Design and Optimization of Buffer Chains and Logic Circuits in a BiCMOS Environment

Muhammad S. Elrabaa, *Student Member, IEEE*, and Mohamed I. Elmasry, *Fellow, IEEE*

Abstract—**In this work the design and optimization of BiCMOS buffer chains and multilevel logic circuits are reported. BiCMOS speedup contours are introduced and analytical expressions for the delay are obtained. The speedup contours and the delay expressions were used in the design and optimization of BiCMOS buffer chains. Also, general design guidelines, which can be easily automated, for circuit design in a BiCMOS environment are given. Designing multistage mixed CMOS/BiCMOS buffers, BiCMOS complex logic gates, and multilevel CML gates is also studied and results are reported.**

Nomenclature

C_{in}, C_L	The input and the load capacitances.
L_{eff}, V_{th}, I_{sat}	MOS minimum channel length, threshold voltage, and saturation current.
$C_{\text{S/D}}$, C_{ovl}	MOS source/drain depletion and overlap capacitances.
λ	Minimum feature size.
β_m	Ratio between the PMOS and the NMOS devices in a CMOS buffer stage.
a	Ratio between the length of the S/D diffusions and λ.
f, N	Tapering factor of a CMOS buffer chain and the number of stages.
β, R_B, I_K, τ_F	Bipolar gain, base resistance, knee current, and transit time.
C_{be}, C_{jc}, C_{js}	Bipolar emitter–base junction, collector junction, and substrate capacitances.

I. Introduction

BiCMOS circuit designers are usually faced with the task of selecting the best combination of CMOS/BiCMOS/bipolar circuit structures for the design of critical paths that would render the optimum system performance in terms of speed, power, and area. The answer to this problem is not easy and straightforward in most of the design situations. This is because the BiCMOS technology offers circuit designers an environment that is very rich with different circuit structures to implement buffer chains and logic circuits. This, in turn, complicates both the selection and design processes. This paper attempts to provide a design methodology for BiCMOS circuit design that is comprehensive and yet easy to implement. More specifically it addresses the design and optimization of buffer chains and logic circuits in a BiCMOS environment.

Many researchers compared the performance of CMOS and BiCMOS gates for equal input capacitance [1], [2] and some of them provided design methodologies [2]. However, they either have limited the CMOS buffer to a single stage or equated the area of the CMOS to that of the BiCMOS buffer. However, the design of mixed CMOS/BiCMOS buffer chains has not been explored.

In Section II of this work, a general comparison between optimized CMOS buffer chains and different types of BiCMOS buffers is carried out with out any restrictions. The speedup factor of BiCMOS over CMOS is reported on speedup contours for different BiCMOS technologies. From the results of this comparison, general design guidelines are given. The effects of scaling are studied, and analytical expressions to calculate the crossover capacitance and the speedup contours are provided. Multistage mixed CMOS/BiCMOS buffers are also explored.

In Section III, the performances of complex logic implemented in CMOS and BiCMOS are compared for various conditions of complexity and loading. Other design options for implementing multilevel logic such as CML are considered. Finally, the applicability of the proposed design procedure in Section II to complex logic is evaluated.

II. Buffering Circuits in a BiCMOS Environment

Three types of buffering circuits are available in a BiCMOS technology: a CMOS buffer chain, a single-stage partial-swing BiCMOS buffer (PSBiCMOS) [1], and a single-stage full-swing BiCMOS buffer (FSBiCMOS) [1]. The schematics of the three buffering circuits are shown in Fig. 1. In this work a generic noncomplementary BiCMOS technology was assumed, hence the two types of chosen BiCMOS buffers do not contain p-n-p BJT's. The design problem is stated as follows: given a C_{in} and C_L, design a buffer (or a buffer chain) using CMOS and/or BiCMOS to achieve a given delay and/or area. The design and performance evaluation of the three types are

Manuscript received August 27, 1991; revised December 25, 1991. This work was supported in part by NSERC, ITRC, and MICRONET research grants.

The authors are with the VLSI Research Group, Department of Electrical and Computer Engineering, University of Waterloo, Waterloo, Ont., Canada N2L 3G1.

IEEE Log Number 9106950.

