Performance Computation for Precharacterized CMOS Gates with *RC* Loads

Florentin Dartu, Noel Menezes, and Lawrence T. Pileggi, *Senior Member, IEEE*

***Abstract*— For efficiency, the performance of digital CMOS gates is often expressed in terms of empirical models. Both delay and short-circuit power dissipation are sometimes characterized as a function of load capacitance and input signal transition time. However, gate loads can no longer be modeled by purely capacitive loads for high performance CMOS due to the *RC* metal interconnect effects. This paper presents a methodology for interfacing empirical gate models to reduced order *RC* interconnect models in terms of a nonlinear iteration procedure. The delay and power are calculated with errors on the same order as those for the original empirical equations. Moreover, a linear equivalent gate model is generated which accurately captures the delays at the interconnect fan-out nodes.**

I. Introduction

TO SHORTEN design cycles, the digital systems are often designed at the gate and/or cell level. In contrast to designing at the transistor level, the performance of the gates and cells can be precharacterized to significantly speed-up the performance analysis process. Specifically, gate delays are precharacterized for static timing analysis, and short-circuit power dissipation (the power dissipation due to the short time period for which the p- and n-transistors are simultaneously "on") is modeled empirically for power analysis. These delay and power measures are usually expressed as functions of the input-signal transition time and load capacitance in terms of table look-up models or fitted equations (often called k-factor equations because the polynomial coefficients are k's [1]).

As the minimum feature sizes for integrated circuits scale downward, the resistive component of the RC interconnect load is now comparable to the gate output impedance and a single capacitor is no longer a valid gate load model [2]. Precharacterizing gates and cells for more complex admittance load models, even a second-order π-load [2], would be prohibitively expensive. In [3] the delay k-factors were derived as a function of a single load capacitance, and the compatibility with the second order π-load was obtained in terms of an *effective capacitance* C_{eff}. This model and extensions of it [4] have been successfully used in the design of high-speed microprocessors [5].

There are times, however, when load models beyond second-order are required. Moreover, to accurately calculate the subsequent gate delays, a waveform approximation is needed, too. Due to the low-pass nature of the RC loads, the "digital" signals exhibit nondigital behavior, and the performance, especially the delay, becomes largely affected by the waveshapes. Therefore, a more general compatibility with RC loads and better waveform estimation is essential for maintaining accuracy.

For estimating the power dissipation, the energy stored on the load is easily calculated, $\frac{1}{2}CV^2$, but the short-circuit component of power dissipation varies as a function of gate capacitance-load. However, the resistive component of the RC load, in addition to dissipating some of the $\frac{1}{2}CV^2$ heat, shields some of the load capacitance from the gate and thereby increases the short circuit power dissipation.

In this paper, we present a simple CMOS gate model that accurately approximates the voltage and current waveforms throughout the circuit by generalizing the effective capacitance model for nth order driving point admittance loads. In contrast with the previous C_{eff} approaches, this model is independent of the input signal threshold points. Furthermore, with the use of a comprehensive linear gate-interconnect-load model, we are able to more efficiently compute the waveforms and, therefore, the delays, not only at the gate output but at the RC interconnect fanout points. Observing that the short-circuit power computation problem is equivalent to the delay case from the RC load point of view, we present the first generalization of the effective capacitance methodology to this problem.

The paper is organized as follows. In Section II, we review some previous work for gate and cell level static timing analysis and power estimation. In Section III, we introduce a time-varying Thevenin equivalent gate/cell model and explain its use. Section IV describes the various considerations that must be taken into account for the computation of the Thevenin resistance and the implications on static timing analysis. In Section V, we extend the model to consider estimation of the short-circuit power dissipation. Then, Section VI offers some examples followed by future work and conclusions in Section VII.

II. Background

There are two approaches to gate delay modeling which have gained widespread acceptance: 1) empirically derived

Manuscript received March 7, 1995; revised September 9, 1995. This work was supported in part by the Semiconductor Research Corporation under Contract 94-DJ-343, by the National Science Foundation under Contract MIP-9157263, and by the IBM Corporation. This paper was recommended by Associate Editor R. A. Saleh.

F. Dartu was with the Department of Electrical and Computer Engineering, The University of Texas, Austin, TX 78712 USA. He is now with the Department of Electrical and Computer Engineering, Carnegie Mellon University, Pittsburgh, PA 15213 USA.

N. Menezes was with the Department of Electrical and Computer Engineering, The University of Texas, Austin, TX 78712 USA. He is now with Intel Corporation, Hillsboro, OR 97124 USA.

L. T. Pileggi (formerly L. Pillage) was with the Department of Electrical and Computer Engineering, The University of Texas, Austin, TX 78712 USA. He is now with the Department of Electrical and Computer Engineering, Carnegie Mellon University, Pittsburgh, PA 15213 USA.

Publisher Item Identifier S 0278-0070(96)04631-3.

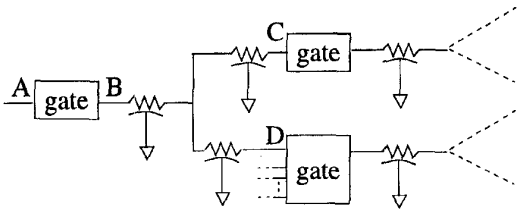


Fig. 1. A typical gate-delay problem.

expressions for delay, output-signal transition time and short-circuit power dissipation as a function of load capacitance and input-signal transition time (k-factor equations), and 2) a switch-resistor model comprised of a linear resistor and a step function of voltage. Both methods are empirically-based, since even the second method requires empirical fitting to approximate the resistance value.

When the load is purely capacitive, one can completely precharacterize a gate's delay and output signal behavior as a function of input signal transition time t_{in} and load capacitance C_L [1]. The experimental data for the delay t_d and the gate-output waveform transition time t_r or t_f are generally fitted to k-factor equations

$$t_d = k(t_{\text{in}}, C_L) \quad \text{and} \quad t_{r/f} = k'(t_{\text{in}}, C_L). \tag{1}$$

The power dissipation can be empirically precharacterized in a similar manner. For static CMOS circuits, it is recognized [6] that the power consumption of a gate is mainly a dynamic process. Low voltage designs with low threshold voltages may have significant subthreshold currents which make the standby power dissipation relatively sizable [7]. However, since this component is independent of output loading, it will not be considered here.

For our purposes, we consider the power dissipation as comprised of two basic components

- the switching component

$$P_{\text{sw}} = \frac{1}{2} C_L V_{\text{DD}}^2 f = E_{\text{st}} f \tag{2}$$

- the short-circuit component

$$P_{\text{sc}} = E_{\text{sc}} f = \left(V_{\text{DD}} \int_0^{\infty} i_{\text{sc}}(t)\, dt \right) f = g(t_{\text{in}}, C_L) f \tag{3}$$

where f is the frequency at which the gate switches, $i_{\text{sc}}(t)$ is the current that flows through the n- and p-channel transistors during the short time period when both are on, E_{st} is the energy stored in the capacitive elements, and E_{sc} is the energy dissipated by the short-circuit current for one transition at the gate input. The charging/discharging component of power dissipation from (2) is easily calculated in terms of the total capacitance load C_L. However, the short circuit power dissipation must be empirically characterized in terms of k-factor type equations similar to those in (1) [8].

We should point out, however, that there is sometimes a charging/discharging component of power dissipation associated with the short-circuit power empirical equations. That is, the precharacterization tools do not generally extract the gate output parasitic capacitance and add it to the load capacitance.

