ICECS' 97, December 15-18, 1997, Cairo, EGYPT

DELAY EVALUATION OF STATIC CMOS GATES FOR SHORT-CHANNEL DEVICES

L. Bisdounis S. Nikolaidis[1] O. Koufopavlou

VLSI Design Laboratory,
Department of Electrical & Computer Engineering, University of Patras,
26500 Patras, Greece

[1] Electronics & Computers Division,
Department of Physics, Aristotle University of Thessaloniki,
54006 Thessaloniki, Greece

Abstract: In this paper an accurate, analytical model for the evaluation of the CMOS inverter transient response and propagation delay for short-channel devices is presented. An exhaustive analysis of the inverter operation is provided which results to accurate expressions of the output response to an input ramp. These analytical expressions are valid for all the inverter operation regions and input waveform slopes and take into account the influences of the short-circuit current during switching, and the gate-to-drain coupling capacitance. The α-power law MOS model which considers the carriers velocity saturation effects of short-channel devices, is used. The final results are in excellent agreement with SPICE simulations.

I. INTRODUCTION

Since, propagation delay is one of the most critical performance parameters in CMOS digital circuits, much effort has to be devoted for the extraction of accurate, analytical expressions for timing models of basic circuits. Using transistor level simulators with continuous-time modeling of the devices, like SPICE, can be very expensive in terms of storage and computation time. Hence, much of past research has addressed the development of analytical delay models, without the necessity of expensive numerical iterations.

The main goal of this work is the analytical evaluation of the propagation delay in a CMOS inverter. To achieve this, analytical expressions of the output waveform must be derived, directly from the differential equation describing the temporal evolution of the inverter output. It is important to model accurately the inverter operation, since several fast methods for reducing a CMOS gate to an equivalent inverter have been proposed [1].

Analytical expressions for the output waveform and the propagation delay, including the effect of the input waveform slope, was presented in [2] and [3], where the influence of the short-circuit current was neglected. These works are based on the Shichman-Hodges square-law MOS model [4] that ignores the carriers velocity saturation effect, which becomes prominent in short-channel devices. In [5], the differential equation describing the discharge of the load capacitor was solved for a rising input ramp considering the current through both transistors and the gate-to-drain capacitance. However, fitting methods are used, resulting in a semi-empirical model, which is still based on the square-law model.

Nabavi-Lishi and Rumin [6] presented a method for the calculation of the inverter delay, where a linear approximation of the output waveform based on empirical factors produced from SPICE simulations, is used. Moreover, an approximated version of the SPICE level-3 MOS model is used, where the reduction of the transistors saturation voltage due to the velocity saturation effect, is neglected.

Sakurai and Newton [7],[8] presented closed-form delay expressions for the CMOS inverter, based on the α-power (n-power in [8]) law MOS model which includes the carriers velocity saturation effect of short-channel devices. For the derivation of the output expression in [7] the short-circuit current is neglected and the delay expression is valid only for fast input ramps, while in [8] a fictitious input ramp is used which is clamped to ground for ramp voltages less than the switching voltage in order to approximate the CMOS inverter by a NMOS circuit. Also, in [7] and [8], the influence of the gate-to-drain coupling capacitance, is neglected. An extension in the delay expression of [7] for the case of very lightly loaded inverter and/or slow input signals is presented in [9] where a table of coefficients produced from SPICE simulations is used, but still for negligible short-circuit current. The delay model presented in [10] uses the α-power MOS model taking into account the short-circuit current of the CMOS inverter, through a two-step iterative approach. In this approach the output voltage and the currents through both transistors are assumed to be piece-wise linear. However, modeling the nonlinear behavior of the transistors using linear approximations contributes to inaccuracies.

In this paper, analytical expressions for the CMOS inverter output response to an input voltage ramp are derived which overcome the weaknesses of previous works. Based on these expressions, accurate analytical formulae for the evaluation of the propagation delay for all the cases of input ramps, are produced. The derived timing model takes into account the influences of the current through both transistors and the gate-to-drain coupling capacitance, without using empirical approaches based on simulation results or numerical methods. The presented model clearly shows the influence of the inverter design characteristics, the load capacitance, and the slope of the input waveform driving the inverter on the propagation delay. The $\acute{a}$-power law MOS model [7] which includes the carriers velocity saturation effect of short-channel devices, is used.

