Delay Models for CMOS Circuits

Grant McFarland and Michael Flynn

Technical Report CSL-TR-95-672

June 1995

This work was supported by the NSF under contract MIP93-13701 and by fellowship support from the IBM/CIS Fellow Mentor Advisor Program.

Delay Models for CMOS Circuits

by

Grant McFarland and Michael Flynn

Technical Report CSL-TR-95-672

June 1995

Computer Systems Laboratory

Departments of Electrical Engineering and Computer Science

Stanford University

Stanford, California 94305-4055

pubs@shasta.stanford.edu

Abstract

Four different CMOS inverter delay models are derived and compared. It is shown that inverter delay can be estimated with fair accuracy over a wide range of input rise times and loads as the sum of two terms, one proportional to the input rise time, and one proportional to the capacitive load. Methods for estimating device capacitance from HSPICE parameters are presented, as well as means of including added delay due to wire resistance and the use of series transistors.

Key Words and Phrases: CMOS Delay Models, Inverter Delay, Capacitance Models, Wire Delay

Contents

1 Introduction 1

2 Estimating Capacitance 1

3 Analytical Delay Models 3

4 One Region Model 4

5 Two Region Model 6

6 Three Region Model 9

7 Alpha-Power Law Model 12

8 Model Summary 13

9 Buffer Delay 16

10 Series Transistors Delay 19

11 Wire Delay 21

12 Conclusion 23

Appendix A – HSPICE Models 24

Appendix B – Limits of the Two Region Model 24

Appendix C – Curve Fitting the Three Region Model 25

List of Figures

1 Current for 1 Region Model 5

2 Voltage for 1 Region Model 5

3 Current for 2 Region Model 8

4 Voltage for 2 Region Model 8

5 Current for 3 Region Model 11

6 Voltage for 3 Region Model 11

7 Delay vs Tin (Falling Input) 14

8 Delay vs Tin (Rising Input) 14

9 Delay vs Fanout (Falling Input) 15

10 Delay vs Fanout (Rising Input) 15

11 Buffer Delay (Falling Input) 17

12 Buffer Delay (Rising Input) 17

13 Buffer Delay (Falling Input) 18

14 Buffer Delay (Rising Input) 18

15 NAND Delay vs Fanout 20

16 NOR Delay vs Fanout 20

17 Wire Delay (Falling Input) 22

18 Wire Delay (Rising Input) 22

List of Tables

1 Introduction

Modern computer design would not be possible without the extensive simulation tools employed by today's engineers. These tools allow enormously complex circuits to be designed entirely by simulation with reasonable hope that the first fabricated version will work correctly. However, these design tools always make tradeoffs between speed, accuracy, and easy of use. Circuit simulators such as HSPICE are commonly used when accuracy is needed, but these simulators are very slow when dealing with large circuits and often difficult to use. Switch level simulators may be fast and easy to use, but they often lack sufficient accuracy to make them truly useful.

This gap creates a need for simple yet relatively accurate analytical models for circuit delay. Fast and easy to use models can allow computer designers to quickly assess the impact of different architectural choices where delay is a factor. Simple delay models also can identify a small number of critical paths to be simulated in more detail and allow CAD tools to perform basic optimization and sizing of many circuits. Compared to empirical models or simulations analytical models often provide a deeper understanding of the tradeoffs in a particular circuit.

To perform these functions delay models need sufficient accuracy to base design decisions upon, but they need not replace more intensive simulations. A model should be as simple as possible in order to reduce computation time and should include a specific scheme for choosing values of curve fitting parameters. This paper presents derivations for four different circuit delay models and compares these models using these criteria.

2 Estimating Capacitance

An important part of any circuit delay model is a means of estimating the various capacitances of the circuit. The capacitance of MOSFET devices is due to the gate capacitance and the source/drain junction capacitance. The gate capacitance is composed of the capacitance due to the gate oxide overlap of the highly doped source/drain regions and the capacitance to the inverted channel region. The gate capacitance varies with the transistor's region of operation which determines what fraction of the channel is inverted. For fairly well balanced circuits in the time for an output to change by $V_{DD}/2$ NMOS devices driven by a rising signal are primarily in the saturation region, and PMOS devices are primarily in the linear region. For a falling signal these regions are reversed. In the linear region the entire channel is inverted, and its capacitance is assumed to be shared equally between the source and drain. In the saturation region approximately two thirds of the channel is inverted, and it is entirely controlled by the source. The gate to source and gate to drain capacitances per unit width of a device of length L are estimated using these assumptions and the HSPICE parameters for the gate oxide capacitance (COX) and for the overlap capacitances ($CGSO$ and $CGDO$).

$$C_{GS} = \quad CGSO + \tfrac{1}{2}COX \times L \quad \text{Linear Region} \tag{1}$$
$$C_{GS} = \quad CGSO + \tfrac{2}{3}COX \times L \quad \text{Saturation Region} \tag{2}$$

$$
\begin{aligned}
C_{DS} = &\quad CGDO + \tfrac{1}{2}COX \times L \quad \text{Linear Region} & (3) \\
C_{DS} = &\quad CGDO \quad \text{Saturation Region} & (4)
\end{aligned}
$$

The junction capacitance of the source and drain is due to the bottom junction area (CJ_{AREA}), the sidewall junction perimeter (CJ_{PERM}), and the gate-edge sidewall junction (CJ_{GATE}). These capacitances are modeled as functions of the bias voltage (V_A) and the HSPICE parameters for bottom junction capacitance (CJ), bottom junction grading (MJ), sidewall junction capacitance ($CJSW$), sidewall grading ($MJSW$), gate-edge sidewall capacitance ($CJGATE$), and the bulk junction potential (PB).

$$
\begin{aligned}
CJ_{AREA} &= CJ(1 + V_A/PB)^{-MJ} & (5) \\
CJ_{PERM} &= CJSW(1 + V_A/PB)^{-MJSW} & (6) \\
CJ_{GATE} &= CJGATE(1 + V_A/PB)^{-MJSW} & (7)
\end{aligned}
$$

At a node rising from 0 to $V_{DD}/2$ NMOS diffusion goes from a bias of 0 to a bias of $V_{DD}/2$, and PMOS diffusion goes from a bias of V_{DD} to $V_{DD}/2$. At a falling node these biases are reversed. The effective junction capacitance for rising and falling transitions is estimated by averaging the capacitance at each of these biases.

Another means of estimating effective capacitances is to compare the delays of circuits with parasitic capacitances to those of circuits with the parasitic capacitors removed and replaced with discrete capacitors. The values of the discrete capacitors which give equal delays are used as the effective capacitances. The following table shows both the calculated values and the simulated values for the device models used in this paper. In all cases the calculated and simulated values are within 10% of each other.

	Rising Voltage				Falling Voltage			
	NMOS		PMOS		NMOS		PMOS	
	Calc	Sim	Calc	Sim	Calc	Sim	Calc	Sim
$C_{GS} + C_{GD}(fF/\mu m)$	1.54	1.56	2.21	2.25	2.09	2.22	1.66	1.54
$CJ_{AREA}(fF/\mu m^2)$	0.158	0.142	0.247	0.244	0.104	0.103	0.388	0.350
$CJ_{PERM}(fF/\mu m)$	0.338	0.318	0.237	0.236	0.258	0.256	0.329	0.306
$CJ_{GATE}(fF/\mu m)$	0.262	0.246	0.184	0.183	0.200	0.198	0.255	0.238

3 Analytical Delay Models

The basic approach of each of the four delay model derivations presented in this paper is the same. For a given load capacitance (C_L), the output voltage ($V_{out}(t)$) is related to the output current ($I_{out}(t)$) by the following differential equation:

$$\frac{dV_{out}(t)}{dt} = \frac{-I_{out}(t)}{C_L} \tag{8}$$

Each model chooses a different analytical expression for the output current of a device and solves this differential equation for the output voltage as a function of time. The expression of output voltage is then solved for an equation of delay.

All the following derivations are for a simple inverter driven by an input which begins to rise at time 0. All of the expressions of voltage are normalized to the supply voltage so that $V_{out}(t) = 0.5$ indicates that the output voltage is at half the supply voltage. Delay is always measured from the time the input begins to change to when the output has changed by the fraction ΔV_{out}. The equations for output voltage as a function of time are only valid for a rising input, but the equations for delay as a function of ΔV_{out} are identical for rising and falling inputs. All the models in this paper are fit to the HSPICE Level 3 device models found in Appendix A.

