silicon Graphics, Inc.
Mountain View, CA 94039 USA
muddu@mti.sgi.com

Abstract

With submicron technologies, gate delays are dominated by *gate load delays* rather than intrinsic gate delays. While the common approach for computing gate load delay (or total gate delay) is through delay tables (or *k*-factor equations), there are important methodology problems associated with the delay table approach. In this paper, we propose a gate driver model with a Thevenin equivalent circuit consisting of a ramp voltage source whose slew time is obtained from the gate slew tables, and a driver resistance in series with the gate load. We then develop analytical gate delay formulas using this Thevenin driver model and modeling the load with various gate load models under both rising and falling ramp input.

1 Introduction

With submicron technologies the overall path delay between gates is dominated by interconnect delays (including both the effect of interconnect on the driving gate and the pure interconnect propagation delay), rather than intrinsic gate delays. To compute pure interconnect propagation delay, various techniques based on either *simulation* [10, 12, 14] or *analytical formulas* [1, 3, 4] have been proposed. However, total gate delay between an input and output pin pair must still be accurately determined. We express this *total gate delay* (D_{AB} in Figure 1) as the sum of *intrinsic gate delay* and *gate load delay*:

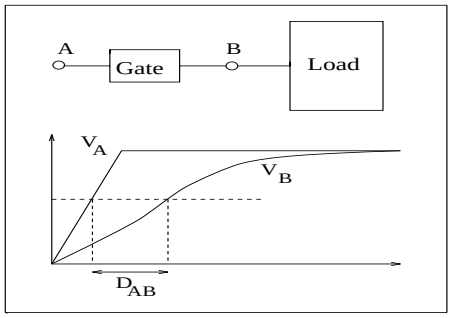

Figure 1: Total gate delay is the sum of intrinsic gate delay and gate load delay.

- *Intrinsic gate delay* is delay due to physical devices (e.g., transistors) in the gate. Intrinsic gate delay can be thought of as total gate delay with infinite load at the output.
- *Gate load delay* is the delay due to the load connected to the output of the gate.

Two popular approaches to gate delay computation are (i) computation of delay through *delay tables* (or *k*-factor equations), and (ii) computation of delay by modeling the gate with a Thevenin equivalent circuit of voltage source and a resistance in series with the gate load. It turns out that the Thevenin equivalent model is a more effective delay model when the load is not purely capacitive, since it naturally captures the interaction of the gate's output resistance and the *RC*/*RLC* load. However, both models are empirical; in particular, the Thevenin equivalent model requires empirical fitting [9] to approximate the resistance value as a function of input slew rate and output load.

In practice, total gate delay or gate load delay for various load values is usually stored for each gate/cell in the library in *delay table* format, which we describe next. The intrinsic gate delay is also known for each gate/cell in the library. While the delay table format has become quite popular, note that the total gate delay can also be computed directly by modeling each gate as a linear resistor with voltage source and then modeling the (discrete or distributed) load using various *gate load models*.

Delay Table Format. Delay tables specify the total gate delay and the output slew rate (rise and/or fall time) for each gate in the library. There exists at least one pair of tables (one table for delay, one table for slew rate) for each cell/gate in the library. Typically, delay tables are developed/characterized as functions of only the input slew rate (rise or fall time) and a single capacitance value which represents the effect of the load. This delay table format is equivalent to the so-called empirical "k-factor" formulas for delay and output rise time.[1]

There are important methodology questions associated with the delay table approach. In practice, delay and rise-times may be obtained by loading the gates with a *discrete load capacitor* and then changing both the load capacitance and input rise-times. But in reality, the output of the gate is connected via interconnects to other gate inputs. Modeling the load at the gate as a single load capacitor may work well for technologies and designs where the area of interconnect at the gate output is small or the interconnect parameters are not dominant compared to gate parameters. With sub-micron technologies the interconnect resistance, capacitance, and inductance must be considered in the delay table characterization (inductance effects will definitely be an issue in the next process generation).

