Design of Tapered Buffers with Local Interconnect Capacitance

Brian S. Cherkauer and Eby G. Friedman

***Abstract*—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. 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%, as compared with tapered buffers which neglect local interconnect capacitance, are exhibited for an example buffer system.**

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]–[4].

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 design methodologies may have tens of femtofarads of local interconnect capacitance between stages. A tapered buffer system optimally designed assuming no local interconnect capacitance may be suboptimal 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.

The paper is composed of the following sections. In Section II, standard techniques for the optimal design of tapered buffers neglecting local interconnect capacitance are summarized. The focus of this paper is the sizing technique for optimizing the design of tapered buffers assuming local interconnect capacitance. This topic is described in Section III. Experimental results verifying this sizing method are presented in Section IV. Finally, a summary with some conclusions is presented in Section V.

Manuscript received March 7, 1994; revised September 27, 1994. This work was supported by the National Science Foundation, Grant MIP-9208165.

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

IEEE Log Number 9408064.

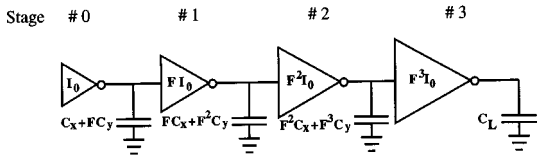

Fig. 1. The split-capacitor model of a tapered buffer.

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 capacitance ratio of each stage in the buffer remains fixed. For this case, 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 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 considers 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., $F = e$), and the total number of stages in the buffer system, N, is $\ln C_L/C_0$, where C_L is the load capacitance being driven by the tapered buffer system, and C_0 is the input gate capacitance of the minimum sized buffer stage.

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 proposed by Lin and Linholm and by Jaeger. With the split-capacitor model notation presented in [3], the load capacitance of the i^{th} stage of the buffer, C_{L_i}, numbered from the input stage as illustrated in Fig. 1, is

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

where C_x represents the output capacitance of a minimum sized inverter (shown as stage 0 in Fig. 1), C_y represents the

input gate capacitance of a minimum sized inverter, and F is the tapering factor.

The capacitive load to current drive ratio of each inverter stage is constant, as shown in (2) where I_0 is the current drive of a minimum sized inverter, ensuring that the propagation delay of each stage is also constant [1], [3].

$$\frac{C_{L_i}}{I_i} = \frac{C_x + FC_y}{I_0} \tag{2}$$

The number of stages, N, calculated from the split-capacitor model developed by Li, *et al.*, in [3], is

$$N = \frac{\ln \frac{C_L}{C_y}}{\ln F}. \tag{3}$$

The tapering factor for minimum delay is determined from the transcendental relationship in F shown below [3].

$$F[\ln(F) - 1] = \frac{C_x}{C_y} \tag{4}$$

These equations are used in Section III to develop an initial design of a tapered buffer system in order to estimate the local interconnect capacitance between buffer stages.

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.

Since the split-capacitor model only considers the input and output transistor capacitances, local interconnect capacitance adds to the capacitive load of each stage. This has the effect of altering the stage-dependent load capacitance to current drive ratio. As the local interconnect capacitance is neither proportional to the geometric size of the stages nor constant for each stage, each stage of the buffer has a different load capacitance to current drive ratio if a fixed-taper methodology is used. The sizing method presented in this paper determines the optimal geometric size of each buffer stage once the local interconnect capacitance is determined, 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 (5), to minimize the delay of the tapered buffer system, where K is the constant capacitance to current drive ratio in units of seconds/volt.

$$\frac{C_{L_i}}{I_i} = K \quad \forall i \tag{5}$$

K is related to the rise and fall time of an inverter, as shown in (6).

$$\tau = KV_{\mathrm{DD}} \tag{6}$$

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 as 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 (7).

$$\left(\frac{W}{L}\right)_i = S_i \left(\frac{W}{L}\right)_0 \tag{7}$$

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

$$I_i = S_i I_0, \tag{8}$$

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

$$C_{L_i} = \begin{cases} S_i C_x + S_{i+1} C_y + C_{\mathrm{int}_i}, & 0 \le i < N \\ S_i C_x + C_L + C_{\mathrm{int}_i}, & i = N, \end{cases} \tag{9}$$

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 (8) and (9) into (5) for all $N+1$ stages produces the matrix equation shown in (10).