4 One Region Model

The simplest way to model the delay of a CMOS gate is to replace each transistor with an equivalent resistor. This is called the one region model because each device has only one region of operation. The current drawn by a device of width W is written as the output voltage across an equivalent resistance R_F/W.

$$I_{out}(t) = \frac{V_{out}(t)W}{R_F} \tag{9}$$

For a given load capacitance C_L the characteristic time constant T_F and differential equation are written as follows:

$$T_F = \frac{R_F C_L}{W} \tag{10}$$

$$\frac{dV_{out}}{dt} = \frac{-I_{out}(t)}{C_L} = \frac{-V_{out}(t)W}{R_F C_L} = \frac{-V_{out}(t)}{T_F} \tag{11}$$

Solving this equation for the output voltage as a function of time and delay (t_D) as a function of ΔV_{out} gives the following:

$$V_{out}(t) = \exp\left(\frac{-t}{T_F}\right) \tag{12}$$

$$t_D = T_F \log\left(\frac{1}{1 - \Delta V_{out}}\right) \tag{13}$$

$V_{out}(t)$ is a decaying exponential function normalized to V_{DD}, and the delay t_D is measured from the time the input begins to change to when the output has changed by a fraction of V_{DD} equal to ΔV_{out}. The one region model completely ignores the shape and slew rate of the input signal.

To fit the single curve fitting parameter R_F, the delay to when the output has changed 50% of V_{DD} (written as T_{50}) is measured and equation 13 is solved for R_F.

$$R_F = \frac{T_{50} W}{C_L \log 2} \tag{14}$$

This gives the following values for the device models used in this paper.

$$R_{FN} = 18.28\, kOhm \cdot \mu m \qquad\qquad R_{FP} = 33.00\, kOhm \cdot \mu m$$

On the following page figure 1 shows the one region model current compared to that of the HSPICE model. This model does a very poor job of accurately matching the output current. Figure 2 shows the rising input voltage and the falling output voltages of the one region model and the HSPICE model. The poor match of the model output current to HSPICE makes the model output voltage a poor fit and only accurate for a narrow range of ΔV_{out}. Since the one region model ignores the slope of the input signal, changes in the input increase these inaccuracies.

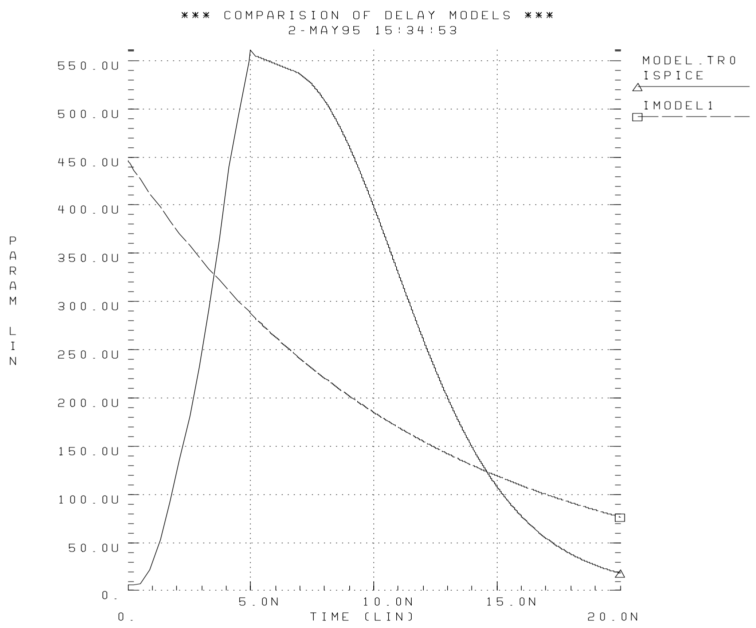

Figure 1: Current for 1 Region Model

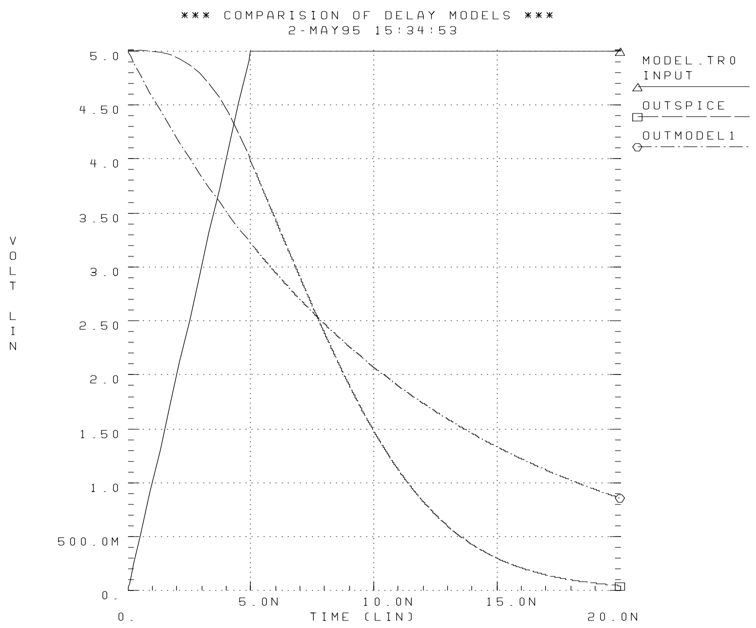

Figure 2: Voltage for 1 Region Model

5 Two Region Model

The accuracy of the one region model can be improved by noting that the output current of the inverter initially increases as the input voltage increases, while the output current of the one region model is always decreasing. Making the output current first proportional to the input voltage and then proportional to the output voltage gives a better match of output currents and a more accurate model. This is the approach taken by Horowitz [1] and is called the two region model because each device is now assumed to have two regions of operation.

In the first region the output current is proportional to some input waveform. For simplicity this paper assumes a normalized ramp input which goes from 0 to 1 in time T_{IN}.

$$V_{in}(t) = t/T_{IN} \tag{15}$$

The first region roughly corresponds to when the switching device is in the saturation regime. The output current is written as a function of the device width W, the input voltage $V_{in}(t)$, an equivalent resistance R_M, and the gate's switching voltage V_T. The second region corresponds to when the switching device is in the linear regime, and the output current is modeled by a simple resistor just as in the one region model.

$$I_{out}(t) = \min\left(\frac{(V_{in}(t) - V_T)W}{R_M}, \frac{V_{out}(t)W}{R_F}\right) \tag{16}$$

Solving for the output voltage and delay gives solutions with two regions in terms of two characteristic time constants T_M and T_F, and the time t_v when the output current first becomes greater than 0.

$$t_v = T_{IN}V_T \qquad T_M = R_M C_L/W \qquad T_F = R_F C_L/W \tag{17}$$

$$V_{out}(t) = \begin{cases} 1 - \dfrac{(t-t_v)^2}{2T_M T_{IN}} & t_v < t < t_s \\ \left(1 - \dfrac{(t_s - t_v)^2}{2T_M T_{IN}}\right)\exp\left(\dfrac{t_s - t}{T_F}\right) & t_s < t \end{cases} \tag{18}$$

$$t_D = \begin{cases} t_v + \sqrt{2T_M T_{IN}\Delta V_{out}} & 0 < t_D < t_s \\ t_s + T_F \log\left(\dfrac{(t_s - t_v)T_F}{T_M T_{IN}(1 - \Delta V_{out})}\right) & t_s < t_D \end{cases} \tag{19}$$

In the first region $V_{out}(t)$ is a quadratic function, and in the second region it is a decaying exponential. The time t_s, when the model switches from the first region to the second, is found by setting equal the two expressions within the minimum operator of equation 16 and setting $V_{out}(t)$ equal to the first region solution.

$$t_s = t_v + \sqrt{T_F^2 + 2T_M T_{IN}} - T_F \tag{20}$$

In equation 15 there is no limit on the value of $V_{in}(t)$. This approximation treats $V_{in}(t)$ as a continually increasing function. This leads to poor approximations if t_s is greater than T_{IN}, and the model current continues to increase after a real input voltage would have reached V_{DD} and stopped increasing. However, if R_M and R_F are chosen appropriately t_s is always be less than T_{IN} (see Appendix B).

