PIECEWISE LINEAR DELAY MODELING OF

DIGITAL VLSI CIRCUITS

By

JIAN CHANG

Bachelor of Science
Beijing Institute of Machinery Industry
Beijing, P. R. China
1993

Master of Science
Oklahoma State University
Stillwater, OK
2001

Submitted to the Faculty of the
Graduate College of the
Oklahoma State University
in partial fulfillment of
the requirements for
the Degree of
DOCTOR OF PHILOSOPHY
May, 2006

PIECEWISE LINEAR DELAY MODELING OF

DIGITAL VLSI CIRCUITS

Dissertation Approved:

Dr. Louis G. Johnson

Dissertation Adviser

Dr. Weili Zhang

Dr. Yumin Zhang

Dr. Nohpill Park

A. Gordon Emslie

Dean of the Graduate College

PREFACE

One of the most important performance measures of digital logic circuits is the delays of switching signals propagating through the logic gates of the circuit. Circuit simulators such as SPICE can find the delay by solving for the current and voltage waveforms as functions of time. Although SPICE can handle the complex, nonlinear behavior of the transistors, it takes a significant amount of computations. Usually no more than a few thousand transistors may be simulated in a reasonable amount of computation time. Simulator such as IRSIM uses the switch model to find the delay, which greatly improves the simulation speed and can process hundreds of thousands of transistors in a reasonable amount of time. But IRSIM predicts delays much less accurate than SPICE because of its delay model inaccuracies.

In this paper, a piecewise linear delay model which can evaluate the propagation delay of a CMOS VLSI circuitry with a wide range of input slope is presented. The model also takes into account the influences of short circuit current and dynamic channel charges. By using simple piecewise linear current model and piecewise linear channel charge storage model, it is possible to simulate the modern digital logic circuits in a reasonable amount of time. Excellent agreements with SPICE simulation have been observed in a CMOS inverter, a two-input NAND gate, and an OAI gate cases. This model is applicable not only to propagation delay calculation of simple gates but also to that of any general circuit topology.

ACKNOWLEDGMENTS

I am finally working on this last piece of my Ph. D. dissertation after four and a half years' hard working. I would like to take this opportunity to sincerely thank all the people for every bit of their contributions whatsoever, to this dissertation.

First and foremost I would like to express my gratitude to my adviser, Dr. Louis G. Johnson, for his continuous support and guidance for my Ph. D. research. It is his insightful understanding of the digital VLSI theory that comes up with the topic of this dissertation and the piecewise linear model thereafter. I would like to sincerely thank him for sharing his understanding and knowledge in this area.

I am greatly thankful to Dr. Weili Zhang, Dr. Yumin Zhang, and Dr. Nohpill Park for taking time out of their busy schedules and providing valuable feedback while serving as members of my Ph. D. committee. Their thoughtful comments and questions were greatly valued.

Special thanks to Caifeng, wife of Dr. Weili Zhang, for helping me and my family in the most difficult times in my Ph.D. study. My gratefulness also goes to all of my research group members, to Cheng Liu, for discussing the numerous questions I met in my research.

Last but not least, I thank my dad and mom, the dearest parents in the world, my lovely wife, Jean, for their supports and encouragements. A special thank-you must go to my son Matthew and daughter Daisy. They give me many joyous moments in my life.

TABLE OF CONTENTS

Chapter Page

LIST OF FIGURES

LIST OF TABLES

Table Page

CHAPTER I

INTRODUCTION

Modern digital logic circuits have millions of transistors on a single silicon chip. It is desirable to learn the circuit performance during the design stage. One of the most important performance measures of digital logic circuits is the delays of switching signals propagating through the logic gates of the circuit. Circuit simulators such as SPICE can find the delay by solving for the current and voltage waveforms as functions of time. Although SPICE can handle the complex, nonlinear behavior of the transistors, it takes a significant amount of computations. Usually no more than a few thousand transistors may be simulated in a reasonable amount of computation time. Simulator such as IRSIM uses the switch model to find the delay, which greatly improves the simulation speed and can process hundreds of thousands of transistors in a reasonable amount of time. But IRSIM predicts delays much less accurate than SPICE because of its delay model inaccuracies.

This dissertation proposes a new delay model to bridge the gap between the IRSIM and SPICE. By applying the new delay model, the computation efficiency is retained about the same order of the magnitude as IRSIM and the delay calculation accuracy is improved to within 10% to 15% of SPICE.

The following section gives a brief review about the delay modeling history.

1.1 Delay Models Review

The usual definition of delays is the time required for the signal response to reach half its final value. Elmore [1] pointed out that although this definition is useful in the laboratory, it is extremely awkward for making computations. Instead, he suggested that it is reasonable to measure the delay, T_D to the centric of area of the curve $e'(t)$, that is,

$$T_D = \int_0^\infty te'(t)dt \tag{1.1}$$

where $e(t)$ is the output waveform and

$$\int_0^\infty e'(t)dt = 1 \tag{1.2}$$

Definition in Eq. (1.1) is in fact the first moment of the impulse response. The Elmore delay provides a single value for the delay estimation. The Elmore delay, or the first moment of the impulse response, is a good approximation for the dominant time constant of the step response. However, such estimation does not consider the logic thresholds of actual MOS device. To do so require finding an approximating response waveform and determining the time at which the logic threshold is crossed [2].

The first analytical delay model is proposed by [3]. The output waveform of a step input feeding into an inverter is obtained by solving the first order differential equations. A closed form expressions for the rise time and fall time are also provided. It is noticed that the transient response improves as the amount of current available from the transistors increases. To get a better feeling for the actual time involved, a characteristic time constant of the device is introduced which is a function of transistor geometry and the load capacitance. By observing the plot where the pair delay is plotted as a function of the threshold parameters, an approximate solution for the delay is given by

$$T_D \approx 0.9\tau\left(\frac{1}{(1-\alpha_N)^2}+\frac{1}{\beta(1-\alpha_P)^2}\right) \tag{1.3}$$

In Eq. (1.3), delay is proportional to the time constant τ and is related to the maximum values of normalized currents for NMOS and PMOS respectively.

Delay model (1.3) is intuitively correct in that the delay must depend on the response of two stages, one of which is primarily determined by the N transistor, the other by the P transistor. If one of the transistors is inherently much faster (higher current), the performance of the circuit is determined by the slower of the two transistors. The delay model is derived by an analytical methods rather a curve fitting. Thus, it is possible to generalize the delay model to more complicated circuits. And a closed form solution makes it possible to include the model into a simulator. However, this model does not consider the short circuit current and it does not include the input transition time.

Circuit simulator, RSIM [4], uses a simple lumped RC model. In the lumped RC model, the delay through a stage is computed by lumping all of the resistances and capacitances together using the product as the delay, where the delay at node n is estimated to be:

$$\tau = \left(\sum_{k=1}^{n} R_k\right)\left(\sum_{k=1}^{n} C_k\right) \tag{1.4}$$

The resistances are computed as the summation of the resistance for each transistor in the stage and the parasitic resistances in each node. The capacitances are of those between the switching transistor and the output node. The lumped RC model is also included in circuit simulator Crystal [5] as one of the three delay models. However, there are two sources of error in the RC model: One is the lumping of resistances and capacitances.

This tends to overestimate the delays since it assumes that all capacitances must discharge through all resistances. The second and the most significant one comes from its inability to deal with waveform shape. On average, the lumped RC model can usually estimate the delay through a path to within 25 percent of SPICE.

Reference [6] defines three quantities with dimensions of time:

$$T_P = \sum_k R_{kk} C_k \tag{1.5}$$

$$T_{Di} = \sum_k R_{ki} C_k \tag{1.6}$$

$$T_{Ri} = \left(\sum_k R_{ki}^2 C_k \right) / R_{ii} \tag{1.7}$$

None of the quantities above is equal to delay although Eq. (1.6) is what is called "delay" by Elmore. Upper and lower bounds are given by the quantities defined above for output voltages, or, equivalently, to find upper and lower bounds for the delay associated with each output. From calculation, the upper bounds for the unit step response are:

$$V_i(t) \le 1 - \frac{T_{Di} - t}{T_P} \tag{1.8}$$

$$V_i(t) \le 1 - \frac{T_{Ri}}{T_P} e^{(T_{Di} - T_{Ri})/T_{Ri}} e^{-t/T_{Ri}} \tag{1.9}$$

and the lower bounds for the unit step response are:

$$V_i(t) \ge 0 \tag{1.10}$$

$$V_i(t) \ge 1 - \frac{T_{Di}}{t + T_{R_i}} \tag{1.11}$$

$$V_i(t) \ge 1 - \frac{T_{Di}}{T_P} e^{(T_P - T_{Ri})/T_P} e^{-t/T_P} \tag{1.12}$$

So instead of solving the exact output waveform, which is difficult in most cases, the bound results can be used 1) to bound the delay, given the signal threshold, or 2) to bound the signal voltage, given a delay time, or 3) certify that a circuit is “fast enough,” given both the maximum delay and the voltage threshold [2]. However, in this model only step inputs are illustrated and initial conditions are assumed to be zero. Limits exist when analyzing the interconnections, in particular to include pass transistors in the interconnections. One pole model is incapable of evaluation the delay for non-monotone input response waveforms. Moreover, although the waveform bounding is theoretically interesting, it is difficult to be adopted in circuit simulators.

IRSIM is a widely used switch-level simulator. It is developed based upon RSIM and uses the delay model proposed by [7], [8] and an improved algorithm [9]. IRSIM uses a single time constant model to find the delay, which is equal to the Elmore delay. Not all of the output waveforms can be modeled successfully with a single time constant due to charge sharing. For example, a circuit in Fig. 1.1 shows a simple RC tree where node a, b, and c are charged high initially. When the switch closes, the voltage at node b initially falls at its own rate; however, the rate of decay is eventually controlled by the dominant time constant caused by the capacitor at node c. The circuit has a low frequency pole-zero pair in its frequency response and the low frequency zero partially cancels the dominant pole. This causes the output to have a two-time-constant behavior.

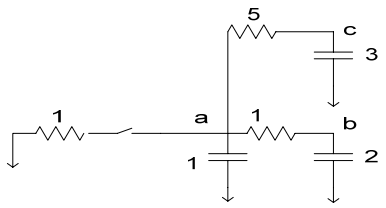

Fig.1. 1 A network with a low-frequency pole-zero pair

In [10] a two time constants model is proposed for this case, one to model the decay of the initial transient and the other to model the slow decay of the output tail. The second order estimate has three parameters: the time constants of the two poles and the one zero. These time constants can be related to three characteristic time constants of the output: the sum of the open circuit time constants, the first moment of the impulse response, and the second moment of the impulse response. The time constants can be uniquely determined by matching the coefficients of the first three terms in the node's network transfer function and the moments of the real waveform.

IRSIM groups charge sharing into two different categories: pure charge sharing and charge sharing with a driven path. Pure charge sharing refers to two non-driven RC subnets that are connected through a switch. Charge sharing with a driven path is the same as pure charge sharing except that a resistive path to a driving source exists for one of the subnets. The pure charge sharing problem was modeled in a way similar to a single time constant model since the transient waveform is usually dominated by a single pole. And the result turns out to be a degenerate version of the standard two-pole-one-zero model with one pole is located at the origin. For charge sharing with a driven path, the similar matching method as [10] is used but focuses on the two-pole-one-zero-at-the-origin system. Delay is meaningless for nodes in driving trees because the initial and final voltage will be the same (grounded). Peak amplitude of the output voltage (voltage spike) is what is interested and can be expressed by two time constants model. The charge sharing models have been successfully fulfilled in the circuit simulator IRSIM. The delay model in IRSIM pays particular attention to the charge sharing problems in digital VLSI circuits which are either ignored or incorrectly computed in most simulators. But the

delay model ignores the signal input transition time which also has a significant impact on the circuit performance evaluation. And other important issues for modern VLSI technology, like short channel effects and velocity saturation, are not covered in the model.

Delay models presented above are limited to the step input waveform only. These delay models are generally insufficient since they do not consider a realistic input waveform and consequently do not take into account the influence of the input waveform on the propagation delay. In a typical CMOS circuit application, the input-dependent propagation delay may well account for as much as 50-100 percent of the total delay.

Reference [11] is the first to introduce the input transition time into the delay modeling of a CMOS inverter. The input is assumed to be a voltage ramp. To calculate an analytical expression for the output ramp response, the same differential equations as in the case of an input step voltage [12] are used. The closed form analytical delay expressions for fast input transitions and slow input transitions are derived.

$$t_{dHL} = A_N \left(\frac{C_L}{k_N}\right)_n + \frac{B_N}{\beta \frac{\mu_p}{\mu_n}} \left(\frac{C_L}{k_N}\right)_{n-1} \tag{1.13}$$

$$t_{dLH} = \frac{A_p}{\beta \frac{\mu_p}{\mu_n}} \left(\frac{C_L}{k_N}\right)_n + B_p \left(\frac{C_L}{k_N}\right)_{n-1} \tag{1.14}$$

where B_N and B_P are dependent on the input waveform.

This model yields a better understanding of the switching behavior of the CMOS inverter than the step response model by considering the slope of the input waveform. Essentially, the propagation delay is shown to be the sum of the step response delay and an input dependent delay. The matching between the ramp input and the characteristic

input waveforms is shown to be easily performed for excellent agreement in output response and propagation delay. However, this model neglects the influence of the short circuit current by assuming that the "short-circuiting" transistor is weak compared to the charging/discharging transistor and did not describe the delay when the weak transistor is the charging/discharging transistor [13]. Therefore only one of the rise and fall delays can be modeled when the inverter is asymmetric. Only for the limited case of a symmetric inverter where both transistors have equal current driving capability, are both the rise and fall delays modeled. This delay model also neglects the voltage-current characteristics of the short-channel MOSFETs, mainly because it does not include the velocity saturation effects of carriers, which become eminent in the sub-micrometer regime. Consequently, this model is not satisfactory when applied to short-channel MOSFET circuits.

The delay model in [11] is based on Shockley's MOSFET model [14]. Since Shockley's transistor model is simple, many formulas have been derived based on this model and the derived formulas are used quite frequently in VLSI initial designs and CAD programs. However, as discussed in the previous section, Shockley's MOSFET model is not suitable for short channel devices. Sakurai et al [15] proposed a new MOSFET model, α-power law model, which is simple enough to be applied to the analytical treatments of the MOS circuits but includes the velocity saturation effects.

α-power law model is based on four parameters: V_{TH} (threshold voltage), α (velocity saturation index), V_{D0} (drain saturation voltage at $V_{GS} = V_{DD}$), and I_{D0} (drain current at $V_{GS} = V_{DS} = V_{DD}$). Better agreement is observed in the saturation region than the Shockley model. By using the α-power law model, the delay of a CMOS inverter is derived.

$$t_{pHL,}t_{pLH} = (\frac{1}{2} - \frac{1-v_T}{1+\alpha})t_T + \frac{C_L V_{DD}}{2I_{D0}} \quad (1.15)$$

The detailed description of the parameters in Eq. (1.15) can be found in [15]. The delay model is a linear combination of two terms. The first term is the input waveform dependent term, which is proportional to the input waveform transition time, and the second term is the output capacitance dependent term, which is proportional to the output capacitance. Sakurai gives an intuitive understanding of the inverter operation and derives closed form equations for evaluating the propagation delay and the output rise and fall time. However, these equations are valid only if the input slope exceeds one-third of the output slope, i.e., when the inverter are reasonable loaded and when the input waveforms are fast-rising ramps. For relatively slow inputs and/or very low fanouts, Sakurai's assumption ceases to be valid, thereby leading to inaccuracies. And short circuit current issue is not attacked in Sakurai's delay model.

References [16] and [17] extend Sakurai's work on delay modeling of inverters to slow input ramps. Two different mechanisms, which can be adequately modeled analytically, govern the delay and the output transition time of an inverter in two extreme cases: infinitely fast and infinitely slow inputs. These extreme points are joined by a curve that can predict the delay and the output transition time for any input.

$$t_d = (-1 + \frac{a}{\tau^b}) \times tr_{in} \times D_{slope} \quad (1.16)$$

where $\tau = tr_{in} / fo$ and D_{slope} is the slope of the dc asymptote that the delay curve follows at infinite input rise times. Curve fitting method is used to get the delay and output transition time equation. Thus, the delay model may not be suitable to be used as a general equation to calculate the delay. In [17], an analytical delay model is presented for

an inverter with fast and slow inputs. But this delay model expression is very complicated and is not easy to be adopted into a simulation tool. Further more, a lot of preliminary work need to be done to find the appropriate set of parameters through simulations and curve fitting methods. Again, neither model includes the short circuit current which is important when the input transition time is slow.

In [18] the slope model is suggested. The slope model incorporates information about waveform shape in order to make more accurate delay estimates. The key to the slope model is that it includes the input rise-time, the output load, and the transistor size into a single factor, rise-time ratio, which determines the transistor's effective resistance. First, the output load and transistor size are combined into a single value called the intrinsic rise-time of the stage; this is the rise time that would occur at the output if the input were driven by a step function. The input rise-time of a stage is then divided by the intrinsic rise-time of the stage's output to produce the rise time ratio for the stage. The rise-time ratio gives an estimate of how fully turned-on the trigger transistor is when it is doing its work. SPICE simulations showed that the rise-time ratio is an accurate predictor of the effective resistance of a transistor, independent of the specific input rise-time, transistor size, or output load. Reference [19] combines slope model and Penfield-Rubenstein delay model into Penfield-Rubenstein slope model. In this model, instead of using the waveform boundary definitions, the average of the upper and lower bounds is used and is simplified to Elmore delay:

$$delay = \sum_i R_i C_i \tag{1.17}$$

This means that each separate capacitor is weighted only by the resistance between it and the signal source. This is an improvement comparing the delay model

used in RSIM, which weights all the capacitance by all the resistance. Computation results by Crystal [20] using the Penfield-Rubenstein slope model show that the average error for a single stage is 20%, and the average error over paths containing several stages is 6%.

Reference [21] presented a piecewise expression for the propagation delay based on BSIM MOSFET model by neglecting the short circuit current. In [22], a piecewise solution is developed with seven operation regions for the transient response of a CMOS inverter based upon the α-power law model. However, no general circuitry model is proposed.

In this dissertation, we propose a piecewise linear delay model which can evaluate the propagation delay of a circuit with a wide range of input slope. The model takes into account the influences of short circuit current and dynamic channel charges. This model is applicable not only to propagation delay calculation of a simple gate, such as an inverter, but also to that of any general circuit topology.

1.2 Organization

The next chapter describes the MOSFET device models used in the piecewise linear delay model. A simple piecewise linear current model which includes transistor saturation operation and a finite input transition time is proposed. A piecewise linear approximation of the parasitic channel transistor capacitances model is presented. The charge stored in the source/drain diffusions is also non-linear. We linearized it by taking the average of the BSIM3 diffusion capacitance model.

Chapter 3 discusses the way to group the circuit into different resistance connected regions to reduce the circuit complexity. General RC circuit dynamic equations are given in each piecewise linear region. Quasi-steady state solutions are used to derive a general solution of the output waveform at any nodes. The time constant in each piecewise linear region is derived based on Elmore delay. The computation of boundary points between different operation regions using Newton-Raphson method is discussed in this chapter.

The piecewise linear delay model is applied to the inverter analysis in Chapter 4. Details are given for a rising input transitions. Various inverter circuit configurations are used to calibrate and verify the delay model. Comparison results of SPICE simulations and delay model computations are shown graphically.

Chapter 5 extends the model application to a two-input NAND gate and an OAI gate. Extensive comparisons are given for different input slopes and output loading factors.

Finally, Chapter 6 summarizes the contributions of this dissertation and describes areas for future efforts.

CHAPTER II

PIECEWISE LINEAR TRANSISTOR MODEL

Device modeling plays a very important role in the circuit performance evaluation. The accuracy of circuit simulators is mainly determined by the accuracy of the device models since the simulation algorithms and convergence techniques in circuit simulators have become mature. Rohrer [23] argued that device models, not internal algorithms, were the keys to the success of a circuit simulation program. A criterion to evaluate a circuit simulator is its simulation speed and accuracy. Thus, a simple and but accurate enough device model is desired in the circuit simulation.

BSIM3 [24] device model used in the SPICE simulator covers the important physical effects in MOSFETs. It tries to illustrate all of the physical effects of MOSFET in a mathematical way. However, BSIM3 is too complicated to be used in a fast simulator. A simple piecewise linear device model [25] is proposed in this dissertation for transistors operating in different piecewise linear regions. This model is derived based upon the physical meanings of the transistor device. Simplified BSIM3 equations are used in modeling the source/drain diffusion capacitances and the channel storage charges.

2.1 Piecewise Linear Current Model

2.1.1 Current Model

A popular technique to approximate the transistors is a voltage controlled switch with a series resistance. In switch model, a switching off device is represented by an open circuit and a switching on device is rendered as a single resistor in series connected with a voltage controlled switch. The switch model is shown in Fig. 2.1.

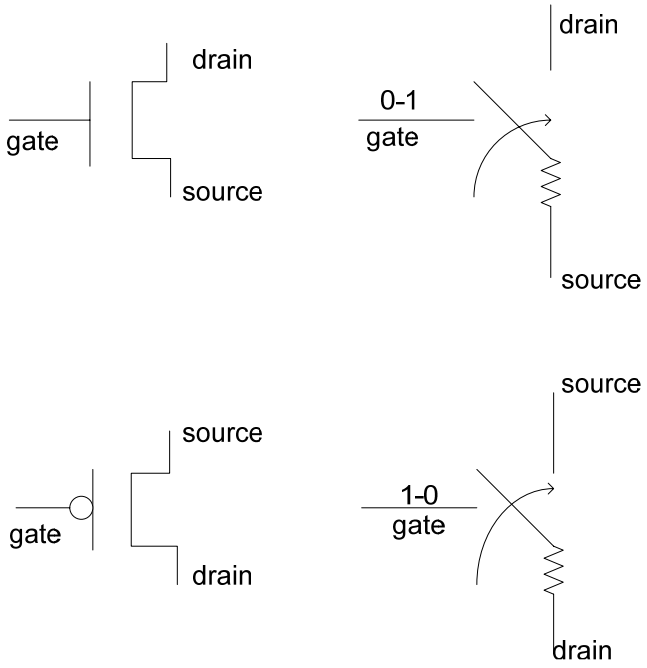

Fig.2. 1 Transistor switch model

This allows simple linear models to be used for the circuit dynamics once the switches have reached their final states. The switch model has been used very successfully in circuit simulation and performance optimization. However, switch model is not adequate enough for circuit simulation since it does not consider the transition time needed to switch a device on and off. Moreover, saturation behavior of the transistors has a significant impact on the switching waveforms. Unfortunately, there is no linear model that can include cutoff, ohmic, and saturation behaviors together. A simple piece-wise linear model is proposed as shown in Fig. 2.2.

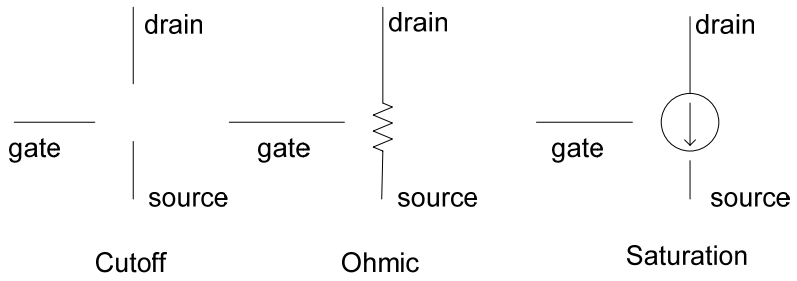

Fig.2. 2 Piecewise linear current model

An open switch is used to represent the transistor in the cutoff region. When the transistor operates in the ohmic region, it is modeled as a linear resistor whose conductance is a function of the transistor's type and size. A current source that is a function of the input voltage is used to model the transistor in the saturation region. For NMOS transistors, the drain current flows from the drain to the source. The current flows from the source to the drain instead in a PMOS transistor. The mathematical expressions of the piece-wise linear model are shown in (2.1) and (2.2):

$$I_{Dn} = \begin{cases} 0 & V_{GS} < V_{Tn} & \textit{Cutoff} \\ a_n G_n V_{DS} & V_{GS} > V_{Tn}, V_{DS} < V_{DSsatn} & \textit{Ohmic} \\ G_n (V_{GS} - V_{Tn}) & V_{GS} > V_{Tn}, V_{DS} > V_{DSsatn} & \textit{Saturation} \end{cases} \tag{2.1}$$

$$I_{Dp} = \begin{cases} 0 & V_{GS} > V_{Tp} & \textit{Cutoff} \\ a_p G_p V_{DS} & V_{GS} < V_{Tp}, V_{DS} > V_{DSsatp} & \textit{Ohmic} \\ G_p (V_{GS} - V_{Tp}) & V_{GS} < V_{Tp}, V_{DS} < V_{DSsatp} & \textit{Saturation} \end{cases} \tag{2.2}$$

where

$$V_{DSsatn} = \frac{V_{GS} - V_{Tn}}{a_n} \tag{2.3}$$

$$V_{DSsatp} = \frac{V_{GS} - V_{Tp}}{a_p} \tag{2.4}$$

All parameters except V_{Tp} are positive, and V_{Tp} is negative. V_{Tn} and V_{Tp} are the threshold voltages for NMOS and PMOS respectively. a_n, a_p, V_{Tn}, and V_{Tp} are process constants. G_n and G_p are the transistor conductance parameters and are determined by the transistor size and can be different for each transistor. V_{DSsatn} and V_{DSsatp} are the drain to source saturation boundary voltages, which determine when the NMOS transistor and the PMOS transistor go to saturation region and are functions of the input voltage.

Fig. 2.3 and 2.4 are the $V_{DS} - I_D$ characteristics of the piecewise linear model for a NMOS and a PMOS transistor of width $2.4um$ and length $0.6um$ using American Microsystems Inc. (AMI) 0.6 micron CMOS technology. The corresponding SPICE simulation results are overlapped in the same figures.

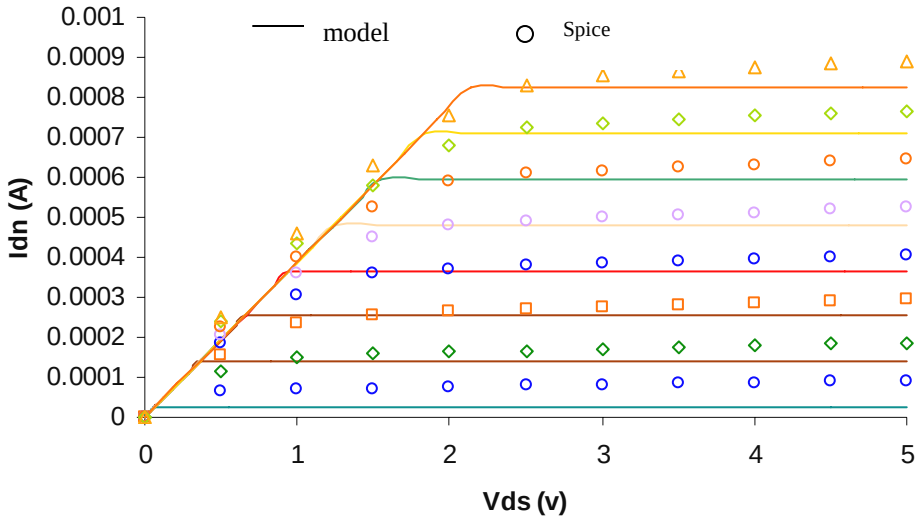

Fig.2. 3 NMOS FET $V_{DS} - I_{Dn}$ characteristics ($W_n = 2.4um$ $L = 0.6um$).

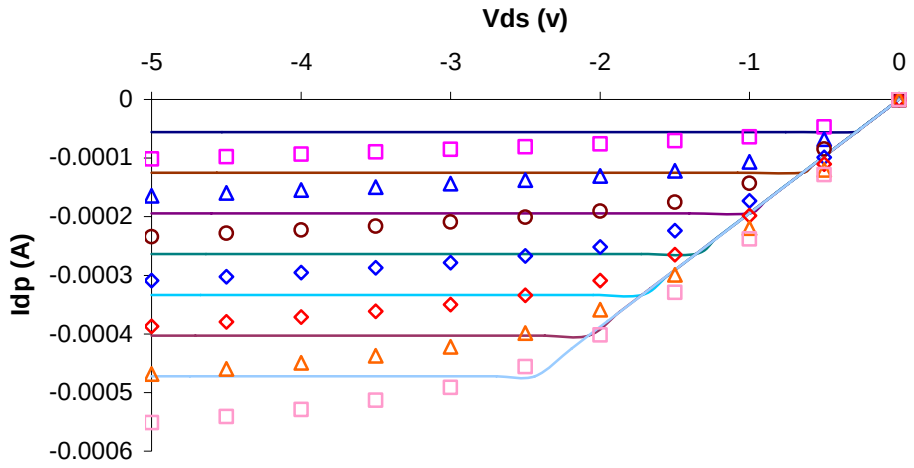

Fig.2. 4 PMOS FET $V_{DS} - I_{Dp}$ characteristics ($Wp = 4.8um\ L = 0.6um$)

The figures above show that the piecewise linear current model is a good approximation to the drain current of sub-micron devices in the saturation region. In the saturation region, the transistor is modeled as an ideal current source which is independent of the drain to source voltage. The channel length modulation effect and body effect are ignored for simplicity purpose. However, the model is not very accurate in the ohmic region. Because in the piecewise linear model, a simple resistor is assumed when the transistor is operating in the linear region. But this is not true in reality. When V_{DS} is small, the inversion channel behaves like a simple resistor. The drain current I_D increases linearly as the drain voltage V_{DS} increases. However, when V_{DS} is larger it will cause an increase of the voltage in the inversion layer at all points along the channel except the source. This reduces the voltage across the gate capacitor and the inversion charge density is reduced. The smaller amount of mobile inversion charges results in a decrease in channel conductance, which leads to a smaller slope in the $I_D - V_{DS}$

characteristics as V_{DS} increases. Fig. 2.4 also shows that the model parameters chosen do not fit the SPICE simulation very well. From the later chapters, we can see that the piecewise linear model is not sensitive to the value of the parameters, which makes the model robust. Better fitting can be achieved by: 1. be more carefully when choosing the model parameters. 2. adding more piecewise linear regions to better approximate the transistor current. Adding more regions to the model makes solving for the circuit dynamics more difficult because the node voltages must be checked at each moment in time to decide which piecewise model to use for each transistor in the circuit. BSIM3v3 uses a much more complicated linear region current model which is shown in Eq. (2.5):

$$I_{ds} = \mu_{eff} C_{ox} \frac{W_{eff}}{L_{eff}} \frac{1}{1 + V_{ds} / E_{sat} L_{eff}} (V_{gs} - V_{th} - A_{bulk} V_{ds} / 2) V_{ds} \tag{2.5}$$

Although the BSIM model is more accurate in the ohmic region than the piecewise linear model, here the accuracy is traded off by the speed and computation complexity consideration. The piecewise linear current model is at least more accurate than the traditional switched resistor model.

2.1.2 Model Parameter Extraction

The piecewise linear current model has total of six parameters: a_n, a_p, V_{Tn}, V_{Tp}, G_n, and G_p. They can be extracted directly from the SPICE $I_D - V_{GS}$ and $I_D - V_{DS}$ plots.

Fig. 2.5 gives a graphical illustration of the extraction of V_{Tn} and V_{Tp} from the SPICE $I_D - V_{GS}$ plot. A series of curves can be drawn in the $I_D - V_{GS}$ coordinates by sweeping drain to source voltage V_{DS}. The tangent lines of each of theses curves crosses

the axis of V_{GS} at V_{T1}, V_{T2} , V_{T3},....The threshold voltage is computed by taking the average of the intersection points:

$$V_T = \frac{V_{T1} + V_{T2} + \ldots + V_{Tn}}{n} \tag{2.6}$$

where n is the number of curves in the $I_D - V_{GS}$ plot.

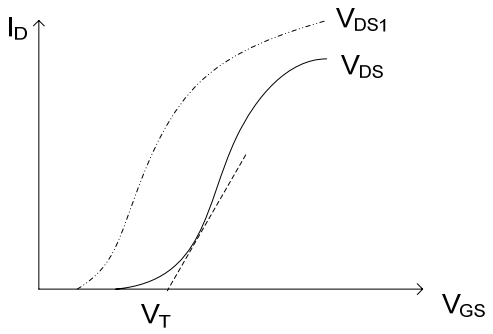

Fig.2. 5 Model parameters extraction

According to Eq. (2.1) and (2.2), the slope of the $I_D - V_{GS}$ curve is the transconductance of the transistor in the saturation region. Similar to the way to find the threshold voltage, transistor transconductance in saturation region can be obtained by taking the average of the slopes of the tangent lines.

$$G_{n(p)sat} = \frac{Slope_1 + Slope_2 + \ldots + Slope_n}{n} \tag{2.7}$$

For transistor operating in the ohmic region, its conductance can be found from SPICE $I_D - V_{DS}$ plot. By estimating the slope of each curve in the low V_{DS} region and taking the average of them, we can find $G_{n(p)ohm}$. Then the value of a_n and a_p can be easily decided by:

$$a_{n(p)} = \frac{G_{n(p)ohm}}{G_{n(p)sat}} \tag{2.8}$$

Appendix A lists the parameters extracted from the SPICE simulations for AMI $0.6um$ using the method illustrated above.