Given the disadvantages of the delay table approach, this paper explores the computation of gate delay by modeling the gate with a Thevenin equivalent circuit of voltage source and a resistance in series to the gate load. We propose a gate driver model with a Thevenin equivalent circuit of ramp voltage source having slew time derived from the gate slew tables and driver resistance in series to the gate load. We then develop analytical gate delay formulas using this Thevenin driver model in conjunction with various gate load models.

2 Review of Gate Load Models

Various *load models* have been proposed for modeling the driving point admittance at the gate output. The gate delays are estimated using these models either through the delay table methodology or through an explicit *simulation* of the gate with the given load model. Before we discuss our analytical (closed-form) expressions for threshold gate delay, we briefly summarize a range of existing gate load models.

2.1 Lumped Models

The simplest approximation of the driving point admittance of the load interconnect tree is the total capacitance of the tree (C_{tot}), which is a (pessimistic) first-order approximation. The actual delay is much smaller than that derived from the lumped capacitance model, because the interconnect resistance acts as a shield to reduce the load capacitance seen by the gate driver. Another simple approximation

*This work was supported by NSF Young Investigator Award MIP-9257982. ABK is currently Visiting Scientist at Cadence Design Systems, Inc. (on sabbatical leave from UCLA), email: ak@cadence.com.

[1] Standard industry delay calculators use 2-dimensional tables for delay and output slew rate of gates. Synopsys [15] uses a similar format for characterizing delays during logic synthesis.

is the lumped *RC* segment model with resistance equal to the total interconnect resistance (R_{tot}) and capacitance equal to the total interconnect capacitance (C_{tot}). This yields an optimistic delay estimate because the total interconnect resistance is lumped together and shields the total capacitance.[2]

2.2 O'Brien/Savarino Π Model

With thinner interconnect geometries, the resistive component of the gate load is comparable to or larger than the gate output resistance, and the gate does not "see" all of the capacitance loading since the metal resistance "shields" some capacitance [11]. For example, if we increase the interconnect resistance of the load and keep the gate output resistance constant then the total gate delay at the output will *decrease* since the interconnect resistance will tend to shield some of the load capacitance. (In this case, while the total gate delay decreases, the increase in interconnect resistance would increase the interconnect propagation delay.)

O'Brien and Savarino [7, 8] proposed using a one-segment Π model to approximate the load at the gate output while still considering resistance shielding effects. Their model approximates the load interconnect at the gate by matching the first three moments of the driving point admittance of the interconnect load. The disadvantage of the Π model is that delay tables need to be expanded to four dimensions: rise time of input voltage, and the three Π model parameters R_1, C_1, C_2.

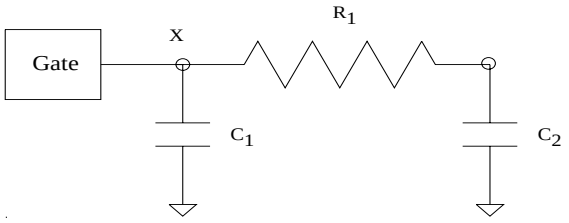

Figure 2: One-segment Π model for matching the first three moments of the driving point admittance of a load interconnect tree.

Let the driving point admittance at the gate output (X) be represented by

$$Y_L(s) = \sum_{i=1}^{\infty} A_i s^i = sA_1 + s^2A_2 + s^3A_3 + \ldots \quad (1)$$

The parameters of the equivalent circuit are obtained by matching the first three moments of the admittance with corresponding coefficients of the driving point admittance of the Π load model in Figure 2, i.e.,

$$R_1 = \frac{-A_3^2}{A_2^3} \quad C_1 = A_1 - \frac{A_2^2}{A_3} \quad C_2 = \frac{A_2^2}{A_3}. \quad (2)$$

2.3 Effective Capacitance Model

The cell tables (or k-factor formulas) for delay and output rise time of gates depend only on the input slew rate and a single load capacitance, which represents the effect of the load. There are two different approaches in the literature for computing such an *effective capacitance*: (i) McCormick's Effective Capacitance Model [6], and (ii) Pillage et al.'s Effective Capacitance Model [11, 13]. The aim of each approach is to approximate the load at gate output using a single effective capacitance.

