Alpha-Power Law MOSFET Model and its Applications to CMOS Inverter Delay and Other Formulas

TAKAYASU SAKURAI, MEMBER, IEEE, AND A. RICHARD NEWTON, FELLOW, IEEE

Abstract —**A simple yet realistic MOS model, namely the α-power law MOS model, is introduced to include the carrier velocity saturation effect, which becomes eminent in short-channel MOSFET's. The model is an extension of Shockley's square-law MOS model in the saturation region. Since the model is simple, it can be applied for handling MOSFET circuits analytically and can predict the circuit behavior in the submicrometer region. Using the model, closed-form expressions are derived for the delay, the short-circuit power, and the transition voltage of CMOS inverters. The resultant delay expression includes input waveform slope effects and parasitic drain/source resistance effects and can be used in simulation and/or optimization CAD tools. It is concluded that the CMOS inverter delay becomes less sensitive to the input waveform slope and short-circuit dissipation increases as the carrier velocity saturation effects get severer in short-channel MOSFET's.**

I. Introduction

Conventionally, Shockley's MOSFET model [12] is widely used in treating MOSFET circuits analytically. Since the model is simple, many formulas have been derived based on the model and the derived formulas are used quite frequently in VLSI initial designs and CAD programs. However, the Shockley model cannot reproduce the voltage–current characteristics of the recent short-channel MOSFET's, mainly because it does not include the velocity saturation effects of carriers, which become eminent in the submicrometer regime. Consequently, the Shockley model is not satisfactory when applied to short-channel MOSFET circuits. In this paper, a new MOSFET model is proposed which is simple enough to be applied to the analytical treatments of the MOS circuits but includes the velocity saturation effects.

As applications of the model, closed-form analytical expressions are derived for the delay, short-circuit power, and logic threshold voltage of CMOS inverters. An expression for the CMOS inverter delay was first introduced by Burns [1] and Hedenstierna and Jeppson extended the work to include the input waveform slope effect [2]. Since both works were based on the Shockley model, more work is required to know the circuit behavior in the submicrometer region. Source and drain resistance is also considered in the delay expression. The source/drain resistance effect is important in estimating delay degradation by the contact resistance, the parasitic diffusion resistance of MOSFET's, and the hot-carrier degradation effect [3], [4], [12].

First, the necessity for a new short-channel MOS model is mentioned and a new model is introduced in Section II. Then in Section III, a delay formula for a CMOS inverter is derived using the proposed model. The effects of the input waveform and the power supply voltage V_{DD} on delay are also discussed in the section. The effect of the source and drain resistance on delay, the short-circuit power, and the logic threshold voltage are treated analytically using the model in Sections IV, V, and VI, respectively. Section VII is dedicated to conclusions.

II. Simple Short-Channel MOSFET Model

In the Shockley model, the drain current I_D is expressed as follows:

$$I_D = \begin{cases} 0 & (V_{GS} \leqslant V_{TH}\text{: cutoff region}) \\ K\{(V_{GS}-V_{TH})V_{DS} - 0.5V_{DS}^2\} & \\ & (V_{DS} < V_{D\mathrm{SAT}}\text{: linear region}) \\ 0.5K(V_{GS}-V_{TH})^2 & \\ & (V_{DS} \geqslant V_{D\mathrm{SAT}}\text{: saturation region}) \end{cases} \tag{1}$$

where $V_{D\mathrm{SAT}}(=V_{GS}-V_{TH})$ is drain saturation voltage and V_{TH} is threshold voltage. K is a drivability factor and equals $\mu(\epsilon_{ox}/t_{ox})(W/L_{\mathrm{eff}})$, where μ denotes an effective mobility, ϵ_{ox} a dielectric constant of a gate oxide, t_{ox} a gate oxide thickness, W a channel width, and L_{eff} an effective channel length. Fig. 1 shows a comparison between the Shockley model and the measured $V_{DS}-I_D$ characteristics for a 1-μm n-channel MOSFET. It is obvious that the Shockley model fails to reproduce the static char-

Manuscript received July 31, 1989; revised December 12, 1989.

T. Sakurai is with the Department of Electrical Engineering and Computer Sciences, University of California, Berkeley, CA 94720 on leave from the Semiconductor Device Engineering Laboratory, Toshiba Corporation, Kawasaki 210, Japan.

A. R. Newton is with the Department of Electrical Engineering and Computer Sciences, University of California, Berkeley, CA 94720.

IEEE Log Number 8934077.

0018-9200/90/0400-0584$01.00 ©1990 IEEE

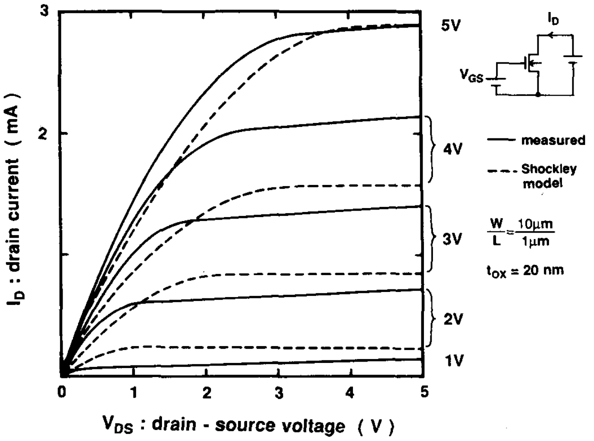


Fig. 1. Measured $V_{DS}-I_D$ characteristics and the Shockley model. The drain current at $V_{GS}=V_{DS}=5$ V and V_{TH} are fitted to obtain the parameters for the Shockley model.

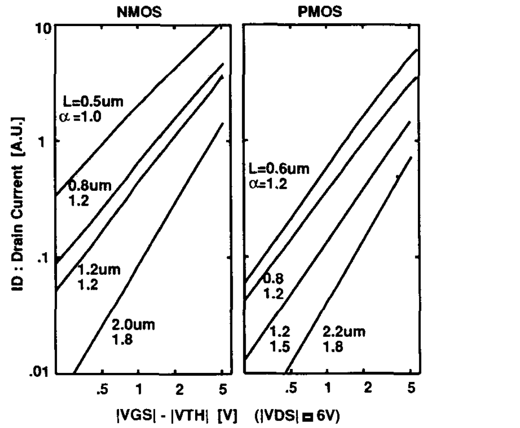


Fig. 3. Measured α's for various short-channel MOSFET's. The linearity shows the validity of the α-power law approximation of the saturation current.

