Correspondence

CMOS Tapered Buffer

N. C. LI, MEMBER, IEEE, GENE L. HAVILAND, MEMBER, IEEE, AND A. A. TUSZYNSKI, SENIOR MEMBER, IEEE

Abstract—Jaeger's buffer comprises a string of tapered inverters. Each inverter is modeled by a capacitor and a conductor. We split the capacitor into inherent and load components (C_x and C_y), and show that the value of the optimal taper depends on the C_x/C_y ratio: the best taper exceeds Jaeger's 2.72 slope, but only moderately.

I. Background

The need for buffers at chip-crossing boundaries of MOS IC's has been highlighted by Weste and Eshraghian [1] as well as Mead and Conway [2]. The wherewithal for the design of such buffers has been scrutinized by Lin and Linholm [3], Jaeger [4], Veendrick [5], Hedenstierna and Jeppson [6], Nemes [7], and Kanuma [8]. Several topics, which bear upon approximations employed in buffer design, have been discussed by Greenbaum [9], as well as Arnout and De Man [10]. Improvements attainable by recourse to BiCMOS have been examined by Rosseel and Dutton [11], as well as De Los Santos and Hoefflinger [12].

A severe mismatch between off-chip loads and on-chip logic devices prevails in high-density CMOS circuits. In the interest of speed and power considerations, MOS transistors are laid out to minimal geometries and W/L ratios close to 1. With gate oxides of about 250 Å, the on-chip capacitance of logic devices amounts to several tens of femtofarads against an off-chip load capacitance of 50 pF or more. Thus, a speed degradation factor of three orders of magnitude would result, if the loads were connected directly to logic-level transistors. Naturally then, guided by past practice, one inserts a tapered buffer between the logic devices and the load.

II. Design of the Tapered Buffer

We begin with the Jaeger version of the Lin–Linholm approach, and then proceed to the split-capacitor modification developed by us. In Jaeger's model, each stage of the buffer is represented by one conductor and one capacitor. We use one conductor but two capacitors. The thrust of our discussion is directed at the optimization of the dynamic response of the buffer.

Jaeger's buffer and its model are shown in Figs. 1 and 2, respectively. There are n stages, numbered 0 to $n-1$. The logic-level capacitance is C_i, the logic-level conductance is g,

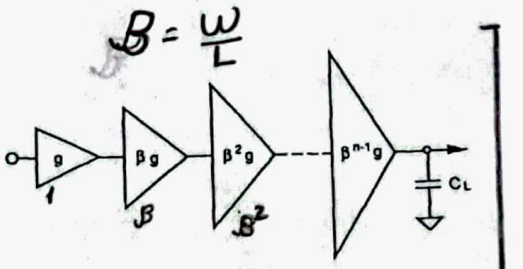

Fig. 1. Circuit configuration explored by Jaeger.

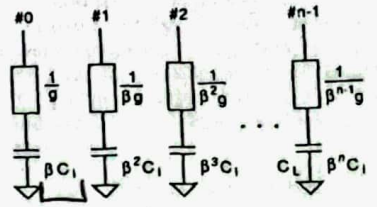

Fig. 2. Model implied in Jaeger's paper.

and the logic-level time constant $\tau_i = C_i/g$. The taper is β, i.e., the W/L ratio of stage $\#(k+1)$ is β times larger than that of stage $\#k$:

$$(W/L)_{k+1} = \beta (W/L)_k. \tag{1}$$

The conductance, capacitance, and time constant of stage $\#k$ are

$$\begin{cases} g_k = \beta^k g \\ C_k = \beta^{k+1} C_i \\ \tau_k = \beta \tau_i. \end{cases} \tag{2}$$

The overall time constant of the buffer (τ_o) is assumed to be equal to the sum of the time constants of the individual stages:

$$\tau_o = \sum_{k=0}^{n-1} (\tau_k) = n\beta\tau_i. \tag{3}$$

The load capacitance at the output stage (C_L) is

$$C_L = \beta^n C_i. \tag{4}$$

The number of stages of the buffer can, therefore, be written as

$$n = \frac{\ln(C_L/C_i)}{\ln(\beta)}. \tag{5}$$

Substitution of (5) into (3) yields

$$\tau_o = \tau_i \cdot \ln(C_L/C_i) \cdot \frac{\beta}{\ln(\beta)} \tag{6}$$

Manuscript received August 23, 1989; revised March 7, 1990.

N. C. Li is with the Department of Electrical and Computer Engineering, Northeastern University, Boston, MA 02115.
G. L. Haviland is with the Solid-State Division, Naval Ocean System Center, San Diego, CA.
A. A. Tuszynski is with the Department of Electrical and Computer Engineering, San Diego State University, San Diego, CA 92182.

