Delay Analysis of Series-Connected MOSFET Circuits

Takayasu Sakurai, *Member, IEEE*, and A. Richard Newton, *Fellow, IEEE*

Abstract—**CMOS gate delay is analyzed using a new realistic short-channel MOS model. Closed-form delay formulas are obtained for CMOS inverters and series-connected MOSFET structures (SCMS's), which include short-channel effects. It is shown that the ratio of the delay of NAND/NOR to the delay of inverter becomes smaller in the submicrometer region. This is because the V_{DS} and V_{GS} of each MOSFET in the SCMS are smaller than those of an inverter MOSFET. The smaller voltages in turn mitigate and relax the severe carrier velocity saturation in miniaturized MOSFET's. This result encourages more extensive use of NAND/NOR/complex gates, cascode voltage switch logic [1], and hot-carrier resistant logic [2] in the submicrometer circuit design. The results of the analysis are informative for submicrometer VLSI designs. For example, if the maximum number of series-connected MOSFET's was considered to be five in 2-μm designs, then the number can be increased to six or seven in the submicrometer circuit design. In the typical cases in VLSI designs, the delay ratio for N-SCMS is much less than N^2. The delay dependence on input terminal position for SCMS structures is also described.**

I. Introduction

A SERIES-connected MOSFET structure (SCMS) appears in NAND/NOR gates, more complex gates, and PLA's and is widely used in VLSI designs. However, little has been known about the behavior of the SCMS because of its relatively complicated nature and the nonlinearity of MOSFET's. The main purpose of this paper is to shed light on the behavior of the SCMS. It is shown that the ratio (delay of NAND/NOR)/(delay of inverter) becomes smaller in the submicrometer region. There are cases where N series-connected MOSFET's show only $N/2$ times as long a delay as a single MOSFET.

In order to derive analytical delay expressions for CMOS gates in the submicrometer region, a realistic yet simple MOS model is required. Analytical delay models are sometimes better than circuit simulations, since they give insight into the delay dependence on parameters and they are faster in calculation. For the realistic yet simple MOS model, an nth power law MOS model [3] is used, which is briefly described in Section II. In Section III, delay expressions suitable for analyzing the SCMS are derived and applied to a logic circuit. Section IV describes the delay ratio of SCMS to that of an inverter for a simple case. The more complex cases are presented in Sections V and VI. In Section VII, delay dependence on input terminal position is described. The final section is dedicated to conclusions.

Manuscript received April 11, 1990; revised October 18, 1990. This work was supported by a grant from Toshiba Corporation.

T. Sakurai was with the Department of Electrical Engineering and Computer Sciences, University of California, Berkeley, CA 97420 on leave from the Semiconductor Device Engineering Laboratory, Toshiba Corporation, Komukai-Toshiba-cho 1, 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 9041244.

II. A Short-Channel MOS Model

The following nth power law MOS model [3] is used in this paper as a short-channel MOS model:

$$V_{TH} = V_{T0} - \gamma_1 V_{BS} \tag{1}$$

$$V_{DSAT} = K(V_{GS} - V_{TH})^m \tag{2}$$

$$I_D = I_{DSAT} = \frac{W_{EFF}}{L_{EFF}} B(V_{GS} - V_{TH})^n \cdot (V_{DS} \geq V_{DSAT}\text{: saturated region}) \tag{3}$$

$$I_D = I_{DSAT}\left(2 - \frac{V_{DS}}{V_{DSAT}}\right)\frac{V_{DS}}{V_{DSAT}} (V_{DS} < V_{DSAT}\text{: linear region}) \tag{4}$$

$$I_D = 0 (V_{GS} < V_{TH}\text{: cutoff region}). \tag{5}$$

In these model equations, W_{EFF} and L_{EFF} are effective channel width and effective channel length, respectively. V_{T0} stands for zero back-gate bias threshold voltage and I_D is the drain current. n, m, K, and B are constants which describe the short-channel effects in an empirical manner. Other notations are as usual. If n is set equal to 2, m to one, K to one, and B to $1/2\beta$, then the model equations are reduced to the Schockley model equations except for the body effect.

This simple model can reproduce the measured characteristics even in the short-channel region, as shown in Fig. 1. The body effect is approximated by a linear form, the meaning of which is illustrated in Fig. 2. The back-gate bias is normally less than 2.5 V in analyzing the SCMS.

In Fig. 1, I_{D0} is defined as the drain current observed when $V_{GS} = V_{DS} = V_{DD}$ and is a good index of the drivability of a MOSFET. V_{D0} is defined as the drain saturation voltage when $V_{GS} = V_{DD}$. These two quantities together with the velocity saturation index n play an essential role in determining circuit behavior.

III. Delay Expressions for CMOS Gates

Using the above model, delay formulas for a CMOS inverter can now be derived. These formulas are also effective in analyzing the SCMS because the MOS model is general enough to express the equivalent I–V characteristics of the SCMS as shown in Fig. 3.

The derivation begins by setting up the differential equation which governs gate operation. This equation is then

0018-9200/91/0200-0122$01.00

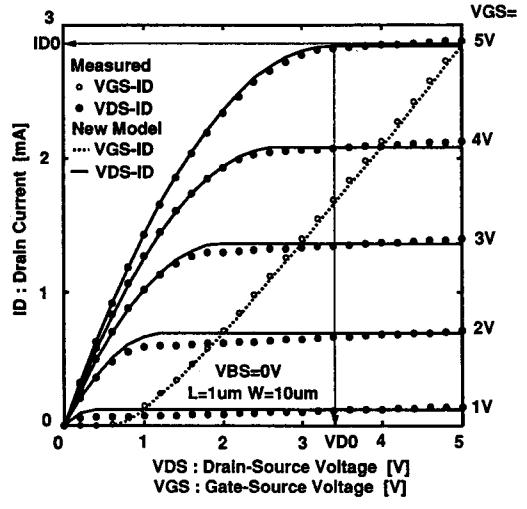


Fig. 1. NMOS $I-V$ curves with the new MOS model.

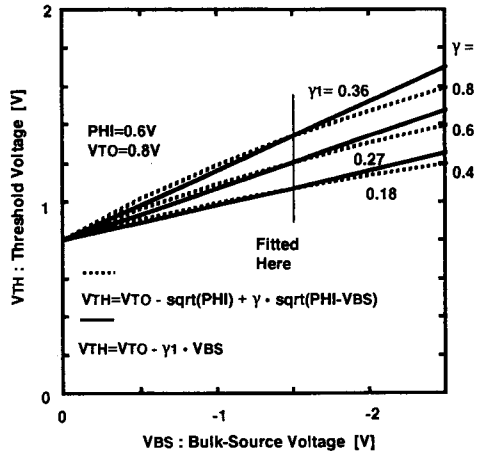


Fig. 2. Linear approximation of body effect.