$$\begin{bmatrix} J & C_y & 0 & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & 0 \\ 0 & J & C_y & 0 & \cdot & \cdot & \cdot & \cdot & \cdot & 0 \\ 0 & 0 & J & C_y & 0 & \cdot & \cdot & \cdot & \cdot & 0 \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ 0 & \cdot & \cdot & \cdot & \cdot & 0 & J & C_y & 0 & 0 \\ 0 & \cdot & \cdot & \cdot & \cdot & \cdot & 0 & J & C_y & 0 \\ 0 & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & 0 & J & C_y \\ 0 & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & 0 & J \end{bmatrix} \times \begin{bmatrix} 1 \\ S_1 \\ S_2 \\ \cdot \\ \cdot \\ \cdot \\ \cdot \\ \cdot \\ \cdot \\ S_N \end{bmatrix} = - \begin{bmatrix} C_{\mathrm{int}_0} \\ C_{\mathrm{int}_1} \\ C_{\mathrm{int}_2} \\ \cdot \\ \cdot \\ \cdot \\ \cdot \\ \cdot \\ \cdot \\ C_{\mathrm{int}_N} + C_L \end{bmatrix}, \tag{10}$$

where

$$J = C_x - KI_0. \tag{11}$$

It is important to note that J in (10) is an unknown, as the value of K has not as 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 (12), permitting the direct determination of an optimal value of J.

$$J^{N+1}+\left[\sum_{i=0}^{N-1}(-C_y)^i C_{\text{int}_i} J^{N-i}\right]+(-C_y)^N(C_{\text{int}_N}+C_L)=0 \tag{12}$$

Solution of (12) for J may be accomplished by using a numerical technique, such as the Newton-Raphson method [6]. 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.

Once J is determined, the values of S_1 through S_N are derived through substitution into (10), resulting in (13). 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, $S_0 = 1$.

$$S_{i+1}=\frac{-C_{\text{int}_i}-JS_i}{C_y} \qquad \text{for } 0 \le i < N \tag{13}$$

It is important to note that very large interconnect capacitances located between the early stages of the tapered buffer could result in a nonphysically realizable solution. If this occurs, this sizing method produces one or more S_i's which are less than one, implying an inverter smaller than minimum size. For a physically realizable system, the inequality shown in (14) must hold for $0 \le i \le N-1$. This inequality results from holding $S_{i+1} \ge 1$ in (13). When (14) is not satisfied, the buffer circuit must be reorganized to either reduce the large interconnect capacitances or to shift the large local interconnect capacitances to the latter stages of the buffer system. An additional buffer stage may be necessary to accomplish this. Alternatively, minimum sized buffers may be used for stages where $S_i < 1$, though this will not preserve the constant capacitance-to-current ratio.

$$S_{i+1}=\frac{-C_{\text{int}_i}-JS_i}{C_y} \ge 1 \tag{14}$$

Also noteworthy is that there is no fixed relationship between the sizes of adjacent stages in a C^{3}RT buffer. Rather, the tapering factor depends upon the magnitudes of the local interconnect capacitances. This is unlike the fixed-taper methods of [1]–[3], and also unlike the variable-taper method presented in [7] and [8], in which the expression $S_{i+1} = F^{i+1}S_i$ is used with constant F to determine the sizes of successive buffer stages. The variable-taper method of [7] and [8] produces tapered buffers with nonminimal propagation delays. Typical propagation delays are 10–15% greater than the fixed-taper method, with a hybrid variable-taper fixed-taper approach producing propagation delays 2% greater than the fixed-taper method [7],[8]. As is shown in Section IV, the C^{3}RT method produces tapered buffers with propagation delays less than the FT method. The local interconnect capacitances are assumed to be zero in [7] and [8]; however, the C^{3}RT method presented here is extendable to the variable-taper method of [7] and [8], permitting the effects of the local interconnect capacitances to be considered.

It is possible for the C^{3}RT 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. A tapering factor less than one is not possible with either the FT or the variable-taper method of [7] and [8].

Implicit in the C^{3}RT 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],[3] are applied to an exploratory design of the tapered buffer system. Furthermore, the C^{3}RT sizing method presented here is equally applicable to optimizing other criteria in tapered buffer design [9], such as power dissipation [10] and reliability [11], which also require the capacitance to current drive ratio of each stage to remain constant.

IV. Experimental Results

The importance of local interconnect 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 latter stages is larger. In general, a C^{3}RT 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 C^{3}RT 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 [12], 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 can be much greater than 10 fF. Thus, the results presented in Table I are conservative since the larger the local interconnect capacitance, the more advantageous the C^{3}RT design method becomes. The percentages shown in Table I indicate the relative magnitudes of each performance characteristic: propagation delay, power dissipation, and active area of the C^{3}RT buffer as compared with the FT buffer. Thus, a value less than 100% indicates an improvement in the performance

TABLE I
COMPARISON OF BUFFER SYSTEM CHARACTERISTICS NEGLECTING INTERCONNECT CAPACITANCE (FT) AND INCLUDING INTERCONNECT CAPACITANCE (C^3RT).

