Unification of Speed, Power, Area, and Reliability in CMOS Tapered Buffer Design

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

ABSTRACT

Circuit speed, power dissipation, physical area, and system reliability are the four performance criteria of concern in tapered buffer design. Each places a separate, often conflicting constraint on the design of a tapered buffer. Enhanced short-channel tapered buffer design equations are developed for propagation delay and power dissipation, as well as a new split-capacitor model of hot-carrier reliability and a two-component physical area model. Each performance criterion is independently investigated and analyzed, and the interaction of the four criteria is examined to develop both a qualitative and a quantitative understanding of the various design tradeoffs. These disparate approaches to tapered buffer design are unified into a convenient, integrated design methodology.

INTRODUCTION

In CMOS integrated circuit design, large capacitive loads occur both on-chip, where high, localized fan-out 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 are able to 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].

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, quite recently, circuit reliability. Traditional 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 result of this paper is the unification of these seemingly independent criteria into a single, integrated design methodology for determining an application-specific tapering factor and number of stages of a tapered buffer system for driving a wide range of capacitive load.

PROPAGATION DELAY

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 a buffer delay equation. The transistor model used in this paper is the α-power I–V relationship developed by Sakurai and Newton [3].

The split-capacitor model was developed by Li, Haviland, and Tuszynski [4] and is illustrated in Figure 1. With this split-capacitor model, C_L for the ith stage of the buffer, numbered from the input stage as illustrated in Figure 1, 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, and F is the tapering factor.

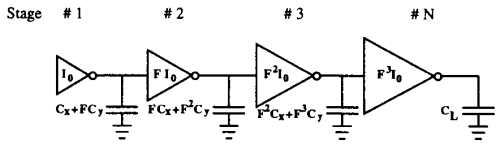

Figure 1. The split-capacitor model of a tapered buffer

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

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

With this relationship, the propagation delay through a tapered buffer can be written in terms of N as

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

where

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

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

This expression is normalized to remove process constants, thereby expressing delay as a function of only those variables which are controlled during the design process. This results in the normalized delay expression shown in (6), which is illustrated graphically in Figure 2.

$$\left(\frac{2\,I_{D0_1}}{V_{DD}(K_{HL} + K_{LH})}\right) t_{buffer} = N\left(C_x + \left(\frac{C_L}{C_y}\right)^{\frac{1}{N}} C_y\right) \tag{6}$$

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

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

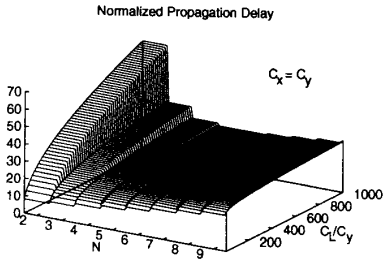

Figure 2. Normalized propagation delay of a tapered buffer system

The number of stages required to minimize the propagation delay through a tapered buffer system increases slowly as a function of C_L/C_y, as shown by the shape of the graph in Figure 2. It should also be noted that the propagation delay dramatically increases for values of N much smaller than the choice of N where delay is a minimum, which is hereafter referred to as N_D. It is also worth observing that only a small increase in delay occurs when using more stages than is delay optimal, i.e., for $N > N_D$. These results agree with those shown for long-channel devices [5].

POWER DISSIPATION

Assuming that the static power dissipated due to leakage current is negligible as compared with 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, as

$$P_{tot} = V_{DD}^2 f\left(1 + K_{PSC}\right)\left(C_x + \left(\frac{C_L}{C_y}\right)^{\frac{1}{N}} C_y\right)\left(\frac{\frac{C_L}{C_y} - 1}{\left(\frac{C_L}{C_y}\right)^{\frac{1}{N}} - 1}\right), \quad (7)$$

where

$$K_{PSC} = \left(\frac{0.9}{0.8} + \frac{V_{D0}}{0.8\,V_{DD}} ln\frac{10\,V_{D0}}{e\,V_{DD}}\right)\frac{1}{(\alpha+1)}\,\frac{1}{2^{\alpha-1}}\,\frac{(1-2\nu_T)^{\alpha+1}}{(1-\nu_T)^{\alpha}} . \quad (8)$$

It is important to note that (7) has no global minimum. It is a continuously increasing function of N. Equation (7) demonstrates that using fewer buffer stages, and consequently larger values of F, reduces both short-circuit and dynamic power dissipation within the buffer. This conclusion is similar to that drawn by Veendrick [6], although he addressed only short-circuit power dissipation and long-channel devices.

