NBSIR 87-3668

The Effect of the Gate Oxide Thickness on the Speed of MOS Integrated Circuits

J. S. Kim

U.S. DEPARTMENT OF COMMERCE
National Bureau of Standards
National Engineering Laboratory
Center for Electronics and Electrical Engineering
Semiconductor Electronics Division
Gaithersburg, MD 20899

December 1987

U.S. DEPARTMENT OF COMMERCE
NATIONAL BUREAU OF STANDARDS

NBSIR 87-3668

THE EFFECT OF THE GATE OXIDE THICKNESS ON THE SPEED OF MOS INTEGRATED CIRCUITS

J. S. Kim

U.S. DEPARTMENT OF COMMERCE
National Bureau of Standards
National Engineering Laboratory
Center for Electronics and Electrical Engineering
Semiconductor Electronics Division
Gaithersburg, MD 20899

December 1987

U.S. DEPARTMENT OF COMMERCE, C. William Verity, *Secretary*
NATIONAL BUREAU OF STANDARDS, Ernest Ambler, *Director*

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

THE EFFECT OF THE GATE OXIDE THICKNESS ON THE SPEED OF MOS INTEGRATED CIRCUITS

J. S. KIM

IC Technology Group
Semiconductor Electronics Division
National Bureau of Standards
Gaithersburg, Maryland 20899

ABSTRACT

A simple analysis is presented for the effect of the gate oxide thickness on the circuit speed in a short-channel CMOS inverter delay circuit. The present analysis is performed within the first-order theory of the MOS transistor. The result of the analysis shows that an optimum value of the gate oxide thickness exists, beyond which a further scaling of the gate oxide will not improve, but will degrade, the circuit speed. The circuit speed corresponding to this optimum oxide thickness is the ultimate upper limit theoretically possible in a given MOS integrated circuit. The optimum value of the gate oxide thickness, to the first order approximation, is proportional to the channel width W, but it is independent of the channel length L. In particular, for wide channel devices, this optimum value exceeds the 5-nm to 30-nm range, which is of practical significance in the design and processing of advanced VLSI circuits. At the optimum oxide thickness, the square-root of the net propagation delay resolves into two components; the square-root of the net delay is the sum of the square-root of a purely parasitic component and that of a device-dependent component. Also it is shown that the transverse field dependence of the channel mobility enhances the parasitic effect; i.e., it tends to increase further the the optimum thickness of the gate oxide.

Key words: device optimization; oxide scaling; propagation delay; sealing effect; thin gate oxide; VLSI.

I. INTRODUCTION

The speed of the signal propagation is a key parameter in the performance of MOS integrated circuits. The primary objective of the present VLSI or VHSIC effort is to obtain the circuit speeds and the device packing densities as high as possible. This will result in the smallest possible size of the individual device. The scaling of the device geometry, on the other hand, imposes many difficulties on both device processing and modeling. In particular, the scaling of the gate oxide thickness causes severe yield and reliability problems. Since the supply voltage cannot be reduced by the same scaling factor as other parameters, the electric field in the gate oxide of a scaled MOS device is very high under normal operating conditions. Aside from these reliability and yield problems, the scaling of the gate oxide thickness may affect the circuit speed of MOS integrated circuits.

Traditionally, the circuit speed analysis has been focused mainly on the effect of the channel length but seldom on the scaling effect of the gate oxide thickness [1]. It is generally

assumed that the thinner the gate oxide, the better the circuit performance. However, this assumption is valid only for an ideal case where each individual transistor does not have any parasitics. In real circuits, the individual transistor always is associated with parasitics such as the series resistance or the stray capacitance. The inclusion of these parasitics in the analysis of the circuit speed has seldom been reported.

The purpose of the present paper is to show that inclusion of the transistor parasitics results in a finite lower bound of the gate oxide thickness, beyond which a further scaling of the gate oxide thickness does not improve, but degrades, the circuit speed. The present analysis is performed within the first-order theory of the MOS transistor, neglecting second-order effects such as the short-channel effect, the channel-length modulation, and the charge sharing. But the effect of the mobility degradation arising from the high transverse electric field is considered in this work.