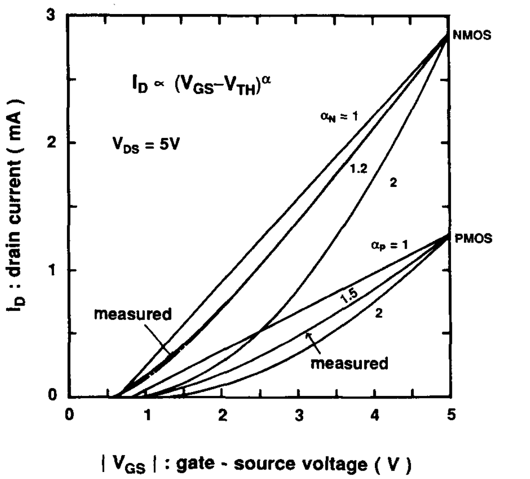


Fig. 2. Measured and model calculation of the $V_{GS}-I_D$ characteristics for short-channel MOSFET's.

acteristics of the recent MOSFET. There are two main discrepancies. One is that the drain saturation voltage $V_{D\text{SAT}}$ is different from the predicted value. The other is that the drain current in the saturation region (pentode region) does not show Shockley's square-law dependence on gate–source voltage. These two discrepancies, that is, the shift of the drain saturation voltage and the discrepancies in the saturation region $I-V$ curves, both come from the velocity saturation effects observed in short-channel MOSFET's.

In Fig. 2, the discrepancy in the saturation region will be shown more clearly. This figure shows $V_{GS}-I_D$ characteristics in the saturation region. As seen from Fig. 2, the drain current I_D is proportional to $(V_{GS}-V_{TH})^\alpha$. The Shockley model claims that $\alpha=2$, whereas the measured value of α for around 1-μm gate length is 1.2 for an n-channel MOSFET and 1.5 for a p-channel MOSFET [5]. These 1-μm MOSFET's are designed for use with 5-V supply voltage. Although it is found that the αth power law describes the measured data well for 1-μm MOSFET's, is this expression valid for general MOSFET's? Fig. 3 shows a log–log plot of I_D versus $V_{GS}-V_{TH}$ for various MOSFET's from 2-μm down to 0.5-μm gate length. They are made with various process technologies, so that the oxide thickness is different for each. For example, the data of a 2-μm MOSFET are taken from the 2-μm process, made several years ago, when the 2-μm design rule was the most advanced process. Here, the MOSFET's from 2.2-μm gate length down to 0.8-μm gate length were optimized for use under around 5-V supply voltage and the MOSFET's with 0.5–0.6-μm gate lengths were optimized for use under around 3.3-V V_{DD}.

Two important points should be mentioned here. The first is that the α-power approximation is generally very good since all of the curves are linear in the log–log plot. The second point is that the α changes from about 2 to 1 as the carrier velocity saturation gets severer. So, if some quantity, for example a delay, is expressed in terms of α, the behavior of that quantity in the short-channel region can be predicted just by changing the α from 2 to 1. Since the index α is closely related with the velocity saturation of carriers, α can be called a velocity saturation index.

Historically, from 2-μm rule down to 0.8-μm design rule, a constant voltage scaling paradigm has been adopted. As a result, the internal electric field increased as the feature size decreased. This forced α to be decreased from 2 to about 1 monotonically. Now in the further miniaturization, a constant electric field scaling might be adopted because of the hot-carrier-related problems. Then, α might not decrease so drastically and will remain essentially constant. However, α will not go back to the classical value of 2, because the technology tends to adopt the shortest gate length possible and consequently the internal electric field in a MOSFET will be kept quite high. This tendency can be assured by the 0.5-μm gate length MOSFET in Fig. 11, which is optimized for use under around 3.3-V V_{DD}; the MOSFET shows an α value of around 1.

A new MOSFET model, namely the α-power law model, is proposed. A full description of the proposed model in equation form is given below:

$$I_D = \begin{cases} 0 & (V_{GS} \leqslant V_{TH}\text{: cutoff region}) \\ (I'_{D0}/V'_{D0})V_{DS} & (V_{DS} < V'_{D0}\text{: triode region}) \\ I'_{D0} & (V_{DS} \geqslant V'_{D0}\text{: pentode region}) \end{cases} \qquad (2)$$

where

$$I'_{D0} = I_{D0}\left(\frac{V_{GS}-V_{TH}}{V_{DD}-V_{TH}}\right)^{\alpha} \left(=\frac{W}{L_{\text{eff}}} P_C (V_{GS}-V_{TH})^{\alpha}\right) \qquad (3)$$

$$V'_{D0} = V_{D0}\left(\frac{V_{GS}-V_{TH}}{V_{DD}-V_{TH}}\right)^{\alpha/2} \left(=P_V (V_{GS}-V_{TH})^{\alpha/2}\right) \qquad (4)$$

where V_{DD} signifies a supply voltage and P_C and P_V are parameters. Two sets of expressions are given in (3) and (4). Although the first expressions are used throughout this paper, there may be cases where the expressions in parenthesis are more suitable. In either set of expressions, the drain current in the saturation region is written in a single-term expression. This single-term nature seems important to facilitate the treatments of circuits analytically. Hereafter only the first set of expressions is described in this paper.

The 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 $V_{GS} = V_{DS} = V_{DD}$). I_{D0} is often used by VLSI designers as an index of MOSFET drivability. It should be noted that all four parameters are easily obtained from the measured data (see Appendix A for more details). When α is set to unity, that is, when the ultimately short channel is considered, V_{D0} becomes proportional to $(V_{GS} - V_{TH})^{1/2}$, which is the same feature as the model prediction of [11].

In Fig. 4, an example of this model is shown graphically for a 1-μm NMOSFET. Better agreement is observed in the pentode region than the Shockley model. The drain saturation voltage V_{D0} is treated as a parameter, because the Shockley model fails to predict the value, as was mentioned before. The linear region is approximated by linear lines. This approximation is suitable in investigating the parasitic resistance effects mentioned in Section IV. Although refinement is preferable in linear region modeling [6], most of the formulas and conclusions of this paper depend little on the linear region modeling.

It should be noted that when a MOSFET is scaled, not only α but also the drive current I_{D0} changes. In the following sections, most of the quantities are normalized with I_{D0}, and most of the discussions are independent of I_{D0}. That is, much stress is put on the relationship between the carrier velocity saturation in the short-channel MOSFET and the circuit behaviors. However, to obtain the real value of the delay, for example, the value of the drive current I_{D0} should be taken into account.

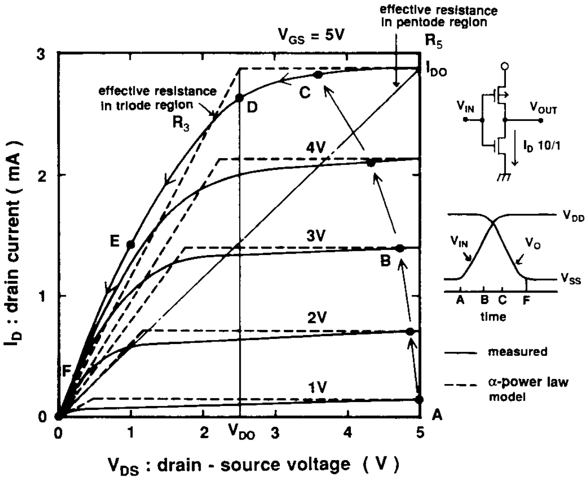


Fig. 4. Proposed α-power law MOS model. The solid lines are the measured $I-V$ curves and the broken lines are the present model calculation. The $A-B-C-D-E-F$ trajectory corresponds to the inverter operation in Fig. 5.