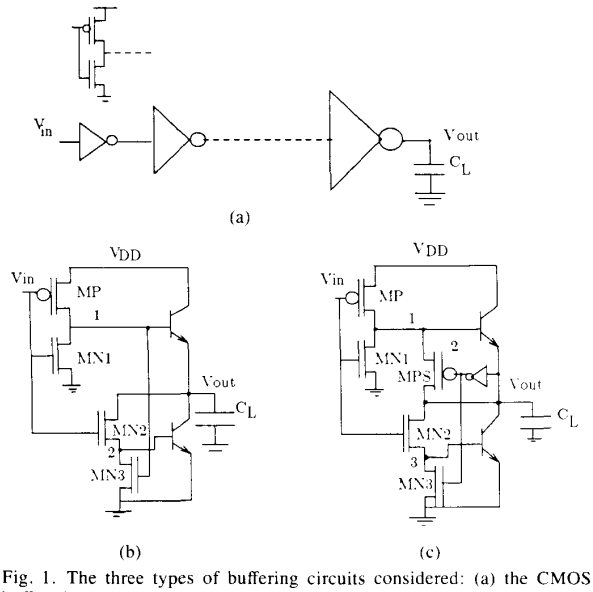


Fig. 1. The three types of buffering circuits considered: (a) the CMOS buffer chain, (b) the partial-swing BiCMOS buffer (PSBiCMOS), and (c) the full-swing BiCMOS buffer (FSBiCMOS).

presented in Sections II-A–E while multistage BiCMOS buffers are considered in Section II-F.

A. The CMOS Buffer Chain

The design procedure used for the CMOS buffer chain is the one given in [6]. The CMOS chain area was calculated as the sum of the channels area and the S/D diffusions area; hence, for N stages

$$A_{\text{CMOS}} = \lambda W_1(\beta_m + 1)(1 + 2a) \sum_{i=0}^{N-1} f^i \qquad (1)$$

where W_1 is the width of the first-stage NMOS device.

B. The BiCMOS Buffers

The sizing of the devices in the considered BiCMOS buffers was based on extensive HSPICE simulations. The bipolar transistors sizes were chosen to give a minimum delay for a load capacitance in the midrange and were maintained constant for the different values of C_{in}, i.e., the BiCMOS buffers' area is constant for the same C_{in}. This choice was based on three reasons; to avoid the overhead of sizing the emitter area at each value of C_L considered, the speed of the BiCMOS buffer saturates after a certain emitter area is reached [7], [8], [13] and the optimum emitter area does not change significantly with increasing the input capacitance above a certain value, which is usually very small [2].

C. Simulations Results

The delay of each type of buffering circuits was calculated as the average of the 50% rise and fall times measured from HSPICE [3] simulations. The input capacitance range taken was from 0.05 to 1.0 pF. For each value of the input capacitance, the load capacitance C_L was changed from 0.05 to 5 pF. The simulations were carried out for three generations of BiCMOS technologies and supply voltages. The HSPICE parameters of the reference technology (1-μm L_{eff} and V_{DD} = 5 V) are in Table I. Two generations of BiCMOS technology were generated (0.56 μm, V_{DD} = 3 V and 0.2 μm, V_{DD} = 2 V) by successively scaling the reference technology according to the general BiCMOS scaling rules described in [4] and [5]. The scaling factors for the horizontal and vertical dimensions, base–collector voltage, and the supply voltage (and the threshold voltages), denoted as K_h, K_v, K_m, and K_u, respectively, for the second and third generations are shown in Table II. An identical version of the second-generation technology with unscaled MOS threshold voltage was used for the PSBiCMOS buffer simulations. This is because PSBiCMOS has to be used in a technology with either a MOS threshold voltage $\geq V_{BE(\text{on})}$ or have two types of MOS devices, one with a threshold voltage $\geq V_{BE(\text{on})}$ and the other with a lower threshold value to eliminate the static power dissipation in the CMOS gates driven by the PSBiCMOS. Also, the PSBiCMOS buffer circuit is not considered for the third generation (V_{DD} = 2 V) due to the large deterioration of its performance.