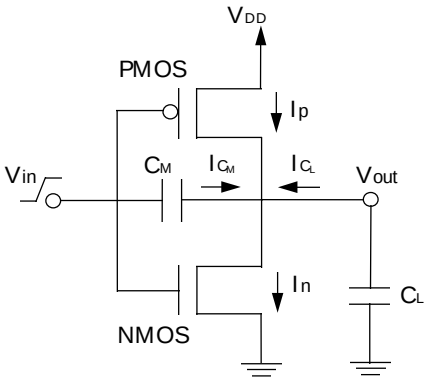

Fig.1: The CMOS Inverter

II. INVERTER TRANSIENT RESPONSE ANALYSIS

The derivations presented in the following are for a rising input ramp

$$V_{in} = \begin{cases} 0, & t \le 0 \\ V_{DD}\cdot(t/\tau), & 0 \le t \le \tau \\ V_{DD}, & t \ge \tau \end{cases} \quad (1)$$

where τ is the input rise time. The analysis for a falling input ramp is symmetrical, and similar results can be obtained by appropriate substitutions in the derived equations. The differential equation which describes the discharge of the load capacitance C_L for the CMOS inverter (Fig.1), taking into account the current through the gate-to-drain coupling capacitance (C_M), is derived from the application of the Kirchoff's current law to the output node

$$I_{C_L} + I_{C_M} + I_p - I_n = 0,$$

$$C_L \frac{dV_{out}}{dt} = C_M\left(\frac{dV_{in}}{dt} - \frac{dV_{out}}{dt}\right) + I_p - I_n. \quad (2)$$

The equivalent gate-drain capacitance C_M is the sum of the gate-to-drain capacitances of both transistors. The gate-to-drain capacitance of a transistor is the sum of the gate-to-drain overlap capacitance and a part of the gate-to-channel capacitance. The overlap part is voltage independent, and is given by: $C_{gd} = W\cdot C_{gdo}$, where W is the effective width of the transistor and C_{gdo} is the gate-drain overlap capacitance per unit channel width which is determined by the process technology. In the cutoff region of the transistor there is no conducting channel and in the saturation region the channel does not extend to the drain. Therefore, the gate-to-drain capacitance due to the channel charge is equal to zero. In the linear region the distributed gate-to-channel capacitance may be viewed as being shared equally between the source and the drain. In this case: $C_{gd\text{-}channel} = 0.5\cdot C_{ox}\cdot W\cdot L$, where C_{ox} is the gate-oxide capacitance per unit area and L is the effective length of the transistor.

For the expressions of the transistor currents the four-parameter α-power law MOS model is used. The parameters are the velocity saturation index (α), the drain current (I_{D0}) at $V_{GS} = V_{DS} = V_{DD}$, the drain saturation voltage (V_{D0}) at $V_{GS} = V_{DD}$ and the threshold voltage (V_{TH}). After normalizing voltages with respect to V_{DD}, i.e. $u_{in}=V_{in}/V_{DD}$, $u_{out} = V_{out}/V_{DD}$, $n=V_{THn}/V_{DD}$, $p = |V_{THp}|/V_{DD}$, $u_{don} = V_{D0n}/V_{DD}$, $u_{dop} = V_{D0p}/V_{DD}$ and using the variable x ($x = t/\tau$), the NMOS and PMOS device currents [7] of the CMOS inverter are given by the following equations:

$$I_n = \begin{cases} 0, & x \le n, & \text{Cutoff} \\ k_{sn}(x-n)^{\alpha_n}, & u_{out} \ge u'_{don}, & \text{Saturation,} \\ k_{ln}(x-n)^{\alpha_n/2}\,u_{out}, & u_{out} < u'_{don}, & \text{Linear} \end{cases} \quad (3)$$

where $k_{sn} = \dfrac{I_{D0n}}{(1-n)^{\alpha_n}}$, $k_{ln} = \dfrac{I_{D0n}}{u_{don}(1-n)^{\alpha_n/2}}$, and