For investigating such circuits whose operation is mainly determined by the small-signal behavior of the triode region of the MOSFET, this model is inadequate. It is also to be noted that the model does not reproduce the characteristics of the region near and below the threshold voltage well. Near- and subthreshold region modeling is not important in calculating the delay of most VLSI's. The modeling of the region is important in estimating the charge decay characteristic of charge storage nodes, but in this case a statistical model should be used since it is very sensitive to process variation. If the main interest of one's analysis is in these regions, this model should not be used.

III. Input Waveform Slope and Delay

By using the α-power law model, an expression for delay is derived for a CMOS inverter. First, consider the case of discharging the output capacitance with NMOS as shown in Fig. 5, where the input voltage is varied linearly in transient time of t_T. In this case, the effect of PMOS can be neglected as is pointed out in [2]. This neglect of PMOSFET is not valid when the input ramp is very slow compared with the output waveform. The approximation is considered to be valid if the input slope exceeds one-third of the output slope [2], which is usually true in VLSI's.

Since the trajectory of the inverter operation on the I_D-V_{DS} plane is like the path $A-B-C-D-E-F$ in Fig. 4, this part of the characteristics should be modeled correctly in order to model the inverter delay well. Fig. 6 shows a comparison of the waveforms calculated using the SPICE MOS level 3 model [7], the α-power law model, and the Shockley model. The better agreement is seen between the SPICE calculation and the proposed model calculation. In all calculations, I_{D0} is matched to the measured value at $V_{GS} = V_{DS} = 5$ V.

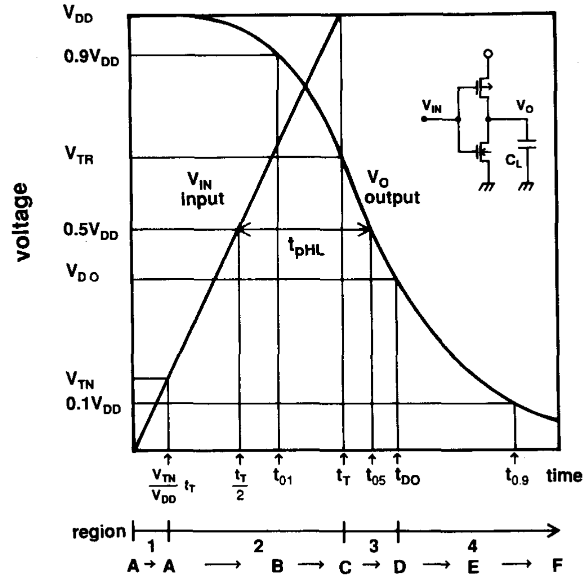


Fig. 5. Discharging waveform and notation. The notations are extensively used in Appendix B.

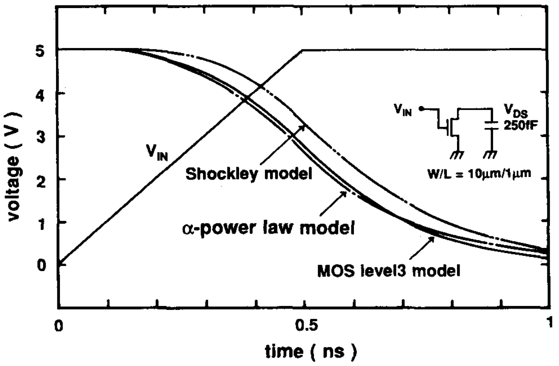


Fig. 6. Comparison of discharging waveforms calculated with the Shockley model, the present model, and SPICE MOS level 3 model.

The time from a half-V_{DD} point of the input to a half-V_{DD} point of the output is defined as a delay, t_{pHL}, in this discharging case. In the charging-up case, the delay t_{pLH} is defined in the same way. It is possible to calculate the delay through an inverter tree by simply adding t_{pLH} and t_{pHL}.

After the conventional manipulation of differential equations, the delay t_{pHL} and t_{pLH} can be expressed as follows (see Appendix B for the detailed derivation):

$$t_{pHL}, t_{pLH} = \left(\frac{1}{2} - \frac{1-\nu_T}{1+\alpha}\right)t_T + \frac{C_L V_{DD}}{2I_{D0}}, \quad \nu_T = \frac{V_{TH}}{V_{DD}} \tag{5}$$

where C_L is the output capacitance of a CMOS inverter. It is to be noted that the delay 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 t_T, and the second term is the output capacitance dependent term, which is proportional to the output capacitance C_L. This expression is independent of the linear region model of the MOSFET when V_{D0} is less than 0.6 V_{DD}, which is normally observed in submicrometer MOSFET's. For a typical short-channel MOSFET case, ν_T and α can be assumed to be 0.2 and 1, respectively. In this case, the above formula becomes

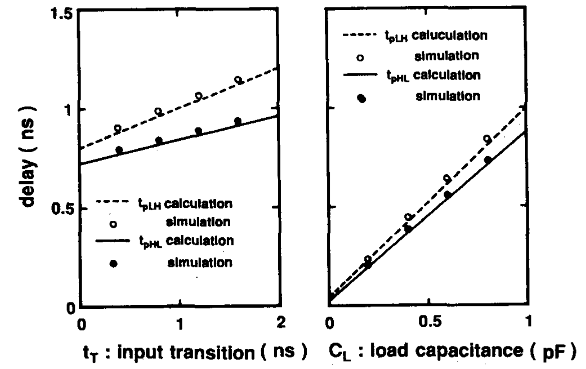


Fig. 7. Calculated and simulated value for t_{pLH} and t_{pHL}. It should be noted that t_{pHL}, which is determined by NMOS, is less sensitive to t_T than t_{pLH}, which is determined by PMOS. α for NMOS (t_{pHL}) is 1.2 and α for PMOS (t_{pLH}) is 1.5.

$$t_{pHL} = 0.1t_T + 0.5\frac{C_L V_{DD}}{I_{D0}}. \tag{6}$$

The first term signifies the input slope contribution to the delay and the second term is the time required to discharge the output capacitance to a half-V_{DD} level by the constant current I_{D0}.

A comparison is made in Fig. 7 between a SPICE simulation and a calculation with the above formula. It is interesting to note that the delay becomes less sensitive to the input slope when the carrier velocity saturation effect gets severer and α becomes smaller, because the factor $(1/2-(1-\nu_T)/(1+\alpha))$ decreases monotonously as α decreases. This tendency is seen also in Fig. 7, where t_{pHL}, which is determined by NMOS, is less sensitive to t_T than is t_{pLH}, which is determined by PMOS. α for NMOS (t_{pHL}) is 1.2 and α for PMOS (t_{pLH}) is 1.5.

This phenomenon is easily understood if the following two extreme cases are considered. Suppose that α is equal to zero. Then, very small V_{GS} can turn on the MOSFET completely and the drain current reaches the maximum value very quickly; hence the input slope does not affect the delay, even though it is slow. That is, the delay does not depend on the input transition time. On the other hand, if α is large, the small V_{GS} is not enough to turn on the MOSFET completely and only a small amount of drain current flows through the MOSFET. So in this case, it takes time for the MOSFET to charge or discharge the output capacitance when the input is slowly varying. That is, the delay depends much on the input transition time.