Horowitz approximates both regions of the equation for delay with the following [1]:

$$t_D \approx t_v + \sqrt{(T_F \log(1 - \Delta V_{out}))^2 + 2 T_{IN} T_M \Delta V_{out}} \tag{21}$$

To fit the two region model V_T is set to the logic threshold of the gate while R_M and R_F are varied to find the minimum percent error over a wide range of input slew times and fanouts. This gives the following values for the device models used in this paper.

$V_{TN} = 0.5$	$V_{TP} = 0.5$
$R_{MN} = 3.96\, kOhm \cdot \mu m$	$R_{MP} = 10.10\, kOhm \cdot \mu m$
$R_{FN} = 10.63\, kOhm \cdot \mu m$	$R_{FP} = 22.88\, kOhm \cdot \mu m$

On the following page figure 3 shows the two region model current compared to that of the HSPICE model. The model current now increases linearly as the input increases and then drops off exponentially. This is a much better fit to the HSPICE model which shows the device current increasing until the input voltage reaches its maximum and then decaying. Figure 4 shows the rising input voltage and the falling output voltages of the two region model and the HSPICE model. The two region does a much better job of approximating the output voltage than the one region model (compare figure 2). Also since the two region model takes into account the slope of the input signal, it retains better accuracy as the input waveform varies.

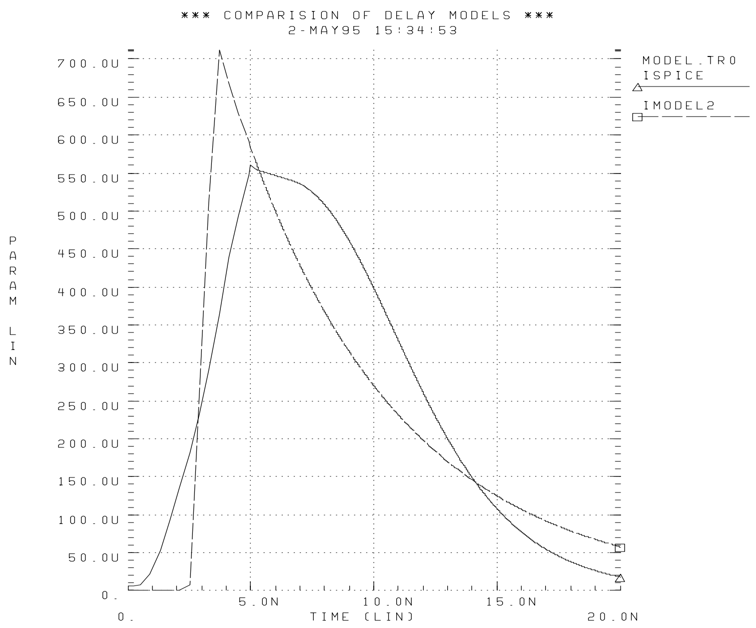

Figure 3: Current for 2 Region Model

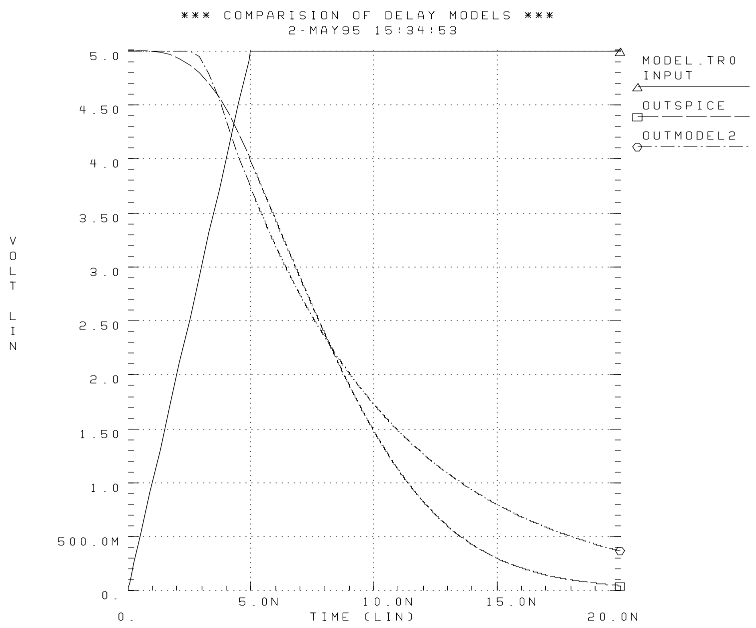

Figure 4: Voltage for 2 Region Model

6 Three Region Model

The two region model can be further improved by looking again at the HSPICE current in figure 3. After reaching its maximum the current does not immediately begin to decay exponentially as is assumed by the two region model. For a short time after it reaches its maximum, the device current is relatively constant. This corresponds to the time period when the device is in the saturation regime, but the input waveform is no longer changing. A more accurate model is created by adding another region corresponding to this period. This is the approach taken by Sakurai [2] and in this paper is called the three region model.

To model three regions the input voltage is written as a function which goes from 0 to 1 in time T_{IN} and then remains constant. The explicit maximum value for $V_{in}(t)$ is the only initial assumption which is different from the two region model.

$$V_{in}(t) = \min\left(t/T_{IN}, 1\right) \tag{22}$$

The output current is written as being first a function of the input voltage and then a function of the output voltage. This expression is identical to the one used in the two region model, only now V_T corresponds more closely to the switching device threshold than to the gate's logic threshold.

$$I_{out}(t) = \min\left(\frac{(V_{in}(t) - V_T)W}{R_M}, \frac{V_{out}(t)W}{R_F}\right) \tag{23}$$

Solving for the output voltage and delay gives solutions with three regions in terms of the time constants t_v, T_M, and T_F. The three regions roughly correspond to the saturation regime with an increasing input voltage, the saturation regime with a constant input voltage, and the linear regime.

$$t_v = T_{IN}V_T \qquad T_M = R_M C_L/W \qquad T_F = R_F C_L/W \tag{24}$$

$$V_{out}(t) = \begin{cases} 1 - \dfrac{(t-t_v)^2}{2T_M T_{IN}} & t_v < t < T_{IN} \\[2ex] 1 - \dfrac{(1-V_T)(2t - T_{IN}(1+V_T))}{2T_M} & T_{IN} < t < t_s \\[2ex] \left(1 - \dfrac{(1-V_T)(2t_s - T_{IN}(1+V_T))}{2T_M}\right)\exp\left(\dfrac{t_s - t}{T_F}\right) & t_s < t \end{cases} \tag{25}$$

$$t_D = \begin{cases} t_v + \sqrt{2T_M T_{IN}\Delta V_{out}} & 0 < t_D < T_{IN} \\[2ex] \dfrac{T_{IN}(1+V_T)}{2} + \dfrac{T_M \Delta V_{out}}{1-V_T} & T_{IN} < t_D < t_s \\[2ex] t_s + T_F \log\left(\dfrac{2T_M - (1-V_T)(2t_s - T_{IN}(1+V_T))}{2T_M(1-\Delta V_{out})}\right) & t_s < t_D \end{cases} \tag{26}$$

In the first region $V_{out}(t)$ is a quadratic function. In the second region, which is the region not found in the two region model, $V_{out}(t)$ is linear. In the third region it is a decaying exponential. The transition from the first to the second region occurs at time T_{IN}, when the input voltage reaches its maximum. The time t_s when the model switches from the second region to the third is found by setting equal the two expressions within the minimum operator of equation 23 and setting $V_{out}(t)$ equal to the linear region solution.

$$t_s = \frac{T_{IN}(1+V_T)}{2} + \frac{T_M}{1-V_T} - T_F \tag{27}$$

When approximating delays to $V_{DD}/2$ ($\Delta V_{out} = 0.5$), the result most often falls in the linear region. Therefore, as a simple approximation the other two regions are ignored and the linear region equation is used alone.

$$t_D \approx \frac{T_{IN}(1+V_T)}{2} + \frac{T_M \Delta V_{out}}{1-V_T} \tag{28}$$

To fit the three region model V_T, R_M, and R_F are all varied in order to minimize percent error in delay over a wide range of input slew times and fanouts. This gives the following values for the device models used in this paper.

$$V_{TN} = 0.313 \qquad V_{TP} = 0.308$$

$$R_{MN} = 11.51\, kOhm \cdot \mu m \qquad R_{MP} = 27.41\, kOhm \cdot \mu m$$

$$R_{FN} = 5.31\, kOhm \cdot \mu m \qquad R_{FP} = 14.24\, kOhm \cdot \mu m$$

On the following page figure 5 shows the three region model current compared to that of HSPICE. The model current now increases linearly, then remains constant for a time, and then decays exponentially. This is a better fit to the HSPICE current than the two region model, and figure 6 shows that this current gives a voltage curve which is an excellent approximation of the HSPICE voltage.

