CMOS Circuit Speed and Buffer Optimization

NILS HEDENSTIERNA AND KJELL O. JEPPSON, SENIOR MEMBER, IEEE

Abstract—**An improved timing model for CMOS combinational logic is presented. The model is based on an analytical solution for the CMOS inverter output response to an input ramp. This model yields a better understanding of the switching behavior of the CMOS inverter than the step-response model by considering the slope of the input waveform. Essentially, the propagation delay is shown to be the sum of the step-response delay and an input dependent delay that may account for as much as 50–100 percent of the total delay. The matching between the ramp input and the characteristic input waveforms is shown to be easily performed for excellent agreement in output response and propagation delay. Even though the short-circuit current is neglected, its influence is shown to be small and may be corrected. As an example, the timing model is used to optimize CMOS output buffers for minimum delay. If the intrinsic output load capacitance is included in the model, the optimum tapering factor is shown to be not e but a value in the range 3–5 depending on process parameters and design style. Also, due to the input dependence of the propagation delay, the last inverter stage in the buffer should have a larger tapering factor than the other stages for minimum delay.**

I. INTRODUCTION

DURING the 1980's, CMOS technology has evolved as a major technology for VLSI design. The use of algorithms to optimize circuit performance and software simulations to verify logic and timing operation have become major tools in VLSI design. As a result, there is a strong need for accurate analytical models to describe circuit operation in general and CMOS circuit operation in particular.

One important problem for VLSI design verification is to find precise models for the propagation delay. Circuit simulators, e.g., SPICE, consume too much CPU time to be practical for other than small circuits with less than a few hundred transistors, and logic simulators, e.g., TEGAS, that can handle up to several tens of thousands of gates usually rely on insufficient delay models. Most textbook analytical models [1] for the transient response of CMOS inverters rely on step input waveforms. These delay models are generally insufficient since they do not consider a realistic input waveform and consequently do not take into account the influence of the input waveform on the propagation delay. In real circuit applications, the input waveform will depend on the fan-out and the driving capability of the preceding stage and, therefore, the propagation delay will also depend on these parameters. Recently, some more precise delay models have been presented for NMOS logic [2]–[4].

In this paper, an analytical solution for the CMOS inverter output response to an input voltage ramp is presented in Section II. The model yields an analytical expression for the propagation delay as a function of the input ramp rise (fall) time with excellent agreement with SPICE simulations. The analytical model neglects the short-circuit power dissipation but, in a typical situation with equal input and output slopes, the error in propagation delay will be less than a few percent. Also, after comparisons with SPICE simulations, we are able to conclude that an average input voltage ramp may be used as a very good approximation for most typical input waveforms. Therefore, we may generalize our analytical expression for characteristic input waveforms. The propagation is then a function not only of the fan-out and the driving capability of the stage itself, as in the step-response model, but also a function of the fan-out and the driving capability of the preceding stage. In a typical CMOS circuit application, the input-dependent propagation delay may well account for as much as 50–100 percent of the total delay.

Even if the short-circuit power dissipation is neglected in the analysis, the expression for the output waveform may be used to approximately estimate the short-circuit dissipation as long as it is small compared to the dynamic power dissipation. This is done in Section III. The propagation delay may then be corrected to improve agreement with SPICE simulations.

The capacitive loads of the CMOS inverter are studied in Section IV in order to include the intrinsic delay of the inverter itself. Using this model, the optimum ratio between the widths of the P- and N-channel transistors in the inverter may be determined. The analysis is performed for inverters but may easily be extended to more complex logic gates.

In Section V, we give a few examples on how to apply the propagation delay model to buffers driving large capacitive loads in order to optimize the buffer for minimum delay. As a result, the optimum tapering factor between two individual inverter stages may be determined as a process and design-style-dependent constant. The main result of the buffer optimization is that, when the intrinsic delay is taken into account, the optimum tapering factor is approximately 3–5, depending on the processing parameters and the design style, and not e (the base of the natural logarithm) as shown by Mead and Conway [5]. Also, due to the input dependence of the delay, the ta-

Manuscript received March 3, 1986; revised November 14, 1986.

The authors are with the Department of Solid State Electronics, School of Electrical and Computer Engineering, Chalmers University of Technology, S-412 96 Göteborg, Sweden.

IEEE Log Number 8612996.

pering factor in the last inverter stage should be larger than in the other inverters.

Finally, in Section VI, the trade-off situation between speed and area is illustrated. In conclusion, the analysis in this paper yields an optimizing algorithm which is of great help in buffer compilers, both for output buffers and critical internal drivers of large capacitive loads, e.g., clock drivers.

II. Analysis

A. Step Response

The speed of CMOS inverters is usually calculated as the step-response propagation delay. The fall t_{dHL} and rise t_{dLH} propagation delays are then calculated as the time to discharge the load capacitor C_L through the N-channel transistor and the time to charge the load capacitor through the P-channel transistor, respectively. The input voltage V_{in} to the CMOS inverter is then assumed to be an ideal voltage step [1].

In both cases, the step-response output voltage V_{out} is first calculated and the propagation delay is determined as the time needed for the output voltage to reach a certain voltage level, e.g., $V_{DD}/2$. The output voltage response for a rising and falling step, respectively, is then determined from the differential equations

$$-C_L \frac{dV_{out}}{dt} = \begin{cases} k_N\left[(V_{in} - V_{TN})\, V_{out} - V_{out}^2/2\right] & \text{for } V_{in} - V_{TN} > V_{out} \\ \dfrac{k_N}{2}(V_{in} - V_{TN})^2 & \text{for } V_{in} - V_{TN} < V_{out} \end{cases} \quad (1)$$

and

$$C_L \frac{dV_{out}}{dt} = \begin{cases} k_P\left[(V_{in} - V_{DD} - V_{TP})(V_{out} - V_{DD}) - (V_{out} - V_{DD})^2/2\right] & \text{for } V_{in} - V_{TP} < V_{out} \\ \dfrac{k_P}{2}(V_{in} - V_{DD} - V_{TP})^2 & \text{for } V_{in} - V_{TP} > V_{out} \end{cases} \quad (2)$$

where k_N and k_P are the N- and P-channel transistor constants. V_{TN} and V_{TP} are the N- and P-channel transistor threshold voltages, respectively.

It is well known [1] that the fall and rise propagation delays may be written

$$t_{dHL} = \frac{C_L/k_N}{V_{DD}(1-n)}\left[\frac{2n}{1-n} + \ln\left(\frac{2(1-n) - v_0}{v_0}\right)\right] \equiv \frac{C_L}{k_N} A_N \quad (3)$$

and

$$t_{dLH} = \frac{C_L/k_P}{V_{DD}(1+p)}\left[\frac{-2p}{1+p} + \ln\left(\frac{2(1+p) - v_0}{v_0}\right)\right] \equiv \frac{C_L}{k_P} A_P \quad (4)$$

respectively, where $v_0 = V_{out}/V_{DD}$ is the normalized output voltage and $n = V_{TN}/V_{DD}$ and $p = V_{TP}/V_{DD}$ are the normalized threshold voltages of the N- and P-channel transistors.

A_N and A_P may be regarded as process constants for a certain supply voltage and with the propagation delay defined at a fixed level. Some typical values for A_N and A_P, for $n = -p = 0.12$ and $V_{DD} = 5$ V at the 50-percent level, are $A_N = A_P = 0.27$.