The next step is to approximate the real input waveform by a ramped waveform to obtain effective t_T. As seen from Fig. 8, a good approximation is achieved by connecting

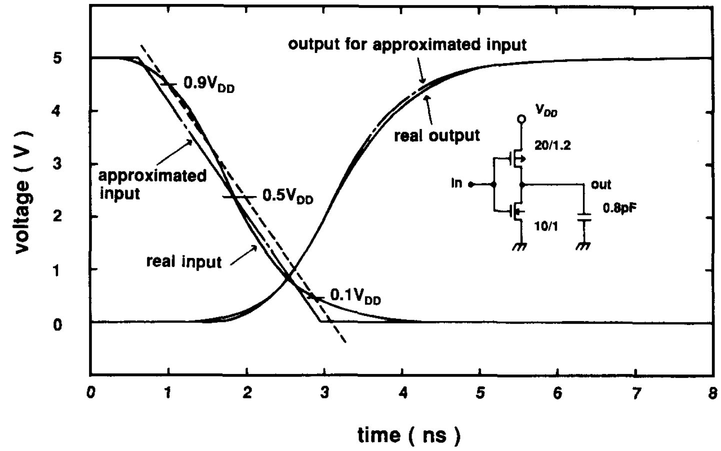


Fig. 8. Approximation of the input waveform by a ramp waveform. A good approximation of a slope is achieved by connecting $0.1V_{DD}$ point and $0.9V_{DD}$ point (broken line).

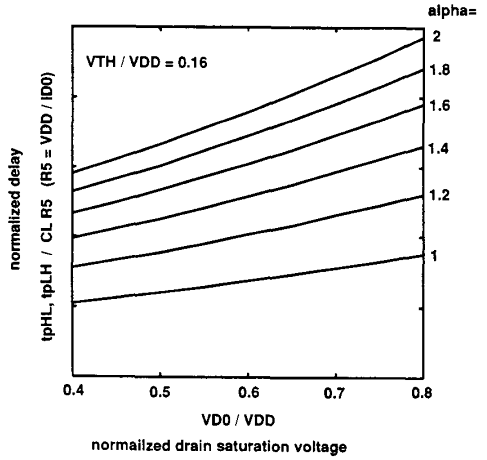
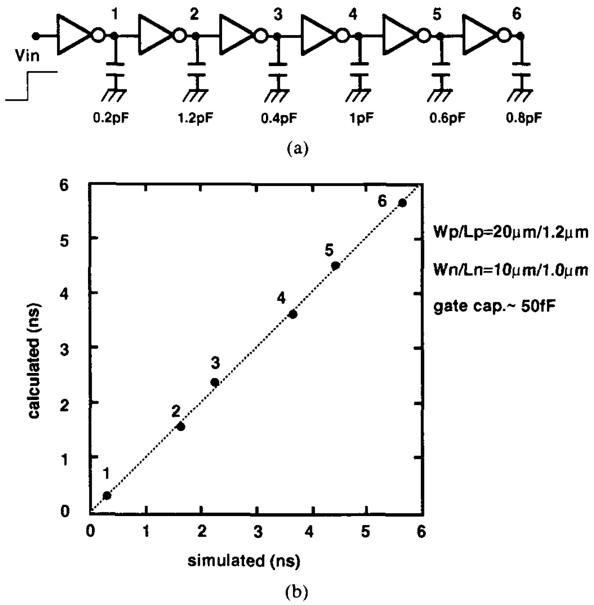


Fig. 9. Calculated value for t_{pLH} and t_{pHL} using (7).



(a)

(b)

Fig. 10. (a) Example of a CMOS inverter chain and (b) the calculated and the simulated delay for the circuit. At time zero, a step-up function from 0 to 5 V is applied to V_{in}. Delay is either t_{pLH} or t_{pHL} defined in Fig. 5. V_{DD} is 5 V, I_{D0N} is 2.87 mA, I_{D0P} is 2.54 mA, V_{TN} is 0.6 V, V_{TP} is 0.8 V, and α's for NMOS and PMOS are 1.2 and 1.5, respectively. V_{D0}'s for NMOS and PMOS are both 0.55.

$0.1V_{DD}$ point and $0.9V_{DD}$ point, when the input slope is similar to the output slope, which is often the case in real VLSI's. Using the approximation, t_T is expressed as follows:

$$t_T = \frac{t_{0.9} - t_{0.1}}{0.8} = \frac{C_L V_{DD}}{I_{D0}} \left(\frac{0.9}{0.8} + \frac{V_{D0}}{0.8V_{DD}} \ln \frac{10V_{D0}}{eV_{DD}} \right). \quad (7)$$

The normalized delay calculated by substituting (7) into (5) is plotted in Fig. 9. Fig. 9 is effective for the case where the input transition time is similar to the output transition time, which is often observed in VLSI's. If, for example, V_{D0}/V_{DD} is set equal to 0.5 and α is set to 1, the delay of (5) is simplified as $\sim 0.6C_L R_5$ using Fig. 9. $R_5(\equiv V_{DD}/I_{D0})$ is an effective pentode resistance of MOSFET as shown in Fig. 4.

A delay estimation is carried out for the inverter chain of Fig. 10(a), together with the result in Fig. 10(b). The agreement between the simulation and the calculation is good.

Using the above formulas, the V_{DD} dependence of the delay is calculated for various values of α and the results are shown in Fig. 11. The horizontal axis is a percent change in V_{DD} and the vertical axis is a percent change in delay. Suppose $I_{D0,\text{REF}}$ and $V_{D0,\text{REF}}$ are the values measured at the reference supply voltage of $V_{DD,\text{REF}}$. In order to calculate V_{D0} and I_{D0} at a general supply voltage V_{DD},

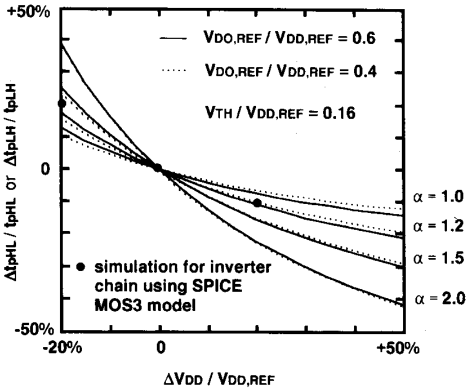


Fig. 11. Calculated delay dependence on a supply voltage V_{DD}. The horizontal axis is a percent change in V_{DD} and the vertical axis is a percent change in delay. At the reference supply voltage $V_{DD,\mathrm{REF}}$, it is assumed that V_{D0} is $V_{D0,\mathrm{REF}}$. V_{TH} is assumed to be kept constant at $0.16V_{DD,\mathrm{REF}}$. It should be noted that the delay becomes less sensitive to the change of V_{DD} as α gets smaller.

the following formulas were used, which are directly derivable from (3) and (4):