2.1.3 Effective Channel Length and Width

Electrical channel length and width of a MOSFET are different from the drawn channel length L_{drawn} and width W_{drawn} because of processing related reasons. As a result, effective channel length L_{eff} and effective channel width W_{eff} are used instead. In BSIM3v3, the definition of W_{eff} is:

$$W_{eff} = W_{drawn} - 2dW \tag{2.9}$$

where

$$dW = W_{INT} + \frac{W_L}{L^{W_{LN}}} + \frac{W_W}{L^{W_{WN}}} + \frac{W_{WL}}{L^{W_{LN}} W^{W_{WN}}} \tag{2.10}$$

W_{INT} is a parameter extracted from experimental results. W_L, W_W, W_{LN}, W_{WN}, and W_{WL} are additional fitting parameters available to improve the model accuracy. The piecewise linear model does not include the effects of those additional fitting parameters. In the proposed model, W_{eff} is approximated as:

$$W_{eff} = W_{drawn} - 2W_{INT} \tag{2.11}$$

In AMI $0.6um$ technology, $W_{INT} = 2.0 \times 10^{-7}$ for NMOS transistor and $W_{INT} = 2.6 \times 10^{-7}$ for PMOS transistor.

The BSIM3 model shows that I_D is approximately proportional to $\frac{W_{eff}}{A_{bulk}}$. If the short channel effect is ignored and plugging in the corresponding BSIM3 parameters, we found the following empirical transistor width dependences in computing I_D :

$$W'_{neff} = W_{neff} \frac{W_{neff} + 1.41\times10^{-6}}{W_{neff} + 1.79\times10^{-6}} \tag{2.12}$$

$$W'_{peff} = W_{peff} \frac{W_{peff} + 5\times10^{-6}}{W_{peff} + 5.51\times10^{-6}} \tag{2.13}$$

Above equations are used to calculate the transistor conductance for a given sheet resistance.

Similarly, effective channel length is model as:

$$L_{eff} = L_{drawn} - 2L_{INT} \tag{2.14}$$

In AMI $0.6um$ technology, $L_{INT} = 3.034496\times10^{-8}$ for NMOS transistor and $L_{INT} = 7.205014\times10^{-8}$ for PMOS transistor.

2.1.4 Transistor Sheet Resistance

Transistor conductance $G_{n(p)ohm}$ found in section 2.1.2 is dependent on the transistor width and length. Although a unique conductance parameter for each transistor will make the piecewise linear model have a better agreement with SPICE, this dependency makes the model implementation much more complicated. Instead, four transistor sheet resistance parameters represent PMOS and NMOS transistor sheet resistances of different transition state (rising or falling), R_{snr} , R_{snf} , R_{spr} , and R_{spf} are

used. Thus, transistor conductance can be expressed as proportionally related to the effective channel width and inversely proportional to effective channel length:

$$G_{n(p)sat} = \frac{1}{R_s}\frac{W'_{eff}}{L_{eff}} \tag{2.15}$$

where R_s represents R_{snr}, R_{snf}, R_{spr}, and R_{spf}, rise and fall sheet resistances of NMOS and PMOS respectively.

Fig. 2.6 and 2.7 compare the conductance found in section 2.1.2 from $I_D - V_{GS}$ characteristic curves and the method provided by Eq. (2.15). The linear approximation using sheet resistances shows a great agreement with the $I_D - V_{GS}$ curve extraction results for a wide range of the transistor widths for a rising input case. However, in Fig. 2.7, it shows some discrepancies for NMOS transistors with a falling input. The errors exist mainly around the NMOS transistor width of $4.8um$. The errors very much possibly come from the human errors. Because the conductance found in section 2.1.2 are through the manually drawn tangent lines of the $I_D - V_{GS}$ curve, which is inevitable to introduce some errors into the final results. However, the sheet resistance method is a good approximation overall. The errors like those in Fig. 2.7 can be calibrated and corrected by running more simple circuit simulations.

In the piecewise linear delay model, the following sheet resistances are used to calculate the transistor conductance:

TABLE 2. 1 Transistor Sheet Resistance.

Transistor	Fall ($\Omega/\square$)	Rise($\Omega/\square$)
NMOS	1.7×10^4	1.3×10^4

PMOS	2.1×10^4	3.3×10^4

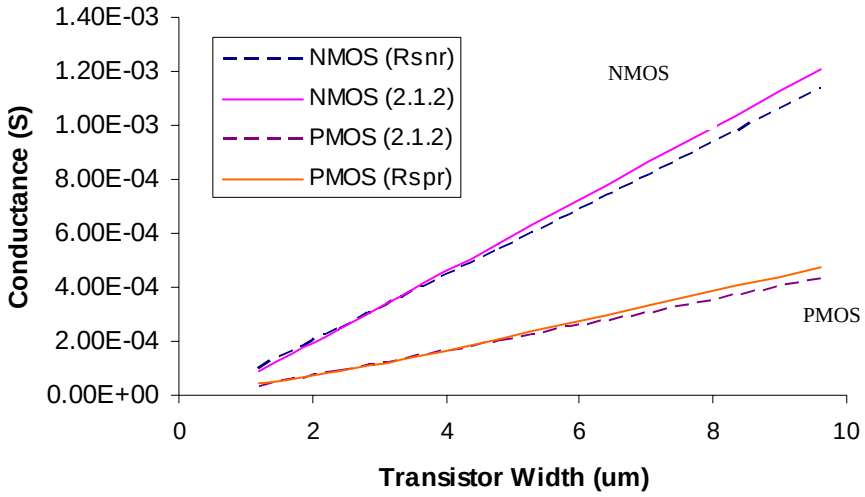

Fig.2. 6 Transistor conductance for transistor with rising input.

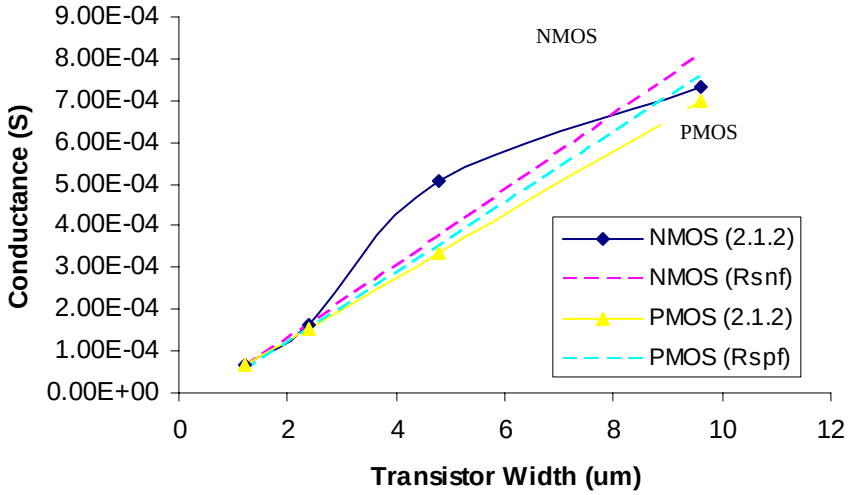

Fig.2. 7 Transistor conductance transistor with falling input.

2.2 Channel Charge Storage Model

The charge stored in the transistor channel is a non-linear function of the transistor terminal voltages. The channel charge is partitioned between the source and drain to obtain a simple lumped parameter model for the dynamic behavior of the transistor. The total charges on both side of the gate oxide are neutral. This requires,

$$Q_G + Q_S + Q_D + Q_B = 0 \tag{2.16}$$

where Q_G is the charge stored on the gate, Q_S and Q_D are the mobile carrier charges in the inversion channel, and Q_B is the bulk charge in the depletion layer under the channel. It should be emphasized that these are not normal capacitances since the charges are functions of all of the transistor terminal voltages (not just a pair of them). The capacitive currents can be written as:

$$i_G = \frac{dQ_G}{dt} = C_{GS}\frac{dV_{GS}}{dt} + C_{GD}\frac{dV_{GD}}{dt} + C_{GB}\frac{dV_{GB}}{dt} \tag{2.17a}$$

$$i_S = \frac{dQ_S}{dt} = C_{SG}\frac{dV_{SG}}{dt} + C_{SD}\frac{dV_{SD}}{dt} + C_{SB}\frac{dV_{SB}}{dt} \tag{2.17b}$$

$$i_D = \frac{dQ_D}{dt} = C_{DG}\frac{dV_{DG}}{dt} + C_{DS}\frac{dV_{DS}}{dt} + C_{DB}\frac{dV_{DB}}{dt} \tag{2.17c}$$

$$i_B = -(i_G + i_S + i_D) \tag{2.17d}$$

where C_{ij} are all independent non-linear functions of the terminal voltages and are defined by:

$$C_{ij} = \frac{\partial Q_i}{\partial V_{ij}} \quad i, j = G, D, S, B \tag{2.18}$$

In digital applications, the substrate terminal is biased at a constant voltage so that the derivative of the bulk voltage can be ignored during analysis.

Johnson [25] proposed a piecewise linear approximation of the BSIM3 channel storage charge model. Since it uses the same definitions of V_T and V_{DSsat} , the model has the same regions of validity as the piecewise linear current model. The channel storage charge models in different transistor operation regions are:

Ohmic region:

$$Q_{Gohm} = C_{ox}(V_{GS} - V_T) - (1-b)aC_{ox}V_{DS} + bC_{ox}(V_{SB} + V_T - V_{FB}) \tag{2.19a}$$

$$-Q_{Sohm} = \frac{1}{2}C_{ox}(V_{GS} - V_T) + a(b - \frac{1}{2} - x_{part}b)C_{ox}V_{DS} \tag{2.19b}$$

$$-Q_{Dohm} = \frac{1}{2}C_{ox}(V_{GS} - V_T) - a(\frac{1}{2} - x_{part}b)C_{ox}V_{DS} \tag{2.19c}$$

Saturation region:

$$Q_{Gsat} = bC_{ox}(V_{GB} - V_{FB}) \tag{2.20a}$$

$$-Q_{Ssat} = (1 - x_{part})bC_{ox}(V_{GS} - V_T) \tag{2.20b}$$

$$-Q_{Dsat} = x_{part}bC_{ox}(V_{GS} - V_T) \tag{2.20c}$$

Cutoff region:

$$Q_{Goff} = bC_{ox}(V_{GB} - V_{FB}) \tag{2.21a}$$

$$-Q_{Soff} = 0 \tag{2.21b}$$

$$-Q_{Doff} = 0 \tag{2.21c}$$

Taking the derivatives of Eq. (2.19) to (2.21) with respect to t and comparing the charge storage currents to Eq. (2.17), gives the following linearized results for the transistor capacitances.

	Cutoff	Ohmic	Saturation
C_{GG}	$bC_{ox}+C_{GBO}$ $+C_{GDO}+C_{GSO}$	$C_{ox}+C_{GBO}+C_{GDO}+C_{GSO}$	$bC_{ox}+C_{GBO}$ $+C_{GDO}+C_{GSO}$
C_{GS}	C_{GSO}	$-(1-b)(a-1)C_{ox}+C_{GSO}$	C_{GSO}
C_{GD}	C_{GDO}	$(1-b)aC_{ox}+C_{GDO}$	C_{GDO}
C_{SG}	C_{GSO}	$\frac{1}{2}C_{ox}+C_{GSO}$	$(1-x_{part})bC_{ox}$ $+C_{GSO}$
C_{SS}	C_{GSO}	$[\frac{1}{2}+a(b-\frac{1}{2}-x_{part}b)]C_{ox}+C_{GSO}$	$(1-x_{part})bC_{ox}$ $+C_{GSO}$
C_{SD}	0	$a(b-\frac{1}{2}-x_{part}b)C_{ox}+C_{GSO}$	0
C_{DG}	C_{GDO}	$\frac{1}{2}C_{ox}+C_{GDO}$	$x_{part}bC_{ox}+C_{GDO}$
C_{DS}	0	$[-\frac{1}{2}+a(\frac{1}{2}-x_{part}b)]C_{ox}$	$-x_{part}bC_{ox}$
C_{DD}	C_{GDO}	$a(\frac{1}{2}-x_{part}b)C_{ox}+C_{GDO}$	C_{GDO}

TABLE 2. 2 Linearized Parasitic Transistor Capacitances with Overlap Capacitances

C_{GSO}, C_{GDO}, and C_{GBO} are gate to source, gate to drain, and gate to bulk overlap capacitances. As the transistor size shrinks, the overlap capacitances are becoming more significant in modern processes. The BSIM model for the overlap capacitance is

$$C_{GSO}=WC_{gso} \tag{2.22}$$

$$C_{GDO}=WC_{gdo} \tag{2.23}$$

$$C_{GBO}=2LC_{gbo} \tag{2.24}$$

where C_{gso}, C_{gdo}, and C_{gbo} are overlap capacitance per unit width of the MOS device.

The inversion channel charge is partitioned into the drain and source charges. There are three different charge partition schemes existing in BSIM3's charge-based models, 50/50, 40/60, and 0/100, which are distinguished in circuit simulation using a model parameter called x_{part}. x_{part} is equal to 0.5 in 50/50 partition, is less than 0.5 in 40/60 partition, and is greater than 0.5 in 0/100 partition. In the piecewise linear model, x_{part} is chosen as a fitting parameter.

A more detailed derivation of capacitance in TABLE 2.2 can be found in Appendix B.

2.3 Source/Drain Diffusion Capacitance Model

As shown in Fig. 2.8, source/drain junction capacitance (normalized to per unit length) is composed of three components: the bottom junction capacitance C_{jbs}, the sidewall periphery junction capacitance of the field oxide edge C_{jbssw}, and the gate-edge periphery junction capacitance C_{jbsswg}.

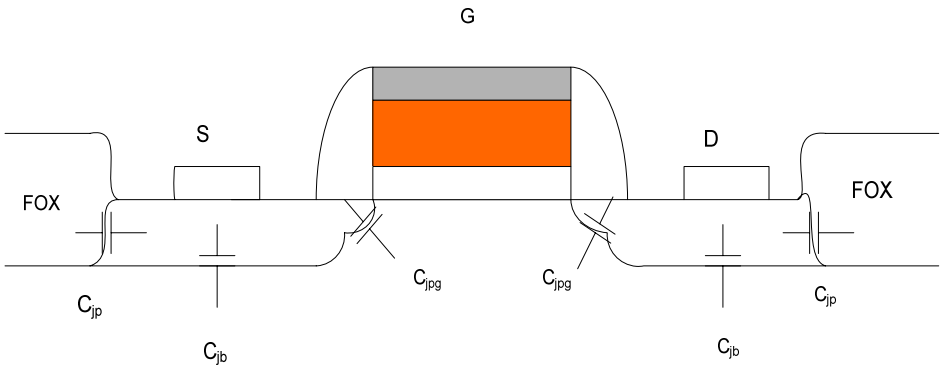

Fig.2. 8 Capacitance components of the source/drain junctions

2.3.1 Source/Drain Diffusion Capacitance Model

The BSIM model for the source diffusion capacitance is

$$\begin{aligned} C_{dS} &= A_S C_{jbs} + W_{eff} C_{jbsswg} + (P_S - W_{eff}) C_{jbssw} \\ &= A_S C_j (1 + \frac{V_{SB}}{\phi_B})^{-m_j} + W_{eff} C_{jswg} (1 + \frac{V_{SB}}{\phi_{Bswg}})^{-mjswg} + (P_S - W_{eff}) C_{jsw} (1 + \frac{V_{SB}}{\phi_{Bsw}})^{-m_{jsw}} \end{aligned} \quad (2.25)$$

where C_j is the unit area bottom capacitance at the zero bias, ϕ_B is the built-in potential of the bottom junction, m_j is capacitance grading coefficient of the bottom junction, C_{jsw} is the unit length periphery capacitance at the field oxide edge at zero bias, ϕ_{Bsw} is the built-in potential of the sidewall junction at the field oxide edge, m_{jsw} is the capacitance grading coefficient of the sidewall junction at the field oxide edge, C_{jswg} is the unit length periphery capacitance at the gate edge at zero bias, ϕ_{Bswg} is the built-in potential of the sidewall junction at the gate edge, and m_{jswg} is capacitance grading coefficient of the sidewall junction at the gate edge.

And similarly the computation of drain diffusion capacitance just simply replace *S* with *D*. It will be convenient to model these capacitances as non-linear charge storage elements as we did for the channel charge. The charge is partitioned corresponding to each of the three terms above.

$$Q_{dS} = Q_{jS} + Q_{jswS} + Q_{jswgS} \quad (2.26)$$

where

$$\frac{dQ_{jS}}{dV_{SB}} = A_S C_j (1 + \frac{V_{SB}}{\phi_B})^{-m_j}$$

and so forth. Assume

$$Q_{jS}(-\phi_B) = 0$$

$$0 < m_j < 1$$

and similarly for the other charges, then

$$Q_{dS} = \frac{A_S C_j \phi_B}{1-m_j}(1+\frac{V_{SB}}{\phi_B})^{1-m_j} + \frac{W_{eff} C_{jswg} \phi_{Bswg}}{1-m_{jswg}}(1+\frac{V_{SB}}{\phi_{Bswg}})^{1-m_{jswg}} + \frac{(Ps-W_{eff})C_{jsw}\phi_{Bsw}}{(1-m_{jsw})}(1+\frac{V_{SB}}{\phi_{Bsw}})^{1-m_{jsw}} \tag{2.27}$$

The charge equation is computational expensive when m_j is not a multiple of 1/2. Detailed circuit simulator, like SPICE, can model this. In the piecewise linear model, we average the diffusion capacitance over the ranges $V_{SB} = -\phi_B$ to 0, $V_{SB} = 0$ to $V_{DD}/2$, and $V_{SB} = V_{DD}/2$ to V_{DD}.

$$\bar{C}_{dS} = \begin{cases} \dfrac{Q_{dS}(0)}{\phi_B} & -\phi_B < V_{SB} < 0 \\ \dfrac{Q_{dS}(V_{DD}/2) - Q_{dS}(0)}{V_{DD}/2} & 0 < V_{SB} < V_{DD}/2 \\ \dfrac{Q_{dS}(V_{DD}) - Q_{dS}(V_{DD}/2)}{V_{DD}/2} & V_{DD}/2 < V_{SB} < V_{DD} \end{cases} \tag{2.28}$$

This is a simple solution because all we need to know is just the operation range of the transistor. The exact value of V_{SB} does not need for $\bar{C}_{dS}$ calculation. Thus, the average diffusion capacitance is a constant in each piecewise linear region. The overhead is that we have to learn the switching state of the transistor, i.e., rise input or fall input, rise output or fall output in order to decide which equation to use. The three components of $\bar{C}_{dS}$ can be calculated as follows:

$$\bar{C}_{dSj} = \frac{A_S C_j}{1-m_j} \begin{cases} 1 & -\phi_B < V_{SB} < 0 \\ \frac{2\phi_B}{V_{DD}} \left[(1+\frac{V_{DD}}{2\phi_B})^{1-m_j} - 1 \right] & 0 < V_{SB} < V_{DD}/2 \\ \frac{2\phi_B}{V_{DD}} \left[(1+\frac{V_{DD}}{\phi_B})^{1-m_j} - (1+\frac{V_{DD}}{2\phi_B})^{1-m_j} \right] & V_{DD}/2 < V_{SB} < V_{DD} \end{cases} \tag{2.29}$$

$$\bar{C}_{dSjsw} = \frac{(P_S - W) C_{jsw}}{1-m_{jsw}} \begin{cases} 1 & -\phi_{Bsw} < V_{SB} < 0 \\ \frac{2\phi_{Bsw}}{V_{DD}} \left[(1+\frac{V_{DD}}{2\phi_{Bsw}})^{1-m_{jsw}} - 1 \right] & 0 < V_{SB} < V_{DD}/2 \\ \frac{2\phi_{Bsw}}{V_{DD}} \left[(1+\frac{V_{DD}}{\phi_{Bsw}})^{1-m_{jsw}} - (1+\frac{V_{DD}}{2\phi_{Bsw}})^{1-m_{jsw}} \right] & V_{DD}/2 < V_{SB} < V_{DD} \end{cases}$$

$$\bar{C}_{dSjswg} = \frac{W C_{jswg}}{1-m_{jswg}} \begin{cases} 1 & -\phi_{Bswg} < V_{SB} < 0 \\ \frac{2\phi_{Bswg}}{V_{DD}} \left[(1+\frac{V_{DD}}{2\phi_{Bswg}})^{1-m_{jswg}} - 1 \right] & 0 < V_{SB} < V_{DD}/2 \\ \frac{2\phi_{Bswg}}{V_{DD}} \left[(1+\frac{V_{DD}}{\phi_{Bswg}})^{1-m_{jswg}} - (1+\frac{V_{DD}}{2\phi_{Bswg}})^{1-m_{jswg}} \right] & V_{DD}/2 < V_{SB} < V_{DD} \end{cases}$$

This can be explained by considering the first half transition of an inverter since for a delay prediction we are only interested in the output waveform up to the $V_{DD}/2$ point. For a rising input, V_{DB} of the PMOS transistor will vary from 0 to $V_{DD}/2$ and V_{DB} of NMOS will change from V_{DD} to $V_{DD}/2$. For a falling input, V_{DB} of the PMOS transistor will vary from V_{DD} to $V_{DD}/2$ and V_{DB} of the NMOS transistor will operate on the low range, from 0 to $V_{DD}/2$. Thus, different equations in Eq. (2.29) are used to attack the drain/source diffusion capacitance for different transition waveform. We generalize the above conclusions to any transistor drain/source diffusion capacitance computation by

making the assumption that there is no difference between the source and drain when calculating the junction capacitance to the bulk.

2.3.2 Transistor Geometry Parameters Selection

The BSIM3 model does not implement the calculation for A_S , A_D , P_S , and P_D inside the model. In order to have a fair comparison to the SPICE simulation, the geometry parameters used in the piecewise linear mode are extrapolated directly from the Cadence layout. Fig. 2.9 shows the geometry information used in Eq. (2.29) to calculate A_S , A_D , P_S , and P_D .

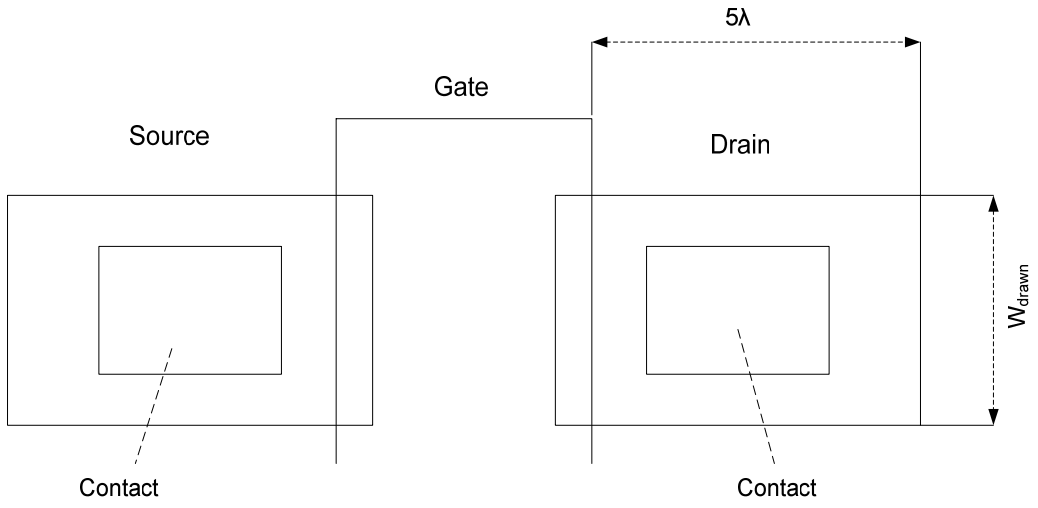

Fig.2. 9 A simple device layout plot showing geometry parameters

Below are equations used by the piecewise linear model to find the transistor geometry parameters. Same equations are also used in SPICE netlist file.

$$A_S = A_D = 5\lambda W_{drawn} \tag{2.30}$$

$$P_S = P_D = 5\lambda \times 2 + W_{drawn} \tag{2.31}$$

where W_{drawn} is the transistor drawn width in the Cadence layout plot; λ is the lambda-based design rules parameter, which characterizes the linear feature, the resolution of the complete wafer implementation process, and permits first-order scaling. For an AMI 0.6*um* process, lambda is equal to 0.3um.

2.4 Summary

A new device model, piecewise linear model, is proposed. From the drain current versus the drain-to-source voltage characteristic plot, it is shown that the piecewise linear current model has a good agreement with SPICE when the MOSFET is operating in the saturation region. But it shows discrepancy when the MOSFET is in the ohmic region. More accuracy can be obtained if adding more piece-wise linear regions, but this will greatly increase the complexity of the model.

Linearized transistor channel charge equations are used to model the charge stored in the transistor. This piece-wise linear model has the same regions of validity as the piece-wise linear current model since it uses the same definitions of V_T and V_{DSsat}. Linearized transistor parasitic capacitances are derived.

Source/drain diffusion capacitances are computed using the BSIM model. In stead of tracking the values of V_{SB} during the transition, the average diffusion capacitance over the variation of V_{SB} is found. The selection of the geometry parameters are also introduced in this chapter.

CHAPTER III

CIRCUIT EQUATIONS AND OUTPUT VOLTAGE APPROXIMATION

When implementing the piecewise linear delay model, a large circuit is divided into smaller sub-circuits according to different resistance connected regions. RC circuit dynamic equations in each sub-circuit are solved without considering the changes happened in other sub-circuits. Output delay is defined as the time difference between the mid-point of the output voltage waveform and the mid-point of the input voltage waveform. Since the model takes into account of the input transition time, it is more accurate than the IRSIM whose delay model simply assumes a step function as the input waveform.

3.1 Resistance Connected Regions

Modern digital logic circuits have millions or even billions of transistors on a single silicon chip. Circuit simulator, like SPICE, solves the whole circuit as a big matrix and gets the node outputs by solving for the current and voltage waveforms as function of time. Usually no more than a few thousand transistors may be simulated in a reasonable amount of computation time.

Instead of solving for all circuit node voltages at once, it is almost always possible to subdivide the problem into smaller pieces which are easier to solve. Define a resistive

connected region as a set of circuit nodes connected by paths through the source or drain terminals of transistors in the ohmic region of operation. In the ohmic region the transistor is modeled as a linear resistance (conductance) between the source and drain. When the transistor is in saturation, the transistor is modeled as a transconductance which has the effect of decoupling the source and drain. Instead of solving for all circuit node voltages at once, the complexity of the problem is reduced by approximating the solution in each resistance connected region separately.

3.2 Circuit Dynamic Equations

3.2.1 Circuit Dynamic Equations

It is assumed that the circuit of interest consists of transistors and capacitors only. The *m-th* transistor is connected to circuit nodes S_m, D_m, G_m, and B_m. The *c-th* capacitor is connected to nodes A_c and B_c as shown in Fig. 3.1:

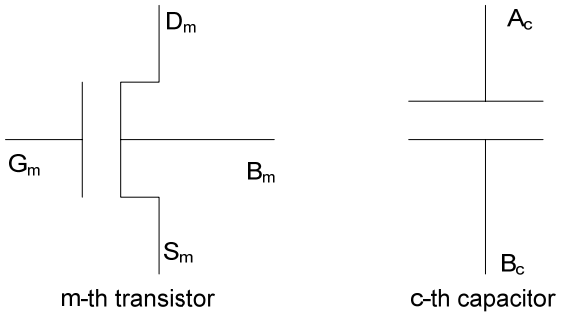

Fig.3. 1 Circuit components

where S_m, D_m, G_m, and B_m are the nodes connected to transistor's source, drain, gate, and bulk respectively. The transistors in each region of operation are approximated by the linear circuit as discussed in the previous section and the capacitors are approximated as

linear. Currents flowing through each node include the static current from Eq. (2.1) and Eq. (2.2), the channel storage current from Eq. (2.17), and the current charging or discharging the node capacitance. KCL at node *i* requires that

$$\sum_m I_{D_m}(\delta_{i,D_m} - \delta_{i,S_m}) + \sum_m (i_{G_m}\delta_{i,G_m} + i_{S_m}\delta_{i,S_m} + i_{D_m}\delta_{i,D_m}) + \sum_c I_c(\delta_{i,A_c} - \delta_{i,B_c}) = 0 \qquad (3.1)$$

where

$$\delta_{i,j} = \begin{cases} 1 & i = j \\ 0 & i \neq j \end{cases} \qquad (3.2)$$

Plugging in the static current and channel charge storage current equations derived in chapter 2, Eq. (3.1) can be rewritten in matrix form as:

$$C\frac{dV}{dt} + C_{in}\frac{dV_{in}}{dt} + GV + G_{in}V_{in} + G_{Tn}V_{Tn} + G_{Tp}V_{Tp} + G_{DD}V_{DD} = 0 \qquad (3.3)$$

where G is initially a conductance matrix formed by grouping together the rows and columns in the same resistance connected region. The conductance matrix, *G*, can be written in block upper triangular form. The non-zero entries in the off-diagonal blocks come from the gate terminals of the transistors in the resistance connected region (the inputs to the logic gate). In Eq. (3.3), *G* is further reduced by taking out the rows and columns for V_{DD}, V_{in}, V_{Tn}, and V_{Tp}. The saturation models make the conductance matrix non-symmetric. *G* is now the on-diagonal sub-block for a single resistance connected region:

$$G_{ii} = \sum_{\{m\}_{ohm}} a_m G_m(\delta_{i,D_m} + \delta_{i,S_m}) + \sum_{\{m\}_{sat}} G_m\delta_{i,S_m} \qquad (3.4)$$

$$G_{ij} = -\sum_{\{m\}_{ohm}} a_m G_m(\delta_{i,S_m}\delta_{j,D_m} + \delta_{i,D_m}\delta_{j,S_m}) + \sum_{\{m\}_{sat}} G_m(\delta_{i,D_m} - \delta_{i,S_m})\delta_{j,G_m} - \sum_{\{m\}_{sat}} G_m\delta_{i,D_m}\delta_{j,S_m} \tag{3.5}$$

G_{Tn} and G_{Tp} are the parts of the original column vectors inside the resistance connected region. They are formed from the original equation Eq.(3.1) by grouping together the conductance items inside the resistance connected region that are related to the threshold voltages V_{Tn} and V_{Tp} respectively.

$$G_{i,Tn} = -\sum_{\{m\}_{n,sat}} G_m(\delta_{i,D_m} - \delta_{i,S_m}) \tag{3.6}$$

$$G_{i,Tp} = -\sum_{\{m\}_{p,sat}} G_m(\delta_{i,D_m} - \delta_{i,S_m}) \tag{3.7}$$

G_{DD} is also a column vector that is generated from the original G matrix. It represents the cross-coupling conductance between the nodes inside the source-drain connected region with the power supply node.

$$G_{i,DD} = -\sum_{\{m\}_{ohm}} a_m G_m(\delta_{i,S_m}\delta_{DD,D_m} + \delta_{i,D_m}\delta_{DD,S_m}) + \sum_{\{m\}_{sat}} G_m(\delta_{i,D_m} - \delta_{i,S_m})\delta_{DD,G_m} - \sum_{\{m\}_{sat}} G_m\delta_{i,D_m}\delta_{DD,S_m} \tag{3.8}$$

V_{in} is the column vector of inputs outside the resistance connected region. G_{in} is an off-diagonal sub-block of the larger G matrix. Since G_{in} represents coupling between nodes inside a resistance connected region to nodes outside that region through transistors, the nodes outside the region must be terminals of transistors with at least one terminal inside the region. This leaves only the following terms in G_{in}.

$$G_{i,in} = \sum_{\{m\}_{sat}} G_m(\delta_{i,D_m} - \delta_{i,S_m})\delta_{in,G_m} - \sum_{\{m\}_{sat}} G_m\delta_{i,D_m}\delta_{in,S_m} \tag{3.9}$$

The C matrix is rectangular in general, and includes capacitive coupling from nodes inside the resistance connected region to all other circuit nodes including those inside and outside the resistance connected region. The rectangular C matrix prevents an exact solution of the node voltages within a resistance connect region without knowing the voltage derivatives at all other circuit nodes.

$$C_{ii} = \sum_c C_c(\delta_{i,A_c} + \delta_{i,B_c}) + \sum_m (C_{GG_m}\delta_{in,G_m} + C_{SS_m}\delta_{in,S_m} + C_{DD_m}\delta_{i,D_m}) \tag{3.10}$$

$$\begin{aligned} C_{ij} = &-\sum_c C_c(\delta_{i,B_c}\delta_{j,A_c} + \delta_{i,A_c}\delta_{j,B_c}) - \sum_m (C_{GS_m}\delta_{i,G_m}\delta_{j,S_m} + C_{SG_m}\delta_{i,S_m}\delta_{j,G_m}) \\ &-\sum_m (C_{GD_m}\delta_{i,G_m}\delta_{j,D_m} + C_{DG_m}\delta_{i,D_m}\delta_{j,G_m}) - \sum_m (C_{DS_m}\delta_{i,D_m}\delta_{j,S_m} + C_{SD_m}\delta_{i,S_m}\delta_{j,D_m}) \end{aligned} \tag{3.11}$$

We make the assumption that node voltages outside the resistance connected region do not change at the same time as node voltages inside the resistance connected region except for the input nodes driving the sources and gates of the transistors in saturation. We also have $C_{SD} = 0$ in saturation which prevents coupling from the drain back to the source. This makes the C matrix square and leaves only the following terms in C_{in}.

$$\begin{aligned} C_{i,in} = &-\sum_c C_c(\delta_{i,B_c}\delta_{in,A_c} + \delta_{i,A_c}\delta_{in,B_c}) \\ &-\sum_m [(C_{GS_m} + C_{GSO_m})\delta_{i,G_m}\delta_{in,S_m} + (C_{SG_m} + C_{GSO_m})\delta_{i,S_m}\delta_{in,G_m}] \\ &-\sum_m [(C_{GD_m} + C_{GDO_m})\delta_{i,G_m}\delta_{in,D_m} + (C_{DG_m} + C_{GDO_m})\delta_{i,D_m}\delta_{in,G_m}] \\ &-\sum_m (C_{DS_m} + \delta_{i,D_m}\delta_{in,S_m} + C_{SD_m}\delta_{i,S_m}\delta_{in,D_m}) \end{aligned} \tag{3.12}$$

V is a $n \times n$ matrix in Eq. (3.3) that represents the voltage outputs of each node in the resistance connected region. n is the quantity of the nodes inside the resistance

connected region except the power supply and ground nodes. A single changing input is assumed in the model. However, in theory the model can handle more than one switching input by making V_{in} a column vector. The rest of the terms in Eq. (3.3), V_{Tn}, V_{Tp}, and V_{DD} are constant scalars.

