A Unified Design Methodology for CMOS Tapered Buffers

Brian S. Cherkauer, *Student Member, IEEE,* and Eby G. Friedman, *Senior Member, IEEE*

Abstract— **In this paper, the various disparate approaches to CMOS tapered buffer design are unified into an integrated design methodology. Circuit speed, power dissipation, physical area, and system reliability are the four performance criteria of concern in tapered buffers, and each places a separate, often conflicting, constraint on the design of a tapered buffer. Enhanced short-channel tapered buffer design equations are presented for propagation delay and power dissipation, as well as a new split-capacitor model of hot-carrier reliability of tapered buffers and a two-component physical area model. Each performance criterion is individually investigated and analyzed with respect to the number of stages and tapering factor, and the interaction of the four criteria is examined to develop both a qualitative and a quantitative understanding of the various design tradeoffs. The creation of process dependent look-up tables for optimal buffer design is described, and a methodology to apply these look-up tables to application-specific tapered buffers for both unconstrained and constrained systems is developed. Summarizing, the methodology described in this paper simultaneously considers the interrelated issues of circuit speed, power dissipation, physical area, and system reliability, permitting the efficient design of tapered buffers.**

I. Introduction

In CMOS integrated circuits, large capacitive loads are often encountered. These large loads occur both on-chip, where high, localized fan-out and long global interconnect lines are common, and off-chip, where highly capacitive chip-to-chip communication lines exist. In order to drive these large capacitive loads at high speeds, buffer circuits are required which must quickly source and sink relatively large currents while not degrading the performance of previous stages. In CMOS, a tapered buffer system is often used to perform this task, particularly when the load is predominantly capacitive [1], [2]. When the load is resistive, typically a long interconnect line, repeaters, a form of distributed buffer, rather than tapered buffers are used [3], [4]. This paper, however, focuses on tapered buffers. Thus, only capacitive load impedances with negligible resistance are considered.

The basic problem that the tapered buffer must solve is illustrated in Fig. 1. High output impedance logic and/or registers must drive a large capacitive load with acceptable speed. The tapered buffer is placed between the logic/registers and the large capacitive load. The tapered buffer provides a high impedance input, so as not to load down the logic/registers, and sources (sinks) high current to quickly charge (discharge) the large capacitive load. Thus, the buffer isolates the logic/registers from the load, amplifying the signal along the way.

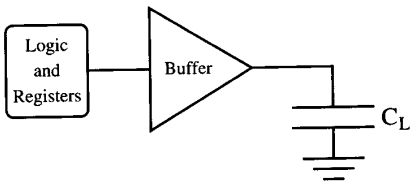

Fig. 1. Buffer used to drive capacitive load.

The tapered buffer is a well-known CMOS circuit structure. Many different approaches to tapered buffer design have been described in the literature, focusing on a variety of performance aspects. The most commonly addressed criteria in tapered buffer design are propagation delay, power dissipation, physical area, and, quite recently, circuit reliability. Previous design methods utilize analytic expressions to determine the tapering factor and the number of stages of a tapered buffer system; these parameters are the two primary variables in the design of tapered buffers. Unfortunately, the methods developed to deal with these different design constraints are quite diverse, do not deal with all four issues simultaneously, and often provide solutions which are in direct conflict. The primary objective of this paper is to unify these seemingly independent criteria by providing a single, integrated design methodology for determining an application-specific tapering factor and number of stages of a tapered buffer system necessary to drive a wide range of capacitive loads. This unification is accomplished by utilizing consistent models to derive analytic expressions for all four performance criteria. Thus, expressions with similar form are produced, permitting the various criteria to be combined into a single tapering methodology which simultaneously considers all four criteria.

In Section II, a brief background of CMOS tapered buffer design is provided. An analytic expression using the alpha-power short-channel MOSFET model for the calculation of propagation delay through a tapered buffer system is presented in Section III, permitting the development of an expression to minimize the buffer system delay. In Section IV, analytic expressions are presented for determining both the short-circuit and dynamic power dissipation of tapered buffers, and the implications of power dissipation on tapered buffer design are discussed. In Section V, an analytic expression for the physical area of a tapered buffer is presented, and the significance of

Manuscript received December 20, 1993; revised August 23, 1994. This work was supported in part by the National Science Foundation under Grant MIP-9208165.

The authors are with the Department of Electrical Engineering, University of Rochester, Rochester, NY 14627 USA.

IEEE Log Number 9408374.

area to buffer design is summarized. An analytic expression is described in Section VI which is used as a measure of hot-carrier degradation of inverter performance, a relationship which is inversely proportional to buffer reliability. This hot-carrier degradation behavior is described in terms of the reliability of tapered buffers. In Section VII, these four design criteria are unified into a single tapered buffer design strategy. The application of this unified methodology to the design of practical buffer circuits is discussed in Section VIII. Finally, some conclusions are drawn in Section IX. A pseudocode implementation of an algorithm for generating technology dependent look-up tables for optimal tapered buffer design is provided in the appendix.

II. Overview of Tapered Buffer Design

The tapered buffer structure was first proposed by Lin and Linholm in 1975 [1]. This structure consists of a series of inverters where each transistor channel width is a fixed multiple, F, larger than that of the previous inverter. The output current drive to output capacitance ratio remains fixed for each stage in the buffer, therefore each inverter stage has equal rise, fall, and delay times. Assuming a simplified capacitance model in which the interstage capacitance is directly proportional to the size of the input capacitance of the inverter, and an area model in which the area of each inverter is directly proportional to the channel widths of the transistors in the inverter, Lin and Linholm developed an analytical optimization scheme based on a "figure of merit," which is a weighted product of the delay and area requirements of a tapered buffer.

Immediately following Lin and Linholm, Jaeger proposed a modification of the optimization process which considered only speed optimization [2]. Jaeger showed that, in contrast to the assumptions used by Lin and Linholm, the total delay through a tapered buffer system is minimized when the entire system delay is considered, rather than the delay of the individual inverter stages. He further showed that the minimum system delay is achieved when the ratio between the transistor channel widths in adjacent stages, F, is exponentially tapered. F and the number of stages N, as described in [2], are defined by (1) and (2), respectively, where C_L is the load capacitance, C_y is the input gate capacitance of the minimum sized buffer stage, and W_i is the width of the devices in the ith stage of the tapered buffer system

$$F = \frac{W_i}{W_{i-1}} = e \approx 2.72 \tag{1}$$

$$N = \ln \frac{C_L}{C_y}. \tag{2}$$

Since the early work developed by Lin and Linholm and Jaeger, both published in 1975, other aspects of buffer design have been addressed, such as power dissipation [5]–[8], circuit area [7]–[11], and system reliability [12], many of which have led to significantly different results than that of Jaeger's original solution. Additionally, Jaeger's approach has been extended to include more accurate capacitance models [9], [13]–[16] and improved delay models [13], [14], [17], [18].

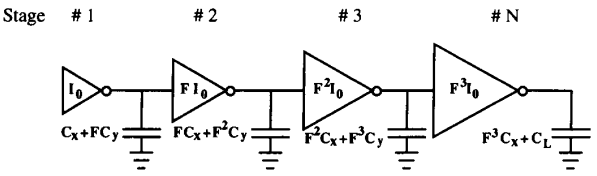

Fig. 2. The split-capacitor model of a tapered buffer [15].

The split-capacitor model, introduced implicitly by Kanuma [13], explicitly by Li, Haviland, and Tuszynski [15], and reviewed in [19], in which the input gate capacitance C_y and the output diffusion capacitance C_x of the inverter are modeled as separate capacitors, offers improved accuracy as compared to the single capacitor model utilized by Lin and Linholm and Jaeger. The split-capacitor model is illustrated in Fig. 2 and is utilized throughout this work.

