Tapered Buffers for Gate Array and Standard Cell Circuits

Brian S. Cherkauer and Eby G. Friedman

Department of Electrical Engineering
University of Rochester
Rochester, New York 14627

Abstract – **Tapered buffer systems are often used in CMOS circuits to drive large capacitive loads. Well accepted tapered buffer design practices neglect the effects of local interconnect capacitance between the buffer stages. However in many design methodologies, particularly semi-custom ASICs based on gate array or standard cell circuits, buffer stages may not be physically abutted, and therefore significant stage-to-stage local interconnect capacitance may exist. This paper presents a design methodology and analytic relationships for the optimal tapering of cascaded buffers which consider the effects of local interconnect capacitance. The method, *constant capacitance-to-current ratio tapering* (C^3RT), is based on maintaining the capacitive load to current drive ratio constant, and therefore the propagation delay of each buffer stage also remains constant. Significant reductions in power dissipation and active area, as well as reduced propagation delay, are exhibited as compared with tapered buffers which neglect local interconnect capacitance.**

I. Introduction

Large capacitive loads are common within CMOS integrated circuits, particularly at output pads and on-chip circuitry driving large fanout and/or long interconnect lines. Drivers are therefore required to source and sink relatively large currents while not degrading the performance of the signal path by placing too large a capacitive load on previous stages. In CMOS, a tapered buffer system is often used to perform this task, particularly when the load is predominantly capacitive [1–5]. When significant resistance is also associated with the load, such as might be encountered in a highly resistive interconnect line, a repeater, which is a form of distributed buffer, may be used rather than a tapered buffer [6,7].

Standard practice in CMOS tapered buffer design is to assume negligible internal local interconnect capacitance between stages. However, in circuit implementations where large capacitive loads must be driven, such as in global clock distribution or cross-chip data paths, local interconnect capacitance between buffer stages may significantly alter the performance characteristics of the tapered buffer system. In design methodologies based on channel routing, such as in gate array or standard cell circuits, local interconnect capacitance between buffer stages may be on the order of tens to hundreds of femtofarads. Even physically abutted buffer stages in structured custom designs may have tens of femtofarads of local interconnect capacitance between stages. A tapered buffer system optimally designed assuming no local interconnect capacitance may be sub-optimal when stage-to-stage interconnect capacitance is considered, even for those cases where the local interconnect capacitance is small. This paper presents a design methodology to determine the transistor sizes within a tapered buffer system which minimizes propagation delay while reducing the power dissipation and physical area once the local interconnect capacitance is determined.

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

II. Buffer Design Neglecting Local Interconnect Capacitance

Lin and Linholm first introduced the CMOS tapered buffer in 1975 [1]. This structure consists of a series of CMOS inverters, where each transistor channel width is a fixed multiple, F, larger than that of the previous inverter. Lin and Linholm show that for a buffer system consisting of N cascaded inverters, the minimum propagation delay through the buffer system is achieved when the output current drive to output capacitive ratio of each stage in the buffer remains fixed. Assuming a simplified capacitance model in which the interstage capacitance is directly proportional to the size of the input capacitance of the following inverter, each stage is a fixed ratio larger than the previous stage, a configuration referred to as a *fixed-taper buffer* (FT).

Immediately following [1], Jaeger proposed a modification of the optimization process which considered only speed optimization [2]. He demonstrated that the minimum system delay is achieved when the ratio between transistor channel widths, W_i, in adjacent stages, F, is exponentially tapered (i.e., $F = e \approx 2.72$), and the total number of stages in the buffer system, N, is $ln\ C_L/C_0$, where C_L is the load capacitance and C_0 is the input gate capacitance of a minimum sized inverter.

Jaeger's optimization scheme was enhanced with the development of the split-capacitor model [3–5]. This model provides greater accuracy than the single capacitor model originally used by Lin and Linholm and by Jaeger. With the split-capacitor model notation presented in [5], the load capacitance of the i^{th} stage of the buffer, C_{L_i}, numbered from the input stage as illustrated in Figure 1, is

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

where C_x represents the output capacitance of a minimum sized inverter (shown as stage 0 in Figure 1), C_y represents the input gate capacitance of a minimum sized inverter, and F is the tapering factor.