In most switching transitions, the source/drain voltages do not change significantly until the gate voltage is at or near its final value. Only very small errors result if we assume that the source/drain voltages are constant while the gate voltage switches. This eliminates coupling from source/drain nodes outside the resistance connected region through transistor gate terminals. The reverse is not the case since a changing gate voltage can have significant effects through capacitive coupling on the source and drain voltages. Just as we did for the conductance, we can treat the gate voltage as an external input to a resistance connected region which drives the nodes within the resistance connected region through a coupling capacitance instead of a transconductance.

3.2.2 Ramp Input Approximation

Rather than attempt an exact solution of (3.3) for an arbitrary input, the inputs to each resistance connected region will be approximated as a simple ramp as shown below.

$$V_{in}(t) = \begin{cases} V_{in}(t_{in0}) & t < t_{in0} \\ V_{in}(t_{in0}) + \dot{V}_{in}(t - t_{in0}) & t_{in0} < t < t_{in0} + t_{Tin} \\ V_{in}(t_{Tin}) & t > t_{in0} + t_{Tin} \end{cases} \tag{3.13}$$

where

$$\dot{V}_{in} = \frac{V_{in}(t_{in0} + t_{Tin}) - V_{in}(t_{in0})}{t_{Tin}}$$

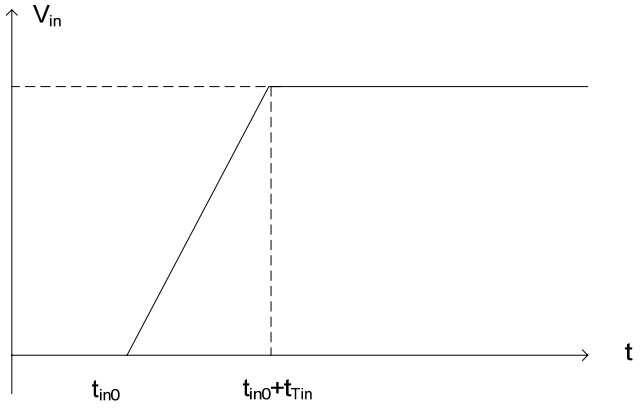

Fig.3. 2 Input ramp approximation

For more general input approximation, the inputs to each resistance connected region will be approximated as a series of piece-wise linear segments as shown

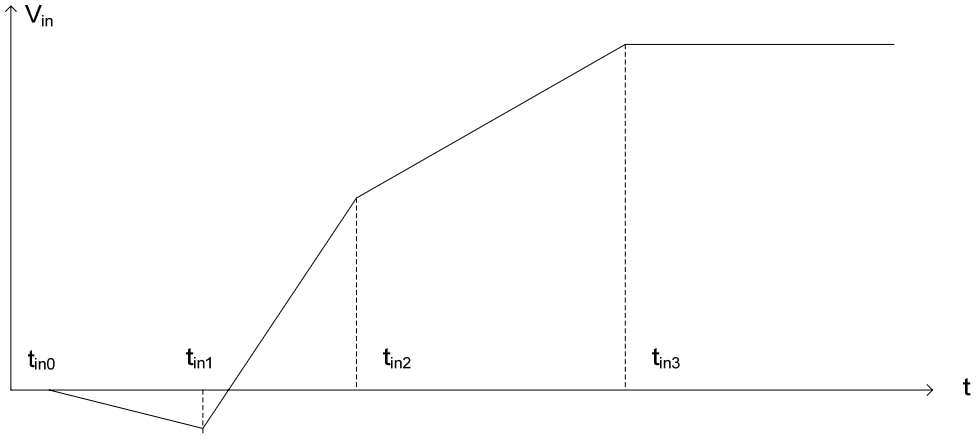

Fig.3. 3 Piecewise linear approximation of input waveform

$$V_{in}(t) \approx \begin{cases} V_{in}(t_{in_0}) & t < t_{in_0} \\ V_{in}(t_{in_k}) + \dfrac{V_{in}(t_{in_{k+1}}) - V_{in}(t_{in_k})}{t_{in_{k+1}} - t_{in_k}}(t - t_{in_k}) & t_{in_k} < t < t_{in_k} + t_{in_{k+1}} \\ V_{in}(t_{in_{k\max}}) & t > t_{in_{k\max}} \end{cases} \tag{3.14}$$

$V_{in}(t_{in_k})$ are determined by approximating the input waveform at a finite number of times, t_{in_k}. The accuracy of the approximation increases with the number of time points. Adding time points increases the amount of calculation necessary to determine the voltages in the resistance connected region.

If a node, *i*, in one region is to be used as an input into another region, it is necessary to find the piece-wise approximation that best fits V_i without making the amount of calculation too large. First, we will find the solution for V_i and then find piecewise approximations in a latter section.

3.3 Steady State Solution

Using the ramp approximation in Eq. (3.13) for the input nodes, the dynamic equations (3.3) become

$$C\frac{dV}{dt} + GV + I_{in}(t_0) + \dot{I}_{in}(t - t_0) = 0 \tag{3.15}$$

where

$$I_{in}(t_0) = C_{in}\dot{V}_{in} + G_{in}V_{in}(t_0) + G_{Tn}V_{Tn} + G_{Tp}V_{Tp} + G_{DD}V_{DD} \tag{3.16}$$

$$\dot{I}_{in} = G_{in}\dot{V}_{in} \tag{3.17}$$

The steady state solution (after the exponential terms die out), $\tilde{V}$, is of the form

$$\tilde{V}(t) = \tilde{V}(t_0) + \dot{\tilde{V}}(t - t_0) + \ddot{\tilde{V}} \frac{(t - t_0)^2}{2} \tag{3.18}$$

where the second order term is only necessary when G is singular as discussed later. The steady state solution must also satisfy the dynamic equations which can be rewritten as

$$G\tilde{V} = -(C \frac{d\tilde{V}}{dt} + I_{in}) \tag{3.19}$$

where collecting terms of the same power of *t* gives

$$G\ddot{\tilde{V}} = 0 \tag{3.20}$$

$$G\dot{\tilde{V}} = -(C\ddot{\tilde{V}} + \dot{I}_{in}) \tag{3.21}$$

$$G\tilde{V}(t_0) = -(C\dot{\tilde{V}} + I_{in}(t_0)) \tag{3.22}$$

The conductance matrix G can be singular or non-singular. Thus, the steady state solution is derived according to these two cases.

3.3.1 Non-Singular Conductance Matrix

The conductance matrix for a resistance connected region will be non-singular as long as the region includes the power or ground node. When *G* is non-singular, then G^{-1} can be used to find

$$\ddot{\tilde{V}} = 0 \tag{3.23}$$

$$\dot{\tilde{V}} = -G^{-1}(C\ddot{\tilde{V}} + \dot{I}_{in}) \tag{3.24}$$

$$\tilde{V}(t_0) = -G^{-1}(C\dot{\tilde{V}} + I_{in}(t_0)) \tag{3.25}$$

which can be rewritten as

$$\dot{\tilde{V}} = -G^{-1}\dot{I}_{in} \tag{3.26}$$

$$\tilde{V}(t_0) = -G^{-1}[CG^{-1}\dot{I}_{in} + I_{in}(t_0)] \tag{3.27}$$

Note that G^{-1} is not difficult to compute since G is a relatively small dimension matrix describing a single resistance connected region. It is possible to find a steady state solution in each resistance connected region knowing only the ramp approximation to V_{in}.

3.3.2 Singular Conductance Matrix

The conductance matrix for a resistance connected region will be singular whenever the region does not include the power or ground node or the source node of any transistor in saturation. It is easy to show that

$$\sum_j G_{ij} = \sum_i G_{ij} = 0 \tag{3.28}$$

for a singular conductance matrix.

It can be found that

$$\ddot{\tilde{V}} = \frac{-\sum_i [\dot{I}_{in}]_i}{\sum_j (\sum_i C_{ij})} \tag{3.29}$$

The rest of the steady state solution can be found by picking any node, *r*, in the resistance connected region that has a non-zero capacitance $\sum_i C_{ir}$. The charge conservation equation can be used to define $\dot{\tilde{V}}_r$ in terms of the other node voltages. The boundary condition of the initial charge can be used to define $\tilde{V}_r(t_0)$ in terms of the other node voltages.

The steady state solutions of the arbitrary reference node *r* are:

$$\dot{\tilde{V}}_r = -\frac{\sum_i [I_{in}(t_0)]_i}{\sum_j (\sum_i C_{ij})} - \frac{1}{(\sum_i C_{ir})} \sum_{j\neq r} \sum_{i'\neq r} (\sum_i C_{ij}) \Gamma^{-1}{}_{ji'} (\sum_{j'} C_{i'j'} \frac{\sum_{j'} [\dot{I}_{in}]_{j'}}{\sum_{j'} \sum_{i''} C_{i''j'}} - [\dot{I}_{in}]_{i'}) \quad (3.30)$$

$$\begin{aligned} \tilde{V}_r(t_0) = {} & \frac{\sum_j (\sum_{i'} C_{i'j}) V_j(t_0)}{\sum_j (\sum_{i'} C_{i'j})} - \sum_{j\neq r} \frac{(\sum_i C_{ij})}{(\sum_i C_{ir})} [\sum_{i\neq r} \Gamma^{-1}{}_{ji} (\sum_{j'} C_{i'j'} \frac{\sum_{i'} [I_{in}(t_0)]_{i'}}{\sum_{j'} \sum_{i'} C_{i'j'}} - [I_{in}(t_0)]_{i'}) \\ & - \sum_{i\neq r} [\Gamma^{-1} \chi \Gamma^{-1}]_{ji} (\sum_{j'} C_{ij'} \frac{\sum_{j'} [\dot{I}_{in}]_{j'}}{\sum_{j'} \sum_{i'} C_{i'j'}} - [\dot{I}_{in}]_i)] \end{aligned} \quad (3.31)$$

where

$$\Gamma_{ij} = G_{ij} - \frac{G_{ir} \sum_{i'} C_{i'j}}{\sum_{i'} C_{i'r}} \quad i \neq r, j \neq r \quad (3.32)$$

$$\chi_{ij} = C_{ij} - \frac{C_{ir} \sum_{i'} C_{i'j}}{\sum_{i'} C_{i'r}} \quad i \neq r, j \neq r \quad (3.33)$$

The steady state solutions of non-reference nodes are

$$\dot{\tilde{V}}_{i\neq r} = -\frac{\sum_{i'} [I_{in}(t_0)]_{i'}}{\sum_{j'} (\sum_{i'} C_{i'j'})} + \sum_{j\neq r} \Gamma^{-1}{}_{ij} (\sum_{j'} C_{jj'} \frac{\sum_{i'} [\dot{I}_{in}]_{i'}}{\sum_{j'} \sum_{i'} C_{i'j'}} - [\dot{I}_{in}]_j) \quad (3.34)$$

$$\begin{aligned} \tilde{V}_{i\neq r}(t_0) = {} & \frac{\sum_j (\sum_{i'} C_{i'j}) V_j(t_0)}{\sum_j (\sum_{i'} C_{i'j})} + \sum_{j\neq r} \Gamma^{-1}{}_{ji} [(\sum_{j'} C_{jj'} \frac{\sum_{i'} [I_{in}(t_0)]_{i'}}{\sum_{j'} \sum_{i'} C_{i'j'}} - [I_{in}(t_0)]_j) \\ & - \sum_{j\neq r} [\Gamma^{-1} \chi \Gamma^{-1}]_{ij} (\sum_{j'} C_{jj'} \frac{\sum_{j'} [\dot{I}_{in}]_{j'}}{\sum_{j'} \sum_{i''} C_{i''j'}} - [\dot{I}_{in}]_j] \end{aligned} \quad (3.35)$$

Detailed derivation can be found in Appendix C.

3.4 Generalized Elmore Delay

3.4.1 Elmore Delays for A Step Input [26]

The usual definition of delay is the time difference between the midpoint of the input waveform to the midpoint of the output waveform. Thus, it is impossible to calculate the delay without knowing the exact input and output waveform. The Elmore delay can be found independent of the exact waveforms. Assume that V_i, the voltage at node i, falls from some high value to a low value:

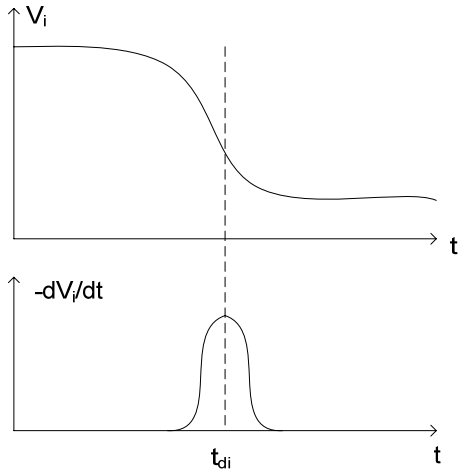

Fig.3. 4 Elmore delay for a falling waveform

Elmore defines the delay, t_{di}, at node i as the centroid of $-\frac{dV_i}{dt}$,

$$t_{di} = \frac{\int_0^\infty t(-\frac{dV_i}{dt})dt}{\int_0^\infty (-\frac{dV_i}{dt})dt} = \frac{1}{\Delta V_i}\int_0^\infty t\frac{dV_i}{dt}dt \tag{3.36}$$

where

$$\Delta V_i = V_i(\infty) - V_i(0) \tag{3.37}$$

This definition is valid only if the output voltage waveform is a monotonic function of *t*, i.e., the derivative does not change. If rewrite Eq. (3.36) by integrating by parts, the Elmore delay can be written as:

$$t_{di} = -\frac{1}{\Delta V_i}\int_0^{\infty}(V_i - V_i(\infty))dt \tag{3.38}$$

This means an alternate equivalent way of defining the Elmore delay for a falling waveform is the area between $V_i(0)$ and $V_i(\infty)$. Same result can be obtained for a rising waveform as Eq. (3.38) except for an opposite sign.

The usual Elmore delay in a circuit made from an arbitrary interconnection of MOSFET is:

$$\tau_i = \frac{[G^{-1}C\Delta V]_i}{\Delta V_i} \tag{3.39}$$

When using Eq. (3.39), several points need to be mentioned:

1. The solution is exact for any general circuit containing only resistors and capacitors only. (perhaps some DC voltage sources.)

2. For nodes with $\Delta V_i = 0$, it is not possible to solve for Elmore delay using the equation above. But when circuit goes into and out of different piece-wise linear regions, the variation of the node voltage may not be zero. A new mathematical expression for Elmore delay used in MOSFET circuits is introduced in next section.

3. Elmore delay in Eq. (3.39) does not take into account of the input rising or falling transition time.

3.4.2 Generalized Elmore Delays

The normal definition of Elmore is not suitable for the response of the resistance connected regions to a ramp input since the node voltages do not always approach a constant steady state value. Instead, we propose the following generalized Elmore delay for each node *i*.

$$\tau_i = \frac{\int_{t_0}^{\infty} \left| V_i(t) - \tilde{V}_i(t) \right| dt}{\left| V_i(t_0) - \tilde{V}_i(t_0) \right|} \tag{3.40}$$

3.4.2.1 Non-Singular Conductance Matrix

The particular solution, $\tilde{V}_i(t)$, satisfies the dynamic equations:

$$C\frac{d}{dt}(V - \tilde{V}) + G(V - \tilde{V}) = 0 \tag{3.41}$$

We can use (3.41) to evaluate (3.40) gives

$$\tau_i = \frac{-\int_{t_0}^{\infty} [G^{-1}C\frac{d}{dt}\left| V_i(t) - \tilde{V}_i(t) \right|]_i dt}{\left| V_i(t_0) - \tilde{V}_i(t_0) \right|} \tag{3.42}$$

$$\tau_i = \frac{-\int_{t_0}^{\infty} [G^{-1}C\left| V_i(t_0) - \tilde{V}_i(t_0) \right|]_i}{\left| V_i(t_0) - \tilde{V}_i(t_0) \right|} \tag{3.43}$$

Unfortunately, G and C are not constant, but changes as the switching transistors go into their various regions of operation. We assume that the solution beginning at t_0 does not know that it will become invalid later, but can be extended indefinitely forward in time with a constant G and C. This allows upper limit in the integrals to be extended to infinity.

3.4.2.2 Singular Conductance Matrix

When the conductance matrix is singular, we cannot use G^{-1} to define the Elmore delay. Proceeding as we did for the steady state solution, we use charge conservation to define the transient for an arbitrary node *r* in terms of the other node transients. The Elmore delay for all nodes except node *r*, is

$$\tau_{i\neq r} = \frac{[\Gamma^{-1}\chi\left|V(t_0)-\tilde{V}(t_0)\right|]_{i\neq r}}{\left|V_{i\neq r}(t_0)-\tilde{V}_{i\neq r}(t_0)\right|} \tag{3.44}$$

The Elmore delay for node *r* is

$$\tau_r = \frac{-\sum_{j\neq r} C_{j0}\left|V_j(t_0)-\tilde{V}_j(t_0)\right|\tau_j}{C_{r0}\left|V_r(t_0)-\tilde{V}_r(t_0)\right|} \tag{3.45}$$

Detailed derivation can be found in Appendix D.

3.5 Approximate Circuit Dynamics

The solution of the circuit dynamics Eq. (3.15) is approximated by the solution of the equivalent first order system in each piece-wise linear region,

$$\tau_i \frac{d}{dt}(V_i-\tilde{V}_i)+V_i-\tilde{V}_i = 0 \quad t_0 < t < t_1 \tag{3.46}$$

t_0 and t_1 are boundary points when the circuit enters a piecewise linear region at time t_0 and leaves the region at time t_1. We know t_0 from solving the previous piecewise linear region. They are the intersection points where the output waveform crosses the boundary line, which separates the whole transition into different regions. Newton-Raphson method is used to find the boundary points. In next section, details are given for the Newton-Raphson method implemented in the model. t_1 is when the circuit leaves the current region and starts operating in the next region.

The solutions to the equivalent first order system are

$$V_i(t) - \tilde{V}_i(t) = [V_i(t_0) - \tilde{V}_i(t_0)]e^{-\frac{t-t_0}{\tau_i}} \quad t > t_0 \tag{3.47}$$

3.6 Boundaries of Piecewise Linear Regions

The solutions in Eq. (3.47) are valid as long as all the transistors remain in the same region and the input slopes $\dot{V}_{in}$ do not change. When a transistor changes state or the input slopes change, the circuit will leaves the current region and enters into a new region. The boundary of the two different regions is determined by the switching transistors turning on or off and going in or out of saturation region. From (3.47), the starting time of each region must be known in order to find out the output voltage waveform in each region. In the piecewise linear delay model, these critical points are solved by Newton-Raphson method.

The boundaries are the same as those defined in Eq. (2.1) to (2.4). Basically they correspond to the following two equations for each transistor in the source-drain connected region:

$$V_{GS} = V_T \tag{3.48}$$

$$V_{DS} = V_{DSsat} = \frac{V_{GS} - V_T}{a} \tag{3.49}$$

In theory, any one of the four terminals of the transistor could be the output node. And in most circuits, more than one terminal voltage output waveforms need to know in order to find the boundaries if those terminals are changing simultaneously. Hence, a general solution to find the boundary points for each node at a particular region is

desired. It would be more convenient to write the boundary equations in terms of the individual terminal voltages:

$$V_G - V_S = V_T \tag{3.50}$$

$$V_G + (a-1)V_S - aV_D = V_T \tag{3.51}$$

From Eq. (3.50) and (3.51), the boundary points formed by the boundary lines crossing the terminal nodes voltage output waveforms can be calculated by Newton-Raphson method. The idea behind the Newton-Raphson method is that usually only one iteration is necessary to obtain accuracy to better than a few percent. For using the Newton method, an initial guess $t = t_{Fin}$ is made. Using this initial guess, the time at the boundary line, t_B, can be found by:

$$t_{Boff} = t_{Fin} + \frac{V_T - \left[V_G(t_{Fin}) - V_S(t_{Fin})\right]}{\dot{V}_G(t_{Fin}) - \dot{V}_S(t_{Fin})} \tag{3.52}$$

$$t_{Bsat} = t_{Fin} + \frac{V_T - \left[V_G(t_{Fin}) - (1-a)V_S(t_{Fin}) - aV_D(t_{Fin})\right]}{\dot{V}_G(t_{Fin}) - (1-a)\dot{V}_S(t_{Fin}) - a\dot{V}_D(t_{Fin})} \tag{3.53}$$

where t_{Boff} is time when the transistor turns on or off and t_{Bsat} is the time when the transistor goes in or leaves the saturation region. The node voltages and their derivatives can be obtained from

$$V_i(t_{Fin}) = \tilde{V}_i(t_{Fin}) + [V_i(t_0) - \tilde{V}_i(t_0)]e^{-\frac{t_{Fin}-t_0}{\tau_i}} \tag{3.54}$$

$$\dot{V}_i(t_{Fin}) = -\frac{1}{\tau_i}[V_i(t_0) - \tilde{V}_i(t_0)]e^{-\frac{t_{Fin}-t_0}{\tau_i}} \tag{3.55}$$

where $i = G,\ S,$ and D respectively.

In order to achieve a better accuracy, usually 3 to 4 iterations are needed to use Newton method. We substitute the values of t_{Boff} and t_{Bsat} back into the equation (3.52) and (3.53):

$$t_{Boff1} = t_{Boff} + \frac{V_T - \left[V_G(t_{Boff}) - V_S(t_{Boff})\right]}{\dot{V}_G(t_{Boff}) - \dot{V}_S(t_{Boff})} \tag{3.56}$$

$$t_{Bsat1} = t_{Bsat} + \frac{V_T - \left[V_G(t_{Bsat}) - (1-a)V_S(t_{Bsat}) - aV_D(t_{Bsat})\right]}{\dot{V}_G(t_{Bsat}) - (1-a)\dot{V}_S(t_{Bsat}) - a\dot{V}_D(t_{Bsat})} \tag{3.57}$$

where $V_G(t_{Boff})$, $V_S(t_{Boff})$, $V_G(t_{Bsat})$, $V_S(t_{Bsat})$, $V_D(t_{Bsat})$ and their corresponding derivatives can be found by substitute t_{Boff} and t_{Bsat} for t_{Fin} in (3.34) and (3.35). In the model implementation, the tolerance is set as 0.001% and this is achieved by 3 to 4 iterations of Newton method described above.

Convergence problem met in the Newton method application is handled in the program by approaching the solution from right, i.e., choosing a bigger value of initial guess. In this dissertation, we pick $t_{Fin} = \max\{t_{Tin}, \tau\}$.

If the output transition waveform is beyond the power supply voltages (V_{DD} and GND), a third boundary line $V_i = V_o + \delta$ is defined to limit the overshoot of the output voltage, where $V_o = V_{DD}$ for a rising input transition and $V_o = 0$ for a falling input transition. The following section presents the procedure of finding the boundary point where the output waveform crosses the diode like cutoff boundary line $V_i = V_{DD} + \delta$. Similar method can be used to find the boundary point where the voltage output waveform crosses the boundary line $V_i = \delta$.

First we need to decide whether the output waveform will rise above the boundary $V_i = V_o + \delta$ or not. The maximum value of the output voltage is compared to the diode cutoff boundary. If the maximum value is less than the boundary, the whole output waveform will be under the boundary and we can proceed to find other boundary points as defined in Eq. (3.52) and Eq. (3.53). Otherwise, the output waveform will cross the boundary line $V_i = V_o + \delta$ at $t = t_r$, as is shown in curve I of Fig. 3.5.

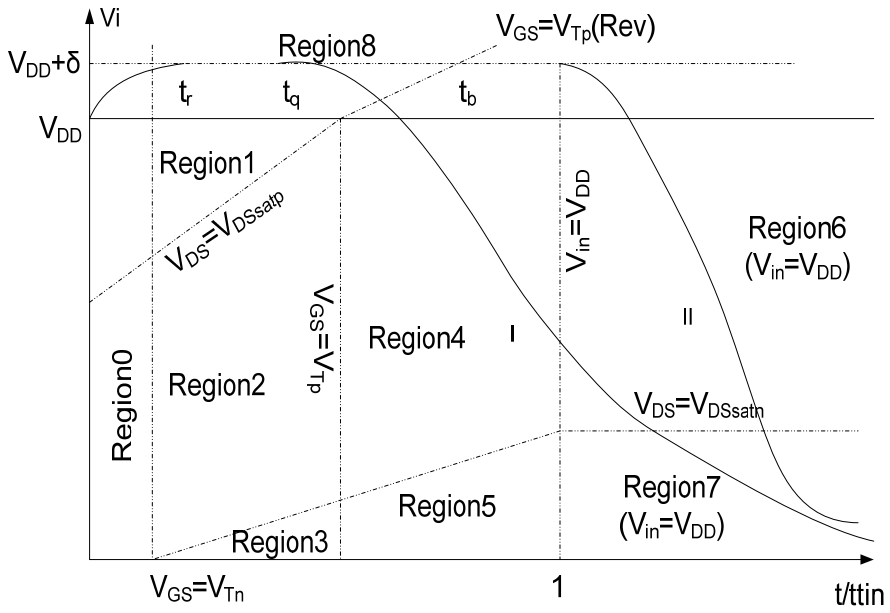

Fig.3. 5 Output falling transition with boundaries and operation regions

The output waveform will stay at $V_i = V_o + \delta$ until $t \geq t_q$, where t_q is the intersection point where the quasi-steady state solution of the output waveform crosses the diode like boundary. t_q can be found from Eq. (3.46) by letting $\frac{dV_i}{dt} = 0$ and $V_i = V_0 + \delta$,

$$t_q = t_r + \frac{-(\tau_i \ddot{\tilde{V}}_i + \dot{\tilde{V}}_i) \pm \sqrt{(\tau_i \ddot{\tilde{V}}_i + \dot{\tilde{V}}_i)^2 - 2\ddot{\tilde{V}}_i(\tau_i \dot{\tilde{V}}_i + \tilde{V}_i(t_0) - (V_0 + \delta))}}{\ddot{\tilde{V}}_i} \tag{3.58}$$

Eq. (3.54) is the general solution to find the intersection point of the quasi-steady state solution of the output voltage waveform and the boundary $V_i = V_o + \delta$. t_q is the point after which the output waveform will start fall. If $t_q > t_b$, the output voltage waveform will stay at $V_o + \delta$ until $t = t_q'$ in the next region, where t_q' is obtained by Eq. (3.54) . For the very fast input transition, $t_q' > t_{Tin}$, the output will start fall at $t = t_{Tin}$ when the input ramp has arrived at its peak value.

3.7 Linear Approximation of Output Waveform

For non-singular conductance, Eq. (3.47) simplifies to

$$V_i(t) = \tilde{V}_i(t_0) + \dot{\tilde{V}}_i(t - t_0) + [V_i(t_0) - \tilde{V}_i(t_0)]e^{-\frac{t-t_0}{\tau_i}} \tag{3.59}$$

when the off-diagonal term G_{ij} of the G matrix is not zero, the steady state solution $\tilde{V}_i(t)$ will be a function of output voltage of node j too. Since $V_j(t)$ is also an exponential function, this will make it very difficult to solve for the boundary points using Newton-Raphson method.

We can linearize Eq. (3.54) if

$$\tilde{V}_i(t_0) + \dot{\tilde{V}}_i\tau_i - [\tilde{V}_i(t_0) + \dot{\tilde{V}}_i\tau_i + [V_i(t_0) - \tilde{V}_i(t_0)]e^{-\frac{\tau_i}{\tau_i}}] < \delta V \tag{3.60}$$

i.e.,

$$[V_i(t_0) - \tilde{V}_i(t_0) < e \bullet \delta V$$

and $V_j(t)$ can be approximated as

$$V_j(t) \approx \begin{cases} V_j(t_0) + [\dot{\tilde{V}}_j - \dfrac{V_j(t_0) - \tilde{V}_j(t_0)}{\tau_j}](t - t_0) & t_0 < t < t_0 + \tau_j \\ \tilde{V}_j(t_0) + \dot{\tilde{V}}_j(t - t_0) & t > t_0 + \tau_j \end{cases} \quad (3.61)$$

3.8 Delay and Output Rise/Fall Time

The piecewise linear model does not have a closed form solution for delay. Delay calculated in the piecewise linear model is the mid-point delay which is the time difference between the half-V_{DD} point of the input voltage waveform and the half-V_{DD} point of the output voltage waveform. The voltage output waveform can be computed from Eq. (3.47) in each piecewise linear region. The output voltage is compared with half-V_{DD} volt in each region and half-V_{DD} point of the output voltage waveform can be found using Newton-Raphson's method. By comparing Eq. (3.13), the delay can be computed by:

$$t_d = t_{50} - t_{in0} - \frac{t_{Tin}}{2} \quad (3.62)$$

The usual definition of rise time or fall time is the duration between the 10% and 90% of the output waveform. The piecewise linear model is pretty accurate for predicting the output waveforms before the mid-point of the output waveform. Thus, the first half of the output waveform computed by the piecewise linear model is used to find the output rise/fall time.

For a falling output, the time duration between the 90% and 50% is used to find the fall time. And for a rising output, the time duration between the 50% and 10% is used to find the rise time. Specifically,

$$t_r = (t_{50} - t_{10}) \times 2 \quad (3.63)$$

$$t_f = (t_{50} - t_{90}) \times 2 \tag{3.64}$$

where t_{50}, t_{10}, and t_{90} are the output transition time at $0.5V_{DD}$, $0.1V_{DD}$, and $0.9V_{DD}$ respectively.

3.9 Summary

By introducing the input transition time into the delay computation, the piecewise linear delay model is supposed to be more accurate than the switch models. In each piecewise linear region, the output is represented by its delayed quasi-steady state solution with an exponential waveform. This chapter introduced the general RC circuit dynamic equation in each region and described in details each element that composes the general solution of the dynamic equation. Although the piecewise linear delay model does not provide a closed form solution for the delay, delay can be obtained by tracking the output voltage waveform. The first half of the output waveform is used to find the output slope since the piecewise linear model is accurate comparing to SPICE simulation results in the first half of the output transition.

CHAPTER IV

ANALYSIS OF A CMOS INVERTER USING PIECEWISE LINEAR MODEL

In the previous chapters a piecewise linear model is introduced. A general solution for the output voltages is derived. The piecewise linear model is applied to the analysis of a CMOS inverter in this chapter. Delays and the output transition rise/fall time predicted by the model are compared to the corresponding SPICE simulation over a wide range of input slopes, transistor sizes, and loading factors.

4.1 CMOS Inverter Transition Analysis

Inverter is the simplest CMOS logic gate. It consists of a PMOS and a NMOS, and in most cases it also has a load capacitor at its output node. Two different transitions are analyzed below: rising input transition and falling input transition. A ramp input is assumed and the delay is computed as the midpoint delay from the output waveform to the midpoint of the input waveform.

4.1.1 Rising Input Transition

Fig. 4.1 shows a CMOS inverter with a rising input waveform. The operation regions of an inverter are shown in Figure 4.2. The separation of the operation regions

corresponds to the different combinations of the operation modes of NMOS and PMOS devices.

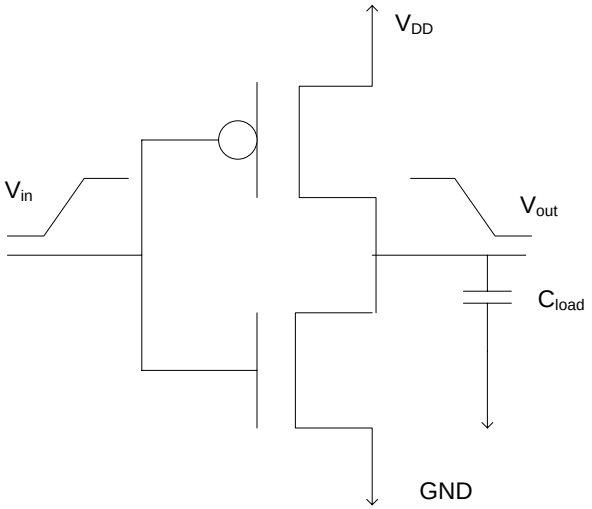

Fig.4. 1 A COMS inverter

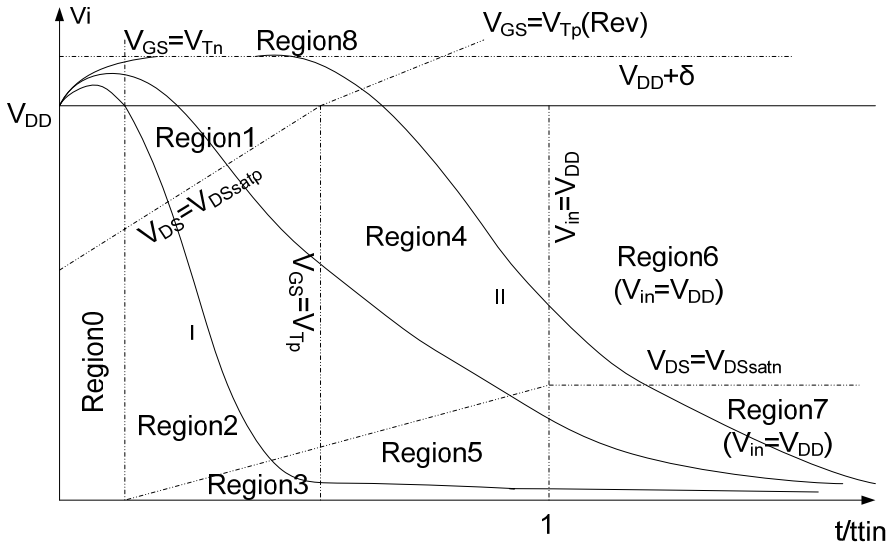

Fig.4. 2 Output falling transition with boundaries and operation regions

The operation modes of NMOS and PMOS transistors in each region are listed in TABLE 4.1:

Region	NMOS	PMOS	Input/Output
Region0	off	ohmic	$V_{in} < V_{DD}$
Region1	saturation	ohmic	$V_{in} < V_{DD}$
Region2	saturation	saturation	$V_{in} < V_{DD}$
Region3	ohmic	saturation	$V_{in} < V_{DD}$
Region4	saturation	off	$V_{in} < V_{DD}$
Region5	ohmic	off	$V_{in} < V_{DD}$
Region6	saturation	off	$V_{in} = V_{DD}$
Region7	ohmic	off	$V_{in} = V_{DD}$
Region8	saturation	ohmic	$V_i > V_{DD}$

TABLE 4. 1 Inverter Falling Output Transition Regions

The regions defined above are formed by nine boundary lines at which the transistor (transistors) changes operating states. (TABLE 4.2) The input transition falls into two categories in the piecewise linear model: fast input and slow input according to whether the PMOS turns off before NMOS goes into the ohmic region. Detailed analysis is given below.