In the split-capacitor model, the capacitance at the output of the ith buffer stage is included, as shown in (3). This expression will be utilized frequently in the following sections

$$C_{L_i} = F^{i-1}(C_x + FC_y). \tag{3}$$

III. Propagation Delay

In modern submicrometer CMOS fabrication technologies, short-channel effects are often quite pronounced. Therefore, an accurate and efficient short-channel transistor model is necessary when developing a buffer delay equation. The short-channel transistor model used in this paper is the alpha-power I-V relationship developed by Sakurai and Newton [20]. In this model, Sakurai and Newton show that the propagation delay (the time from the input signal reaching $V_{DD}/2$ to the output signal reaching $V_{DD}/2$) for an inverter is as shown in (4), and the slope of a straight-line approximation of the output waveform of an inverter, t_T, is shown in (6). In (4), $t_{p\mathrm{HL}}$ is the propagation delay for the output transition from high to low, and $t_{p\mathrm{LH}}$ is the propagation delay for the output transition from low to high

$$t_{p\mathrm{HL}}, t_{p\mathrm{LH}} = \left(\frac{1}{2} - \frac{1-\nu_T}{1+\alpha}\right)t_T + \frac{C_L V_{DD}}{2I_{D0}} \tag{4}$$

where

$$\nu_T = \frac{V_{\mathrm{TH}}}{V_{DD}} \tag{5}$$

$$t_T = \frac{C_L V_{DD}}{I_{D0}}\left(\frac{0.9}{0.8} + \frac{V_{D0}}{0.8V_{DD}} \ln \frac{10V_{D0}}{eV_{DD}}\right). \tag{6}$$

In the alpha-power model, I_{D0} represents the drive current of the MOS device and is proportional to W/L, V_{D0} represents the drain-to-source voltage at which velocity saturation occurs with $V_{GS} = V_{DD}$ and is a process dependent constant, and α models the process dependent degree to which velocity saturation affects the drain-to-source current and is within the range $1 \leq \alpha \leq 2$, where $\alpha = 1$ corresponds to a device operating strongly under velocity saturation, while $\alpha = 2$ represents a device where there is negligible velocity saturation. V_{DD} is the positive supply voltage, and V_{TH} is the MOS threshold voltage.

Under the assumption of symmetric inverters, i.e., inverters with equal rise and fall times, the delay through each stage of a tapered buffer is equal [1]. In addition, the rise and fall times of the signals throughout the tapered buffer system are equal, therefore straight-line approximations for the input and output signals yield equal magnitudes for all signal slopes [14]. With the alpha-power short-channel model, waveform symmetry is demonstrated since the ratio C_L/I_{D0} of each inverter stage remains constant. Thus, (4) and (6) are constant for each individual stage within the tapered buffer system. The ratio C_L/I_{D0} of the ith stage may be calculated as

$$\frac{C_{L_i}}{I_{D0_i}} = \frac{F^{i-1}(C_x + FC_y)}{F^{i-1}(I_{D0_1})} = \frac{(C_x + FC_y)}{I_{D0_1}}. \tag{7}$$

Substituting (6) and (7) into (4) yields an expression for the single stage delay of a tapered buffer system [21]. Since the delay through each stage is equal, the total delay is calculated by multiplying the single stage delay by the number of stages, N [2], [20]. This process results in an expression for the propagation delay through a tapered buffer, as given below in (8)–(10). The subscripts 'n' and 'p' in (9) and (10) designate constants describing the NMOS and PMOS devices, respectively

$$t_{\text{buffer}} = N\frac{V_{DD}(C_x + FC_y)}{I_{D0_1}}\left[\frac{K_{\text{HL}} + K_{\text{LH}}}{2}\right] \tag{8}$$

where

$$K_{\text{HL}} = \left[\left(\frac{0.9}{0.8} + \frac{V_{D0_p}}{0.8V_{DD}}\ln\frac{10V_{D0_p}}{eV_{DD}}\right)\left(\frac{1}{2} - \frac{1-\nu_{T_n}}{1+\alpha_n}\right) + \frac{1}{2}\right] \tag{9}$$

$$K_{\text{LH}} = \left[\left(\frac{0.9}{0.8} + \frac{V_{D0_n}}{0.8V_{DD}}\ln\frac{10V_{D0_n}}{eV_{DD}}\right)\left(\frac{1}{2} - \frac{1-\nu_{T_p}}{1+\alpha_p}\right) + \frac{1}{2}\right]. \tag{10}$$

Note that due to differences in $V_{\text{D0}}, V_{\text{TH}}$, and α between the NMOS and PMOS devices within each inverter, equalizing the drive current I_{D0} of both transistors does not precisely equalize the high-to-low and low-to-high propagation delays of the inverter. Therefore, an average of the high-to-low and low-to-high propagation delays is provided in (8). Equations (9) and (10) describe the dependence of the propagation delay on the individual NMOS and PMOS device parameters.

It is interesting to note that short-channel geometries do not change the form of the propagation delay through a tapered buffer system. Short-channel geometries affect the delay calculation, but not the delay optimization method. Consequently, previous delay optimization work, primarily developed for long-channel devices, is equally applicable to short-channel devices.

The number of stages, N, calculated from the split-capacitor model solution developed by Li *et al.*, in [15], is

$$N = \frac{\ln\frac{C_L}{C_y}}{\ln F}. \tag{11}$$

The choice of tapering factor, F, for minimum delay may also be calculated in the same manner as for the long-channel model, and this transcendental relationship in F is [15]

$$F[\ln(F) - 1] = \frac{C_x}{C_y}. \tag{12}$$

Previous research on the design of tapered buffer circuits has concentrated on choosing an optimal tapering factor, F, for a tapered buffer system to achieve a desired performance goal. In determining an application-specific F, the number of stages, N, can take on only positive integer values. Typically, however, expressions for propagation delay, power dissipation, physical area, and system reliability are developed as continuous functions of F. A drawback to this approach is that it disguises an inherently discrete system as a continuous system, thus greatly increasing the search space for F. Upon selecting F, it is then necessary to convert from a continuous system to a discrete system in order to physically realize the tapered buffer.

In unifying these four performance criteria, the discrete nature of N is used as a simplifying design constraint. Thus, all design equations are expressed as discrete functions of N in addition to continuous functions of F. This transformation reduces the search space for an optimal design to typically fewer than ten values and allows for a quick comparison of the circuit speed, power dissipation, physical area, and device degradation tradeoffs among the fewest possible circuit implementations.

The tapering factor, F, as a discrete function of the number of stages, N, using the Li split-capacitor model, is

$$F = \left(\frac{C_L}{C_y}\right)^{\frac{1}{N}}. \tag{13}$$

With this relationship, the propagation delay through a tapered buffer, given in (8), can be rewritten in terms of N as

$$t_{\text{buffer}} = N\frac{V_{DD}\left(C_x + \left(\frac{C_L}{C_y}\right)^{\frac{1}{N}}C_y\right)}{I_{\text{D0}_1}}\left(\frac{K_{\text{HL}} + K_{\text{LH}}}{2}\right). \tag{14}$$

This expression is normalized to remove process constants, thereby expressing delay as a function of only those variables which may be controlled during the design process. This procedure results in the normalized delay expression shown in (15), which is illustrated graphically in Fig. 3

$$\begin{aligned} t_{\text{norm}} &= \left(\frac{2I_{D0_1}}{V_{DD}(K_{\text{HL}} + K_{\text{LH}})}\right)t_{\text{buffer}}[N] \\ &= N\left(C_x + \left(\frac{C_L}{C_y}\right)^{\frac{1}{N}}C_y\right). \end{aligned} \tag{15}$$

From the shape of the graph shown in Fig. 3, it is seen that the propagation delay of a tapered buffer system as a function of N exhibits upward concavity. Therefore, a minimum delay exists for any specific C_L/C_y. The value of N which produces the minimum propagation delay for a given C_L/C_y is hereafter referred to as N_D. Examination of N_D as a function of C_L/C_y reveals that N_D increases slowly with increasing C_L/C_y. This can be seen in Fig. 3, in which $N_D = 2$ for $C_L/C_y = 10$, and N_D increases to 5 with $C_L/C_y = 1000$. These results agree with those shown in [7] and [10]. The sensitivity of N_D to C_x

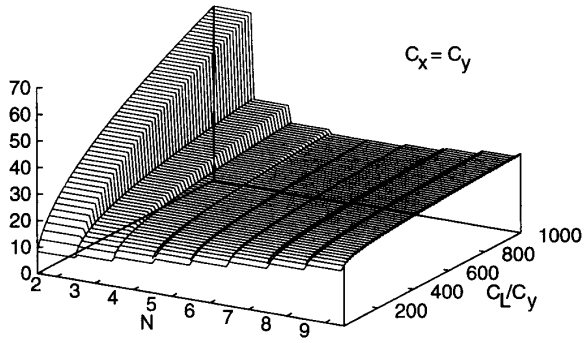

Fig. 3. Normalized propagation delay of a tapered buffer system.

within this range of C_L/C_y is small. It should be noted that the propagation delay dramatically increases when using a value of N which is much smaller than N_D. Furthermore, it is worth observing that the propagation delay increases slowly when using more stages than is delay optimal, i.e., for $N > N_D$.

IV. Power Dissipation

There are two primary mechanisms by which power is actively dissipated in the switching of a CMOS tapered buffer. The dominant component is dynamic power, in which power is dissipated by charging and discharging the load capacitance of each inverter. Dynamic power dissipation of tapered buffer systems is discussed in subsection A. Short-circuit power, the second significant power dissipation component, is dissipated during the time in which the NMOS and PMOS devices in an inverter are both on [5]. The magnitude of the short-circuit current is greatly dependent upon the shape of the input waveform driving the inverter, and a new method to compute the short-circuit power dissipation of tapered buffer systems is presented in subsection B. The integration of dynamic and short-circuit power dissipation into an expression for the total power dissipation and discussion of reduced power dissipation design in tapered buffers is presented in subsection C.