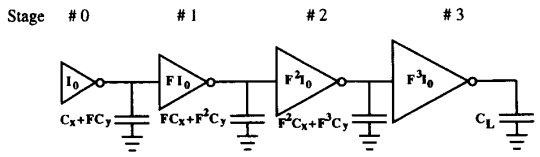

Figure 1. The split-capacitor model of a tapered buffer

III. BUFFER DESIGN WITH LOCAL INTERCONNECT CAPACITANCE

In a buffer system where each buffer stage is physically abutted with its neighboring buffer stages, interconnect capacitance is generally small, though its effects are not necessarily negligible. In many practical situations, the local interconnect capacitance between stages may be large. This situation arises when buffer stages are not physically abutted due to floorplanning considerations, which may occur, for example, when the cascaded buffers are placed in separate functional blocks or different rows of cells within an integrated circuit.

The sizing method presented in this paper determines the optimal geometric size of each buffer stage once the local interconnect capacitance is quantified, such that the load capacitance to current drive ratio of each stage remains constant, ensuring that the propagation delay of the total buffer system is minimal. This tapering methodology is referred to in this paper as *constant capacitance-to-current ratio tapering* (C^3RT).

The capacitance to current drive ratio of each stage must be constant for all stages, as shown in (2), to minimize the tapered buffer system delay, where K is the constant capacitance to current drive ratio in units of seconds/volt.

$$\frac{C_{L_i}}{I_i} = K \quad \forall i \tag{2}$$

As the local interconnect capacitances between stages are independent of the transistor dimensions, it may not be assumed that the geometric width of each stage of the buffer should be a fixed ratio larger than the previous stage. Therefore, a geometric size ratio, S_i, is defined for each stage which is the ratio of the channel width-to-length of the transistors in the i^{th} stage to the channel width-to-length of the initial minimum sized inverter (stage 0) of the tapered buffer system, as shown in (3).

$$\left(\frac{W}{L}\right)_i = S_i \left(\frac{W}{L}\right)_0 \tag{3}$$

Thus, the current drive of the i^{th} stage is

$$I_i = S_i I_0 \ , \tag{4}$$

and the capacitive load of the i^{th} stage is

$$C_{L_i} = \begin{cases} S_i C_x + S_{i+1} C_y + C_{int_i} \ , & 0 \le i < N \\ S_i C_x + C_L + C_{int_i} \ , & i = N \ , \end{cases} \tag{5}$$

where C_{int_i} represents the local interconnect capacitance at the output of the i^{th} stage, and C_L is the load capacitance of the tapered buffer system. Substituting (4) and (5) into (2) for all N stages produces the matrix equation shown in (6).

$$\begin{bmatrix} J & C_y & 0 & \cdots & 0 \\ 0 & J & C_y & 0 & \cdots & 0 \\ \vdots & & \ddots & \ddots & \ddots & \vdots \\ 0 & \cdots & 0 & J & C_y & 0 \\ 0 & \cdots & \cdots & 0 & J & C_y \\ 0 & \cdots & \cdots & \cdots & 0 & J \end{bmatrix} \begin{bmatrix} 1 \\ S_1 \\ S_2 \\ \vdots \\ S_N \end{bmatrix} = - \begin{bmatrix} C_{int_0} \\ C_{int_1} \\ C_{int_2} \\ \vdots \\ C_{int_N} + C_L \end{bmatrix} \tag{6}$$

where

$$J = C_x - K I_0 \ . \tag{7}$$

It is important to note that J in (6) is an unknown, as the value of K has not yet been determined. Thus there are N+1 unknowns in the system: the size of stages 1 through N, represented by S_1 through S_N, and J.