Boundary	Boundary Line	Comments
Boundary1	$V_{GS} = V_{Tn}$	NMOS on/off
Boundary2	$V_{DS} = V_{DSsatp}$	PMOS in/out saturation
Boundary3	$V_{GS} = V_{Tp}$	PMOS on/off

Boundary4	$V_{DS} = V_{DSsatn}$	NMOS in/out saturation
Boundary5	$V_{in} = V_{DD}$	Input done transition
Boundary6	$V_{DS} = V_{DSsatn}$	Input done transition NMOS in/out saturation
Boundary7	$V_{GS} = V_{Tp}$ (Rev)	PMOS s/d reversed
Boundary8	$V_i = V_{DD}$	Power line
Boundary9	$V_i = V_{DD} + \delta$	Diode-like cutoff

TABLE 4. 2 Boundary Lines

4.1.1.1 Slow Input

The curve I in Fig. 4.2 is an output transition for a slow input. As seen from Fig. 4.2, curve I passes Region 0, Region 1, Region 2, Region 3, Region 5, and Region 7. The input changes from low to high for a given transition time t_{Tin}.

In Region 0, the NMOS transistor is off since $V_{GSn} < V_{Tn}$. PMOS transistor is in the ohmic region because $V_{GSp} < V_{Tp}$ and $V_{DSp} > V_{DSsatp}$. Part of the charge from the input which is injected through the coupling capacitance causes an overshoot at the early part of the output voltage waveform. This charge has the major influence on the output in Region0. Because the output voltage is greater than the power supply voltage, the source and drain of the PMOS are reversed, i.e., the output node becomes the source node of the PMOS transistor and the power supply node becomes the drain.

In Region 1, the NMOS transistor is saturated and the PMOS transistor is still in the ohmic region. As the input voltage keeps increasing, the voltage at the output node

decreases, which results in a decreasing of the drain to source voltage of the PMOS. When $V_{DSp} < V_{DSsatp}$, PMOS is saturated. As we can see from Fig. 4.2, when the voltage output waveform hit the boundary line $V_{DS} = V_{DSsatp}$, the PMOS changes its operation region to saturation region. Although the output voltage decreases, the NMOS is still operating in the high V_{DS} region which is far greater than V_{DSsatn} . After PMOS switching the operation state, the inverter works in the transition Region 2 in Fig. 4.2.

In between Region 0 and Region 1, there is another region Region 8. As in Region 1, the NMOS transistor is saturated and the PMOS transistor is in the ohmic region. The only difference between Region 1 and Region 8 is that in Region 8, the output voltage is greater than the power supply voltage, V_{DD}. Thus, PMOS reverses its source and drain.

In Region 2, both transistors are saturated. When the input voltage increases further, the drain voltage of the NMOS will drop more. If $V_{DSn} < V_{DSsatn}$, NMOS will operate in the ohmic region. This is reflected in Fig. 4.2 when the output waveform reaches the boundary line, $V_{DSn} = V_{DSsatn}$. Meanwhile, the drain to source voltage of the PMOS is more negative and much less than V_{DSsatp}. Since the gate to source voltage is less than the threshold voltage of the PMOS, PMOS is still in the saturation region. Based on the analysis above, the inverter leaves Region 2 and enters Region 3.

In Region 3, the NMOS transistor is in the ohmic region and the PMOS transistor is saturated. Since the source of the PMOS is connected to the power supply, the gate to source voltage of the PMOS is increasing when the input voltage waveform rises. When the gate to source voltage is big enough comparing to the threshold voltage V_{Tp}, PMOS is

off. As the output voltage keeps falling, the drain to source voltage of the NMOS decreases too. And since the drain to source voltage of the NMOS is less than the drain to source saturation voltage, the NMOS is still in the ohmic region. In Fig. 4.2, the output voltage waveform leaves Region3 and hits the boundary of $V_{GS} = V_{Tp}$, and enters into Region 5.

In Region 5, the NMOS transistor is saturated and the PMOS transistor is off. When the input voltage has finished its transition at $t = t_{Sin} + t_{Tin}$, the input voltage does not change any more and stays at $V_{in} = V_{DD}$. NMOS continues to operate in the low V_{DS} region. The gate to source voltage for PMOS is 0 which is always greater than the negative V_{Tp}. In Fig. 4.2, the output voltage waveform crosses the boundary of $V_{in} = V_{DD}$, and enters into Region 7.

In Region 7, the operation modes of NMOS and PMOS are the same as region 5 with NMOS in the ohmic region and PMOS in the cutoff region except that the input voltage stops changing. The inverter will stay in Region 7 as long as the input voltage stays at $V_{in} = V_{DD}$.

4.1.1.2 Fast Input

Above is the output transition analysis for an inverter with a slow rising input. When the input transition is very fast or the output load is very big, the PMOS transistor will turn off before the NMOS transistor goes into the ohmic region. We call it fast input transition.

Curve *II* in Fig. 4.2 is the output voltage transition for a fast input. It crosses Region 0, Region 1, Region 2, Region 4, Region 5, and Region 7.

The transition analysis for Region 0, Region 1, and Region 2 are exactly the same as what we did for a slow input. NMOS transistor changes from off state to be saturated and PMOS changes from the ohmic region to the saturation region. Thus, Region 0, Region 1, and Region 2 appear in both of the slow input transition path and fast input transition path I and II in Fig. 4.2.

What make the difference between the slow input and the fast input are the transitions after the Region 2.

For a fast input since the gate input voltage changes very fast comparing to the output voltage drop. Gate to source voltage of the PMOS, V_{GSp}, increases faster than the drain to source voltage drop, V_{DSn}, of the NMOS transistor . PMOS is cut off as long as the gate to source voltage is equal to the PMOS transistor threshold voltage , at which time the drain to source voltage of the NMOS transistor has not dropped below the V_{DSsatn} yet. The inverter circuit thus enters into an operation region that does not exist in the slow input. In this region, Region 4, NMOS is saturated and PMOS is off. In Fig 4.2, the voltage output waveform hits the boundary $V_{GS} = V_{Tp}$ instead of the NMOS saturation boundary $V_{DS} = V_{DSsatn}$.

Due to the different changing rate of NMOS drain to source voltage and PMOS gate to source voltage, the output waveform may enter into two different regions. Curve II in Fig. 4.2 shows one of the two cases. If the output voltage of the NMOS drops much faster than the input transition speed, the NMOS enters into ohmic region once its drain to source voltage is less than its drain to source saturation voltage. Since the gate to source voltage of the PMOS is increasing and is greater than the threshold voltage of the PMOS, PMOS is still in the cutoff region. In Fig. 4.2, the output waveform leaves Region

4 and enters into Region 5. The transition following the Region 5 is the same as slow input transition case. With the input voltage arrives at its final transition value V_{DD} , NMOS operates in the ohmic region and PMOS in the cutoff region.

Another case is that the input voltages reaches to its final value V_{DD} before the drain to source voltage of the NMOS transistor, V_{DSn}, falls below its drain to source saturation voltage value, V_{DSsatn}. The output voltage waveform enters into Region 6 first instead of Region 5 in Fig. 4.2. In Region 6, NMOS transistor is saturated and PMOS transistor stays cutoff. The inverter stays in this region until the drain to source voltage of the NMOS, V_{DSn}, is less than the V_{DSsatn}. When $V_{DS} < V_{DSsatn}$, the NMOS changes its operation region from saturation to ohmic region. This is shown in Fig. 4.2 as the output voltage waveform hits the boundary $V_{DS} = V_{DSsatn}$ and goes into final region Region 7. The rest of the transitions are the same as previous case.

4.1.2 Falling Input Transition

The Analysis of the output transition for a falling input is similar to that of a rising input. Different regions are formed by different boundary lines which are shown in Fig.4.3.

The operating states of the PMOS and NMOS in each region are defined in TABLE 4.3.

The output transition also falls into two categories: fast input and slow input. They are distinguished by whether the NMOS transistor turns off first (fast input) or the PMOS transistor goes into the ohmic region first (slow input). In other words, if the

voltage output waveform hits the boundary line $V_{GS} = V_{Tn}$ first in Fig. (4.3), the input is a fast input. Otherwise, if the output waveform hits the boundary line $V_{DS} = V_{DSsatp}$ first, the input is a slow input.

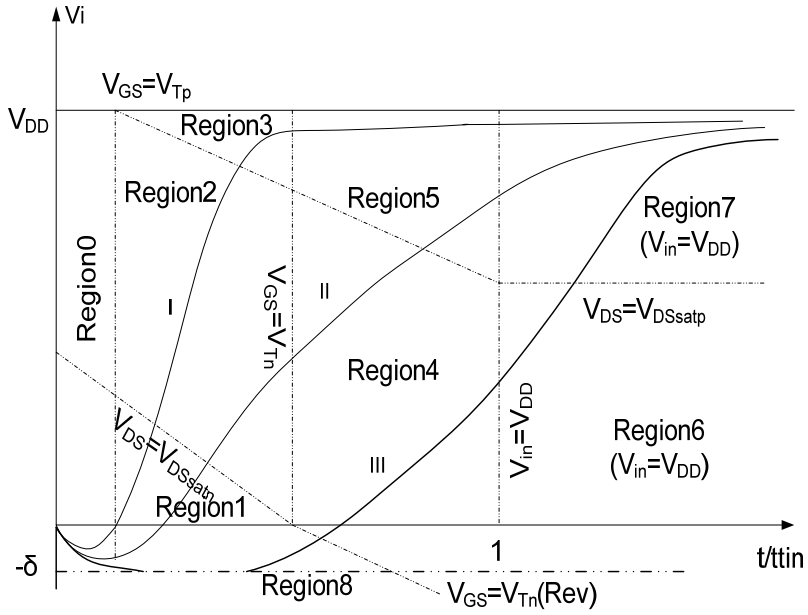

Fig.4. 3 Output rising transition with boundaries and operation regions

Region	NMOS	PMOS	Input/Output
Region0	ohmic	off	$V_{in} > 0$
Region1	ohmic	saturation	$V_{in} > 0$
Region2	saturation	saturation	$V_{in} > 0$
Region3	saturation	ohmic	$V_{in} > 0$
Region4	off	saturation	$V_{in} > 0$

Region5	off	ohmic	$V_{in} > 0$
Region6	off	saturation	$V_{in} = 0$
Region7	off	ohmic	$V_{in} = 0$
Region8	ohmic	saturation	$V_i < 0$

TABLE 4. 3 Inverter Rising Output Transition Regions

In Fig. 4.3, Curve *I* is a slow input transition. It crosses the Region 0, Region 8, Region 1, Region 2, Region 3, Region 5, and Region 7. Curve *II* and *III* are fast input transitions, instead of going to Region 3 as in the slow input case, they pass the Region 4, Region 5, Region 6, and Region 7.

The rest of the analysis is similar to the rising input and will not be stated here.

4.1.3 Special Issues

In the previous two sections, we have discussed the output transitions in details for an inverter with rising and falling input respectively. In this section, we will discuss the diode-like cutoff boundary in Fig. 4.2 and 4.3.

In Fig. 4.2, the starting portion of the transition is above V_{DD} . This is caused by the cross-coupling capacitance between the transistor gate and drain. As shown in Fig. 4.4, for a very fast input, the voltage accumulated on the gate side due to a rising input will also cause a voltage rise on the drain side of the cross-coupling capacitance. So the output node voltage is actually rising at the beginning of the transition for a given rising input.

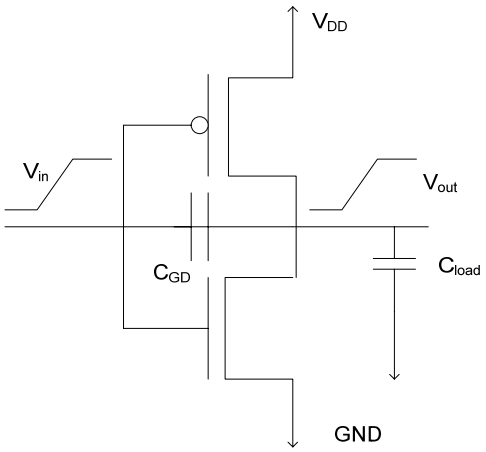

Fig.4. 4 Gate-drain cross-coupling capacitance

In the real circuits, the output voltage swings within some certain voltage range rather than an undefined one. In the piece-wise linear model, two diode-like limits are set to model the finite output voltage swing: $V_i = V_{DD} + \delta$ for a falling output transition and $V_i = -\delta$ for a rising output transition. The idea behind these two boundaries is to provide substrate diodes to limit the drain voltage in the circuit as shown in Fig. 4.5:

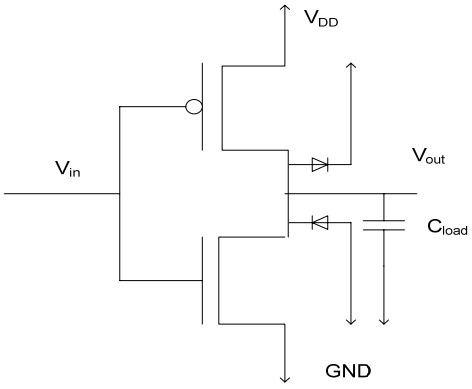

Fig.4. 5 Substrate diodes

The threshold voltage of the diodes in Fig. 4.5 is set to δ. Whenever the output voltage is greater than $V_{DD}+\delta$, the upper diode conducts current and the voltage drops to below $V_{DD}+\delta$. This diode functions like the boundary line $V_i = V_{DD}+\delta$ in Fig. 4.2.

For a rising output, when the output voltage is less than $GND-\delta$, the lower diode conducts current and the output voltage rises to above $GND-\delta$. The lower diode functions like the boundary line $V_i = -\delta$.

4.2 Model Evaluation and Calibration

The merit of the piecewise linear model lies on its introduction of the input transition time. Instead of a simple step input assumption, the circuit delay calculation includes the finite input transition time. Thus, the piecewise linear model is more accurate than the simulator like IRSIM which is based on the switch model. In this section, delays of various inverter circuits are compared between SPICE simulation results and the piecewise linear model predictions. Important factors like different PMOS to NMOS ratios, load conditions, and input slopes that have huge effects on the circuit performance are considered to evaluate the piecewise linear model.

4.2.1 An Inverter Drives a Constant Capacitance Load

The schematic diagram of the circuit is shown in Fig. 4.6:

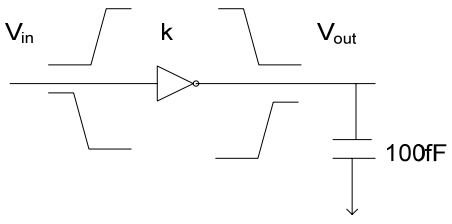

Fig.4. 6 An inverter drives a constant capacitance load.

In this circuit, an inverter is driving a constant capacitance load $100\,fF$. The reason to choose a big constant capacitance load is two folds. First of all, it is used to overshadow the impacts of the source/drain diffusion capacitance and the gate to drain capacitance to the delay. As has mentioned before, the piecewise linear model increases the accuracy of the circuits performance evaluation by taking into account of the input transition time. By reducing the effects of other factors, the impacts of the input transition time to the delay are more prominent. Secondly, a big load capacitance makes it easy to calibrate the transistor conductance. The time constant for a step input can be approximated as $\tau = R(C_{\text{int}} + C_{load})$. If the load capacitance is much bigger than the intrinsic capacitances like source/drain diffusion capacitance, the time constant can be written as $\tau \approx RC_{load}$. Thus, if the delay for a step input is known from the SPICE simulation, a rough estimation of the quantity of transistor resistance can be obtained by dividing the load capacitance. In this way, we can calibrate the model parameters that were found by the transistor $I-V$ characteristics diagram.

k is the ratio of PMOS transistor width to NMOS transistor width with the value of 1, 2, 4, and 0.5. One may choose any widths of PMOS and NMOS for a given ratio. Instead we make some restrictions on the selection of transistors width when doing the simulation and the delay computation for a better model evaluation purpose. For a rising input, the output node capacitance is discharged through the pull down circuit (NMOS). NMOS transistor plays a more important role in the delay calculation. Theoretically speaking, for different ratios if the NMOS sizes are all the same the delay for a step input will be about the same. Thus, the NMOS transistor width is selected as a constant 2.4um while the PMOS transistor width varying from 1.2um, 2.4um, 4.8um, to 9.6um. For a

falling input, the output node gets charged through the pull-up circuit (PMOS) during the transition. PMOS is a dominant factor in delay computation. For a falling step input the delay for all the ratios should be the same if the PMOS width is the same. In the falling input simulation and model calculation, PMOS width keeps constant as 4.8um while the NMOS width varying from 1.2um, 2.4um, 4.8um, to 9.6um. In this way, the piecewise linear model parameter G is calibrated more accurately.

4.2.1.1 Delay Results Comparison

The delay calculated here is the mid-point delay, i.e. the delay from the mid-point of the output voltage waveform to the mid-point of the input voltage waveform. Fig. 4.7 and 4.8 illustrate the SPICE simulation results and the delay computed by the piecewise linear model for circuit in Fig. 4.6. The input transition time varies in a wide range from a step input ($t_{Tin} = 10\,ps$) to a very slow input ($t_{Tin} = 2500\,ps$). The width ratio of PMOS transistor to NMOS transistor changes from 0.5 to 4. The piecewise linear model shows a great agreement overall with respect to the SPICE simulation results. The overall average relative error of the piecewise linear model is -1.34% of SPICE. The maximum relative error of the piecewise linear model is -10.35% of SPICE.

From Fig. 4.7 and 4.8, one can see that it is important to include the effects of input slope in the circuit performance evaluation. Without considering the input slope, the delay for different input transition time is just a constant value which is equal to the delay of a step input (delay in Fig. 4.7 and 4.8 where $t_{Tin} = 0\,ps$).

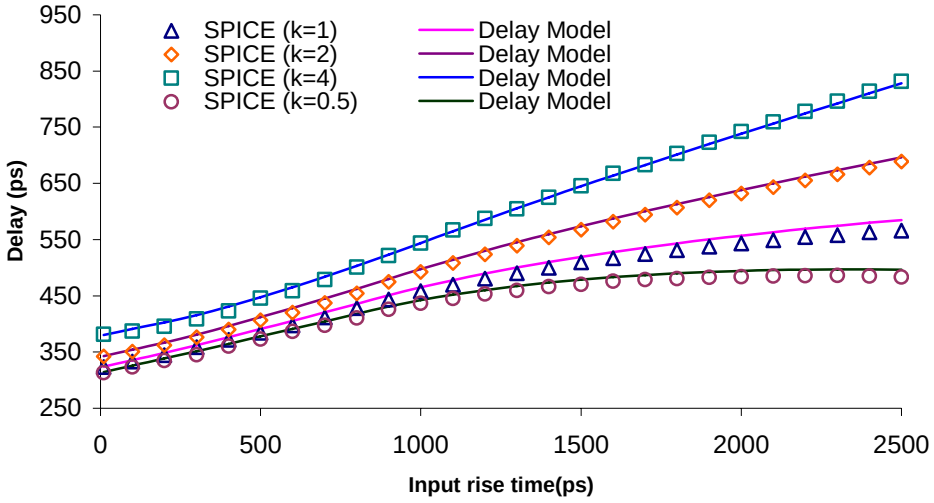

Fig.4. 7 Delay comparison of an inverter with rising input and constant capacitance load.

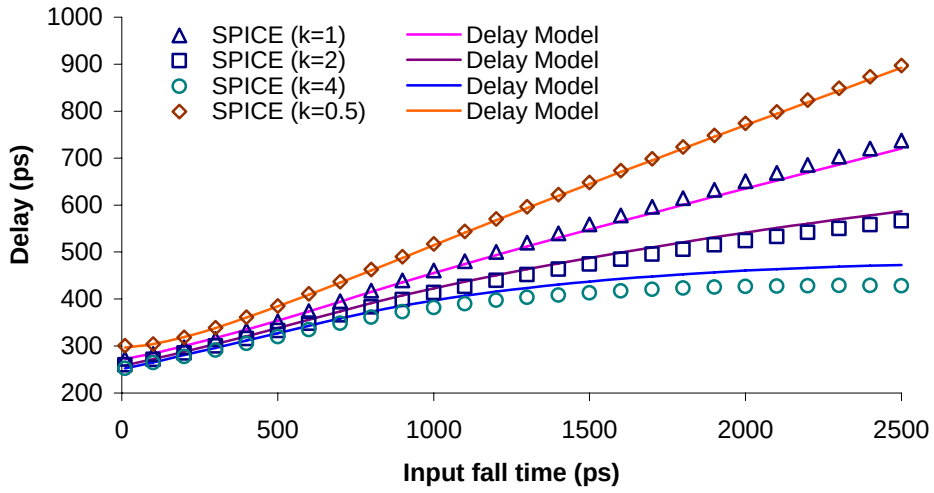

Fig.4. 8 Delay comparison of an inverter with falling input and constant capacitance load

The delays of the step input with different transistor ratios do not converge into a single point in SPICE simulations even though the pull-down transistor sizes are the same. This is because although the load capacitance is relatively big, the contributions of

the gate-drain capacitance and the varying drain diffusion capacitances due to the different PMOS sizes are not trivial. By including the input capacitance model, the piece-wise linear model has a better fit for a step input.

The sheet resistances of the switching on transistors are calibrated according to the SPICE simulation of a step input. Different sets of sheet resistances are used for rising and falling input transitions in order to have a better fit. In the piece-wise linear mode, sheet resistance of NMOS, R_{sn}, for a rising input transition is equal to $1.3\times10^{4}\Omega/\square$. The sheet resistance of PMOS, R_{sp}, for a falling input transition is equal to $2.1\times10^{4}\Omega/\square$.

4.2.1.2 Output Rise/Fall Time

The SPICE simulation results for rise time and fall time shown here are not the traditional definitions. Instead, they follow the definitions in Eq. (4.8) and (4.9) for the piecewise linear model. Fig. 4.9 and 4.10 plot the output rise time and fall time calculated by the piecewise linear model and its corresponding SPICE simulation results for the circuit illustrated in Fig. 4.6. The model agrees well in general with the SPICE with average relative error of -4.81% and maximum relative error of -10.24% in the range of $t_{Tin}=0\,ps$ to $t_{Tin}=1200\,ps$.

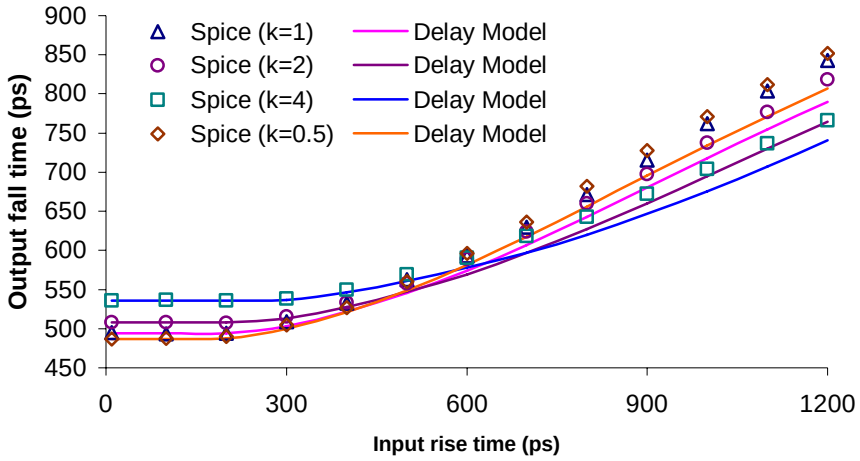

Fig.4. 9 Output fall time comparison of an inverter with falling input and constant capacitance load.

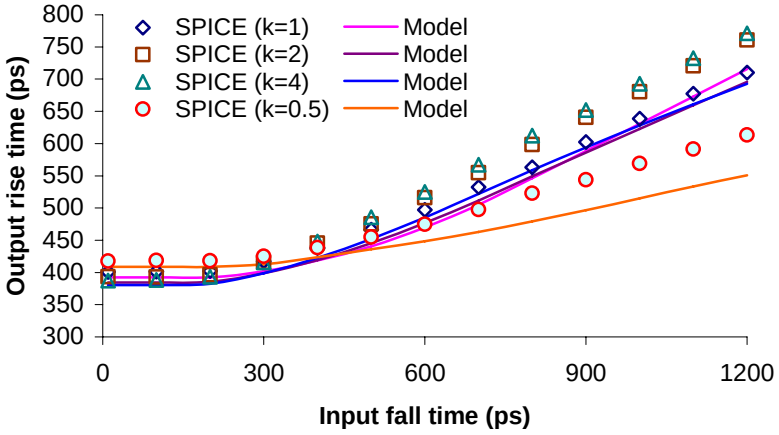

Fig.4. 10 Output rise time comparison of an inverter with falling input and constant capacitance load.

4.2.1.3 Output Waveform

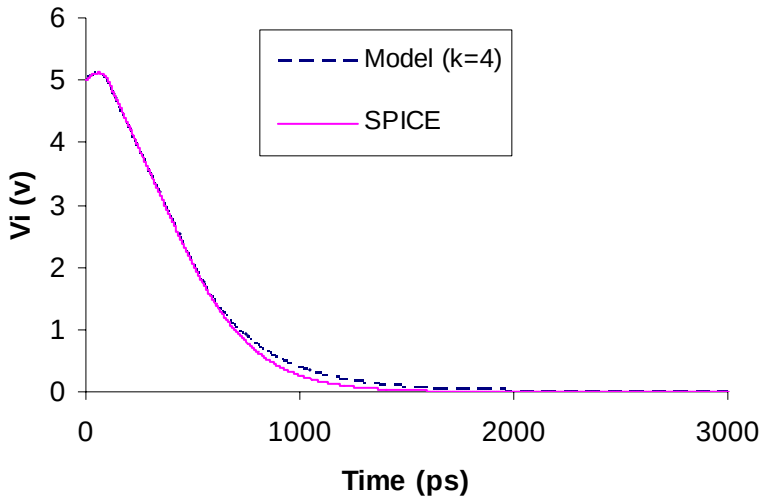

Fig.4. 11 Output waveform comparison of an inverter with rising input and constant capacitance load. $t_{Tin} = 100\,ps$.

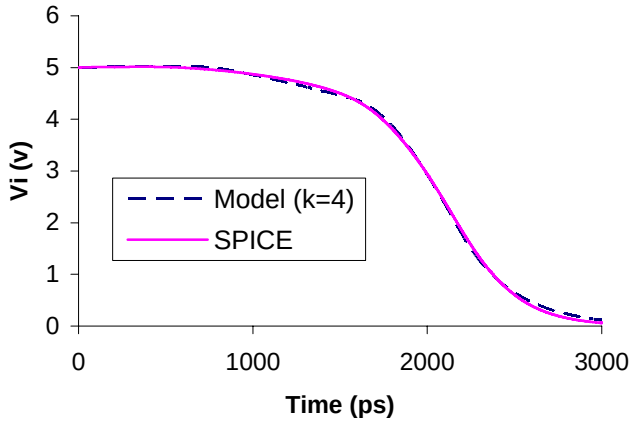

Fig.4. 12 Output waveform comparison of an inverter with rising input and constant capacitance load. $t_{Tin} = 2500\,ps$.

Fig. 4.11 and Fig. 4.12 plot the output waveforms of SPICE simulation and piecewise linear model prediction for the circuit shown in Fig. 4.6. The width ratio of PMOS transistor to NMOS transistor is $k = 4$. Two extreme cases are plot here, one is with very fast input transition ($t_{Tin} = 100\,ps$) and another is with very slow input

transition ($t_{Tin} = 2500\,ps$). The piecewise linear model is pretty accurate in predicting the output waveform for a very large range of rising input transition.

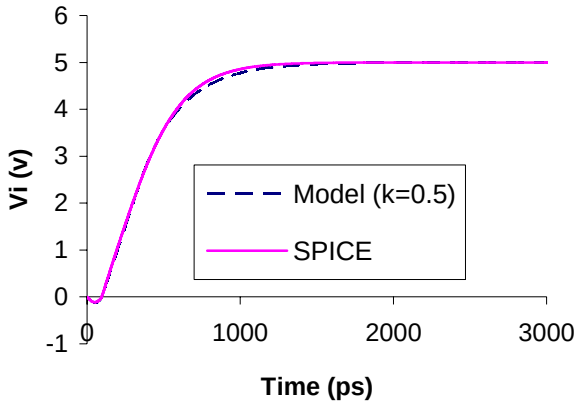

Fig.4. 13 Output waveform comparison of an inverter with falling input and constant capacitance load. $t_{Tin} = 100\,ps$

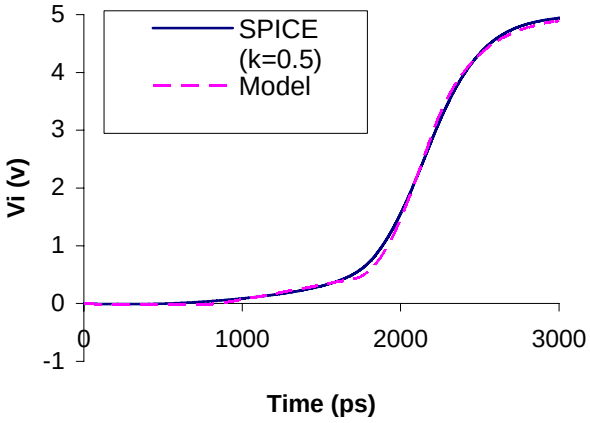

Fig.4. 14 Output waveform comparison of an inverter with falling input and constant capacitance load. $t_{Tin} = 2500\,ps$

In Fig. 4.13 and Fig. 4.14, we plot the output waveforms of a very fast falling input transition (100ps) and a very slow falling input transition (2500ps) which are predicted by the piecewise linear model together with their corresponding SPICE

simulation results. The width ratio of PMOS transistor to NMOS transistor is $k = 0.5$. The piecewise linear model is pretty accurate in predicting the output waveform for a very large range of rising input transition. Again, the model is very accurate in predicting the output waveform just like the rising input transition case.

4.2.2 An Inverter Drives a Constant Inverter Load

The schematic diagram of the circuit is shown in Fig. 4.15:

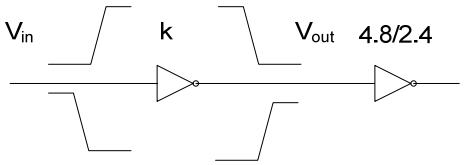

Fig.4. 15 An inverter drives a constant inverter load

Just like the previous case, the driver inverter is varying the ratio k . And for a rising input transition, the pull-down circuits are of the same size for different k values. For a falling input transition, the pull-up circuits are of the same size. The load inverter used has a constant PMOS to NMOS width ratio $k = \frac{4.8um}{2.4um}$. The load capacitance is not a dominant factor any more since the load inverter size is comparable to the driver inverter size. The drain diffusion capacitance now has more impacts to the delay than the previous case. Thus, the circuit in Fig. 4.15 is suitable for calibrating the source/drain diffusion capacitance model.