IEEE Log Number 9036486.

which leads to

$$\beta\ (\text{optimum}) = e = 2.72. \tag{7}$$

Thus one arrives at an overall delay of

$$\tau_o = e \cdot [\ln(C_L/C_i)] \cdot \tau_i \tag{8}$$

and a buffer insertion penalty factor

$$B_P = e \cdot \ln(C_L/C_i). \tag{9}$$

Transition from femtofarad logic to picofarad loads incurs the still surprisingly high penalty factor of almost twenty.

III. Split-Capacitor Solution

We adopt the equivalent circuit and the summation of time constants used by Jaeger, but we split the capacitor into two parts: an inherent output capacitance C_x and an incidental load capacitance C_y (Fig. 3). The logic-level value of $C_x + C_y$ is C_i. The load capacitance of the last stage is C_L. To be included in C_x is C_e, an equivalent short-circuit current capacitance, whose maximum value is

$$C_e \approx \frac{I_P}{3V_{DD}}(\tau_r + \tau_f) \tag{10}$$

where I_P is the peak short-circuit current of the inverter, while τ_r and τ_f stand for rise and fall times, respectively. See [5] for background to (10).

The new definitions read as follows. The logic-level time constant is

$$\tau_i = (C_x + C_y)/g$$

and the time constant of stage #k is

$$\tau_k = \frac{\beta^k C_x + \beta^{(k+1)} C_y}{\beta^k g_m} \tag{11}$$

$$= \frac{C_x + C_y + (\beta - 1)C_y}{g_m} \tag{12}$$

$$= [1 + (\beta - 1)p]\tau_i \tag{13}$$

where

$$p = \frac{C_y}{C_x + C_y}. \tag{14}$$

The total delay through the buffer is

$$\tau_o = n\tau_k \tag{15}$$

where

$$n = \frac{\ln(C_L/C_y)}{\ln \beta}. \tag{16}$$

Substituting now (13) and (16) into (15), we get

$$\tau_o = \tau_i \cdot \ln(C_L/C_y) \cdot \frac{[1 + (\beta - 1)p]}{\ln \beta}. \tag{17}$$

Finally, differentiating (17) with respect to β, and invoking (14) we arrive at

$$\beta[\ln(\beta) - 1] = \frac{C_x}{C_y}. \tag{18}$$

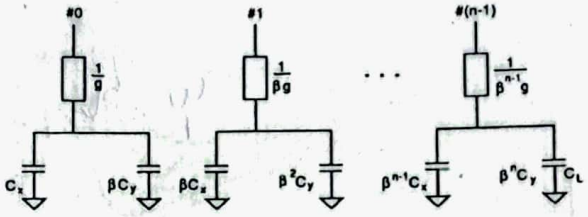

Fig. 3. Split-capacitor model.

TABLE I
TAPER AS A FUNCTION OF C_x/C_y

C_x/C_y	0	0.1	0.2	0.3	0.4	0.5	0.6	0.8	1.0	3.0
β	2.72	2.82	2.91	3.00	3.09	3.19	3.27	3.43	3.59	4.97

Optimum beta is now seen to depend on the relative magnitudes of C_x and C_y (Table I and Fig. 4). As was to be expected, if C_x is negligibly small compared to C_y, then the optimum slope reduces to $e = 2.72$, in correspondence to Jaeger's solution. Conversely, if C_x is much larger than C_y, then β may exceed 2.72 by a considerable margin (Table II). Typically, β is moderately larger than 2.72.

IV. The Fan-Out Decision

In general layout work, of special interest are nodes with low to moderate fan-out. Faced with a fan-out of k, do we or don't we use a buffer? Obviously enough, where this question arises, reference is made to a single-stage buffer, scaled as shown in Fig. 5(b). That buffer is to be compared with the straight inverter in Fig. 5(a).

Retaining the technique of linear addition of time constants, we write the overall delay in Fig. 5(b) as

$$\tau_o = \frac{C_x + \beta C_y}{g} + \frac{\beta C_x + kC_y}{\beta g} = \left[2C_x + \left(\beta + \frac{k}{\beta}\right)C_y\right]\Big/ g. \tag{19}$$