2.4 Open-Ended *RC* Π Model

In a pre-routing timing analysis, exact routing topology is not available. The paper [2] approximates an estimated interconnect tree by

[2]The lumped capacitance and lumped *RC* models are referred to as *Wire Load Model1* and *Wire Load Model2* in Synopsys manuals [15]. Similar lumped models are available with other industry timing analysis tools.

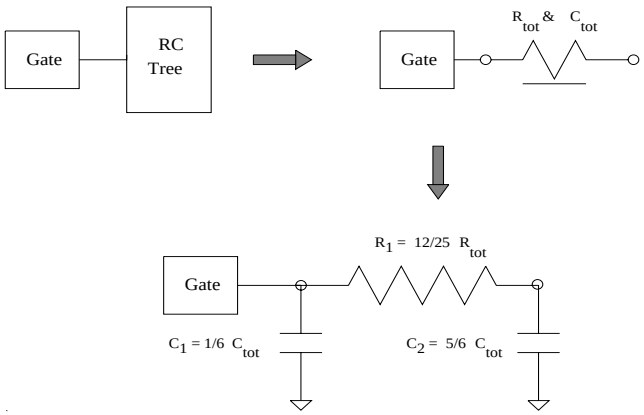

Figure 3: An open-ended *RC* line to capture an *RC* interconnect tree, and the *RC* Π model.

an equivalent open-ended *RC* line whose resistance and capacitance are equal to the (estimated) total interconnect resistance and capacitance, as shown in Figure 3. The open-ended *RC* line approximation still considers the distributed nature of the load interconnect in the calculation of model parameters, but is more efficient since only R_{tot} and C_{tot} values are used. The admittance of an open-ended *RC* line is $Y(s) = \frac{\tanh(\theta)}{Z_0} = sC_{tot} - s^2\frac{R_{tot}C_{tot}^2}{3} + s^3\frac{2R_{tot}^2C_{tot}^3}{15} + \ldots$, where $\theta = \sqrt{R_{tot}sC_{tot}}$ and $Z_0 = \sqrt{\frac{R_{tot}}{sC_{tot}}}$. The three moments of the admittance function when substituted into Equation (2), yield Π model circuit parameters $R_1 = \frac{12R_{tot}}{25}$, $C_1 = \frac{C_{tot}}{6}$ and $C_2 = \frac{5C_{tot}}{6}$. Comparisons of the open-ended model and various other load models are given in [5]. The open-ended Π model can be extended to include inductance effects in the gate load delay computation which will be an issue in the next process generation.

3 Driver Model

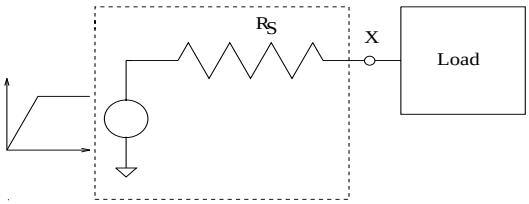

Figure 4: Driver model with ramp input, whose slew time is obtained from the gate slew tables, and series resistance connected to the load.

To model the gate driver, a Thevenin equivalent circuit model has been used with a step input voltage source [6] and a driver resistance which is computed using the transistor linear region resistance value [16]. For complex gates, estimating the driver resistance using the transistor linear region model gives inaccurate delay values. Also, assuming a step input as the voltage source can introduce considerable error. To eliminate the latter inaccuracy, we propose a new driver model for the gate with an equivalent circuit consisting of a linear source resistance (R_S) and a *ramp* input source ($v_{in}(t)$) whose slew time is equal to the output slew time from the cell tables; see Figure 4. (If the model uses gate input slew time as the slew time for the source voltage, the size of the gate will have no effect on the gate delay. Hence, we use the output slew time from the cell tables as the input slew time for the voltage source in the model.) In practice the driver resistance can be computed to reasonable accuracy by taking an average over a range of possible input slew times and effective capacitance values [5].