4.2.2.1 Delay Results Comparison

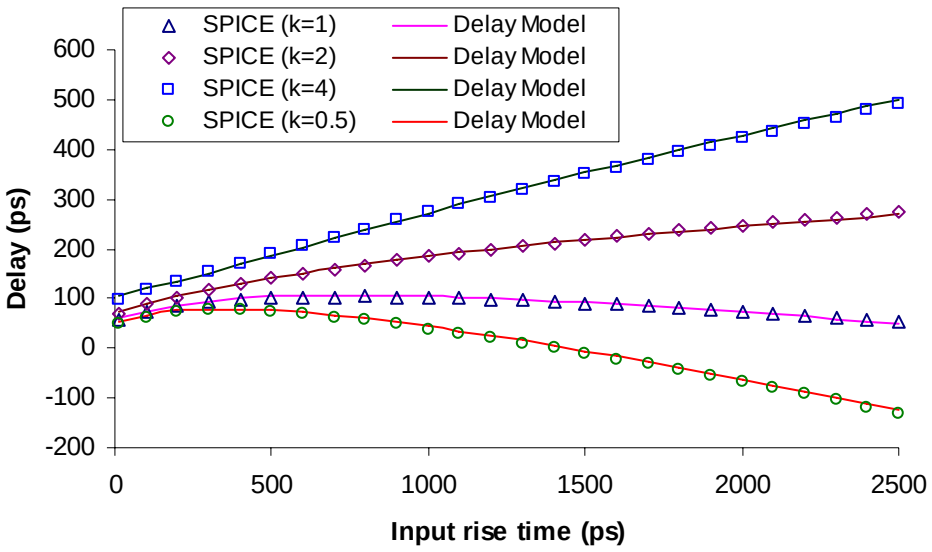

Fig.4. 16 Delay comparison of an inverter with rising input and constant inverter load

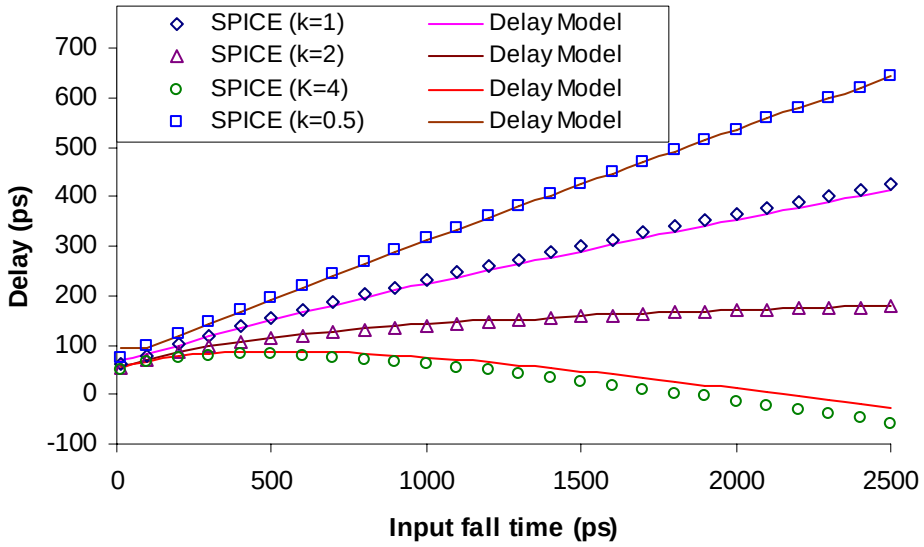

Fig.4. 17 Delay comparison of an inverter with falling input and constant inverter load.

Fig. 4.16 and 4.17 are the delay results obtained by the SPICE simulation and the piecewise linear model computation for the circuit described in Fig. 4.15. It is obvious that one cannot ignore the input slope effect when models the delay. It is impossible for a

simple step input assumption to model the delays which span from negative -100*ps* to 600*ps*. Even the linear assumption that the delay is just proportional to the input slope like most delay model did is not accurate. The piecewise linear model can: 1. predict the delay in a large input transition range. 2. predict the negative delay. 3. predict the curvature of the delay curve. The average absolute error predicted by the piecewise linear model for the circuit in Fig. 4.15 is 1.83*ps* and the maximum absolute error of the model is -30.34*ps*.

4.2.2.2 Output Rise/Fall Time

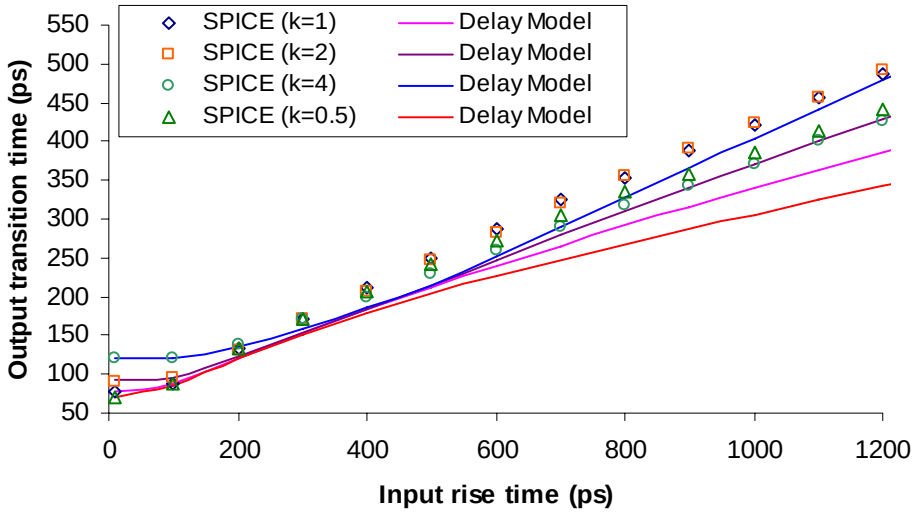

Fig.4. 18 Output fall time comparison of an inverter with falling input and constant inverter load.

Fig. 4.18 and Fig. 4.19 show the output fall time and rise time of the circuit in Fig. 4.15. The input transition is in the range of 10*ps* to 1200*ps*. The average relative output rise/fall time error is -9.74% and the maximum relative error is -27% with respect to SPICE simulation.

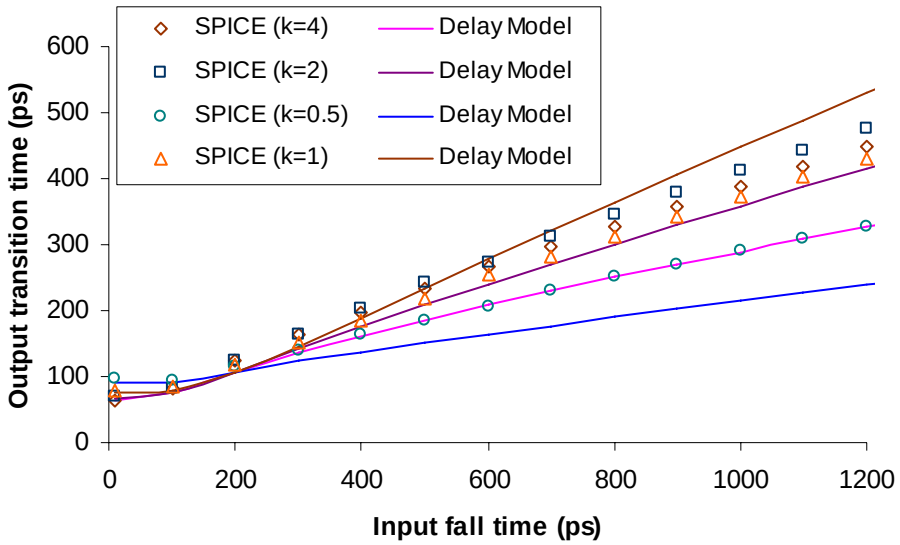

Fig.4. 19 Output rise time comparison of an inverter with falling input and constant inverter load.

4.2.2.3 Output Waveform

Fig. 4.20 through Fig. 4.23 plot the output rise and fall waveforms of the circuit in Fig. 4.15. As in the previous case, a very fast input transition (100ps) and a very slow input transition (2500ps) are used to test the accuracy of the piecewise linear model. The width ratio of PMOS transistor to NMOS transistor is $k = 0.5$ and $k = 4$ respectively. The piecewise linear model fits the SPICE simulation very well for a very large range of input slopes.

4.3 Summary

The input transitions fall into two categories: fast input and slow input, which are distinguished by different output transition regions they underwent. These regions are formed by the boundary lines which decide the transistors' operation state. The circuit

dynamics in each region is approximated by its quasi-steady state solution and the initial conditions in each region.

The simplest CMOS circuit, an inverter, is utilized to fulfill two purposes: 1. to check the accuracy of the piecewise linear model. 2. to calibrate the piecewise linear model so that it will fit the SPICE simulation results within a reasonable error range.

An inverter drives a big constant capacitance load, Fig. 4.6, gives us an overview of the importance of the input slope to the delay modeling and also provides a rough estimation on the sheet resistances. The source/drain diffusion capacitance model is calibrated via the circuit shown in Fig. 4.15. The piecewise linear model agrees well with the SPICE simulation. The model shows that the errors are within 5% or 17ps of the SPICE simulations.

The piecewise linear model shows relatively bigger errors when calculates the output rise time and fall time. This is because the rise time and fall time are defined by the first half of the output transition in the piecewise linear model since the model has a better fit during this portion of the transition. This is not the normal definition of the rise time and fall time though. A better approach needs to be found in this category.

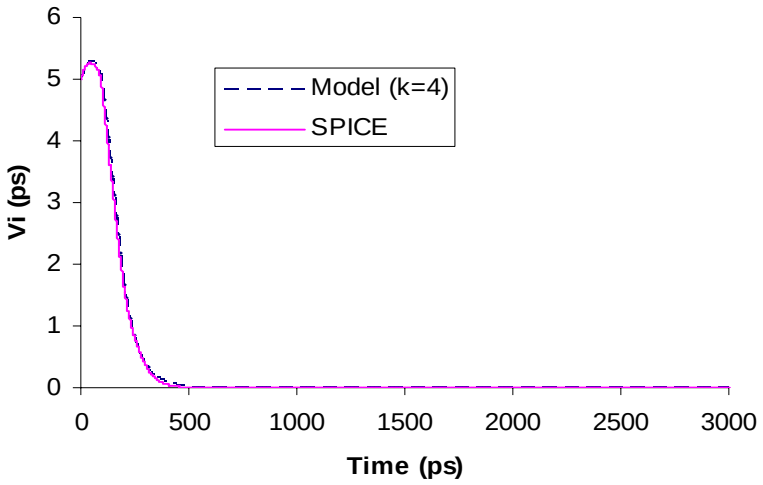

Fig.4. 20 Output waveform comparison of an inverter with rising input and constant inverter load. $t_{Tin} = 100\,ps$

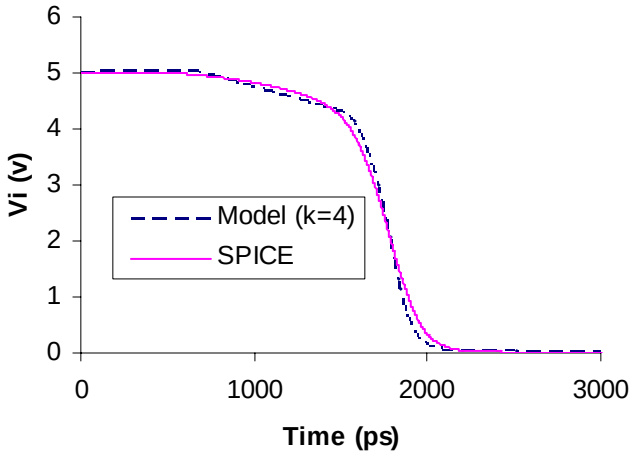

Fig.4. 21 Output waveform comparison of an inverter with rising input and constant inverter load. $t_{Tin} = 2500\,ps$

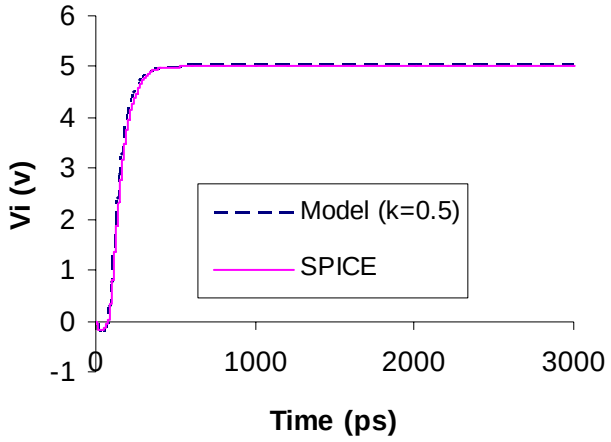

Fig.4. 22 Output waveform comparison of an inverter with falling input and constant inverter load. $t_{Tin} = 100\,ps$

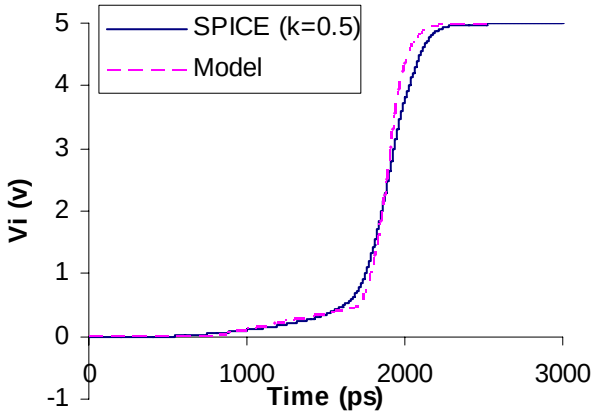

Fig.4. 23 Output waveform comparison of an inverter with falling input and constant inverter load. $t_{Tin} = 2500\,ps$

CHAPTER V

MORE CIRCUITS ON MODEL APPLICATION

In Chapter IV, the piecewise linear model is applied to the inverter delay analysis. The model shows great agreement with regard to the SPICE simulation. The piecewise linear model is applicable not only to propagation delay of a simple gate, such as an inverter but also to that of any general circuit topology. The model is also scalable by using different set of model parameters for different technologies. And the model is suitable to a fast simulator application.

In this chapter, we extend the application of piecewise linear model to two-input NAND gate and OAI gate. Similar to the inverter analysis, delays and output rise and fall time predicted by the model are compared to the corresponding SPICE simulation by varying the input slope, switching gate, transistor size, and the loading factors.

5.1 Two-Input NAND Gate Analysis

5.1.1 Two-Input NAND Gate with Constant Capacitance Load

Fig. 5.1 shows a CMOS implementation of two-input NAND gate with input signals *A* and *B*. The gate drives a constant capacitance load, $C_{load} = 100\,fF$. We will assume that only one gate is switching at any time. Multiple inputs can be handled by making V_{in} a column matrix. The pull-up transistors are of the same size and so are the

pull-down transistors. The width ratio of PMOS transistors to NMOS transistors varies from $k = 0.5$, 1, 2 to 4.

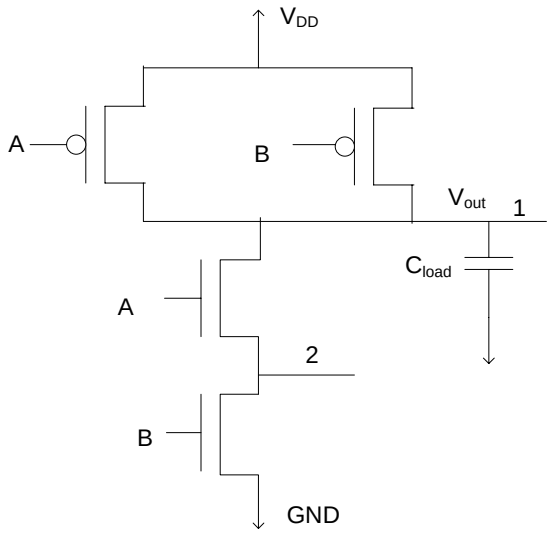

Fig.5. 1 Two-input NAND gate with constant capacitance load

5.1.1.1 $A = 0 \rightarrow 1,\ B = 1$

In this case, the input node *A* of the upper NMOS transistor is switching from low to high, a rising input transition, and the node B of the lower NMOS transistor stays at $V_B = V_{DD}$.

Fig. 5.2 shows the delay versus the input rise time of the circuit described above. The input rise time varies from $10\,ps$ to $2500\,ps$. The width ratio of PMOS transistors to NMOS transistors varies from $k = 0.5$, 1, 2 to 4. The average relative error of the piecewise linear model is -1.69% with respect to SPICE simulation. The maximum relative error is 6.88%.

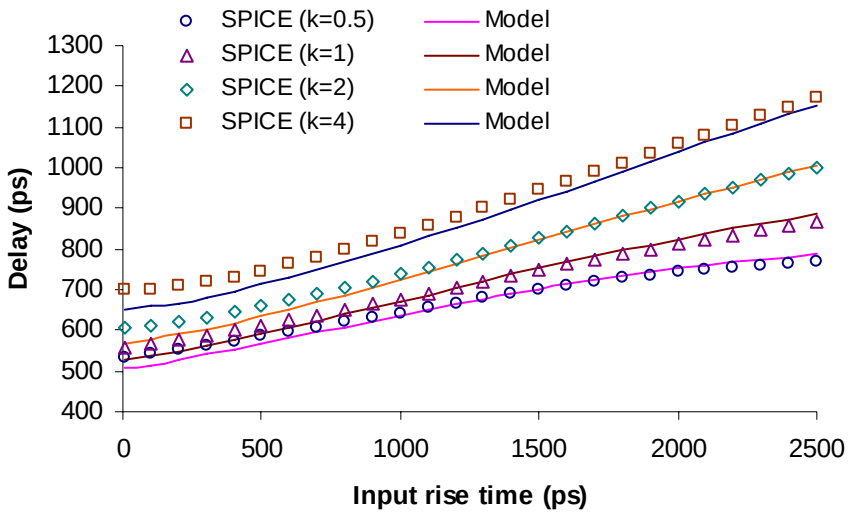

Fig.5. 2 Delay comparison a two-input NAND with rising input and constant capacitance load, $A = 0 \rightarrow 1,\ B = 1$

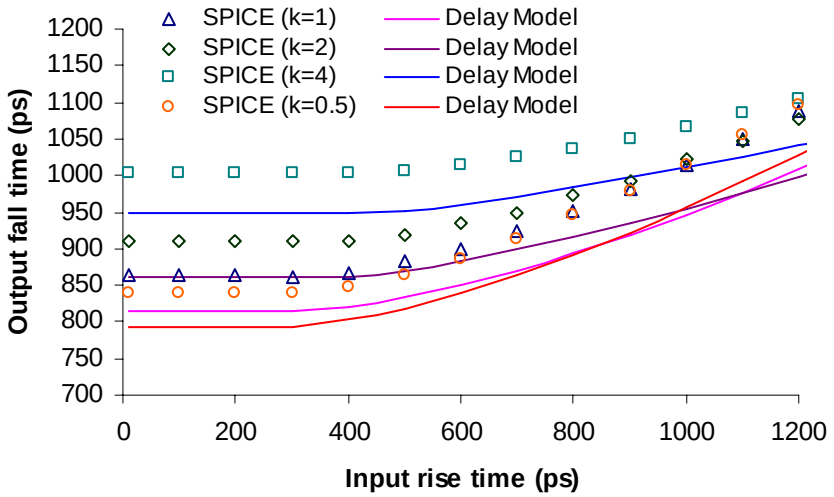

Fig.5. 3 Output fall time comparison of a two-input NAND with rising input and constant capacitance load, $A = 0 \rightarrow 1,\ B = 1$.

The output fall time versus the input rise time of the same transition is shown in Fig. 5.3. The input rise time varies from $10\,ps$ to $1200\,ps$. The average relative error of

the piecewise linear model is -5.68% with respect to SPICE simulation. The maximum relative error is -7.34%.

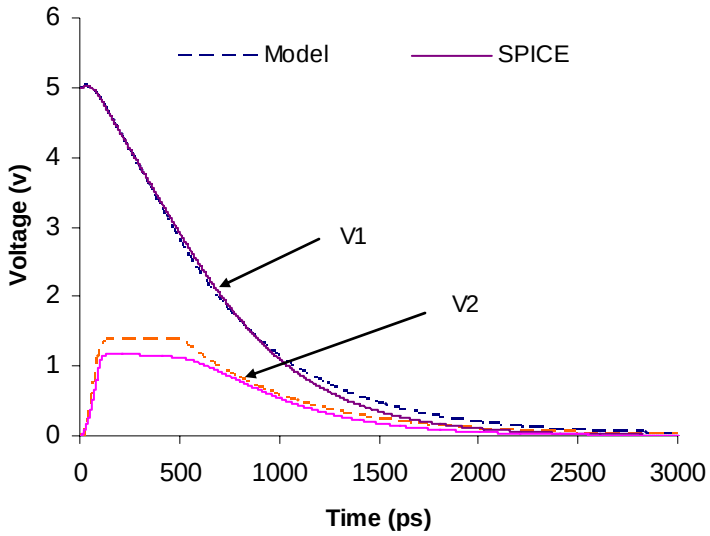

Fig.5. 4 Output waveforms comparison of a two-input NAND with rising input and constant capacitance load. $t_{Tin} = 100\,ps$. $A = 0 \rightarrow 1,\ B = 1$

In Fig. 5.4 and 5.5, the output waveforms of the output node ($V1$) and the internal node ($V2$) of a fast input transition ($t_{Tin} = 100\,ps$) and a slow input transition ($t_{Tin} = 1500\,ps$) are plot. The piecewise linear model is very accurate in predicting the waveforms of the circuit nodes over a wide range.

5.1.1.2 $A = 1 \rightarrow 0,\ B = 1$

In this case, the input node A of the upper NMOS transistor is switching from high to low, a falling input transition, and the node B of the lower NMOS transistor stays at $V_B = V_{DD}$.

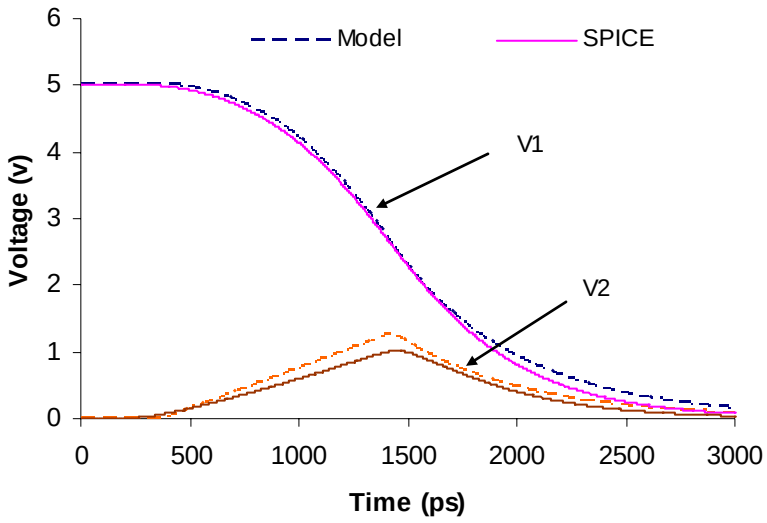

Fig.5. 5 Output waveforms comparison of a two-input NAND with rising input and constant capacitance load. $t_{Tin} = 1500\,ps$. $A = 0 \rightarrow 1,\ B = 1$

Fig. 5.6 shows the delay versus the input fall time of the circuit described above. The input fall time varies from $10\,ps$ to $2500\,ps$. The widths of the PMOS transistors are chosen as $4.8um$ and the widths of the NMOS transistors vary from $W_n = 1.2um$, $2.4um$, $4.8um$, and $9.6um$. Hence, the width ratio of PMOS transistors to NMOS transistors varies from $k = 0.5$, 1, 2 to 4. The average relative error of the piecewise linear model is 0.38% with respect to SPICE simulation. The maximum relative error is 11.4%.

The output rise time versus the input fall time of the same transition is shown in Fig. 5.7. The input fall time varies from $10\,ps$ to $1200\,ps$. The average relative error of the piecewise linear model is -5.58% with respect to SPICE simulation. The maximum relative error is -10.5%.

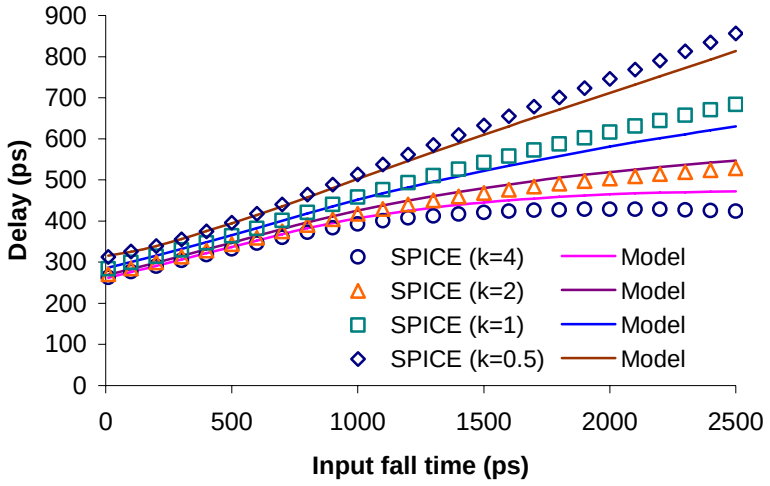

Fig.5. 6 Delay comparison of two-input NAND with falling input and constant capacitance load, $A=1\rightarrow 0$, $B=1$

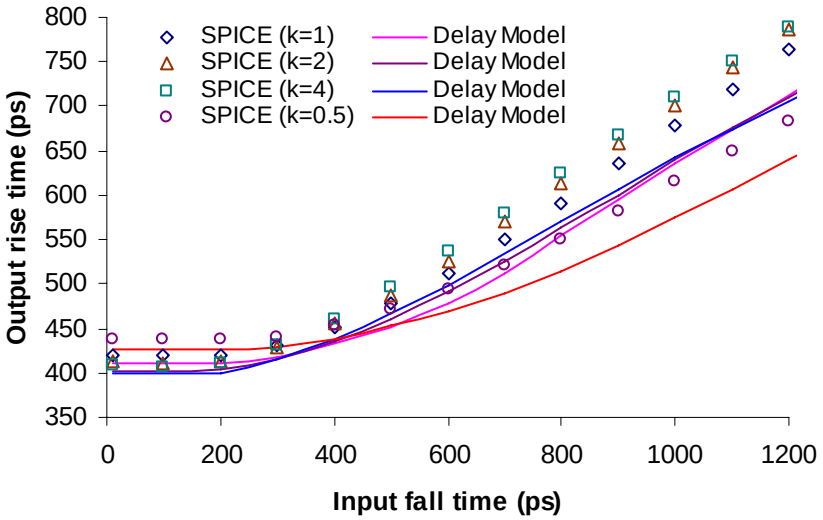

Fig.5. 7 Output rise time comparison of a two-input NAND with falling input and constant capacitance load, $A=1\rightarrow 0$, $B=1$.

Fig. 5.8 and 5.9 plot the output waveforms of the output node ($V1$) and the internal node ($V2$) of a fast input transition ($t_{Tin}=100\,ps$) and a slow input transition

($t_{Tin} = 2500\,ps$). The piecewise linear model is capable of predicting the waveforms of the circuit nodes over a wide range of input slope.

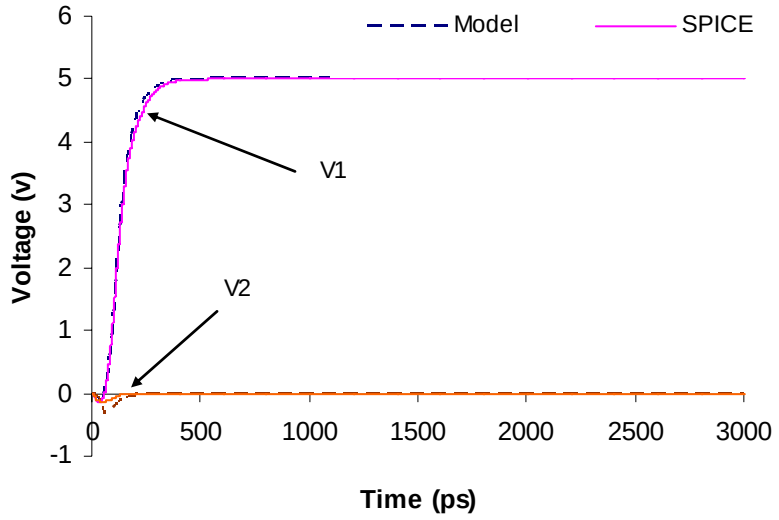

Fig.5. 8 Output waveforms comparison of a two-input NAND with falling input and constant capacitance load. $t_{Tin} = 100\,ps$. $A = 1 \rightarrow 0$, $B = 1$

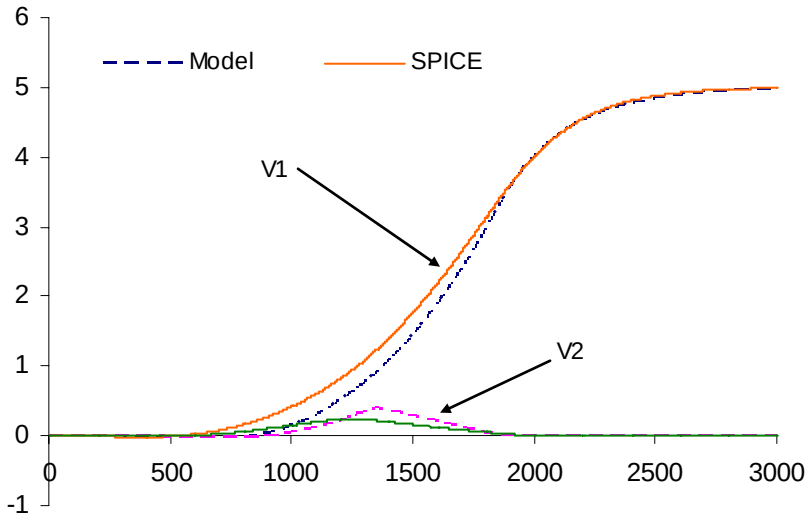

Fig.5. 9 Output waveforms comparison of a two-input NAND with falling input and constant capacitance load. $t_{Tin} = 2500\,ps$. $A = 1 \rightarrow 0$, $B = 1$

5.1.1.3 $A=1,\ B=0\rightarrow 1$

The input node *A* of the upper NMOS transistor stays at $V_A = V_{DD}$. The node B of the lower NMOS transistor switches from low to high.

Fig. 5.10 shows the delay versus the input rise time of the transition described above. The input rise time varies from $10\,ps$ to $2500\,ps$. The widths of the NMOS transistors are chosen as $2.4um$ and the widths of the PMOS transistors vary from $W_p = 1.2um$, $2.4um$, $4.8um$, and $9.6um$. Hence, the width ratio of PMOS transistors to NMOS transistors varies from $k = 0.5$, 1, 2 to 4. The average relative error of the piecewise linear model is -7.31% with respect to SPICE simulation. The maximum relative error is -9.44%.

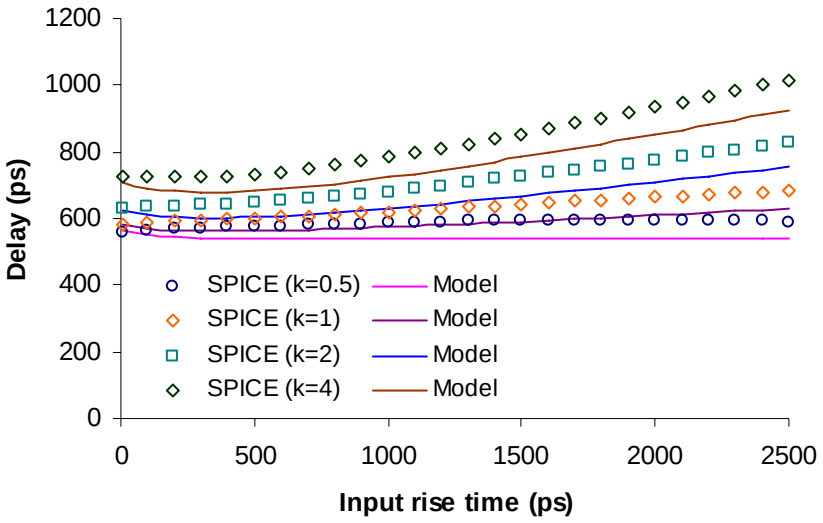

Fig.5. 10 Delay comparison of a two-input NAND with rising input and constant capacitance load, $A=1,\ B=0\rightarrow 1$.

The output fall time versus the input rise time of the same transition is shown in Fig. 5.11. The input fall time varies from $10\,ps$ to $1200\,ps$. The average relative error of

the piecewise linear model is -7.31% with respect to SPICE simulation. The maximum relative error is -10.8%.

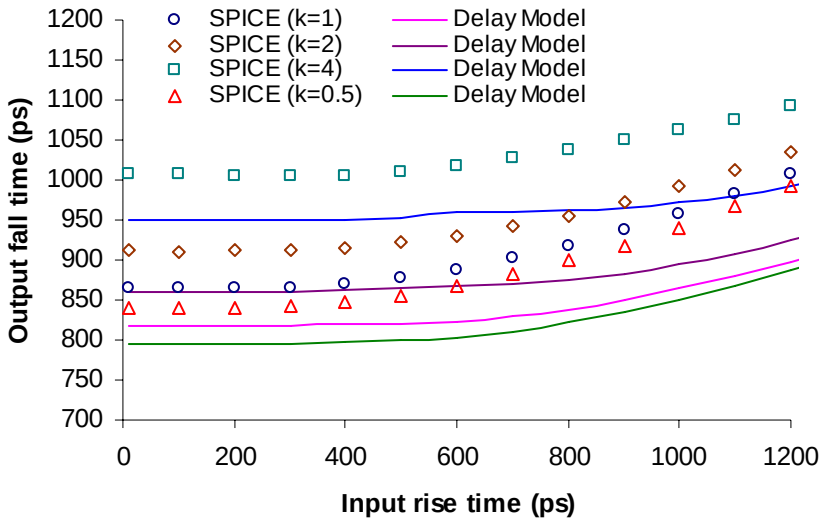

Fig.5. 11 Output fall time comparison of a two-input NAND with rising input and constant capacitance load, $A=1$, $B=0\rightarrow 1$.

