

Analytical Transient Response and Propagation Delay Evaluation of the CMOS Inverter for Short-Channel Devices

L. Bisdounis, S. Nikolaidis, and O. Koufopavlou

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 in accurate expressions of the output response to an input ramp. Most of the factors which influence the inverter operation are taken into account. 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.

Index Terms— Circuit transient analysis, delay effects, delay estimation, inverters, short-channel MOSFET's.

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 the 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. The first closed-form delay expression based on the output response which was obtained directly from the differential equation describing the CMOS inverter operation was derived in [1] for a step input. 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. The differential equation describing the discharge of the load capacitor was solved in [5] for a rising input ramp considering the current through both transistors and the gate-to-drain coupling capacitance. However, in [5], fitting methods were used, resulting in a semi-empirical model, which is still based on the square-law MOS

model. Nabavi–Lishi and Rumin [6] presented a method for the calculation of the inverter delay, where a linear approximation of the output 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 drain saturation voltage due to the velocity saturation 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. 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 an 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. It is based on piecewise linear approximation of the output voltage and the currents through both transistors, resulting in inaccurate modeling of the nonlinear transistor's behavior.

In this paper, analytical expressions for the CMOS inverter output response to an input voltage ramp, which overcome the weaknesses of previous works, are derived. Based on these expressions, accurate analytical formulas 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 presimulation tables 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 α -power law MOS model [7], which includes the carriers' velocity saturation effect of short-channel devices, is used.

II. INVERTER TRANSIENT RESPONSE ANALYSIS

The derivations presented in the following are for a rising input ramp: $V_{in} = V_{DD} \cdot (t/\tau)$ for $0 \leq t \leq \tau$, $V_{in} = 0$ for $t \leq 0$, and $V_{in} = V_{DD}$ for $t \geq \tau$, where τ is the input rise time. The analysis for a falling input is symmetrical, and similar results are obtained by appropriate substitutions in the derived equations. The differential equation which describes

Manuscript received January 3, 1997; revised June 12, 1997.

L. Bisdounis and O. Koufopavlou are with the VLSI Design Laboratory, Department of Electrical and Computer Engineering, University of Patras, 26500 Patras, Greece.

S. Nikolaidis is with the Electronics and Computers Division, Department of Physics, Aristotle University of Thessaloniki, 54006 Thessaloniki, Greece.

Publisher Item Identifier S 0018-9200(98)00733-1.

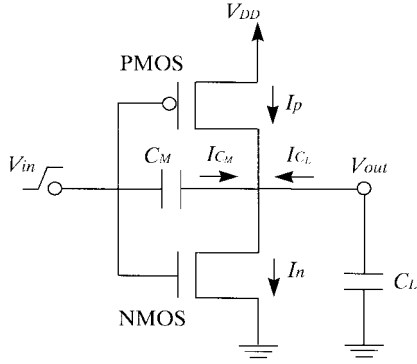


Fig. 1. The CMOS inverter.

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 at the output node:

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

For the rising input voltage ramp, the above equation becomes

$$\frac{dV_{out}}{dt} = \begin{cases} \frac{I_p - I_n}{C_L + C_M}, & t \leq 0 \text{ or } t > \tau \\ \frac{c_m V_{DD}}{\tau} + \frac{I_p - I_n}{C_L + C_M}, & 0 < t \leq \tau \end{cases} \quad (1)$$

where

$$c_m = \frac{C_M}{C_L + C_M}.$$

The output load C_L consists of the inverter drain junction capacitances, the gate capacitances of fanout gates, and the interconnect capacitance. The equivalent gate-drain capacitance C_M is the sum of the gate-to-drain capacitances of both transistors, which consist of the gate-to-drain overlap capacitance and a part of the gate-to-channel capacitance. It is calculated using the parameters C_{ox} (gate-oxide capacitance per unit area) and C_{gdo} (gate-drain overlap capacitance per unit channel width) [11].