The simulation results are presented as contours of speedup factors of the BiCMOS single buffers over the optimized CMOS buffer chain. Thus, the speedup factor is defined as

$$\text{Speedup factor} = \frac{\text{Delay of Optimized CMOS Buffer Chain}}{\text{Delay of BiCMOS Buffer}}. \qquad (2)$$

For the 1-μm, V_{DD} = 5-V BiCMOS technology, the speedup contours for the PSBiCMOS and the FSBiCMOS are shown in Figs. 2 and 3, respectively. Also shown in the same figures are the contours of equal areas of CMOS and BiCMOS and when the CMOS area is ten times that of BiCMOS. Fig. 4 shows the delay of the PSBiCMOS buffer as a function of C_{in} and C_L. Finally, to examine the effects of scaling, Fig. 5 shows the contour lines of unity speedup factor for both types of BiCMOS buffers for the three generations of BiCMOS technology. Also shown is the area contour where $A_{\text{CMOS}} = A_{\text{PSBiCMOS}} \approx A_{\text{FSBiCMOS}}$. It should be noted that, by definition, the load capacitance C_L at which the speedup factor is unity is referred to as the speed crossover capacitance C_{xov}, while the load capacitance C_L at which the area of the CMOS buffer chain is equal to that of the BiCMOS buffer is referred to as the area crossover capacitance C_{Axov}. The following are concluded from Figs. 2–5.

1) The speed of the BiCMOS buffers saturates after a certain value of C_{in} is reached, and this value increases with increasing C_L (Fig. 4).

2) The speedup factor at a certain value of C_{in} increases with increasing C_L but it begins to decrease as C_L increases further. This is specially obvious from Figs. 2 and 3 at low C_{in}. This is because as the number of stages in the CMOS chain increases (i.e., as C_L increases), its

TABLE I
HSPICE PARAMETERS OF THE REFERENCE TECHNOLOGY (1 μm, 5 V)

	V_{th}	L_{eff}	$C_{S/D}$	C_{ovl}	Bipolar
NMOS	0.85 V	1.0 μm	4E-4 F/m	2E-10 F/m	C_{jc} = 10.0 fF, C_{js} = 25.0 fF
PMOS	−0.85 V	1.0 μm	5E-4 F/m	2E-10 F/m	C_{bc} = 7.5 fF, τ_F = 12 ps
					R_B = 200 Ω, β = 100

TABLE II
THE SCALING FACTORS FOR THE SECOND AND THIRD GENERATIONS OF BiCMOS TECHNOLOGY

	K_h	K_r	K_u	K_m	λ (μm)	L_{eff} (μm)
2nd generation	0.667	0.850	0.600	0.600	0.40	0.56
3rd generation	0.375	0.630	0.667	0.667	0.15	0.20

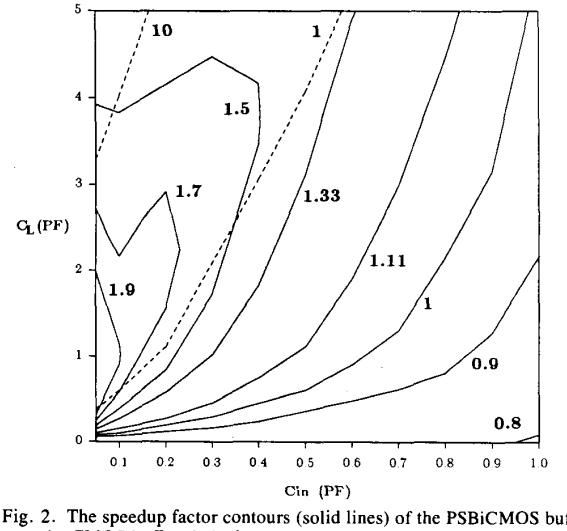


Fig. 2. The speedup factor contours (solid lines) of the PSBiCMOS buffer over the CMOS buffer chain for a 1-μm, 5-V BiCMOS technology plotted on a C_L versus C_{in} plot. The area contours $A_{\text{CMOS}} = A_{\text{PSBiCMOS}}$ and $A_{\text{CMOS}} = 10A_{\text{PSBiCMOS}}$ are also shown (dashed lines).