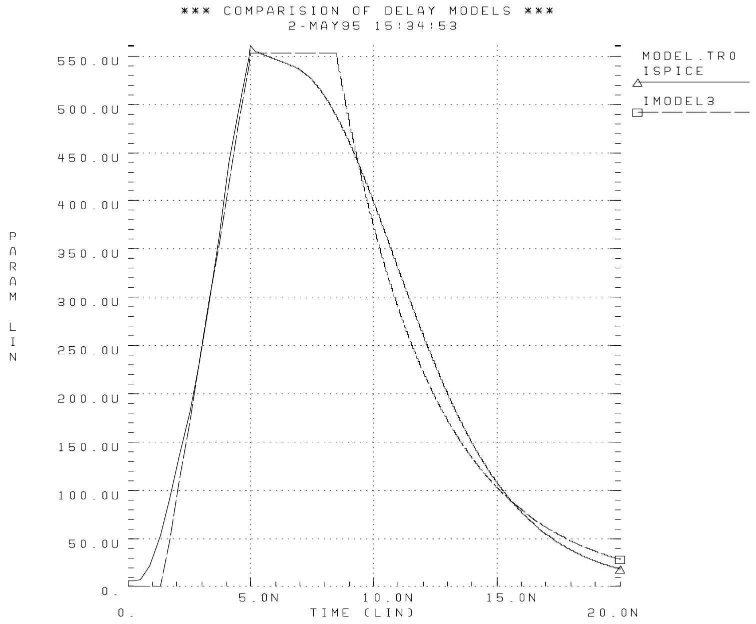

Figure 5: Current for 3 Region Model

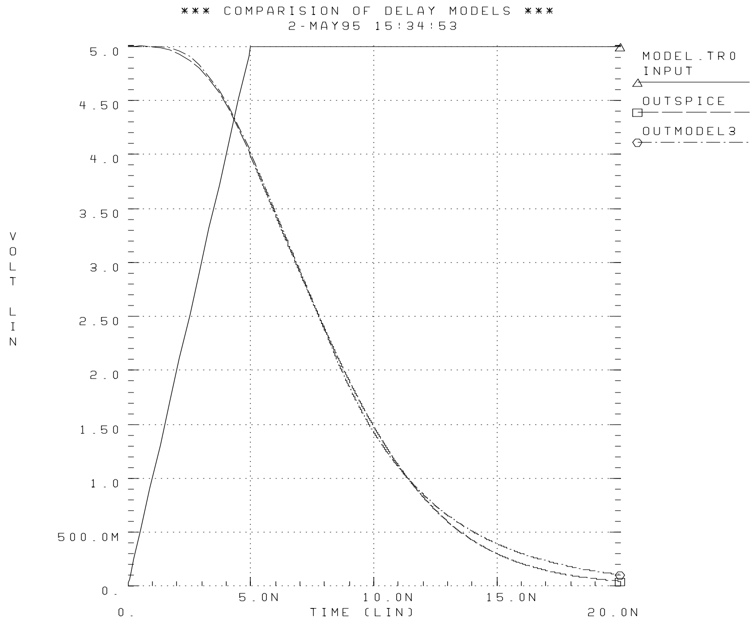

Figure 6: Voltage for 3 Region Model

7 Alpha-Power Law Model

The three region model assumes that the output current is a linear function of the input voltage. This is actually only the case for a completely velocity saturated device. For a device with no velocity saturation the current is a function of the square of the input voltage. Most modern devices with fall somewhere in between these two extremes. This is taken into account by introducing a new curve fitting parameter α and writing the output current as follows:

$$I_{out}(t) = \min\left(\frac{(V_{in}(t) - V_T)^{\alpha}\, W}{R_M}, \frac{V_{out}(t) W}{R_F}\right) \qquad 1 \le \alpha \le 2 \tag{29}$$

This is the approach used by Nabavi-Lishi and is called the alpha-power law model [3]. The same input wave form is assumed as for the three region model, and $V_{out}(t)$ and delay are solved for in the same way. The use of α in the expression for current is the only initial assumption which is different from the three region model.

$$t_v = T_{IN} V_T \qquad T_M = R_M C_L / W \qquad T_F = R_F C_L / W \tag{30}$$

$$V_{out}(t) = \begin{cases} 1 - \dfrac{(t - t_v)^{\alpha+1}}{(1+\alpha)(1-V_T)^{\alpha-1} T_M T_{IN}^{\alpha}} & t_v < t < T_{IN} \\ 1 - \dfrac{(1-V_T)((1+\alpha)t - T_{IN}(\alpha + V_T))}{(1+\alpha)T_M} & T_{IN} < t < t_s \\ \left(1 - \dfrac{(1-V_T)((1+\alpha)t_s - T_{IN}(\alpha+V_T))}{(1+\alpha)T_M}\right) \exp\left(\dfrac{t_s - t}{T_F}\right) & t_s < t \end{cases} \tag{31}$$

$$t_D = \begin{cases} t_v + \sqrt[\alpha+1]{(1+\alpha)(1-V_T)^{\alpha-1} T_M T_{IN}^{\alpha} \Delta V_{out}} & 0 < t_D < T_{IN} \\ \dfrac{T_{IN}(\alpha+V_T)}{1+\alpha} + \dfrac{T_M \Delta V_{out}}{1 - V_T} & T_{IN} < t_D < t_s \\ t_s + T_F \log\left(\dfrac{(1+\alpha)T_M - (1-V_T)((1+\alpha)t_s - T_{IN}(\alpha+V_T))}{(1+\alpha)T_M(1-\Delta V_{out})}\right) & t_s < t_D \end{cases} \tag{32}$$

$$t_s = \frac{T_{IN}(\alpha+V_T)}{1+\alpha} + \frac{T_M}{1-V_T} - T_F \tag{33}$$

For the case where $\alpha = 1$ the alpha-power law model is identical to the three region model. When approximating delays to $V_{DD}/2$ ($\Delta V_{out} = 0.5$), the result most often falls in the linear region. Therefore, as a simple approximation the other two regions are ignored and the linear region equation is used alone.

$$t_D \approx \frac{T_{IN}(\alpha+V_T)}{1+\alpha} + \frac{T_M \Delta V_{out}}{1 - V_T} \tag{34}$$

This single equation and the three region approximation given in equation 28 are both are simply the sum of two terms, one proportional to the input slew time, and one proportional to the capacitive load. For any value of α the only difference between these equations is the values of V_T and R_M used to fit a given set of data. Therefore, when using this single equation approximation the alpha-power law model and the three region model are identical.

8 Model Summary

One Region Model

$$t_D \approx T_F \log\left(\frac{1}{1-\Delta V_{out}}\right) \tag{35}$$

The one region model is simple, easy to use and requires only a single curve fitting parameter (R_F). However, it fails to take into account the slope of the input waveform, and therefore has very poor accuracy.

Two Region Model

$$t_D \approx t_v + \sqrt{(T_F \log(1-\Delta V_{out}))^2 + 2T_{IN}T_M\Delta V_{out}} \tag{36}$$

The two region approximation is a more complex equation which requires three curve fitting parameters (V_T, R_M, R_F), but it includes the effect of the input slope and gives better accuracy than the one region model.

Three Region Model

$$t_D \approx \frac{T_{IN}(1+V_T)}{2} + \frac{T_M\Delta V_{out}}{1-V_T} \tag{37}$$

The three region approximation (which is identical to the alpha-power law approximation) is a simple equation with only two curve fitting parameters (V_T, R_M). Like the two region model it takes into account the input slope to provide better accuracy.