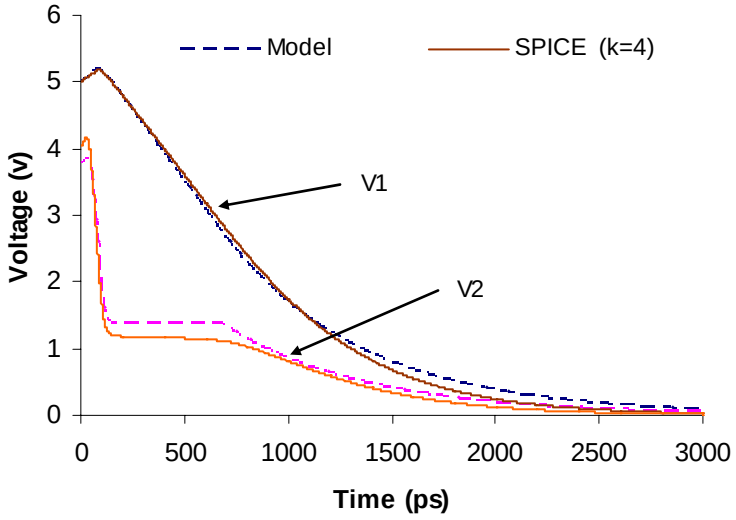

Fig.5. 12 Output waveforms comparison of a two-input NAND with rising input and constant capacitance load. $t_{Tin}=100\,ps$. $A=1$, $B=0\rightarrow 1$.

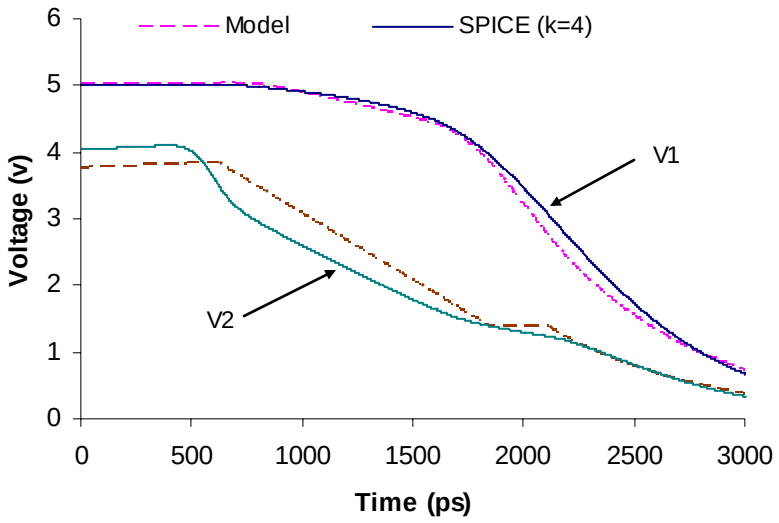

Fig.5. 13 Output waveforms comparison of a two-input NAND with rising input and constant capacitance load. $t_{Tin} = 2500\,ps$. $A = 1$, $B = 0 \rightarrow 1$.

Fig. 5.12 and 5.13 plot the voltage output transient response of the output node ($V1$) and the internal node ($V2$) of a fast input transition ($t_{Tin} = 100\,ps$) and a slow input transition ($t_{Tin} = 2500\,ps$).

5.1.1.4 $A = 1$, $B = 1 \rightarrow 0$

The input node A of the upper NMOS transistor stays at $V_A = V_{DD}$. The node B of the lower NMOS transistor switches from high to low.

Fig. 5.14 shows the delay versus the input fall time of the transition described above. The input fall time varies from $10\,ps$ to $2500\,ps$. The widths of the PMOS transistors are chosen as $4.8um$ and the widths of the NMOS transistors vary from $W_n = 1.2um$, $2.4um$, $4.8um$, and $9.6um$. Hence, the width ratio of PMOS transistors to NMOS transistors varies from $k = 0.5$, 1, 2 to 4. The average relative error of the

piecewise linear model is -0.28% with respect to SPICE simulation. The maximum relative error is 9.7%.

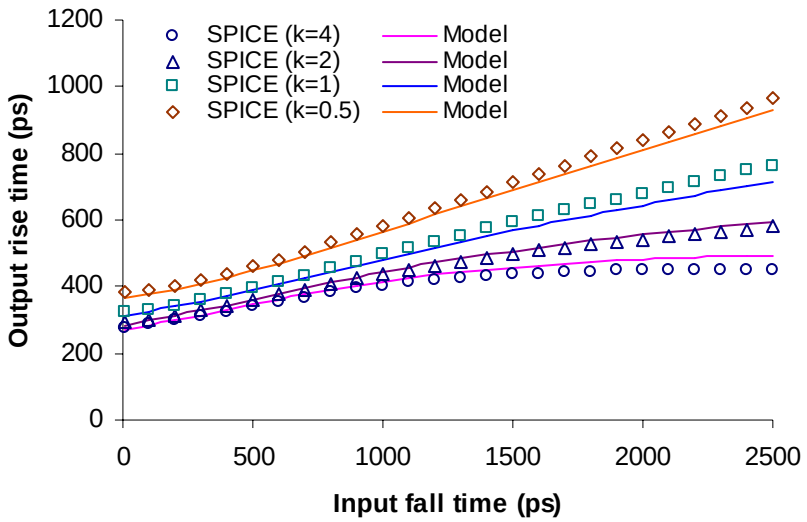

Fig.5. 14 Delay comparison of a two-input NAND with falling input and constant capacitance load, $A=1$, $B=1\rightarrow 0$

The output rise time versus the input fall time of the same transition is shown in Fig. 5.15. The input fall time varies from $10\,ps$ to $1200\,ps$. The average relative error of the piecewise linear model is -4.14% with respect to SPICE simulation. The maximum relative error is -10.11%.

Fig. 5.16 and 5.17 plot the voltage output transient response of the output node ($V1$) and the internal node ($V2$) of a fast input transition ($t_{Tin}=100\,ps$) and a slow input transition ($t_{Tin}=2500\,ps$).

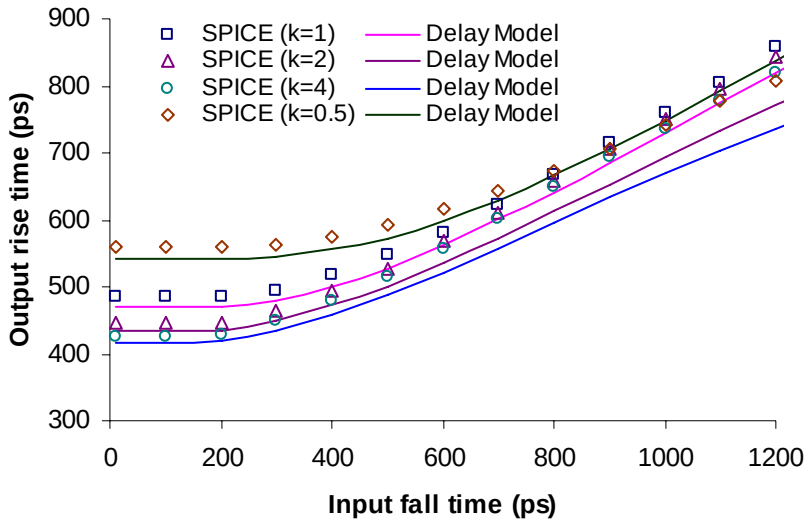

Fig.5. 15 Output rise time comparison of a two-input NAND with falling input and constant capacitance load, $A=1$, $B=1\rightarrow 0$.

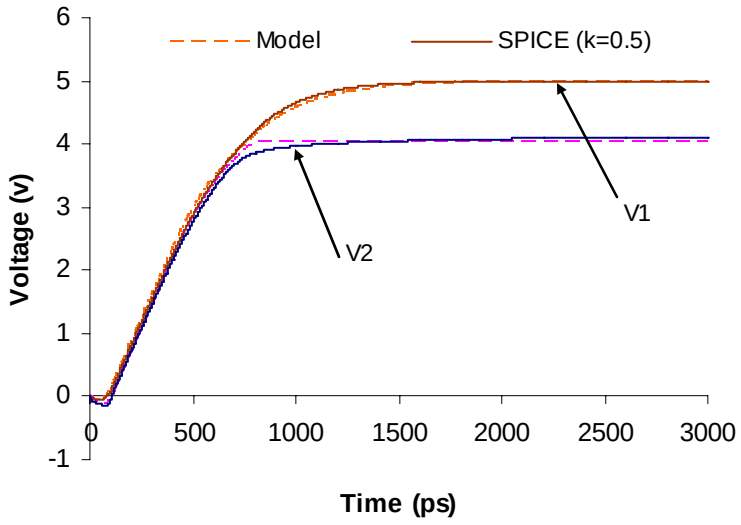

Fig.5. 16 Output waveforms comparison of a two-input NAND with rising input and constant capacitance load. $t_{Tin}=100\,ps$. $A=1$, $B=1\rightarrow 0$.

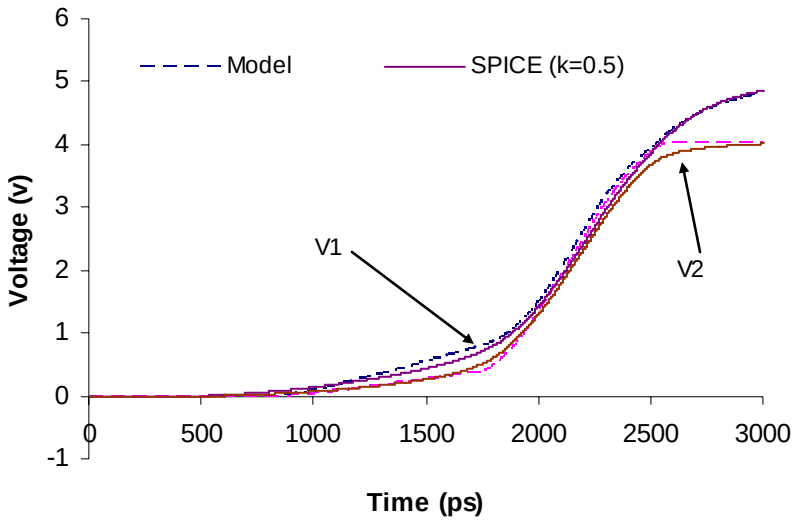

Fig.5. 17 Output waveforms comparison of a two-input NAND with rising input and constant capacitance load. $t_{Tin} = 2500\,ps$. $A = 1$, $B = 1 \rightarrow 0$.

5.1.2 Two-Input NAND Gate with Constant Inverter Load

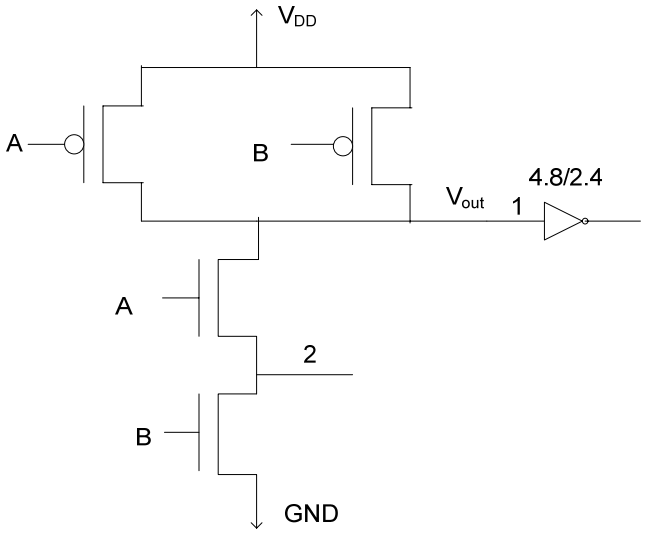

Fig.5. 18 Two-input NAND gate with constant inverter load

Fig. 5.18 shows a CMOS implementation of two-input NAND gate with input signals *A* and *B*. The gate drives a constant inverter load, $W_p / W_n = 4.8um / 2.4um$. This

circuit gives a more realistic load effect than the constant capacitance case. We will assume that only one gate is switching at any time. Multiple inputs can be handled by making V_{in} a column matrix. The pull-up transistors are of the same size and so are the pull-down transistors. The width ratio of PMOS transistors to NMOS transistors varies from $k = 0.5$, 1, 2 to 4. The test procedures are the same as the previous circuit. We alternate the switching gates, vary the transition (rise or fall), and change the input slope.

We introduce the absolute error in this section to evaluate the model accuracy because for a small inverter load, the delay is a relative small number (in picoseconds). It makes more sense to find the absolute error between model and SPICE simulation in this case. For comparison purpose, we will list both the relative error and absolute error.

5.1.2.1 $A = 0 \rightarrow 1,\ B = 1$

The input node *A* of the upper NMOS transistor switches from low to high. The node B of the lower NMOS transistor stays at $V_B = V_{DD}$.

Fig. 5.19 shows the delay versus the input rise time of the transition described above. The input rise time varies from $10\,ps$ to $2500\,ps$. The widths of the PMOS transistors are chosen as $4.8um$ and the widths of the NMOS transistors vary from $W_n = 1.2um$, $2.4um$, $4.8um$, and $9.6um$. Hence, the width ratio of PMOS transistors to NMOS transistors varies from $k = 4$, 2, 1 to 0.5. The average relative error of the piecewise linear model is -3.63% with respect to SPICE simulation. The maximum relative error is 50.7%. The average absolute error of the piecewise linear model is 3.20ps with respect to SPICE simulation. The maximum absolute error is 19.3ps.

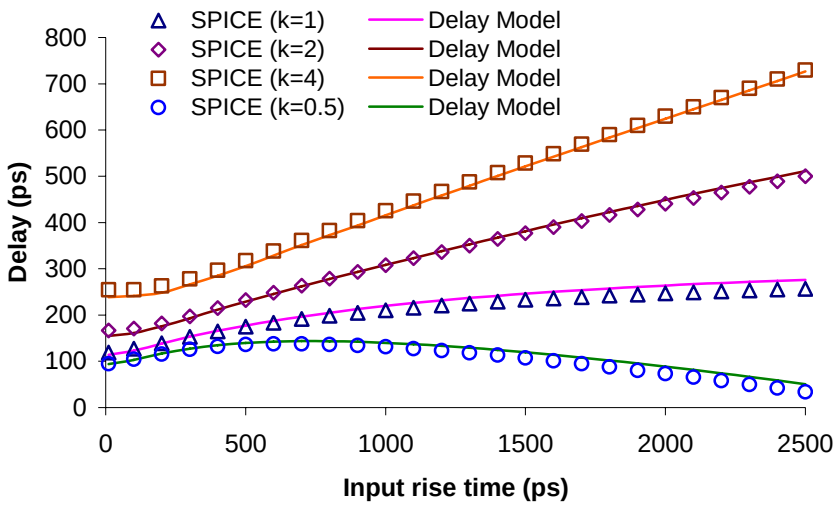

Fig.5. 19 Delay comparison of a two-input NAND with rising input and constant inverter load, $A = 0 \rightarrow 1,\ B = 1$

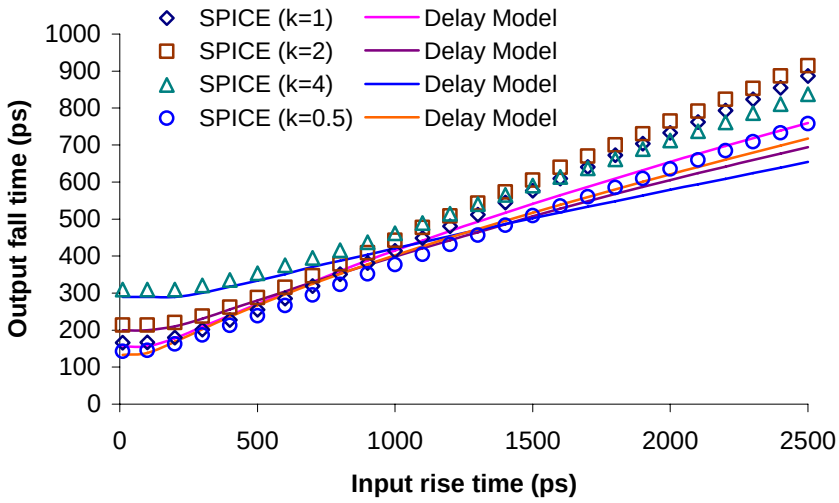

Fig.5. 20 Output fall time comparison of a two-input NAND with rising input and constant inverter load, $A = 0 \rightarrow 1,\ B = 1$.

The output fall time versus the input rise time of the same transition is shown in Fig. 5.20. The input rise time varies from $10\,ps$ to $2500\,ps$. The average relative error of the piecewise linear model is -1.73% with respect to SPICE simulation. The maximum

relative error is -12.9%. The average absolute error of the piecewise linear model is -8.1*ps* with respect to SPICE simulation. The maximum absolute error is -65.8*ps*.

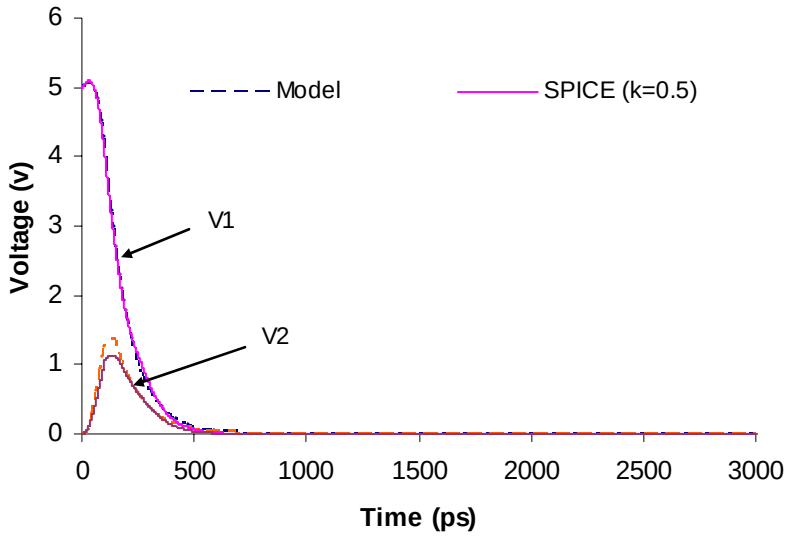

Fig.5. 21 Output waveforms comparison of a two-input NAND with rising input and constant inverter load. $t_{Tin} = 100\,ps$. $A = 0 \rightarrow 1,\ B = 1$.

Fig. 5.21 plots the voltage output transient response of the output node ($V1$) and the internal node ($V2$) of an input transition time of $t_{Tin} = 100\,ps$.

5.1.2.2 $A = 1 \rightarrow 0,\ B = 1$

The input node A of the upper NMOS transistor switches from high to low, a falling input transition. The node B of the lower NMOS transistor stays at $V_B = V_{DD}$.

Fig. 5.22 shows the delay versus the input fall time of the transition described above. The input fall time varies from $10\,ps$ to $2500\,ps$. The widths of the PMOS transistors are chosen as $4.8um$ and the widths of the NMOS transistors vary from $W_n = 1.2um,\ 2.4um,\ 4.8um$, and $9.6um$. Hence, the width ratio of PMOS transistors to

NMOS transistors varies from $k = 4$, 2, 1 to 0.5. The average relative error of the piecewise linear model is 8.48% with respect to SPICE simulation. The maximum relative error is 814%. The average absolute error of the piecewise linear model is -10.69*ps* with respect to SPICE simulation. The maximum absolute error is 100.8*ps*.

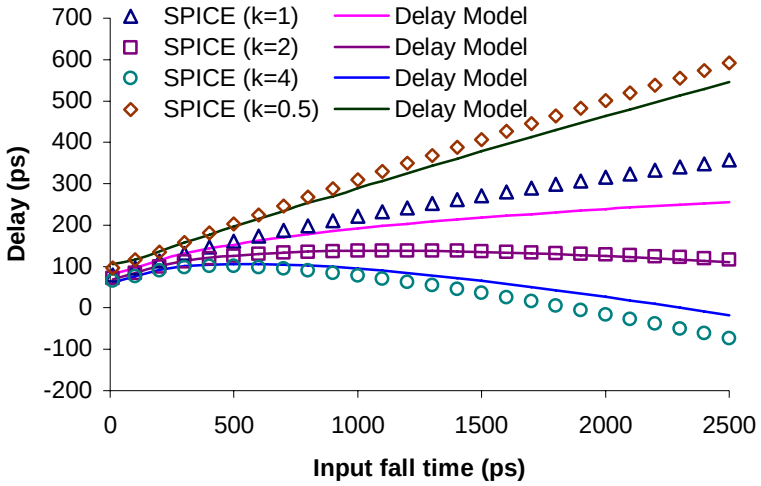

Fig.5. 22 Delay comparison of a two-input NAND with falling input and constant inverter load, $A = 1 \rightarrow 0$, $B = 1$.

The output rise time versus the input fall time of the same transition is shown in Fig. 5.23. The input fall time varies from $10\,ps$ to $1200\,ps$. The average relative error of the piecewise linear model is -12.87% with respect to SPICE simulation. The maximum relative error is -27.2%. The average absolute error of the piecewise linear model is -48.3*ps* with respect to SPICE simulation. The maximum absolute error is -125.3*ps*.

Fig. 5.24 plots the voltage output transient response of the output node ($V1$) and the internal node ($V2$) of an input transition time of $t_{Tin} = 100\,ps$. $k = 4$.

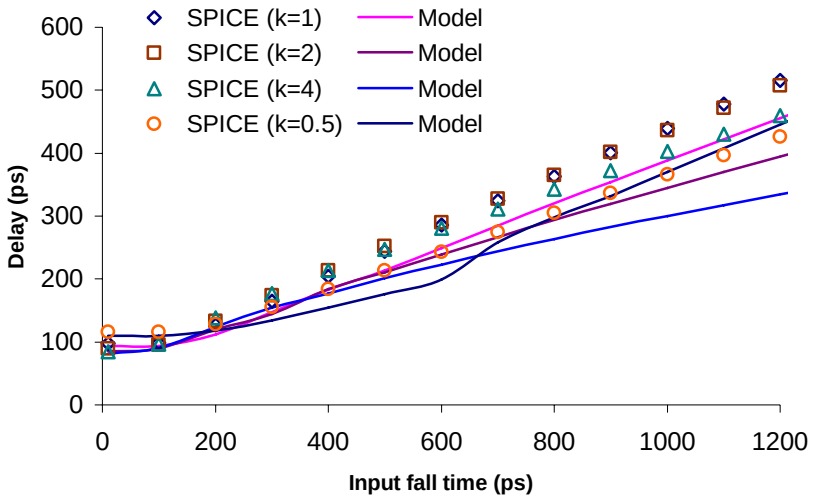

Fig.5. 23 Output fall time comparison of a two-input NAND with rising input and constant inverter load, $A=1\rightarrow 0$, $B=1$.

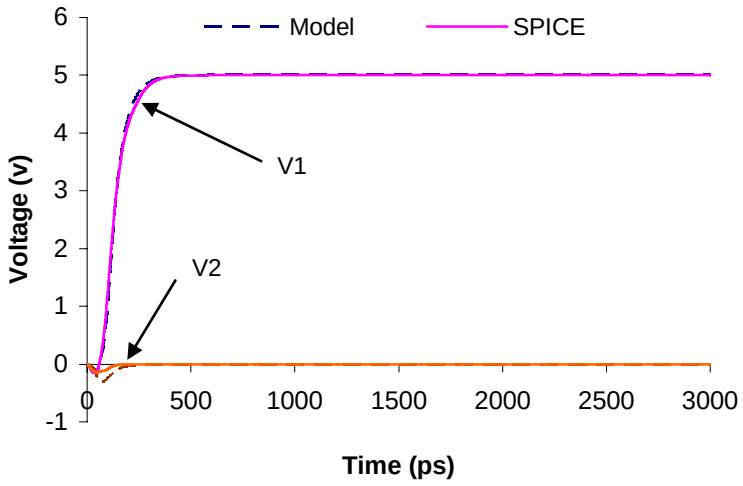

Fig.5. 24 Output waveforms comparison of a two-input NAND with falling input and constant inverter load. $t_{Tin}=100\,ps$. $A=1\rightarrow 0$, $B=1$.

5.1.2.3 $A=1$, $B=0\rightarrow 1$

The input node *A* of the upper NMOS transistor stays at $V_A=V_{DD}$. The node B of the lower NMOS transistor switches from low to high, a rising input transition.

Fig. 5.25 shows the delay versus the input rise time of the transition described above. The input rise time varies from $10\,ps$ to $2500\,ps$. The widths of the NMOS transistors are chosen as $2.4um$ and the widths of the PMOS transistors vary from $W_p = 1.2um$, $2.4um$, $4.8um$, and $9.6um$. Hence, the width ratio of PMOS transistors to NMOS transistors varies from $k = 0.5$, 1, 2 to 4. The average relative error of the piecewise linear model is -4.46% with respect to SPICE simulation. The maximum relative error is 110.1%. The average absolute error of the piecewise linear model is -14.8*ps* with respect to SPICE simulation. The maximum absolute error is 36.2*ps*.

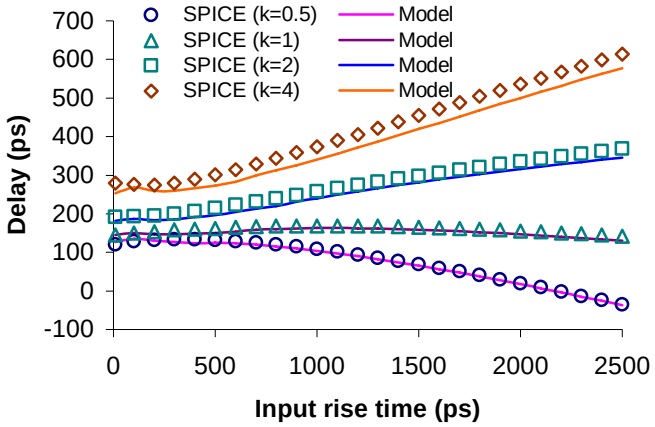

Fig.5. 25 Delay comparison of a two-input NAND with rising input and constant inverter load, $A = 1$, $B = 0 \rightarrow 1$

The output fall time versus the input rise time of the same transition is shown in Fig. 5.26. The input rise time varies from $10\,ps$ to $1200\,ps$. The average relative error of the piecewise linear model is -11.17% with respect to SPICE simulation. The maximum relative error is -17.2%. The average absolute error of the piecewise linear model is -37.12*ps* with respect to SPICE simulation. The maximum absolute error is -68.8*ps*.

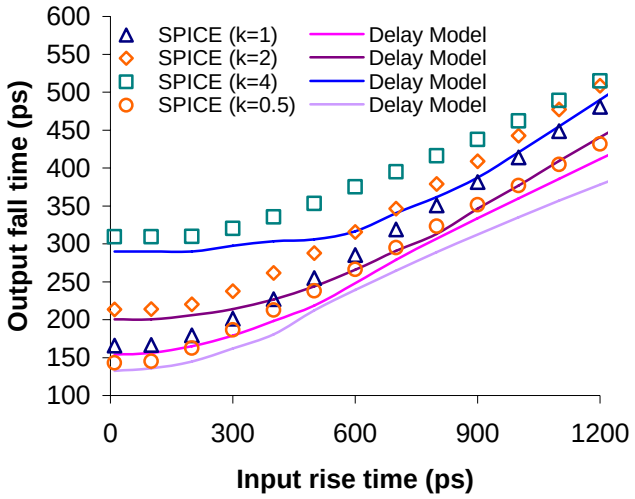

Fig.5. 26 Output fall time comparison of a two-input NAND with rising input and constant inverter load, $A=1$, $B=0\rightarrow 1$.

Fig. 5.27 plots the voltage output transient response of the output node ($V1$) and the internal node ($V2$) of an input transition time of $t_{Tin}=100\,ps$. $k=4$.

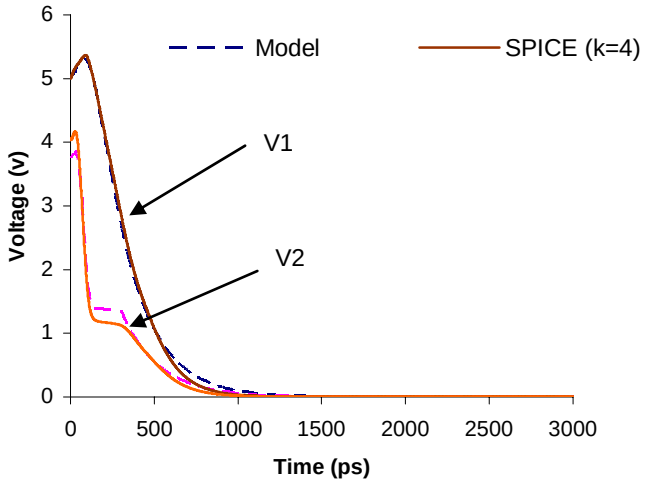

Fig.5. 27 Output waveforms comparison of a two-input NAND with rising input and constant inverter load. $t_{Tin}=100\,ps$. $A=1$, $B=0\rightarrow 1$.

5.1.2.4 $A=1$, $B=1\rightarrow 0$

The input node *A* of the upper NMOS transistor stays at $V_A = V_{DD}$. The node B of the lower NMOS transistor switches from high to low, a falling input transition.

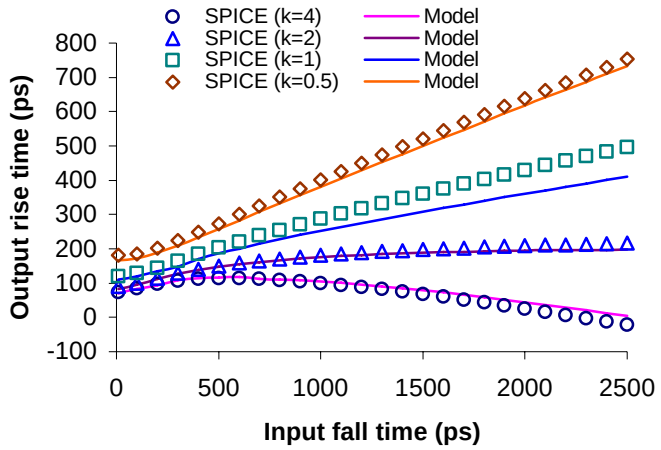

Fig.5. 28 Delay comparison of a two-input NAND with falling input and constant inverter load, $A = 1$, $B = 1 \rightarrow 0$.

Fig. 5.28 shows the delay versus the input fall time of the transition described above. The input fall time varies from $10\,ps$ to $2500\,ps$. The widths of the PMOS transistors are chosen as $4.8um$ and the widths of the NMOS transistors vary from $W_n = 1.2um$, $2.4um$, $4.8um$, and $9.6um$. Hence, the width ratio of PMOS transistors to NMOS transistors varies from $k = 4$, 2, 1, to 0.5. The average relative error of the piecewise linear model is -8.93% with respect to SPICE simulation. The maximum relative error is -790.3%. The average absolute error of the piecewise linear model is -15.46*ps* with respect to SPICE simulation. The maximum absolute error is 26.08*ps*.

The output rise time versus the input fall time of the same transition is shown in Fig. 5.29. The input fall time varies from $10\,ps$ to $1200\,ps$. The average relative error of the piecewise linear model is -9.38% with respect to SPICE simulation. The maximum

relative error is -27.7%. The average absolute error of the piecewise linear model is -31.81*ps* with respect to SPICE simulation. The maximum absolute error is -143.8*ps*.

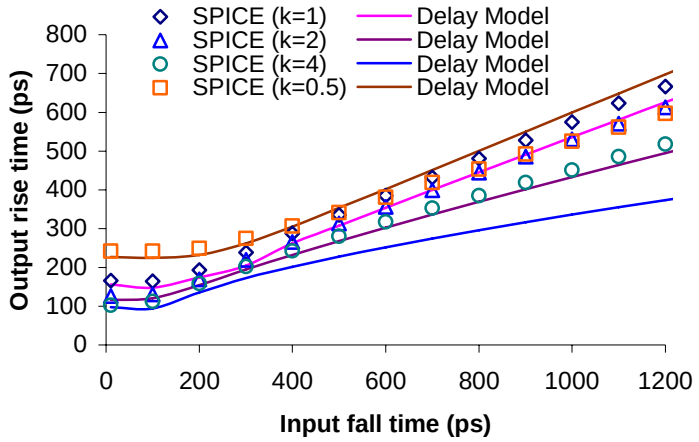

Fig.5. 29 Output rise time comparison of a two-input NAND with falling input and constant inverter load, $A=1$, $B=1\rightarrow 0$.

Fig. 5.30 plots the voltage output transient response of the output node (*V1*) and the internal node (*V2*) of an input transition time of $t_{Tin}=100\,ps$. $k=0.5$.

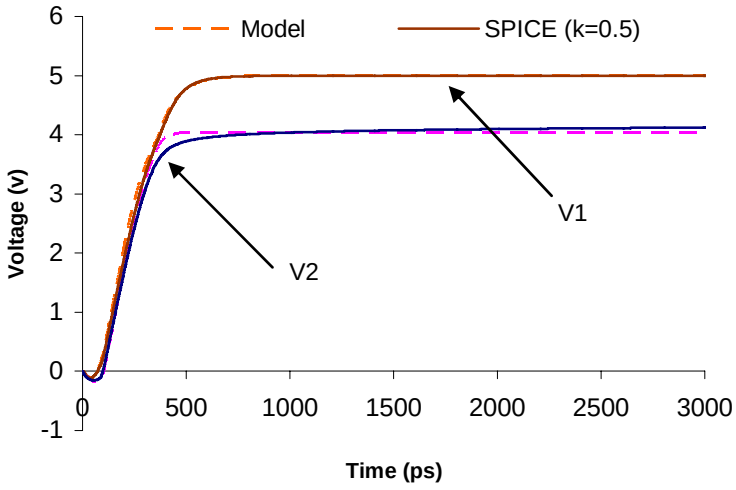

Fig.5. 30 Output waveforms comparison of a two-input NAND with falling input and constant inverter load. $t_{Tin}=100\,ps$. $A=1$, $B=1\rightarrow 0$.