$$u'_{don} = u_{don}\left(\frac{x-n}{1-n}\right)^{\frac{\alpha_n}{2}},$$

$$I_p = \begin{cases} k_{lp}(1-x-p)^{\alpha_p/2}(1-u_{out}), & 1-u_{out} < u'_{dop}, & \text{Linear} \\ k_{sp}(1-x-p)^{\alpha_p}, & 1-u_{out} \ge u'_{dop}, & \text{Saturation,} \\ 0, & x \ge 1-p, & \text{Cutoff} \end{cases} \quad (4)$$

where $k_{sp} = \dfrac{I_{D0p}}{(1-p)^{\alpha_p}}$, $k_{lp} = \dfrac{I_{D0p}}{u_{dop}(1-p)^{\alpha_p/2}}$, and

$$u'_{dop} = u_{dop}\left(\frac{1-x-p}{1-p}\right)^{\frac{\alpha_p}{2}}.$$

Since the input ramp will reach its final value with the NMOS device either in saturation or in the linear region two main cases of input ramps must be considered, in order to give a complete analysis of the output waveform. For *fast* input ramps, the NMOS device is still saturated while for *slow* input ramps the NMOS is in its linear region, when the input voltage ramp reaches its final value. The operation regions of the inverter are shown in Fig.2. The separation of the operating area in regions corresponds to the different combinations of the operation modes of the NMOS and PMOS devices (i.e. linear, saturation, cutoff).

Case A - Fast input ramps:

Region 1, $0 \le x \le n$: The NMOS transistor is off, and the PMOS transistor is in the linear region. Part of the charge from the input which injected through the coupling capacitance causes an overshoot at the early part of the output voltage waveform (Fig.2). During the overshoot the

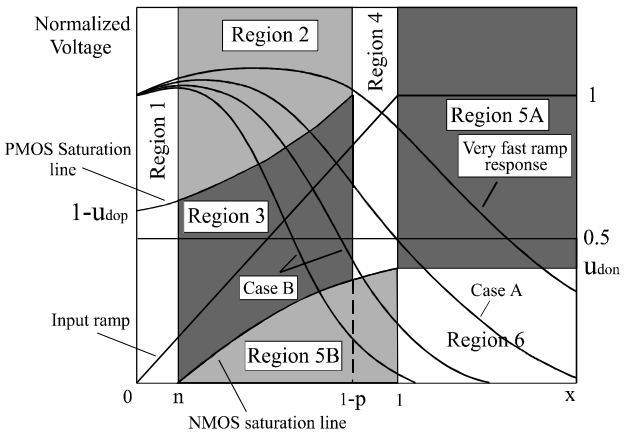

Fig.2: Operation regions of the inverter

PMOS device operates in a reversed linear mode because the output voltage is greater than the supply voltage. Since, in this region the differential equation (2) cannot be solved analytically, an average value of x (x_{av} = n /2) is used in the expression of the PMOS current, resulting in the following solution

$$u_{out} = 1 + c_m\, y_n^{-1}\left(1 - e^{-y_n x}\right), \quad \text{where} \qquad (5)$$

$$c_m = \frac{C_M}{C_L + C_M}\ ,\ y_n = A_{lp}\left(1 - p - \frac{n}{2}\right)^{\frac{\alpha_p}{2}}\ ,\ A_{lp} = \frac{k_{lp}\,\tau}{V_{DD}\left(C_L + C_M\right)}\ .$$

Region 2, $n \le x \le x_{satp}$:
The NMOS transistor is saturated and the PMOS transistor is still in the linear region. x_{satp} is the normalized time when the PMOS device enters saturation. In order to give a solution in the differential equation describing the discharge of the output load in this region an approximation for the PMOS current, is used (Fig.3). If we assume that the minimum of the PMOS current appears when the input voltage arrives at the NMOS threshold voltage (x = n), then we can approximate the PMOS current by a linear function of the normalized time

$$I_p = I_{pmin} + S\left(x - n\right), \qquad (6)$$

where I_{pmin} is calculated using the PMOS current equation in the linear region and the value of the normalized output voltage at x = n in equation (5), and S is the PMOS current slope. S is calculated by equating the exact PMOS current in the linear region as given from the *á*-power law MOS model in equation (4) with the approximated PMOS current in equation (6), at the point $x_c = (1-p)/2$ which is close to the middle of region 2 (see Fig.3). After the above approximation the solution of the equation (2) becomes

$$u_{out} = 1 + c_m\left(x - n + R\right) + I_{pmin}\, d\,\left(x - n\right) + \frac{S\, d\left(x-n\right)^2}{2} - \frac{A_{sn}\left(x-n\right)^{\alpha_n+1}}{\alpha_n + 1}\ , \qquad (7)$$

where $A_{sn} = \frac{k_{sn}\,\tau}{V_{DD}\left(C_L + C_M\right)}$, $d = \frac{\tau}{V_{DD}\left(C_L + C_M\right)}$, and $R = y_n^{-1}\left(1 - e^{-n y_n}\right)$.