A. Dynamic Power Dissipation

Dynamic power dissipation is typically the dominant component of power dissipation in CMOS circuits. It is therefore important to closely examine the dynamic power dissipation of tapered buffer circuit structures.

Dynamic power dissipation is a well-understood phenomenon arising from the charging and discharging of node capacitances as signal lines shift logic levels. The power itself is dissipated in the channel on-resistance of the charging or discharging transistors. Each switching cycle consumes $C_L V_{DD}^2$ joules. If a switching cycle occurs with frequency f, the dynamic power dissipated in the ith stage of a tapered buffer may be expressed in the classical form

$$P_{\mathrm{Dyn}_i} = C_{Li} V_{DD}^2 f. \tag{16}$$

Assuming that the system frequency is independent of the buffer design, the switching frequency, f, in (16) may be considered constant for different tapered buffer implementations. This is also a realistic assumption, for if a system cannot tolerate a somewhat slower than minimal delay buffer, then the option for design tradeoffs does not exist. This would render moot the discussion of power dissipation as a parameter to minimize in tapered buffer design. Thus, with f considered a constant, the dynamic power dissipation for the entire tapered buffer system is the sum of the dynamic power dissipation of each stage, as shown in (17)

$$P_{\mathrm{Dyn}_{\mathrm{total}}} = V_{DD}^2 f \sum_{i=1}^{N} F^{i-1}(C_x + FC_y). \tag{17}$$

Performing the summation in (17) and choosing N as specified by (11) results in

$$P_{\mathrm{Dyn}_{\mathrm{total}}} = V_{DD}^2 f (C_x + FC_y)\left(\frac{\frac{C_L}{C_y} - 1}{F - 1}\right), \quad F > 1. \tag{18}$$

Note in (18) that dynamic power dissipation as a function of tapering exhibits no global minimum. As F increases, $P_{\mathrm{Dyn}_{\mathrm{total}}}$ continuously decreases. This phenomenon is discussed further in subsection C.

B. Short-Circuit Power Dissipation

Sakurai and Newton compute the short-circuit power dissipated in an inverter using the alpha-power model [20]. This expression is based on assumptions similar to those used by Veendrick [5], but with the alpha-power transistor model replacing the Shichman-Hodges model [22] used by Veendrick. It is assumed that the transistor which is switched from cutoff to saturation remains in saturation during the entire time short-circuit current is conducted, and that the short-circuit current waveform is mirror symmetric about a central vertical axis. The expression for short-circuit power dissipation during one switching cycle (a switching cycle is two logic level transitions, high-to-low and low-to-high) developed in [20] is presented in (19), where t_T is the input signal transition time, as previously shown in (6)

$$P_{\mathrm{SC}} = V_{DD} f t_T I_{D0} \frac{1}{(\alpha+1)} \frac{1}{2^{\alpha-1}} \frac{(1-2\nu_T)^{\alpha+1}}{(1-\nu_T)^{\alpha}}. \tag{19}$$

Substituting the expression for input waveform transition time in (6) into (19) yields the following expression for the short-circuit power dissipated in the ith stage of a buffer

$$P_{\mathrm{SC}_i} = V_{DD}^2 f \left(\frac{0.9}{0.8} + \frac{V_{D0}}{0.8 V_{DD}} \ln \frac{10 V_{D0}}{e V_{DD}}\right) \frac{1}{(\alpha+1)} \times \frac{1}{2^{\alpha-1}} \frac{(1-2\nu_T)^{\alpha+1}}{(1-\nu_T)^{\alpha}} C_{L_i}. \tag{20}$$

Summing (20) for each stage of the buffer system and consolidating the constants provides the total short-circuit power dissipation, $P_{\mathrm{SC}_{\mathrm{total}}}$, for one switching cycle of a tapered buffer

$$P_{\mathrm{SC}_{\mathrm{total}}} = K_{P_{\mathrm{SC}}} V_{DD}^2 f \sum_{i=1}^{N} F^{i-1}(C_x + FC_y) \tag{21}$$

where

$$K_{P_{SC}} = \left(\frac{0.9}{0.8} + \frac{V_{D0}}{0.8V_{DD}} \ln \frac{10V_{D0}}{eV_{DD}}\right) \times \frac{1}{(\alpha+1)} \frac{1}{2^{\alpha-1}} \frac{(1-2\nu_T)^{\alpha+1}}{(1-\nu_T)^{\alpha}}. \quad (22)$$

Performing the summation, (21) may be expressed as

$$P_{SC_{total}} = K_{P_{SC}} V_{DD}^2 f(C_x + FC_y)\left(\frac{\frac{C_L}{C_y} - 1}{F-1}\right), \quad F > 1. \quad (23)$$

This expression for short-circuit power dissipation is of the same form as that of dynamic power dissipation shown in (18). Thus, short-circuit power dissipation, like dynamic power dissipation, has no global minimum but is continually decreasing with F increasing. This result is somewhat counter-intuitive, as larger F leads to slower signal transitions, and thus more short-circuit current. However, due to the interdependence of F and N, increasing F simultaneously reduces N. Though the short-circuit power dissipation of each stage may increase with F increasing, there are fewer stages, and the resulting total short-circuit power dissipation of the tapered buffer system decreases. This result is consistent with the findings of Veendrick [5]. In subsection C, (18) and (23) are combined, and the total power dissipation of a tapered buffer system is discussed.

C. Total Power Dissipation

Assuming that the static power dissipated due to leakage current is negligible compared with the dynamic and short-circuit power dissipation, the total power dissipated in a tapered buffer system, P_{total}, may be expressed as the sum of the individual dynamic and short-circuit power dissipation components, described in (18) and (23), respectively, and is

$$P_{total} = V_{DD}^2 f(1+K_{P_{SC}})(C_x + FC_y)\left(\frac{\frac{C_L}{C_y} - 1}{F-1}\right), \quad F > 1. \quad (24)$$

Note that (24) has no global minimum. It is a continuously decreasing function of F. This demonstrates that with increasing F, the power dissipation of the tapered buffer system decreases. Given the dependent nature of N on F, as shown in (11), as F approaches infinity, N approaches zero, and P_{total} approaches zero. Intuitively, the tapered buffer that dissipates the least power is the buffer which consists of no buffer stages. However, this limit is not useful from a design standpoint, and thus designing a tapered buffer system with minimum power as a single constraint is not a meaningful process, unlike designing a tapered buffer system for minimum delay.

Equation (24), however, demonstrates that using larger values of F, and consequently fewer buffer stages, reduces both short-circuit and dynamic power dissipation within the buffer. This conclusion is similar to that drawn by Veendrick, although only short-circuit power dissipation and long-channel devices are addressed in [5]. Therefore, the observation described in [5] that the tapered buffer with the minimum power consists of the fewest stages necessary to meet any remaining design criteria still remains when dynamic power dissipation and short-channel devices are both considered.

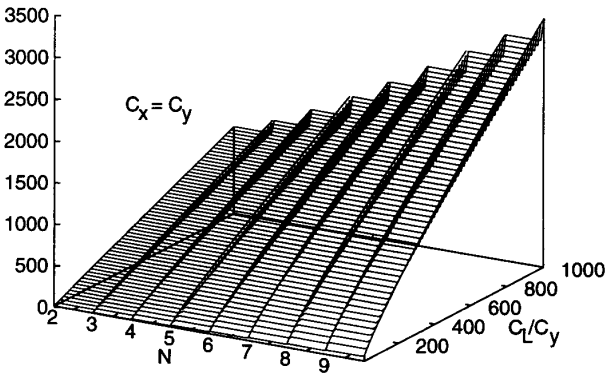

Fig. 4. Normalized total power dissipation of a tapered buffer system.

Examining the first derivative of (24) with respect to F provides further insight into the sensitivity of the total power dissipation to variations in the tapering factor, F. This relationship is shown below

$$\frac{d(P_{total})}{dF} = -V_{DD}^2 f(1+K_{P_{SC}}) \times \frac{(C_x + C_y)\left(\frac{C_L}{C_y} - 1\right)}{(F-1)^2}, \quad F > 1. \quad (25)$$

Note in (25) that the sensitivity of the power dissipation to F exhibits a $\frac{1}{(F-1)^2}$ dependence. Thus, the sensitivity of the power dissipation diminishes as F is increased. A practical limit to reducing power dissipation is reached when F grows large enough to reduce N to a single stage, as that is the minimum number of stages necessary to realize a tapered buffer system.