For the expressions of the transistor currents, the four-parameter α -power law MOS model [7] 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 $x = t/\tau$, the NMOS device current is given by the following equation:

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

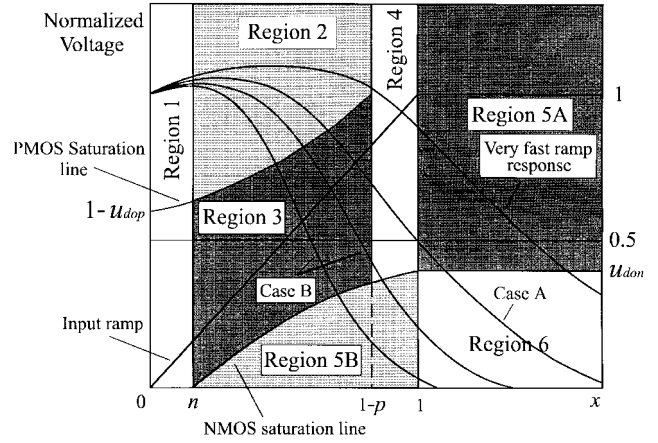


Fig. 2. Operation regions of the inverter.

where

$$k_{sn} = \frac{I_{D0n}}{(1-n)^{\alpha_n}}, \quad k_{ln} = \frac{I_{D0n}}{u_{don}(1-n)^{\frac{\alpha_n}{2}}}, \quad u'_{don} = u_{don} \left(\frac{x-n}{1-n} \right)^{\frac{\alpha_n}{2}}$$

and the PMOS current is given in a similar way.

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 are 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: In region 1 ($0 \leq x \leq n$), the NMOS transistor is off and the PMOS transistor is in the linear region. 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 (Fig. 2). This charge has the major influence on the output in region 1. Since in this region the differential equation (1) 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:

$$u_{out} = 1 + c_m y_n^{-1} (1 - e^{-y_n x}) \quad (3)$$

where

$$y_n = A_{lp} \left(1 - p - \frac{n}{2} \right)^{\frac{\alpha_p}{2}}$$

and

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

In region 2 ($n \leq x \leq x_{satp}$), the NMOS transistor is saturated and the PMOS transistor is still in the linear region.

$x_{\text{sat}p}$ is the normalized time value when the PMOS transistor is entering the saturation region. In order to give a solution in the differential equation describing the discharge of the output load in this region, an approximation of the PMOS current is used (Fig. 3). Assuming that the minimum of the PMOS current appears when the input voltage arrives at the NMOS threshold voltage ($x = n$), then the PMOS current can be approximated by a linear function of the normalized time: $I_p = I_{p\text{min}} + S(x - n)$. $I_{p\text{min}}$ is calculated using the PMOS current equation in the linear region and the value of the normalized output voltage at $x = n$ in (3). The current slope S is calculated by equating the exact PMOS current in the linear region with the approximated one at the point $x_c = (1 - p)/2$. After that, the output voltage waveform is described by

$$u_{\text{out}} = 1 + c_m(x - n + R) + I_{p\text{min}}d(x - n) + \frac{Sd(x - n)^2}{2} - \frac{A_{\text{sn}}(x - n)^{\alpha_n + 1}}{\alpha_n + 1} \quad (4)$$

where

$$R = y_n^{-1}(1 - e^{-ny_n})$$

$$d = \frac{\tau}{V_{\text{DD}}(C_L + C_M)}$$

$$A_{\text{sn}} = \frac{k_{\text{sn}}\tau}{V_{\text{DD}}(C_L + C_M)}.$$

The above equation gives waveforms very close to those derived from SPICE simulations (as shown in Section IV), which indicates the validity of the PMOS current linear approximation. In order to continue the analysis for the next region, the evaluation of the normalized time value $x_{\text{sat}p}$ and the normalized output voltage value $u_{\text{sat}p}$ when the PMOS device saturates is required. These values satisfy the PMOS saturation condition: $u_{\text{out}} = 1 - u'_{\text{dop}}$. In order to solve this equation, a Taylor series expansion [12] around the point $x = 1 - p - n$ up to the second-order coefficient is used, for both u_{out} and u'_{dop} . In the special case of very fast input ramps, the PMOS device is turned off after its linear region without entering saturation (Fig. 2). This occurs because the output voltage overshoot finishes when the PMOS device is already off.

