A Design Methodology for Low Power, Reduced Area, Reliable CMOS Buffers

Brian S. Cherkauer and Eby G. Friedman
Department of Electrical Engineering
University of Rochester
Rochester, New York 14627

Abstract – Circuit speed, power dissipation, physical area, and system reliability are the four performance criteria of primary concern in tapered buffer design. Each places a separate, often conflicting constraint on the design of a tapered buffer. This paper presents a unified design methodology for CMOS tapered buffers which permits trade-offs to be easily made among these four performance criteria. The methodology utilizes analytical expressions for each of the performance criteria. A process dependent look-up table is constructed based on these expressions and is used in conjunction with application-specific performance constraints to efficiently determine the optimal implementation for each particular buffer instantiation.

I. Introduction

In CMOS integrated circuits, large capacitive loads occur both on-chip, where high, localized fan-out is often encountered, 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 source and sink relatively large currents quickly, while not degrading the performance of previous stages. In CMOS, a tapered buffer system is often used to perform this task [1, 2].

The basic problem that the tapered buffer must solve is illustrated in Figure 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 able to quickly charge (and discharge) the large capacitive load. Thus, the buffer isolates the logic/registers from the load, amplifying the signal along the way.

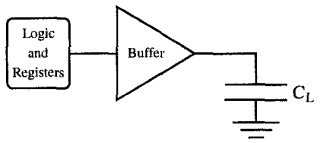

Fig. 1. Buffer used to drive capacitive load

Many different approaches to tapered buffer design have been described in the literature. The most commonly addressed criteria in tapered buffer design are propagation delay, power dissipation, physical area, and circuit reliability. Recently, enhanced expressions have been developed which describe each of these criteria in a form which permits the unification of these expressions into a single design methodology [3]. In this paper, the application of this methodology to the practical design of tapered buffers is presented. This methodology permits a designer to quickly and efficiently make optimal trade-offs among circuit speed, power dissipation, physical area, and system reliability.

This material is based upon work supported by the National Science Foundation under Grant No. MIP-9208165.

II. Performance Criteria

The tapered buffer model and the performance expressions presented in [3] are reviewed in this section. The split-capacitor model of a tapered buffer system, as denoted by Li, Haviland, and Tuszynski [4], is illustrated in Figure 2. This split-capacitor model is used in the development of each of the performance expressions. With this model, the load capacitance of the i^{th} stage of a tapered buffer, numbered from the input stage as illustrated in Figure 2, is

$$C_{L_i} = F^{i-1}(C_x + F\,C_y)\,, \tag{1}$$

where C_x represents the output capacitance of stage 1, C_y represents the input gate capacitance of stage 1, C_L is the capacitance being driven by the buffer, and F is the tapering factor. The tapering factor, F, as a discrete

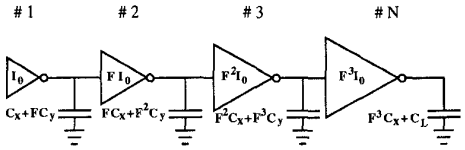

Fig. 2. The split-capacitor model of a tapered buffer

function of the number of stages, N, using the Li split-capacitor model is

$$F = \left(\frac{C_L}{C_y}\right)^{\frac{1}{N}} \tag{2}$$

In modern submicrometer CMOS fabrication technologies, short-channel effects are often quite pronounced. Therefore, an accurate and efficient short-channel transistor model must be included in the buffer delay equation. The transistor model used in this paper is the α-power I–V relationship developed by Sakurai and Newton [5]. Utilizing the Li split-capacitor model with the α-power short-channel transistor model, the following tapered buffer performance expressions, originally described in [3], are presented.

The propagation delay through a tapered buffer system is

$$t_{buffer} = N\frac{V_{DD}\left(C_x + \left(\frac{C_L}{C_y}\right)^{\frac{1}{N}} C_y\right)}{I_{D0_1}}\left(\frac{K_{HL}+K_{LH}}{2}\right), \tag{3}$$

where

$$K_{HL} = \left[\left(\frac{0.9}{0.8} + \frac{V_{D0_p}}{0.8\,V_{DD}}\,ln\frac{10\,V_{D0_p}}{e\,V_{DD}}\right)\left(\frac{1}{2} - \frac{1-\nu_{T_n}}{1+\alpha_n}\right) + \frac{1}{2}\right], \tag{4}$$

$$K_{LH} = \left[\left(\frac{0.9}{0.8} + \frac{V_{D0_n}}{0.8\,V_{DD}}\,ln\frac{10\,V_{D0_n}}{e\,V_{DD}}\right)\left(\frac{1}{2} - \frac{1-\nu_{T_p}}{1+\alpha_p}\right) + \frac{1}{2}\right]. \tag{5}$$