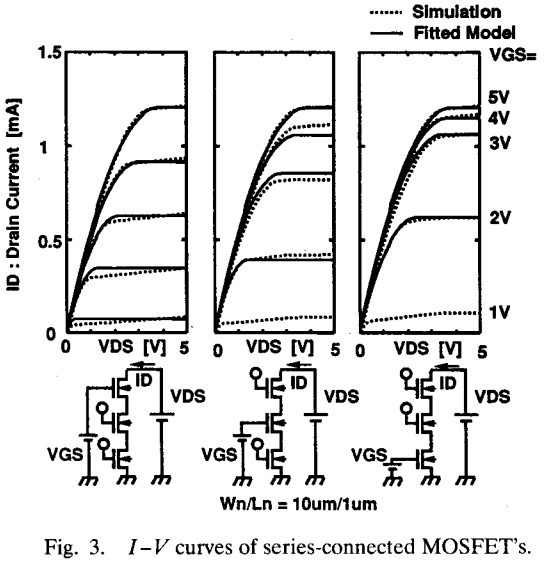


Fig. 3. $I-V$ curves of series-connected MOSFET's.

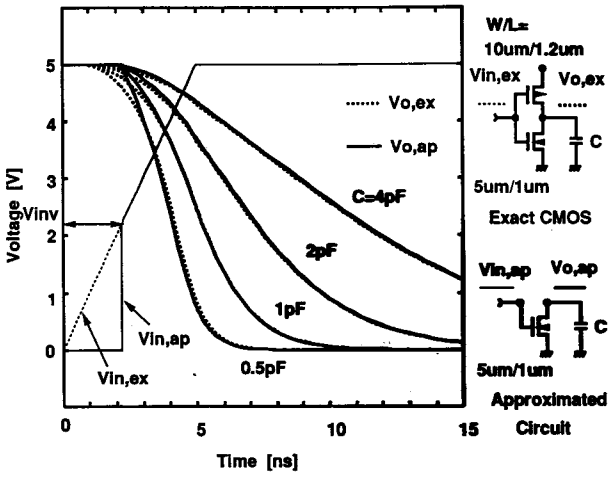


Fig. 4. Approximating CMOS by NMOS.

solved for the very fast input case and for the very slow input case, and the two solutions are connected smoothly. The ramp input waveform can be approximated by $V_{\text{in, ap}}$ which stays at 0 V until the ramp input goes across the logic threshold V_{INV}, abruptly rises up at that point, and coincides with the ramp input waveform therafter as shown in Fig. 4. The logic threshold voltage is the gate input voltage, which makes the output voltage equal to a half V_{DD}. The detailed derivation of the delay expressions can be found in Appendix A. First, define a critical input transition time t_{T0}:

$$t_{T0} = \frac{C_O V_{DD}}{2I_{D0}} \frac{(n+1)(1-v_T)^n}{(1-v_T)^{n+1} - (v_V - v_T)^{n+1}} \tag{6}$$

where $v_T = V_{T0}/V_{DD}$ and $v_V = V_{\text{INV}}/V_{DD}$. Then the delay t_d, the delay from $0.5V_{DD}$ of input to $0.5V_{DD}$ of output, and the effective output transition time $t_{T\text{OUT}}$ can be expressed as follows. In calculating $t_{T\text{OUT}}$, the output waveform slope is approximated by 70% of its derivative at the half-V_{DD} point [4]. $t_{T\text{OUT}}$ can be used as t_T for the next logic gate:

($t_T \leqslant t_{T0}$: for the faster input)

$$t_d = t_T \left\{ \frac{1}{2} - \frac{1-v_T}{n+1} + \frac{(v_V - v_T)^{n+1}}{(n+1)(1-v_T)^n} \right\} + \frac{1}{2} \frac{C_O V_{DD}}{I_{D0}} \tag{7}$$

$$t_{T\text{OUT}} = \frac{C_O V_{DD}}{0.7 I_{D0}} \frac{4v_{D0}^2}{(4v_{D0} - 1)} \tag{8}$$

($t_T > t_{T0}$: for the slower input)

$$t_d = t_T \left[v_T - \frac{1}{2} + \left\{ (v_V - v_T)^{n+1} + \frac{(n+1)(1-v_T)^n}{2t_T I_{D0}/C_O V_{DD}} \right\}^{1/n+1} \right] \tag{9}$$

$$t_{T\text{OUT}} = \frac{C_O V_{DD}}{0.7 I_{D0}} \left(\frac{1-v_T}{t_d/t_T + 1/2 - v_T} \right)^n \tag{10}$$

where C_O is an output capacitance and $v_{D0} = V_{D0}/V_{DD}$.

To apply the above-mentioned formulas to a circuit of the form of Fig. 5(a), quantities including effective I_{D0}, n, and V_{D0} are required for the n series-connected MOSFET structure. One way of obtaining these values is by extracting them by fitting models to all possible compound $I-V$ curves, as

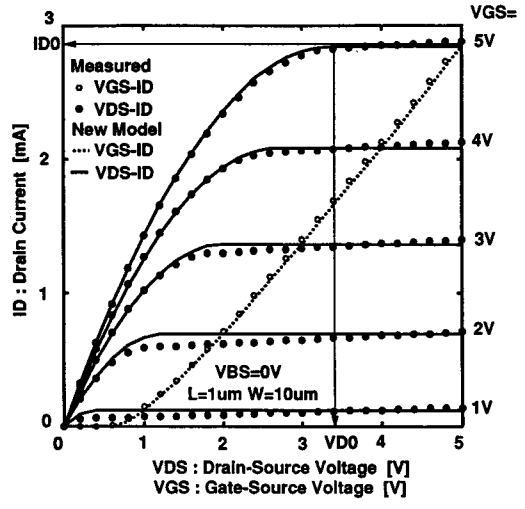


Fig. 1. NMOS $I-V$ curves with the new MOS model.

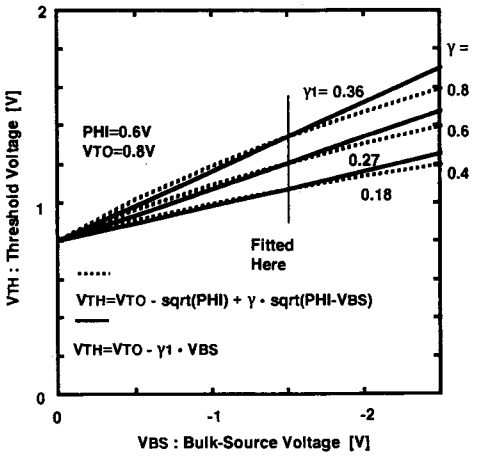


Fig. 2. Linear approximation of body effect.