B. Ramp Response

The purpose of this paper is to get a more accurate picture of the output voltage response by letting the input voltage be a voltage ramp. The normalized input voltage may then be written

$$v_{in} = \begin{cases} 0, & t < 0 \\ st, & 0 < t < \tau \\ 1, & t > \tau \end{cases} \quad (5)$$

where $\tau = 1/s$ is the rise time of the input voltage ramp. A similar equation is valid for a negative input ramp.

To calculate an analytical expression for the output ramp response, we use the same differential equations (1) and (2) as in the case of an input step voltage. Thereby, we neglect the short-circuit current flowing through the N-channel transistor for a negative input ramp and through the P-channel transistor for a positive input ramp, respectively. This approximation will be justified in Section III. To solve the differential equation, the operation of the CMOS inverter must be divided into different regions. Since the input voltage ramp will reach its final value with the transistor either in saturation or in the linear region, we get two main cases which will each be divided into three regions. In the beginning, as long as $0 < v_{in} - n < v_0$, the N-channel transistor is saturated. We now define the normalized output saturation voltage v_1 when the transistor leaves saturation and enters the linear region. This voltage is given by $v_1 = v_{in} - n$. We also define v_2 as the output voltage when the input voltage reaches its

final value. The two main cases are then defined by $v_2 > v_1$ and $v_1 > v_2$.

Case A: $v_2 > v_1$: The first case to be studied is for fast input ramps such that the transistor is still saturated when the input voltage ramp reaches its final value, i.e., $v_2 > v_1$. In this case, the saturation condition is given by $v_1 = 1 - n$.

Region 1: $v_0 > v_2$: As long as the transistor is saturated and the input voltage is a ramp, the differential equation becomes

$$C_L \frac{dv_0}{dt} = -\frac{k_N}{2} V_{DD}(st - n)^2 \tag{6}$$

and with the initial condition $v_0 = 1$ for $st = n$, integration yields

$$v_0 = 1 - \frac{k_N V_{DD}}{6sC_L}(st - n)^3. \tag{7}$$

Since the transistor is still saturated when the input voltage reaches its final value, the output voltage at this time

$$v_2 = 1 - \frac{k_N V_{DD}}{6sC_L}(1 - n)^3 \tag{8}$$

is larger than the saturation voltage, i.e., $v_2 > v_1 = 1 - n$. This condition may be formulated as

$$\frac{k_N V_{DD}}{sC_L} < \frac{6n}{(1 - n)^3} \tag{9}$$

which gives the minimum input slope valid in Case A.

Region 2: $v_1 < v_0 < v_2$: With the input voltage at its final value and the N-channel transistor still saturated, the output voltage is given by

$$v_0 = v_2 - \frac{k_N V_{DD}}{2C_L}(1 - n)^2 (t - \tau) \tag{10}$$

This is equivalent to the step response, the only difference being that the initial condition for the normalized output voltage is v_2 instead of 1 for $t = \tau$.

Region 3: $v_0 < v_1$: For $v_0 < v_1 = 1 - n$, the transistor enters the linear region, and the output voltage (also in similarity with the step response) is given by

$$t - t_{\text{sat}} = \frac{C_L}{k_N V_{DD}(1 - n)} \ln\left(\frac{2(1 - n) - v_0}{v_0}\right) \tag{11}$$

where t_{sat} is the time when $v_0 = v_1$. Insertion of this initial condition into (10) yields

$$t_{\text{sat}} = \tau + \frac{2C_L(v_2 - (1 - n))}{k_N V_{DD}(1 - n)^2}. \tag{12}$$

Using the expression for v_2 in (8), we obtain

$$t_{\text{sat}} = \tau + \frac{2C_L n}{k_N V_{DD}(1 - n)^2} - \frac{1}{3s}(1 - n). \tag{13}$$

The time delay from saturation to $V_{DD}/2$ can now be calculated from (11), yielding a delay of

$$\Delta t = \frac{C_L}{k_N V_{DD}(1 - n)} \ln\left(\frac{2(1 - n) - 0.5}{0.5}\right). \tag{14}$$

The total propagation delay at the 50-percent level may then be written

$$t_{dHL} = t_{\text{sat}} + \Delta t - \frac{\tau}{2} = \frac{1}{6s}(1 + 2n) + t_{dHL,\text{step}} \tag{15}$$

where $t_{dHL,\text{step}}$ is the step response delay ($s = \infty$).

The ramp response output voltage for this case is shown in Fig. 1. According to (9), the input ramp is fast compared to the output. The agreement with SPICE simulations on level 1 (neglecting the P-channel transistor) is exact.

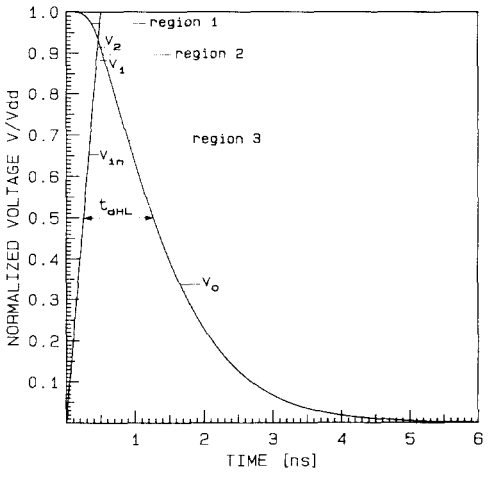

Fig. 1. Normalized output ramp response for a relatively fast input ramp (Case A). The input ramp reaches its final value ($v_{in} = 1$, $v_0 = v_2$) before the N-channel transistor leaves saturation and enters the linear region ($v_0 = v_1$); hence, $v_2 > v_1$. The fall propagation delay at the 50-percent level is also defined ($k_N V_{DD}/C_L = 1.5 \cdot 10^9\ \text{s}^{-1}$).

Case B: $v_1 > v_2$: In the second case, the N-channel transistor leaves saturation while the input voltage is still a ramp. This is true for input slopes implicitly given by, e.g., (9):

$$\frac{k_N V_{DD}}{sC_L} > \frac{6n}{(1 - n)^3}. \tag{16}$$

Region 1: $v_0 < v_1$: For output voltages $v_0 > v_1$, the solution in (7) is valid. The output voltage v_1 when the N-channel transistor leaves saturation is then determined by insertion of the saturation condition $v_1 = st - n$, i.e.,

$$v_1 = 1 - \frac{k_N V_{DD}}{6sC_L} v_1^3. \tag{17}$$

Region 2: $v_2 < v_0 < v_1$: When the input is a ramp voltage and the N-channel transistor is in the linear region, the differential equation (2) becomes

$$C_L \frac{dv_0}{dt} = -k_N V_{DD}[(st - n)v_0 - v_0^2/2]. \tag{18}$$

This differential equation with the initial conditon $v_0 = v_1$ for $t = (v_1 + n)/s$ has the solution

$$\frac{1}{v_0} = \exp\left[\frac{k_N V_{DD}}{2sC_L}(st - n)^2\right]\left[\frac{1}{v_1 \exp\left(\frac{k_N V_{DD} v_1^2}{2sC_L}\right)} + \sqrt{\frac{\pi k_N V_{DD}}{8sC_L}}\left(erf\sqrt{\left(\frac{k_N V_{DD} v_1^2}{2sC_L}\right)} - erf\sqrt{\left(\frac{k_N V_{DD}(st - n)^2}{2sC_L}\right)}\right)\right]. \quad (19)$$

When the input voltage reaches its final value, the output voltage v_2 may be determined from (19) after insertion of $st = 1$.