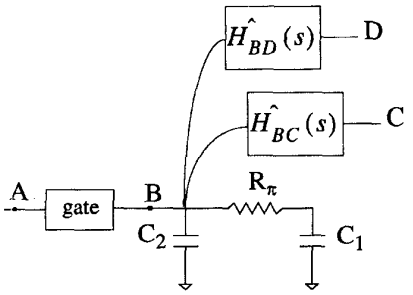


Fig. 2. Reduced order models for the driving point admittance (second-order) and the transfer functions of the RC interconnect, Fig. 1.

For delay characterization, this capacitance is included as part of the intrinsic gate delay in the k-factors, which is an accurate depiction. So for power analysis we must add this dissipation to the short-circuit empirical equation, or we must extract this device capacitance which appears at the gate output and add it to the total load capacitance when evaluating (2).

As shown in Fig. 1, however, the gates are loaded by RC circuits for today's technologies. Referring to this figure, one approach has been to capture the "gate delay" from A to B and the output signal rise or fall time at B using the total load capacitance (the sum of all of the interconnect capacitance and the gate input capacitances at C and D) and the signal transition time at A in (1). The RC delay would then be analyzed separately using the waveform at B, which is characterized by the result from (1).

This two-step delay approximation works well when the load "seen by the gate" is accurately approximated by the total capacitance of the net. That is, the procedure described for Fig. 1 assumes that the *driving point admittance* of the interconnect is equal to the total capacitance. It has been recognized, however, that this is an invalid assumption for today's high speed CMOS circuits [4]. This was first recognized for ECL gates in [2] where the second-order driving point admittance, modeled as a π-circuit (as shown in Fig. 2), was offered as a better approximation.

In [9], it was shown that while these second-order input admittance models are often accurate enough for CMOS technologies, nth order driving point models can be constructed for extremely large RC loads using a moment matching technique such as Asymptotic Waveform Evaluation (AWE) [10]. These reduced order input admittance models are generated at the same time as the approximate transfer function models for the responses for the fanout nodes, as shown in Fig. 2.

Evaluating the complete delay in Fig. 2 is predicated on the ability to analyze the gate delay when driving an RC load, which is minimally the π-load shown. One obvious approach would be to precharacterize the gates in terms of k-factor equations which are a function of the π-circuit components C_2, R_π, and C_1. Even for a π-load, this would be prohibitively expensive since it would increase what was a 2-D delay table into a 4-D table. But perhaps more importantly, unlike for capacitance loads, it is practically impossible to predict the range of interest for the π-circuit component values. Of course, there is no way of handling instances when higher-order driving point approximations are required.

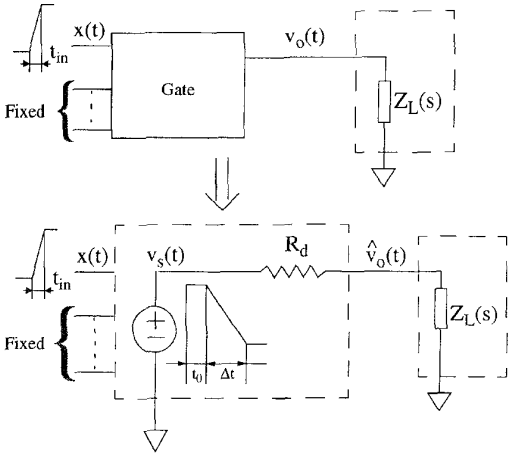


Fig. 3. The proposed model for CMOS gate model.

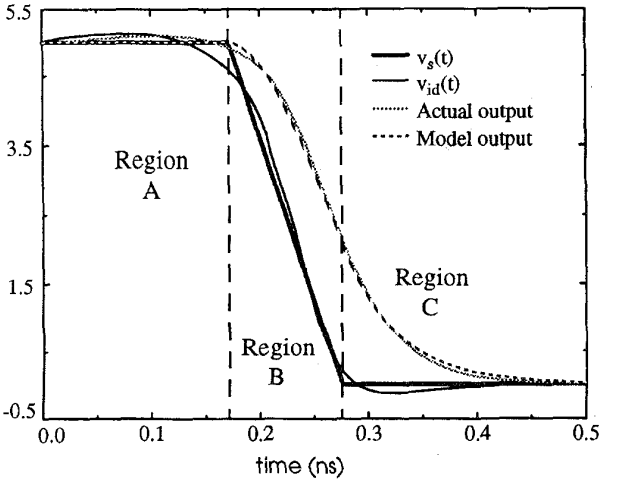


Fig. 4. Results obtained for an inverter with a 0.6 pF load in a 3μm CMOS technology, emphasizing the model regions of operation.

It might appear that the switch-resistor model is a more effective delay model when the load is not purely capacitive. That is, the resistance model is able to capture the interaction of the gate's output resistance and the RC load. Timing analysis tools such as TV [11] and Crystal [12] were developed using switch-resistor models to analyze the transistor level circuit descriptions. The main difficulty with these approaches is calculating a single linear resistor which captures the switching behavior of a CMOS gate. Recognizing that this resistance is a function of the gate's input signal transition time and output load, in [12] a single output resistance for the gate is empirically derived. That is, the resistance is calculated as the average output impedance over a range of input signal transition times and output loads. Therefore, since this resistance value must be characterized as a function of output load capacitance seen by the gate, it has the same limitations as the k-factor model.

To provide compatibility between these empirical gate level models and second-order driving point admittance models, the concept of an effective capacitance was introduced in [3]. While the C_{eff} model in [3] could capture the gate delay, it did not capture the output waveform which is necessary to capture the RC propagation delay to the fan-out nodes using the approximation transfer functions in Fig. 2. So this C_{eff} model was extended in [4] so as to capture the gate output waveshape in terms of a two-piece waveform approximation.

While this two-piece waveform approximation is reasonably accurate, it complicates the calculation of the delay and the rise/fall time at the fan-out nodes as compared to a single Thevenin equivalent driver model. Moreover, the regions of this C_{eff} model in [4] are distinguished by the completion of the gate's input signal waveform. However, for complex cells which have intrinsic delays comprised of multiple gates, it is not possible to distinguish the regions of the model in [4] by monitoring the completion time of the input signal to the gate. The model proposed in [13] was toward a more general CMOS gate level model which works for all size cells and k-factors, and is readily compatible with the transfer function approximations used to calculate the RC interconnect propagation delays. The following describes an extension of the model in [13].

III. A Generalized CMOS Gate Delay Model

The proposed CMOS gate-delay model is a time-varying Thevenin equivalent model as shown in Fig. 3. Consider $Z_L(s)$ as some nth order driving point impedance model for the RC interconnect circuit, while a ramp-like gate input signal is assumed. The linear resistor R_d can be selected independent of the load and input waveform or updated based on the actual gate load and input transition time. Obtaining the value(s) for R_d will be described in Section IV.