The above equation gives waveforms very close to those derived from SPICE simulations (as shown in section IV), which indicates the validation of the PMOS current linear approximation. In order to continue the analysis for the next region, the evaluation of the normalized time value x_{satp} and the normalized output voltage value u_{satp} when the PMOS device saturates, is required. These values satisfy the PMOS saturation condition: $u_{out} = 1 - u'_{dop}$. In order to solve this equation a Taylor series expansion around the point $x = 1 - p - n$ up to the second order coefficient is used, for both u_{out} and u'_{dop}. After that, the PMOS saturation condition becomes

$$z_0 + z_1 x + z_2 x^2 = m_0 + m_1 x + m_2 x^2\,, \qquad (8)$$

where z's and m's are the Taylor series coefficients of u_{out} and u'_{dop} respectively. Standard ways of finding those coefficients can be found in most mathematical handbooks. The only root of the quadratic equation (8) which belongs in the interval $[n,\ 1-p]$, is x_{satp}. The error which is inserted in the evaluation of x_{satp} due to the above method in most cases is up to 0.3%. By substituting x_{satp} in equation (7) the normalized output voltage u_{satp} is evaluated.

As we can see in Fig.2, in the special case of very fast input ramps, the PMOS device is turned off after its linear region without enters saturation. This occurs because the output voltage overshoot finishes when the PMOS is already off. Hence, the inverter does not enter in region 3 and the calculation of x_{satp} and u_{satp} is not required.

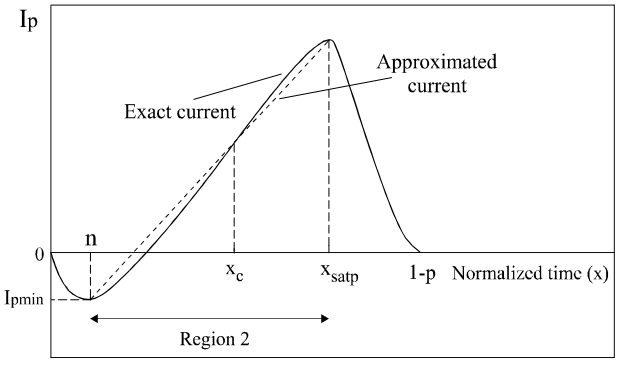

Fig.3: Approximation of the PMOS current in region 2

Region 3, $x_{satp} \le x \le 1-p$:
Both transistors are saturated. The analytical solution of the differential equation (2) becomes

$$u_{out} = u_{23} + c_m x - \frac{A_{sn}}{\alpha_n + 1}\left(x - n\right)^{\alpha_n+1} - \frac{A_{sp}}{\alpha_p + 1}\left(1 - x - p\right)^{\alpha_p+1}\ , \qquad (9)$$

where $A_{sp} = \frac{k_{sp}\,\tau}{V_{DD}\left(C_L + C_M\right)}$.

The integration constant u_{23} is inserted to ensure continuity with respect to region 2, and is given by

$$u_{23} = u_{satp} - c_m x_{satp} + \frac{A_{sn}}{\alpha_n + 1}\left(x_{satp} - n\right)^{\alpha_n+1} + \frac{A_{sp}}{\alpha_p + 1}\left(1 - x_{satp} - p\right)^{\alpha_p+1}\ .$$

Region 4, $1\text{-}p \le x \le 1$:
The NMOS transistor is saturated and the PMOS transistor is off. The analytical solution of the equation (2) is

$$u_{out} = u_{23} + c_m x - \frac{A_{sn}}{\alpha_n + 1}\left(x - n\right)^{\alpha_n+1}. \qquad (10)$$