Region 3: $v_0 < v_2$*:* For times larger than τ or, equivalently, output voltages lower than v_2, the output voltage may be derived from

$$t - \tau = \frac{C_L}{k_N V_{DD}(1 - n)}\left[\ln\left(\frac{2(1 - n) - v_0}{v_0}\right) - \ln\left(\frac{2(1 - n) - v_2}{v_2}\right)\right]. \quad (20)$$

The ramp response output voltage for this case is shown in Fig. 2. The agreement with SPICE simulations on level 1 is exact.

The fall propagation delay t_{dHL} for these input slopes is then

$$t_{dHL} = t(0.5) - \frac{\tau}{2} \quad (21)$$

where $t(0.5)$ is the time when the output voltage reaches the 50-percent level. $t(0.5)$ is given by (19) if $v_2 < 0.5$ and by (20) if $v_2 > 0.5$. The voltages v_1 and v_2, indicating the different regions of operation for the N-channel transistor, are plotted in Fig. 3. Also included are the different regions of operation for the P-channel transistor.

Similar expressions may be derived for negative input ramps, when the output is going high, by solving the corresponding differential equations for the P-channel transistor.

The analytical expression for the propagation delay is shown in Fig. 4 as a function of input ramp rise time together with SPICE simulation data with different ratios β of the P-channel to N-channel widths as parameter. As may be seen, a straight line is a very good approximation even for reasonably slow input ramps. In most cases, the expression in (15) may be used as a good approximation. Actually, the approximation is better when the influence of the P-channel transistor is included than when it's neglected as in the analytical expression.

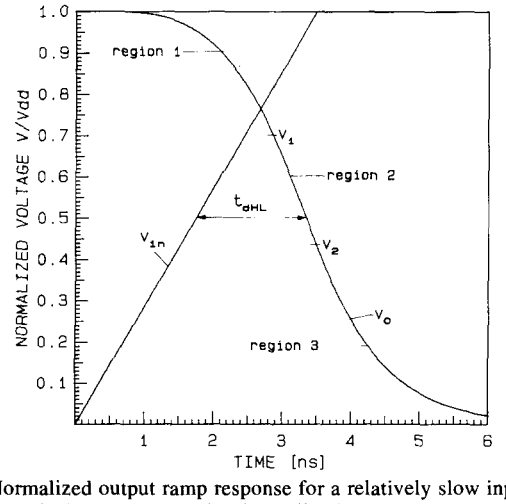

Fig. 2. Normalized output ramp response for a relatively slow input ramp (Case B). The input ramp reaches its final value ($v_{in} = 1$, $v_0 = v_2$) with the N-channel transistor already in the linear region ($v_0 = v_1$); hence, $v_2 < v_1$. The fall propagation delay at the 50-percent level is also defined ($k_N V_{DD}/C_L = 1.5 \cdot 10^9\ s^{-1}$).

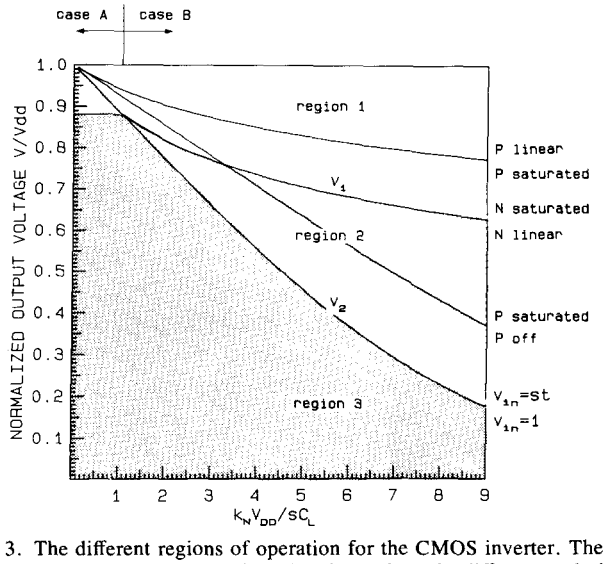

Fig. 3. The different regions of operation for the CMOS inverter. The v_1 and v_2 curves determine the shaded regions where the different analytical equations for the output ramp response are valid as indicated by the region number.

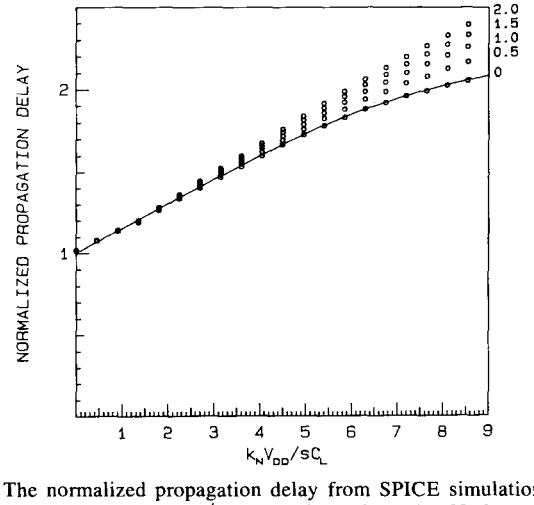

Fig. 4. The normalized propagation delay from SPICE simulations (circles) as a function of $k_N V_{DD}/sC_L$ with the P-channel to N-channel width ratio β, as parameter together with the analytical expression (solid line).

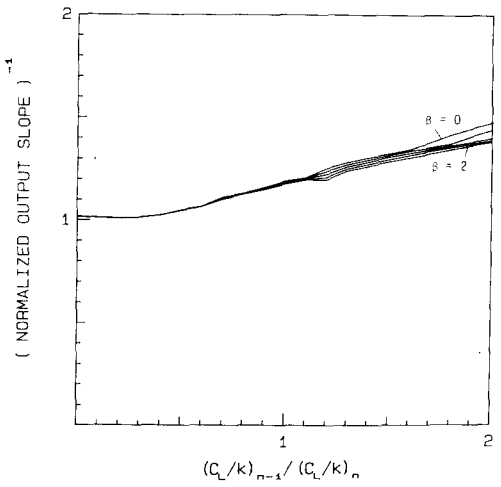

Fig. 5. The inverse of the normalized output slope at the 50-percent level from SPICE simulations with the P-channel to N-channel width ratio β, as parameter (β = 0, 0.5, 1.0, 1.5, 2.0). The figure illustrates that the output slope for a certain inverter with a certain load is almost constant independent of the input load and the input waveform as long as it is a characteristic input waveform. The normalization is made with respect to the analytical value.

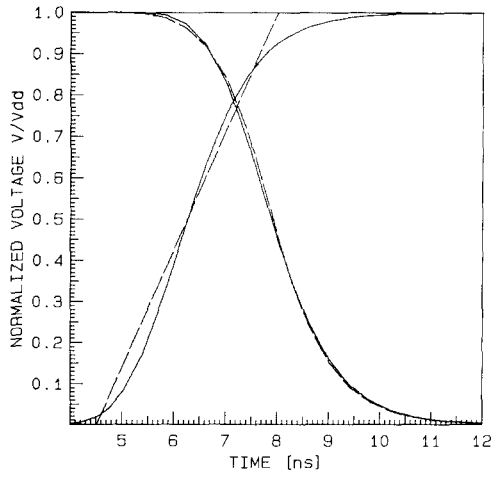

Fig. 6. Comparison between the ramp response and the characteristic waveform response from SPICE simulations on a symmetric inverter with the same P-channel and N-channel widths (β = 1). The best fit input ramp approximation has a slope approximately 70 percent of the characteristic input waveform slope at the 50-percent level ($k_N V_{DD}/C_L = 1.5 \cdot 10^9\ s^{-1}$).