Qualitatively, the existence of a lower bound for the optimum oxide thickness can be explained as follows. The propagation delay is approximately proportional to the RC time constant of the individual MOSFET. The capacitive and resistive components of the individual MOSFET should be considered to be a combination of parasitics and components resulting from the device itself. In the region where the gate oxide thickness is very small, the gate capacitance increases toward infinity as the oxide thickness decreases toward zero, but the resistance component remains as a non-zero, finite quantity. Therefore, the RC time constant, which is the product of these two quantities, increases. In the other region, where the thickness is very large, the channel resistance of the transistor increases with increasing oxide thickness, but the capacitance component remains as a non-zero value due to the parasitic capacitance. Thus the circuit speed again degrades as the oxide thickness increases. Therefore, an optimum thickness for the gate oxide must exist in the intermediate region.

The propagation delay corresponding to this optimum value of the gate oxide thickness gives the fastest circuit speed theoretically possible in a given MOS integrated circuit. It is very important for the design and processing of integrated circuits to determine this optimum value, for it determines not only the optimum circuit performance, but also relaxes many constraints which are imposed on the present device processing by the stringent device scaling.

II. GENERAL CONSIDERATION

In this section, a simple analytical expression is formulated for the signal propagation delay in a CMOS inverter chain. Figure 1 shows the equivalent circuit for the first several stages of the delay chain. P, N, V, I, and τ denote the p-channel transistor, the n-channel transistor, the node voltage, the current, and the delay, respectively, in each stage. The subscript of these quantities identifies the corresponding stage. C_p is the parasitic capacitance. The series resistance R_s, which may include all the resistive components except the transistor channel resistance in each stage, is assumed to be distributed symmetrically at both ends of the transistor. Considering the carrier mobility ratio and the resulting design optimization, we may assume:

$$
\begin{aligned}
L_N &= L_P = L \\
W_N &= \frac{W_P}{2} = W \\
\mu_N &= 2\mu_P = \mu \\
V_{TN} &= -V_{TP} = V_T \\
V_{ON} &= -V_{OP} = V_O,
\end{aligned}
\tag{1}
$$

where L, W, μ, V_T, and V_O are the channel length, the channel width, the carrier mobility, the threshold voltage, and the switching voltage, respectively. The switching voltage V_O is defined as the voltage which triggers the switching action of the next stage transistor and can be set $V_O \approx 2V_T$. The subscript P and N denote the p- and n-channel device, respectively. From eq (1), the gate capacitance of the transistors is:

$$
\begin{aligned}
C_{gN} &= WLC_o \qquad (n\text{-channel device}) \\
C_{gP} &= 2WLC_o \qquad (p\text{-channel device})
\end{aligned}
\tag{2}
$$

where $C_o = \epsilon / X_o$, $\epsilon \approx 3.4 \times 10^{-13} F/cm$; C_{gN} and C_{gP} are the gate capacitance of n- and p-channel MOSFET, respectively.

The net propagation delay in each stage can be considered to be a combination of two components, namely, the inversion layer build-up time and the gate capacitance charging time as follows:

$$
\begin{aligned}
\tau_i &= \tau_{inv} + \int_{\tau_{inv}}^{\tau_i} dt \\
&= \tau_{inv} + \int_{V_u}^{V_l} \frac{C_p + 3C_o WL}{I_i} dV_i,
\end{aligned}
\tag{3}
$$

where τ_i, I_i, and V_i are the propagation delay, the current, the voltage in the i^{th} stage, respectively; V_u and V_l are appropriate integration limits; and τ_{inv} is the time for the inversion layer build-up. Note that, for a given delay circuit, I_i is uniquely determined by V_{i-1} and V_i which determine the gate voltage and the drain voltage of the active transistor in the i^{th} stage.

τ_i is defined as the time interval between the switching onset time of the i^{th} stage and that of the $(i+1)^{th}$ stage. Generally, the delay corresponding to a particular stage may be different from that of an adjacent stage because of gradual degradation of the input signal propagating along the delay chain. In particular, for a step input

$$
V_{in} = \begin{cases} 0 & (t < 0) \\ V_{DD} & (t \geq 0), \end{cases}
\tag{4}
$$

the series $\{\tau_{2i+1}\}$ and $\{\tau_{2i}\}$ converge to an upper limit τ_N and τ_P, respectively, as i increases. Figure 2 shows the evolution of the step input as a function of time t for the first three stages. If the convergence is rapid enough, the propagation delay can be

defined as the average of these two characteristic time constants [2]. Table I shows the inter-relationship of relevant quantities between adjacent stages.