Region 5A, $1 \le x \le x_{satn}$:
The input ramp has reached its final value with the NMOS transistor still in saturation and the PMOS transistor off. x_{satn} is the normalized time value when the NMOS transistor leaves saturation i.e. $u_{out} = u'_{don}$ (see equation (3)). The analytical solution of the differential equation (2) in this region is

$$u_{out} = u_{23} + c_m - \frac{A_{sn}}{\alpha_n + 1}(1-n)^{\alpha_n+1} - A_{sn}(1-n)^{\alpha_n}(x-1) \quad (11)$$

Region 6, $x \geq x_{satn}$:
The NMOS device enters in its linear region, and the PMOS is off. The analytical solution of the equation (2) is

$$u_{out} = u_{don}\, e^{-A_{ln}(1-n)^{\alpha_n/2}(x-x_{satn})}, \quad (12)$$

where $A_{ln} = \frac{k_{ln}\,\tau}{V_{DD}(C_L + C_M)}$.

x_{satn} is calculated from equation (11) for $u_{out} = u_{don}$.

Case B - Slow input ramps:
In the second case, slow input ramps such that the NMOS device leaves saturation while the input voltage is still a ramp, are studied. This occurs if the value of the normalized output voltage when the input ramp reaches its final value is lower than u_{don} (Fig.2). The output expressions for the regions 1, 2, 3 and 4 are the same with those of the previous case. As we can see in Fig.2, the normalized time value x_{satn} must be calculated from equation (10) for $u_{out} = u'_{don}$, which corresponds to the NMOS saturation condition (see equation (3)). In the case of slower input ramps the inverter doesn't enter in region 4. This occurs when the PMOS transistor is turned off and the NMOS transistor is already in the linear region $(x_{satn} < 1-p)$. In this case x_{satn} is calculated from equation (9) for $u_{out} = u'_{don}$. In order to solve those two equations a Taylor series expansion around the point $x = 1-p$ up to the second order coefficient is used for u'_{don}, and two more around the point $x = 1-p-n$ for the output expressions (9) and (10). Hence, the normalized time value x_{satn} becomes the root of a simple quadratic equation. The error which is inserted in the evaluation of x_{satn} due to the above method in both cases is up to 0.3%.

Region 5B, $x_{satn} \leq x \leq 1$:
The NMOS transistor is in the linear region and the PMOS transistor is either off or so poorly conducting that its influence can be neglected. SPICE simulations indicate that the PMOS device current in this region (for $x < 1-p$) is up to 2-3% of the NMOS device current. Neglecting the charging current through the gate-to-drain coupling capacitance an approximated solution of the equation (2) is

$$u_{out} = u_{satn}\, e^{-\frac{2A_{ln}}{\alpha_n+2}\left[(x-n)^{(\alpha_n+2)/2} - (x_{satn}-n)^{(\alpha_n+2)/2}\right]}, \quad (13)$$

where $u_{satn} = u_{don}\left[(x_{satn}-n)/(1-n)\right]^{\alpha_n/2}$ is the value of the normalized output voltage when the NMOS transistor leaves saturation.

Region 6, $x \geq 1$:
The input ramp has reached its final value, the NMOS device is still in the linear region and the PMOS device is off. The analytical solution of (2) becomes

$$u_{out} = u_{[1]}\, e^{-A_{ln}(1-n)^{\alpha_n/2}(x-1)}, \quad (14)$$

where $u_{[1]}$ is the value of the normalized output voltage when the input ramp reaches its final value and is calculated if we set x=1 in equation (13).

III. PROPAGATION DELAY EVALUATION

The fall propagation delay at the 50% voltage level is written as

$$t_{delay} = t_{0.5} - \frac{\tau}{2} = x_{0.5}\,\tau - \frac{\tau}{2}, \quad (15)$$

where $x_{0.5}$ is the normalized time value when $u_{out} = 0.5$. Thus, for the evaluation of the propagation delay, the normalized time value $x_{0.5}$ must be determined for both cases of input ramps.