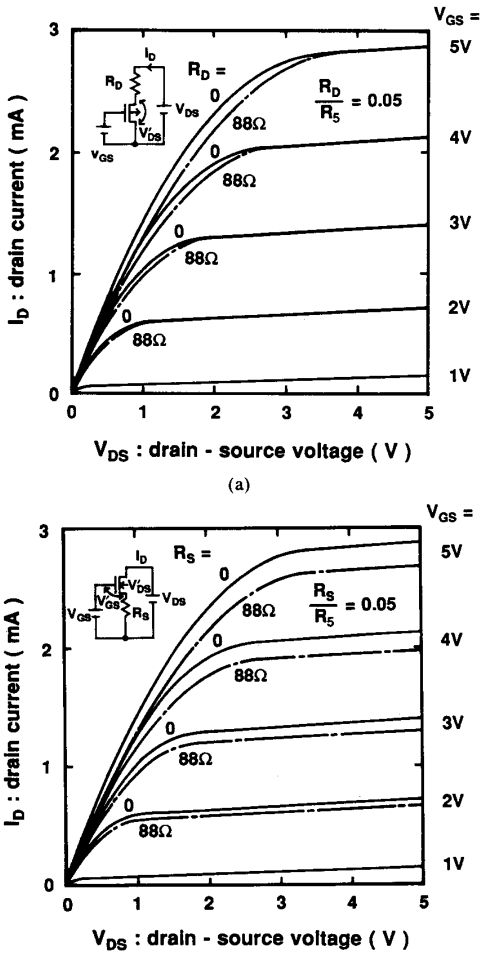
$$I_{D0}=\left(\frac{V_{DD}-V_{TH}}{V_{DD,\mathrm{REF}}-V_{TH}}\right)^{\alpha} I_{D0,\mathrm{REF}}$$

$$V_{D0}=\left(\frac{V_{DD}-V_{TH}}{V_{DD,\mathrm{REF}}-V_{TH}}\right)^{\alpha/2} V_{D0,\mathrm{REF}}.$$

The SPICE simulation for 1-μm MOSFET's differs from the Shockley model calculation, where α is set to 2. The delay variation shows strong dependence on α. It is interesting to note that the delay becomes less sensitive to the change of V_{DD} as α gets smaller. That is, with short-channel MOSFET's, delay shows a weaker dependence on power supply voltage than the classic Shockley MOSFET.

IV. Effect of Source and Drain Resistance on Delay

In the submicrometer MOSFET, a contact resistance and a diffusion resistance give rise to parasitic source and drain resistance. Hot-carrier degradation is another cause of parasitic drain resistance [3], [4], [12]. In this sense, it is important to know what happens to MOS circuits if a resistance is inserted in series with a MOSFET. For example, Fig. 12(a) shows the static characteristics of MOSFET's with and without a drain resistance R_D and Fig. 12(b) shows the counterpart of a source resistance R_S. The drain resistance only affects the linear region characteristics while the source resistance affects both the linear and the saturation region characteristics. Fig. 13 shows switching waveforms with drain and source resistance.



(a)

(b)

Fig. 12. Simulated static characteristics of the MOSFET with and without (a) a drain resistance R_D and (b) a source resistance R_S. The drain resistance only affects the linear region characteristics while the source resistance affects both the linear and the saturation region characteristics.

If the following substitutions are made for I_{D0} and V_{D0}, all the delay formulas described above are valid. A detailed derivation of these substitutions is given in Appendix C:

$$\frac{V_{D0}}{V_{DD}} \rightarrow \frac{V_{D0}}{V_{DD}}+\frac{R_D}{R_5}+\frac{R_S}{R_5} \tag{8}$$

$$I_{D0} \rightarrow I_{D0}\frac{1}{1+\dfrac{\alpha}{1-v_T}\cdot\dfrac{R_S}{R_5}}, \quad v_T=\frac{V_{TH}}{V_{DD}} \tag{9}$$

where R_5 denotes an effective pentode resistance of a MOSFET whose graphical interpretation is depicted in Fig. 4. To show the validity of the above formulas, the simulated $I-V$ curve change by the inserted resistance is

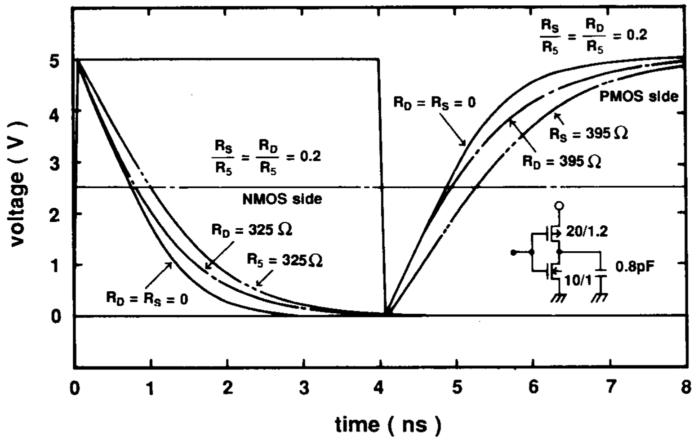


Fig. 13. Switching waveforms with and without the drain and source resistance. The solid line is the waveform without resistance, the dash–dot–dash line is that with a drain resistance, and the dash–dot–dot–dash line is that with a source resistance. The source resistance gives the more significant effect.

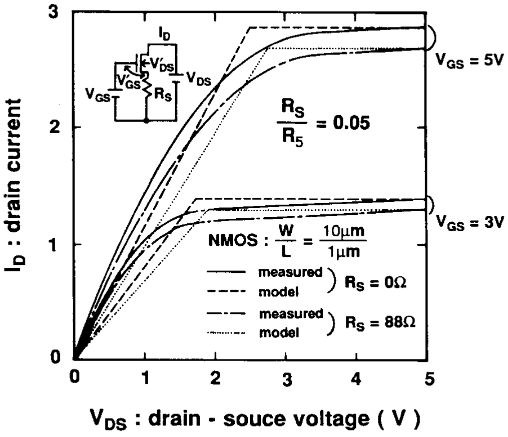


Fig. 14. Simulated $I-V$ curve change by a source resistance and the model calculation. The model calculation uses the substitutions (8) and (9).

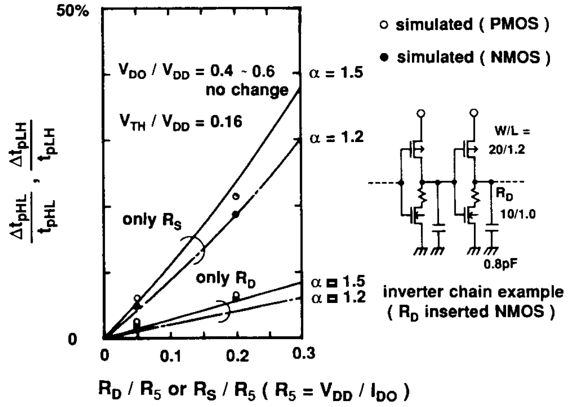


Fig. 15. Calculated and simulated delay with source/drain resistance.

shown in Fig. 14 together with the model calculation. The model calculation follows the change well.