Expressing total power dissipation, as given in (24), as a discrete function of the number of stages results in (26). Assuming that switching frequency, f, is independent of buffer design, as discussed previously in subsection A, a normalized version of (26) is depicted in Fig. 4

$$P_{total} = V_{DD}^2 f(1+K_{P_{SC}})\left(C_x + \left(\frac{C_L}{C_y}\right)^{\frac{1}{N}} C_y\right) \times \left(\frac{\frac{C_L}{C_y} - 1}{\left(\frac{C_L}{C_y}\right)^{\frac{1}{N}} - 1}\right). \quad (26)$$

Unlike Fig. 3, in which the normalized propagation delay of a tapered buffer system is displayed, the total power dissipation graph shown in Fig. 4 depicts no local minima. The shape of the graph demonstrates a steady increase in power dissipation for increasing values of N. It may be observed that the propagation delay penalty incurred by reducing N by many stages below N_D is not mitigated by a substantial reduction in power dissipation, and therefore, minimal incentive exists to decrease N far below N_D.

V. Physical Area

The relationship between the physical area of a tapered buffer and the tapering factor closely tracks the relationship between dynamic power dissipation and the tapering factor. This similarity is due to the strong dependence of dynamic power dissipation on the transistor gate oxide capacitance, the magnitude of which is defined by the active area of each device. The primary difference between the dependence on tapering of dynamic power dissipation and physical area is that the total area overhead required to construct an inverter is significant in comparison with the active area, which is the component that scales with tapering factor [8]. Thus, only a certain percentage of the physical area, typically 20–50%, increases as the geometric device width of each inverter is made larger.

The physical area model of a buffer stage consists of two components: the area overhead, A_{OH}, which is constant for all stages of the buffer, and the active area, A_{ctv}, which scales with F. Thus, the physical area required for the ith stage may be expressed as

$$A_i = A_{\mathrm{OH}} + F^{i-1}A_{\mathrm{ctv}}. \tag{27}$$

The total area of a tapered buffer as a function of F is expressed in (28) by summing A_i for N stages, and substituting (11) into (27)

$$A_{\mathrm{total}} = A_{\mathrm{OH}}\left(\frac{\ln \frac{C_L}{C_y}}{\ln F}\right) + A_{\mathrm{ctv}}\left(\frac{\frac{C_L}{C_y} - 1}{F - 1}\right), \quad F > 1. \tag{28}$$

Both area terms in (28) decrease with increasing F. This trend occurs since N decreases with increasing F. Equation (28) is misleading in that it treats N as a continuous variable, whereas N may only assume integer values. Increasing F without decreasing N has the effect of increasing area. As with power dissipation, the optimal area of a tapered buffer is zero. Therefore, a tapered buffer with the fewest stages and which satisfies any remaining design criteria is the preferred physical area of the buffer.

The total area requirement of a tapered buffer given in (28) may be expressed as a discrete function of N, as shown in (29). This expression is dependent upon the relative magnitudes of A_{OH} and A_{ctv}. Assuming the area overhead of a minimum sized inverter is three times the active area, i.e., $A_{\mathrm{OH}} = 3 \times A_{\mathrm{ctv}}$, a graph depicting the physical area as a function of N and C_L/C_y is shown in Fig. 5

$$A_{\mathrm{total}} = N \cdot A_{\mathrm{OH}} + \left(\frac{\frac{C_L}{C_y} - 1}{\left(\frac{C_L}{C_y}\right)^{\frac{1}{N}} - 1}\right) A_{\mathrm{ctv}}. \tag{29}$$

As with power dissipation, the area requirement steadily increases for increasing N. The relative difference between the integer values of N is larger when considering physical area than with power dissipation, mainly due to the first component of (29) which displays a linear increase in area overhead with N. This increase in area penalty with increasing N provides an additional incentive not to increase N beyond N_D. Fig. 5 also graphically demonstrates that, as with power dissipation, reducing N by many stages below N_D increases propagation delay significantly while decreasing total area only approximately linearly.

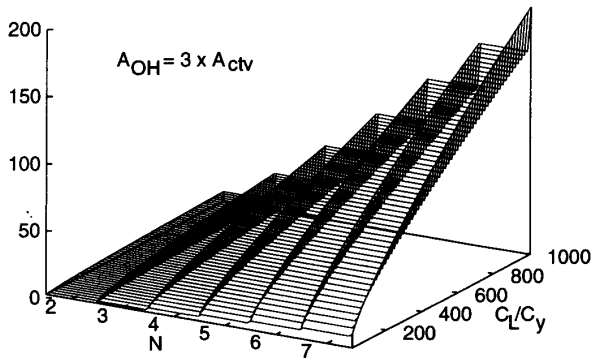

Fig. 5. Normalized physical area of a tapered buffer system.

VI. System Reliability

An important and only recently considered criterion in tapered buffer design is reliability. The failure mechanism of concern in tapered buffer design is hot-carrier degradation of the NMOS devices due to injected charge being trapped in the gate oxide of the NMOS devices within the buffer [23], [24]. The degradation experienced by the NMOS devices in an inverter is typically much greater than that experienced by the PMOS devices. This difference in degradation occurs since the substrate currents in the PMOS devices are smaller due to the lower mobility of holes in comparison with electrons [25]. Therefore, only the degradation of the NMOS devices is considered here.

The average bond-breaking current density in the NMOS transistors is a measure of hot-carrier degradation experienced by the NMOS transistors [12], [26]. The average bond-breaking current density as a function of tapering factor is shown in (30), where $\langle J_{\mathrm{BB}_0}\rangle$ is a process constant describing the average bond-breaking current density of the saturated NMOS transistor

$$\langle J_{\mathrm{BB}}\rangle = f(C_x + FC_y)\langle J_{\mathrm{BB}_0}\rangle. \tag{30}$$

Note that the formula for $\langle J_{\mathrm{BB}}\rangle$ has been extended in this paper from that presented in [12], [26] to include the split-capacitor model. The complete derivation for (30) is provided in [27].

The average bond-breaking current density is a measure of expected device lifetime. Lifetime may be considered a constant performance objective in the manner that propagation delay, power dissipation, and physical area are considered constant performance objectives. By using device degradation, which is a long-term effect, to predict lifetime, reliability may be treated in the same manner as the remaining three design criteria.

Hot-carrier degradation, as measured by average bond-breaking current density, increases with F. Thus, reliability improves with decreasing F. This behavior is due to the strong dependence of degradation upon signal slope and the weak dependence upon capacitive load [23], [26]. Smaller values of

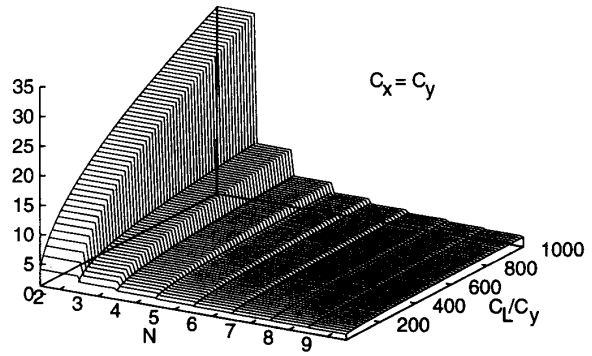

Fig. 6. Normalized degradation of a tapered buffer system.

F lead to greater waveform slopes within the tapered buffer. This behavior may be shown by substituting (7) into the expression for the waveform transition time, t_T, shown in (6)

$$t_T = \frac{(C_x + FC_y)}{I_{D0_1}}\left(\frac{0.9}{0.8} + \frac{V_{D0}}{0.8V_{\mathrm{DD}}}\ln\frac{10V_{D0}}{eV_{DD}}\right). \quad (31)$$

The most reliable design and the shortest waveform transition time correspond to a value of $F = 0$, however, a zero tapering factor again has no physical meaning. A fundamental design constraint for tapered buffer systems requires

$$F^N C_y = C_L \quad (32)$$

thus constraining $F > 1$ for all C_L greater than C_y. A lower bound on F for maximum reliability therefore exists at $F = 1$. However, this bound does not have any practical value in that it does not reduce the number of possible circuit implementations since as F approaches one, N becomes infinite for any $C_L > C_y$.

Converting (30) to a discrete function of N, the total device degradation of a tapered buffer system is described as shown in (33). Again, assuming that frequency is independent of the tapered buffer system and normalizing the overall function, the degradation experienced by the NMOS devices within a tapered buffer is shown in Fig. 6 as a function of N and C_L/C_y

$$\langle J_{\mathrm{BB}}\rangle = f\left(C_x + \left(\frac{C_L}{C_y}\right)^{\frac{1}{N}} C_y\right)\langle J_{\mathrm{BB}_0}\rangle. \quad (33)$$

The shape of the degradation graph depicted in Fig. 6 is similar to that of the graph illustrating propagation delay in Fig. 3 in that it exhibits a dramatic increase in degradation for small values of N. However, unlike propagation delay, degradation continuously decreases with increasing values of N, showing no local minima nor a strong dependence on load capacitance. An observation from Fig. 6 is that, from the perspective of reliability and for moderate-to-large sized load capacitance, N should be chosen to be greater than two.