Assuming that switching frequency, f, is independent of buffer design, a normalized version of (7) is depicted in Figure 3. Unlike the propagation delay of a tapered buffer system, the total power dissipation graph shown in Figure 3 depicts no local minima. The shape of the graph demonstrates a steady increase in power dissipation for increasing values of N.

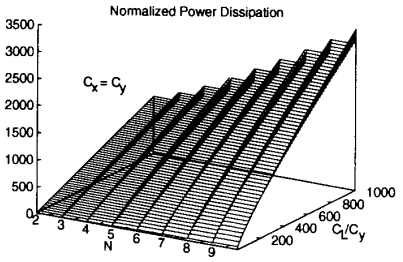

Figure 3. Normalized total power dissipation of a tapered buffer system

PHYSICAL AREA

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 [7]. Thus, the physical area required for the i^{th} stage may be expressed as

$$A_i = A_{OH} + F^{i-1} A_{ctv} . \quad (9)$$

The total area of a tapered buffer as a function of N is expressed in (10) by summing A_i for N stages. Note that both area terms in (10) increase with increasing N.

$$A_{total} = N \cdot A_{OH} + \left(\frac{\frac{C_L}{C_y} - 1}{\left(\frac{C_L}{C_y}\right)^{\frac{1}{N}} - 1}\right) A_{ctv} \quad (10)$$

Assuming the area overhead of a minimum sized inverter is three times the active area, i.e., A_{OH} = 3 x A_{ctv}, a graph depicting the physical area as a function of N is shown in Figure 4.

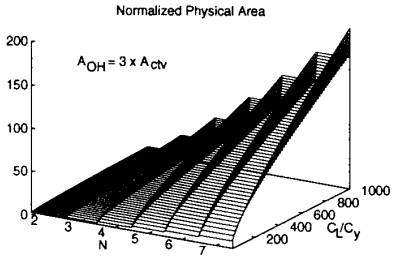

Figure 4. Normalized physical area of a tapered buffer system

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 [8]. The degradation experienced by the NMOS devices in an inverter is typically much greater than that experienced by the PMOS devices. Therefore, only the degradation of the NMOS devices is considered here.

Sun, Leblebici, and Kang describe an analytic expression for the hot-carrier degradation of the NMOS devices within a tapered buffer [8]. Utilizing their expressions, the average bond-breaking current density in the NMOS devices of a tapered buffer may be expressed as

$$\langle J_{BB} \rangle = f\left(C_x + \left(\frac{C_L}{C_y} \right)^{\frac{1}{N}} C_y \right) \langle J_{BB_0} \rangle , \qquad (11)$$

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]. Note that the formula for $\langle J_{BB} \rangle$ has been extended in this paper from that presented in [8] to include the split-capacitor model.

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 Figure 5.

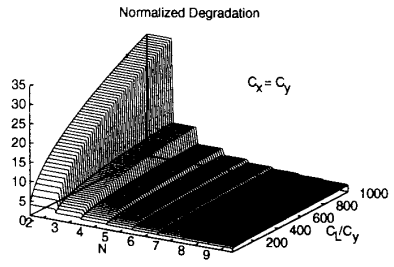

Figure 5. Normalized degradation of a tapered buffer system

The shape of the degradation graph shown in Figure 5 is similar to that of the graph depicting propagation delay in Figure 2 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.

UNIFICATION

In the previous sections of this paper, analytical expressions for propagation delay, power dissipation, physical area, and hot-carrier degradation are presented using a single nomenclature. The unification of these performance criteria will be examined in this section.