4 Gate Delay Computation for Lumped Capacitance Load Model

The simplest approximation models the entire load at the gate output with a single lumped capacitance. The gate is modeled with a source ramp input of rise time T_R and a series source resistance R_S (see Figure 4).

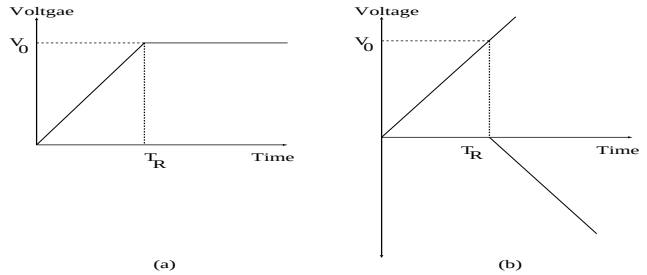

Figure 5: A ramp input function: (a) finite ramp with rise time T_R, and (b) finite ramp decomposed into two shifted infinite ramps.

Figure 5 shows that a finite rising ramp input can be expressed in the time domain as

$$v_{in}(t) = \frac{V_0}{T_R}[tU(t) - (t-T_R)U(t-T_R)] \quad \text{for all } t \geq 0$$

where $U(t)$ denotes the step function. The voltage at the gate output (X) in the transform domain is

$$\begin{aligned} V_X(s) &= \frac{V_{in}(s)}{(1+sR_SC_{tot})} \\ &= \frac{V_0}{T_R}(1-e^{-sT_R})\left[\frac{1}{s^2} - \frac{R_SC_{tot}}{s} + \frac{R_SC_{tot}}{(s+1/R_SC_{tot})}\right] \end{aligned}$$

and the time-domain response is

$$\begin{aligned} v_X(t) &= V_0\left(\frac{t}{T_R} - \frac{R_SC_{tot}}{T_R} + \frac{R_SC_{tot}}{T_R}e^{\frac{-t}{R_SC_{tot}}}\right) \quad \text{for } t \leq T_R \\ &= V_0\left(1 + \frac{R_SC_{tot}}{T_R}(e^{\frac{-t}{R_SC_{tot}}} - e^{\frac{-(t-T_R)}{R_SC_{tot}}})\right) \quad \text{for } t > T_R \end{aligned} \tag{3}$$

The threshold delay is

$$\begin{aligned} T_{RD1} &= R_SC_{tot}\left|\ln\left(\frac{1}{1+\frac{v_{th}T_R}{R_SC_{tot}} - \frac{T_{RD1}}{R_SC_{tot}}}\right)\right| \quad \text{for } t \leq T_R \\ &= R_SC_{tot}\left|\ln\left(\frac{R_SC_{tot}}{T_R}\cdot\frac{(e^{\frac{1}{R_SC_{tot}/T_R}}-1)}{(1-v_{th})}\right)\right| \\ & \qquad \text{for } t > T_R \end{aligned} \tag{4}$$

where v_{th} is the threshold voltage at which delay is computed. For the case of $t \leq T_R$ the threshold delay equation needs to be solved iteratively (typically less than 10 iterations of simple back-substitution are sufficient). Approximate formulas can alternatively be obtained by substituting an upper bound for T_{RD1} in the logarithmic expression and then fitting against SPICE data. The voltage response and threshold delay expressions given an effective capacitance model (C_{eff}) have the same form; simply replace C_{tot} by C_{eff}.

5 Gate Delay Computation for Π Load Model

A second approximation models the entire load at the gate output with the Π model of [7], which captures the load admittance up to the third moment. The driver is again modeled with a Thevenin equivalent circuit consisting of a source ramp input with rise time T_R and a series source resistance R_S. Recall that the parameters of the Π equivalent circuit can be expressed in terms of the driving point admittance at gate output as given in Equation (2). The voltage at the gate output (X) in the transform domain is