In region 3 ($x_{\text{sat}p} \leq x \leq 1 - p$), both transistors are saturated. The analytical solution of (1) is

$$u_{\text{out}} = u_{23} + c_mx - \frac{A_{\text{sn}}}{\alpha_n + 1}(x - n)^{\alpha_n + 1} - \frac{A_{\text{sp}}}{\alpha_p + 1}(1 - x - p)^{\alpha_p + 1} \quad (5)$$

where

$$A_{\text{sp}} = \frac{k_{\text{sp}}\tau}{V_{\text{DD}}(C_L + C_M)}.$$

The integration constant u_{23} is inserted to ensure continuity with respect to region 2.

In region 4 ($1 - p \leq x \leq 1$), the NMOS transistor is saturated, the PMOS transistor is off, and the analytical

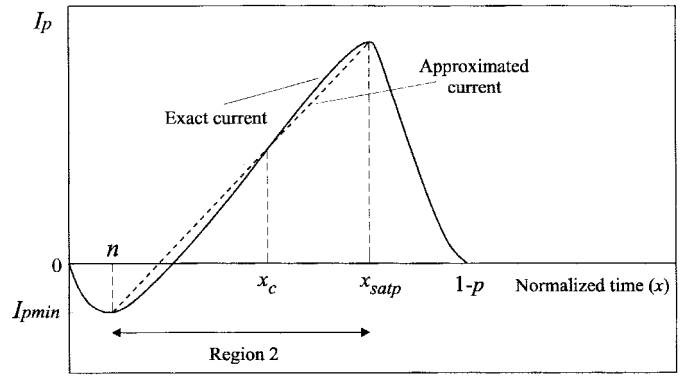


Fig. 3. Linear approximation of the PMOS current in region 2.

solution of the differential equation (1) is

$$u_{\text{out}} = u_{23} + c_mx - \frac{A_{\text{sn}}}{\alpha_n + 1}(x - n)^{\alpha_n + 1}. \quad (6)$$

In the next region, 5A ($1 \leq x \leq x_{\text{sat}n}$), the input ramp has reached its final value without any changes in the operating mode of the transistors. The analytical solution of the differential equation (1) in this region is

$$u_{\text{out}} = u_{23} + c_m - \frac{A_{\text{sn}}}{\alpha_n + 1}(1 - n)^{\alpha_n + 1} - A_{\text{sn}}(1 - n)^{\alpha_n}(x - 1). \quad (7)$$

$x_{\text{sat}n}$ is the normalized time value where the NMOS transistor leaves saturation and is calculated from the above equation for $u_{\text{out}} = u_{\text{don}}$.

Finally, in region 6 ($x \geq x_{\text{sat}n}$), the NMOS device enters in its linear region and the PMOS device remains off. The output voltage waveform is described by

$$u_{\text{out}} = u_{\text{don}}e^{-A_{\text{ln}}(1 - n)^{\alpha_n/2}(x - x_{\text{sat}n})} \quad (8)$$

where

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

Case B—Slow Input Ramps: The output expressions for regions 1, 2, 3, and 4 are the same as those of the previous case. In this case the normalized time value $x_{\text{sat}n}$ should be calculated from (6) for $u_{\text{out}} = u'_{\text{don}}$, and in the case of slower inputs (Fig. 2) from (5). 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 Taylor series expansions around the point $x = 1 - p - n$ for the output expressions (5) and (6).

In region 5B ($x_{\text{sat}n} \leq x \leq 1$), 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 (1) in this region is

$$u_{\text{out}} = u_{\text{sat}n}e^{-\frac{2A_{\text{ln}}}{\alpha_n + 2}[(x - n)^{(\alpha_n + 2)/2} - (x_{\text{sat}n} - n)^{(\alpha_n + 2)/2}]} \quad (9)$$

where

$$u_{satn} = u_{don} \left(\frac{x_{satn} - n}{1 - n} \right)^{\frac{\alpha_n}{2}}.$$

In the last region ($x \geq 1$), the input ramp has reached its final value, the NMOS device is still in the linear region, and the PMOS device is off. The output voltage is given by

$$u_{out} = u_{[1]} e^{-A_{In}(1-n)^{\alpha_n/2}(x-1)} \quad (10)$$

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 (9).