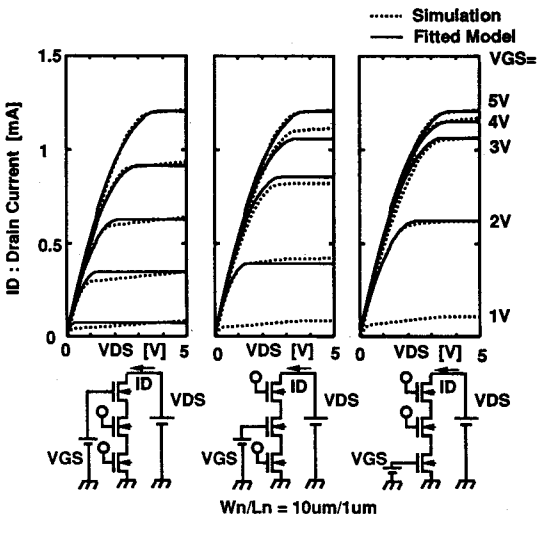


Fig. 3. $I-V$ curves of series-connected MOSFET's.

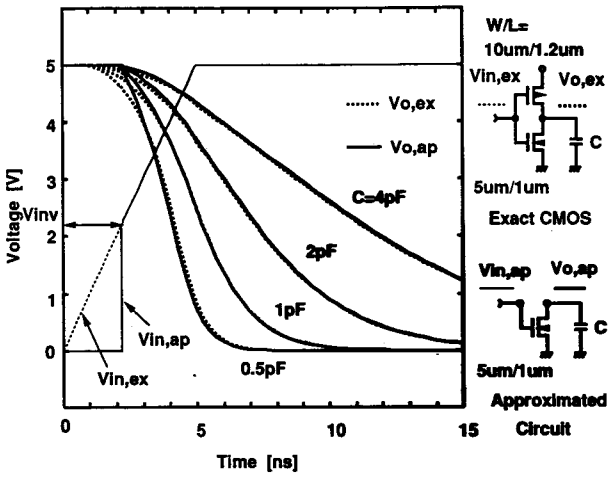


Fig. 4. Approximating CMOS by NMOS.

solved for the very fast input case and for the very slow input case, and the two solutions are connected smoothly. The ramp input waveform can be approximated by $V_{\text{in, ap}}$ which stays at 0 V until the ramp input goes across the logic threshold V_{INV}, abruptly rises up at that point, and coincides with the ramp input waveform therafter as shown in Fig. 4. The logic threshold voltage is the gate input voltage, which makes the output voltage equal to a half V_{DD}. The detailed derivation of the delay expressions can be found in Appendix A. First, define a critical input transition time t_{T0}:

$$t_{T0} = \frac{C_O V_{DD}}{2I_{D0}} \frac{(n+1)(1-v_T)^n}{(1-v_T)^{n+1} - (v_V - v_T)^{n+1}} \tag{6}$$

where $v_T = V_{T0}/V_{DD}$ and $v_V = V_{\text{INV}}/V_{DD}$. Then the delay t_d, the delay from $0.5V_{DD}$ of input to $0.5V_{DD}$ of output, and the effective output transition time $t_{T\text{OUT}}$ can be expressed as follows. In calculating $t_{T\text{OUT}}$, the output waveform slope is approximated by 70% of its derivative at the half-V_{DD} point [4]. $t_{T\text{OUT}}$ can be used as t_T for the next logic gate:

($t_T \leqslant t_{T0}$: for the faster input)

$$t_d = t_T \left\{ \frac{1}{2} - \frac{1-v_T}{n+1} + \frac{(v_V - v_T)^{n+1}}{(n+1)(1-v_T)^n} \right\} + \frac{1}{2} \frac{C_O V_{DD}}{I_{D0}} \tag{7}$$

$$t_{T\text{OUT}} = \frac{C_O V_{DD}}{0.7 I_{D0}} \frac{4v_{D0}^2}{(4v_{D0} - 1)} \tag{8}$$

($t_T > t_{T0}$: for the slower input)

$$t_d = t_T \left[v_T - \frac{1}{2} + \left\{ (v_V - v_T)^{n+1} + \frac{(n+1)(1-v_T)^n}{2t_T I_{D0}/C_O V_{DD}} \right\}^{1/n+1} \right] \tag{9}$$

$$t_{T\text{OUT}} = \frac{C_O V_{DD}}{0.7 I_{D0}} \left(\frac{1-v_T}{t_d/t_T + 1/2 - v_T} \right)^n \tag{10}$$

where C_O is an output capacitance and $v_{D0} = V_{D0}/V_{DD}$.

To apply the above-mentioned formulas to a circuit of the form of Fig. 5(a), quantities including effective I_{D0}, n, and V_{D0} are required for the n series-connected MOSFET structure. One way of obtaining these values is by extracting them by fitting models to all possible compound $I-V$ curves, as

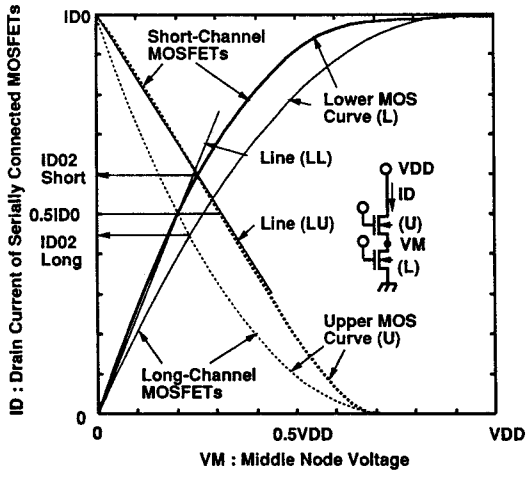


Fig. 8. Drain current of SCMS in long-channel device and short-channel device.

An RC model would predict that N series-connected MOSFET's would show approximately N times the delay compared with a single MOSFET, when C_O is dominant, because R is multiplied by N while C is kept constant. This is accurate for ideally long MOSFET's without body effect, where $n=2$, $v_{D0}=1-v_T$, and $\gamma_1=0$, because F_D becomes exactly equal to N. For shorter MOSFET's, however, the approximation is not valid.

For a long-channel MOSFET without body effect, the relation $F_D=N$ can be shown in a more rigorous way as follows. In this ideal case, the drain current I_D in the linear mode can be decomposed into $f(V_D)-f(V_S)$ [8], where V_D and V_S are the drain and source potential, respectively, and $f(V)=\beta\{(V_{DD}-V_{T0})V-0.5V^2\}$, where β is a constant. In this decomposition, the gate voltage of all transistors is assumed to be biased V_{DD}, which is true in the case of a NMOS SCMS. A PMOS case can be considered similarly.