Fig. 15 shows a delay comparison between the calculations by the formula and the SPICE simulation. R_S gives a stronger degradation. For around 1-μm MOSFET's, the delay degradation is approximated by the following formula, when the series resistance is small compared to the effective resistance of the MOSFET:

$$\frac{\Delta t_{pHL}}{t_{pHL}}, \frac{\Delta t_{pLH}}{t_{pLH}} \approx \frac{R_S}{R_5} + \frac{1}{3}\frac{R_D}{R_5} \approx \frac{1}{2}\frac{R_S}{R_3} + \frac{1}{6}\frac{R_D}{R_3} \tag{10}$$

where R_3 ($\equiv V_{D0}/I_{D0}$) denotes an effective triode resistance of MOSFET. It should be noted that if the resistance is inserted only in series with the NMOS and not with PMOS, then the inverter-chain delay degradation is about a half of the above formula. This is because half of the inverters in the inverter chain which charge up the output capacitance through the PMOSFET are not affected by the inserted resistance.

V. Short-Circuit Power in Static CMOS Circuit

In a CMOS inverter switching, when an input voltage is around the middle between V_{DD} and V_{SS}, there is a direct current path from V_{DD} to V_{SS}. The power consumed in this mode is called the short-circuit power in a static CMOS inverter. The formula for this short-circuit power is first given by Veendrick [8] based on the Shockley model, and used in some CAD tools such as VLSI power estimators. By replacing the Shockley model with the α-power law MOS model, and using the same assumptions as Veendrick, the short-circuit power per switching, P_S, is expressed as follows (the current expression in the saturation

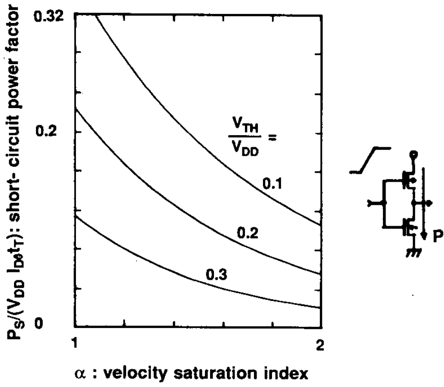


Fig. 16. Short-circuit power of a CMOS inverter per switching. The short-circuit power increases as α decreases to one, if the drivability of the MOSFET, I_{D0}, is kept constant. This means that even if MOSFET's of the same drivability are compared, the short-circuit power increases as a MOSFET gets smaller and the velocity saturation gets severer.

region should be used):

$$P_S = 2 \cdot V_{DD} \cdot 2 \int_0^{t_T/2} I_D \left(V_{GS} = V_{DD} \frac{t}{t_T} \right) dt \tag{11}$$

$$= V_{DD} t_T I_{D0} \frac{1}{\alpha+1} \frac{1}{2^{\alpha-1}} \frac{(1-2v_T)^{\alpha+1}}{(1-v_T)^{\alpha}},$$

$$v_T = \frac{V_{TH}}{V_{DD}}. \tag{12}$$

The first factor 2 in (11) comes from the fact that the short-circuit current flows twice per one switching. The second factor 2 and the integration over 0 to $t_T/2$ is due to Veendrick's approximation that the current waveform is mirror symmetric with $t = t_T/2$ as a symmetric axis. Formula (12) coincides with Veendrick's formula if α is set to 2. Therefore, the formula can be said to be a direct extension of Veendrick's formula. This formula is independent of the linear region model.

A plot of the formula is shown in Fig. 16. The short-circuit power increases as α decreases to one, if the drivability of the MOSFET, I_{D0}, is kept constant. This means that even if MOSFET's of the same drivability are compared, the short-circuit power increases as the carrier velocity saturation gets severer in short-channel MOSFET's. This is understandable because when α gets smaller, the drain current with V_{GS} around $V_{DD}/2$ is larger compared with the larger α case, when I_{D0} is kept equal.

VI. Logic Threshold Voltage

The logic threshold voltage or the inverting voltage of a CMOS inverter is another important quantity [9]. It is, for example, used in designing interface circuits where the threshold voltage as a gate is of interest. The logic threshold voltage of an inverter, V_{INV}, is defined as the input and output voltage when they are equal. The formula for the logic threshold voltage is derived by using the α law MOS model. The formula can be derived by equating the PMOS drain current and the NMOS drain current, when both are in the saturation region. For simplicity it is assumed that the threshold voltages of PMOS and NMOS are equal to V_{TH} and the PMOS and the NMOS have the same α:

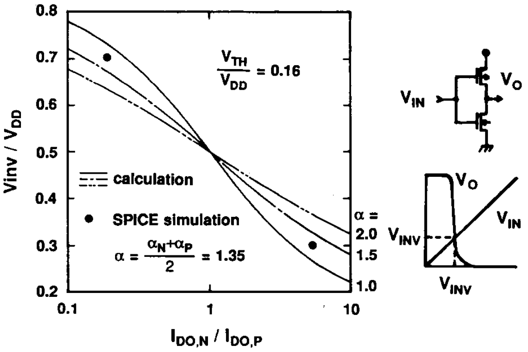


Fig. 17. Logic threshold voltage of a CMOS inverter. It is seen that as α becomes small, the logic threshold voltage becomes more sensitive to the gate width ratio of PMOS and NMOS, that is, I_{DOP}/I_{DON}.

$$V_{INV} = \frac{I_{DON}^{1/\alpha} \cdot V_{TH} + I_{DOP}^{1/\alpha}(V_{DD} - V_{TH})}{I_{DON}^{1/\alpha} + I_{DOP}^{1/\alpha}} \tag{13}$$

where I_{DOP} and I_{DON} stand for the I_{D0} of PMOS and NMOS, respectively. If the velocity saturation index for the PMOS (α_P) and that for the NMOS (α_N) are different but similar, an approximation of $\alpha = (\alpha_P + \alpha_N)/2$ turns out to be good. The result is graphically shown in Fig. 17. As seen from the figure, the simulated logic threshold voltages differ from the predicted value by the Shockley model, that is, α being equal to 2. It is seen that as α becomes small, the logic threshold voltage becomes more sensitive to the gate width ratio of PMOS and NMOS, that is, I_{DOP}/I_{DON}. This result is not dependent on the triode model.

VII. Conclusions

A new MOS model is introduced to overcome the shortcomings of the Shockley MOS model in the submicrometer region. The new model can express the salient features of the short-channel MOSFET $I-V$ characteristics. The model is simple and suitable for circuit analysis.

Useful expressions are derived for the delay, short-circuit power, and logic threshold voltage with the new MOS model. It has been shown that with the short-channel MOSFET's the CMOS inverter delay becomes less sensitive to the input waveform slope and to the V_{DD} variation than with the classic MOSFET's whose I_D shows square-law dependence on V_{GS}. In addition, short-circuit dissipation increases, and transition voltage becomes more sensitive to the gate width ratio of PMOS and NMOS.