VII. Unification

In Sections III–VI, analytical expressions for propagation delay, power dissipation, physical area, and hot-carrier degradation are presented using a single nomenclature, allowing for the unification of these four criteria. This unification is examined qualitatively in subsection A, followed by a quantitative analysis in subsection B. In subsection C, the unified tapered buffer is compared to tapered buffers designed for minimum delay.

A. Graphical Interpretation

In Figs. 3–6, the behavior of each of the four performance criteria with respect to variations in N and C_L/C_y is graphically presented. Utilizing these figures, the effects of deviating from the value of N which produces the minimum propagation delay, N_D, on the propagation delay, power dissipation, physical area, and hot-carrier degradation of a tapered buffer system may now be summarized.

Three of the criteria, propagation delay, power dissipation, and physical area, provide penalties for increasing N beyond N_D, and the reduced hot-carrier degradation benefit of increasing N beyond N_D is minimal in comparison to the increases in both power dissipation and physical area. Thus, it may be concluded that there is no compelling reason to increase N beyond N_D.

Propagation delay and hot-carrier degradation both exhibit dramatic increases for small values of N. These increases are not mitigated by substantial reductions in either physical area or power dissipation. It may therefore be concluded that N should be chosen large enough such that the substantial penalties in propagation delay and hot-carrier degradation for small N are not incurred.

Within the region between N_D and the small values of N where propagation delay and hot-carrier degradation exhibit dramatic increase, both power dissipation and physical area exhibit substantial reduction with decreasing N. Simultaneously, propagation delay and hot-carrier degradation increase moderately, but not prohibitively. Given this analysis, it is empirically concluded that the optimal value of N, considering all four factors, should be less than N_D, but not so low as to incur the tremendous propagation delay and system reliability penalties that occur with very small values of N.

In investigating the effects of these criteria on the performance of a tapered buffer system, equal weighting of all four design criteria has been assumed, i.e., all criteria are of equal importance in the design of a tapered buffer system. Clearly, for those application-specific circuits in which a subset of these four design criteria are emphasized, the observations regarding an optimal choice of N are skewed from those discussed here. However, the same general approach to choosing N may be applied, with the choice of N increasing for systems where propagation delay and/or reliability are of greater importance, and decreasing for systems where power dissipation and/or physical area are of greater importance. Furthermore, a methodology for evaluating those applications with unequal weighting is described in Section VIII.

B. Delay-Power-Area-Degradation Product

One strategy to permit further examination of these conflicting behaviors is to investigate the integrated effects of

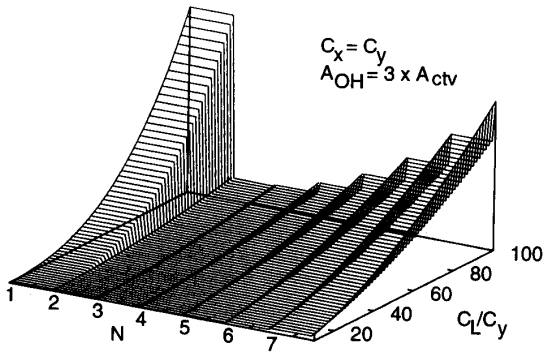

Fig. 7. Delay-power-area-degradation product of a tapered buffer with small load capacitance ($10 \leq C_L/C_y \leq 100$).

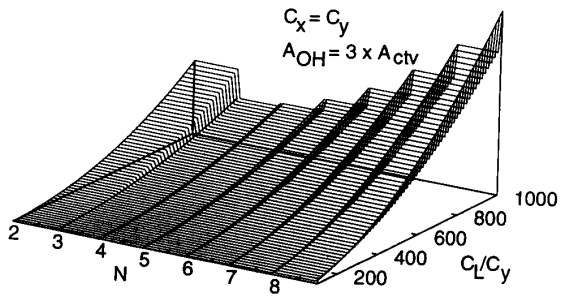

Fig. 8. Delay-power-area-degradation product of a tapered buffer with medium sized load capacitance ($100 \leq C_L/C_y \leq 1000$).

propagation delay, power dissipation, physical area, and hot-carrier degradation depicted by the product of (14), (26), (29), and (33). This product gives a figure of merit based on equal weighting of all four design criteria. The choice of N for which this product is a minimum provides the optimal buffer implementation [28]. The delay-power-area-degradation product as a function of N is (34), shown at the bottom of this page, where K' represents the collection of process constants. The delay-power-area-degradation product as a function of F is of similar complexity. Due to the form of the delay-power-area-degradation product, an analytic solution for its minimum is intractable. In this section, therefore, graphical solutions for the minimum delay-power-area-degradation product are investigated while a table look-up strategy for analyzing the delay-power-area-degradation product is presented in Section VIII. Since the delay-power-area-degradation product grows much larger for increasing C_L/C_y, the graph is broken into three separate graphs in order to preserve the information within each figure.

In Fig. 7, the delay-power-area-degradation product for $10 \leq C_L/C_y \leq 100$ is depicted. From this graph it is shown that over much of this range of C_L/C_y there is minimal difference between $N = 2$ and $N = 3$, thus the optimal number of stages is two when logical inversion is not desired, and three when logical inversion is preferred.

The symbol N_{opt} is used to represent both the number of stages which produces the minimum delay-power-area-degradation product and the number of stages which produces the near-minimum product. The notation in (35) represents these two approximately equivalent choices for the optimal value of N, one with logical inversion and one without

$$N_{\text{opt}} = (i, j). \tag{35}$$

Once the optimal number of stages, N_{opt}, is chosen, the optimal tapering factor, F_{opt}, may be computed directly from (36). F_{opt} also has two values, each of which corresponds to one of the two possible values of N_{opt}

$$F_{\text{opt}} = \left(\frac{C_L}{C_y}\right)^{\frac{1}{N_{\text{opt}}}}. \tag{36}$$

In Fig. 8, the delay-power-area-degradation product for $100 \leq C_L/C_y \leq 1000$ is depicted. This graph shows that for most of this range, there is minimal difference between $N = 3$ and $N = 4$. This leads to the conclusion that $N = 3$ is the optimum number of stages when logical inversion is desired, and $N = 4$ is the optimum when logical inversion is not preferred. The transition from $N_{\text{opt}} = (2, 3)$ to $N_{\text{opt}} = (3, 4)$ occurs in the lower end of the C_L/C_y range, though the exact transition is dependent upon the Cy/C_x ratio, which for these graphs, $C_y/C_x = 1$. Note that $N = 1$ is not shown on this graph, as it produces a much higher delay-power-area-degradation product that does not easily fit onto the vertical axis of Fig. 8.

In Fig. 9, the delay-power-area-degradation product for $1000 \leq C_L/C_y \leq 10\,000$ is shown. Over most of this range, there is minimal difference between $N = 4$ and $N = 5$. Thus, when logical inversion is not desired, a four stage buffer system is optimal, and when logical inversion is desired, a five stage buffer is preferable.

C. Comparison of N_{opt} with N_D

The number of stages and tapering factor which produce the minimum delay (N_D and F_D) is compared with the optimal number of stages and tapering factor (N_{opt} and F_{opt}) in Table I, where optimal is defined by the minimum delay-power-area-degradation product. The conditions $C_x = C_y$ and $A_{\text{OH}} = 3 \times A_{\text{ctv}}$ are assumed in Table I.

From Table I, it is shown that N_{opt} does not increase with increasing load capacitance (C_L/C_y) as quickly as N_D does. This behavior is due to one factor of the delay-power-

$$\text{Prod} = K' \frac{N(C_L - C_y)\left(C_L^{\frac{1}{N}} C_y + C_x C_y^{\frac{1}{N}}\right)^3 \left(A_{\text{ctv}} C_y^{(1+\frac{1}{N})} - A_{\text{ctv}} C_L C_y^{\frac{1}{N}} - A_{\text{OH}} C_L^{\frac{1}{N}} C_y N + A_{\text{OH}} C_y^{(1+\frac{1}{N})}\right)}{C_y^{(2+\frac{2}{N})}\left(-C_L^{\frac{1}{N}} + C_y^{\frac{1}{N}}\right)^2} \tag{34}$$

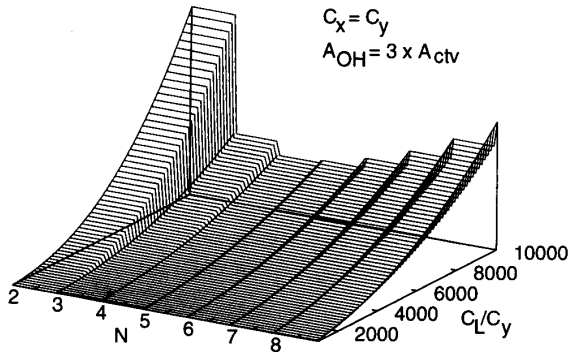

Fig. 9. Delay-power-area-degradation product of a tapered buffer with large load capacitance ($1000 \leq C_L/C_y \leq 10\,000$).