$$\begin{aligned} V_X(s) &= V_{in}(s)\frac{(1+sR_1C_2)}{1+s(R_SC_1+R_SC_2+R_1C_2)+s^2R_SR_1C_1C_2} \\ &= \frac{V_0(1-e^{-sT_R})}{T_R}\left[\frac{1}{s^2} - \frac{R_S(C_1+C_2)}{s}\right. \\ &\quad \left. + \frac{(1+s_1R_1C_2)}{b_2s_1^2(s_1-s_2)}\frac{1}{(s-s_1)} - \frac{(1+s_2R_1C_2)}{b_2s_2^2(s_1-s_2)}\frac{1}{(s-s_2)}\right] \end{aligned}$$

where $b_1 = R_S(C_1+C_2)+R_1C_2$, $b_2 = R_SR_1C_1C_2$, *and*

$$s_{1,2} = \frac{-b_1 \pm \sqrt{b_1^2-4b_2}}{2b_2}$$

The time-domain response for $t \leq T_R$ is

$$\begin{aligned} v_X(t) &= \frac{V_0}{T_R}\left[t - R_S(C_1+C_2) + \frac{(1+s_1R_1C_2)}{b_2s_1^2(s_1-s_2)}e^{s_1t}\right. \\ &\quad \left. - \frac{(1+s_2R_1C_2)}{b_2s_2^2(s_1-s_2)}e^{s_2t}\right] \end{aligned}$$

and the response for $t > T_R$ is

$$\begin{aligned} v_X(t) &= \frac{V_0}{T_R}\left[T_R + \frac{(1+s_1R_1C_2)(1-e^{-T_Rs_1})}{b_2s_1^2(s_1-s_2)}e^{s_1t}\right. \\ &\quad \left. - \frac{(1+s_2R_1C_2)(1-e^{-T_Rs_2})}{b_2s_2^2(s_1-s_2)}e^{s_2t}\right] \end{aligned}$$

Depending on the sign of $(b_1^2-4b_2)$ the poles will be either real or complex. We derive threshold delay formulas for the case of real poles only. Using

$$b_1^2 - 4b_2 = R_S^2(C_1+C_2)^2 + R_1^2C_2^2 + 2R_SR_1C_2(C_2-C_1)$$

We see that the load admittance parameters should satisfy $(C_2-C_1) = (2A_2^2 - A_1A_3) \geq 0$ for the poles to be real.[3] Since the magnitude $|s_2|$ is greater than $|s_1|$, the second term in the time-domain response decreases rapidly compared to the first term. Hence, for the threshold delay computation we neglect the e^{s_2t} term in the response. Therefore, an approximation for threshold delay is

$$\begin{aligned} T_{RD2} &\approx \frac{1}{|s_1|}\left|\ln\left(\frac{(1+s_1R_1C_2)}{b_2s_1^2(s_1-s_2)(v_{th}T_R + R_S(C_1+C_2) - T_{RD2})}\right)\right| \\ & \qquad \text{for } t \leq T_R \\ &\approx \frac{1}{|s_1|}\left|\ln\left(\frac{(1+s_1R_1C_2)(e^{T_R|s_1|}-1)}{b_2s_1^2(s_1-s_2)(1-v_{th})T_R}\right)\right| \\ & \qquad \text{for } t > T_R \end{aligned} \tag{5}$$

6 Gate Delay Computation for Open-Ended Π Load Model

Most generally, we can model the load at the gate output by a driving-point admittance Y_L, then approximate the first few terms of the infinite series expansion depending on the accuracy required, $Y_L(s) = sA_1 + s^2A_2 + \ldots + s^kA_k + \ldots$ where A_k is the k^{th} moment of the load admittance at gate output (see Figure 6). For the open-ended RC Π model the first three moments of the driving point admittance are expressed in terms of the total resistance and capacitance of the load

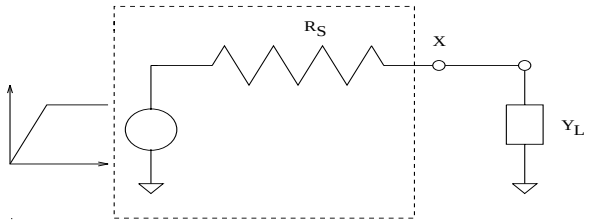

Figure 6: Ramp input and series resistance connected to a general admittance load model.

network. We now use these moments to compute an analytical formula for gate load delay using the same driver model as above.