III. EFFECT OF SCALING ON PROPAGATION DELAY

Since the first stage delay τ_1 is a good indicator of the net delay τ_{pd}, only the first stage delay τ_1 is evaluated, which is denoted simply as τ from now on. At $t = 0$, the n-channel transistor N_1 is turned on, hence the drain current I_1 of transistor N_1 begins to discharge the net gate capacitance of the first stage, as shown in figures 1 and 2. Note that this net capacitance $3WLC_o + C_p$ is the sum of the gate capacitances of transistor N_2 and transistor P_2, and the parasitic capacitance. Subsequently, as V_1 decreases from V_{DD}, the effect gate voltage of transistor P_2 increases. At $t = \tau$, $V_1 = V_{DD} - V_o$, hence the switching of the second stage begins, i.e., transistor P_2 becomes active. Equation (3) may be rewritten, if the term τ_{inv} is neglected for simplicity, as follows:

$$\tau \approx \int_{V_{DD}}^{V_{DD}-V_O} \frac{(C_p + 3WLC_o)}{I} dV. \tag{5}$$

The subscript 1 is left out in eq (5) for simplicity.

The expression for the drain current I [3] should be modified as follows:

$$I = \mu C_o \left(\frac{W}{L}\right)(V_{DD} - V_T - \frac{V}{2})(V - IR_s). \tag{6}$$

By re-arranging eq (6), we obtain for two operations of the switching transistor:

$$\frac{1}{I} = \begin{cases} \frac{R_s}{V_s} + \frac{2L}{\mu C_o W V_s^2} & \text{(saturation region)} \\ \frac{R_s}{V} + \frac{L}{\mu C_o W (V_s - \frac{V}{2})V} & \text{(linear region),} \end{cases} \tag{7}$$

where $V_s = V_{DD} - V_T$ and $V_O = 2V_T$. Assume that the transistor N_1 of the first stage operates in the saturation region for $t \leq \tau$, since both the effective gate voltage and the drain voltage of transistor N_1 are V_{DD} initially for a step input signal. This may not generally be true as the input signal degrades at a later stage, where the linear region of the switching transistor operation becomes important. Therefore, one can expect a mixture of these two regions of operation in a single switching action and should consider both the linear- and the saturation-region of operation in performing the integral in eq (5). Fortunately, the only difference between these two cases is a slight difference in the numerical scaling factor α contained in the expression of τ, as will be seen shortly.

The final expression for the propagation delay can be obtained by performing the integration from eqs (5) and (7):

$$\tau = \alpha(\tau_o + a_1 L + a_2 L^2) \tag{8}$$

where τ_o, a_1, and a_2 are given by

$$\begin{aligned} \tau_o &= R_s C_p \\ a_1 &= \frac{2C_p}{\mu C_o V_s W} + 3C_o R_s W \\ a_2 &= \frac{6}{\mu V_s} \end{aligned} \tag{9}$$

where $V_s = V_{DD} - V_T$. If the switching transistor operates either in the linear region or in the saturation region only, then from eqs (5) and (7), the scaling factor α is given by:

$$\alpha = \begin{cases} \frac{2V_T}{V_{DD}-2V_T} & \text{(saturation region)} \\ \ln\left(\frac{V_{DD}}{V_{DD}-2V_T}\right) & \text{(linear region).} \end{cases} \tag{10}$$

Equation (10) shows that, as mentioned earlier, the scaling factor α depends only on voltage values through integration limits in eq (5) and the operation mode of the the switching transistor. By superposition, the scaling factor α can be expressed generally as a weighted average of the value of the linear region and that of the saturation region. Therefore, the expression of eq (8) for the propagation delay should be applicable more generally than under specific conditions considered in the present analysis, if the numerical factor α is modified accordingly.

There are two important features in eqs (8) to (10) which are:

1) The propagation delay is of a quadratic form with respect to the channel length L. The constant term $\tau_o = R_s C_p$ depends only on the parasitic components and is completely independent of the transistor parameters. The propagation delay does not have an extremum with respect to the channel length for $L > 0$; hence it increases monotonically with increasing channel length. This is the rationale behind the whole effort on the submicron VLSI technology.
2) The gate oxide thickness dependence of the propagation delay appears only in the coefficient a_1 through $C_o = \epsilon / X_o$. As will be shown later, an optimum value for the gate oxide thickness exists which provides the smallest possible propagation delay. This aspect is as a direct result of including parasitics in the analysis and may have a significant impact on the design and fabrication of scaled-down VLSI circuits.