III. PROPAGATION DELAY EVALUATION

The fall propagation delay at the 50% voltage level is written as: $t_d = t_{0.5} - \frac{\tau}{2} = x_{0.5}\tau - \frac{\tau}{2}$, 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. The normalized drain saturation voltage u_{don} (Fig. 2) is a critical parameter in order to find in which region the 50% level of the output voltage occurs. Hence, it is necessary to consider two possibilities in the delay calculation: $u_{don} \leq 0.5$ and $u_{don} \geq 0.5$. When $u_{don} \leq 0.5$, in the case of fast input ramps, the output voltage reaches the 50% level when the inverter operates in region 5A if $u_{[1]} \geq 0.5$ or in region 4 if $u_{[1]} \leq 0.5$. $u_{[1]}$ is calculated from (6) for $x = 1$. When $u_{out} = 0.5$ occurs in region 5A, $x_{0.5}$ is calculated from (7)

$$x_{0.5} = \frac{c_m + u_{23} - 0.5}{A_{sn}(1-n)^{\alpha_n}} + \frac{\alpha_n + n}{\alpha_n + 1}. \quad (11)$$

In the case where $u_{out} = 0.5$ occurs in region 4, $x_{0.5}$ should be calculated from (6). In order to solve this equation, the Taylor series expansion of the output expression (6), which was also used for the calculation of x_{satn} , is used. After that, $x_{0.5}$ becomes the root of a simple quadratic equation. For slow input ramps, the condition $u_{out} = 0.5$ occurs in region 4 if $u_{[1-p]} \geq 0.5$ or 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 (5) for $x = 1 - p$. In the first case, the normalized time value $x_{0.5}$ is calculated as described above, and in the second one, it is calculated using the Taylor series expansion of the output expression (5). The error introduced in the calculation of $x_{0.5}$ in regions 3 and 4 due to the use of the Taylor series expansions is up to 0.2%.

When $u_{don} \geq 0.5$, for fast input ramps $u_{out} = 0.5$ occurs in region 6 and $x_{0.5}$ is calculated from (8)

$$x_{0.5} = x_{satn} - \frac{\ell n(0.5/u_{don})}{A_{In}(1-n)^{\alpha_n/2}}. \quad (12)$$

In the case of slow input ramps, the output voltage reaches the 50% level when the inverter operates in region 6 except if $u_{[1]} \geq 0.5$, where there are two possibilities. Depending on u_{satn} , the output voltage reaches the 50% level in region 5B ($u_{satn} \geq 0.5$) or in region 3 ($u_{satn} \leq 0.5$). In region 6, $x_{0.5}$ is calculated from (10)

$$x_{0.5} = 1 - \frac{\ell n(0.5/u_{[1]})}{A_{In}(1-n)^{\alpha_n/2}}. \quad (13)$$

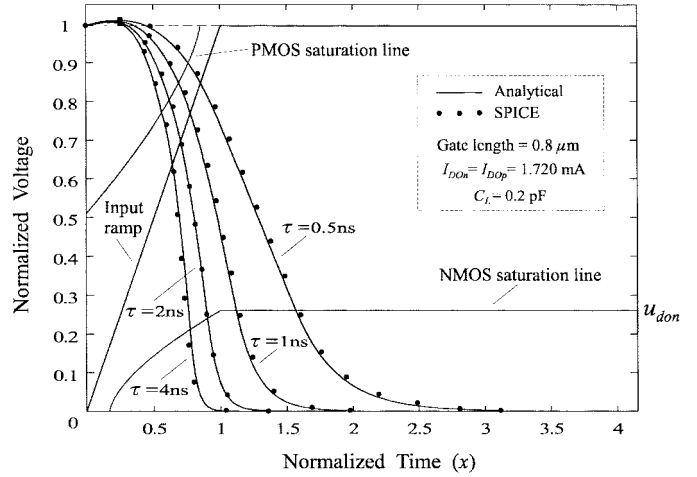


Fig. 4. Inverter output waveforms for several values of input rise time.