C. Matching Between Ramp Response and Response of Characteristic Input Waveform

In real circuit applications, the input waveforms to the inverter is not a ramp but the output waveform from a previous inverter in the circuit. Then, instead of looking at the propagation delay as a function of the input ramp rise time, we would prefer to regard it as a function of the C_L/k ratio of the preceding stage or, equivalently, as a function of the step-response delay of the preceding stage.

The inverse of the output slope of an inverter for a certain output voltage, e.g., $V_{DD}/2$, may then be written

$$1/s = \frac{C_L}{k_N} B \tag{22}$$

where B is dependent on the input waveform. In Fig. 5, we have plotted the inverse of the output slope at the 50-percent level for different characteristic input waveforms. The characteristic waveform is the definite waveform towards which the waveform converges in a series of identical inverters [6]. In our simulations, the characteristic waveforms are varied by changing the capacitive loading of the preceding inverters relative to the loading of the studied inverter. As may be seen, the inverse of the slope is approximately constant indicating that B may be regarded as a constant independent of the input waveform. Once the value of B has been found to be constant, it may be redefined to instead represent an average slope of the characteristic waveform. This average slope may then be used as an input to the next stage. In Fig. 6, we have compared the ramp response with the characteristic waveform response in SPICE simulations and found that the best value of B corresponds to a slope 70-percent of the value at $v_0 = 0.5$.

By choosing the proper value of B and inserting into (15) $1/s$ from (22) and the step-response delay according to (3), the fall delay of the nth inverter stage in a series of inverters may now be written, remembering that the input delay is determined by a P-channel transistor, as

$$\begin{aligned}(t_{dHL})_n &= \left(\frac{C_L}{k_N}\right)_n A_N + \left(\frac{C_L}{k_P}\right)_{n-1} B_N \\ &\equiv (t_{dHL,\,\mathrm{step}})_n + \frac{B_N}{A_P}(t_{dLH,\,\mathrm{step}})_{n-1}\end{aligned} \tag{23}$$

where $B_N = (1 + 2n)B/6$. A typical value for B_N is 0.21. Essentially, the ramp response propagation delay is the sum of the step-response delay and a certain fraction (B_N/A_P) of the step-response delay of the preceding stage. Similar empirical models have been used in different timing simulators, either as analytical expressions [7] or as look-up tables [8].

Replacing the P-channel transistor k_P with

$$k_P = \beta \frac{\mu_p}{\mu_n} k_N \tag{24}$$

where μ_p and μ_n are the hole and electron mobilities, respectively, and β is the ratio between the P- and N-channel widths, the delay may be rewritten as

$$t_{dHL} = A_N \left(\frac{C_L}{k_N}\right)_n + \frac{B_N}{\beta \dfrac{\mu_p}{\mu_n}} \left(\frac{C_L}{k_N}\right)_{n-1}. \tag{25}$$

The fall propagation delay of a CMOS inverter, as simulated by SPICE with characteristic input waveforms [6], is plotted as a function of $(C_L/k_N)_{n-1}/(C_L/k_N)_n$ in Fig. 7 together with the ramp response delay as a function of $k_N V_{DD}/sC_L$. Effectively, what we do by choosing the value of B is to determine the relation between the upper

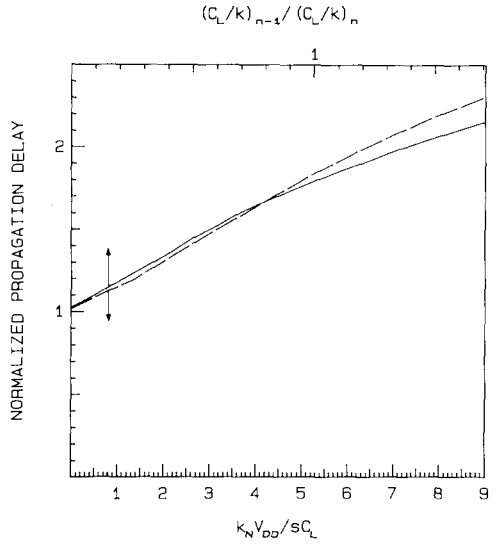

Fig. 7. The ramp-response propagation delay as a function of $k_N V_{DD}/sC_L$ (dashed) together with the characteristic waveform propagation delay as a function of $(C_L/k)_{n-1}/(C_L/k)_n$. The constant B has been chosen for best fit between the upper and lower scales. The curves are from SPICE simulations but the normalization is made with respect to the analytical step-response delay.

and lower scales for the best fit over the most important range.

In the same manner, the rise propagation delay may be written as

$$t_{dLH} = \frac{A_P}{\beta \dfrac{\mu_p}{\mu_n}} \left(\frac{C_L}{k_N}\right)_n + B_P \left(\frac{C_L}{k_N}\right)_{n-1} \tag{26}$$

where B_P is a constant equivalent to B_N.

The average propagation delay is then taken as

$$\begin{aligned} t_d &= \frac{1}{2}\left(A_N + \frac{A_P}{\beta \dfrac{\mu_p}{\mu_n}}\right)\left(\frac{C_L}{k_N}\right)_n \\ &\quad + \frac{1}{2}\left(B_P + \frac{B_N}{\beta \dfrac{\mu_p}{\mu_n}}\right)\left(\frac{C_L}{k_N}\right)_{n-1} \\ &= a\left(\frac{C_L}{k_N}\right)_n + b\left(\frac{C_L}{k_N}\right)_{n-1}. \end{aligned} \tag{27}$$

In this equation, $a(C_L/k_N)_n$ is the average step response propagation delay of the inverter stage itself, while $b(C_L/k_N)_{n-1}$ is a weighted fraction (b/a) of the average step-response propagation delay of the preceding stage.

Some typical values, for $n = -p = 0.12$, $V_{DD} = 5$ V, and $\beta\ \mu_p/\mu_n = 1$ at the 50-percent level, are $a = 0.27$ and $b = 0.21$. Typically, the input dependent delay then contributes with 75 percent of the step-response delay for an inverter loaded with an identical inverter at both input and output.

III. CMOS Power Dissipation

When the output voltage response for a positive input voltage ramp was calculated in the previous section, the P-channel transistor current was neglected. Although that means that we have neglected the short-circuit power dissipation when both transistors are conducting, it is an acceptable approximation as long as the short-circuit power dissipation is small compared to the power needed to charge the capacitor. We may then use the output voltage response to approximately calculate this power dissipation.