In order to assess quantitatively the effect of scaling of the gate oxide, note that the dependence of the propagation delay on the thickness of the gate oxide X_o is contained only in the coefficient a_1 in eq (9). Therefore, one obtains from eqs (8) and (9)

$$\begin{aligned} \frac{\partial \tau}{\partial X_o} &= \frac{\partial a_1}{\partial X_o} \\ &= \left(\frac{\partial a_1}{\partial C_o}\right)\left(\frac{\partial C_o}{\partial X_o}\right) \\ &= \left[-\frac{2C_p}{\mu V_s W {C_o}^2} + 3R_s W\right]\frac{\partial C_o}{\partial X_o}. \end{aligned} \tag{11}$$

In performing partial differentials in eq (11), it is assumed that the threshold voltage V_T will be independent of the oxide thickness, since the threshold voltage of the transistors is usually predetermined by the circuit optimization and is kept constant accordingly during the device fabrication process. The dependence of the channel mobility on the gate oxide thickness, which manifests itself as the mobility degradation due to the increased transverse field for thinner gate oxides, is also neglected for the time being, but its net effect will be treated qualitatively later. By equating the right-hand-side of eq (11) to 0, one obtains

$$-\frac{2C_p}{\mu V_s W C_o^2} + 3R_s W = 0. \tag{12}$$

Equation (12) has only one positive root with respect to C_o. The corresponding value of the gate oxide thickness X_m is given by

$$X_m = \epsilon W \sqrt{\frac{3\mu V_s R_s}{2C_p}}, \tag{13}$$

where the relationship $C_o = \epsilon / X_o$ is used. Since the second derivative of a_1, hence τ, with respect to X_o is always positive as can be shown from eqs (8) and (9), the propagation delay takes its minimum value at $X_o = X_m$ if plotted against X_o. From eqs (8) and (13), this minimum delay τ_m is given by

$$\sqrt{\tau_m} = \sqrt{\alpha}(\sqrt{\tau_o} + \sqrt{\tau_d}) \tag{14}$$

where

$$\begin{aligned} \tau_o &= R_s C_p \\ \tau_d &= \frac{6L^2}{\mu V_s} \end{aligned} \tag{15}$$

The sum rule as shown in eq (14) implies that the square-root of the net propagation delay can be resolved into two components: a purely parasitic term which is entirely independent of the transistor used, and a purely device dependent term associated only with the carrier transport in the channel.

Figure 3 shows the propagation delay per stage, τ, as a function of the gate oxide thickness X_o for several different values of relevant device parameters, each curve showing a minimum. The values of optimum oxide thickness for various combinations of relevant parameters are listed in Table II. Note that X_m varies from several nanometers to several tens of nanometers, which typically represent the thickness of the gate oxide used in the present and future VLSI circuits. The existence of an optimum value X_m has a very important bearing in microelectronics technology. The circuit speed cannot be improved indefinitely by scaling-down the gate oxide thickness, but a further scaling of the gate oxide thickness beyond the optimum value X_m will degrade the circuit speed.

By adjusting the geometrical parameters L and W of MOS transistors, X_m and τ_m can be varied. Therefore, the existence of the optimum value X_m has a very important bearing on

the design of the individual MOSFET in an MOS integrated-circuit. Note again that the optimization of the circuit speed through the gate oxide thickness adjustment affects only the linear term in the net propagation — the term which is proportional to the channel length of the MOSFET as shown in eq (8). One of the main objectives of the future VLSI technology is to increase the circuit speed by reducing the channel length. In eq (8), the first term τ_o in the parenthesis on the right-hand-side will not be affected by the device scaling and the third term will be negligible compared to the linear term as the channel length becomes small, thus leaving the linear term as the only meaningful component for scaled-down devices. The present analysis shows that a situation may arise, particularly for scaled-down, wide-channel MOS integrated circuits, where a further scaling of the gate oxide is both undesirable and deleterious to the circuit performance.