TABLE I
MOSFET MODEL PARAMETERS USED IN CALCULATIONS

Device	L (μm)	W (μm)	I_{D0} (mA)	α	$ V_{D0} $ (Volts)	$ V_{T1} $ (Volts)	C_{ox} (fF/ μm^2)	C_{gd} (fF/ μm)
NMOS	0.8	4	1.720	1.29	1.30	0.844	2.18	0.35
PMOS	0.8	6.55	1.720	1.41	2.45	0.734	2.18	0.35
		4	1.063	1.42				
		2	0.534	1.43				

In region 5B, $x_{0.5}$ is calculated from (9)

$$x_{0.5} = n + \left[(x_{satn} - n)^{(\alpha_n+2)/2} - \frac{(\alpha_n + 2)\ell n(0.5/u_{satn})}{2A_{In}} \right]^{\frac{2}{\alpha_n+2}} \quad (14)$$

and in region 3 is calculated using the Taylor series expansion of the output expression (5).

IV. RESULTS AND CONCLUSIONS

Fig. 4 shows some typical output waveforms produced from the expressions of Section II. A CMOS process technology of 0.8- μm 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 I. For the extraction of the results in Fig. 4, the transistor widths have been selected in order to achieve equal drain currents (I_{D0}) at $V_{GS} = V_{DS} = V_{DD}$. The output waveforms produced by SPICE level 3 simulations are added for comparison. A supply voltage of 5 V and an output load of 0.2 pF 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 of 0.5 ns and 1 ns correspond to case A, while those for input rise times of 2 ns and 4 ns correspond to case B. In Fig. 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 [6]–[8] and [10] are also given. The presented model gives results closer to those derived from

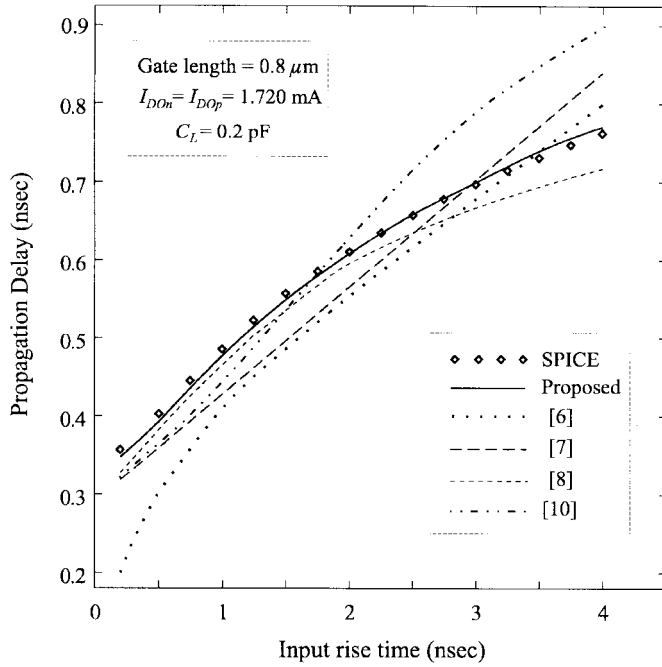


Fig. 5. Comparison of simulated and calculated inverter propagation delay from the proposed and previous models.

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. In Fig. 6 the inverter propagation delay is plotted as a function of input rise time for several values of the W_p/W_n ratio. Results from SPICE simulations are also given for comparison. It can be observed that the formulas for the propagation delay evaluation of the previous section are valid in a wide range of input transition times and the channel-width ratio of PMOS and NMOS devices. As shown in Fig. 6, the nonlinearity of the delay curves which appear for slow input ramps ($\tau > 1.5$ ns) is modulated by the W_p/W_n ratio. When the W_p/W_n ratio is increased, the PMOS device current during the regions where PMOS device is *on* increases, growing the reaction to the output node discharge. This is due to the reduction of the discharge current, which results in an increase of the propagation delay. For $W_p/W_n = 0.5$, a decrease of the propagation delay is caused for quite slow input transitions ($\tau > 3$ ns), while for $W_p/W_n \geq 1$, the delay is still growing for all the input transitions considered. The reduction of the delay in the first case occurs due to the asymmetry in the inverter which results in a logic threshold voltage sufficiently lower than $V_{DD}/2$, where the delay is measured.