C_{int_1}	Propagation Delay			Power Dissipation			Active Area		
	FT (ns)	C^3RT (ns)	relative value	FT (mW)	C^3RT (mW)	relative value	FT (μm^2)	C^3RT (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%

characteristic of the C^3RT buffer as compared with the FT buffer. Propagation delay and power dissipation values are derived from SPICE [13].

It is demonstrated in Table I that, for a specific example, the C^3RT buffer has improved performance characteristics over the FT buffer. The propagation delay of the C^3RT buffer exhibits a small improvement of up to 2%, with larger values of local interconnect capacitance tending to exhibit increasing improvement in propagation delay. Power dissipation reductions in the C^3RT buffer 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 C^3RT 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 C^3RT buffer also decreases with increasing interconnect capacitance, and this leads to a reduction in overall capacitance, and hence a reduction in overall power dissipation. Active area reductions of up to 46% are shown for this example, with area improvements increasing with increasing local interconnect capacitance. In general, a fixed-taper buffer implementation based on the split-capacitor model is nonoptimal when interconnect capacitance is not considered. Also noteworthy is that with $C_{\text{int}_1} = 500$ fF, stage 2 in this example circuit is actually smaller than stage 1. Thus, as described in Section III, tapering factors less than unity are possible with the C^3RT method.

In Fig. 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 the five-stage buffer of Table I with $C_{\text{int}_1} = 250$ fF. As shown in the graph, the 250 fF capacitive load between stages 1 and 2 dramatically reduces the tapering factor from 3.49 to 1.02 between those two stages. This reduction in tapering factor occurs since 250 fF is comparable in magnitude to the sum of 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 (the input of stage 2). 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 local interconnect capacitance [14].

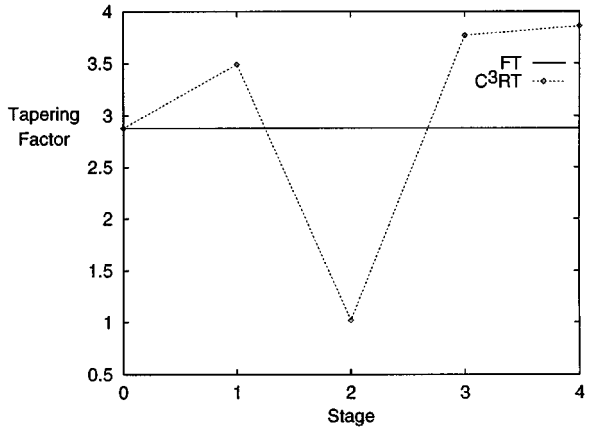

Fig. 2. Comparison of tapering factor of a five-stage buffer for FT and C^3RT design methods.

V. CONCLUSION

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, C^3RT, 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 C^3RT 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 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] 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.
[4] 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.
[5] A. Kanuma, "CMOS circuit optimization," *Solid-State Electron.*, vol. 26, no. 1, pp. 47–58, 1983.
[6] W. H. Press, B. P. Flannery, S. A. Teukolsky, and W. T. Vetterling, *Numerical Recipes in C, The Art of Scientific Computing*. Cambridge Univ. Press, 1988.

[7] 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.
[8] S. R. Vemuru and E. D. Smith, "Split-capacitive load variable taper buffer design," in *Proc. IEEE Midwest Symp. Circuits Syst.*, May 1991, pp. 815–818.
[9] B. S. Cherkauer and E. G. Friedman, "A unified design methodology for CMOS tapered buffers," *IEEE Trans. VLSI Syst.*, vol. 3, no. 1, Mar. 1995.
[10] 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.
[11] 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.
[12] J.-H. Chern, J. Huang, L. Arledge, P.-C. Li, and P. Yang, "Multilevel metal capacitance models for CAD design synthesis systems," *IEEE Electron Dev. Lett.*, vol. 13, no. 1, pp. 32–34, Jan. 1992.
[13] S. M. Kang, "Accurate simulation of power dissipation in VLSI circuits," *IEEE J. Solid-State Circuits*, vol. SC-21, no. 5, pp. 889–891, Oct. 1986.
[14] B. S. Cherkauer and E. G. Friedman, "Tapered buffers for gate array and standard cell circuits," in *Proc. IEEE Int. ASIC Conf.*, Sept. 1994, pp. 96–99.