It should be noted that with the gate bias condition mentioned above and with the drain voltage of the topmost transistor (V_N in Fig. 9(a)) being equal to V_{DD}, the topmost MOSFET is operating in the saturated mode and the other transistors are operating in the linear mode. In the following derivation, however, it is assumed that all MOSFET's are operating in the linear mode. This condition is true when the output node voltage V_N is less than $V_{DD}-V_{T0}$. When V_N becomes V_{DD}, the excessive voltage (V_{T0}) is consumed only by the topmost MOSFET, which is now operating in the saturated mode, but the drain current does not change with the excessive voltage because of the nature of the saturated mode. For a single MOSFET case, it is also true that the current does not change while the output node voltage is between $V_{DD}-V_{T0}$ and V_{DD}. Therefore, the result of the following derivation ($N\cdot I_{D0N}=I_{D0}$) holds even when the output node is biased at V_{DD} and the topmost MOSFET of the SCMS is operating in the saturated region.

Using the notation of Fig. 9(a), the following equations hold:

$$I_{D0N}=f(V_N)-f(V_{N-1}) \tag{13}$$

$$\cdots$$

$$I_{D0N}=f(V_2)-fV_1) \tag{14}$$

$$I_{D0N}=f(V_1)-f(V_0). \tag{15}$$

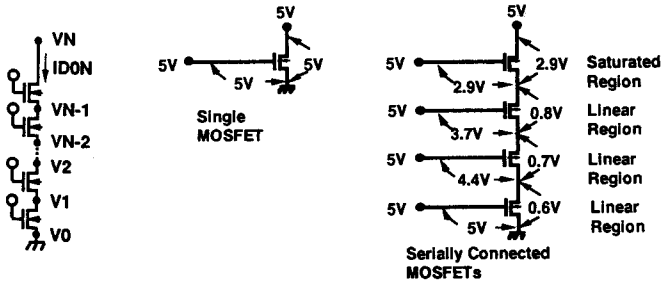


Fig. 9. (a) Notations for SCMS. (b) V_{GS} and V_{DS} observed in the SCMS for $V_{DD}=5$ V.

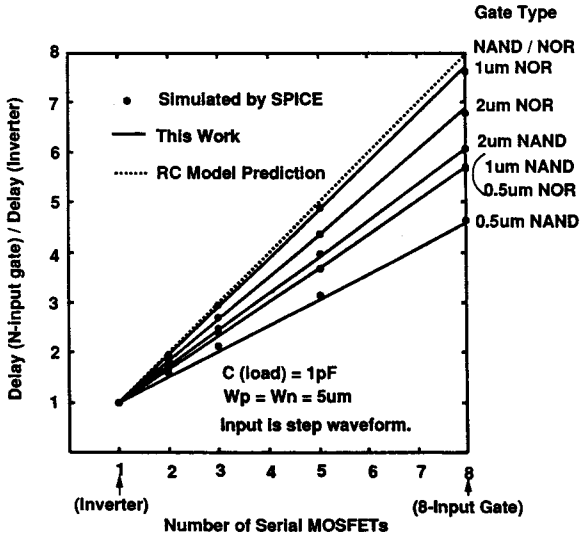


Fig. 10. Delay degradation factor of SCMS (C_O: large).

Summing these equations leads to

$$N\cdot I_{D0N}=f(V_N)-f(V_0)=I_{D0},\ \rightarrow F_D=I_{D0}/I_{D0N}=N. \tag{16}$$

The relation $F_D=N$ is rather surprising considering the nonlinear nature of MOSFET's. The naive understanding that an N-connected MOSFET shows N times as long a delay as a single MOSFET is true for the long-channel ideal case.

In Fig. 10, calculated results using (11) are compared with simulation for various generations of MOSFET's and excellent agreement can be seen. In Fig. 10, pull-down delay is plotted for the NAND gates and pull-up for NOR gates. The figure clearly shows the improvement of F_D in the submicrometer region. Equation (12) can be used as a simple index to estimate the delay degradation of the SCMS over a single MOSFET and provide insight into SCMS operation in the submicrometer region.

V. Step Input with Small Output Capacitance

An example for this case is shown in Fig. 11. An RC model predicts N series-connected MOSFET's with small output capacitance would show N^2 times the delay compared with a single MOSFET. However, the real situation is more favorable to the SCMS. Fig. 13 also shows that the

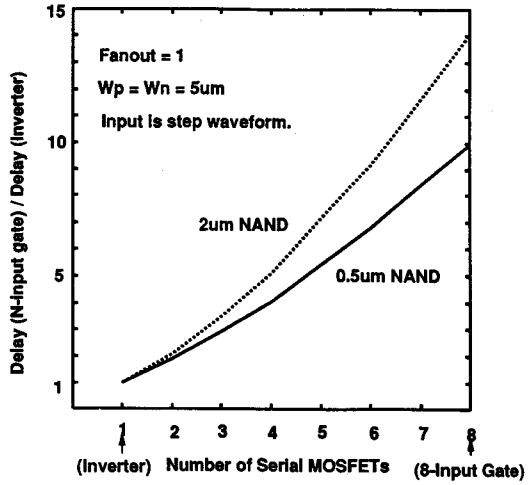


Fig. 11. Delay degradation factor of SCMS (C_O: small, input is applied to the lowest input terminal).

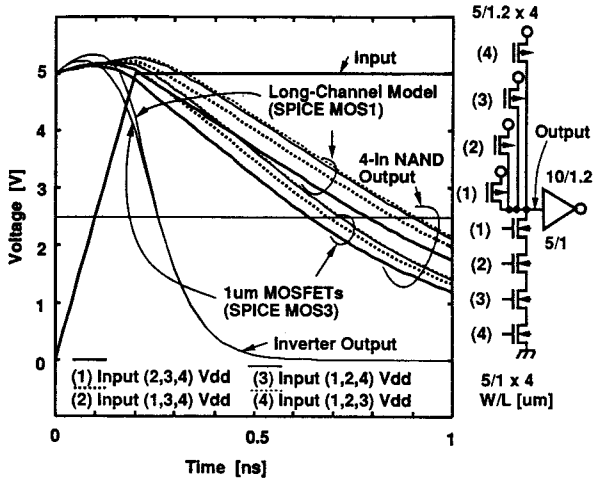


Fig. 12. Inverter and NAND gate behavior with long-channel and short-channel MOS model.

degradation factor gets better in the submicrometer region, the physical explanation of which is given in the next section.

VI. The General Case and Physical Interpretation

In the general case, where C_O may not dominate and the input waveform has a finite slope, the analysis becomes more complex. However, the claim that F_D decreases in the submicrometer region is still true, an example of which is shown in Fig. 12. This is because the capacitance ratio of the SCMS to the single MOSFET is basically unchanged even if the feature size is changed, while the current ratio of the SCMS to a single MOSFET is improved in the submicrometer region.

The physical interpretation of the improvement in the current ratio is as follows. In the SCMS, the V_{DS} of each MOSFET is smaller than that of an inverter MOSFET since the output voltage is spread across multiple MOSFET's. The V_{GS} of each MOSFET is also smaller because the source voltage is raised from ground (or lowered from V_{DD} in the PMOS case). This situation is illustrated in Fig. 9(b) assuming $V_{DD} = 5$ V. Because of the reduced V_{DS} and V_{GS}, the carriers feel less electric field both parallel to and perpendicular to the channel. Consequently, velocity saturation is mitigated in the SCMS compared with an inverter and a relatively large current flows in the SCMS in the submicrometer region.