With some value for R_d, we search for the Thevenin voltage which will produce an accurate gate-output waveform and delay. To begin, we note that the transfer function of the model in Fig. 3 is

$$H(s) = \frac{\hat{V}_o(s)}{V_s(s)} = \frac{Z_L(s)}{Z_L(s) + R_d}. \tag{4}$$

The ideal waveshape for this voltage source model $v_{\text{id}}(t)$ would be the one which produces the same output waveform as the actual gate. This waveshape can be observed by applying the actual gate output waveform $v_o(t)$ to the inverse transfer function

$$H^{-1}(s) = 1 + \frac{R_d}{Z_L(s)}. \tag{5}$$

As an example, Fig. 4 shows the actual response output and the corresponding $v_{\text{id}}(t)$ waveshape for a 3 μm CMOS technology inverter driving a 0.6 pF capacitance load. Also shown in Fig. 4 is a *saturated* ramp approximation $v_s(t)$ for $v_{\text{id}}(t)$ and the response waveform (labeled "Model output") that it produces. For this example, it is apparent that the $v_{\text{id}}(t)$ waveshape can be accurately modeled by a simple, linearized waveshape. Unfortunately, for deep submicron CMOS technologies we have found that this linear approximation for $v_{\text{id}}(t)$ is somewhat less accurate, as shown in Fig. 5.

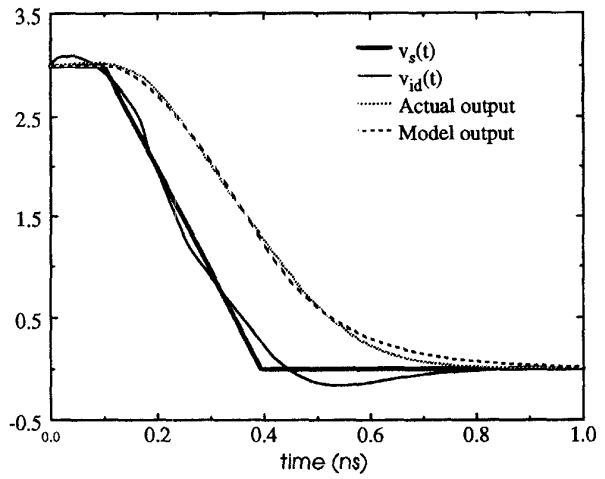


Fig. 5. Results obtained for an inverter with a 0.6 pF load in 0.8 μm MOSIS technology.

While it is possible to model $v_{\rm id}(t)$ by some other (nonlinear or piecewise linear) function, it will be at the expense of modeling simplicity. Therefore, we keep the model simple and determine the saturated ramp approximation for $v_{\rm id}(t)$ which guarantees a match with the 20% and 50% points of the actual gate response. This is the definition we are giving for the model voltage source (MVS) shown in Fig. 3 as $v_s(t)$. These two particular time points are chosen because of the importance of the 50% delay and the need of another point from Region B (as defined in Fig. 4). With such a fit for the example in Fig. 5, it can be seen that even though $v_s(t)$ does not appear to be an ideal match for $v_{\rm id}(t)$, the output model waveform is practically as good as that from Fig. 4, and the 50% delay matches exactly. The definitions of the MVS parameters t_0 and Δt are given in Fig. 6.

Since the 20% point cannot be determined from k-factor equations like those in (1), and for reasons that will become clear later in this paper, we are using a gate precharacterization for three time points from the output waveform, all measured with respect to the mid point of the input signal

$$\begin{aligned} t_{20} &= k(t_{\rm in}, C_L) \\ t_{50} &= k'(t_{\rm in}, C_L) \\ t_{90} &= k''(t_{\rm in}, C_L). \end{aligned} \tag{6}$$

To define the equations for fitting the 20% and 50% points of the actual gate response to those of the model response, $\hat{v}_o(t)$ in Fig. 3, we first observe that $\hat{v}_o(t)$ is a saturated ramp response due to the input voltage $v_s(t)$. Therefore, we can express this output voltage as an analytical function of the parameters of $v_s(t)$, namely t_0 and Δt

$$\hat{v}_o(t) \equiv y(t, t_0, \Delta t). \tag{7}$$

Specifically, for a falling output transition

$$y(t, t_0, \Delta t) = \begin{cases} V_{\rm DD} & t - t_0 < 0 \\ V_{\rm DD} - y_0(t, t_0, \Delta t) & 0 \le t - t_0 < \Delta t \\ V_{\rm DD} - y_0(t, t_0, \Delta t) + y_0(t - \Delta t, t_0, \Delta t) & t - t_0 \ge \Delta t \end{cases} \tag{8}$$

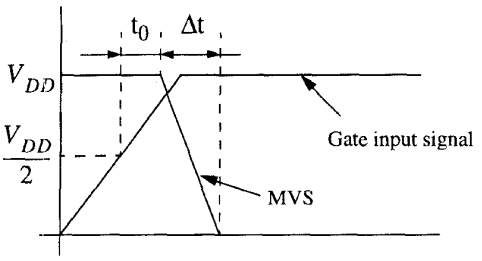


Fig. 6. MVS parameter definition.

and for a rising output transition

$$y(t, t_0, \Delta t) = \begin{cases} 0 & \text{if } t - t_0 < 0 \\ y_0(t, t_0, \Delta t) & \text{if } 0 \le t - t_0 < \Delta t \\ y_0(t, t_0, \Delta t) - y_0(t - \Delta t, t_0, \Delta t) & \text{if } t - t_0 \ge \Delta t. \end{cases} \tag{9}$$

Where $y_0(t, t_0, \Delta t)$ is the single positive ramp response for a capacitive load, C_L

$$y_0(t, t_0, \Delta t) = V_{\rm DD} \cdot \left[\frac{t - t_0}{\Delta t} - \frac{R_d C_L}{\Delta t} \cdot \left(1 - e^{-\frac{t - t_0}{R_d C_L}} \right) \right]. \tag{10}$$

To solve for the t_0 and Δt which yield an output waveform that matches the 20% and 50% points of the actual response, we solve the following system of equations via Newton–Raphson iteration

$$\begin{cases} y(t_{50}(C_{\rm load}), t_0, \Delta t) = 0.5 \cdot V_{\rm DD} \\ y(t_{20}(C_{\rm load}), t_0, \Delta t) = \begin{cases} 0.2 \cdot V_{\rm DD} & \text{for a rising transition} \\ 0.8 \cdot V_{\rm DD} & \text{for a falling transition.} \end{cases} \end{cases} \tag{11}$$

Where $t_{50}(C_{\rm load})$ and $t_{20}(C_{\rm load})$ are computed from the k-factors in (6) (for a specified input transition time). This would result in the MVS parameters for $v_s(t)$, which could be used to determine the responses at the fan-out nodes using RC circuit transfer function approximations.

More importantly, when the load is not purely capacitive, we can still write an analytical expression for $y(t, t_0, \Delta t)$. For example, given a π-load, $y(t, t_0, \Delta t)$ is the saturated ramp response solution of an RC–RC circuit for an input voltage $v_s(t)$. For an nth order driving point impedance model, we can also obtain an analytical expression for $y(t, t_0, \Delta t)$ using AWE. But since the k-factors are not defined for noncapacitive loads, we must resort to an *effective capacitance* type of approximation to evaluate t_0 and Δt using (11).

This effective capacitance $C_{\rm eff}$ would be similar to that defined in [3] and [4]. In this paper, we maintain the principle of equating the average current drawn by the effective capacitance and the driving point impedance model. A second-order driving point impedance model, a π-circuit load, is illustrated in Fig. 7. The current averaging is for a finite period of time $t_{\rm av}$ which we will refer to as the *active region*. The determination of this active region is critical to an accurate approximation. Moreover, selection of this active region will be based on whether we are interested in determining the delay or the short-circuit power—more on this later.