The elimination of S_1 through S_N from the above system results in (8), permitting the direct determination of an optimal value of J.

$$J^{N+1} + \left[\sum_{i=0}^{N-1} (-C_y)^i C_{int_i} J^{N-i}\right] + (-C_y)^N (C_{intN} + C_L) = 0 \tag{8}$$

Solution of (8) for J may be accomplished by using a numerical technique, such as the Newton-Raphson method [8]. This approach produces N+1 possible solutions, of which only one value is of practical use. In order to physically realize a tapered buffer system, J must be a negative real number, empirically shown to be of the same order of magnitude as the local interconnect capacitances.

Once J is determined, the values of S_1 through S_N are derived through substitution into (6), resulting in (9). The values of S_1 through S_N are used to size the tapered buffer system, such that the load capacitance to current drive ratio is constant for each stage. Note that since stage 0 is a minimum sized buffer, S_0 = 1.

$$S_{i+1} = \frac{-Cint_i - JS_i}{C_y} \tag{9}$$

Implicit in the C^3RT methodology is the assumption that N and the local interconnect capacitances are known. In order to determine these values, the classical techniques described in Section II [2,5] are applied to an exploratory design of the tapered buffer system. Furthermore, the C^3RT sizing method presented here is equally applicable to optimizing other criteria in tapered buffer design, such as low power dissipation [9] and reliability [10], which also require the capacitance to current drive ratio of each stage remaining constant.

The C^3RT methodology is similar to the design of repeaters [7] in that it considers interstage capacitance. However, it differs from repeater design in two significant ways. First, in the C^3RT methodology the interstage resistance is assumed to be negligible compared with the transistor on-resistance of the buffer stages, whereas with the long lines being driven by repeaters, the distributed RC impedance of the interconnect line must be considered. Second, in the design of repeater circuits, an unconstrained ability to partition the interconnect lines into lower impedance sections is assumed. In the C^3RT methodology, the capacitive load is not partitionable, and the interstage capacitance results from placement and routing constraints. Thus, the C^3RT methodology is suited for driving large capacitive loads where interstage capacitance results from buffer stage placement and routing, while repeaters are suited for driving distributed RC interconnect lines.

IV. EXPERIMENTAL RESULTS

The importance of local interconnect capacitance in the design of tapered buffers varies from small to significant, depending upon the relative magnitude of the local interconnect capacitances. Large interconnect capacitance and proximity to the input of the buffer system have greater significance than small capacitances or proximity to the output of the buffer system, as the interconnect capacitance is proportionally less significant closer to the output since the input gate capacitance of the following stage is larger. In general, a C³RT implementation results in a buffer which is faster, dissipates less power, and requires less physical area than a fixed-taper buffer which neglects the effects of stage-to-stage local interconnect capacitance.

This sizing technique has been applied to an example five-stage tapered buffer system designed using both the FT method and the C³RT method and compared in Table I. In this example, C_x = 10 fF, C_y = 25 fF, C_L = 5 pF, f = 10 MHz, and the local interconnect capacitance between stage 1 and stage 2 of the buffer (C_{int_1}) is varied. Considering parallel plate and fringing capacitance [11], the interconnect capacitance between physically abutted buffer stages with minimum width (3 μm) interconnect lines in 2.0 μm technology is approximately 10 fF. Therefore, the local interconnect capacitance between the remaining stages is assumed to be 10 fF. In circuits which utilize channel routing, such as gate array or standard cell circuits, or when greater than minimum width interconnect lines are used to reduce electromigration failure in these high current drive buffer systems, the interconnect capacitance between stages is much greater. Thus, the results presented in Table I are conservative since the larger the local interconnect capacitance, the more advantageous the C³RT design method becomes. The percentages shown in Table I indicate the relative magnitude of each performance characteristic: propagation delay, power dissipation, and active area of the C³RT buffer as compared with the FT buffer. Thus, a value less than 100% indicates an improvement in the performance characteristics of the C³RT buffer as compared with the FT buffer. Propagation delay and power dissipation values are derived from SPICE [12].