The situation might change because the SCMS but not the inverter suffers from the body effect. However, as shown in Fig. 12, the current improvement induced by the mitigated velocity saturation dominates the current degradation induced by the body effect. Moreover, there are technologies like p-pocket which can suppress the body effect while the velocity saturation gets severer in the actual devices as is seen, for example, in [9].

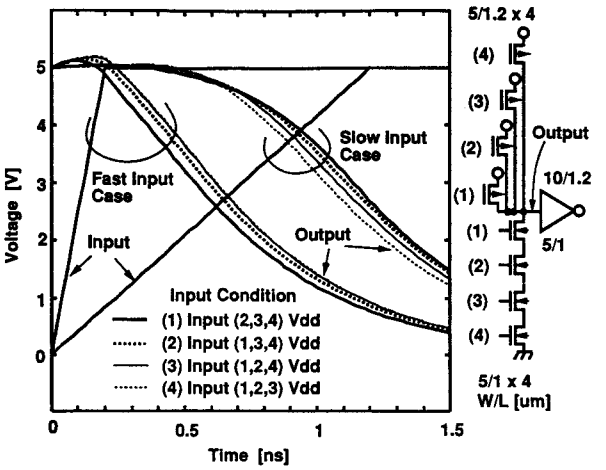


Fig. 13. Delay comparison among various input terminals of four-input NAND gate.

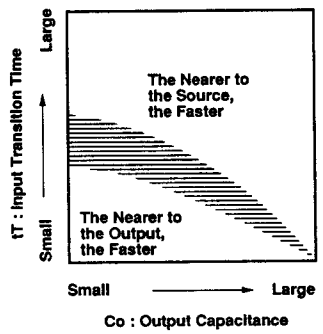


Fig. 14. Delay dependence on input terminal in NAND/NOR/complex gates.

VII. Delay Dependence on Input Terminal Position

Which input of a four-input NAND/NOR has the shortest delay to the output? Consider the NAND case since the NOR case follows from symmetry. When the output capacitance C_O is very large compared with the capacitance of the logic gate itself, the lower (nearer to ground) terminal shows the shorter delay. This is because n becomes smaller for these lower terminals, as shown in Fig. 3. This means that the drain current quickly approaches its final value when changing V_{GS} and enables the faster discharging of the output capacitance.

If the output capacitance C_O is small, there are two cases to consider, depending on the value of t_T, as shown in Fig. 13. When t_T is large, the lower terminals show faster operation because the logic threshold is lowered and is achieved faster (n is smaller and only small V_{GS} is need to turn the device on hard). When t_T is small, the lower MOSFET must discharge the upper MOSFET capacitances so the upper terminal shows a faster delay. These situations are illustrated in Fig. 14.

VIII. Conclusion

The SCMS is analyzed. It has been shown that the ratio of the delay of NAND/NOR to the delay of inverter becomes smaller in the submicrometer region. There are cases where N series-connected MOSFET's show only $N/2$ times as long a delay as a single MOSFET. This result encourages the use of NAND/NOR/complex gates, PLA's, CVSL [1], and hot-carrier resistant logic [2] in submicrometer circuit design.

The result also suggests the reexamination of the VLSI design/optimization in the submicrometer region. For example, the logic threshold voltage of a NAND gate becomes much lower than 0.5 V_{DD} in the submicrometer region, if the W_P/W_N ratio is chosen the same as in a long-channel MOSFET generation (see (A15)). It has been shown that the accuracy of an RC-based model is deteriorated for carrier velocity saturated MOSFET's.

Appendix A
Derivation of the Delay Expression

In most practical cases, the channel-length modulation effect is small because MOSFET's are usually engineered in such a way that the channel-length modulation is minimized. In some cases, however, channel-length modulation is eminent. Considering this situation, the channel-length modulation effect is included in this appendix by modifying (3) and (4) as follows:

$$I_D = I_{D5} = I_{DSAT}(1+\lambda V_{DS}) (V_{DS} \geq V_{DSAT}: \text{saturated region}) \tag{A1}$$

$$I_D = I_{D3} = I_{D5}\left(2 - \frac{V_{DS}}{V_{DSAT}}\right)\frac{V_{DS}}{V_{DSAT}} \cdot (V_{DS} < V_{DSAT}: \text{linear region}) \tag{A2}$$

where λ is a widely used channel-length modulation parameter. The subscripts 3 and 5 for I_D denote a triode and a pentode operating region, respectively.

In this appendix, the discharging of an output capacitance through NMOS's is explained since the discussion for the charging by PMOS's is symmetric. As seen from Fig. 4, a CMOS inverter with a ramp input can be approximated by an NMOS circuit with an input waveform like $V_{\text{in, ap}}$. $V_{\text{in, ap}}$ is the same as the real ramp input except that it remains zero until the input reaches the logic threshold voltage. The logic threshold voltage is the gate input voltage which makes the output voltage equal to half V_{DD}.

For the extreme cases, this approximation is exact. That is, for the ultimately fast input case, the ramp input becomes a step function and $V_{\text{in, ap}}$ also becomes the step function and the current through PMOS can be completely neglected. For the extremely slow input, the output changes abruptly and comes down to 0.5 V_{DD} when the input goes across the logic threshold voltage. The approximated circuit shows the same delay. The intermediate case is shown in Fig. 4 and this approximation greatly reduces the complexity of the system and make it possible to treat the CMOS inverter delay analytically.

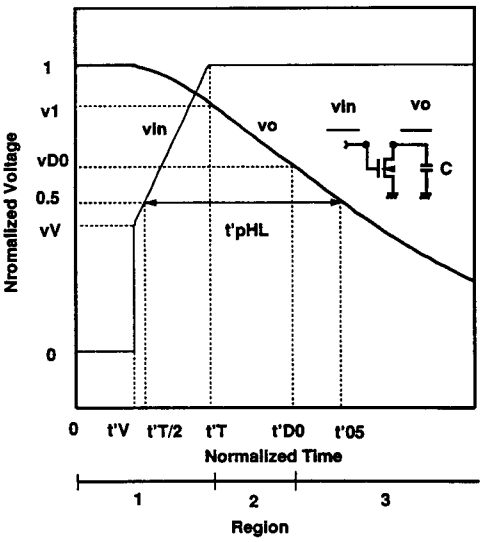


Fig. 15. Input/output waveforms of fast input case.

The strategy for solving the differential equation which governs the discharging process is to solve it for the very fast input case and for the very slow input case separately as is mentioned in the text. The two solutions for the two extreme cases happen to be connected smoothly.