In the case of fast input ramps the output voltage reaches the 50% level when the inverter operates in region 5A if $u_{[1]} \geq 0.5$ and in region 4 if $u_{[1]} \leq 0.5$. $u_{[1]}$ is the value of the normalized output voltage when the input ramp reaches its final value and is calculated from equation (10) for x = 1. When $u_{out} = 0.5$ occurs in region 5A, $x_{0.5}$ is calculated from equation (11)

$$x_{0.5} = \frac{c_m + u_{23} - 0.5}{A_{sn}(1-n)^{\alpha_n}} + \frac{\alpha_n + n}{\alpha_n + 1}. \quad (16)$$

In the case where $u_{out} = 0.5$ occurs in region 4, $x_{0.5}$ should be calculated from equation (10). In order to solve this equation the Taylor series expansion of the output expression (10), which was used for the calculation of x_{satn}, is used resulting to the following simple solution

$$x_{0.5} = \frac{-y_1 - \sqrt{y_1^2 - 2y_2(2y_0 - 1)}}{2y_2}, \quad (17)$$

where y_0, y_1, y_2 are the coefficients in the Taylor series expansion.

For slow input ramps the condition $u_{out} = 0.5$ occurs in region 4 if $u_{[1-p]} \geq 0.5$ and in region 3 if $u_{[1-p]} \leq 0.5$. $u_{[1-p]}$ is the value of the normalized output voltage when the PMOS device enters the cutoff region, and is calculated from equation (9) for x = 1–p. In the first case the normalized time value $x_{0.5}$ is given by equation (17), and in the second one is calculated using the Taylor series expansion of the output expression (9)

$$x_{0.5} = \frac{-w_1 - \sqrt{w_1^2 - 2w_2(2w_0 - 1)}}{2w_2}, \quad (18)$$

where w_0, w_1, w_2 are the coefficients in the Taylor series expansion. The error which is inserted in the calculation of x_{05} in regions 3 and 4 due to the use of the Taylor series expansions is up to 0.2%.

IV. RESULTS AND CONCLUSIONS

Figure 4 show some typical output waveforms produced from the expressions of section II. A 1.2μm CMOS process technology has been used to validate the accuracy of the presented inverter output waveform expressions. The model parameters and the dimensions of both transistors are listed in Table 1. The transistor widths have been selected in order to achieve equal drain currents (I_{D0}) at $V_{GS}=V_{DS}=V_{DD}$

Table 1: MOSFET parameters used in calculations

Parameter	*NMOS*	*PMOS*
L (μm)	1.2	1.2
W (μm)	5	11.75
I_{D0} (mA)	1.53	1.53
α	1.43	1.54
$\lvert V_{D0} \rvert$ (Volts)	1.70	2.50
$\lvert V_{TH} \rvert$ (Volts)	0.736	0.751
C_{ox} (fF/μm^2)	1.45	1.45
C_{gdo} (fF/μm)	0.30	0.30

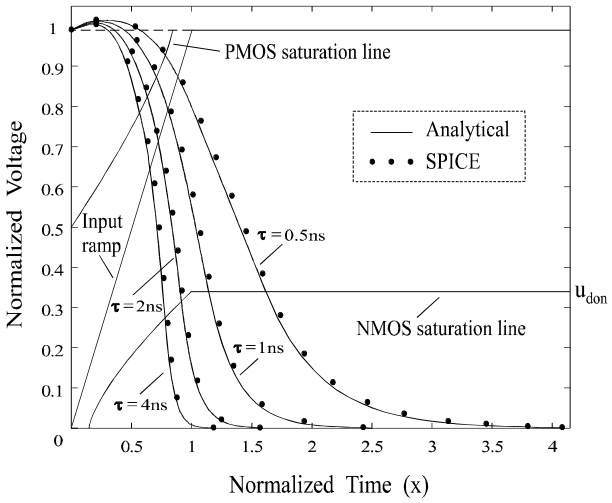

Fig.4: Inverter output waveforms

The output waveforms produced by SPICE simulations are added for comparison. A supply voltage of 5Volts and an output load of 0.2pF, have been used. It can be observed, that the analytical waveforms are very close to those produced by SPICE simulations. The output waveforms for input rise times 0.5ns and 1ns correspond to case A of the previous section analysis, while those for input rise times 2ns and 4ns correspond to case B.