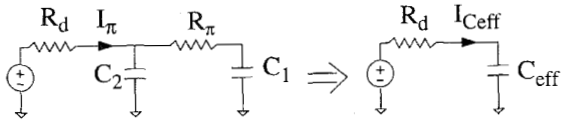


Fig. 7. The averaging of the load currents.

We will describe the complete procedure for calculating the effective capacitance and the voltage source parameters for the case of a second-order driving point impedance model. All of these steps are extendible to an nth order driving point model as shown in Appendix A.

To compute the MVS parameters and the effective capacitance simultaneously, we solve the following set of equations

$$\begin{cases} I_\pi(\Delta t, t_{\rm av}) = I_{C_{\rm eff}}(\Delta t, t_{\rm av}, C_{\rm eff}) \\ y(t_{50}(C_{\rm eff}), t_0, \Delta t) = 0.5 \cdot V_{\rm DD} \\ y(t_{20}(C_{\rm eff}), t_0, \Delta t) = \begin{cases} 0.2 \cdot V_{\rm DD} & \text{if rising transition} \\ 0.8 \cdot V_{\rm DD} & \text{if falling transition.} \end{cases} \end{cases} \tag{12}$$

The first equation in (12) matches the average curents in the π-circuit and $C_{\rm eff}$, as shown in Fig. 7. Where the average currents for the two are

$$I_\pi(\Delta t, t_{\rm av}) = \left[A t_{\rm av} + \frac{B}{p_1}\left(1 - e^{-p_1 t_{\rm av}}\right) + \frac{D}{p_2}\left(1 - e^{-p_2 t_{\rm av}}\right) \right] \frac{V_{\rm DD}}{R_d t_{\rm av} \Delta t} \tag{13}$$

and

$$I_{C_{\rm eff}}(\Delta t, t_{\rm av}, C_{\rm eff}) = \left[R_d C_{\rm eff} t_{\rm av} - (R_d C_{\rm eff})^2 \left(1 - e^{-\frac{t_{\rm av}}{R_d \cdot C_{\rm eff}}}\right) \right] \frac{V_{\rm DD}}{R_d t_{\rm av} \Delta t} \tag{14}$$

with

$$A = \frac{z}{p_1 p_2} \quad B = \frac{z - p_1}{p_1(p_1 - p_2)} \quad D = \frac{z - p_2}{p_2(p_2 - p_1)}. \tag{15}$$

The z, p_1, p_2 parameters represent the zero and the two poles of the transfer function $\frac{V_o(s) - V_s(s)}{V_s(s)}$ for the circuit defined in Fig. 3, Z_L being a π-circuit. The t_{50} and t_{20} terms are expressed only as a function of the load capacitance ($C_{\rm eff}$ in this case) because the gate input transition time is a known (determined by analysis of the prior logic stage).

We define the active region to be Region B from Fig. 4. This is the time when the MVS is in transition, which represents the natural time period for averaging the currents (replace $t_{\rm av}$ in (12) by Δt). Region C from Fig. 4 will represents the time period for which the gate behaves like a resistor, which is modeled by R_d.

We apply the following initial guesses for $C_{\rm eff}$ and the MVS parameters

$$C_{\rm eff} = \min(C_1 + C_2, C_{\max}) \tag{16}$$

$$\begin{cases} \Delta t = \frac{10}{3} \cdot (t_{50}(C_{\rm eff}) - t_{20}(C_{\rm eff})) \\ t_0 = t_{50}(C_{\rm eff}) - 0.69 \cdot R_d \cdot C_{\rm eff} - \frac{\Delta t}{2} \end{cases} \tag{17}$$

Note that $C_{\max}$ is the maximum capacitance for which the k-factor equations are valid (this limit is generally specified by the empirical precharacterization process).

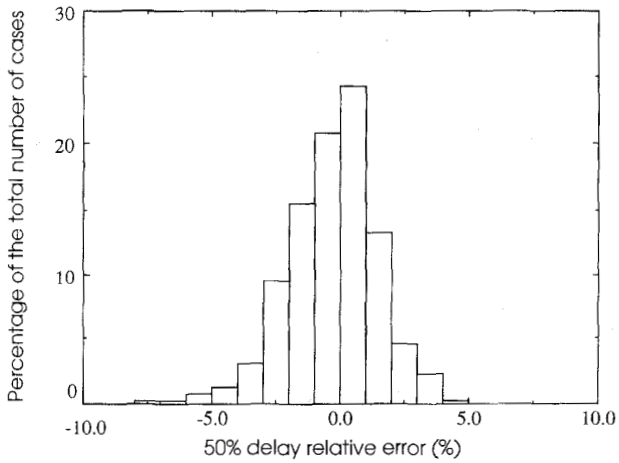


Fig. 8. Histogram for the 50% delay relative errors obtained using the basic model for an inverter in 0.8 μm MOSIS technology, having $l = 0.8$ μm, $w_N = 24$ μm, and $w_P = 60$ μm and different loads.

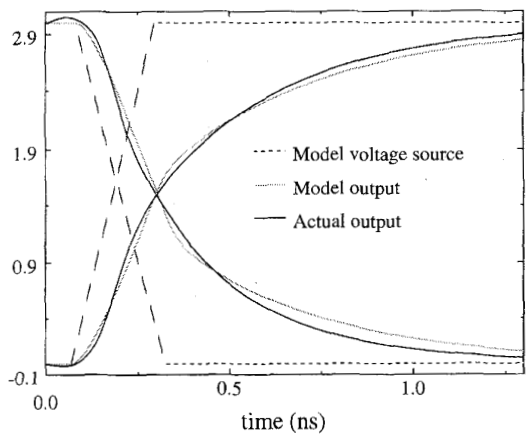


Fig. 9. Typical results for experiments included in Fig. 8 statistic. In this example, $C_1 = 0.71$ pF, $C_2 = 0.24$ pF, $R_p = 312$ Ω, $t_{\rm in} = 0.2$ ns.

Using these initial guesses, we evaluate (12) via Newton–Raphson iteration. This nonlinear problem is well behaved and, therefore, convergence is extremely fast. For example, from a set of 10 000 examples solved, 98% required four or less iterations, for an 0.1% tolerance specification.

From the accuracy point of view, Fig. 8 presents a histogram of the t_{50} relative errors for this model given a set of 400 gate-delay problems with various π-loads as compared with HSPICE.[1] The mean error for the k-factor equations used in this example was 1.3%. One fairly typical driving point waveform approximation from this sample of 400 gate-delay problems is shown in Fig. 9.

A. Precharacterizing the Voltage Source Parameters

Because we precharacterize gate models which are used hundreds of thousands of times, it is preferable to spend more time in the precharacterization process for the k-factors if the solution of (12) (that is solved for each gate in the circuit)

[1] HSPICE is a Registered Trademark of Meta-Software, Inc.

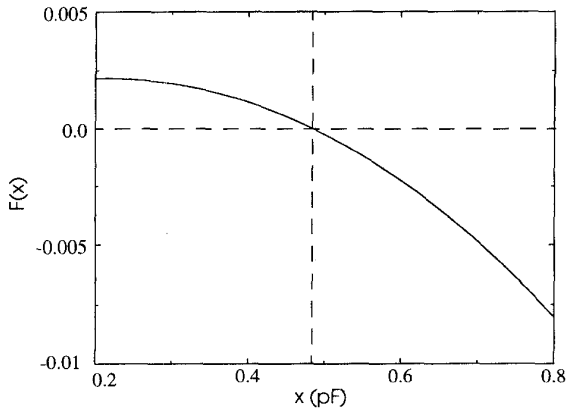


Fig. 10. The typical form of $F(x)$. In this case, $C_2 = 0.2$ pF, $C_1 = 0.6$ pF, $R_d = 250\,\Omega$, $R_p = 200\,\Omega$, and the gate used for the results in Fig. 8.