The voltage at the gate output (X) for a finite rising ramp input is written in the transform domain as

$$\begin{aligned} V_X(s) &= V_{in}(s)\frac{1}{1+R_S Y_L} \\ &= \frac{V_0(1-e^{-sT_R})}{T_R}\left[\frac{k_1}{s-s_1}+\frac{k_2}{s-s_2}+\frac{k_3}{s}+\frac{k_4}{s^2}\right] \end{aligned}$$

where s_1 and s_2 are the poles of the transfer function. Solving for all the variables, the time-domain response is

$$\begin{aligned} v_x(t) &= \frac{V_0}{T_R}\left[-b_1+t+\frac{1+b_1s_2}{s_2-s_1}e^{s_1t}+\frac{1+b_1s_1}{s_1-s_2}e^{s_2t}\right] \\ &\qquad \text{for } t \le T_R \\ &= \frac{V_0}{T_R}\left[T_R+\frac{(1+b_1s_2)(e^{s_1t}-e^{s_1(t-T_R)})}{s_2-s_1}\right. \\ &\quad \left.+\frac{(1+b_1s_1)(e^{s_2t}-e^{s_2(t-T_R)})}{s_1-s_2}\right] \quad \text{for } t > T_R \qquad (6) \end{aligned}$$

Depending on the sign of $(b_1^2-4b_2)$ the poles will be either real or complex. We now derive threshold delay formulas assuming the poles are real (if the poles are complex, a similar analysis can be applied [3]). Again, $|s_2| > |s_1|$ allows us to neglect the e^{s_2t} term in the response, and an approximation for threshold delay is

$$\begin{aligned} T_{RD3} &= \frac{1}{|s_1|}\left|\ln\left(\frac{(1+b_1s_2)}{(s_2-s_1)(b_1+u_{th}T_R-T_{RD3})}\right)\right| \\ &\qquad \text{for } t \le T_R \\ &= \frac{1}{|s_1|}\left|\ln\left(\frac{(1+b_1s_2)(e^{|s_1|T_R}-1)}{(s_2-s_1)T_R(1-v_{th})}\right)\right| \\ &\qquad \text{for } t > T_R \end{aligned}$$

For $t \le T_R$ the threshold delay equation must again be solved iteratively. However, approximate formulas can be obtained by substituting an upper bound for T_{RD3} in the logarithmic expression and then fitting against SPICE data. Since the threshold delay computed from the case $t > T_R$ is greater than T_R, an alternative formula for the threshold delay can be obtained as $T_{RD4} = T_R + \tau_{RD4}$. Substituting into Equation (6), the threshold delay is

$$T_{RD4} = T_R + \frac{1}{|s_1|}\left|\ln\left(\frac{(1+b_1s_2)(1-e^{-|s_1|T_R})}{(s_2-s_1)T_R(1-v_{th})}\right)\right| .$$

Our approach can be applied to compute delay for falling ramp input for all the cases above. For the present load model, the finite falling ramp input can be expressed in the time domain as

$$v_{in}(t) = \frac{V_0}{T_F}\left[T_F U(t) - tU(t) + (t-T_F)U(t-T_F)\right] \quad \text{for all } t \ge 0$$

[3]For most practical cases the value of C_2 is greater than C_1 (refer to [2]) and hence the poles are real. Also, $C_2 > C_1$ in the open-ended load model of [2] for the driving point admittance.

where T_F is the fall time. The voltage at the output node of the gate (X) for falling ramp input in the transform domain is

$$V_X(s) = \frac{V_0}{T_F}\left(\frac{T_F}{s}-\frac{1}{s^2}(1-e^{-sT_F})\right)\frac{1}{1+sb_1+s^2b_2}$$