The presented timing model can be used for more complex static gates, since several fast methods [13] have been proposed for reducing a gate to an equivalent inverter. Using these so-called “collapsing” techniques, the propagation delay of a gate can be computed quickly and accurately using the inverter

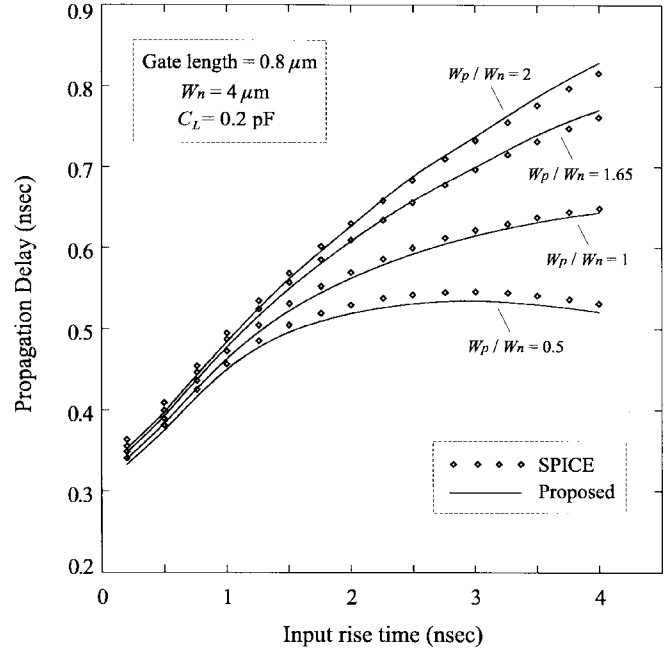


Fig. 6. Inverter propagation delay for several values of W_p/W_n .

timing model and without the complications associated with trying to generalize the inverter-based model to complex gates.

REFERENCES

- [1] J. R. Burns, “Switching response of complementary symmetry MOS transistor logic circuits,” *RCA Rev.*, vol. 25, pp. 627–661, Dec. 1964.
- [2] N. Hedenstierna and K. O. Jeppson, “CMOS circuit speed and buffer optimization,” *IEEE Trans. Computer-Aided Design*, vol. CAD-6, pp. 270–281, Mar. 1987.
- [3] A. I. Kayssi, K. A. Sakallah, and T. M. Burks, “Analytical transient response of CMOS inverters,” *IEEE Trans. Circuits Syst.-I*, vol. 39, pp. 42–45, Jan. 1992.
- [4] H. Shichman and 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 and N. C. Rumin, “Simultaneous delay and maximum current calculation in CMOS gates,” *IEE Electron. Lett.*, vol. 28, pp. 682–684, Mar. 1992.
- [7] T. Sakurai and 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, Apr. 1990.
- [8] —, “Delay analysis of series-connected MOSFET circuits,” *IEEE J. Solid-State Circuits*, vol. 26, pp. 122–131, Feb. 1991.
- [9] S. Dutta, S. S. M. Shetti, and S. L. Lusky, “A comprehensive delay model for CMOS inverters,” *IEEE J. Solid-State Circuits*, vol. 30, pp. 864–871, Aug. 1995.
- [10] S. H. K. Embabi and 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.
- [11] N. H. E. Weste and K. Eshraghian, *Principles of CMOS VLSI Design: A Systems Perspective*. New York: McGraw-Hill, pp. 183–191, 1993.
- [12] M. R. Spiegel, *Mathematical Handbook of Formulas and Tables*. New York: McGraw-Hill, p. 110, 1968.
- [13] A. Nabavi-Lishi and 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.