For anything but the most general estimations the one region model does not have sufficient accuracy. To choose between the two and three region models their accuracy is compared over changing input slopes and capacitive loads. In the following two pages figures 7 through 10 show that the two and three region models have roughly equal accuracy, with the two region model doing slightly better for large input slew times and fanouts, and the three region model doing slightly better for small input slew times and fanouts. The three region model provides roughly the same accuracy as the two region model with a simpler equation and fewer curve fitting parameters. This makes curve fitting the three region model much easier, and therefore it is used in the rest of this paper. A simple method for fitting the three region model is presented in appendix C.

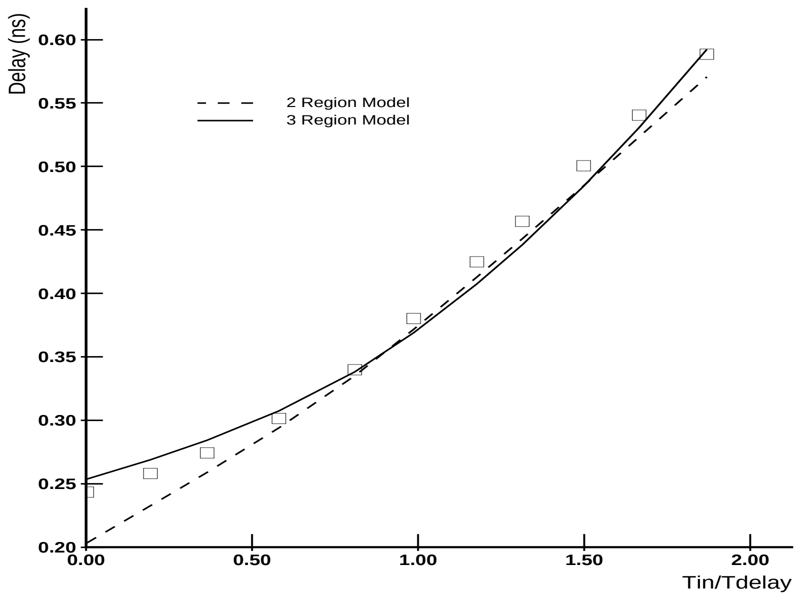

Figure 7: Delay vs Tin (Falling Input)

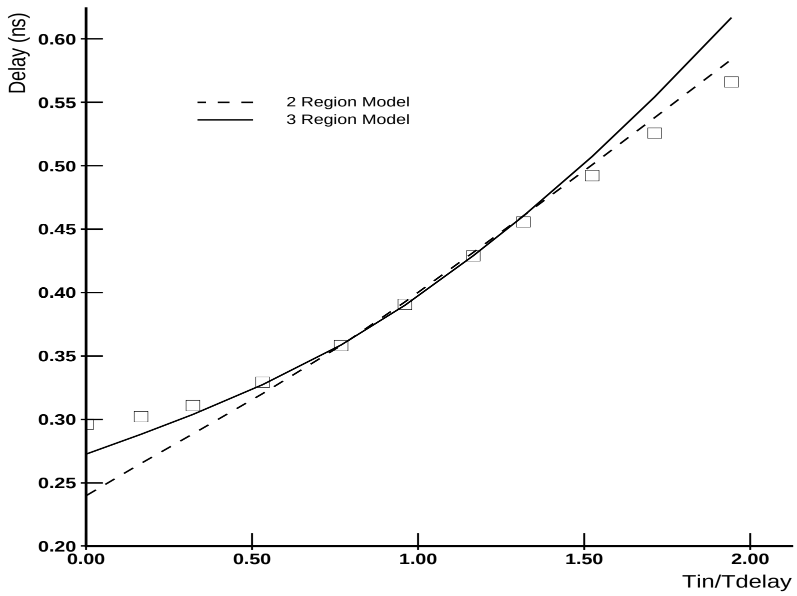

Figure 8: Delay vs Tin (Rising Input)

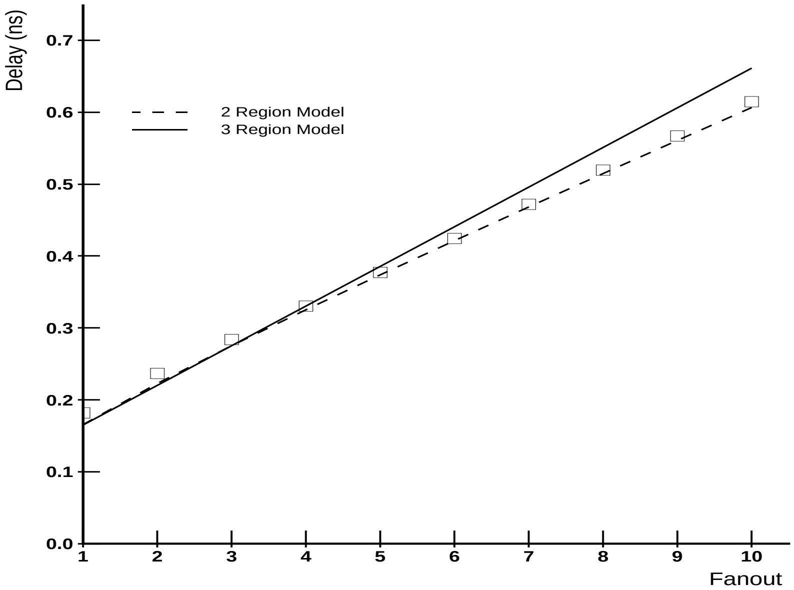

Figure 9: Delay vs Fanout (Falling Input)

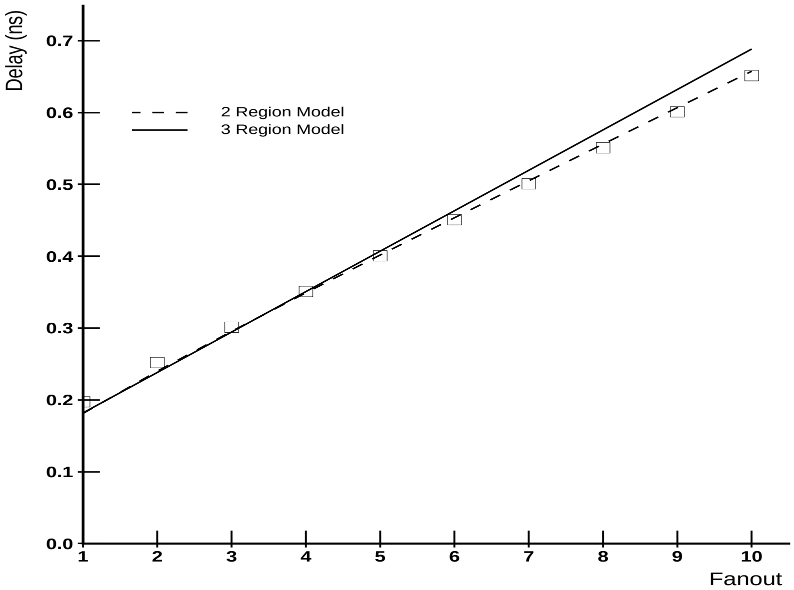

Figure 10: Delay vs Fanout (Rising Input)

9 Buffer Delay

Predicting the delay of a buffer comprised of multiple stages of inverters requires using the input slew time and delay of one stage to calculate the effective input slew time into the next stage. If T_{IN1} is the input slew time into the first stage, then the output of the first stage begins to change after a time $V_T \times T_{IN1}$. Let T_{D1} represent the delay from when the input to the first stage begins to change to when the output of the first stage reaches $V_{DD}/2$. If the output were changing linearly, the input slew time to the next stage would be twice the interval between T_{D1} and $V_T \times T_{IN1}$. However, the effective input slew time is less than this because the output begins by changing quadratically. To take this into account a curve fitting parameter S_{IN} is used to calculate the effective input slew time to the second stage (T_{IN2}). In this paper a value of $S_{IN} = 0.79$ is used.

$$T_{IN2} = 2(T_{D1} - V_T T_{IN1})S_{IN} \tag{38}$$

Another concern in real circuits is that delay is often greatly reduced by using circuits with reduced noise margins. The delay of static CMOS circuits is greatly reduced by setting the ratio of the PMOS and NMOS devices to favor a single transition. However, this reduces the noise margin and the circuit's cycle time. Timed circuits such as domino or post-charge logic use timed reset devices to improve the circuit's cycle time, but they are still making the same tradeoff of reducing the delay of a single transition by reducing the noise margin. In this paper the noise margin of a gate is defined as the minimum voltage from either V_{DD} or ground which produces an output voltage of $V_{DD}/2$. The greatest noise margin possible is achieved by the ratio of device widths where an input of $V_{DD}/2$ produces an output of $V_{DD}/2$. Expressed as a faction of V_{DD} this corresponds to a noise margin of 0.5. Figures 11 through 14 show a good match between the delay of the three region model and the delay found with HSPICE for an optimum number of inverters of various noise margins and fanouts. The ratios of the inverter stages alternate so that if one stage favors a rising transition the next stage favors a falling transition.