5.2 OAI Gate Analysis

5.2.1 OAI Gate with Constant Capacitance Load

Fig. 5.31 shows a CMOS implementation of OAI gate with input signals *A*, *B*, and *C*. The gate drives a constant capacitance load, $C_{load} = 100\,fF$. We will assume that only one input signal is switching at any time. Multiple inputs can be handled by making V_{in} a column matrix. The pull-up transistors are of the same size and so are the pull-down transistors. The width ratio of PMOS transistors to NMOS transistors varies from $k = 0.5$, 1, 2 to 4.

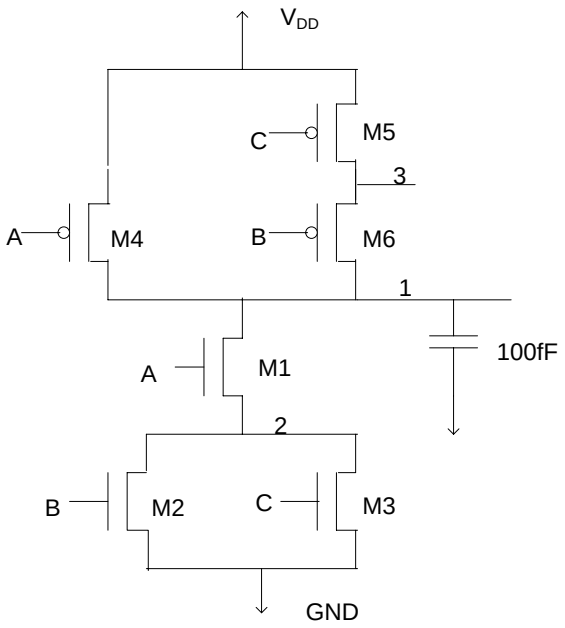

Fig.5. 31 OAI gate with constant capacitance load

5.2.1.1 $A = 0 \rightarrow 1,\ B = 1$, and $C = 0$

The input node *A* of the transistor M1 switches from low to high, a rising transition. M1 switches on and M4 switches off. The node B stays at $V_B = V_{DD}$ so the

NMOS transistor M2 is on and the PMOS transistor M6 if off. The node C connects to the ground so the NMOS transistor M3 is off and PMOS transistor M5 is on.

Fig. 5.32 shows the delay versus the input rise time of the transition described above. The input rise time varies from $10\,ps$ to $2500\,ps$. The widths of the NMOS transistors are chosen as $2.4um$ and the widths of the NMOS transistors vary from $W_p = 1.2um$, $2.4um$, $4.8um$, and $9.6um$. Hence, the width ratio of PMOS transistors to NMOS transistors varies from $k = 0.5$, 1, 2 to 4. The average relative error of the piecewise linear model is -2.03% with respect to SPICE simulation. The maximum relative error is -8.4%.

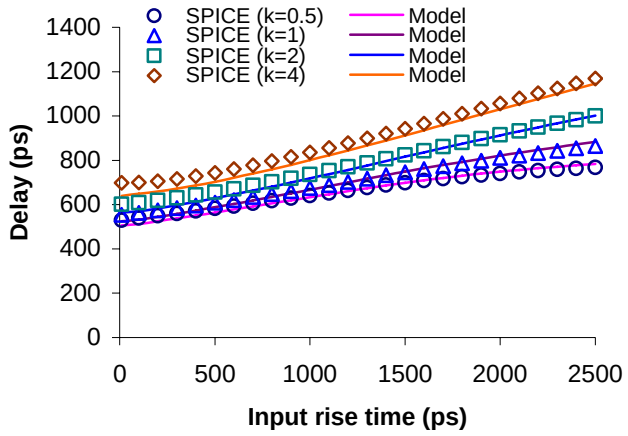

Fig.5. 32 Delay comparison of an OAI gate with rising input and constant capacitance load, $A = 0 \rightarrow 1$, $B = 1$, and $C = 0$.

The output fall time versus the input rise time of the same transition is shown in Fig. 5.33. The input rise time varies from $10\,ps$ to $1200\,ps$. The average relative error of the piecewise linear model is -7.27% with respect to SPICE simulation. The maximum relative error is -7.68%.

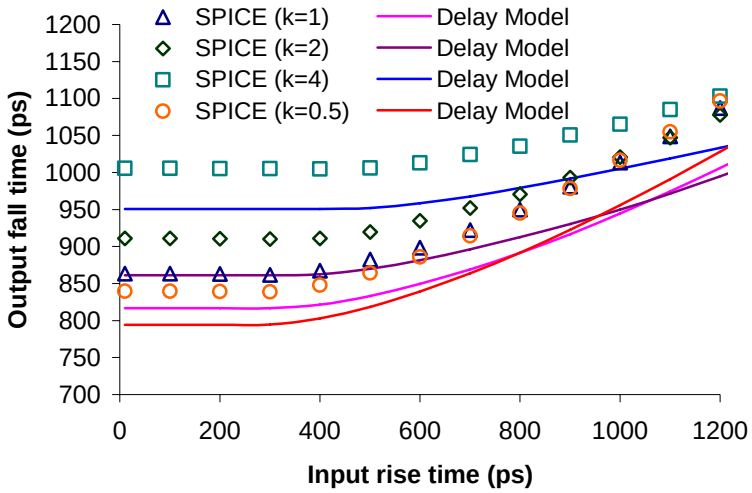

Fig.5. 33 Output fall time comparison of an OAI gate with rising input and constant capacitance load, $A = 0 \rightarrow 1$, $B = 1$, and $C = 0$.

Fig. 5.34 plots the voltage output transient response of the output node ($V1$) and the internal node ($V2$) of an input transition time of $t_{Tin} = 100\,ps$. $k = 4$.

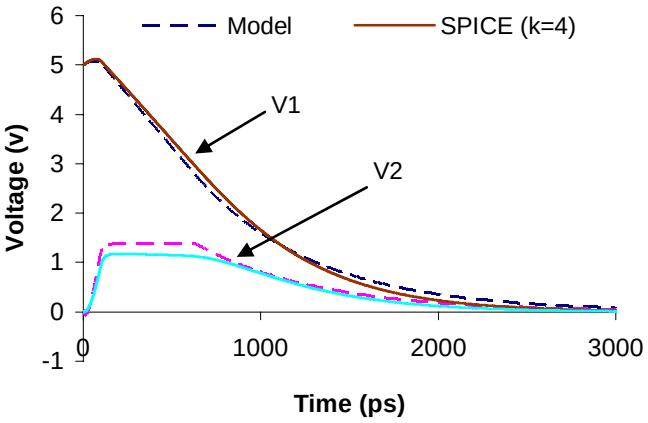

Fig.5. 34 Output waveforms comparison of an OAI gate with rising input and constant capacitance load. $t_{Tin} = 100\,ps$. $A = 0 \rightarrow 1$, $B = 1$, and $C = 0$.

5.2.1.2 $A = 1 \rightarrow 0$, $B = 1$, and $C = 0$

The input node A of the transistor M1 switches from high to low, a falling transition. M1 switches off and M4 switches on. The node B stays at $V_B = V_{DD}$ so the NMOS transistor M2 is on and the PMOS transistor M6 if off. The node C connects to the ground so the NMOS transistor M3 is off and PMOS transistor M5 is on.

Fig. 5.35 shows the delay versus the input fall time of the transition described above. The input fall time varies from $10\,ps$ to $2500\,ps$. The widths of the PMOS transistors are chosen as $4.8um$ and the widths of the NMOS transistors vary from $W_n = 1.2um$, $2.4um$, $4.8um$, and $9.6um$. Hence, the width ratio of PMOS transistors to NMOS transistors varies from $k = 4$, 2, 1 to 0.5. The average relative error of the piecewise linear model is -0.63% with respect to SPICE simulation. The maximum relative error is 10.39%.

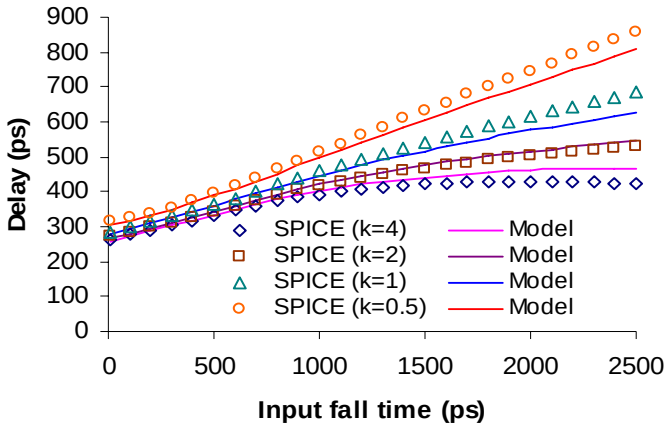

Fig.5. 35 Delay comparison of an OAI gate with falling input and constant capacitance load, $A = 1 \to 0$, $B = 1$, and $C = 0$.

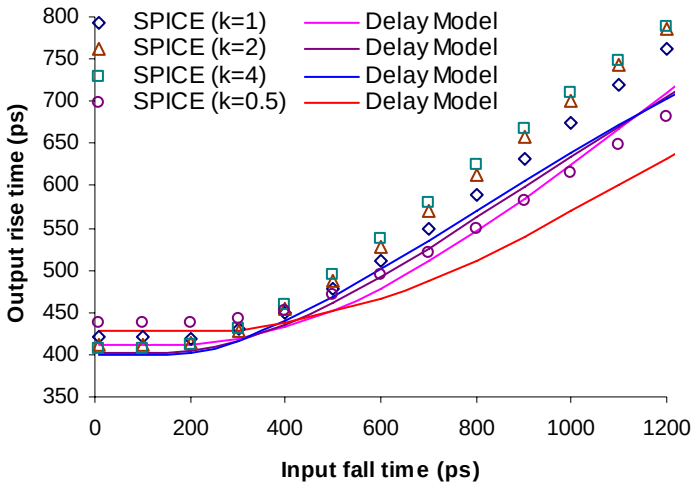

Fig.5. 36 Output rise time comparison an OAI gate with falling input and constant capacitance load, $A=1\rightarrow 0$, $B=1$, and $C=0$.

The output rise time versus the input fall time of the same transition is shown in Fig. 5.36. The input fall time varies from $10\,ps$ to $1200\,ps$. The average relative error of the piecewise linear model is -5.71% with respect to SPICE simulation. The maximum relative error is -10.7%.

Fig. 5.37 plots the voltage output transient response of the output node ($V1$) and the internal node ($V2$) of an input transition time of $t_{Tin}=100\,ps$. $k=4$.

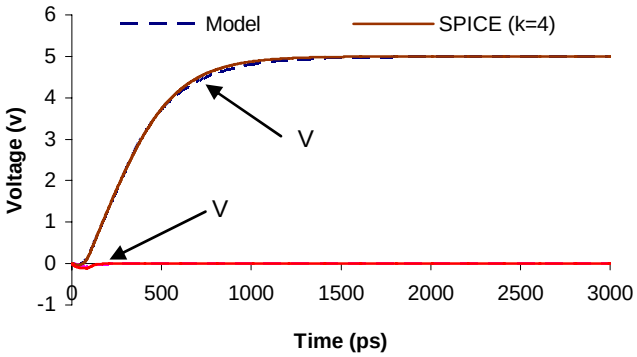

Fig.5. 37 Output waveforms comparison of an OAI gate with falling input and constant capacitance load. $t_{Tin}=100\,ps$. $A=1\rightarrow 0$, $B=1$, and $C=0$.

5.2.1.3 $A=1$, $B=0$, and $C=0\rightarrow 1$

The input node A of the transistor M1 stays at $V_A = V_{DD}$ so the NMOS transistor M1 is on and the PMOS transistor M4 is off. The node B connects to the ground so the NMOS transistor M2 is off and PMOS transistor M6 is on. The node C switches from low to high, a rising transition. M3 switches on and M5 switches off.

Fig. 5.38 shows the delay versus the input rise time of the transition described above. The input rise time varies from $10\,ps$ to $2500\,ps$. The widths of the NMOS transistors are chosen as $2.4um$ and the widths of the PMOS transistors vary from $W_p = 1.2um$, $2.4um$, $4.8um$, and $9.6um$. Hence, the width ratio of PMOS transistors to NMOS transistors varies from $k=0.5$, 1, 2 to 4. The average relative error of the piecewise linear model is -8.0% with respect to SPICE simulation. The maximum relative error is -13.4%.

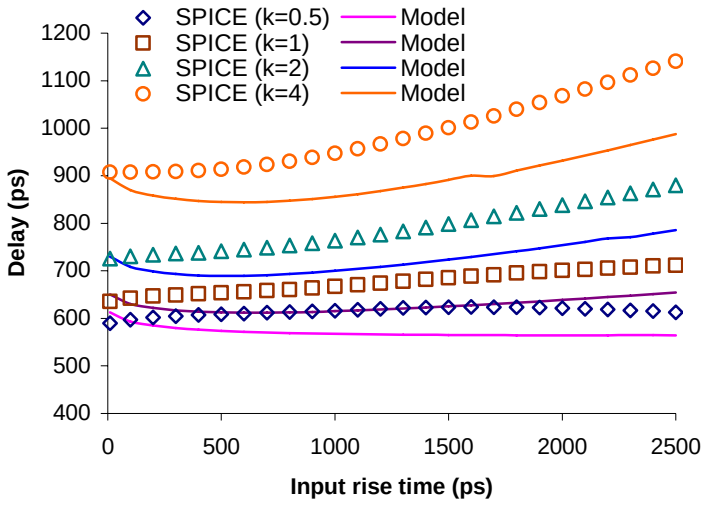

Fig.5. 38 Delay comparison of an OAI gate with rising input and constant capacitance load, $A=1$, $B=0$, and $C=0\rightarrow 1$.

The output fall time versus the input rise time of the same transition is shown in Fig. 5.39. The input rise time varies from $10\,ps$ to $1200\,ps$. The average relative error of the piecewise linear model is -6.92% with respect to SPICE simulation. The maximum relative error is -9.27%.

Fig. 5.40 plots the voltage output transient response of the output node $V1$ and the internal nodes $V2$ and $V3$ of an input transition time of $t_{Tin} = 1000\,ps$. $k = 4$.

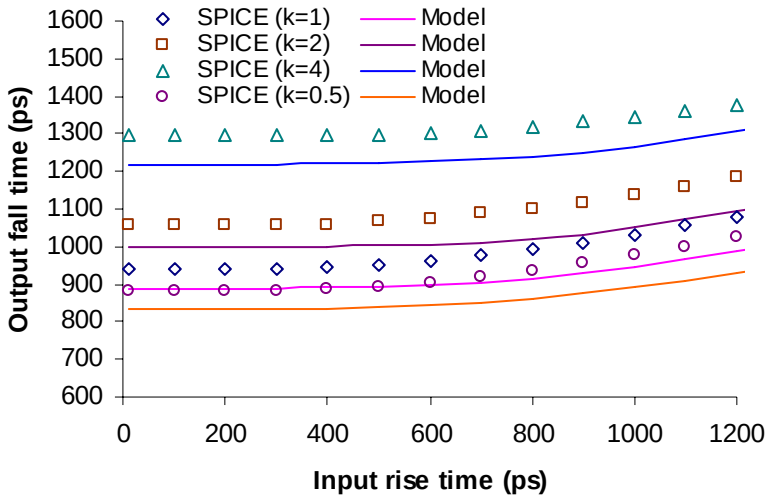

Fig.5. 39 Output fall time comparison of an OAI gate with rising input and constant capacitance load, $A = 1$, $B = 0$, and $C = 0 \rightarrow 1$.

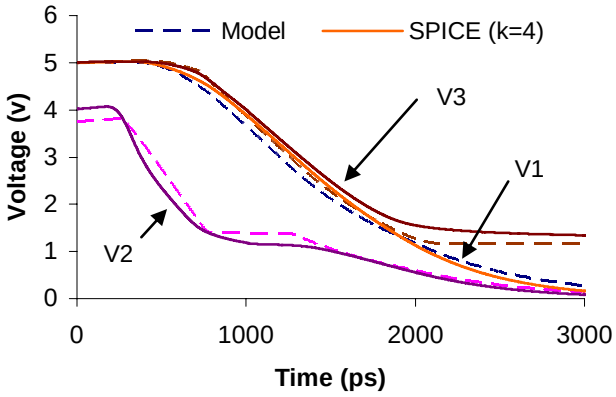

Fig.5. 40 Output waveforms comparison of an OAI gate with rising input and constant capacitance load. $t_{Tin} = 1000\,ps$. $A = 1$, $B = 0$, and $C = 0 \rightarrow 1$.

5.2.1.4 $A=1$, $B=0$, and $C=1\rightarrow 0$

The input node A of the transistor M1 stays at $V_A = V_{DD}$ so the NMOS transistor M1 is on and the PMOS transistor M4 is off. The node B connects to the ground so the NMOS transistor M2 is off and PMOS transistor M6 is on. The node C switches from high to low, a falling transition. M3 switches off and M5 switches on.

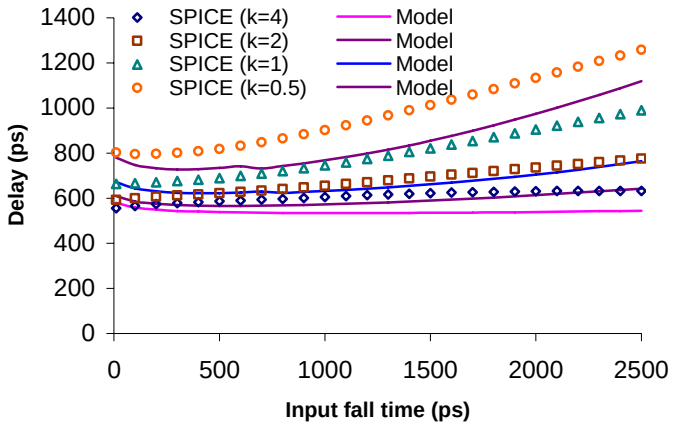

Fig.5. 41 Delay comparison of an OAI gate with falling input and constant capacitance load, $A=1$, $B=0$, and $C=1\rightarrow 0$.

Fig. 5.41 shows the delay versus the input fall time of the transition described above. The input fall time varies from $10\,ps$ to $2500\,ps$. The widths of the PMOS transistors are chosen as $4.8um$ and the widths of the NMOS transistors vary from $W_n = 1.2um$, $2.4um$, $4.8um$, and $9.6um$. Hence, the width ratio of PMOS transistors to NMOS transistors varies from $k=4$, 2, 1 to 0.5. The average relative error of the piecewise linear model is -12.9% with respect to SPICE simulation. The maximum relative error is -22.8%.

The output rise time versus the input fall time of the same transition is shown in Fig. 5.42. The input fall time varies from $10\,ps$ to $1200\,ps$. The average relative error of the piecewise linear model is -8.54% with respect to SPICE simulation. The maximum relative error is -15.45%.

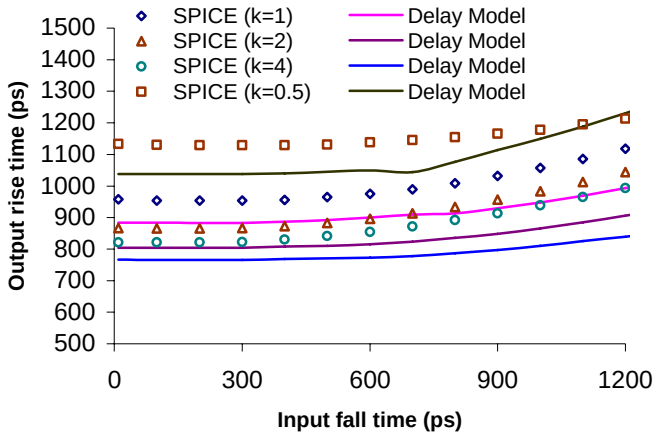

Fig.5. 42 Output rise time comparison of an OAI gate with falling input and constant capacitance load, $A=1$, $B=0$, and $C=1\rightarrow 0$.

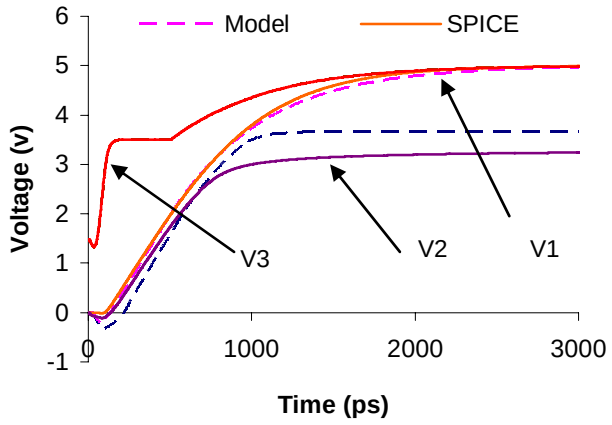

Fig.5. 43 Output waveforms comparison of an OAI gate with falling input and constant capacitance load. $t_{Tin}=100\,ps$. $A=1$, $B=0$, and $C=1\rightarrow 0$.

Fig. 5.43 plots the voltage output transient response of the output node *V*1 and the internal nodes *V*2 and *V*3 of an input transition time of $t_{Tin} = 100\,ps$. $k = 4$.

5.2.2 Special Issue on the Selection of the Substrate Cutoff Boundary

As we have noticed, when modeling the OAI gate with bottom NMOS switching, during the fast input transition region, the delay curve predicted by the piecewise linear model shows concave up behavior. Although this behavior also shows in the SPICE simulation, the model seems to overestimate this effect. This can be shown more clearly in the following circuit:

Fig. 5.44 shows a CMOS implementation of OAI gate with input signals *A*, *B*, and *C*. The gate drives a constant inverter load, $W_p / W_n = 4.8um / 2.4um$. We will assume that only one input signal is switching at any time. Multiple inputs can be handled by making V_{in} a column matrix. The pull-up transistors are of the same size and so are the pull-down transistors. The width ratio of PMOS transistors to NMOS transistors varies from $k = 0.5$, 1, 2 to 4.

The input transition is $A = 1$, $B = 0$, and $C = 1 \rightarrow 0$. The input node *A* of the transistor M1 stays at $V_A = V_{DD}$ so the NMOS transistor M1 is on and the PMOS transistor M4 is off. The node B connects to the ground so the NMOS transistor M2 is off and PMOS transistor M6 is on. The node C switches from high to low, a falling transition. M3 switches off and M5 switches on. The delay curve is shown in Fig. 5.45.

The concave up behavior is very obvious in Fig. 5.45. Further investigation found that the one of the model parameters δ, the value of the substrate cutoff boundary is somehow related to this behavior or at least has the effect on the very fast input

transition. Fig. 5.45 is the plot with $\delta = \infty$, i.e., no cutoff boundary is set. The output waveforms look like Fig. 5.46.

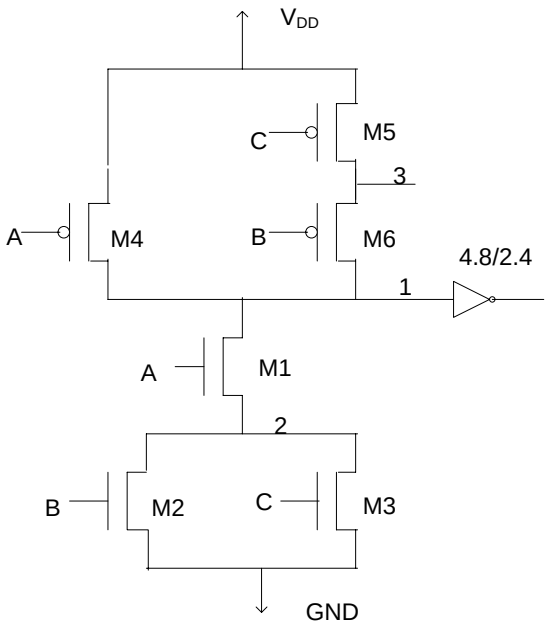

Fig.5. 44 An OAI gate with constant inverter load

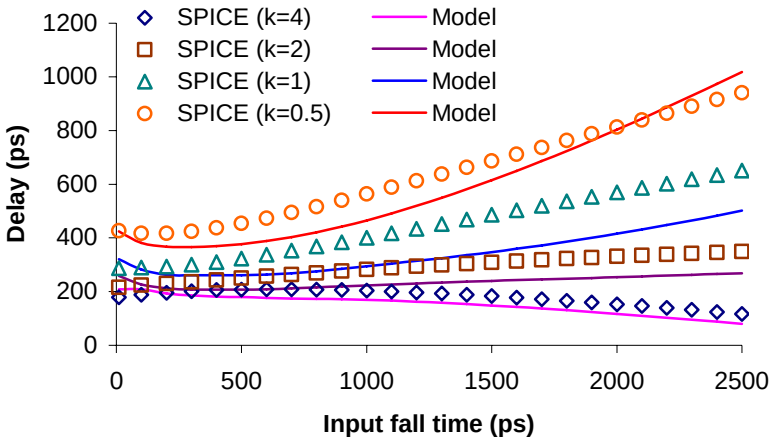

Fig.5. 45 Delay comparison of an OAI gate with falling input and constant inverter load, $A = 1$, $B = 0$, and $C = 1 \rightarrow 0$. No cutoff boundary.

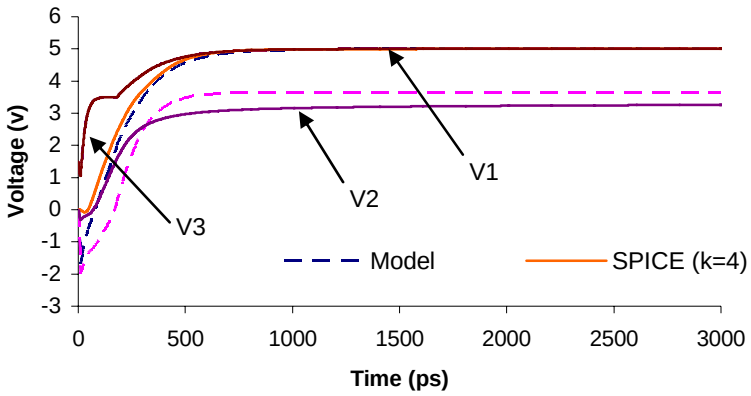

Fig.5. 46 Output waveforms comparison of an OAI gate with falling input and constant capacitance load. $t_{Tin} = 10ps$. $A = 1$, $B = 0$, and $C = 1 \rightarrow 0$. No cutoff boundary.

It can be seen that the node voltages of node 1 and node 2 drop as low as about -2 *volts* during the transition which are unrealistic in real circuits. Normally due to the substrate diode, the overshoot cannot be over 0.6 *volts*. In the following case, we set the cutoff boundary $\delta = 0.6v$, the delay curve now looks like:

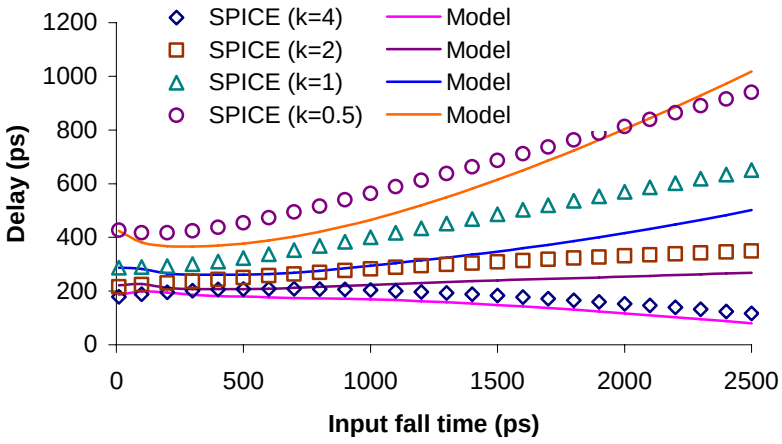

Fig.5. 47 Delay comparison of an OAI gate with falling input and constant inverter load, $A = 1$, $B = 0$, and $C = 1 \rightarrow 0$. $\delta = 0.6$.

And the output waveforms are shown in Fig. 5.48.

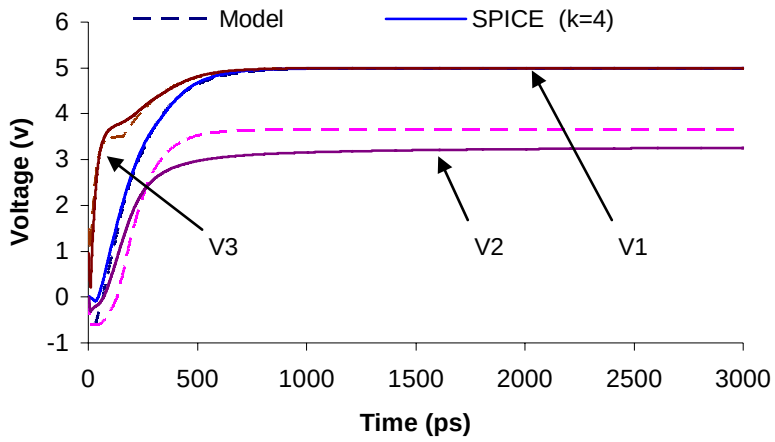

Fig.5. 48 Output waveforms comparison of an OAI gate with falling input and constant capacitance load. $t_{Tin} = 10ps$. $A = 1$, $B = 0$, and $C = 1 \to 0$. $\delta = 0.6v$.

In Fig. 5.47, after adding the cutoff boundary, the model fits better in the very fast input region. The voltage overshoot in Fig. 5.48 is only as small as -0.6 *volt* which is closer to the SPICE simulation results. Although the concave up behavior still exists in Fig. 5.47, further improvement can be made by: 1. further decrease the cutoff boundary and make it comparable to SPICE. 2. add more piecewise linear regions .

The average relative error of the piecewise linear model is -16.78% with respect to SPICE simulation. The maximum relative error is -31.09%. The average absolute error of the piecewise linear model is -59.56*ps* with respect to SPICE simulation. The maximum absolute error is -154.02*ps*.

Observing the SPICE simulation, we found that the maximum overshoot of the SPICE simulation is around $-0.3v$. In order to test the sensitivity of the model parameter δ, we further reduced the cutoff boundary to $\delta = 0.3v$, the delay curve looks like

And the output waveforms are shown in Fig. 5.50.

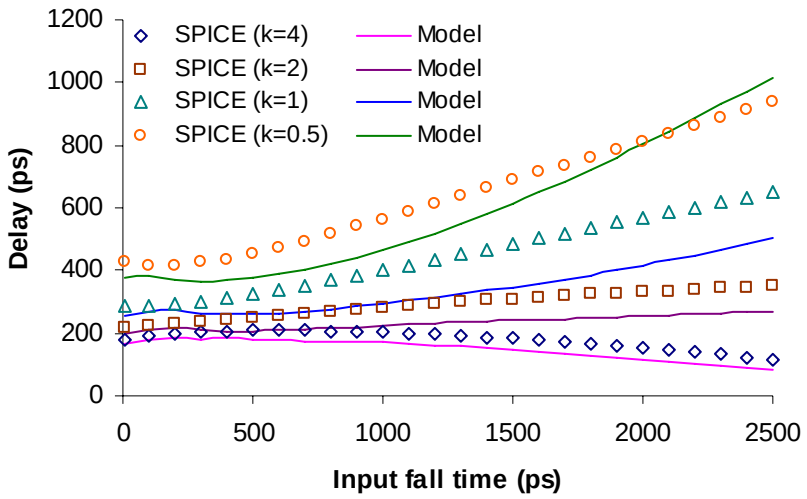

Fig.5. 49 Delay comparison of an OAI gate with falling input and constant inverter load, $A=1$, $B=0$, and $C=1\rightarrow 0$. $\delta=0.3$.

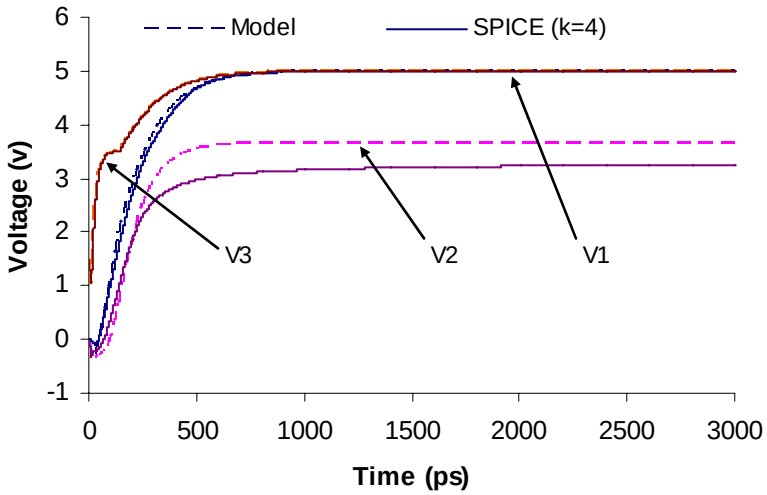

Fig.5. 50 Output waveforms comparison of an OAI gate with falling input and constant capacitance load. $t_{Tin}=10ps$. $A=1$, $B=0$, and $C=1\rightarrow 0$. $\delta=0.3v$.

As expected, further decreasing the value of δ, the predicted delay value for the very fast input transitions decreases. The delay curves in Fig. 5.49 even show the property of concave down in the fast transition regions. The minimal output voltage values shown in Fig. 5.50 are limited to $-0.3v$ due to the cutoff boundary.

Pushing the cutoff boundary further to the extremity at $\delta = 0v$, i.e., no output curves allow rise or fall beyond V_{DD} and GND, this behavior of delay dropping in the fast input transition regions is getting more obvious.