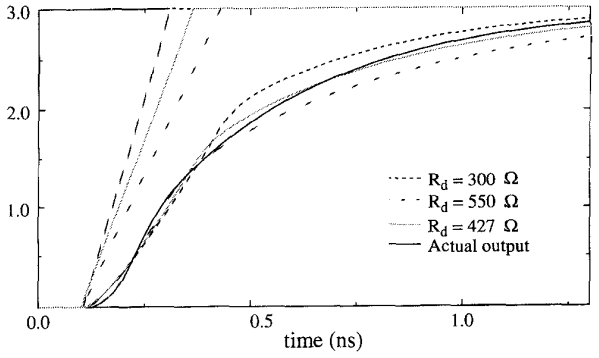


Fig. 11. Results obtained for a NAND gate driving a π-load with $C_1 = 0.5$ pF, $C_2 = 0.2$ pF, and $R_p = 400\,\Omega$ when the model is precharacterized using different resistance values.

can be simplified. For this reason, we propose to replace or to augment the delays and output transition time k-factors with similar k-factors for t_0 and Δt of the MVS

$$\begin{aligned} t_0 &= f(t_{\text{in}}, C_L) \\ \Delta t &= g(t_{\text{in}}, C_L). \end{aligned} \tag{18}$$

That is, for each pair of values for C_L and t_{in}, we have a unique Δt and t_0 for the voltage source model. Therefore, to reduce the overall runtime costs of this model which require Newton–Raphson iterations to solve for t_0 and Δt, we instead calculate these parameters during precharacterization and store them as k-factors such as (18). Using (18) would simplify (12) from a system of three nonlinear equations in three unknowns to a single nonlinear equation in one unknown

$$F(C_{\text{eff}}) = 0 \tag{19}$$

or

$$\begin{aligned} F(C_{\text{eff}}) = A\Delta t(C_{\text{eff}}) &+ \frac{B}{p_1}\left(1 - e^{-p_1 \Delta t(C_{\text{eff}})}\right) \\ &+ \frac{D}{p_2}\left(1 - e^{-p_2 \Delta t(C_{\text{eff}})}\right) - R_d C_{\text{eff}} \Delta t(C_{\text{eff}}) \\ &+ (R_d C_{\text{eff}})^2 \left(1 - e^{-\frac{\Delta t(C_{\text{eff}})}{R_d \cdot C_{\text{eff}}}}\right). \end{aligned} \tag{20}$$

In Fig. 10, we display the general shape of the function $F(x)$ for $x \in (C_2, C_1 + C_2)$. Using Newton–Raphson and an initial guess $(C_1 + C_2)$, it is apparent from Fig. 10 that the solution is approached from a capacitance value greater than the final C_{eff} value. That is, each C_{eff} solution is greater than the exact value, thereby assuring a "pessimistic" C_{eff} [4].

IV. Computation of the Driver Output Resistance

In the model presented in Fig. 3, we model the last operating region (Region C in Fig. 4) as a linear resistor connected to ground (for a falling transition) or V_{DD} (for a rising transition). This is a refinement over switch-level simulators that use a linear resistor to model the gate over all operating regions. Our solution is more accurate due to the reduced swing in the output voltage over which we linearize the output resistance.

The basic choice to be made is whether the resistor value is unique for a gate or if it should depend upon the input signal and the effective load. Since each case has its own advantages we present both scenarios here.

A. Single R_d Value

We can compute R_d during the k-factor generation process based on the largest possible load capacitance $C_{\max}$ that can be encountered by the driver in an actual circuit (as discussed in [13]). This computation, which is performed only once for a gate, allows for a more sophisticated approach employing least-mean square and iterative adjustment methods [13].

We should point out that the delay approximation for the driving point is fairly insensitive to the value of R_d, as shown in Fig. 11. Using a larger or a smaller R_d will mainly impact the tail portion of the waveform (Region C).

Observing that until now we have used only two out of three functions from (6) and a single R_d value to characterize a gate, it is apparent that the main advantage of a single resistance model is its simplicity. For timing analysis we would need only two k-factor equations (as before), a resistor value, and a procedure to map a complex load to a single capacitance value.

B. Using Multiple Values of R_d

It is possible to simplify the selection of the resistance value by an exponential fit of the output waveform in two points. We presume the output waveshape to be of the form

$$v(t) = V_0 e^{-\frac{t}{R_d C_L}}. \tag{21}$$

The first fitting point is obviously the 50% time point. For the second time point we must choose a point near the end of the waveform transition. Experimentally, we have found the 90% point to be the most suitable. We have also observed that taking a point closer to the mid point tends to overestimate the resistor value. Since the ideal time-point value for the gate resistance estimation may be technology dependent, running a couple of experiments before the actual precharacterization will help. Therefore, in our case

$$R_d = \frac{t_{90}(C_L, t_{\text{in}}) - t_{50}(C_L, t_{\text{in}})}{C_L \cdot \ln 5}. \tag{22}$$

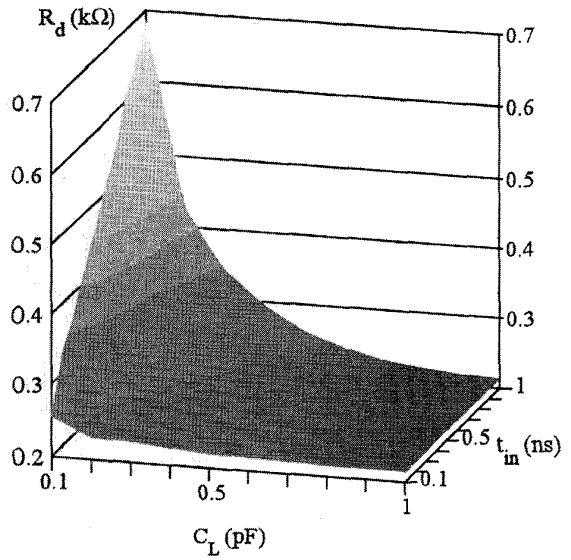


Fig. 12. R_d computed with (22) as a function of C_L and input transition time.

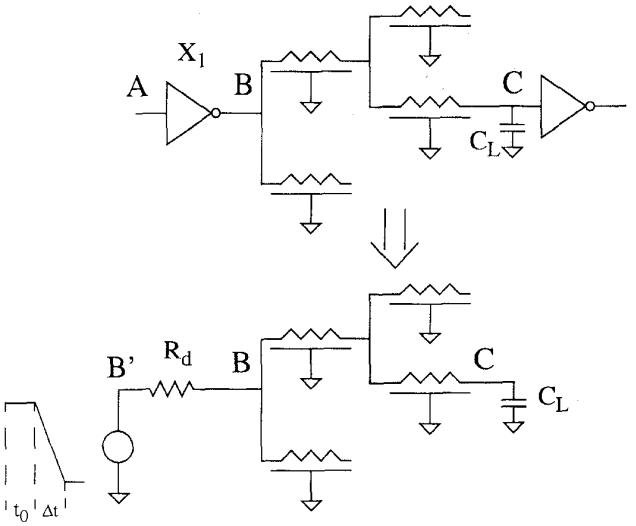


Fig. 13. A typical problem for a static timing analyzer.