Scrutinizing (19) for best β, one arrives at

$$\beta = k^{1/2} \tag{20}$$

in confirmation of the uniform taper approach. The total delay is

$$\tau_o(\min) = 2(C_x + \sqrt{k}C_y)/g \tag{21}$$

which is to be compared with the "no buffer" delay

$$\tau_o' = (C_x + kC_y)/g. \tag{22}$$

The former is smaller than the latter when

$$2(C_x + \sqrt{k}C_y) < C_x + kC_y \tag{23}$$

that is when

$$k - 2\sqrt{k} > \frac{C_x}{C_y}. \tag{24}$$

If C_x is very small compared to C_y, then the critical value of the

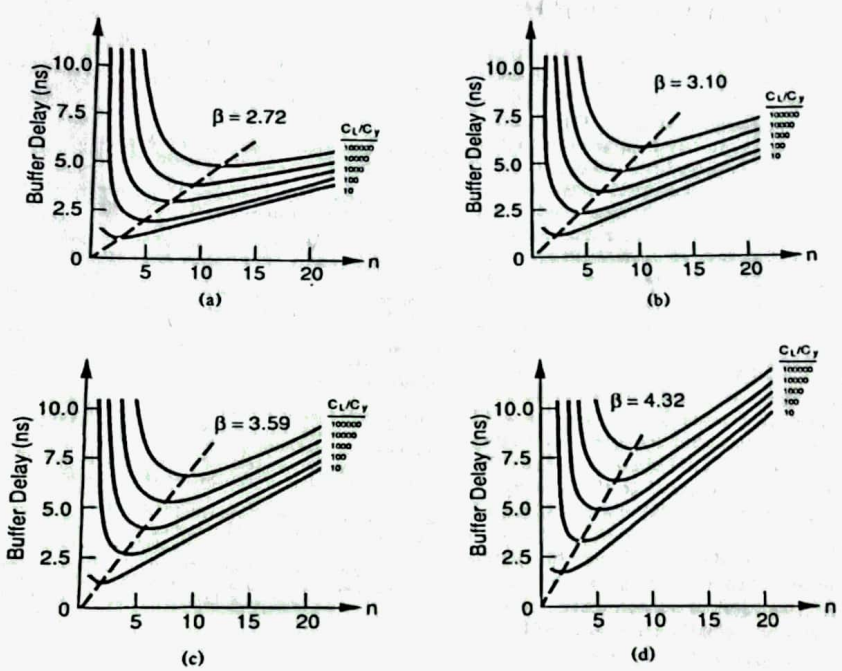

Fig. 4. Taper versus C_x / C_y: (a) $\beta = 2.72$, (b) $\beta = 3.10$, (c) $\beta = 3.59$, and (d) $\beta = 4.32$.

TABLE II
EQUATION (15) AND SPICE SIMULATION RESULTS

number of stages n	4	5	6	7	8
taper β	7.45	4.99	3.82	3.15	2.73
B_P per (15)	13.3	12.2	12.1	12.6	13.0
B_P per SPICE*	13.8	12.6	11.9	12.2	12.4

*For $C_i = 38.9$ fF, $C_L = 50$ pF, and MOSIS 1.2-μm CMOS SPICE parameters.

TABLE III
SLOPE VERSUS C_x / C_y

C_x / C_y	0	0.25	0.50	0.75	1.0	2.0	3.0
k_{cr}	4	4.49	4.95	5.39	5.83	6.46	9.0
β	2	2.1	2.2	2.3	2.4	2.5	3.0

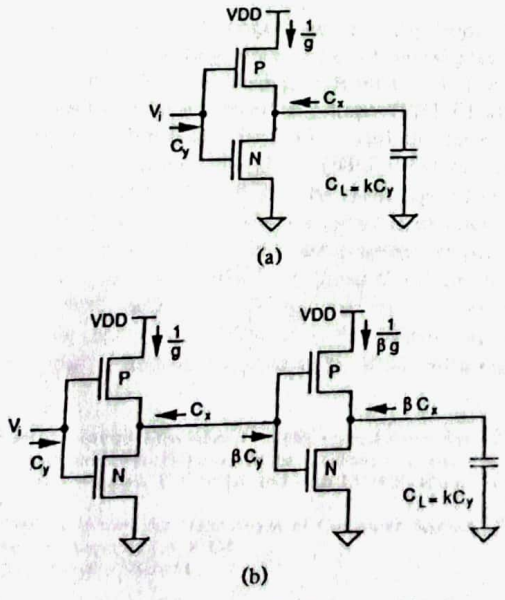

Fig. 5. The fan-out decision: (a) direct hookup and (b) buffered connection.

fan-out is

$$k_{cr}(0) = 4. \tag{25}$$

Otherwise

$$k_{cr} = 2 + 2\sqrt{(1 + C_x / C_y)} + C_x / C_y. \tag{26}$$

Equation (26) and Table III reveal that the answer to the buffer question depends on both C_x / C_y and C_L / C_y. As a rule, a buffer should be used only when C_L / C_y is larger than four.

V. CONCLUSION

The split-capacitor model leads to the conclusion that the taper is a function of C_x / C_y and, therefore, a matter of technology, i.e., it depends on feature size, gate-oxide thickness, junction capacitances, etc. For any particular load capacitance, there exists a best taper and a corresponding best number of stages, but the law relating the delay penalty to the taper of the buffer is not very strong. At on-chip distribution points, buffers are justified only where fan-out exceeds a factor of 4. However, the chip-crossing penalty of MOS implementations is severe even in the best case; therein lies a strong argument in favor of BiCMOS I/O.