TABLE I
COMPARISON OF N AND F FOR MINIMUM PROPAGATION DELAY VERSUS UNIFIED METHODOLOGY $C_x = C_y$, $A_{OH} = 3 \times A_{ctv}$

C_L/C_y	Minimum delay		Unified methodology	
	N_D	F_D	N_{opt}	F_{opt}
10	2	3.16	(1, 2)	(10, 3.16)
100	4	3.16	(2, 3)	(10, 4.64)
1000	5	3.98	(3, 4)	(10, 5.62)
10000	7	3.72	(4, 5)	(10, 6.31)
100000	9	3.59	(5, 6)	(10, 6.81)

area-degradation product, the physical area, being independent of capacitive load. The result is that N_{opt} is less sensitive to variations in load than the choice of N based on solely minimizing propagation delay. Also noteworthy is that for $C_x \approx C_y$, a lower bound on N_{opt} appears as approximately $N = \lfloor \log_{10} \frac{C_L}{C_y} \rfloor$. With the powers of 10 for C_L/C_y that are used in Table I, this manifests itself as $F_{\text{opt}} = 10$ appearing in each entry. Additionally, note that $F_D > e$. This behavior is the result of utilizing the more realistic split-capacitor model rather than the single capacitor model applied by Jaeger [2]. The split-capacitor model results in tapering factors larger than e for minimum delay.

VIII. APPLICATION OF UNIFIED METHODOLOGY

In Section VII, the propagation delay, power dissipation, physical area, and system reliability of a tapered buffer are unified both qualitatively and quantitatively. As the intent of this paper is to provide a unified methodology for the design of tapered buffers which is easily applied to practical systems, the design of an application-specific tapered buffer system utilizing this process is addressed in this section.

Only a single degree of freedom, the choice of N, exists in the design of a tapered buffer system. Once N is determined, the tapering factor, F, is uniquely derived from the relationship shown in (13) and repeated below. Due to this interdependence of F on N, the tapering factor, F, does not provide an additional degree of freedom in the design of tapered buffers

$$F = \left(\frac{C_L}{C_y}\right)^{\frac{1}{N}}. \tag{13}$$

TABLE II
NUMBER OF STAGES FOR VARYING LOAD CAPACITANCE, $C_x = 10$ fF, $C_y = 15$ fF, $A_{OH} = 150\ \mu m^2$, $A_{ctv} = 50\ \mu m^2$

Load Capacitance (C_L)	Number of Stages (N_{opt})
$\leq$ 225 fF	(1, 2)
225 fF $\rightarrow$ 3 pF	(2, 3)
3 pF $\rightarrow$ 40 pF	(3, 4)
40 pF $\rightarrow$ 300 pF	(4, 5)
300 pF $\rightarrow$ 7 nF	(5, 6)

The parameters C_x, C_y, A_{OH}, and A_{ctv} in (14), (26), (29), and (33) all depend upon the layout and fabrication technology of the tapered buffer system. However, for a given technology and buffer layout, these values can be considered as constants. Therefore, the design of a tapered buffer reduces to choosing N, and therefore F, based on a specific load capacitance, C_L.

The application of the delay-power-area-degradation product to buffer system design in which no specific performance constraints exist is described in subsection A. The application of the delay-power-area-degradation product to buffer design in systems where specific performance constraints exist is described in subsection B.

A. Unconstrained Systems

The circuit attributes of propagation delay, power dissipation, physical area, and hot-carrier degradation are unified and strategies for determining the optimal choice of N are described in Section VII. This unifying process produces an optimal buffer implementation given a design space which is unconstrained other than by the tapered buffer relationship of (13) and with the assumption that all four design criteria are of equal importance. This permits a straightforward minimization of the delay-power-area-degradation product in order to determine the optimal number of stages.

Since the delay-power-area-degradation product is transcendental in both N and F, a generalized analytical solution for its minimum value is not provided, and the minimum value is determined using simple numerical techniques. If the product is expressed as a discrete function of N, as suggested in Section VII, this method quickly converges to the minimum product value in only a few iterations.

It is possible, therefore, to construct a look-up table for a specific technology for a broad range of load capacitance. Pseudocode of an algorithm for generating this look-up table is provided in the appendix. Given a circuit specification requiring a buffer to drive a large capacitive load, the optimal number of stages of a tapered buffer system can be immediately determined from a look-up table, permitting the tapering factor to be calculated directly from (13). An example of such a look-up table is shown in Table II.

As shown in Table II, very few entries are required to cover the expected range of load capacitance. The first two rows in the table represent typical on-chip loads, and the use of one to three stage tapered buffers for these capacitive loads agrees

TABLE III
NUMBER OF STAGES FOR MINIMUM DELAY-POWER-DEGRADATION PRODUCT, $C_x = 10$ fF, $C_y = 15$ fF

Load Capacitance (C_L)	Number of Stages (N_{opt})
≤ 85 fF	(1, 2)
85 fF → 290 fF	(2, 3)
290 fF → 900 fF	(3, 4)
900 fF → 2.8 pF	(4, 5)
2.8 pF → 8.5 pF	(5, 6)
8.5 pf → 25.5 pF	(6, 7)
25.5 pf → 75 pF	(7, 8)
75 pF → 250 pF	(8, 9)
250 pF → 700 pF	(9, 10)

with standard practice. The third and fourth rows represent larger capacitive loads typically encountered when driving off-chip. Finally, the fifth row represents highly atypical capacitive loads which might exist when driving board level system-wide interconnections, e.g., clock distribution networks or data busses. Also note that the table contains a solution for both inverted and noninverted logic polarities for each capacitive range, permitting the selection of the logic polarity best suited for the circuit implementation.

It is important to note that the delay-power-area-degradation product assumes all four design criteria are of equal importance. In systems where certain criteria are of greater importance than others, a similar process may be applied using a weighted product or other combination of (14), (26), (29), and (33) in order to reflect the relative importance of these design criteria.

In some systems, one or more of the four performance criteria may be of negligible or minor importance. It is therefore useful to construct look-up tables for subsets of the four criteria. There are 15 possible products, however four of these products have a minimum at either $N = 0$ or $N = \infty$, which are physically unrealizable. These four nonphysically realizable products are power, area, degradation, and power-area. However, eleven products of possible interest remain. In order to exemplify the process of evaluating these eleven products, an example table describing the number of stages for varying load capacitance for an equally weighted delay-power-degradation product is shown in Table III.

B. Constrained Systems

In the previous subsection, a buffer design methodology is presented for those systems where the buffer is not constrained to meet specific performance requirements. However, often an application-specific system places particular performance constraints upon a buffer. In modern CMOS-based systems, propagation delay and power dissipation are often both of primary concern, placing limits on the range of N which may be considered during the design of a tapered buffer system. Physical area and system reliability may place additional constraints upon N, and the same process described below for just propagation delay and power dissipation is applicable with those criteria.

As power dissipation monotonically increases with N, a specification on the maximum power dissipation, $P_{\max}$, has the effect of placing an upper bound on N, $N_{P_{\max}}$, at or below which the power dissipation of the buffer system is within specification. In order to determine the maximum number of stages of a tapered buffer system which will satisfy the maximum power dissipation requirement, (26) is set equal to $P_{\max}$ and solved for $N_{P_{\max}}$. This transformation results in

$$N_{P_{\max}} = \left\lfloor \frac{\ln\left(\frac{C_L}{C_y}\right)}{\ln\left[\frac{K_p C_x\left(\frac{C_L}{C_y}-1\right)+P_{\max}}{P_{\max}-K_p C_y\left(\frac{C_L}{C_y}-1\right)}\right]} \right\rfloor \tag{37}$$

where

$$K_P = V_{DD}^2 f(1 + K_{P_{SC}}) \tag{38}$$

and $K_{P_{SC}}$ is defined in (22). Thus, the maximum number of stages and tapering factor of a buffer system can be directly determined from a specified maximum power dissipation requirement.

Similarly, (14) may be set equal to a specified delay value and solved for N. However, unlike the expression for power dissipation, this process results in a transcendental expression in which the two solutions for N are not analytically obtainable. Therefore, a preferred method is to numerically solve for N using the discrete nature of N to limit the granularity, and thus the solution time.

Due to the upward concavity of the delay function shown in Fig. 3, a maximum delay constraint places both upper and lower bounds on the value of N between which the delay constraint is satisfied. If N_{opt} falls within the bounds established by the propagation delay and power dissipation requirements, then N_{opt} satisfies both the propagation delay and the power dissipation requirements of the system and is the recommended number of stages for the particular application-specific tapered buffer.