Since the supply voltage in integrated circuits cannot be scaled as other device parameters, reducing the thickness of the gate oxide increases the effective transverse electric field in the channel region. This increased transverse field is known to degrade the channel mobility [4]. Therefore, the partial derivative $\partial\mu/\partial X_o$, which has been assumed to be 0 in all previous steps, must be taken into account in deducing the expression for the optimum oxide thickness shown in eq (13) in order to include this X_o-dependence of the mobility. This will lead to the following expression for the modified optimum oxide thickness X_μ in place of eq (13):

$$X_\mu = fX_m, \tag{16}$$

where X_m is the optimum oxide thickness for $\beta = \partial\mu/\partial X_o = 0$, and the correction factor f is given by

$$f = \left[1 - \left(X_\mu + \frac{3\epsilon WL}{C_p}\right)\frac{\beta}{\mu}\right]^{-\frac{1}{2}}. \tag{17}$$

Since $\beta = \partial\mu/\partial X_o \geq 0$, one always gets $f \geq 1$. This implies that the effect of the mobility degradation is always to increase the optimum thickness of the gate oxide.

IV. CONCLUSION

The effect of scaling of the gate oxide thickness on the propagation delay has been analyzed. The result shows that a lower bound in the gate oxide thickness always exists which gives the maximum circuit speed. The existence of this lower bound is a direct result of transistor parasitics interacting with the circuit. The position of this lower bound depends on the series resistance, parasitic gate capacitance, the channel width, the channel length, the channel carrier mobility, the turn-on voltage, and the supply voltage. The value of the lower bound increases linearly with increasing channel width but does not depend on the channel length within the approximation regime used in the present analysis where higher-order effects such as short-channel effect have been ignored for simplicity. It is also shown in this work that inclusion of the mobility degradation resulting from the gate oxide scaling tends to increase the optimum oxide thickness toward values larger than those shown in Table II. Theoretically, inclusion of higher order effects such as short-channel effect is a

worthy effort. Experimentally, a test structure having delay chains can be fabricated with varying gate oxide thickness and the propagation delay can be measured.

ACKNOWLEDGMENT

The author would like to thank H. S. Bennett, D. L. Blackburn, H. C. Lin, and L. W. Linholm for their critical reading of the manuscript. He also acknowledges the comments on the manuscript made by M. H. White.

REFERENCES

1. M. H. White, Characterization of CMOS Devices for VLSI, *IEEE Trans. Electron Devices* ED-29, 578 (1982).
2. J. Milman, *Microelectronics* (McGraw-Hill, 1979), p. 161.
3. S. M. Sze, *Physics of Semiconductor Devices* (Wiley, 1969), pp. 519-520.
4. O. Leistiko, A. S. Grove, and C. T. Sah, Electron and Hole Mobilities in Inversion Layers on Thermally Oxidized Silicon Surfaces, *IEEE Trans. Electron Devices* ED-12, 248 (1965).