The P-channel transistor current is given by

$$I_p = \begin{cases} \dfrac{k_p}{2} V_{DD}^2 (v_{\text{in}} - 1 - p)^2 \\ \qquad \text{for } v_{\text{in}} - p > v_0 \\ k_p V_{DD}^2 [(v_{\text{in}} - 1 - p)(v_0 - 1) - (v_0 - 1)^2/2] \\ \qquad \text{for } v_{\text{in}} - p < v_0 \end{cases} \tag{28}$$

The input voltage is $v_{\text{in}} = st$ and the P-channel transistor is conducting as long as $v_{\text{in}} < 1 + p$. The input voltage v_{p1}, when the P-channel transistor is entering the linear region, is determined by $v_{p1} - p = v_0$ and

$$v_0 = 1 - \frac{k_N V_{DD}}{6sC_L}(v_{p1} - n)^3. \tag{29}$$

The short-circuit energy dissipation per transition may then be written

$$\begin{aligned} P &= V_{DD} \int I\, dt \\ &= k_p V_{DD}^3 \Bigg[\int_{n/s}^{v_{p1}/s} [(v_{\text{in}} - 1 - p)(v_0 - 1) \\ &\quad - (v_0 - 1)^2/2]\, dt + \int_{v_{p1}/s}^{(1+p)/s} (v_{\text{in}} - 1 - p)^2\, dt \Bigg] \end{aligned} \tag{30}$$

where the P-channel transistor is linear and the N-channel transistor saturated in the first integral and the P-channel transistor is saturated in the second integral. The expression for v_0 in (7) may then be inserted and the integration yields

$$\begin{aligned} P = C_L V_{DD}^2 d \Bigg[&\frac{d}{24}(1 - n + p)(v_{p1} - n)^4 \\ &- \frac{d}{30}(v_{p1} - n)^5 \\ &- \frac{d^2}{504}(v_{p1} - n)^7 + \frac{1}{6}(1 - v_{p1} + p)^3 \Bigg] \end{aligned} \tag{31}$$

where $d = k_P V_{DD}/sC_L$.

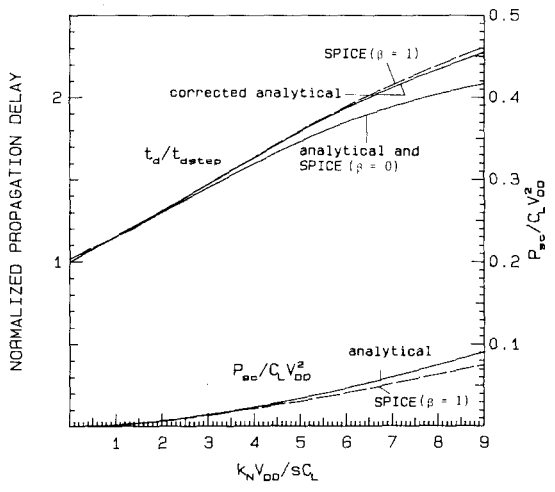

Fig. 8. Comparison between the ramp-response propagation delay for a symmetric inverter with the same P-channel and N-channel widths ($\beta = 1$) as simulated with SPICE and as an analytical expression with and without correction for the short-circuit current. This correction is derived as the short-circuit energy dissipation percentage of the capacitive energy dissipation. This percentage is also shown with the lower curves on the right-hand scale.

A similar expression may be derived for the current through the N-channel transistor while the load capacitor is charged by the P-channel transistor.

The energy dissipation per transition is plotted as a function of d in Fig. 8 for identical N- and P-channel transistors with $n = -p = 0.12$. For identical input and output slopes ($d = 3.175$), the energy dissipation is 1.5 percent of the capacitive energy dissipation. SPICE simulations for ramp input voltages are shown for comparison.

Since the short-circuit energy dissipation is directly proportional to the average short-circuit current and this current is not contributing to the discharge of the load capacitor, we approximately know with how many percent we, on the average, have overestimated the discharging current when we calculated the propagation delay in the previous section. As long as this percentage is small, it is reasonable to believe that the propagation delay is underestimated by the same percentage. This is also shown in Fig. 8.

Veendrick [9] has shown that the short-circuit energy dissipation per cycle in a symmetrical CMOS inverter ($k_N = k_P$) without capacitive load is

$$P = \frac{k_N V_{DD}^3}{12s}(1 - 2n)^3. \tag{32}$$

Our expression above with $d = 3.175$ for identical in- and out-slopes gives a value that is about 30 percent of the maximal short-circuit power dissipation as calculated by Veendrick.

IV. CMOS Load Capacitances

A. Linearized Capacitance Model

Assuming that all the capacitances in the inverter are linearized to constant values, let's relate them to the gate

TABLE I
Typical SPICE Parameters for Capacitive Load Calculations [11]

	N-channel	*P-channel*	
TOX	50	50	[nm]
CGSO, CGDO	3.0 E-10	3.0 E-10	[F/m]
CGBO	1.0 E-9	1.0 E-9	[F/m]
CJ	0.3 E-3	0.2 E-3	[F/m^2]
CJSW	0.5 E-9	0.4 E-9	[F/m^2]

capacitance C_{gn} of the N-channel transistor. The P-channel gate capacitance may be written

$$C_{gp} = \delta_1 \beta C_{gn} \tag{33}$$

where δ_1 is a process parameter and β is the channel width ratio between the P- and the N-channel transistors. Usually, both transistors have the same gate oxide thickness and the same channel length so that $\delta_1 = 1$.

In similarity with Kanuma [10], the output capacitance of the N-channel transistor may be written

$$C_n = \gamma C_{gn} \tag{34}$$

and the output capacitance of the P-channel transistor may be similarly written

$$C_p = \gamma \delta_2 \beta C_{gn} \tag{35}$$

where γ and δ_2 are process and layout dependent parameters.

The load capacitance C_L of an inverter with a fan-out of n similar inverters may then be written

$$C_L = C_{gn}[(1 + \delta_1\beta)n + (1 + \delta_2\beta)\gamma]. \tag{36}$$

With the inverter input gate capacitance C_g equal to the sum of the N- and P-channel gate capacitances

$$C_g = C_{gn} + C_{gp} = C_{gn}(1 + \delta_1\beta) \tag{37}$$

the output capacitance is given by

$$C_L = C_g(n + g\gamma) \tag{38}$$

where

$$g = (1 + \delta_2\beta)/(1 + \delta_1\beta). \tag{39}$$

Hence, $g\gamma$ is the ratio between the intrinsic output capacitance and the input gate capacitance of the inverter.

The typical SPICE parameters related to capacitive loads are shown in Table I. Using these values, we may calculate

$$\delta_1 = 1$$

$$\delta_2 = 0.72$$

$$\gamma_1 = 1.57$$

and

$$g\gamma = 1.35 \quad \text{for} \quad \beta = 1.$$

Here, the value of $g\gamma$ may be reduced to about half when large multiple-gate transistors are used and two gates are sharing the same collector contact, e.g., in output buffers.

Using (27), the propagation delay for an inverter with a fan-out of n identical inverters and a fan-out of m identical inverters in the preceding stage (including the inverter itself) may now be written

$$t_d = \frac{C_g}{k_N} a\left[(n + g\gamma) + \frac{b}{a}(m + g\gamma)\right]. \quad (40)$$

The propagation delay may be divided into three different delays. These are the intrinsic delay

$$t_{di} = \frac{C_g}{k_N}(a + b)g\gamma \quad (41)$$

the fan-out dependent delay

$$t_{d0} = \frac{C_g}{k_N} an \quad (42)$$

which is usually specified as a delay per unit external capacitance load in custom-design cell libraries, and finally the input load dependent delay

$$t_{\text{din}} = \frac{C_g}{k_N} bm \quad (43)$$

which is usually forgotten in cell library specifications but which may add important contributions to the total delay.