An example of such a constrained system, utilizing the process parameters shown in Table IV, is as follows. Assume an application requires that a 45 pF load be driven at 25 MHz with a buffer dissipating no more than 300 mW (continuous operation), has a propagation delay of no more than 15 ns, and an inverting polarity is preferred. No specific constraints exist for physical area or system reliability, though it is desired to optimize these within the specified speed and power constraints. Therefore, all four of the performance criteria are of concern, and the delay-power-area-degradation product used to generate Table II may be applied. From (37), the maximum allowable number of stages, assuming a maximum power dissipation of 300 mW, is $N_{P_{\max}} = 6$. Equation (14) is used to numerically determine that the propagation delay is less than 15 ns for $4 \leq N \leq 13$. Thus, the range of N which satisfies both the propagation delay and the power dissipation constraints is $4 \leq N \leq 6$. This "design space" is shown as the gray area in Fig. 10. The optimal number of stages for

TABLE IV
EXAMPLE PROCESS PARAMETERS

Parameter	Value
$\alpha_n = \alpha_p$	1
$V_{THn} = \lvert V_{THp} \rvert$	0.5 V
$V_{D0n} = \lvert V_{D0p} \rvert$	3 V
V_{DD}	5 V
$I_{D0n} = \lvert I_{D0p} \rvert$	100 μA
C_x	10 fF
C_y	15 fF
A_{ctv}	50 μm^2
A_{OH}	150 μm^2

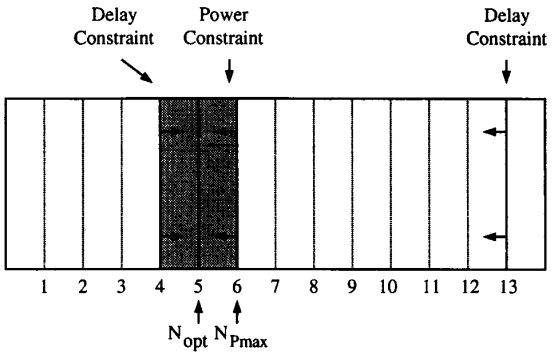

Fig. 10. N_{opt} falls within propagation delay and power dissipation constraints.

$C_L = 45$ pF, derived from Table II, is $N_{opt} = (4, 5)$. The logically inverted value, $N = 5$, falls within the design space. Thus, a choice of $N = 5$ and therefore a tapering factor of $F = 4.96 \approx 5.0$ is recommended for the tapered buffer circuit described in this example.

However, N_{opt} may fall outside the constraints imposed by the propagation delay and/or power dissipation specifications. If this occurs, N is chosen to be the value closest to N_{opt} while remaining within the range established by these constraints.

As an example of this process, again utilize the process parameters given in Table IV, and assume that a different application requires a buffer to drive a 45 pF load at 25 MHz with a maximum power dissipation of 450 mW, a maximum propagation delay of 12 ns, and no logical inversion. Applying (37) to this example, $N \leq N_{P_{max}} = 10$. Equation (14) is used to determine that the propagation delay is less than 12 ns for $6 \leq N \leq 8$. Thus, the design space which meets these constraints is $6 \leq N \leq 8$. From Table II, $N_{opt} = (4, 5)$, neither value of which falls within the preferred design space. In this case, $N = 6$ should be chosen since it is the value closest to N_{opt} which is within the design space and satisfies the noninverted logic polarity requirement. From $N = 6$, a tapering factor of $F = 3.80$ is directly determined. This example is illustrated in Fig. 11.

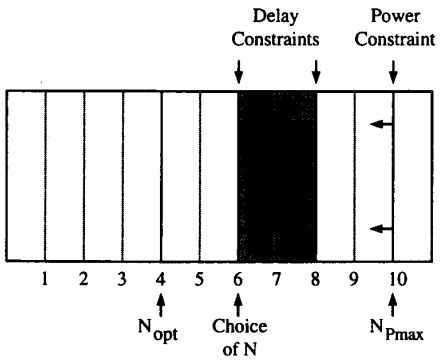

Fig. 11. N_{opt} falls outside propagation delay and power dissipation constraints.

In this manner, a tapered buffer system may be designed with all four design criteria unequally weighted. One or more criteria are used to establish a region in which equal weighting is applied locally. Due to the upward concavity of the delay-power-area-degradation product, choosing the value of N closest to N_{opt} within the constrained region guarantees that the delay-power-area-degradation product has the minimum possible value within the permitted design space. Thus, the equally weighted product is applied locally to the design space in order to determine appropriate application-specific values of N and F.

IX. CONCLUSION

A CMOS integrated circuit designer is often faced with multiple, conflicting, design criteria when confronted with the task of driving a large capacitive load with a tapered buffer system. This paper provides analytical expressions for the four primary criteria typically encountered in tapered buffer design: propagation delay, power dissipation, physical area, and system reliability. It is preferable to consider these four design issues as discrete functions of N, rather than as continuous functions of F, since the design space is vastly reduced. Each of these design criteria is graphically illustrated as a discrete function of N, and the shapes of these graphs are used to develop a unifying strategy for choosing N. The behavior of each design criterion as a function of N leads to the conclusion that the optimal number of stages, for equal weighting of all four criteria, is less than the number of stages which produces the minimum propagation delay. This result is due to the significant increase in power dissipation and physical area outweighing the increase in reliability for increasing N.

The delay-power-area-degradation product is investigated to examine this conclusion. This product provides a measure of the simultaneous tradeoffs that exist among all four design criteria of the tapered buffer system. It is shown that for a wide range of load capacitance, there exist an optimal and a nearly optimal value of N whose difference is one. This result describes both logically inverted and noninverted tapered buffer systems which are, for all practical purposes, equivalent in delay-power-area-degradation product.

```
/* As initialization, one may assume that for CL = Cy, Nopt = (1,2)
*/
initialize brackets
while upper bracket capacitance is less than maximum capacitance
    lower bracket capacitance <- upper bracket capacitance
    central bracket N <- lower bracket N + 1
    upper bracket N <- central bracket N + 1
    evaluate lower bracket product
/* With the lower bracket representing a CL for which Nopt = (i, i+1),
   the bracketing function searches for an upper bracket CL for which
   Nopt = (i+1, i+2) or higher.  This reduces to finding a value
   of CL for which the product for
   N = i+2 is less than the product for N = i.
*/
   do                          /* bracket transition */
       increase upper bracket capacitance
       evaluate upper bracket product
   while lower bracket product < upper bracket product
/* The upper and lower brackets now contain values of CL between which
   Nopt transitions from (i, i+1) to (i+1, i+2).  Because of the well
   behaved nature of the Nopt function, simple bisection may be
   applied to find this transition point within a specified tolerance.
*/
   do                          /* bisection */
       central bracket cap. <- average upper and lower bracket caps.
       Product1 <- product of central cap. with lower N
       Product2 <- product of central cap. with upper N
       if Product1 < Product2
            lower bracket <- central bracket
       else
            upper bracket <- central bracket
   while upper and lower bracket difference > tolerance
/* The transition from Nopt = (i, i+1) to (i+1, i+2) has now been
   located.  Repeat algorithm for next transition using old upper
   bracket as the new lower bracket until maximum capacitance is
   reached.
*/
endwhile
```

Fig. 12. Pseudocode for look-up table generation algorithm.

These results are used to develop a design strategy for both unconstrained and constrained tapered buffer systems. A technology dependent look-up table is used in the design of tapered buffer systems where the performance specifications are unconstrained. For those applications where specified performance attributes constrain the design, the relationships developed in this paper are used to establish additional limits on the number of stages. The technology dependent look-up tables are then used to determine the final choice of the appropriate number of stages, and therefore the tapering factor, once the necessary performance specifications have been satisfied.

This paper describes a unified design methodology for tapered buffer systems which simultaneously considers propagation delay, power dissipation, physical area, and hot-carrier system reliability. This methodology integrates these until now disparate performance criteria, permitting the optimal design of application-specific CMOS tapered buffers.

APPENDIX

Generation of N_{opt} Tables

Look-up tables for N_{opt}, as exemplified by Tables II and III, may be generated through the use of a simple algorithm, a pseudocode version of which is depicted in Fig. 12. Given target process parameters and a maximum load capacitance to serve as a termination point, the algorithm brackets the capacitance transition values between $N_{\text{opt}} = (i, i+1)$ and $N_{\text{opt}} = (i+1, i+2)$ and applies bisection to locate these capacitance transition values within a certain error tolerance.

REFERENCES

[1] H. C. Lin and L. W. Linholm, "An optimized output stage for MOS integrated circuits," *IEEE J. Solid-State Circuits*, vol. SC-10, no. 2, pp. 106–109, Apr. 1975.