The power dissipation of a tapered buffer system, including both dynamic and short-circuit [6] power dissipation, is

$$P_{tot} = V_{DD}^2 f\left(1 + K_{P_{SC}}\right)\left(C_x + \left(\frac{C_L}{C_y}\right)^{\frac{1}{N}} C_y\right)\left(\frac{\frac{C_L}{C_y} - 1}{\left(\frac{C_L}{C_y}\right)^{\frac{1}{N}} - 1}\right), \quad (6)$$

where

$$K_{P_{SC}} = \left(\frac{0.9}{0.8} + \frac{V_{D0}}{0.8\, V_{DD}} ln \frac{10\, V_{D0}}{e\, V_{DD}}\right) \frac{1}{(\alpha+1)} \frac{1}{2^{\alpha-1}} \frac{(1-2\nu_T)^{\alpha+1}}{(1-\nu_T)^{\alpha}} . \quad (7)$$

The physical area model of a buffer stage consists of two components: the area overhead, A_{OH}, which is constant for every buffer stage, and the active area, A_{ctv}, which scales with F [7]. Thus, the physical area of a tapered buffer system may be expressed as

$$A_{total} = N \cdot A_{OH} + \left(\frac{\frac{C_L}{C_y} - 1}{\left(\frac{C_L}{C_y}\right)^{\frac{1}{N}} - 1}\right) A_{ctv} . \quad (8)$$

The long term hot carrier degradation experienced by the NMOS devices in the tapered buffer system is directly related to the average bond breaking current density [8]. The average bond-breaking current density in the NMOS devices of a tapered buffer may be expressed as

$$\langle J_{BB} \rangle = f\left(C_x + \left(\frac{C_L}{C_y}\right)^{\frac{1}{N}} C_y\right) \langle J_{BB_0} \rangle , \quad (9)$$

where $\langle J_{BB_0} \rangle$ is a process constant describing the average bond-breaking current density of the saturated NMOS transistor, which is a measure of device lifetime [8].

III. Application of Unified Methodology

In Section II, 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 is addressed in this section. It is suggested in [3] that a weighted delay-power-area-degradation product, the product of (3), (6), (8), and (9), be used to determine the optimal tapered buffer implementation. In this section, the design of an optimally tapered buffer which utilizes this process is described.

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 (2). 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.

The parameters C_x, C_y, A_{OH}, and A_{ctv} in (3) – (9) 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 minimum delay-power-area-degradation product may be used to determine the optimal tapered buffer implementation in *unconstrained* systems. That is, the product may be used in those systems where specific performance constraints for the tapered buffer system are unspecified.

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 [3], 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. Given a circuit specification requiring a buffer to drive a large capacitive load, as shown in Figure 1, 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 (2). An example of such a look-up table is shown in Table I.

Table I
Number of Stages for Varying Load Capacitance,
C_x = 10 fF, C_y = 15 fF, A_{OH} = 150 μm^2, A_{ctv} = 50 μm^2

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

As shown in Table I, 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 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 non-inverted logic polarities for each capacitive range, permitting the selection of the logic polarity best suited for the particular circuit implementation. The symbol N_{opt} is used to represent these two nearly-equivalent values for the optimal number of stages of a tapered buffer system.

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

(3), (6), (8), and (9) 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 non-physically realizable products are power, area, degradation, and power-area. However, eleven remaining 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 II.

Table II
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 → 75.0 pF	(7, 8)
75.0 pF → 250 pF	(8, 9)
250 pF → 700 pF	(9, 10)

B. Constrained Systems

In the previous subsection, a buffer design methodology is presented for those systems where the buffer is not constrained to meet a specific performance requirement. 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_{Pmax}, at or below which the power dissipation of the buffer 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, (6) is set equal to P_{max} and solved for N_{Pmax}. This transformation results in

$$N_{Pmax} = \left\lfloor \frac{ln\left(\frac{C_L}{C_y}\right)}{ln\left[\frac{K_p C_x\left(\frac{C_L}{C_y}-1\right)+P_{max}}{P_{max}-K_p C_y\left(\frac{C_L}{C_y}-1\right)}\right]} \right\rfloor , \quad (10)$$

where

$$K_P = V_{DD}^2 f(1 + K_{P_{SC}}) , \quad (11)$$

and K_{Psc} is defined in (7). 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, (3) 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 [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 III, which were used to generate Table I, 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 I may be applied. From (10), the maximum allowable number of stages, assuming a maximum power dissipation of 300 mW, is $N_{Pmax} = 6$. Equation (3) is used to numerically determine that the propagation delay is less than 15 ns for $4 \le N \le 13$. Thus, the range of N which satisfies both the propagation delay and the power dissipation constraints is $4 \le N \le 6$. This "design space" is shown as the gray area in Figure 3. The optimal number of stages for C_L = 45 pF, derived from Table I, 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.