In the following, voltages are normalized by V_{DD}, currents by I_{D0}, and time by $\tau = C_O V_{DD}/I_{D0}$. The normalized voltage is denoted as v instead of V, the normalized current i instead of I, and the normalized time t' instead of t. λ' denotes λV_{DD}. First, consider a very fast input case as shown in Fig. 15 (see this figure also for notations used in the following). There are three regions: Region 1, the time before the input reaches V_{DD}, Region 2, the time before the output reaches V_{D0}, and Region 3, the time after the output reached V_{D0}.

In Region 1, the differential equation which governs the discharging process can be written as

$$\frac{dv_O}{dt'} = -i_5 = -\left(\frac{t'/t'_T - v_T}{1-v_T}\right)^n \frac{1+\lambda' v_O}{1+\lambda'} \tag{A3}$$

which should be solved with the initial condition of $v_O = 1$ at $t' = t'_V$. The solution is

$$\frac{1+\lambda'}{\lambda'} \log \frac{1+\lambda' v_O}{1+\lambda'} = -\frac{t'_T}{(1-v_T)^n (n+1)} \cdot \left\{\left(\frac{t'}{t'_T} - v_T\right)^{n+1} - (v_V - v_T)^{n+1}\right\}. \tag{A4}$$

v_1 is obtained by substituting t' by t'_T.

In Region 2, the differential equation is simple since the input is constant V_{DD}:

$$\frac{dv_O}{dt'} = -i_5 = \frac{1+\lambda' v_O}{1+\lambda'}. \tag{A5}$$

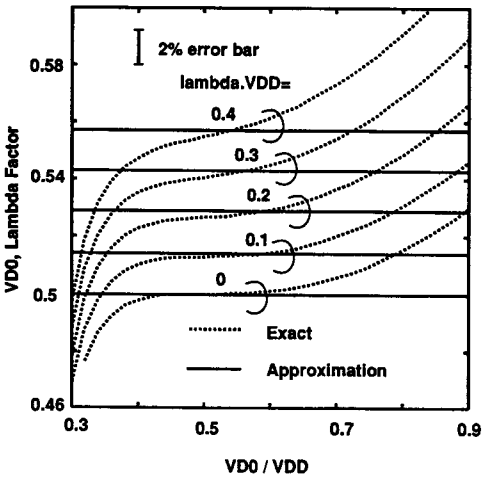


Fig. 16. Approximation of V_{D0} and λ term.

The initial condition is $v_O = v_1$ at $t' = tT$ and the solution is

$$\frac{1+\lambda'}{\lambda'}\log\frac{1+\lambda' v_O}{1+\lambda'} = -\frac{t'_T}{(1-v_T)^n(n+1)} \cdot\left\{(1-v_T)^{n+1}-(v_V-v_T)^{n+1}\right\}-(t'-t'_T). \quad \text{(A6)}$$

t'_{D0} is obtained by letting v_O to v_{D0} and is written as follows:

$$t'_{D0} = t'_T\left\{1-\frac{1-v_T}{n+1}+\frac{(v_V-v_T)^{n+1}}{(n+1)(1-v_T)^n}\right\} -\frac{1+\lambda'}{\lambda'}\log\frac{1+\lambda' v_{D0}}{1+\lambda'}. \quad \text{(A7)}$$

In Region 3, where the MOSFET is operating in the linear region,

$$\frac{dv_O}{dt'} = -i_3 = -\left(2-\frac{v_O}{v_{D0}}\right)\frac{v_O}{v_{D0}}\frac{1+\lambda' v_O}{1+\lambda'} \quad \text{(A8)}$$

is the differential equation to be solved with the initial condition of $v_O = v_{D0}$ at $t' = t'_{D0}$. The solution is

$$t' = t'_{D0}+(1+\lambda')v_{D0}\left\{\frac{\lambda' v_{D0}}{1+2\lambda' v_{D0}}\log\frac{1+\lambda' v_O}{1+\lambda' v_{D0}} +\frac{1}{2(1+2\lambda' v_{D0})}\log\left(2\frac{v_O}{v_{D0}}\right)-\frac{1}{2}\log\frac{v_O}{v_{D0}}\right\}. \quad \text{(A9)}$$

Therefore, the delay $t'_d (= t'_{pHL})$ can be expressed as follows:

$$t'_d = t'\left(v_O=\frac{1}{2}\right)-\frac{t'_T}{2} \approx t'_T\left\{\frac{1}{2}-\frac{1-v_T}{n+1}+\frac{(v_V-v_T)^{n+1}}{(n+1)(1-v_T)^n}\right\} +\frac{1}{2}+\frac{\lambda'}{7}. \quad \text{(A10)}$$

To derive this expression, the complicated term of v_{D0} and λ' in (A7) and (A9) is approximated by $1/2+\lambda'/7$. The error of this approximation is less than 2% when $0.5 < v_{D0} < 0.7$, $0 \leqslant \lambda' < 0.25$, and less than 4% when $0.4 < v_{D0} < 0.8$, $0 \leqslant \lambda' < 0.4$, as shown in Fig. 16. The transition time of the output waveform, $t'_{T\mathrm{OUT}}$, is calculated as

$$t'_{T\mathrm{OUT}} = \frac{1}{0.7}\frac{dt'}{dv_O}\Big]_{v_O=0.5} = \frac{8v_{D0}^2(1+\lambda')}{0.7(4v_{D0}-1)(2+\lambda')}. \quad \text{(A11)}$$

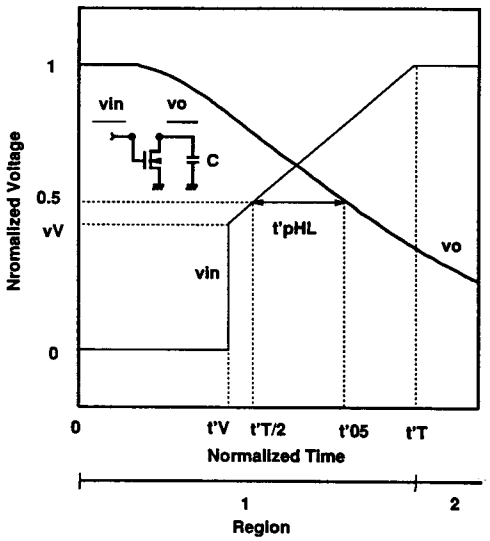


Fig. 17. Input/output waveforms of slow input case.