Neglecting the term with pole s_2, we obtain threshold delay

$$\begin{aligned} T_{FD3} &\approx \frac{1}{|s_1|}\left|\ln\left(\frac{1+b_1s_2+T_Fs_2}{(s_1-s_2)(v_{th}T_F-b_1-T_F+T_{FD3})}\right)\right| \\ &\qquad \text{for } t \le T_F \\ &\approx \frac{1}{|s_1|}\left|\ln\left(\frac{T_Fs_2+(1+b_1s_2)(1-e^{-s_1T_F})}{(s_1-s_2)v_{th}T_F}\right)\right| \\ &\qquad \text{for } t > T_F \qquad (7) \end{aligned}$$

7 Conclusions

We have proposed a new gate driver model using a Thevenin equivalent circuit consisting of ramp voltage source with slew time obtained from gate slew tables, and driver resistance in series with the gate load. We have also developed analytical gate delay formulas using this Thevenin driver model and modeling the load with various gate load models. These analytical gate delay formulas can be used at various stages of the synthesis/layout optimization loop to speed up the delay analysis and provide insight on how gates and interconnects together determine performance.

REFERENCES

[1] W. C. Elmore, "The Transient Response of Damped Linear Networks with Particular Regard to Wideband Amplifiers", *Journal of Applied Physics* 19, Jan. 1948, pp. 55-63.

[2] A. B. Kahng and S. Muddu, "Efficient gate Delay Modeling for Large Interconnect Loads", *Proc. IEEE Multi-Chip Module Conf.*, Feb. 1996 (Submitted to IEEE Trans. on CAD).

[3] A. B. Kahng and S. Muddu,' "Accurate Analytical Delay Models for VLSI Interconnects", *IEEE Int. Symposium on Circuits and Systems*, May 1996.

[4] A. B. Kahng, K. Masuko, and S. Muddu, "Analytical Delay Models for VLSI Interconnects Under Ramp Input", *To appear in IEEE ICCAD*, Nov. 1996.

[5] A. B. Kahng and S. Muddu, "A Glossary on Analysis and Modeling of VLSI Interconnections", *manuscript at Cadence Design Systems, Inc.*, Feb. 1996.

[6] S. P. McCormick, "Modeling and Simulation of VLSI Interconnections with Moments", *PhD Thesis*, MIT, June 1989.

[7] P. R. O'Brien and T. L. Savarino, "Modeling the Driving-Point Characteristic of Resistive Interconnect for Accurate Delay Estimation", *Proc. IEEE ICCAD*, 1989, pp. 512-515.

[8] P. R. O'Brien and T. L. Savarino, "Efficient On-Chip Delay Estimation for Leaky Models of Multiple-Source Nets", *Proc. IEEE Custom Integrated Circuits Conf.*, 1990, pp. 9.6.1-9.6.4.

[9] J. K. Ousterhout, "A Switch-level Timing Verifier for Digital MOS VLSI", *IEEE Trans. on CAD*, July 1985, pp. 336-349.

[10] L. T. Pillage and R. A. Rohrer, "Asymptotic Waveform Evaluation for Timing Analysis", *IEEE Trans. on CAD* 9, Apr. 1990, pp.352-366.

[11] J. Qian, S. Pullela, and L. Pillage, "Modeling the "Effective Capacitance" for the *RC* Interconnect of CMOS gates", *IEEE Trans. on CAD*, December 1994, pp. 1526-1535.

[12] C. L. Ratzlaff, N. Gopal, and L. T. Pillage, "RICE: Rapid Interconnect Circuit Evaluator", *Proc. 28th ACM/IEEE Design Automation Conf.*, June 1991.

[13] C. Ratzlaff, S. Pullela, and L. Pillage, "Modeling the *RC* Interconnect effects in a Hierarchical Timing Analyzer", *Proc. IEEE Custom Integrated Circuits Conference*, May 1992, pp. 15.6.1-15.6.4.

[14] M. Sriram and S. M. Kang, "Fast Approximation of The Transient Response of Lossy Transmission Line Trees", *Proc. ACM/IEEE Design Automation Conf.*, June 1993, pp. 691-696.

[15] Synopsys, "Design Compiler Family Reference Manual", version 3.3a, March 1995.

[16] N. Weste and K. Eshraghian, *Principles of CMOS VLSI Design: A systems Perspective*, Addison-Wesley, 1988.