[2] R. C. Jaeger, "Comments on 'An optimized output stage for MOS integrated circuits'," *IEEE J. Solid-State Circuits*, vol. SC-10, no. 3, pp. 185–186, June 1975.

[3] H. B. Bakoglu and J. D. Meindl, "Optimal interconnection circuits for VLSI," *IEEE Trans. Electron Devices*, vol. ED-32, no. 5, pp. 903–909, May 1985.

[4] S. Dhar and M. A. Franklin, "Optimum buffer circuits for driving long uniform lines," *IEEE J. Solid-State Circuits*, vol. 26, no. 1, pp. 32–40, Jan. 1991.

[5] H. J. M. 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, no. 4, pp. 468–473, Aug. 1984.

[6] L. A. Glasser and L. P. J. Hoyte, "Delay and power optimization in VLSI circuits," in *Proc. ACM/IEEE Design Automat. Conf.*, June 1984, pp. 529–535.

[7] F. S. Lai, "A generalized algorithm for CMOS circuit delay, power, and area optimization," *Solid-State Electron.*, vol. 31, no. 11, pp. 1619–1627, Nov. 1988.

[8] B. Hoppe, G. Neuendorf, D. Schmitt-Landsiedel, and W. Specks, "Optimization of high-speed CMOS logic circuits with analytical models for signal delay, chip area, and dynamic power dissipation," *IEEE Trans. Comput.-Aided Design*, vol. 9, pp. 236–247, Mar. 1990.

[9] C. M. Lee and H. Soukup, "An algorithm for CMOS timing and area optimization," *IEEE J. Solid-State Circuits*, vol. SC-19, no. 5, pp. 781–787, Oct. 1984.

[10] E. T. Lewis, "Optimization of device area and overall delay for CMOS VLSI designs," *Proc. IEEE*, vol. 72, pp. 670–688, June 1984.

[11] S. R. Vemuru and A. R. Thorbjornsen, "Variable-taper CMOS buffer," *IEEE J. Solid-State Circuits*, vol. 26, no. 9, pp. 1265–1269, Sept. 1991.

[12] W. Sun, Y. Leblebici, and S. M. Kang, "Design-for-reliability rules for hot-carrier resistant CMOS VLSI circuits," in *Proc. IEEE Int. Symp. Circuits. Syst.*, May 1992, pp. 1254–1257.

[13] A. Kanuma, "CMOS circuit optimization," *Solid-State Electron.*, vol. 26, no. 1, pp. 47–58, 1983.

[14] N. Hedenstierna and K. O. Jeppson, "CMOS circuit speed and buffer optimization," *IEEE Trans. Comput.-Aided Design*, vol. CAD-6, no. 2, pp. 270–281, Mar. 1987.

[15] N. C. Li, G. L. Haviland and A. A. Tuszynski, "CMOS tapered buffer," *IEEE J. Solid-State Circuits*, vol. 25, no. 4, pp. 1005–1008, Aug. 1990.

[16] C. Prunty and G. Laszlo, "Optimum tapered buffer," *IEEE J. Solid-State Circuits*, vol. 27, no. 1, pp. 118–119, Jan. 1992.

[17] M. Nemes, "Driving large capacitance in MOS LSI systems," *IEEE J. Solid-State Circuits*, vol. SC-19, no. 1, pp. 159–161, Feb. 1984.

[18] T. Sakurai, "A unified theory for mixed CMOS/BiCMOS buffer optimization," *IEEE J. Solid-State Circuits*, vol. 27, no. 7, pp. 1014–1019, July 1992.

[19] N. Hedenstierna and K. O. Jeppson, "Comments on the optimum CMOS tapered buffer problem," *IEEE J. Solid-State Circuits*, vol. 29, no. 2, pp. 155–159, Feb. 1994.

[20] T. Sakurai and A. R. Newton, "Alpha-power law MOSFET model and its applications to CMOS inverter delay and other formulas," *IEEE J. Solid-State Circuits*, vol. 25, no. 2, pp. 584–594, Apr. 1990.

[21] B. S. Cherkauer and E. G. Friedman, "Unification of speed, power, area, and reliability in CMOS tapered buffer design," in *Proc. IEEE Int. Symp. Circuits, Syst.*, May 1994, pp. 4.111–4.114.

[22] H. Shichman and D. A. Hodges, "Modeling and simulation of insulated-gate field-effect transistor switching circuits," *IEEE J. Solid-State Circuits*, vol. SC-3, no. 3, pp. 285–289, Sept. 1968.

[23] W. Weber, M. Bronx, T. Künemund, H. M. Mühlhoff, and D. Schmitt-Landsiedel, "Dynamic degradation in MOSFET's—Part II: Application in the circuit environment," *IEEE Trans. Electron Devices*, vol. 38, no. 8, pp. 1859–1867, Aug. 1991.

[24] Y. Leblebici and S. M. Kang, "Modeling and simulation of hot-carrier-induced device degradation in MOS circuits," *IEEE J. Solid-State Circuits*, vol. 28, no. 5, pp. 585–595, May 1993.

[25] C. Duvvury and S. Aur, "Hot-carrier degradation effects for DRAM circiuts," in *Hot Carrier Design Considerations for MOS Devices and Circuits*, C. T. Wang, Ed. Princeton, NJ: Van Nostrand Reinhold, 1992, pp. 120–171.

[26] Y. Leblebici, W. Sun, and S. M. Kang, "Parametric macro-modeling of hot-carrier induced dynamic degradation in MOS VLSI circuits," *IEEE Trans. Electron Devices*, vol. 40, no. 3, pp. 673–676, Mar. 1993.

[27] B. S. Cherkauer and E. G. Friedman, "A split-capacitor expression for hot-carrier degradation of CMOS tapered buffers," Tech. Rep. EE-94-11, Dept. Elec. Eng., Univ. Rochester, Rochester, NY, Aug. 1994.

[28] ——, "A design methodology for low power, reduced area, reliable CMOS buffers," in *Proc. IEEE 37th Midwest Symp. Circuits, Syst.*, Aug. 1994.

Brian S. Cherkauer (S'93) was born in Buffalo, NY, in 1968. He received the B.S. degree (summa cum laude) from the State University of New York at Buffalo in 1990 and the M.S. degree from the University of Rochester, Rochester, NY, in 1991, both in electrical engineering. He is working toward the Ph.D. degree in electrical engineering at the University of Rochester.

He was with Mennen Medical, Inc., Clarence, NY, as a Software Engineering Technician, from 1988 to 1990, where he developed electrocardiogram simulation, digitization, and playback software. He has been a Teaching and Research Assistant at the University of Rochester since 1992. His research interests include high-performance digital and analog integrated circuit design techniques; CMOS and BiCMOS integrated circuit design; speed, area, and power tradeoffs; and design for reliability.

Mr. Cherkauer received the Sproull Fellowship from the University of Rochester in 1990.

Eby G. Friedman (S'78–M'79–SM'90) was born in Jersey City, NJ, in 1957. He received the B.S. degree from Lafayette College, Easton, PA, in 1979, and the M.S. and Ph.D. degrees from the University of California, Irvine, in 1981 and 1989, respectively, all in electrical engineering.

He was with Philips Gloeilampen Fabrieken, Eindhoven, The Netherlands, in 1978, where he worked on the design of bipolar differential amplifiers. From 1979 to 1991, he was with Hughes Aircraft Company, rising to the position of Manager of the Signal Processing Design and Test Department, responsible for the design and test of high-performance VLSI/VHSIC CMOS and BIMOS digital and analog IC's, the development of supporting design and test methodologies and CAD tools, and the development of high-performance and high-resolution DSP and oversampled systems. He has been with the Department of Electrical Engineering at the University of Rochester, Rochester, NY, since 1991, where he is an Associate Professor and Director of the High Performance VLSI/IC Design and Analysis Laboratory. He has authored a book chapter and many papers in the fields of high-speed and low-power CMOS design techniques, pipelining and retiming, and the theory and application of synchronous clock distribution networks, and has edited the book, *Clock Distribution Networks in VLSI Circuits and Systems,* (New York: IEEE Press, 1995). His current research and teaching interests are in the areas of high-performance VLSI/IC design and analysis and related system applications.

Dr. Friedman was a recipient of the Howard Hughes Masters and Doctoral Fellowships, an NSF Research Initiation Award, a DoD Augmentation Award for Science and Engineering Research Training, and a University of Rochester College of Engineering Teaching Excellence Award. He is a Member of the editorial board of the IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS II: ANALOG AND DIGITAL SIGNAL PROCESSING, Chair of the VLSI Systems and Applications CAS Technical Committee, a Member of the technical program committee of several conferences (ISCAS, APAW, GLSVLSI, MWSCAS, and ASIC), and an Officer of the Electron Devices Chapter of the IEEE Rochester Section.