When the input is very slow, the output crosses 0.5 V_{DD} in Region 1 as shown in Fig. 17. In this case (A4) is valid, and using (A4) the delay $t'_d (= t'_{05} - 1/2t_T)$ is obtained as

$$t'_d = t'_T\left[v_T-\frac{1}{2}+\left\{(v_V-v_T)^{n+1} -\frac{(1-v_T)^n(n+1)}{t'_T}\frac{1+\lambda'}{\lambda'}\log\frac{2+\lambda'}{2+2\lambda'}\right\}^{1/n+1}\right]$$

$$\approx t'_T\left[v_T-\frac{1}{2}+\left\{(v_V-v_T)^{n+1} +\frac{(1-v_T)^n(n+1)}{t'_T}\left(\frac{1}{2}+\frac{\lambda'}{7}\right)\right\}^{1/n+1}\right]. \quad \text{(A12)}$$

The error of the approximated formula is less than 2% when $0 \leqslant \lambda' < 0.25$, and less than 4% when $0 \leqslant \lambda' < 0.4$. $t'_{T\mathrm{OUT}}$ is calculated as

$$t'_{T\mathrm{OUT}} = \frac{1}{0.7}\frac{dt'}{dv_O}\Big]_{v_O=0.5,\, t'=t'_{05}=t'_d+t'_T/2} = \left(\frac{1-v_T}{t'_d/t'_T+1/2-v_T}\right)^n\frac{2+2\lambda'}{2+\lambda'}. \quad \text{(A13)}$$

The solution for the fast input case, (A10), and that for the very slow input case, (A12), can be connected at the critical input transition time t'_{T0} given below. t'_{T0} can be calculated by equating (A10) and (A12). It should be noted that not only the values of the both equations but also the first derivatives coincide at the critical time:

$$t'_{T0} = \frac{(n+1)(1-v_T)^n}{(1-v_T)^{n+1}-(v_V-v_T)^{n+1}}\left(\frac{1}{2}+\frac{\lambda}{7}\right). \quad \text{(A14)}$$

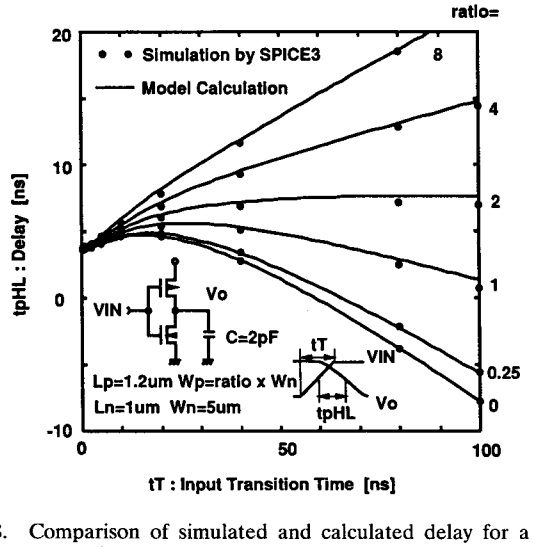


Fig. 18. Comparison of simulated and calculated delay for a CMOS inverter with various t_T and W_p/W_n.

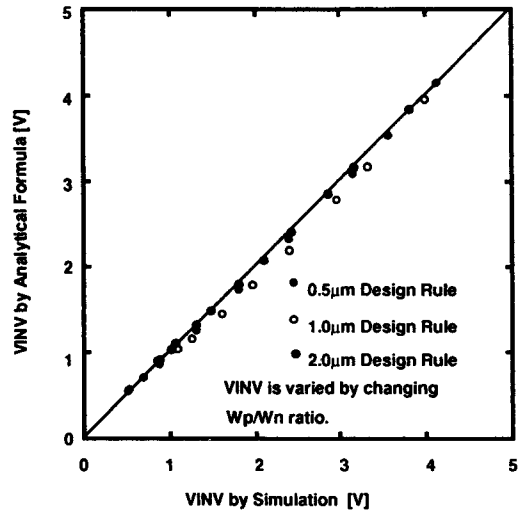


Fig. 19. Comparison of approximated formula and simulation for CMOS inverter logic threshold voltage V_{INV}.

Equations (A14), (A9), (A10), (A12), and (A13) correspond to (6), (7), (8), (9), and (10) of the text, respectively. An example of a calculation using these formulas is demonstrated in Fig. 18. The formulas are valid in the wide range of t_T and the channel-width ratio of PMOS and NMOS. The logic threshold voltage V_{INV} was calculated by the following expression:

$$v_V = \frac{V_{INV}}{V_{DD}} = \frac{I_{D0N}^{1/n} v_{TN} + I_{D0P}^{1/n}(1-v_{TN})}{I_{D0N}^{1/n} + I_{D0P}^{1/n}(1-v_{TN})/(1-v_{TP})},$$

$$n = \frac{n_N + n_P}{2} \quad \text{(A15)}$$

where the subindexes N and P denote NMOS and PMOS, respectively. The accuracy of this formula is shown in Fig. 19 for various generations of design rules.

Appendix B
Expression for the Delay Degradation Factor

In this appendix, the channel-length modulation is assumed not negligible, that is, $\lambda \neq 0$ (see Fig. 8 for the notation). For the upper MOSFET, the drain current I_D is written as

$$I_D = I_{D0}\left(\frac{1-v_M-v_T-\gamma_1 v_M}{1-v_T}\right)^n \frac{1+\lambda'(1-v_M)}{1+\lambda'}. \quad \text{(B1)}$$

This curve (curve U in Fig. 8) goes through (v_M, I_U), defined as

$$v_M = v_U = \frac{1-v_T}{1+\gamma_1}\left(1-\frac{1}{2^{1/n}}\right)$$

$$I_U = \frac{1}{2} I_{D0} \frac{1+\lambda'(1-v_U)}{1+\lambda'}. \quad \text{(B2)}$$

The line LU in Fig. 8 is drawn to pass $(0, I_{D0})$ and (v_M, I_U).

For the lower MOSFET, I_D is expressed as

$$I_D = I_{D0} \frac{1+\lambda' v_M}{1+\lambda'}\left(2-\frac{v_M}{v_{D0}}\right)\frac{v_M}{v_{D0}}. \quad \text{(B3)}$$

Therefore, the curve L goes through (v_L, I_L) defined by the following expressions:

$$v_M = v_L = \left(1-\frac{1}{\sqrt{2}}\right)v_{D0}$$

$$I_L = \frac{1}{2}\frac{1+\lambda' v_L}{1+\lambda'} I_{D0}. \quad \text{(B4)}$$

The line LL is chosen so as to pass $(0,0)$ and (v_M, I_L).

The N-connected MOSFET case can be treated similarly. Since $N-1$ lower MOSFET's are operating in the linear region, the effective resistance of the $N-1$ transistors is $N-1$ times higher than the effective resistance of the lower single transistor (L in Fig. 8). Therefore, when the single MOSFET L in Fig. 8 is replaced by $N-1$ MOSFET's, a curve of v_M versus the effective drain current of the $N-1$ MOSFET's goes through $(0,0)$ and $(v_M, I_L/(N-1))$. That is, the slope of the line LL becomes $1/(N-1)$.