Table I. Comparison of buffer system characteristics neglecting interconnect capacitance (FT) and including interconnect capacitance (C³RT)

C_{int_1}	Propagation Delay			Power Dissipation			Active Area		
	FT (ns)	C³RT (ns)	relative value	FT (mW)	C³RT (mW)	relative value	FT (μm^2)	C³RT (μm^2)	relative value
10 fF	2.29	2.28	99.4%	1.42	1.38	97.2%	1900	1795	94.5%
100 fF	2.56	2.54	99.1%	1.44	1.32	91.7%	1900	1591	83.7%
250 fF	2.87	2.85	99.3%	1.48	1.25	84.4%	1900	1313	69.1%
500 fF	3.38	3.32	98.0%	1.56	1.21	77.6%	1900	1024	53.9%

Table I demonstrates that for a specific example, the C³RT buffer has improved performance characteristics over the FT buffer. The propagation delay exhibits a small improvement of up to 2%, with larger values of local interconnect capacitance tending to exhibit increased improvement in propagation delay. Power dissipation reductions of up to 22% are shown with increasing local interconnect capacitance. Also noteworthy is the steady absolute decrease in power dissipation with increasing local interconnect capacitance for the C³RT buffer. This may appear counter-intuitive, as greater capacitance leads to greater power dissipation, and indeed this is the case with the FT buffer. However, note that the active area of the buffer also decreases with increasing interconnect capacitance, and this leads to a reduction in overall capacitance, and hence a reduction in overall power dissipation. This phenomenon is further examined below. Active area reductions of up to 46% are also shown for this example, with area improvements increasing with increasing interconnect capacitance. In general, a fixed-taper buffer implementation based on the split-capacitor model is non-optimal when interconnect capacitance is not considered.

Also noteworthy is that with C_{int_1} = 500 fF, stage 2 in this example circuit is actually smaller than stage 1. Unlike the FT method, it is possible for the C³RT methodology to produce a tapered buffer in which a particular stage may be smaller than the previous stage, i.e., a tapering factor of less than unity between two stages. This phenomenon occurs when relatively large local interconnect capacitances are present at one or more nodes. However, the overall performance characteristics of the tapered buffer system will be improved despite the less than unity tapering factor.

In Figure 2, the effects of the magnitude of the local interconnect capacitance on tapering factor (the ratio of sizes of adjacent stages), $F_i = S_i/S_{i-1}$, are illustrated for a seven-stage buffer. In this example, all local interconnect capacitances are 10 fF, with the exception of C_{int1} and C_{int4}, both of which are 100 fF. As shown in the graph, the 100 fF capacitive load between stages 1 and 2 dramatically reduces the tapering factor from 3.06 to 1.78 between those two stages. This occurs since 100 fF is comparable in magnitude to the input and output stage capacitances seen at that node. Thus, stage 2 is smaller than the fixed-taper implementation, thereby reducing the gate input capacitance at the output of stage 1. A higher tapering factor than the fixed-taper solution, however, is necessary in the remaining stages. This process can be thought of as shifting the capacitive load toward the output of the chain, where the devices are less sensitive to the interconnect capacitance.

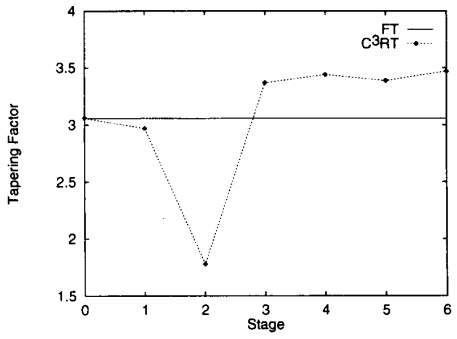

Figure 2. Comparison of tapering factor of a seven-stage buffer for FT and C³RT design methods

Note in Figure 2 that the 100 fF interconnect capacitance between stages 4 and 5 has minimal effect on the tapering factor between these stages, reducing it from 3.44 to 3.39. This small reduction in tapering factor occurs because the 100 fF local interconnect capacitance is much smaller than the input and output capacitances in the latter stages of the tapered buffer system.