B. Optimum β for Minimum Delay

The analytical expression for the propagation delay as formulated in (40) may also be used to find the optimum β for minimum delay. For a given, standard-size N-channel transistor, we may look for optimum β by taking the derivative $dt_d/d\beta$, remembering that C_g, a, b, and g are all β-dependent. Optimum β is then given by

$$\beta_m = \sqrt{\frac{\mu_n}{\mu_p}\frac{A_P(n + \gamma) + B_N(m + \gamma)}{A_N(\delta_1 n + \delta_2\gamma) + B_P(\delta_1 m + \delta_2\gamma)}}. \quad (44)$$

As expected for $A_P = A_N$, $B_N = B_P$, $\delta_1 = \delta_2 = 1$, and $n = m$, we get

$$\beta_m = \sqrt{\frac{\mu_n}{\mu_p}}.$$

The propagation delay is plotted as a function of β in Fig. 9. As may be seen, there is less than 10 percent to gain in propagation delay by optimizing β instead of using standard-size N- and P-channel transistors with $\beta = 1$.

V. Buffer Speed Optimization

In order to drive large off-chip load capacitors with a minimum of propagation delay, it is necessary to use an output buffer consisting of a number of CMOS inverters with gradually increasing driving capability according to Fig. 10. The tapering factor f_n of the nth inverter stage is defined as how much the input capacitance is increased in the following stage, i.e.,

$$f_n = (C_g)_{n+1}/(C_g)_n \quad (45)$$

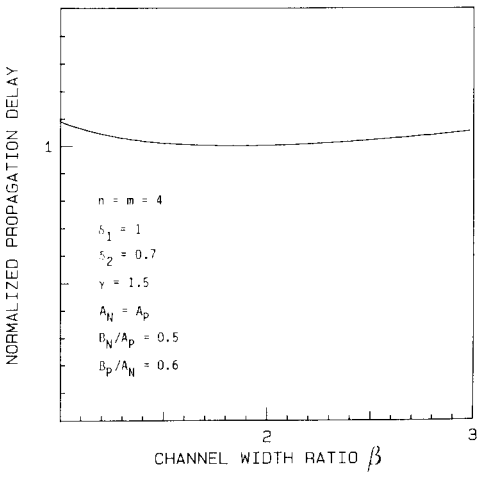

Fig. 9. The propagation delay (normalized with respect to the minimum delay) as a function of the P-channel to N-channel width ratio β.

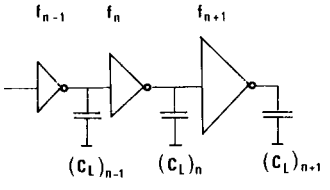

Fig. 10. A CMOS buffer consisting of a series of inverters with gradually increasing driving capability.

where $(C_g)_{n+1}$ and $(C_g)_n$ are the inverter input capacitances for the $n + 1$th and nth stages, respectively. If all the inverters have the same value of β, this definition means that the driving capability is also increased with the tapering factor

$$f_n = (k)_{n+1}/(k)_n \quad (46)$$

where $(k)_{n+1}$ and $(k)_n$ are the transistor constants of the respective stages. With constant β, this relation is valid for both the N- and P-channel transistors. These two equations then yield

$$\left(\frac{C_g}{k}\right)_{n+1} = \left(\frac{C_g}{k}\right)_n = \left(\frac{C_g}{k}\right)_{n-1} = \cdots. \quad (47)$$

The load capacitance of the nth stage may now be written, according to (38), as

$$(C_L)_n = (f_n + g\gamma)(C_g)_n. \quad (48)$$

The average propagation delay of the nth inverter stage may now in the same way be written according to (40) as

$$(t_d)_n = \frac{C_g}{k_N}[a(f_n + g\gamma) + b(f_{n-1} + g\gamma)] \quad (49)$$

where C_g/k_N is the inverter input gate capacitance to N-channel transistor constant ratio for any of the inverters in the buffer chain.

A. "Infinite Buffer"

Now, assuming a constant tapering factor f, the propagation delay may be written as

$$(t_d)_n = \tau_0(f + g\gamma) \tag{50}$$

where $\tau_0 = C_g/k_N(a + b)$. If the buffer chain consists of N stages, the total delay may be written

$$t_B = \tau_0 N(f + g\gamma) \tag{51}$$

where the tapering factor is assumed to be constant such that, by definition, $f^N = Y$, where $Y = C_L/C_0$ is the ratio between the external load capacitance C_L of the last (Nth) inverter and C_0, the gate capacitance of the first inverter. Eliminating N in (51), using $N = \ln Y/\ln f$, yields a buffer delay

$$t_B = \tau_0 \ln Y(f + g\gamma)/\ln f. \tag{52}$$

Looking for optimum f_o for minimum delay, we take the derivative

$$\frac{dt_B}{df} = \tau_0 \ln Y\left[\frac{1}{\ln f} - \frac{f + g\gamma}{f(\ln f)^2}\right] \tag{53}$$

with a minimum for f_o implicitly determined by

$$f_o = e^{(g\gamma + f_o)/f_o}. \tag{54}$$

The optimum tapering factor f_o as a function of $g\gamma$ is plotted in Fig. 11. As may be seen, $g\gamma = 0$ yields $f_o = e$, which is according to Mead and Conway [5] who have done this optimization neglecting the intrinsic load capacitance of the inverter. A similar optimization has also been done by Kanuma [10], but instead of plotting f_o, he has plotted $a = 1/\ln f_o$ without clearly indicating that a is related to the optimum tapering factor, and by Nemes [7] who has plotted f_o as a function of $g\gamma/(1 + g\gamma)$. The buffer propagation delay as a function of f with $g\gamma$ as parameter is shown in Fig. 12.

The optimum tapering factor for use in an output buffer design may be regarded as a constant for any given process, with its design rules, and the design style being used. The designer knows the optimum tapering factor of his/her process and may then easily determine the optimum number of inverters N_o in the output buffer from

$$N_o = \ln Y/\ln f_o. \tag{55}$$

However, the number of inverters N can only be an integer number. We then have to check whether $N = r$ or $N = r + 1$, where $r < N_o < r + 1$, gives the shortest buffer delay.

Using $f = (Y)^{1/N}$ to eliminate f in (51) above, the delay may be rewritten as

$$t_B = \tau_0\left[N\left((Y)^{1/N} + g\gamma\right)\right]. \tag{56}$$

As described by Kanuma [10], the optimum number of inverters is then $N = r$ if

$$r\left[(Y)^{1/r} + g\gamma\right] < (r + 1)\left[(Y)^{1/(r+1)} + g\gamma\right] \tag{57}$$

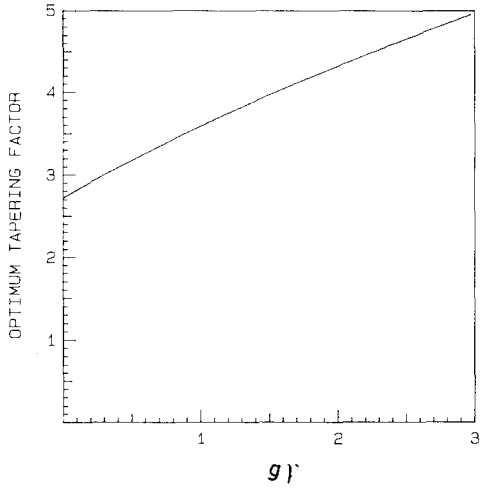

Fig. 11. Optimum tapering factor f_o as a function of $g\gamma$, the ratio between the intrinsic output load capacitance and the input gate capacitance of the inverter.