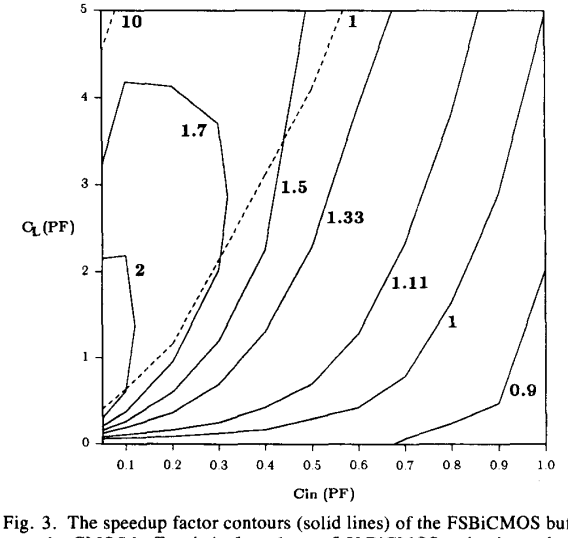


Fig. 3. The speedup factor contours (solid lines) of the FSBiCMOS buffer over the CMOS buffer chain for a 1-μm, 5-V BiCMOS technology plotted on a C_L versus C_{in} plot. The area contours $A_{\text{CMOS}} = A_{\text{FSBiCMOS}}$ and $A_{\text{CMOS}} = 10A_{\text{FSBiCMOS}}$ are also shown (dashed lines).

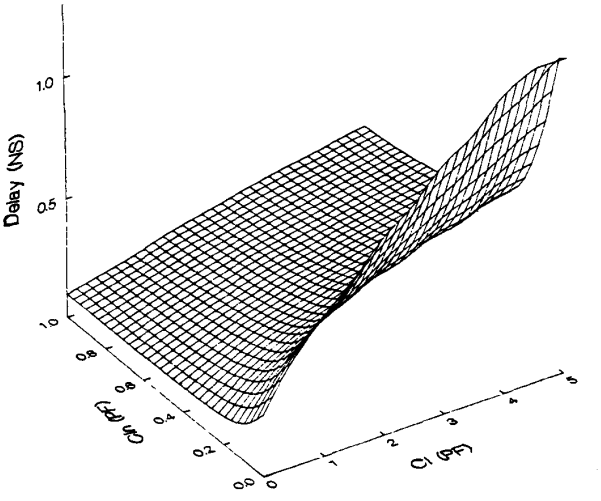


Fig. 4. The delay of the PSBiCMOS buffer versus input capacitance and load capacitance for the 1-μm, 5-V BiCMOS technology.

delay sensitivity to the load capacitance decreases. This is because the delay due to the load is only $1/N$ of the total delay, where N is the number of stages in the CMOS buffer. Fig. 6 shows how while the BiCMOS buffer delay increases linearly with the load capacitance, the CMOS delay increases at a rate that is decreasing with increasing C_L. This is why the BiCMOS buffer speedup starts to decrease at higher C_L. However, the CMOS chain area becomes much larger than that of the BiCMOS MOS buffers.

3) At any value of C_L, the BiCMOS buffers speedup factors decrease as C_{in} increases (Figs. 2 and 3).

4) The FSBiCMOS speedup factor is better than that of PSBiCMOS (Figs. 2 and 3).

5) The crossover capacitance C_{xov} increases rapidly with scaling down of devices and power supplies as shown in Fig. 5. Also, speed degradation with scaling is less in FSBiCMOS than it is in PSBiCMOS. At 2 V the FSBiCMOS is faster than the CMOS chain for small ranges of C_{in} and C_L.

As the above results indicate, it is helpful to the circuit designers to have analytical expressions for the crossover

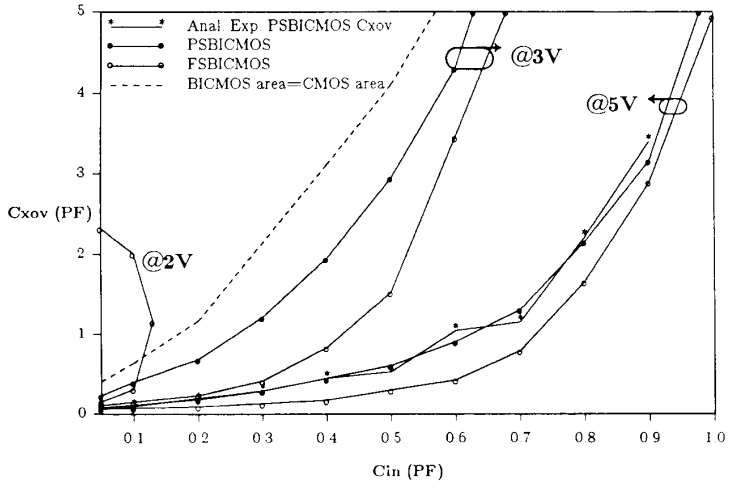


Fig. 5. The crossover capacitance of the PSBiCMOS and the FSBiCMOS buffers versus the input capacitance for the three BiCMOS technologies: 1 μm, 5 V; 0.56 μm, 3 V; and 0.2 μm, 2 V. The area contour $A_{\text{CMOS}} = A_{\text{PSBiCMOS}}$ is also shown (dashed line). Also shown is the PSBiCMOS crossover capacitance as calculated from the analytical expressions obtained.