Figures 2 – 5 graphically present the behavior of each of the four performance criteria with respect to variations in N. Utilizing these figures, the effects of deviating from 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 hot-carrier degradation benefit of increasing N beyond N_D is minimal in comparison to the increased 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. It is therefore 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 for very small values of N.

Delay-power-area-degradation Product

One strategy to permit further examination of these conflicting behaviors is to investigate the integrated effects of propagation delay, power dissipation, physical area, and hot-carrier degradation, depicted by the product of (3), (7), (10), and (11). This product gives a figure of merit based on equal weighting of all four design criteria. The minimum product represents a choice of N that provides the optimal buffer implementation.

Figure 6 depicts the delay-power-area-degradation product for $10 \leq C_L/C_y \leq 100$. From this graph, it is shown that over much of this range, 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. This pairing of optimum and near-optimum N providing for both logic polarities is characteristic of the delay-power-area-degradation product.

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 (12) represents these two approximately equivalent choices for optimal N, one with logical inversion and one without.

$$N_{opt} = (i, j) \qquad (12)$$

Once the optimal number of stages, N_{opt}, is chosen, the optimal tapering factor, F_{opt}, may be computed from (2). F_{opt} will also have two values, each of which corresponds to one of the two values of N_{opt}.

Table I compares the number of stages and tapering factor which produce the minimum delay (N_D and F_D) with the optimal number of stages and tapering factor (N_{opt} and

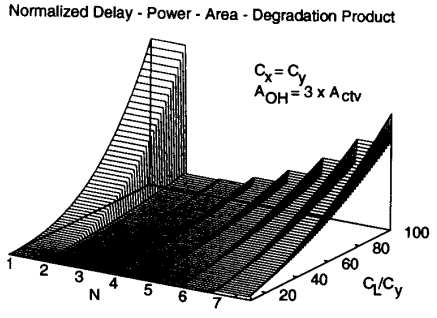

Figure 6. Delay-power-area-degradation product of a tapered buffer with load capacitance, $10 \leq C_L/C_y \leq 100$

Table I. Comparison of N and F for minimum propagation delay versus unified methodology

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)

F_{opt}), where optimal is defined by the minimum delay-power-area-degradation product, under the conditions that $C_x = C_y$ and $A_{OH} = 3 \times A_{ctv}$.

Thus, as shown in Table I, N_{opt} does not increase with increasing load capacitance (C_L/C_y) as quickly as N_D does. This is due to one component of the delay-power-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 solely on minimizing propagation delay.

Also noteworthy is that for $C_x \approx C_y$, a lower bound on N_{opt} appears as approximately $N = \left\lfloor log_{10} \frac{C_L}{C_y} \right\rfloor$. With the powers of 10 for C_L/C_y shown in Table I, this manifests itself as $F_{opt} = 10$ in Table I. Additionally, note that $F_D > e$. This 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.

Using the delay-power-area-degradation product developed in this paper, a tapered buffer may be designed so as to optimize these four performance criteria. The delay-power-area-degradation product is evaluated to discern the number of stages which produces a minimum and near-minimum. Once the number of stages for the tapered buffer implementation is determined, the tapering factor is computed from (2).

A look-up table similar to that shown above may be generated for a given process technology. The delay-power-area-degradation product is evaluated for varying C_L to establish load capacitance ranges over which N_{opt} is constant. This permits the optimal number of stages to be directly determined based on an application-specific load capacitance. The use of this technology-dependent look-up table eliminates the need for the delay-power-area-degradation product to be evaluated for each buffer instantiation.

CONCLUSIONS

A CMOS integrated circuit designer is often faced with multiple, conflicting design criteria when confronted with the task of quickly 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. The behavior of each design criterion as a function of N leads to the important 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.

The delay-power-area-degradation product is investigated to examine this conclusion. 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 provides for both logically inverted and non-inverted tapered buffer systems which are, for all practical purposes, equivalent in delay-power-area-degradation product.

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

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] 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.

[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] F. S. Lai, "A Generalized Algorithm for CMOS Circuit Delay, Power, and Area Optimization," *Solid-State Electronics*, Vol. 31, pp. 1619–1627, November 1988.

[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.