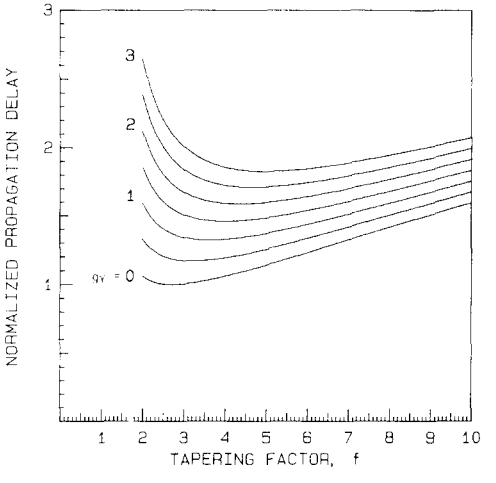

Fig. 12. The propagation delay (normalized with respect to the minimum delay at $g\gamma = 0$) as a function of the tapering factor f, with $g\gamma$ as parameter for a CMOS buffer.

and $N = r + 1$ if

$$r\left[(Y)^{1/r} + g\gamma\right] > (r + 1)\left[(Y)^{1/(r+1)} + g\gamma\right]. \tag{58}$$

However, rarely more than 5 percent will be lost in propagation delay if $N = r$ is chosen instead of $N = r + 1$ in the second case, but considerable savings in area may be achieved as discussed in the next section.

Since the chosen number of inverters may not be exactly equal to the optimum number yielding the optimum tapering factor f_o, the actual tapering factor should be adjusted to

$$f = (Y)^{1/N} \tag{59}$$

B. "Finite Buffer" with Integer Number of Inverters

So far the step-response and ramp-response buffer optimization has yielded the same result. The fact that the capacitive loading of the input of an inverter increases the delay of that inverter stage will, however, be important if

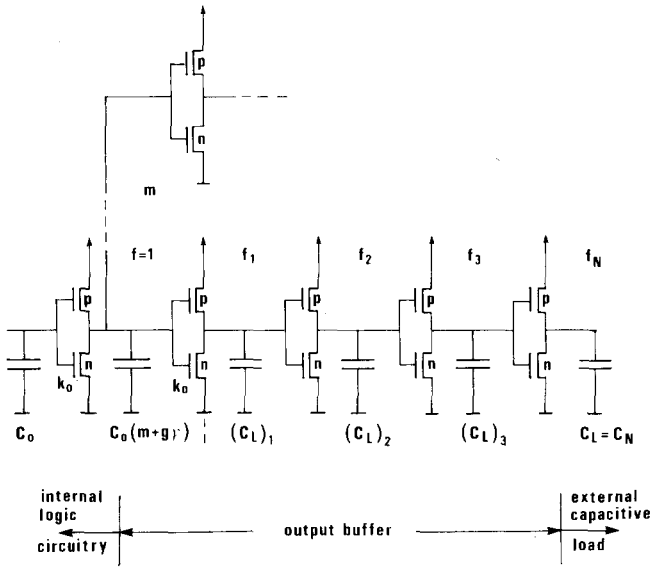

Fig. 13. A CMOS output buffer together with part of the internal circuitry.

we look more carefully at the first and the last buffer stages. An output buffer is shown in Fig. 13 together with part of the internal circuitry. The internal circuitry consists of a number of standard-size inverters. In Fig. 13, the last internal inverter is loaded with m similar inverters, one of these being the first inverter stage in the buffer. This means that the tapering factor between the first inverter in the buffer and the last inverter in the internal circuitry is m instead of f. Using (40), the delay of the first stage is

$$t_{d1} = \frac{C_0}{k_0} a\left[f_1 + g\gamma + \frac{b}{a}(m + g\gamma)\right] \tag{60}$$

where f_1 is the tapering factor for the first stage and C_0 and k_0 are the standard-size inverter input capacitance and standard-size N-channel transistor constant, respectively.

The load of the Nth inverter is the external capacitor C_L. The last stage is different with respect to the other inverter stages since its output waveform will not affect the delay of any succeeding stages. This may be a reason to believe that the last stage should have a different tapering factor f_N. The delay of the last stage may then be written

$$t_{dN} = \frac{C_0}{k_0} a\left[f_N + g\gamma + \frac{b}{a}(f + g\gamma)\right]. \tag{61}$$

The total delay of the output buffer where all stages, except the first and the last, have the same tapering factor f, may then be written

$$t_B = \tau_0\left[\frac{b}{a+b}(m + g\gamma) + f_1 + g\gamma + (N-2)(f + g\gamma) + \frac{a}{b+a}(f_N + g\gamma)\right]. \tag{62}$$

TABLE II
Y-REGIONS WHERE 1 TO 5 BUFFER STAGES GIVE SHORTEST PROPAGATION DELAY

N	Mead, Conway [5] $g\gamma = 0$	"Infinite buffer" $g\gamma = 1$	"Finite buffer" $g\gamma = 1$, $b/a = 0.75$
1	-4	-6	-10
2	4-11	6-22	10-38
3	11-32	22-82	38-143
4	32-87	82-300	143-525
5	87-237	300-1086	525-1900

The relation between the tapering factors is

$$f_1 f^{N-2} f_N = Y. \tag{63}$$

In order to find the optimum f_1 and f_N related to f for any given number of inverters, we let the derivatives equal zero yielding $f_1 = f$ and $f_N = (1 + b/a) f$. This means that for minimum delay, at the 50-percent voltage level, the last inverter stage should have a larger tapering factor than the rest of the inverter stages in the buffer. This result is a consequence of the delay dependence of the input waveform. If this dependence were removed, by letting $b = 0$, the tapering factor should be constant for all stages, including the last stage.

Using these optimum tapering factors, the buffer delay in (62) may be rewritten

$$t_B = \tau_0\left[N(f + g\gamma) + \frac{mb}{a+b}\right]. \tag{64}$$

This expression is very similar to the delay in (51) and derivation yields the same optimum tapering factor f_o as before!

In Table II, we have shown the Y-regions where 1, 2, 3, 4, and 5 buffer inverter stages are most effective for the two cases of our analysis, together with the analysis of Mead and Conway [5]. Typical values of $g\gamma = 1$ and $b/a = 0.75$ have been chosen for our analysis.

C. Buffer with Different β in the Last Inverter Stage

Another interesting case is when the output inverter must have equal rise and fall times so short that a tapering factor smaller than the optimum tapering factor must be used in the last stage. In this case, the N- and P-channel transistors in the last stage must have equal driving capabilities, i.e., $\beta_N = \mu_n/\mu_p$, even if other vlaues of β are used in the internal circuitry and the other buffer stages. In this case, the buffer delay may be written

$$t_B = \frac{1}{2}\frac{C_0}{k_0}\left[(a+b)(f + g\gamma)(N-1) + (f_N + g\gamma)\frac{1 + \delta_1\beta_N}{1 + \delta_1\beta} \cdot (A_N + A_P) + m(B_P + B_N)\right] \tag{65}$$

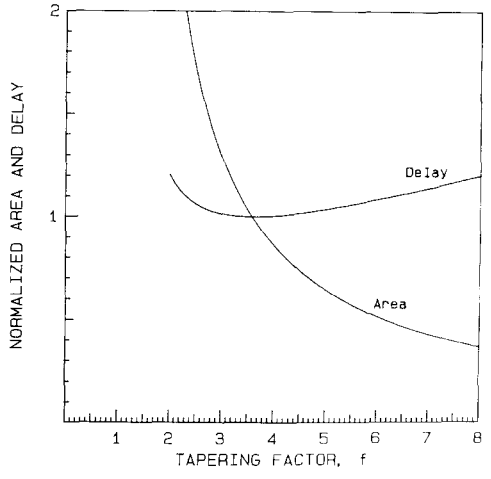

Fig. 14. The normalized buffer area plotted as a function of the tapering factor f, together with the normalized propagation delay. Both quantities are normalized with respect to their values at $f = f_0$ for $g\gamma = 1$.

where the tapering factor f in the first $N - 1$ stages may be optimized while the tapering factor f_N in the last stage is an independent constant. Thus, $f^{N-1}f_N = Y$.