One can apply this simplified computation of R_d for a range of capacitive loads and input transition times. The results are shown for an 0.8 μm inverter as a 3-D plot in Fig. 12.

Modeling the output resistance with a single value is obviously inaccurate, as shown from Fig. 12. However, we can see from Fig. 11 that the response delay at the driving point is fairly insensitive to this variation in resistance. Of course, if extreme accuracy is required, we can use the information from Fig. 12 in one of two ways.

- Use the traditional k-factors from (6), and replace any instance of R_d from (12) with the calculated value from (22).
- Or, store (18) and fit $R_d(C_L, t_{\text{in}})$ by a polynomial approximation (like we did the k-factors). Unfortunately, the shape of the resistance function (as presented in (12)) requires a regionwise polynomial type of fit, and a single low order polynomial is insufficient.

C. The Effect of R_d on the Fanout Point Accuracy

It may appear that a method that computes a capacitance value that yields the same gate delay as the actual complex load is the perfect answer to our timing problems. This is not true if we consider static timing analysis in the presence of the interconnect-dominated circuits, as presented in Fig. 13. The accurate A-B delay approximation may be only a small fraction of the overall A-C delay. To capture the B-C portion of the delay accurately requires an accurate waveform approximation at B. For example, for the case in Fig. 11, all three approximations of the actual waveform are good from a driving-point delay point of view (less than 3% errors). However, at the fan-out point the percentage errors are twice as large!

Observing that the only element varying in these three experiments is the driver resistance value, this suggests a very strong reason for updating the driver resistance value based upon the input transition time and output effective load. Though this requires a new k-factor equation (for the 90% time point), this is the only way for obtaining the required accuracy at the fan-out points.

Given the resistance value and the MVS parameters, we can evaluate the fanout delays and waveshape responses using the approximate transfer functions (refer back to Fig. 2). Since the resistance value may be varied during analyzes, we assume that we have the approximate transfer function between points B and C, $H_{BC}(s)$, in terms of the n dominant poles and residues [3]. But given the MVS parameters and R_d, we really want the approximate transfer function between points B' and C, $H_{B'C}(s)$, where B' is the dummy node at which we connect $v_s(t)$. So we can directly calculate $H_{B'C}(s)$ by convolving $H_{B'B}(s)$ and $H_{BC}(s)$ which introduces k new poles in the transfer function when the driving point admittance is given as a kth order model, and modifies the residues as described in [15].

V. Calculating the Power Consumption

As mentioned previously, the charging/discharging component of power dissipation is evaluated using the total net capacitance and (2). The interconnect resistance only serves to distribute the thermal dissipation among the interconnect segments and the gate resistance. But for the short-circuit component of power dissipation, (3), the load "seen by the gate" affects the regions of operations of the transistors, thereby affecting the power dissipated (this was implied in [6]). Therefore, the resistance of the metal which impacts the effective capacitance, will also affect the short-circuit component of power dissipation.

The empirical calculation and subsequent analysis of the short-circuit power dissipation is identical to that for the delay computation. To illustrate the load capacitance dependence, we experimentally measured the short-circuit power consumption of a buffer as a function of C_L for three different input signal transition times. The HSPICE results are shown in Fig. 14. To

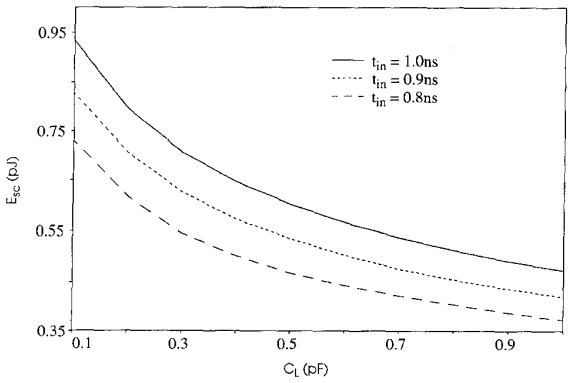


Fig. 14. Variation of the short-circuit power consumption component with the load capacitance for a buffer built in 0.8 μm MOSIS technology with $V_{\rm DD} = 3$ V.

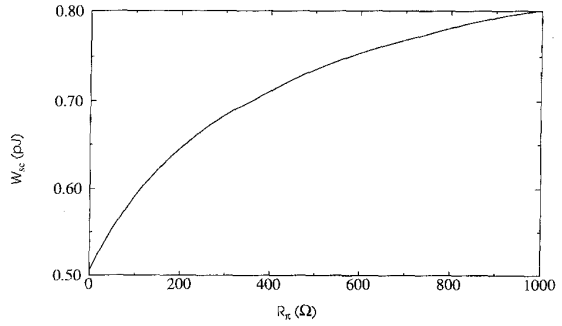


Fig. 15. Variation of the short-circuit power consumption component with the shielding effect (the load is modeled as a π-circuit with $C_2 = 0.1$ pF, $C_1 = 0.8$ pF, $t_{\rm in} = 1$ ns, while the gate is the same with the one used for Fig. 8).

demonstrate the $C_{\rm eff}$ dependance, we measured the variation in short-circuit power for a constant total capacitance load while the π-load resistance is increased. These HSPICE results are presented in Fig. 15.

It can be easily shown that the short-circuit component for a gate driving a complex load modeled as π-circuit, is bounded by the values obtained for the same gate driving a load capacitance equal to C_2 and $C_{\rm tot}$, respectively. Unfortunately, $C_{\rm tot}$, that is usually taken as a conservative approximation of the RC interconnect, offers an important underestimation of the dissipated power. At the same time, C_2 may highly overestimate the power consumption with the plots of Fig. 14 proving to be more abrupt at lower load capacitances.

Before presenting our solution, let us gain some insight into the problem through a simplified example. Consider the case of a single rising transition for a buffer (the case of an falling transition is equivalent). We have

$$E_{\rm sc} = \int_0^{t_W=\infty} i_{\rm sc}(t) v_{\rm out}(t)\, dt \tag{23}$$

where $i_{\rm sc}(t)$ is the short-circuit current as defined in (3) and $v_{\rm out}(t)$ is the output voltage.

Considering that $i_{\rm sc}(t)$ is a function of the gate input voltage $v_{\rm in}(t)$ and $v_{\rm out}(t)$, then a load capacitance C_W approximates a π-load from the $E_{\rm sc}$ point of view if and only if the gate output voltages are close to each other for the t_W period of time. Obviously, this result is not useful if we consider an infinite t_W (as theoretically required). That is, there is no C_W which would provide an accurate voltage and current waveform match to a π-load over an infinite t_W.