Further extension is preferable for the triode region modeling to increase precision [4], although the results obtained here would remain essentially unchanged. None of the derived formulas, except (7) and (8), depends on the triode model. Since the proposed model efficiently models a short-channel MOSFET, it can be used to modify the classical expressions based on the Shockley model. One interesting application is on a CMOS arbiter/synchronizer optimization [10]. In order to make a bridge between the device engineering and the circuit behavior, it is an interesting direction to explore to express α in terms of physical parameters such as device dimensions and doping profiles.

Appendix A
Extraction of Model Parameters

In this appendix, two methods are described to extract the model parameters V_{TH} and α. The first method uses brute force. First, guess a plausible V_{TH}. The guess is not so difficult if there is a $V_{GS}-I_D$ plot. The drain current I_D should be measured in the saturation region. Then, write $\log(V_{GS})-\log(I_D)$ plot. If the curve is linear, the slope is α. If the curve is not linear, modify V_{TH} a little and try the log–log plot again. Repeat the process until the log–log plot gets satisfactorily linear.

The second method involves equation solving, but the equation has only one variable. First, from the measured $V_{GS}-I_D$ plot, pick three points that are to be fitted. Suppose the three points are (V_{G1}, I_{D1}), (V_{G2}, I_{D2}), and (V_{G3}, I_{D3}). V_{TH} can be obtained by solving the following equation:

$$f(V_{TH}) = \log\left(\frac{I_{D1}}{I_{D2}}\right)\log\left(\frac{V_{G2}-V_{TH}}{V_{G3}-V_{TH}}\right) - \log\left(\frac{I_{D2}}{I_{D3}}\right)\log\left(\frac{V_{G1}-V_{TH}}{V_{G2}-V_{TH}}\right) = 0 \quad \text{(A1)}$$

The bisection method [14] is the best choice for solving this equation since it finds out the root without fail within ten iterations. Then, α can be obtained from the following expression:

$$\alpha = \frac{\log(I_{D1}/I_{D2})}{\log((V_{G1}-V_{TH})/(V_{G2}-V_{TH}))}. \quad \text{(A2)}$$

Appendix B
Derivation of the Delay Formula

The case of Fig. 5 is considered in the derivation. Before the input reaches V_{TN}, NMOS is off and the output voltage V_o remains V_{DD} (region 1 in Fig. 5). Then in the region 2, the input ramps up linearly and the NMOS is operated in the saturation region. The output voltage V_o is changed observing the following differential equation:

$$C_L\frac{dV_o}{dt} = -I'_{D0} = -I_{D0}\left(\frac{t/t_T - v_T}{1-v_T}\right)^{\alpha}, \quad v_T = \frac{V_{TH}}{V_{DD}}. \quad \text{(B1)}$$

The solution is

$$V_o = V_{DD} - \frac{I_{D0}t_T}{C_L}\cdot\frac{1}{1+\alpha}\cdot\frac{1}{(1-v_T)^{\alpha}}\left(\frac{t}{t_T}-v_T\right)^{\alpha+1}$$
$$(\text{region 2: } v_T t_T \leqslant t \leqslant t_T). \quad \text{(B2)}$$

In region 3, the input is fixed at V_{DD} and the n-channel MOSFET is operated in the saturation region. Consequently, the output capacitance C_L is discharged by a constant current I_{D0} and the output voltage V_o changes linearly. By connecting the solution at $t=t_T$ with the solution of (B2), we have

$$V_o = V_{DD} - \frac{I_{D0}}{C_L}\left(t - \frac{v_T+\alpha}{1+\alpha}t_T\right)$$
$$(\text{region 3: } t_T \leqslant t \leqslant t_{D0}). \quad \text{(B3)}$$

where t_{D0} is the time when V_o gets equal to V_{D0}, expressed as follows:

$$t_{D0} = \frac{C_L}{I_{D0}}(V_{DD}-V_{D0}) + \frac{v_T+\alpha}{1+\alpha}t_T. \quad \text{(B4)}$$

In the final region 4, the input is still fixed at V_{DD} but the operation mode of the NMOSFET goes into the linear region. As a result, the differential equation that governs the discharging process can be written as

$$C_L\frac{dV_o}{dt} = -\frac{I_{D0}}{V_{D0}}V_o \equiv -\frac{1}{R_3}V_o. \quad \text{(B5)}$$

The solution in this region has an exponential form and goes through the point (t_{D0}, V_{D0}):

$$V_o = V_{D0}e^{-1/C_L R_3 (t-t_{D0})} \quad (\text{region 4: } t_{D0} \leqslant t). \quad \text{(B6)}$$

Denoting t_{05} as the time when the output reaches a half-V_{DD} point, the delay t_{pHL} is calculated by using (B6) and (B4):

$$t_{pHL} = t_{05} - \frac{t_T}{2} = \left(\frac{v_T+\alpha}{1+\alpha} - \frac{1}{2}\right)t_T + \frac{C_L V_{DD}}{2I_{D0}}. \quad \text{(B7)}$$

This is the same formula as (5). For t_{pLH}, the expression is exactly the same but the values of V_{TH}, α, and I_{D0} for the p-channel MOSFET should be used.

Although t_{05} may fall in region 3, the value of t_{pHL} coincides with the above formula (B7) within 3% error when $V_{D0} \leqslant 0.6V_{DD}$. This condition is satisfied in normal short-channel MOSFET's. When the input is very slow, t_{05} falls in region 2, and in this case, the solution becomes very complicated. However, Hedenstierna and Jeppson [2] pointed out that the approximation that the input is sufficiently fast gives a good result when t_T is less than three times the transition time of the output, which is true in most cases.

Appendix C
Substitution Rules for a Resistance Inserted MOSFET

In this appendix, V_{DS} denotes an "apparent" drain–source voltage which is externally applied to the resistance inserted system terminals while V'_{DS} means a "true" drain–source voltage which is really applied to the MOSFET terminals. V_{GS} and V'_{GS} denote the gate–source counterparts of the above quantities.

First, let us consider a change in the triode region when R_S and R_D are inserted. As seen from the trajectory of an inverter operation (Fig. 4), the change in the $I-V$ curve at $V_{GS} = V_{DD}$ is important. So the drain current change in this region is mainly considered. In this case, I'_{D0} becomes I_{D0} and V'_{D0} becomes V_{D0} in (3) and (4), and the following equation holds:

$$I_D = \frac{I_{D0}}{V_{D0}} V'_{DS} = \frac{I_{D0}}{V_{D0}} (V_{DS} - R_D I_D - R_S I_D). \qquad \text{(C1)}$$

Solving in terms of I_D leads to

$$I_D = \frac{I_{D0}}{V_{D0} + R_D I_{D0} + R_S I_{D0}} V_{DS}. \qquad \text{(C2)}$$

This means that the substitution

$$V_{D0} \rightarrow V_{D0} + R_D I_{D0} + R_S I_{D0} \qquad \text{(C3)}$$

is effective. By dividing both sides by V_{DD} and using R_5 ($\equiv V_{DD}/I_{D0}$), the substitution rule of (8) results.