Again, the optimum tapering factor f_o is given by (54), and the optimum number of inverters N_o by

$$N_o = \ln Y'/\ln f_o \tag{66}$$

where $Y' = Y/f_N$.

The tapering factor is still defined as how much the input capacitance is increased from one inverter stage to the next. However, for the last stage, with a different β_N, it is no longer a consequence that the N- or P-channel devices are f_N times larger than their predecessors. Instead, it may be shown that if the inverter input capacitance is scaled up by a factor of f, then the individual N-channel transistor is scaled up with a factor $x = f(1 + \delta_1\beta)/(1 + \delta_1\beta_N)$.

Using an integrated circuit technology with $\delta_1 = 1$ and a mobility ratio of $\mu_n/\mu_p = 2.5$ and choosing $\beta = 1$ in all but the last stage, where $\beta_N = \mu_n/\mu_p$, we find that the N-channel transistor in the last stage is scaled up by a factor of $x = 0.57$f, while the P-channel transistor is scaled up by $\beta_N x = 1.43$f.

VI. Buffer Area Optimization

So far, we have only minimized the buffer propagation delay without any area or power dissipation considerations. The area and also, in a first approximation, the power dissipation of the buffer may be considered to be proportional to the sum of the input capacitances of all the buffer stages. For an output buffer consisting of N inverter stages with a constant tapering factor f, this sum may be written

$$C_{\text{tot}} = C_0 f^{N-1}\left(1 + \frac{1}{f} + \frac{1}{f^2} + \cdots + \frac{1}{f^{N-1}}\right) \tag{67}$$

where C_0 is the input gate capacitance of the first standard-size inverter in the buffer. If we sum all the stage capacitances, the total capacitance may be written

$$C_{\text{tot}} = C_0 \frac{Y - 1}{f - 1}. \tag{68}$$

Both the area and, in a first approximation, also the total power dissipation ($C_{\text{tot}} V_{DD}^2$) is then proportional to $1/(f - 1)$. In Fig. 14, the normalized area is plotted as a function of f together with the normalized propagation delay (for $g\gamma = 1$). In this diagram, we can also see the relative time penalty when the tapering factor instead is chosen to minimize area or power dissipation. Since the propagation delay curve is rather flat, we can see that we are loosing only 10 percent in propagation delay, while more than 50 percent in area or power dissipation is saved. As an example, let's take a buffer with the optimum number of inverters $N_0 = 5$. If we instead choose $N = 4$, the tapering factor must be increased to $f_0^{5/4}$, which gives a loss of about 3 percent in propagation delay, but saves 35 percent in area. If only three inverters are chosen in the buffer, the loss in propagation delay is about 22 percent but 54 percent is saved in area.

VII. Conclusions

By presenting an analytical solution for the output response of a CMOS inverter to a ramp input signal, an improved understanding of the switching behavior of the CMOS inverter is achieved. The model is easily generalized to characteristic input waveforms yielding a propagation delay model that is far superior to the step-response model since it includes the input load dependence of the propagation delay. As a consequence, this input-load dependent delay should be added to the traditional intrinsic and fan-out dependent delays in custom-design cell libraries to improve timing calculations. The model is also ideally suited for inclusion in timing simulators.

As an example of the usefulness of the switching model, we have applied it to the problem of output buffer optimization for minimum propagation delay. It is shown that the model presents an algorithm that should be very useful as an optimizing tool in a CAD work station or silicon compiler.

Acknowledgment

The authors wish to thank Dr. D. Andersson for helpful discussions of the analytical calculations and Drs. S. Christensson and L. Lundgren for valuable criticism of the manuscript.

References

[1] M. I. Elmasry, "Digital MOS integrated circuits: A tutorial," in *Digital MOS Integrated Circuits*. New York: IEEE Press, 1981, pp. 4–27.

[2] E. Seewann, "Switching speeds of MOS inverters," *IEEE J. Solid-State Circuits*, vol. SC-15, pp. 246–252, Apr. 1980.

[3] T. Tokuda, K. Okazaki, K. Sakashita, I. Ohkura, and T. Enomoto, "Delay-time modeling for ED MOS logic LSI," *IEEE Trans. Computer-Aided Design*, vol. CAD-2, pp. 129–134, July 1983.

[4] R. J. Bayruns, R. L. Johnston, D. L. Fraser, Jr., and S-C. Fang, "Delay analysis of Si NMOS Gbit/s logic circuits," *IEEE J. Solid-State Circuits*, vol. SC-19, pp. 755–764, Oct. 1984.

[5] C. Mead and L. Conway, *Introduction to VLSI-Systems*. Reading, MA: Addison-Wesley, 1980.
[6] J. R. Burns, "Switching response of complementary-symmetry MOS transistor logic circuits," *RCA Rev.*, vol. 25, pp. 627–661, Dec. 1964.
[7] C. Bobev, R. Sundblad, and C. Svensson, "Event-driven simulation of general CMOS circuits with delay modeling," to be published in *IEEE Trans. Computer-Aided Design*.
[8] J. Ousterhout, "A switch-level timing verifier for digital MOS VLSI," *IEEE Trans. Computer-Aided Design*, vol. CAD-4, pp. 336–349, 1986.
[9] 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.
[10] A. Kanuma, "CMOS circuit optimization," *Solid-State Electron.*, vol. 26, pp. 47–58, 1983.
[11] *Portable CMOS Design Rules for the Swedish Universities*, Lund, Sweden, 1985.
[12] M. Nemes, "Driving large capacitances in MOS LSI systems," *IEEE J. Solid-State Circuits*, vol. SC-19, pp. 159–161, 1984.

*

Nils Hedenstierna received the M.S. degree in electrical engineering in 1984 from Chalmers University of Technology, Göteborg, Sweden, and has been a member of the research and teaching staff at its Department of Solid State Electronics since then.

He is currently working towards the Ph.D. degree and his main research interest is on various aspects of CMOS VLSI design such as hierarchical design verification, design rule checking, and timing models.

Kjell O. Jeppson (S'68–M'76–SM'83) received the M.S. degree in electrical engineering in 1970 and the Ph.D. in solid-state electronics in 1977 from Chalmers University of Technology, Göteborg, Sweden.

Since 1970, he has been a member of the research and teaching staff at the Department of Solid State Electronics at Chalmers University of Technology. He was appointed Lecturer (Associate Professor) in 1979. He spent the academic year 1973/74 with Rockwell International, Anaheim, CA, and the fall semester 1985 at the Southampton University Microelectronics Centre, England. His main research interest is on MOS devices and CMOS VLSI design. He has published several papers on MNOS nonvolatile memories, narrow-channel width effects, and nonlinear diffusion in silicon. He has authored a textbook on semiconductor devices (*Halvledarteknik-komponenter och teknologi för integrerade kretsar*) and coauthored an exercise book on the same subject (both in Swedish).

Dr. Jeppson is a Technical Editor for the Swedish electronics magazine *Elteknik*. He is also a member of the ISSCC European Program Committee.