In Figure 3, the total geometric width of the split-capacitor fixed-taper buffer system is compared to the C³RT buffer, using the same seven-stage example circuit as in Figure 2. A semi-log scale is used to adequately display the different orders of magnitude of the width between the initial and final stages. The percentages depicted in the figure are the relative physical widths of the C³RT implementation compared with an FT implementation. Note that each stage after the initial minimum sized stage requires less area with the C³RT design method. Thus, increasing the tapering factor after stage 3 does not translate into physically larger stages. This overall reduction in transistor dimensions, and therefore a similar reduction in total device capacitances, is the primary reason for the reductions in physical area and power dissipation shown in Table I that occurs when local interconnect capacitance is considered during the design of the tapered buffer system.

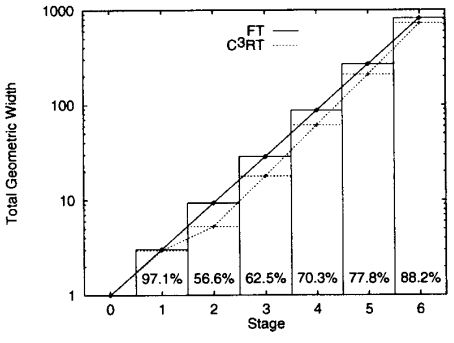

Figure 3. Comparison of transistor channel widths of a seven-stage buffer for FT and C³RT design methods

Note that in Figure 3 the FT method produces a straight line on a log scale, whereas the C³RT method does not. This different behavior occurs since with fixed tapering, each stage is constrained to be F larger than the previous stage, and thus appears linear on a log scale. No such relationship exists with the C³RT method.

V. Conclusions

CMOS tapered buffers are frequently used to drive large capacitive loads which arise from long global interconnect lines, such as clock distribution networks, high capacitance fanout, and off-chip loads. Typically, local interconnect capacitance is assumed to be negligible during the design of these tapered buffer systems. However, interconnect capacitance within the buffer system can be significant, particularly when floorplanning considerations require the buffer to be located in separate functional blocks or different rows of cells. A methodology for designing optimally tapered buffer systems which considers local interconnect capacitance is presented here. This method, C³RT, permits a tapered buffer to be optimized to its specific physical environment.

Tapered buffer systems designed with this method are shown to have improved performance characteristics in the presence of local interconnect capacitance over those buffer systems in which the design method neglects local interconnect capacitance. For a specific example, reductions in power dissipation of up to 22% and reductions in active area of up to 46% coupled with reductions in propagation delay of up to 2% are exhibited using the C³RT method as compared with traditional fixed-tapered buffers. Thus, significant performance improvements can be attained by considering the effects of local interconnect capacitance during the design of 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] A. Kanuma, "CMOS Circuit Optimization," *Solid-State Electronics*, Vol. 26, pp. 47–58, 1983.

[4] N. Hedenstierna and K. O. Jeppson, "CMOS Circuit Speed and Buffer Optimization," *IEEE Transactions on Computer-Aided Design*, Vol. CAD-6, pp. 270–281, March 1987.

[5] 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.

[6] H. B. Bakoglu and J. D. Meindl, "Optimal Interconnection Circuits for VLSI," *IEEE Transactions on Electron Devices*, Vol. ED-32, No. 5, pp. 903–909, May 1985.

[7] S. Dhar and M. A. Franklin, "Optimum Buffer Circuits for Driving Long Uniform Lines," *IEEE Journal of Solid-State Circuits*, Vol. SC-26, pp. 32–40, January 1991.

[8] W. H. Press, B. P. Flannery, S. A. Teukolsky, and W. T. Vetterling, *Numerical Recipes in C, The Art of Scientific Computing*. Cambridge University Press, 1988.

[9] 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.

[10] 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.

[11] J.-H. Chern, J. Huang, L. Arledge, P.-C. Li, and P. Yang, "Multilevel Metal Capacitance Models For CAD Design Synthesis Systems," *IEEE Electron Device Letters*, Vol. EDL–13, No. 1, pp. 32–34, January 1992.

[12] S. M. Kang, "Accurate Simulation of Power Dissipation in VLSI Circuits," *IEEE Journal of Solid-State Circuits*, Vol. SC-21, No. 5, pp. 889–891, October 1986.