On the other hand, in the saturation region,

$$I_D = I_{D0} \left(\frac{V_{GS} - R_S I_D - V_{TH}}{V_{DD} - V_{TH}} \right)^{\alpha} \qquad \text{(C4)}$$

holds. Assuming that $R_S I_D \ll V_{GS} - V_{TH}$, and solving in terms of I_D, we have

$$I_D \approx I_{D0} \frac{1}{1 + \dfrac{\alpha}{V_{GS} - V_{TH}} \cdot I_{D0} R_S \left(\dfrac{V_{GS} - V_{TH}}{V_{DD} - V_{TH}} \right)^{\alpha}} \cdot \left(\frac{V_{GS} - V_{TH}}{V_{DD} - V_{TH}} \right)^{\alpha}. \qquad \text{(C5)}$$

The large I_D region where V_{GS} is near V_{DD} is important in estimating the delay. With this approximation, the following expression is derived:

$$I_D \approx I_{D0} \frac{1}{1 + \dfrac{\alpha}{V_{DD} - V_{TH}} \cdot I_{D0} R_S} \left(\frac{V_{GS} - V_{TH}}{V_{DD} - V_{TH}} \right)^{\alpha}. \qquad \text{(C6)}$$

Comparing this expression with (3), and introducing $R_5(\equiv V_{DD}/I_{D0})$, we have the substitution rule (9). It should be mentioned that the substitution is exact when α is 1, that is, in the case of a typical submicrometer MOSFET.

Acknowledgment

The encouragement throughout the course of the study by Y. Takeishi, Y. Unno, H. Yamada, and T. Iizuka of Toshiba Corporation is appreciated. Discussions with K. Kobayashi, H. Ishiuchi, K. Sawada, M. Kakumu, and K. Nogami were useful in developing the MOSFET model. The support of T. Shirotori and Y. Ito is also appreciated. It should be mentioned that the comments of the reviewers much improved the paper.

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, no. 2, pp. 270–280, Mar. 1987.

[3] K. Nogami, K. Sawada, M. Kinugawa, and T. Sakurai, "VLSI circuit reliability under AC hot-carrier stress," in *Proc. Symp. VLSI Circuits*, May 1987, pp. 13–14.

[4] T. Sakurai, K. Nogami, M. Kakumu, and T. Iizuka, "Hot-carrier generation in submicrometer VLSI environment," *IEEE J. Solid-State Circuits*, vol. SC-21, no. 1, pp. 187–192, Feb. 1986.

[5] T. Sakurai, "CMOS inverter delay and other formulas using α-power law MOS model," in *Proc. IEEE Int. Conf. Computer-Aided Design* (Santa Clara, CA), Nov. 1988, pp. 74–77.

[6] T. Sakurai and A. R. Newton, to be submitted to *IEEE J. Solid-State Circuits.*

[7] A. Vladimirescu and S. Liu, "The simulation of MOS integrated circuits using SPICE2," Univ. of Calif., Berkeley, ERL Memo. M80/7, Oct. 1980.

[8] H. Veendrick, "Short-circuit dissipation of static CMOS circuitry and its impact on the design of buffer circuits," *IEEE J. Solid-State Circuits*, vol. SC-19, pp. 468–473, 1984.

[9] J. Mavor, M. Jack, and P. Denyer, *Introduction to MOS LSI Design*. Reading, MA: Addison-Wesley, 1983.

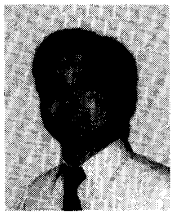
[10] T. Sakurai, "Optimization of CMOS arbiter and synchronizer with submicrometer MOSFET's," *IEEE J. Solid-State Circuits*, vol. 23, no. 4, pp. 901–906, Aug. 1988.

[11] R. S. Muller and T. I. Kamins, *Device Electronics for Integrated Circuits*, 2nd ed. New York: Wiley, 1986, p. 482.

[12] K. K. Ng and W. T. Lynch, "The impact of intrinsic series resistance on MOSFET scaling," *IEEE Trans. Electron Devices*, vol. ED-34, pp. 503–511, Mar. 1987.

[13] W. Shockley, "A unipolar field effect transistor," *Proc. IRE*, vol. 40, pp. 1365–1376, Nov. 1952.

[14] W. H. Press, B. P. Flannery, S. A. Teukolsky, and W. T. Vettering, *Numerical Recipes in C.* London: Cambridge Univ. Press, 1988, p. 261.



Takayasu Sakurai (S'77–M'78) was born in Tokyo, Japan, on January 10, 1954. He received the B.S., M.S., and Ph.D. degrees in electronic engineering from the University of Tokyo, Tokyo, Japan, in 1976, 1978, and 1981, respectively. His Ph.D. work is on electronic structures of a Si–SiO$_2$ interface.

In 1981 he joined the Semiconductor Device Engineering Laboratory, Toshiba Corporation, Kawasaki, Japan, where he was engaged in the research and development of CMOS dynamic RAM and 64-kb SRAM, 256-kb SRAM, 1-Mb virtual SRAM, cache memories, and a RISC with on-chip large cache memory. He also worked on the modeling of wiring capacitance and delay, a new soft-error free memory cell, new memory architectures, new hot-carrier resistant circuits, arbiter optimization, and gate-level delay modeling. Since 1988 he has been a Visiting Scholar at the University of California, Berkeley, doing research in the field of computer-aided design of VLSI's. His present interests are application-specific memories, VLSI processors, and CAD for VLSI's.

Dr. Sakurai is a member of the Institute of Electronics, Information and Communication Engineers of Japan and the Japan Society of Applied Physics.

A. Richard Newton (S'73–M'78–SM'86–F'88) was born in Melbourne, Australia, on July 1, 1951. He received the B.Eng. (elec.) and M.Eng. Sci. degrees from the University of Melbourne, Melbourne, Australia in 1973 and 1975, respectively, and the Ph.D. degree from the University of California, Berkeley, in 1978.

He is a Professor in the Department of Electrical Engineering and Computer Sciences at the University of California, Berkeley, and served as Vice Chair from 1984 to 1988. He has been actively involved as a researcher and teacher in the area of computer-aided design and computer architecture for 13 years. His special interests are synthesis (behavioral, logic, physical), design of integrated circuits, and multiprocessor implementation of algorithms. He has consulted for many companies in the area of computer-aided design for integrated circuit design, including Digital Equipment Corporation, General Electric, Hewlett-Packard, Intel, Synopsys, SDA Systems, Silicon Systems, Tektronix, and Xerox Corporation. In addition, he is a member of the Technical Advisory Boards of Sequent Computers, Candence Incorporated, and Objectivity. In addition, he supervises the research of over a dozen graduate students working in the area of computer-aided design for VLSI systems.

Dr. Newton is a Fellow of the IEEE and the Technical Program Chair of the 1988 and 1989 ACM/IEEE Design Automation Conferences. He was also an Associate Editor for IEEE Transactions on Computer-Aided Design of ICAS from 1985 to 1988 and a member of the Circuits and Systems Society ADCOM. He has received a number of awards, including Best Paper awards at the European Solid State Circuits Conference and 1987 ACM/IEEE Design Automation Conference, and he was selected in 1987 as the national recipient of the C. Holmes McDonald Outstanding Young Professor Award of the Eta-Kappa-Nu Engineering Honor Society.