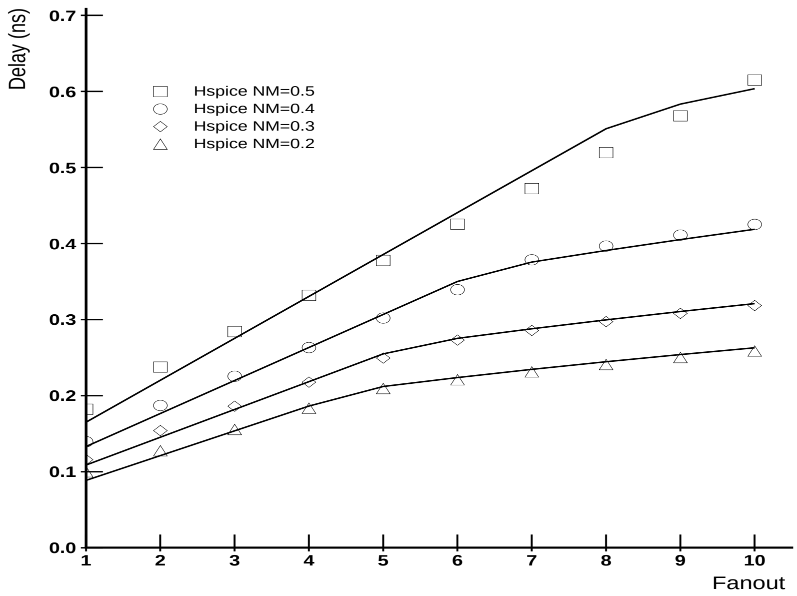

Figure 11: Buffer Delay (Falling Input)

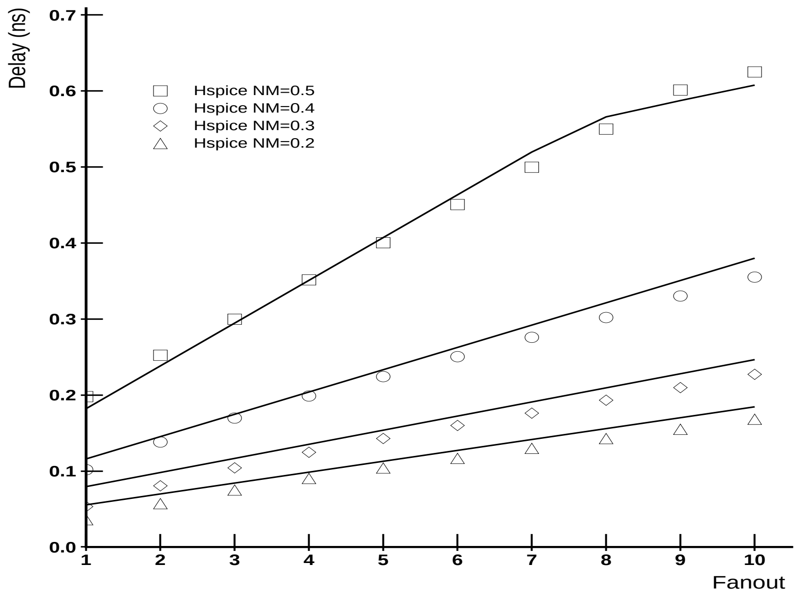

Figure 12: Buffer Delay (Rising Input)

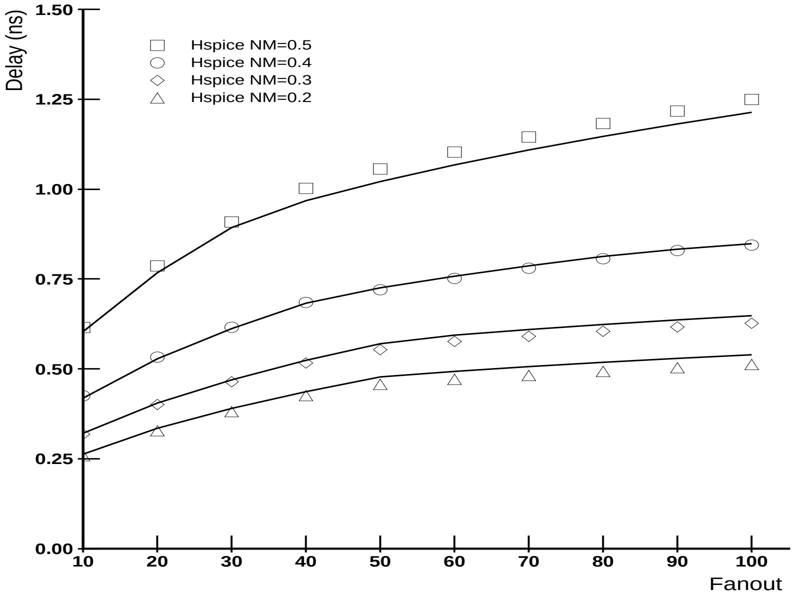

Figure 13: Buffer Delay (Falling Input)

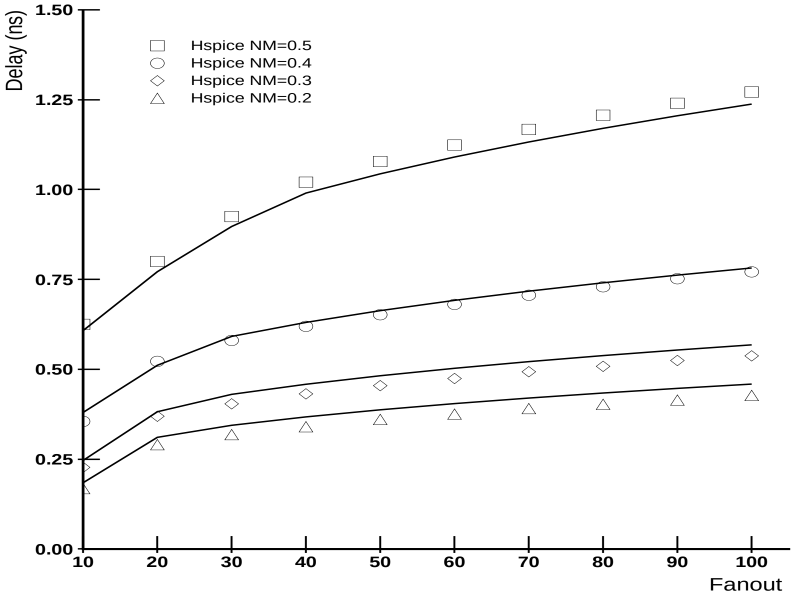

Figure 14: Buffer Delay (Rising Input)

10 Series Transistors Delay

To predict the delay of logic gates besides simple inverters the three region model is made to fit the delay of series transistors [4]. This is done by making the effective resistance R_M and the effective threshold V_T functions of the number of series devices. First N transistors in series of width W are given an effective width of W/N. This is sufficient and no new values for R_M and V_T are needed if the devices display no velocity saturation or body effect.

With velocity saturation the reduction in V_{DS} caused by connecting devices in series fails to reduce the device current as much as expected. This tends to reduce the effective resistance R_M. The body effect causes an increase in the magnitude of the device threshold. This tends to increase the effective threshold V_T needed to fit the three region model. Both of these effects are difficult to model analytically. The simplest approach is to use HSPICE simulations to find new values for R_M and V_T for each number of series devices which is to be used. For the HSPICE models used in this paper the following values are found (R_M has units $kOhm \cdot \mu m$, and V_T is unitless):

	NMOS		PMOS	
	R_M	V_T	R_M	V_T
1 Series	11.51	0.313	27.41	0.308
2 Series	9.13	0.372	21.70	0.414
3 Series	8.26	0.404	19.22	0.484
4 Series	7.78	0.430	19.69	0.488

These values give a reasonable fit of the delay of NAND and NOR gates with various numbers of inputs as shown in figures 15 and 16. These figures assume that all the NAND inputs are rising simultaneously and that all the NOR inputs are falling simultaneously.