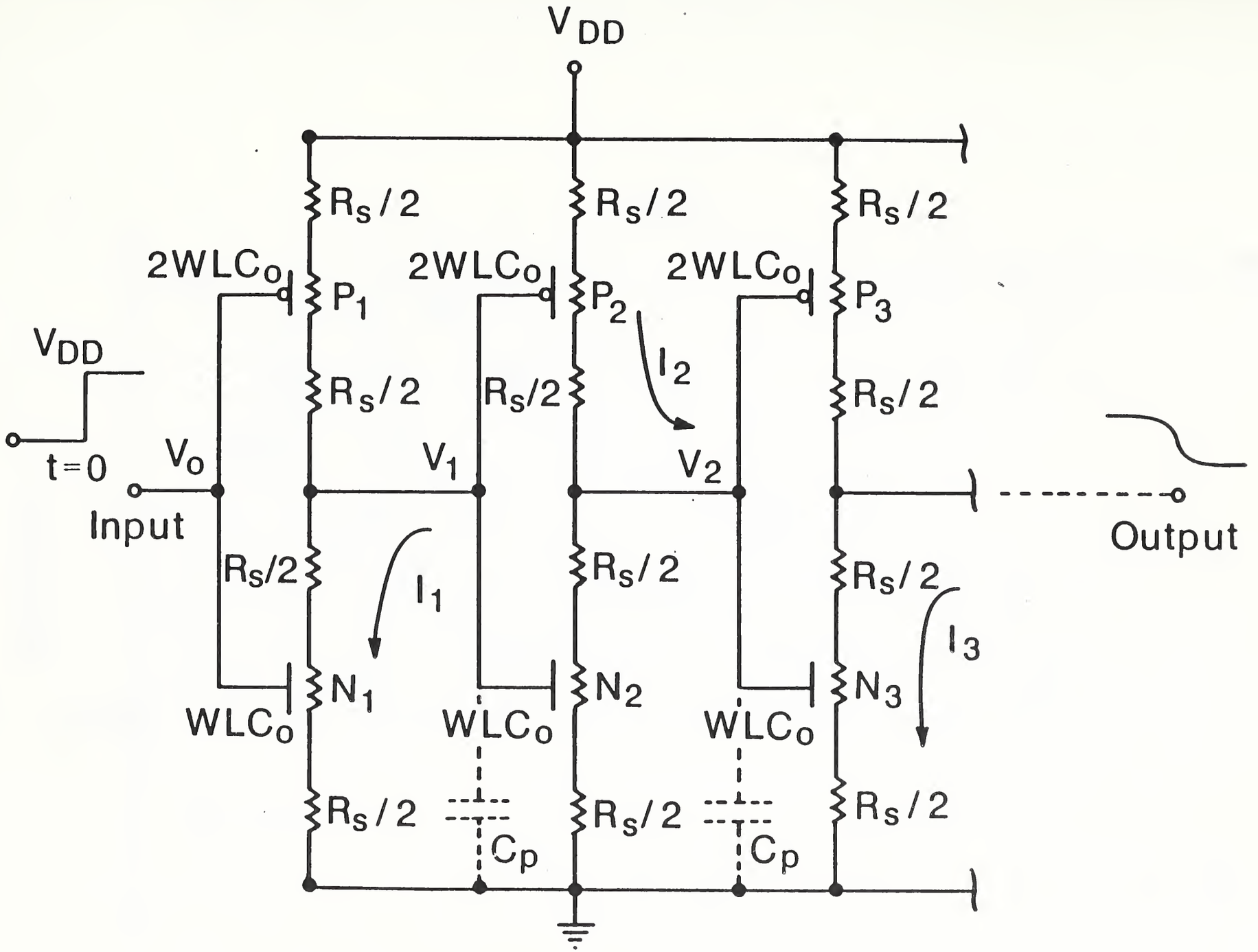

Figure 1. The equivalent circuit for the initial stages of the CMOS inverter delay chain.