By solving the intersection of the line LU and the line LL, I_{D0N} can be obtained:

$$I_{D0} - \frac{I_{D0}-I_U}{v_U} v_M = \frac{I_L}{v_L}\frac{v_M}{N-1} = I_{D0N}. \quad \text{(B5)}$$

Elimination of v_M leads to

$$F_D = \frac{I_{D0}}{I_{D0N}} = 1 + \frac{I_{D0}-I_U}{v_U}\frac{v_L}{I_L}(N-1). \quad \text{(B6)}$$

With the assumption that λ is small and $(v_U - v_L) \ll 1$, F_D is reduced to

$$F_D = 1 + \frac{1-2^{-1/2}}{1-2^{-1/n}}\frac{v_{D0}}{1-v_T}(1+\gamma_1)(1+\lambda')(N-1). \quad \text{(B7)}$$

This formula corresponds to (11) in the text.

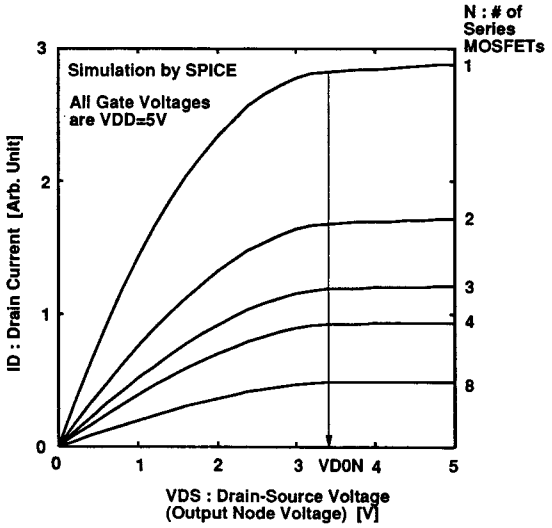


Fig. 20. Effective drain saturation voltage of SCMS.

Appendix C
Approximated Expressions for the Quantities Related to the SCMS

In order to use the delay formulas derived in Appendix A for the SCMS, I_{D0N}, V_{INV}, V_{D0}, and n for the SCMS are required. In this appendix, an approximated way of obtaining these parameters is described. As for I_{D0N} and V_{INV}, they have already been given by (B7) and (A15), respectively. The effective drain saturation voltage V_{D0} of the SCMS is essentially unchanged from the V_{D0} of the single MOSFET, as seen in Fig. 20.

The remaining quantity is n. Let n_{NJ} denote the velocity saturation index n observed when the Jth input terminal counting from the output of the N series-connected MOSFET's is chosen as an input. The NMOS case is explained here, but the PMOS case is symmetric.

First, the case of $N=2$ and $J=2$ is discussed. Suppose $0.5\ V_{DD}$ is applied to the lower MOSFET gate; the drain current I_{DM22} is expressed as follows because the lower MOSFET is operating in the saturated region:

$$I_{DM22} = \frac{W}{L_{EFF}} B \left(\frac{1}{2} V_{DD} - V_{TH} \right)^n . \tag{C1}$$

Knowing that the drain current is I_{D02} at $V_{GS} = V_{DD}$ and I_{DM22} at $V_{GS} = 0.5 V_{DD}$, n_{22} is calculated as

$$n_{22} = \frac{\log(I_{D02}/I_{DM22})}{\log((1-v_T)/(0.5-v_T))} . \tag{C2}$$

The next case is $N=2$ and $J=1$. In this case, the drain current I_{DM21}, which flows when the gate voltage is set to $0.5\ V_{DD}$, can be calculated using a technique similar to that used in Appendix B:

$$I_{DM21} = I_{DM22} \Big/ \left\{ 1 + \frac{1}{2} n \frac{v_{D0}}{0.5 - v_T} (1 + \gamma_1) \right\} . \tag{C3}$$

Then, the following expression holds:

$$n_{12} = \frac{\log(I_{D02}/I_{DM21})}{\log((1-v_T)/(0.5-v_T))} . \tag{C4}$$

For the general N and J, the following empirical formula can be employed:

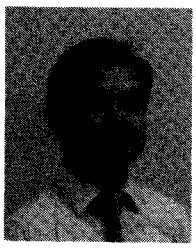
$$n_{NJ} = \frac{n_{21} n_{22}}{(n_{21} - n_{22})(J-1) + n_{22}} \tag{C5}$$

Acknowledgment

The encouragement of Prof. R. Brayton, Prof. A. Sangiovanni-Vincentelli, Y. Unno, Dr. Y. Takeishi, Y. Fukuda, H. Yamada, and Dr. T. Iizuka throughout the course of this work is appreciated. Discussions with H. Ishiuchi and T. Fujii on MOS physics and circuit designs were inspiring and should be acknowledged. Assistance provided by Dr. T. Quarles and Dr. R. Spickelmier concerning SPICE and computer environments is also appreciated. Lastly, the critical reading and useful comments of the reviewers, which greatly improved the quality of the paper, are gratefully appreciated.

References

[1] L. Heller, W. Griffin, J. Davis, and N. Thoma, "Cascode voltage switch logic—A differential logic family," in *ISSCC Dig. Tech. Papers*, Feb. 1984, pp. 16–17.
[2] 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.
[3] T. Sakurai and A. R. Newton, "A simple short-channel MOSFET model and its application to delay analysis of inverters and series-connected MOSFET's," in *Proc. ISCAS*, May 1990, TUAM-3-7; see also "A simple MOSFET model for circuit analysis and its application to CMOS gate delay analysis series-connected MOSFET structure," Dept. EECS, Univ. of Calif., Berkeley, ERL memo. Mar. 1990.
[4] 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.
[5] A. Vladimirescu and S. Liu, "The simulation of MOS integrated circuits using SPICE2," Univ. of Calif., Berkeley, ERL Memo. M80/7, Oct. 1980.
[6] J. K. Ousterhout, "A switch-level timing verifier for digital MOS VLSI," *IEEE Trans. Computer-Aided Design*, vol. CAD-4, no. 3, pp. 336–349, July 1985.
[7] P. Penfield and J. Rubinstein, "Signal delay in RC tree networks," in *Proc. 18th DAC*, June 1981, pp. 613–617.
[8] M. Horowitz, "Timing models for MOS pass networks," in *Proc. ISCAS*, 1983, pp. 198–201.
[9] T. Sakurai and A. R. Newton, "A simple MOSFET model for circuit analysis," to be published in *IEEE Trans. Electron Devices*.



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 University of Tokyo, Tokyo, Japan, in 1976, 1978, and 1981, respectively. His Ph.D work was 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, 64- and 256-Kb SRAM, 1-Mb virtual SRAM, cache memories, and a RISC with on-chip large cache memory. During these developments, he also worked on the modeling of wiring capacitance and delay, new soft-error-free memory cells, new memory architectures, new hot-carrier resistant circuits, arbiter optimization, and gate-level delay modeling. For one year and a half beginning in 1988, he was 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 in application-specific memories, BiCMOS ASIC's, VLSI microprocessors, 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) 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.

Prof. Newton 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. He 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 INTEGRATED CIRCUITS AND SYSTEMS from 1985 to 1988 and a member of the Circuits and Systems Society ADCOM.