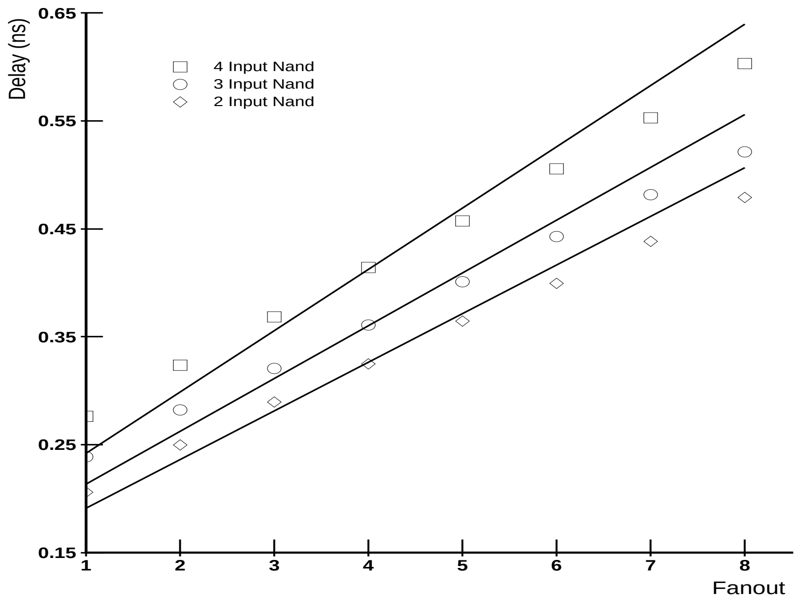

Figure 15: NAND Delay vs Fanout

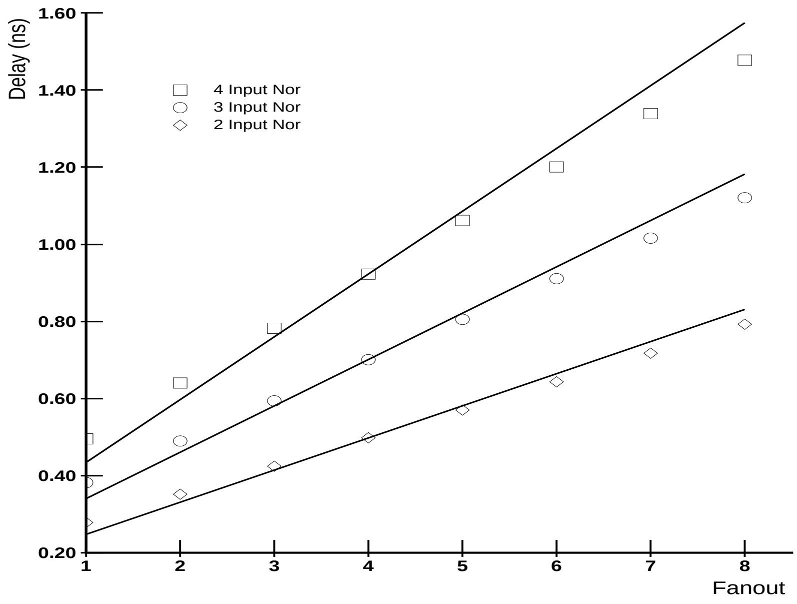

Figure 16: NOR Delay vs Fanout

11 Wire Delay

Real systems also require the modeling of the added delay of driving a load through a wire of significant resistance. First the delay for no wire resistance (t_{DO}) is estimated by using the three region model and adding the capacitance of the wire (C_W) to the load capacitance.

$$T_M = R_M(C_D + C_W + C_L)/W \tag{39}$$

$$t_{DO} = \frac{T_{IN}(1+V_T)}{2} + \frac{T_M \Delta V_{out}}{1-V_T} \tag{40}$$

If the wire resistance (R_W) is not zero but much less than the transistor resistance (R_M), a simple π model is used. This model places half the wire capacitance at the driver output and half at the load separated by the wire resistance. Differential equations are then written for the voltage at the driver output ($V_W(t)$) and the voltage at the load ($V_{out}(t)$).

$$\frac{dV_{out}(t)}{dt} = \frac{V_W(t) - V_{out}(t)}{R_W(0.5C_W + C_L)} \tag{41}$$

$$\frac{dV_W(t)}{dt} = \frac{V_{out}(t) - V_W(t)}{R_W(0.5C_W + C_D)} - \frac{V_{in}(t) - V_T}{R_M(0.5C_W + C_D)} \tag{42}$$

Using the same input as assumed for the three region model, this system of differential equations is solved to give the an expression for delay in the linear region as follows:

$$T_M = R_M(C_D + C_W + C_L)/W \qquad T_{W1} = R_W(0.5C_W + C_D) \qquad T_{W2} = R_W(0.5C_W + C_L) \tag{43}$$

$$t_D = \frac{T_{IN}(1+V_T)}{2} + \frac{T_M \Delta V_{out}}{1-V_T} + \frac{T_{W1}T_{W2}}{T_{W1}+T_{W2}} = t_{DO} + \frac{T_{W1}T_{W2}}{T_{W1}+T_{W2}} \tag{44}$$

The added delay due to the wire resistance (T_{WIRE1}) is simply the final term.

$$T_{WIRE1} = \frac{T_{W1}T_{W2}}{T_{W1}+T_{W2}} \qquad\qquad R_W \ll R_M \tag{45}$$

If the resistance of the wire is much larger than the equivalent resistance of the transistor driving it, the added delay is approximately that of a step input driving a distributed RC. For a wire with capacitance C_W tied to a load of capacitance C_L the added delay to the 50% point is [5]:

$$T_{WIRE2} = R_W(0.4C_W + 0.7C_L) \qquad\qquad R_W \gg R_M \tag{46}$$

An equation for the general case is created by combining these two cases. The term T_{WIRE1} is multiplied by a faction estimating the percentage delay due to device resistance, and the term T_{WIRE2} is multiplied by a fraction estimating the percentage delay due to wire resistance. Adding together both these factors gives a general equation as follows:

$$t_D = t_{DO} + T_{WIRE1}\left(\frac{t_{DO}}{t_{DO}+T_{WIRE1}}\right) + T_{WIRE2}\left(\frac{T_{WIRE1}}{t_{DO}+T_{WIRE1}}\right) = t_{DO} + \frac{T_{WIRE2}+t_{DO}}{1+t_{DO}/T_{WIRE1}} \tag{47}$$

Figures 17 and 18 compare the model delays through a 20mm wire with and without wire resistance to HSPICE simulations for a wide range of driver sizes.

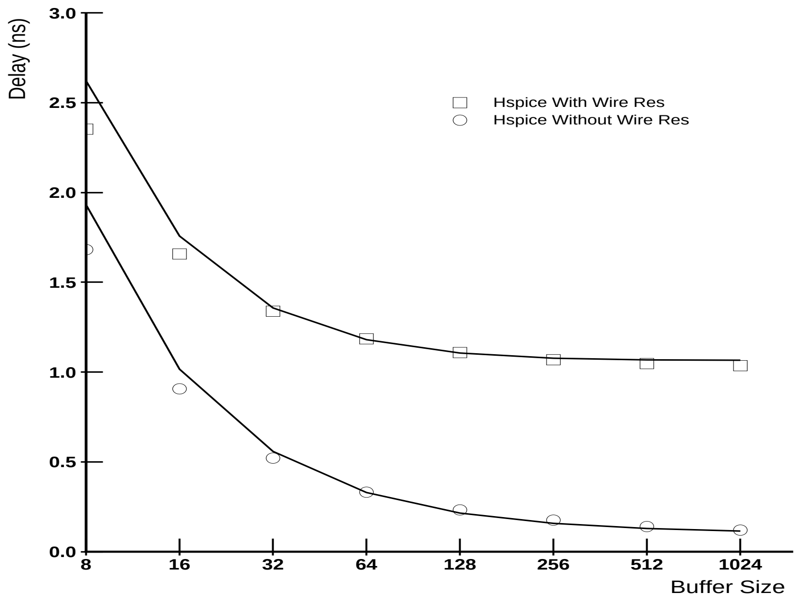

Figure 17: Wire Delay (Falling Input)