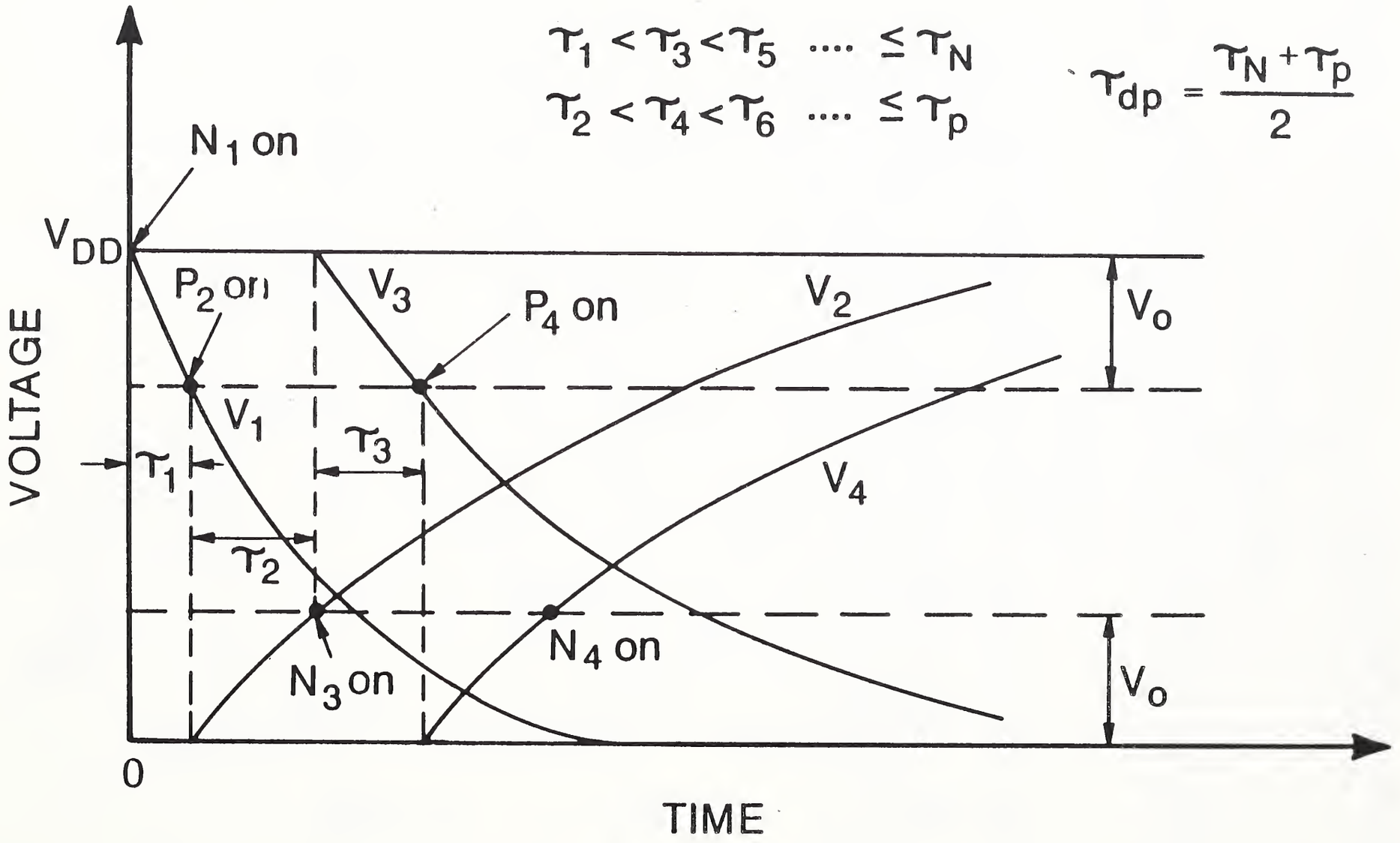

Figure 2. Evolution of a step input signal propagating along the delay chain.

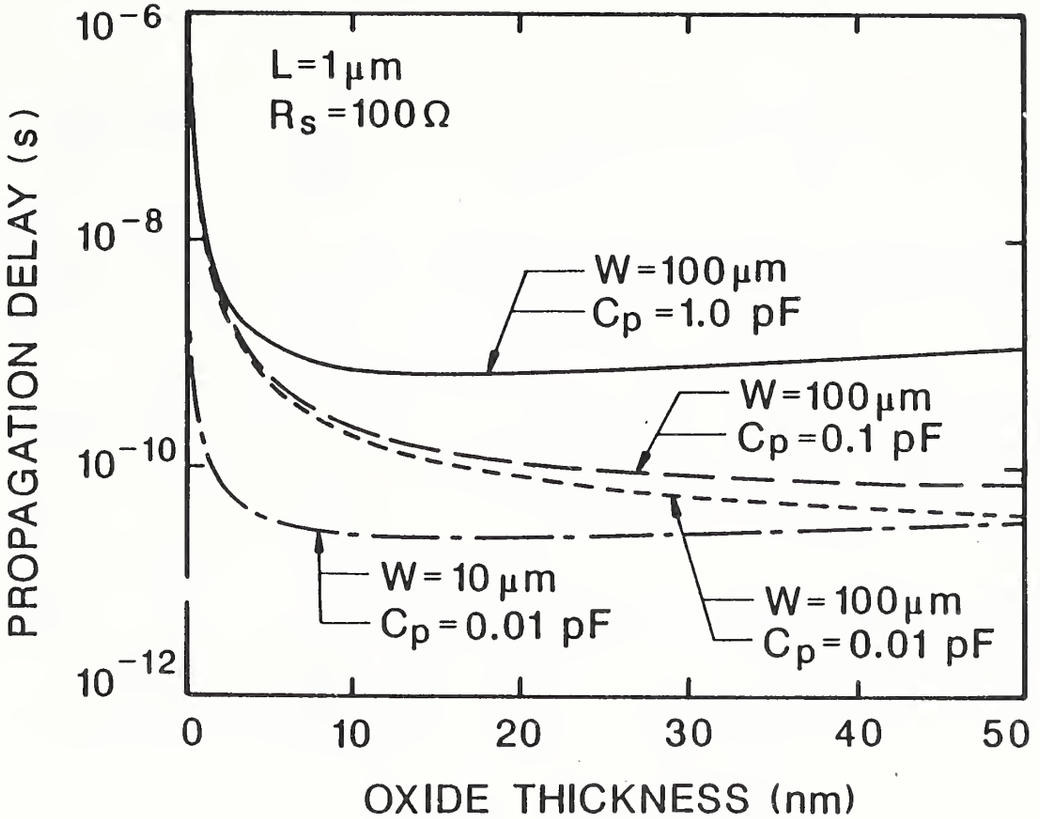

Figure 3. The propagation delay as a function of the oxide thickness for selected values of input parameters.