Fortunately, $i_{\rm sc}(t)$ is nonzero (or better said, nonnegligible) only for a finite period of time, i.e., when $V_{TN} \le v_{\rm in}(t) \le V_{\rm DD} - |V_{TP}|$. Hence, from an $E_{\rm sc}$ estimation point of view, we are interested in $v_{\rm out}(t)$ only up to the point when $v_{\rm in}(t) = V_{TN}$, i.e.,

$$t_W = t_{\rm in} \frac{V_{\rm DD} - V_{TN}}{V_{\rm DD}}. \tag{24}$$

Due to the significant subthreshold currents for today's technologies, V_{TN} and V_{TP} are not the classical intrinsic threshold voltages for the n- and p-channel transistors, but rather indicators of some predetermined level of drain-to-source current flow for the transistor. Selecting a lower value for V_{TN} (implying a bigger value for t_W) will increase the accuracy in computing the integral from (23), but it will decrease the accuracy for $v_{\rm out}(t)$. For this reason V_{TN} should be experimentally determined for optimum performance.

For practical considerations, it is preferable to determine $\frac{t_W}{t_{\rm in}}$ for both transition types in place of V_{TN} and V_{TP}. Considering a typical RC load with a significant resistive component, we have to determine the short-circuit power dissipation for the precharacterized gate and find the value of load capacitance C_L that will cause the same short-circuit power dissipation. Then, we plug C_L in (12) and solve for $t_{\rm av}$ observing from the waveform definitions that

$$t_{\rm av} = t_W - t_0 - \frac{t_{\rm in}}{2}. \tag{25}$$

The quantities in (25) have the same meaning as previously defined. The procedure may be repeated for different loads in order to gain confidence in the solution.

Once we determine the ratio $\frac{t_W}{t_{\rm in}}$, it is possible to use one of the procedures presented in Section III for effective capacitance computation with $t_{\rm av}$ replaced by the value in (25). In Fig. 16, statistical results for the same gate as in Fig. 8 are presented. For this gate, $V_{TN} = 0.69$ V and $V_{TP} = -1.11$ V (while the HSPICE parameter V_{T0} is 0.73 V and -0.97 V, respectively).

For this experiment, we used k-factor characterization for the short circuit power dissipation (containing up to the quadratic terms in C_L and $t_{\rm in}$) with 5.3% mean error. The higher errors for the power case are the combined result of less accurate k-factor equations and the requirement of conservative approximation (that will overestimate the actual results). It is important to note that the effective capacitance and the model parameters values differ for the power estimation and delay approximation cases.

VI. Examples

We further demonstrate the effectiveness of this CMOS gate delay model on the following examples.

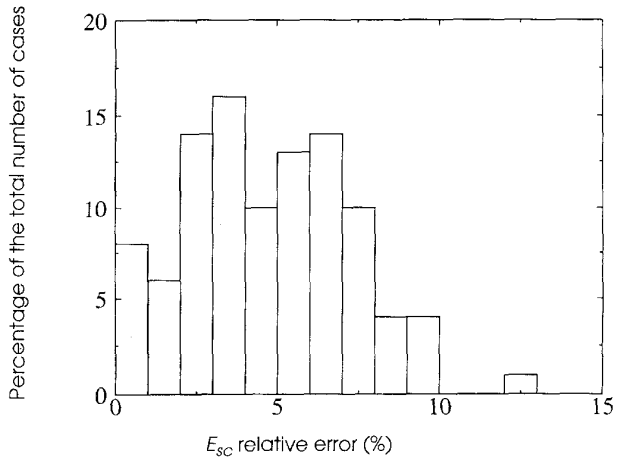


Fig. 16. Histogram for the short-circuit power consumption relative errors for the same gate as in Fig. 8. In this case, the k-factor equations presented an average error of 5.3%.

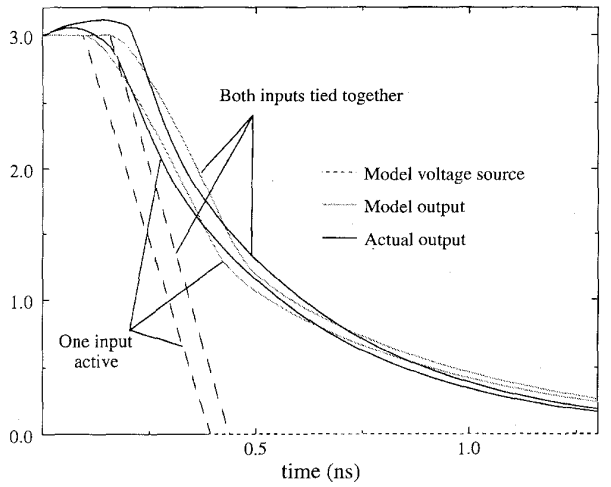


Fig. 18. Results obtained for a NAND gate driving a π-load with $C_1 = 0.7$ pF, $C_2 = 0.15$ pF, and $R_p = 150\,\Omega$.

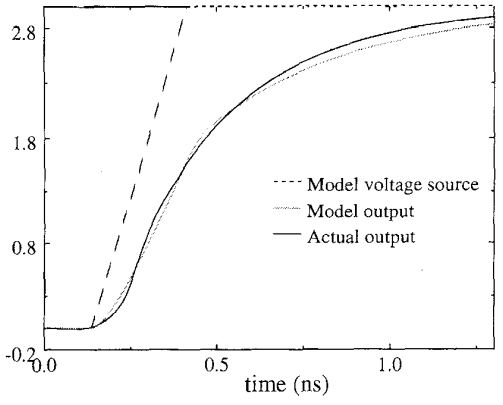


Fig. 17. Results obtained for a D-latch driving a π-load with $C_1 = 0.2$ pF, $C_2 = 0.6$ pF, and $R = 350\,\Omega$ (the presented output is $\overline{Q}$).

A. General CMOS Gates and Cells

There are applications that require empirical models for complex cells (e.g., flip-flops, latches, etc.) and/or multiple inputs gates (e.g., n-inputs NAND gates). Fig. 17 shows the response given by our model as compared with HSPICE for a D-latch. The waveforms depict the output $\bar{Q}$ for a falling edge on D. Three C_{eff} iterations were required.

This is one of the main advantages of this model over the methods presented in [3] and [4]. The C_{eff} obtained with their procedures depends on the input transition time while complex cells have output waveforms practically independent of the input signal parameters. The generalized model presented here overcomes this problem and offers a good approximation of the output signal.

B. Multiple Input Switching Cells

Accurate modeling of simultaneous switching is extremely difficult due to the ambiguity of the signal waveshapes and arrival times. With the model proposed here, however, we can easily bound the simultaneous switching response on both sides, and demonstrate the sensitivity of the gate delay to the number of inputs switching.

As an example, an empirical model was developed (following the steps outlined in Section III) for a two-input NAND gate with both inputs tied together and for one input switching alone (with the other input set to the proper sensitizing value). The voltage source waveshapes and the response waveforms as they compare with HSPICE are shown in Fig. 18.

Both models use the same pull-down resistance, but notice that for a falling transition, the voltage source delay is smaller and the transition time is shorter for the single input transition as compared to the same input signal arriving at both inputs simultaneously. For the arrival of two signals with different waveshapes, the voltage source model for the two inputs tied together is a worst case response if the slower of the two edges is used in the empirical equation. Similarly, the single input transition response with the slower of the two transition times is an optimistic prediction for the gate delay.

The models for multiple input switching of large cells works well when the two bounding voltage-source waveforms are relatively close. However, cells with a large number of inputs can have extremely loose bounds for the earliest and latest possible waveforms. For such cases, the model accuracy will be inconsequential compared to the error due to input-signal uncertainity.