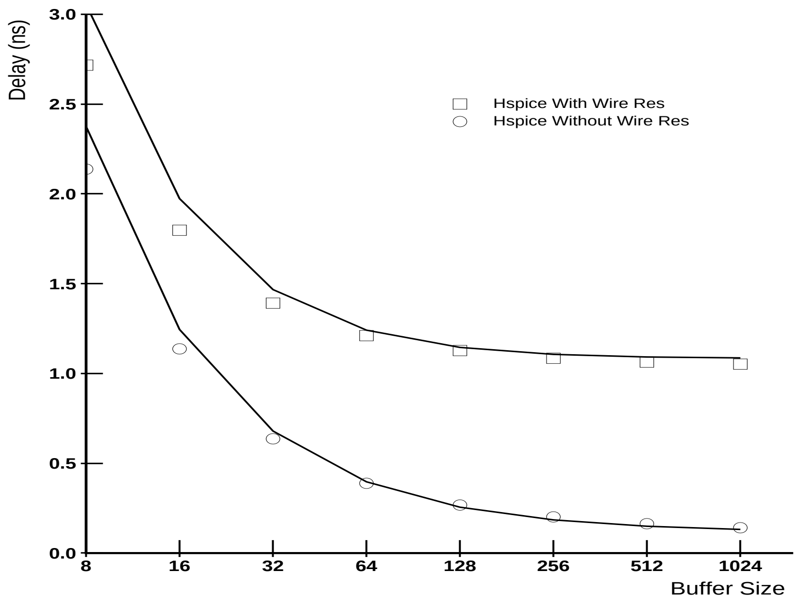

Figure 18: Wire Delay (Rising Input)

12 Conclusion

This paper derives four different CMOS inverter delay models and shows that inverter delay is simply and fairly accurately modeled over a wide range of input slopes and capacitive loads using an equation of the form:

$$t_D = K_1 T_{IN} + K_2 C_{LOAD} \tag{48}$$

where K_1 and K_2 are curve fitting parameters. Logic gate delay through series transistors is estimated by the same equation with different sets of curve fitting parameters for each number of series transistors. Methods for estimating the added delay due to wire resistance are also presented.

Simple analytical models such as this do not replace independent verification by more accurate simulations, but they do possess sufficient accuracy to be used in high level design choices and basic circuit optimizations.

Appendix A – HSPICE Models

```
.OPTIONS DEFL=0.8u DEFW=1.6u

.MODEL TN NMOS LEVEL=3
+ VTO=0.77      TOX=1.65E-8    UO=570          GAMMA=0.80
+ VMAX=2.7E5    THETA=0.404    ETA=0.04        KAPPA=1.2
+ PHI=0.90      NSUB=8.8E16    NFS=4E11        XJ=0.2U
+ PB=0.80       DELTA=0.0      LD=0.0001U      RSH=0.5
+ ACM=3         HDIF=1U        MJ=0.389        MJSW=0.26
+ CGSO=2.1E-10 CGDO=2.1E-10
+ CJ=2E-4       CJSW=4.00E-10  CJGATE=3.1E-10

.MODEL TP PMOS LEVEL=3
+ VTO=-0.87     TOX=1.65E-8    UO=145          GAMMA=0.73
+ VMAX=0.0      THETA=0.233    ETA=0.028       KAPPA=0.04
+ PHI=0.90      NSUB=9.0E16    NFS=4E11        XJ=0.2U
+ PB=0.80       DELTA=0.0      LD=0.0001U      RSH=0.5
+ ACM=3         HDIF=1U        MJ=0.420        MJSW=0.31
+ CGSO=2.7E-10 CGDO=2.7E-10
+ CJ=5E-4       CJSW=4.00E-10 CJGATE=3.1E-10
```

Appendix B – Limits of the Two Region Model

The two region model does not explicitly limit the value of the input voltage $V_{in}(t)$. The function $V_{in}(t)$ increases linearly from 0 to 1 in time T_{IN}, but then continues to increase. Therefore, it is important to choose values for the equivalent resistances R_M and R_F such that the model current stops being a function of $V_{in}(t)$ before time T_{IN}. This is true if the time t_s which marks the boundary between the first and second regions of the model (see equation 20) is less than T_{IN} for any value of T_{IN}. This requirement is written as:

$$\lim_{T_{IN}\to 0}\left(\frac{t_s}{T_{IN}}\right) = \lim_{T_{IN}\to 0}\left(\frac{t_v + \sqrt{T_F^2 + 2T_M T_{IN}} - T_F}{T_{IN}}\right) \leq 1 \tag{49}$$

In the limit both the numerator and denominator of this fraction are zero. Applying L'Hôpital's Rule by taking the derivative of the numerator and the denominator with respect to T_{IN} gives the following:

$$\lim_{T_{IN}\to 0}\left(V_T + \frac{T_M}{\sqrt{T_F^2 + 2T_M T_{IN}}}\right) = V_T + \frac{T_M}{T_F} = V_T + \frac{R_M}{R_F} \leq 1 \tag{50}$$

This limit gives the relation:

$$t_s \leq T_{IN} \qquad \text{for} \qquad \frac{R_M}{R_F} \leq 1 - V_T \tag{51}$$

Therefore, if the ratio of R_M to R_F is less than $1 - V_T$, no limiting value for $V_{in}(t)$ is needed.

Appendix C – Curve Fitting the Three Region Model

Any model which uses empirical parameters requires a scheme for choosing the values of those parameters. To fit the three region model first an input slew time T_{IN} and a capacitive load C_L typical of the circuits to be modeled are chosen. Then $V_{out}(T_{IN})$, the normalized output voltage at time T_{IN}, and T_{50}, the time from when the input begins to change to when the output has changed 50%, are measured. Solving the quadratic region equation for $V_{out}(T_{IN})$ gives:

$$V_{out}(T_{IN}) = 1 - \frac{T_{IN}(1 - V_T)^2}{2T_M} \tag{52}$$

This is rewritten as follows:

$$\frac{(1 - V_T)^2}{R_M} = \frac{2C_L(1 - V_{out}(T_{IN}))}{T_{IN}W} \tag{53}$$

The slope of $V_{out}(t)$ in the linear region of the model is equal to $-(1 - V_T)/T_M$. Assuming that T_{50} falls in the linear region gives:

$$\frac{-(1 - V_T)}{T_M} = \frac{V_{out}(T_{IN}) - V_{out}(T_{50})}{T_{IN} - T_{50}} = \frac{V_{out}(T_{IN}) - 0.5}{T_{IN} - T_{50}} \tag{54}$$

This is rewritten as follows:

$$\frac{1 - V_T}{R_M} = \frac{C_L(V_{out}(T_{IN}) - 0.5)}{(T_{50} - T_{IN})W} \tag{55}$$

Dividing equation 53 by equation 55 and solving for V_T gives:

$$V_T = 1 - \frac{2(1 - V_{out}(T_{IN}))(T_{50} - T_{IN})}{(V_{out}(T_{IN}) - 0.5)T_{IN}} \tag{56}$$

This equation for V_T contains only the value of T_{IN} chosen and the measured values $V_{out}(T_{IN})$ and T_{50}. Having found a value for V_T, equation 52 is solved for R_M.

$$R_M = \frac{T_{IN}W(1 - V_T)^2}{2C_L(1 - V_{out}(T_{IN}))} \tag{57}$$

These last two equations allow values for V_T and R_M to be quickly found with a simple simulation.

References

[1] M. Horowitz. "Timing Models for MOS Circuits", *Stanford University Dissertation*, 1985, Chapter 5.

[2] T. Sakurai. "Alpha-Power Law MOSFET Model and its Applications to CMOS Inverter Delay and Other Formulas", *IEEE Journal of Solid-State Circuits*, 1990, p. 584.

[3] A. Nabavi-Lishi. "Inverter Models of CMOS Gates for Supply Current and Delay Evaluation", *IEEE Transactions on Electron Devices*, 1994, p. 1271.

[4] T. Sakurai. "Delay Analysis of Series-Connected MOSFET Circuits", *IEEE Journal of Solid-State Circuits*, 1991, p. 122.

[5] H. Bakoglu, "Circuits, Interconnections, and Packaging for VLSI", Addison-Wesley, 1990.