In Figure 5 the inverter propagation delay for a rising input ramp, is plotted as a function of the input rise time. Results using the approaches for the evaluation of the propagation delay presented in [2], [6], [7] and [10], are also given. It can be observed, that the presented model gives results closer to those derived from SPICE simulations than the other methods. The error is less than 3%. This occurs because the proposed model includes the influences of the short-circuit current and the gate-to-drain coupling capacitance on the expressions of the inverter output waveform. The presented timing model can be used for more complex static gates, since several fast methods [1] have been proposed for reducing a CMOS gate to an equivalent inverter. The most critical issue in gate modeling is the reduction of the arrays of serial and parallel transistors to single transistors with equivalent drivabilities. Using these called "collapsing" techniques the propagation delay of a gate can be computed quickly and accurately using the inverter timing model and without the complications associated with trying to generalize the inverter model to complex gates. The development of an accurate timing model for the basic CMOS gate (inverter)

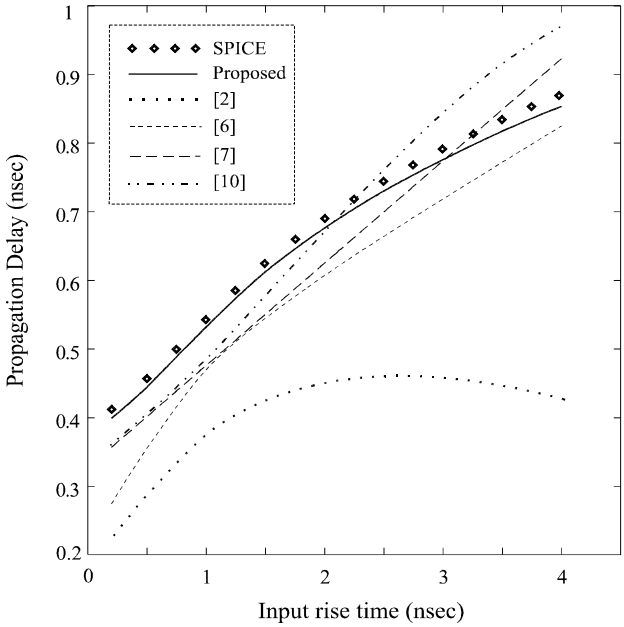

Fig.5: Inverter propagation delay

is of great importance in order to avoid accumulated errors in the procedure of the delay evaluation of static CMOS gates.

V. REFERENCES

[1] A. Nabavi-Lishi, N.C. Rumin, "Inverter models of CMOS gates for supply current and delay evaluation", *IEEE Trans. Computer-Aided Design*, vol. 13, pp. 1271-1279, Oct. 1994.

[2] N. Hedenstierna, K.O. Jeppson, "CMOS circuit speed and buffer optimization", *IEEE Trans. Computer-Aided Design*, vol. CAD-6, pp. 270-281, March 1987.

[3] A.I. Kayssi, K.A. Sakallah, T.M. Burks, "Analytical transient response of CMOS inverters", *IEEE Trans. Circuits and Systems - I*, vol. 39, pp. 42-45, Jan. 1992.

[4] H. Shichman, D.A. Hodges, "Modeling and simulation of insulated-gate field-effect transistor switching circuits", *IEEE J. Solid-State Circuits*, vol. SC-3, pp. 285-289, Sept. 1968.

[5] K.O. Jeppson, "Modeling the influence of the transistor gain ratio, and the input-to-output coupling capacitance on the CMOS inverter delay", *IEEE J. Solid-State Circuits*, vol. 29, pp. 646-654, June 1994.

[6] A. Nabavi-Lishi, N.C. Rumin, "Simultaneous delay and maximum current calculation in CMOS gates", *IEE Electronics Letters*, vol. 28, pp. 682-684, March 1992.

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

[8] T. Sakurai, A.R. Newton, "Delay analysis of series-connected MOSFET circuits", *IEEE J. Solid-State Circuits*, vol. 26, pp. 122-131, Feb. 1991.

[9] S. Dutta, S.S. Mahant Shetti, S.L. Lusky, "A comprehensive delay model for CMOS inverters", *IEEE J. Solid-State Circuits*, vol. 30, pp. 864-871, August 1995.

[10] S.H.K. Embabi, R. Damodaran, "Delay models for CMOS, BiCMOS, BiNMOS circuits and their applications for timing simulations", *IEEE Trans. Computer-Aided Design*, vol. 13, pp. 1132-1142, Sept. 1994.