TABLE I. Status of Selected Transistors in the Initial Stage of the Inverter Circuit

Switching Transistor	V_G	V_D	V_S	I	τ
N_1	V_{DD}	V_1	0	$I_1=\frac{W}{L}\mu C_O(V_{DD}-V_T-\frac{V_1}{2})(V_1-I_1R_S)$	τ_1
P_2	$(V_1$ $V_{DD}-V_1$	V_2 $V_{DD}-V_2$	$V_{DD})$ 0	$I_2=\frac{W}{L}\mu C_O(V_{DD}-V_1-V_T-\frac{V_{DD}-V_2}{2})(V_{DD}-V_2-I_2R_S)$	τ_2
N_3	V_2	V_3	0	$I_3=\frac{W}{L}\mu C_O(V_{DD}-V_T-\frac{V_3}{2})(V_3-I_3R_S)$	τ_3
P_4	$(V_3$ $V_{DD}-V_3$	V_4 $V_{DD}-V_4$	$V_{DD})$ 0	$I_4=\frac{W}{L}\mu C_O(V_{DD}-V_3-V_T-\frac{V_{DD}-V_4}{2})(V_{DD}-V_4-I_4R_S)$	τ_4

TABLE II. Values of X_m (nm)

C_p(pF) \ $R_s(\Omega)$	0.1	1.0	10	100
0.001	1.67	5.27	16.7	52.7
0.01	0.5	1.67	5.27	16.7
0.1	--	0.5	1.67	5.27
1.0	--	--	0.5	1.67

W = 10 μm : Note that $X_m \propto W\sqrt{\frac{R_s}{C_p}}$

NBS-114A (REV. 2-8C)

U.S. DEPT. OF COMM. BIBLIOGRAPHIC DATA SHEET (See instructions)	1. PUBLICATION OR REPORT NO. NBSIR 87-3668	2. Performing Organ. Report No.	3. Publication Date December 1987

4. TITLE AND SUBTITLE

The Effect of the Gate Oxide Thickness on the Speed of MOS Integrated Circuits

5. AUTHOR(S)

J. S. Kim

6. PERFORMING ORGANIZATION (If joint or other than NBS, see instructions)

NATIONAL BUREAU OF STANDARDS
DEPARTMENT OF COMMERCE
WASHINGTON, D.C. 20234

7. Contract/Grant No.

8. Type of Report & Period Covered

9. SPONSORING ORGANIZATION NAME AND COMPLETE ADDRESS (Street, City, State, ZIP)

10. SUPPLEMENTARY NOTES

☐ Document describes a computer program; SF-185, FIPS Software Summary, is attached.

11. ABSTRACT (A 200-word or less factual summary of most significant information. If document includes a significant bibliography or literature survey, mention it here)

A simple analysis is presented for the effect of the gate oxide thickness on the circuit speed in a short-channel CMOS/inverter delay circuit. The present analysis is performed within the first-order theory of the MOS transistor. The result of the analysis shows that an optimum value of the gate oxide thickness exists, beyond which a further scaling of the gate oxide will not improve but degrade the circuit speed. The circuit speed corresponding to this optimum oxide thickness is the ultimate upper limit theoretically possible in a given MOS integrated circuit. The optimum value of the gate oxide thickness, to the first order approximation, is proportional to the channel width W, but it is independent of the channel length L. In particular, for wide channel devices, this optimum value exceeds the 5 nm - 30 nm range, which is of practical significance in the design and processing of advanced VLSI circuits. At the optimum oxide thickness, the square-root of the net propagation delay resolves into two components; the square-root of the net delay is the sum of the square-root of a purely parasitic component and that of a device-dependent component.

12. KEY WORDS (Six to twelve entries; alphabetical order; capitalize only proper names; and separate key words by semicolons)

device optimization; oxide scaling; propagation delay; sealing effect; thin gate oxide; VLSI

13. AVAILABILITY

☒ Unlimited

☐ For Official Distribution. Do Not Release to NTIS

☐ Order From Superintendent of Documents, U.S. Government Printing Office, Washington, D.C. 20402.

☐ Order From National Technical Information Service (NTIS), Springfield, VA. 22161

14. NO. OF PRINTED PAGES

18

15. Price

$9.95