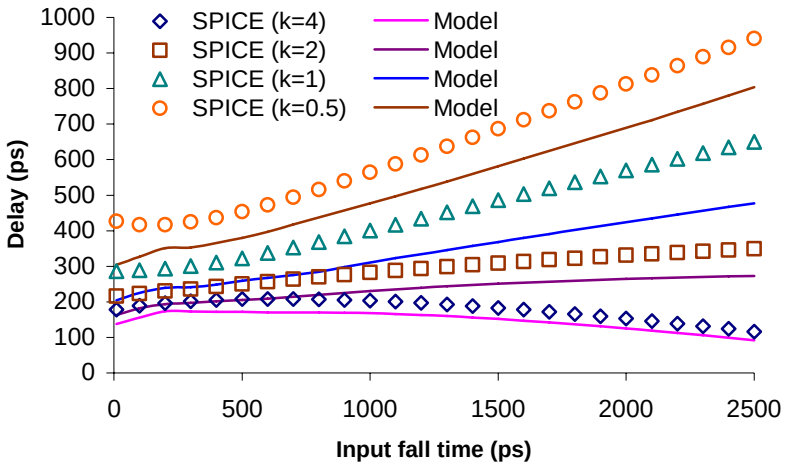

Fig.5. 51 Delay comparison of an OAI gate with falling input and constant inverter load, $A = 1$, $B = 0$, and $C = 1 \rightarrow 0$. $\delta = 0$.

Fig. 5.51 shows that the concave down effect is getting worse if we set the cutoff boundary the same as the ground line. The model prediction has an obvious lead time with regard to the SPICE simulations for a 10*ps* input fall time in Fig. 5.52.

So the piecewise linear model is sensitive to the choice of cutoff boundary. The suitable value of cutoff boundary δ is the one that best fits the SPICE simulations.

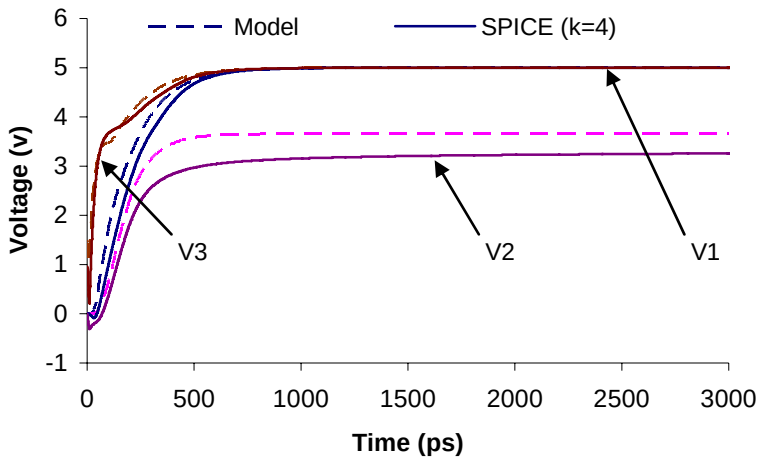

Fig.5. 52 Output waveforms comparison of an OAI gate with falling input and constant capacitance load. $t_{Tin} = 10ps$. $A = 1$, $B = 0$, and $C = 1 \rightarrow 0$. $\delta = 0$.

5.3 Conclusion

In this chapter, extensive tests have been made on a two-input NAND gate and an OAI gate. We compare the piecewise linear model with respect to SPICE simulation by considering the different input transition slopes, direction of transitions (rising input or falling input), pull-up and pull-down transistor ratios, and loading effects. The piecewise linear model shows a good agreement with SPICE simulation in general. The piecewise linear model can also predict the output waveforms of different nodes in the circuit and shows good fit to the SPICE output waveform over a very wide of input transition range. Thus, it is possible to utilize the piecewise linear model to model the circuit power consumptions.

We also discussed the concave up problems shown in the delay characteristic curves and proposed some possible solutions.

CHAPTER VI

CONCLUSION

A VLSI circuit which contains millions of transistors can be divided into smaller resistance connected regions. The delay through the circuit is just the sum of the delays in each resistance connected region. The delay model used in a fast simulator must satisfy two criteria: simple and accurate.

This dissertation presented a piecewise linear delay model which includes the signal input slope in the delay computation. Circuit enters into different piecewise linear regions when the transistors turning on and off or going in and out of the saturation region. Circuit dynamic equations are solved in each piecewise linear region by using steady state solution. The piecewise linear delay model is a single time constant model. In each piecewise linear region, the time constant is a modified version of Elmore delay.

Since the output waveform is known in each piecewise linear region, the circuit delay and the output slope can be easily obtained from the output waveform. The delay computed in the piecewise linear delay model is the mid-point delay which is decided by the time when the output voltage waveform hits the half-V_{DD} point. Output slope is determined by the first half of the output transition waveform.

Extensive comparisons have been made between the piecewise linear model and the SPICE simulations for simple gates, such as an inverter, a two-input NAND, and an

OAI gate. Excellent accuracies have been observed. The piecewise linear model is capable of predicting the output waveforms and the propagation delay over a variety of input and output conditions.

The piecewise linear model can handle large complicated circuits by introducing smaller resistance connected regions. The piecewise linear model is also scalable from one technology to another by choosing different set of technology related model parameters.

The model has the simulation speed advantages over SPICE since usually each resistance connected region has only a limited number of transistors and nodes. Hence the matrix operations based on the smaller resistance connected region are faster than SPICE. The simulation speed of the piecewise linear model will be comparable to switch models like the one used in IRSIM. However, the piecewise linear delay model will be slower than IRSIM since IRSIM has a much simpler timing model.

In general, it is possible to implement a fast circuit simulator based on the piecewise linear model or include the piecewise linear model into an existing circuit simulator such as IRSIM because of its accuracy and speed advantages.

This dissertation is focus on the accuracy of the delay model. A comprehensive test has been made on the simple gates to verify the circuit accuracy. Unfortunately, we have not tested the simulation speed of the piecewise linear model yet. So one of the important research works need to be done next is to compare the simulation speed of the model with respect to the SPICE simulation over some large circuits or test benches.

The piecewise linear model is applicable to any general circuit topology. More works need to be done for the delay of complicated circuits or more general cases.

Another thought is that we claim that the piecewise linear is scalable by picking different model parameters for different technology. However, all the researches done so far are for AMI 0.6u technology, no work has been conducted to verify the scalability of the model.

BIBLIOGRAPHY

[1] W. C. Elmore, "The Transient Response of Damped Linear Networks with Particular Regard to Wideband Amplifers," *Journal of Applied Physics*, vol. 19, no. 1, pp. 55-63, January 1948.

[2] J. Rubinstein, P. Penfield Jr., and M. A. Horowitz, "Signal Delay in RC Tree Networks", *IEEE Trans. Computer-Aided Design*, vol. CAD-2, no. 3, pp. 202-211, July 1983.

[3] J. R. Burns, "Switching Response of Complementary Symmetry MOS Transistor Logic Circuits," *RCA Review*, vol. 25, pp. 627-661, 1964.

[4] C. J. Terman, "Simulation Tools for Digital LSI Design", PhD dissertation, Massachusetts Institute of Technology, September, 1983.

[5] J. K. Ousterhout, "A Switch-Level Timing Verifier for Digital MOS VLSI," *IEEE Trans. Computer-Aided Design*, vol. CAD-4, no. 3, pp. 337-349, July 1985.

[6] P. Penfield and J. Rubinstein, "Signal Delay in RC Tree Networks", *Proc. Eighteenth Design Automation Conference*, pp. 613-617, June 1981.

[7] Chorng-Yeong Chu, "Improved Models for Switch-Level Simulation," PhD dissertation, Stanford University, 1989.

[8] Chorng-Yeong Chu and M. A. Horowitz, "Charge-Sharing Models for Switch-Level Simulation," *IEEE Trans. Computer-Aided Design*, vol. CAD-6, no. 6, pp. 1053-1061, November 1987.

[9] A. Salz and M. A. Horowitz, “IRSIM: An Incremental MOS Switch-Level Simulator,” *Proc. 26th Design Automation Conference*, pp. 173-178, June 1989.

[10] M. A. Horowitz, “Timing Models for MOS Circuits,” PhD dissertation, Stanford University, 1984.

[11] N. Hedenstierna and K. O. Jeppson, “CMOS Circuit Speed and Buffer Optimization,” *IEEE Trans. Computer-Aided Design*, vol. CAD-6, pp. 270-281, March 1987.

[12] M. I. Elmasry, “Digital MOS Integrated Circuits: A Tutorial,” *Digital MOS Integrated Circuits*. New York: IEEE press, 1981, pp. 4-27.

[13] K. O. Jeppson, “Modeling the Influence of the Transistor Gain Ratio and the Input-to-Output Coupling Capacitance on the CMOS Inverter Delay,” *IEEE Journal of Solid-State Circuits*, vol. 29, no. 6, June 1994.

[14] W. Shockley, “A Unipolar Field Effect Transistor,” *Proc. IRE*, vol. 40, pp. 1365-1376, November 1952.

[15] T. Sakurai and A. R. Newton, “Alpha-Power Law MOSFET Model and Its Applications to CMOS Inverter Delay and Other Formulas,” *IEEE Journal of Solid-State Circuits*, vol. 25, no. 2, April 1990.

[16] S. Dutta, S. S. M. Shetti, and S. L. Luskey, “A Comprehensive Delay Model for CMOS Inverters,” *IEEE Journal of Solid-State Circuits*, vol. 30, no. 8, August 1995.

[17] T. Sakurai and A. R. Newton, “Delay Analysis of Series-Connected MOSFET Circuits”, *IEEE Journal of Solid-State Circuits*, vol. 26, no. 2, February 1991.

[18] D. J. Pilling and J. G. Skalnik, “A Circuit Model for Predicting Transient Delays in LSI Logic Systems,” Proc. 6th Asilomar Conference on Circuits and Systems, pp. 424-428, 1972.

[19] J. K. Ousterhout, "Switch-Level Delay Models for Digital MOS VLSI," *Proc. 21st IEEE/ ACM Design Automation Conference*, Alberquerque, N. M., pp. 542-548, June 1984.

[20] J. K. Ousterhout, "Crystal: A Timing Analyzer for nMOS VLSI Circuits," *Proc. Third Caltech Conference on VLSI*, R. Bryant, ed., Computer Science Press, pp. 57-70, 1983.

[21] P. Cocchini, G. Piccinini, and M. Zamboni, "A Comprehensive submicrometer MOST delay model and its application to CMOS buffers", *IEEE Journal of Solid-State Circuits*, vol. 32, no. 8, February 1997.

[22] L. Bisdounis, S. Nikolaidis, and O. Koufopavlou, "Analytical transient response and propagation delay evaluation of the CMOS inverter for short-channel devices", *IEEE Journal of Solid-State Circuits*, vol. 33, no. 2, February 1998.

[23] R. A. Rohrer, "Circuit Simulation - The Early Years," *IEEE Circuits and Devices Magzine*, vol. 8, issue 3, pp. 32-37, May 1992.

[24] Y. Cheng and C. Hu, *MOSFET Modeling & BSIM3 User's Guide*, Reading, Kluwer Academic Publishers, pp. 17-19, 1999.

[25] L. G. Johnson, "Delay of CMOS Logic Gates," Research discussions (unpublished).

[26] L. G. Johnson, "Advanced Digital VLSI Design," Lecture notes for ECEN 6263, 2001.

http://lgjohn.ecen.ceat.okstate.edu/6263/index.html

APPENDICES

APPENDIX A

PIECEWISE LINEAR MODEL PARAMETERS

A.1 Current Model Related Parameters

W_p (um)	R_{sp} ($\Omega/\square$)	V_{Tp} (v)	a_p
1.2	2.1×10^4	-1.8	1.251763
2.4	2.1×10^4	-1.6	1.396283
4.8	2.1×10^4	-1.49	1.348792
9.6	2.1×10^4	-1.386	1.352994

TABLE A. 1 Falling Input PMOS Parameters

W_p (um)	R_{sp} ($\Omega/\square$)	V_{Tp} (v)	a_p
1.2	3.3×10^4	-1.292	2.9
2.4	3.3×10^4	-1.241	2.8
4.8	3.3×10^4	-1.243	2
9.6	3.3×10^4	-1.161	1.2

TABLE A. 2 Rising Input PMOS Parameters

W_n (um)	R_{sn} ($\Omega/\square$)	V_{Tn} (v)	a_n
1.2	1.7×10^4	1.34	1.38442
2.4	1.7×10^4	1.1	1.64235
4.8	1.7×10^4	0.99	1.03714
9.6	1.7×10^4	0.96	1.43759

TABLE A. 3 Falling Input NMOS Parameters

W_n (um)	R_{sn} ($\Omega/\square$)	V_{Tn} (v)	a_n
1.2	1.3×10^4	1.6	1.747667
2.4	1.3×10^4	1.24	1.726266
4.8	1.3×10^4	1.2	1.741734
9.6	1.3×10^4	1.09	1.706108

TABLE A. 4 Rising Input NMOS Parameters

A.2 Other Model Related Parameters

$C_{ox} = 2.449\times10^{-3} F/m^2$ [gate oxide capacitance per unit area]

$x_{partn} = 0.3$ [NMOS channel charge partition parameter]

$x_{partp} = 0.3$ [PMOS channel charge partition parameter]

$\delta = 0.6$ [substrate cutoff boundary]

$V_{DD} = 5v$ [power supply voltage]

APPENDIX B

CHANNEL CHARGE STORAGE MODEL

The capacitive currents into the charge storage elements are:

$$i_G = \frac{dQ_G}{dt} = C_{GS}\frac{dV_{GS}}{dt} + C_{GD}\frac{dV_{GD}}{dt} + C_{GB}\frac{dV_{GB}}{dt} \tag{B.1}$$

$$i_S = \frac{dQ_S}{dt} = C_{SG}\frac{dV_{SG}}{dt} + C_{SD}\frac{dV_{SD}}{dt} + C_{SB}\frac{dV_{SB}}{dt} \tag{B.2}$$

$$i_D = \frac{dQ_D}{dt} = C_{DG}\frac{dV_{DG}}{dt} + C_{DS}\frac{dV_{DS}}{dt} + C_{DB}\frac{dV_{DB}}{dt} \tag{B.3}$$

$$i_B = -(i_G + i_S + i_D) \tag{B.4}$$

where C_{ij} are all independent non-linear functions of the terminal voltages and are defined by the following:

$$C_{ij} = \frac{\partial Q_i}{\partial V_{ij}} \quad i, j = G, D, S, B \tag{B.5}$$

It is most convenient to use these equations rewritten in terms of the individual terminal voltages V_G, V_S , V_D , and V_B.

$$i_G = \frac{dQ_G}{dt} = C_{GG}\frac{dV_G}{dt} - C_{GS}\frac{dV_S}{dt} - C_{GD}\frac{dV_D}{dt} - C_{GB}\frac{dV_B}{dt} \tag{B.5}$$

$$i_S = \frac{dQ_S}{dt} = -C_{SG}\frac{dV_G}{dt} + C_{SS}\frac{dV_S}{dt} - C_{SD}\frac{dV_D}{dt} - C_{SB}\frac{dV_B}{dt} \tag{B.6}$$

$$i_D = \frac{dQ_D}{dt} = -C_{DG}\frac{dV_G}{dt} - C_{DS}\frac{dV_S}{dt} + C_{DD}\frac{dV_D}{dt} - C_{DB}\frac{dV_B}{dt} \tag{B.7}$$

$$i_B = -(i_G + i_S + i_D) \tag{B.8}$$

where

$$C_{GG} = C_{GS} + C_{GD} + C_{GB} \tag{B.9}$$

$$C_{SS} = C_{SG} + C_{SD} + C_{SB} \tag{B.10}$$

$$C_{DD} = C_{DG} + C_{DS} + C_{DB} \tag{B.11}$$

In digital applications, the substrate terminal is biased at a constant voltage so that the last column in Eq. (B.1) to (B.4) can be ignored.

The following is a piece-wise linear approximation of the BSIM model. Since it uses the same definitions of V_T and V_{DSsat}, the model has the same regions of validity as the piece-wise linear current model.

Ohmic region:

$$Q_{Gohm} = C_{ox}(V_{GS} - V_T) - (a_S + a_D)C_{ox}V_{DS} + b_{SB}C_{ox}(V_{SB} + V_T - V_{FB}) \tag{B.12}$$

$$-Q_{Sohm} = \frac{1}{2}C_{ox}(V_{GS} - V_T) - a_S C_{ox}V_{DS} \tag{B.13}$$

$$-Q_{Dohm} = \frac{1}{2}C_{ox}(V_{GS} - V_T) - a_D C_{ox}V_{DS} \tag{B.14}$$

Saturation region:

$$Q_{Gsat} = bC_{ox}(V_{GB} - V_{FB}) + b_{SB}C_{ox}(V_{SB} + V_T - V_{FB}) \tag{B.15}$$

$$-Q_{Ssat} = (1 - x_{part})b_{GS}C_{ox}(V_{GS} - V_T) \tag{B.16}$$

$$-Q_{Dsat} = x_{part}b_{GS}C_{ox}(V_{GS} - V_T) \tag{B.17}$$

Cutoff region:

$$Q_{Goff} = b_{SB} C_{ox} (V_{GB} - V_{FB}) \tag{B.18}$$

$$-Q_{Soff} = 0 \tag{B.19}$$

$$-Q_{Doff} = 0 \tag{B.20}$$

Using continuity of charge at the ohmic-saturation boundary and equating Eq. (B.13) and (B.16), we find:

$$a_S = a(\frac{1}{2} - (1 - x_{part}) b_{GS}) \tag{B.21}$$

Similarly, by equating Eq. (B.14) and (B.17), we find:

$$a_D = a(\frac{1}{2} - x_{part} b_{GS}) \tag{B.22}$$

Putting back the values for V_T, a_S, and a_D back into Eq. (B.12) and (B.15) gives:

$$Q_{Gohm} = C_{ox} (V_{GS} - V_T) - (1 - b) a C_{ox} V_{DS} + b C_{ox} (V_{SB} + V_T - V_{FB}) \tag{B.23}$$

$$Q_{Gsat} = b C_{ox} (V_{GB} - V_{FB}) \tag{B.24}$$

$$Q_{Goff} = b C_{ox} (V_{GB} - V_{FB}) \tag{B.25}$$

The BSIM model uses $b = \frac{2}{3}$ gives the best fit in the ohmic and saturation regions. These values leave Q_G, and therefore the gate current, dependent on V_{DS} and V_{SB} when the transistor is on. This dependence greatly complicates the evaluation of the model. Rather than choosing the values for b which best fit the BSIM model, we choose values that leave Q_G independent of V_{DS} and V_{SB}. Choosing $b = 1$, Eq. (B.12) to (B.20) can be written as:

$$Q_G = C_{ox} (V_{GB} - V_{FB}) \tag{B.26}$$

$$-Q_{Sohm} = \frac{1}{2}C_{ox}(V_{GS} - V_T) + a(b - \frac{1}{2} - x_{part}b)C_{ox}V_{DS} \tag{B.27}$$

$$-Q_{Dohm} = \frac{1}{2}C_{ox}(V_{GS} - V_T) - a(\frac{1}{2} - x_{part}b)C_{ox}V_{DS} \tag{B.28}$$

$$-Q_{Ssat} = (1 - x_{part})bC_{ox}(V_{GS} - V_T) \tag{B.29}$$

$$-Q_{Dsat} = x_{part}bC_{ox}(V_{GS} - V_T) \tag{B.30}$$

Taking the derivative of Eq.(B.26) to (B.30) with respect to t, the charge storage currents are:

$$i_{Goff} = bC_{ox}\frac{dV_G}{dt} \tag{B.31}$$

$$i_{Gohm} = C_{ox}\left[\frac{dV_G}{dt} + (1-b)(a-1)\frac{dV_S}{dt} - (1-b)a\frac{dV_D}{dt}\right] \tag{B.32}$$

$$i_{Sohm} = C_{ox}\left[-\frac{1}{2}\frac{dV_G}{dt} + [\frac{1}{2} + a(b - \frac{1}{2} - x_{part}b)]\frac{dV_S}{dt} - a(b - \frac{1}{2} - x_{part}b)\frac{dV_D}{dt}\right] \tag{B.33}$$

$$i_{Dohm} = C_{ox}\left[-\frac{1}{2}\frac{dV_G}{dt} + [\frac{1}{2} - a(\frac{1}{2} - x_{part}b)]\frac{dV_S}{dt} + a(\frac{1}{2} - x_{part}b)\frac{dV_D}{dt}\right] \tag{B.34}$$

$$i_{Gsat} = bC_{ox}\frac{dV_G}{dt} \tag{B.35}$$

$$i_{Ssat} = C_{ox}\left[-(1 - x_{part})b\frac{dV_G}{dt} + (1 - x_{part})b]\frac{dV_S}{dt}\right] \tag{B.36}$$

$$i_{Dsat} = C_{ox}\left[-x_{part}b\frac{dV_G}{dt} + x_{part}b\frac{dV_S}{dt}\right] \tag{B.37}$$

Comparing with the derivatives in (B.5) to (B.8) gives the linearized results for the transistor capacitances in TABLE 2.2.

APPENDIX C

STEADY STATE SOLUTION FOR SINGULAR CONDUCTANCE MATRIX

The conductance matrix for a resistance connected region will be singular whenever the region does not include the power or ground node or the source node of any transistor in saturation. It is easy to show that

$$\sum_j G_{ij} = \sum_i G_{ij} = 0 \tag{C.1}$$

for a singular conductance matrix. This kind of singular G matrix allows non-zero $\ddot{\tilde{V}}$ solutions have the form

$$[\ddot{\tilde{V}}] = \begin{bmatrix} 1 \\ \dots \\ 1 \end{bmatrix} \ddot{\tilde{V}} \tag{C.2}$$

$\ddot{\tilde{V}}$ is the same for each node i since

$$[G\ddot{\tilde{V}}]_i = \sum_j G_{ij} \begin{bmatrix} 1 \\ \dots \\ 1 \end{bmatrix} \ddot{\tilde{V}} = \ddot{\tilde{V}} \sum_j G_{ij} = 0 \tag{C.3}$$

The capacitance matrix is not singular since there will always be capacitive coupling to ground through the substrate. Taking the sum of all rows of the dynamic equation (3.15) gives

$$\sum_j (\sum_i C_{ij}) \frac{d\tilde{V}_j}{dt} + \sum_i [I_{in}(t_0) + \dot{I}_{in}(t - t_0)]_i = 0 \tag{C.4}$$

which represents charge conservation, the first term being the rate of change of the total charge stored on all of the internal capacitances in the resistance connected region and the second term being current leaving the resistance connected region. Equating terms of the same power of *t* gives

$$\sum_j (\sum_i C_{ij}) \ddot{\tilde{V}}_j = -\sum_i [\dot{I}_{in}]_i \tag{C.5}$$

$$\sum_j (\sum_i C_{ij}) \ddot{\tilde{V}}_j = -\sum_i [I_{in}(t_0)]_i \tag{C.6}$$

$\ddot{\tilde{V}}$ is independent of *j*, then

$$\ddot{\tilde{V}} = \frac{-\sum_i [\dot{I}_{in}]_i}{\sum_j (\sum_i C_{ij})} \tag{C.7}$$

The rest of the steady state solution can be found by picking any reference node, *r*, in the resistance connected region that has a non-zero capacitance $\sum_i C_{ir}$. The charge conservation equation can be used to define $\dot{\tilde{V}}_r$ in terms of the other node voltages. Using terms in (C.4) independent of *t* gives

$$\dot{\tilde{V}}_r = \frac{-\sum_{j \neq r} (\sum_i C_{ij}) \dot{\tilde{V}}_j - \sum_i [I_{in}(t_0)]_i}{(\sum_i C_{ir})} \tag{C.8}$$

The boundary condition of the initial charge

$$\sum_j (\sum_i C_{ij}) \tilde{V}_j(t_0) = \sum_j (\sum_i C_{ij}) V_j(t_0) \tag{C.9}$$

can be used to define $\tilde{V}_r(t_0)$ in terms of the other node voltages.

$$\tilde{V}_r(t_0) = V_r(t_0) + \frac{\sum_{j\neq r}(\sum_i C_{ij})(V_j(t_0) - \tilde{V}_j(t_0))}{(\sum_i C_{ir})} \tag{C.10}$$

Any other node, $i \neq r$, must still satisfy the constraints from the original dynamic equation.

$$G\dot{\tilde{V}} + G_r\dot{\tilde{V}}_r = -(C\ddot{\tilde{V}} + C_r\ddot{\tilde{V}}_r + \dot{I}_{in}) \tag{C.11}$$

$$G\tilde{V}(t_0) + G_r\tilde{V}_r(t_0) = -(C\dot{\tilde{V}} + C_r\dot{\tilde{V}}_r + I_{in}(t_0)) \tag{C.12}$$

where *G* and *C* matrices are now further reduced by removing the row and columns corresponding to node *r*. The reduced *G* matrix is now non-singular and its inverse can be used to find the steady state solution. G_r and C_r are the column matrices for column *r* in the original *G* and *C* matrices.

$$[G_r]_i = G_{ir} = -\sum_{j\neq r} G_{ij} \tag{C.13}$$

$$[C_r]_i = C_{ir} \tag{C.14}$$

Define

$$\Gamma_{ij} = G_{ij} - \frac{G_{ir}\sum_{i'} C_{i'j}}{\sum_{i'} C_{i'r}} \quad i \neq r, j \neq r \tag{C.15}$$

substituting for $\dot{\tilde{V}}_r$ gives the solution for $\dot{\tilde{V}}_{i\neq r}$

$$\dot{\tilde{V}}_{i\neq r} = -\frac{\sum_{i'}[I_{in}(t_0)]_{i'}}{\sum_{j'}(\sum_{i'} C_{i'j'})} + \sum_{j\neq r}\Gamma^{-1}{}_{ij}(\sum_{j'} C_{jj'}\frac{\sum_{i'}[\dot{I}_{in}]_{i'}}{\sum_{j'}\sum_{i'} C_{i'j'}} - [\dot{I}_{in}]_j) \tag{C.16}$$

Plug into (C.8),

$$\dot{\tilde{V}}_r = -\frac{\sum_i [I_{in}(t_0)]_i}{\sum_j (\sum_i C_{ij})} - \frac{1}{(\sum_i C_{ir})} \sum_{j\neq r} \sum_{i'\neq r} (\sum_i C_{ij}) \Gamma^{-1}{}_{ji'} (\sum_{j'} C_{i'j'} \frac{\sum_{j'} [\dot{I}_{in}]_{j'}}{\sum_{j'} \sum_{i''} C_{i''j'}} - [\dot{I}_{in}]_{i'}) \quad \text{(C.17)}$$

The same procedure is repeated to find $\tilde{V}_{i\neq r}(t_0)$,

$$\begin{aligned} \tilde{V}_{i\neq r}(t_0) = & \frac{\sum_j (\sum_{i'} C_{i'j}) V_j(t_0)}{\sum_j (\sum_{i'} C_{i'j})} + \sum_{j\neq r} \Gamma^{-1}{}_{ji} [(\sum_{j'} C_{jj'} \frac{\sum_{i'} [I_{in}(t_0)]_{i'}}{\sum_{j'} \sum_{i'} C_{i'j'}} - [I_{in}(t_0)]_j) \\ & - \sum_{j\neq r} [\Gamma^{-1} \chi \Gamma^{-1}]_{ij} (\sum_{j'} C_{jj'} \frac{\sum_{j'} [\dot{I}_{in}]_{j'}}{\sum_{j'} \sum_{i''} C_{i''j'}} - [\dot{I}_{in}]_j] \end{aligned} \quad \text{(C.18)}$$

where χ is defined similar to Γ as

$$\chi_{ij} = C_{ij} - \frac{C_{ir} \sum_{i'} C_{i'j}}{\sum_{i'} C_{i'r}} \quad i \neq r, j \neq r \quad \text{(C.19)}$$

$$\begin{aligned} \tilde{V}_r(t_0) = & \frac{\sum_j (\sum_{i'} C_{i'j}) V_j(t_0)}{\sum_j (\sum_{i'} C_{i'j})} - \sum_{j\neq r} \frac{(\sum_i C_{ij})}{(\sum_i C_{ir})} [\sum_{i\neq r} \Gamma^{-1}{}_{ji} (\sum_{j'} C_{i'j'} \frac{\sum_{i'} [I_{in}(t_0)]_{i'}}{\sum_{j'} \sum_{i'} C_{i'j'}} - [I_{in}(t_0)]_{i'}) \\ & - \sum_{i\neq r} [\Gamma^{-1} \chi \Gamma^{-1}]_{ji} (\sum_{j'} C_{ij'} \frac{\sum_{j'} [\dot{I}_{in}]_{j'}}{\sum_{j'} \sum_{i'} C_{i'j'}} - [\dot{I}_{in}]_i)] \end{aligned} \quad \text{(C.20)}$$

APPENDIX D

ELMORE DELAY FOR SINGULAR CONDUCTANCE MATRIX

When the conductance matrix is singular, we cannot use G^{-1} to define the Elmore delay. Proceeding as we did for the steady state solution, we use charge conservation to define the transient for an arbitrary node *r* in terms of the other node transients.

$$\sum_i [C\frac{d}{dt}(V-\tilde{V}) + G(V-\tilde{V})_i] = 0 \tag{D.1}$$

$$\sum_j (\sum_i C_{ij})(V_j - \tilde{V}_j) = \sum_j (\sum_i C_{ij})(V_j(t_0) - \tilde{V}_j(t_0)) \tag{D.2}$$

Now use charge conservation to define the transient for an arbitrary node *r* in terms of the other node transients.

$$V_r - \tilde{V}_r = \frac{-\sum_{j\neq r} (\sum_i C_{ij})(V_j - \tilde{V}_j)}{\sum_i C_{ir}} \tag{D.3}$$

substituting for $V_r - \tilde{V}_r$ from Eq. (3.41) gives

$$G(V-\tilde{V}) + G_r(V_r - \tilde{V}_r) = -C\frac{d}{dt}(V-\tilde{V}) - C_r\frac{d}{dt}(V_r - \tilde{V}_r) \tag{D.4}$$

$$\Gamma(V-\tilde{V}) = -\chi\frac{d}{dt}(V-\tilde{V}) \tag{D.5}$$

$$(V-\tilde{V}) = -\Gamma^{-1}\chi\frac{d}{dt}(V-\tilde{V}) \tag{D.6}$$

Then we can get the Elmore delay for all nodes except node *r*,

$$\tau_{i\neq r} = \frac{[\Gamma^{-1}\chi\left|V(t_0)-\tilde{V}(t_0)\right|]_{i\neq r}}{\left|V_{i\neq r}(t_0)-\tilde{V}_{i\neq r}(t_0)\right|} \tag{D.7}$$

and the Elmore delay for node r,

$$\tau_r = \frac{-\sum_{j\neq r} C_{j0}\left|V_j(t_0)-\tilde{V}_j(t_0)\right|\tau_j}{C_{r0}\left|V_r(t_0)-\tilde{V}_r(t_0)\right|} \tag{D.8}$$

VITA

Jian Chang

Candidate for the Degree of

Doctor of Philosophy

Dissertation: PIECEWISE LINEAR DELAY MODELING OF DIGITAL VLSI CIRCUITS

Major Field: Electrical Engineering

Biographical:

Personal Data: Born in Taiyuan, Shanxi, P. R. China, on March 19, 1972, Son of Baolin Chang and Minli Ren.

Education: Received the Bachelor of Science in Mechanical Engineering degree from Beijing Institute of Machinery Industry in July 1993; received the Master of Science in Electrical Engineering degree from Oklahoma State University in December 2001; completed the requirements for the degree of Doctor of Philosophy in Electrical Engineering at Oklahoma State University in May 2006.

Experience: Worked with Beijing Jeep Corporation Ltd., China, as an engineer from August 1993 to July 1999. Employed by the School of Electrical and Computer Engineering as a Graduate Teaching Assistant from August 1999 to December 2001. Employed by the School of Electrical and Computer Engineering as a Graduate Research Assistant from January 2002 to May 2006.

Name: Jian Chang Date of Degree: May, 2006

Institution: Oklahoma State University Location: Stillwater, Oklahoma

Title of Study: PIECEWISE LINEAR DELAY MODELING OF DIGITAL VLSI CIRCUITS

Pages in Study: 137 Candidate for the Degree of Doctor of Philosophy

Major Field: Electrical Engineering

Scope and Method of Study: One of the most important performance measures of digital logic circuits is the delays of switching signals propagating through the logic gates of the circuit. Circuit simulators such as SPICE can find the delay by solving for the current and voltage waveforms as functions of time. Although SPICE can handle the complex, nonlinear behavior of the transistors, it takes a significant amount of computations. Usually no more than a few thousand transistors may be simulated in a reasonable amount of computation time. Simulator such as IRSIM uses the switch model to find the delay, which greatly improves the simulation speed and can process hundreds of thousands of transistors in a reasonable amount of time. But IRSIM predicts delays much less accurate than SPICE because of its delay model inaccuracies. In this paper, a piecewise linear delay model which can evaluate the propagation delay of a CMOS VLSI circuitry with a wide range of input slope is presented. The model also takes into account the influences of short circuit current and dynamic channel charges. By using simple piecewise linear current model and piecewise linear channel charge storage model, it is possible to simulate the modern digital logic circuits in a reasonable amount of time. This model is applicable not only to propagation delay calculation of simple gates but also to that of any general circuit topology.

Findings and Conclusions: Excellent agreements with SPICE simulation have been observed in a CMOS inverter, a two-input NAND gate, and an OAI gate cases. The piecewise linear model is capable of predicting the output waveforms and the propagation delay over a variety of input and output conditions. The piecewise linear model can handle large complicated circuits by introducing smaller resistance connected regions. The model is also scalable from one technology to another by choosing different set of technology related model parameters. In general, it is possible to implement a fast circuit simulator based on the piecewise linear model or include the piecewise linear model into an existing circuit simulator such as IRSIM because of its accuracy and speed advantages.

ADVISER'S APPROVAL: Dr. Louis G. Johnson