Table III
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

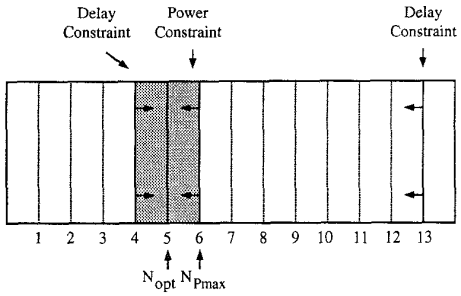

Fig. 3. N_{opt} falls within propagation delay and power dissipation constraints

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 III 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 (10) to this example, $N \leq N_{Pmax} = 10$. Equation (3) 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 I, $N_{opt} = (4, 5)$, 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 non-inverted logic polarity requirement. From $N = 6$, a tapering factor of $F = 3.80$ is directly determined. This example is illustrated in Figure 4.

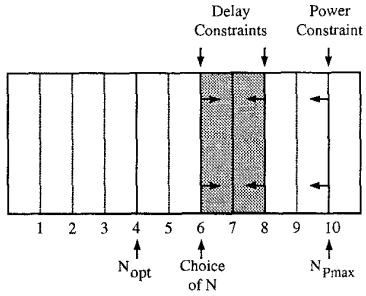

Fig. 4. N_{opt} falls outside propagation delay and power dissipation constraints

In this manner, a tapered buffer system may be designed with all four design criteria unequally weighted. Rather, 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 an appropriate application-specific value of N and F.

IV. Conclusions

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. In this paper, an efficient design methodology for making performance trade-offs among circuit speed, power dissipation, physical area, and hot-carrier reliability is described.

The delay-power-area-degradation product has been investigated for this purpose. It is shown that a process dependent look-up table may be constructed which permits a designer to efficiently determine the optimal number of stages for a tapered buffer system. The tapering factor is then calculated based on a simple expression, uniquely determining the tapered buffer system.

The delay-power-area-degradation product is applied to the design of tapered buffer systems which are either unconstrained or constrained by specified performance requirements. This methodology is adapted to systems where one or more of the four performance criteria have unequal weighting.

Thus, this paper integrates into a single unified design methodology the until now disparate design approaches for CMOS tapered buffer systems. Utilizing this methodology, optimal trade-offs may be made simultaneously among all four performance criteria: circuit speed, power dissipation, physical area, and hot-carrier reliability.

References

[1] H. C. Lin and L. W. Linholm, "An Optimized Output Stage for MOS Integrated Circuits," *IEEE Journal of Solid-State Circuits*, Vol. SC-10, No. 2, pp. 106–109, April 1975.

[2] R. C. Jaeger, "Comments on 'An Optimized Output Stage for MOS Integrated Circuits'," *IEEE Journal of Solid-State Circuits*, Vol. SC-10, pp. 185–186, June 1975.

[3] B. S. Cherkauer and E. G. Friedman, "Unification of Speed, Power, Area, and Reliability in CMOS Tapered Buffer Design," *Proceedings of the IEEE International Symposium on Circuits and Systems*, pp. 4.111–4.114, May 1994.

[4] N. C. Li, G. L. Haviland, and A. A. Tuszynski, "CMOS Tapered Buffer," *IEEE Journal of Solid-State Circuits*, Vol. SC–25, pp. 1005–1008, August 1990.

[5] T. Sakurai and A. R. Newton, "Alpha-Power Law MOSFET Model and its Applications to CMOS Inverter Delay and Other Formulas," *IEEE Journal of Solid-State Circuits*, Vol. SC–25, No. 2, pp. 584–594, April 1990.

[6] H. J. M. Veendrick, "Short-Circuit Dissipation of Static CMOS Circuitry and Its Impact on the Design of Buffer Circuits," *IEEE Journal of Solid-State Circuits*, Vol. SC-19, pp. 468–473, August 1984.

[7] 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 Transactions on Computer-Aided Design*, Vol. CAD–9, pp. 236–247, March 1990.

[8] W. Sun, Y. Leblebici, and S. M. Kang, "Design-For-Reliability Rules for Hot-Carrier Resistant CMOS VLSI Circuits," *Proceedings of the IEEE International Symposium on Circuits and Systems*, pp. 1254–1257, May 1992.