VII. Future Work and Conclusion

We have presented a gate/cell-level delay and power modeling approach that works well for the highly-resistive interconnect loads present in today's CMOS circuits. We observe that this approach accurately captures the nondigital behavior of the output waveforms and the parameters for this model can be precharacterized efficiently. The linear nature of the gate model with the interconnect load allows for efficient computation of the waveforms, not only at the gate output but at the fanout points of the interconnect loads, too. Another advantage over previous models is the ability to model large

cells for which the output signal is independent of the input signal waveshape. Results presented for various examples (buffers, NAND gates, D-latches) demonstrate the accuracy and utility of our model.

This model provides the foundation for future work in several directions. First, we plan to provide more formal assertions regarding the bounds for simultaneous switching. Second, we plan to develop metrics and bounds to facilitate the use of this model with wiring estimates and loading estimates for high-level delay estimation.

APPENDIX

Given any RC circuit that models the driving point impedance of the RC interconnect, we can write an equation equivalent to (13) (i.e., the average current drawn by the circuit from the ramp source through R_d).

Approximating the combined driver resistance R_d and interconnect load admittance with n poles and their corresponding residues from AWE, [10]

$$Y(s) = \left(1 + \sum_{i=1}^{n} \frac{k_i}{s + p_i}\right) \cdot \frac{1}{R_d} \tag{26}$$

(13) becomes

$$I_{\text{av}} = -\frac{V_{\text{DD}}}{R_d t_{\text{av}} \Delta t}\left[t_{\text{av}} \sum_{i=1}^{n} \frac{k_i}{p_i^2} + \sum_{i=1}^{n} \frac{k_i}{p_i^3}(1 - e^{-p_i t_{\text{av}}})\right]. \tag{27}$$

The admittance formula in (26) is easily constructed using RICE, [14]. The π-circuit load in Fig. 2 represents a circuit synthesis of a two-pole model in (26).

REFERENCES

[1] N. H. E. Weste and K. Eshraghian, "Principles of CMOS VLSI Design," *Empirical Delay Models*, 2nd ed. Reading, MA: Addison-Wesley, 1992, pp. 213.

[2] P. R. O'Brien and T. L. Savarino, "Modeling the driving-point characteristic of resistive interconnect for accurate delay estimation," in *Proc. IEEE Int. Conf. Computer-Aided Design*, Nov. 1989, pp. 512–515.

[3] C. L. Ratzlaff, S. Pullela, and L. T. Pillage, "Modeling the RC-interconnect effects in a hierarchical timing analysis," in *Proc. IEEE Custom Integrated Circuits Conf.*, May 1992, pp. 15.6.1–15.6.4.

[4] J. Qian, S. Pullela, and L. T. Pillage, "Modeling the 'effective capacitance' of RC interconnect," *IEEE Trans. Computer-Aided Design*, vol. 13, pp. 1526–1535, Dec. 1994.

[5] R. E. Mains, T. A. Mosher, L. P. P. P. van Ginneken, and R. F. Damiano, "Timing verification and optimization for the PowerPC processor family," in *Proc. Int. Conf. Computer Design*, 1994, pp. 390–393.

[6] H. J. M. Veendrik, "Short-circuit dissipation of static CMOS circuitry and its impact on the design of buffer circuits," *IEEE J. Solid-State Circuits*, vol. SC-19, pp. 468–473, 1984.

[7] A. P. Chandrakasan, S. Sheng, and R. W. Brodersen, "Low-power CMOS digital design," *IEEE J. Solid-State Circuits*, vol. 27, pp. 473–484, 1992.

[8] B. J. George, G. Yeap, M. G. Wloka, S. C. Tyler, and D. Gossain, "Power analysis and characterization for semi-custom design," in *Proc. Int. Workshop Low Power Design*, 1994, pp. 215–218.

[9] N. Gopal, D. Neikirk, and L. T. Pillage, "Evaluating RC interconnect using moment matching approximations," in *Proc. IEEE Int. Conf. Computer-Aided Design*, Nov. 1991, pp. 741–777.

[10] L. T. Pillage and R. A. Rohrer, "Asymptotic waveform evaluation for timing analysis," *IEEE Trans. Computer-Aided Design*, vol. 9, pp. 352–366, 1990.

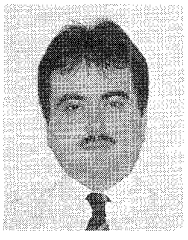
[11] N. Jouppi, "Timing analysis and performance improvement of MOS VLSI designs," *IEEE Trans. Computer-Aided Design*, vol. CAD-6, pp. 650–665, 1987.

[12] J. K. Ousterhout, "A switch-level timing verifier for digital MOS VLSI," *IEEE Trans. Computer-Aided Design*, vol. CAD-4, pp. 336–349, 1985.

[13] F. Dartu, N. Menezes, J. Qian, and L. T. Pillage, "A gate-delay model for high speed CMOS circuits," in *Proc. 31st ACM/IEEE Design Automation Conf.*, 1994, pp. 576–580.

[14] F. Dartu, "A gate-delay model for high-speed CMOS circuits," M.S. thesis, University of Texas, Austin, Dec. 1994.

[15] C. L. Ratzlaff and L. T. Pillage, "RICE: Rapid interconnect circuit evaluation using AWE," *IEEE Trans. Computer-Aided Design*, vol. 13, pp. 763–776, 1994.

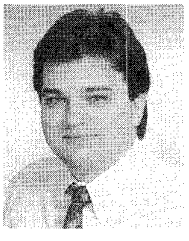


Florentin Dartu received the Diploma of Engineer degree in electrical engineering from Iasi Polytechnic Institute, Iasi, Romania, in 1989, and the M.S. in electrical engineering from the University of Texas, Austin.

He is currently with the Department of Electrical and Computer Engineering, Carnegie Mellon University, Pittsburgh, PA. His present research interests are in modeling and simulation of the mixed-mode systems.

Noel Menezes received the B.E. degree in electronics engineering from the Maharaja Sayajirao University, Baroda, India, in 1988, and the M.S.E. and Ph.D. degrees in electrical engineering from the University of Texas, Austin, in 1991 and 1995, respectively.

He is currently with the Strategic CAD Technology group at Intel Corporation, Hillsboro, OR. His current research interests are interconnect analysis, interconnect-based synthesis, and circuit and timing simulation.



Lawrence T. Pileggi (M'85–S'87–SM'94) (formerly Pillage) received the Ph.D. degree in electrical and computer engineering from Carnegie Mellon University, Pittsburgh, PA, in 1989.

He is an Associate Professor of electrical and computer engineering at Carnegie Mellon University. From 1984 to 1986, he was with Westinghouse Research and Development where he acquired two patents, and in 1986, he was recognized with the corporation's highest engineering achievement award. In 1989, he joined the faculty of the University of Texas, Austin, and in January of 1996, he joined the faculty of Carnegie Mellon University. His research interests include various aspects of circuit-level design analysis and automation. He is a coauthor of *Electronic Circuit and System Simulation Methods* (McGraw-Hill, 1995).

Dr. Pileggi received the IEEE TRANSACTIONS ON COMPUTER-AIDED DESIGN Best Paper Award for 1991, a Presidential Young Investigator Award from the National Science Foundation in 1991, a Semiconductor Research Corporation Technical Excellence Award in 1992, an Invention Award from the SRC in 1993 and subsequently a U.S. Patent for the RICE simulation tool, the University of Texas Parent's Association Centennial Teaching Fellowship for excellence in undergraduate instruction in 1994, and a Faculty Partnership Award from IBM in 1995.