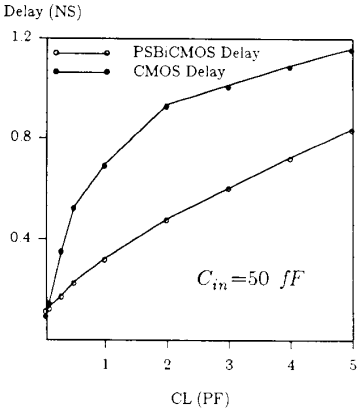


Fig. 6. The delays of the CMOS buffer chain and the PSBiCMOS buffer as a function of the load capacitance C_l. It is clear that as C_l increases, the slope of the CMOS delay decreases, while that of the PSBiCMOS remains constant.

capacitance C_{xov} and the speedup contours. Analytical expressions for the crossover capacitance are obtained next. Similar analysis could be carried out for other speedup contours.

D. Crossover Capacitance Analysis

To find C_{xov} at a certain C_{in}, the delay of the CMOS buffer chain is equated to that of the BiCMOS buffer. Analytical expressions for the delay of a CMOS inverter are given in [6] for long-channel devices and in [9] for short-channel devices. The general form of the delay of stage I is

$$T_{D\text{CMOS}} = \frac{1}{2}\left[\left(\frac{C_L}{K_N}\right)_I G_N + \left(\frac{C_L}{K_P}\right)_{I-1} B_N + \left(\frac{C_L}{K_P}\right)_I G_P + \left(\frac{C_L}{K_N}\right)_{I-1} B_P\right] \quad (3)$$

where $G_{N,P}$ and $B_{N,P}$ are constants and their values depend on whether the devices have long channels [6] or short channels [9]. Also, $K_{N,P}$ is the transconductance constant in the NMOS or PMOS saturation current equation [6], [9]. Note that at each stage both C_L and $K_{N,P}$ are proportional to the devices' widths and input capacitance. Hence, the total delay of a CMOS chain with N stages and a tapering factor f is

$$T_{D\text{chain}} = NC_{in}(g + f)\alpha \quad (4)$$

where g is the ratio of input capacitance to output capacitance of an unloaded CMOS inverter. α is a constant that depends on the technology and the first-stage input capacitance C_{in}:

$$\alpha = \left[\frac{G_N + B_P}{K_N} + \frac{G_P + B_N}{K_P}\right]_{\text{1st stage}}. \quad (5)$$

The delay of the PSBiCMOS buffer has been analyzed extensively [7], [8], [10], [13]. However, due to the complicated nature of the BiCMOS buffer transients, the delay expressions obtained lack precision and simplicity, except for the unified delay model reported in [10] which is simple and shows good correlation with simulations and experiments. But this model still lacks the required accuracy to be used in calculating the crossover capacitance C_{xov}. This is mainly due to the difficulty in choosing the right values of the device parameters such as the gain β, the transit time τ_F, and the different parasitic capacitances. The values of these parameters change during the transients. In this work, the PSBiCMOS delay was fitted by an equation similar to the one in [10]. High-level injection is assumed since this is the case for a wide range of C_{in} and C_L [7]. The general equation for the rise or fall time is given by [10]

$$T_{f(r)} = \frac{\tau_F I_C + A_e C_j V_{BE(\text{on})}}{I_{\text{sat}}} + \frac{C_L V_s}{I_{\text{sat}} + I_C}. \quad (6)$$

I_C is the collector current of the bipolar transistor, which for the high-level injection case equals $\sqrt{\beta I_K I_{sat}}$, A_e is the emitter area, C_j is the junction capacitance at the base, and V_s is the voltage swing. From (6) and noting that I_{sat} is proportional to C_{in}, the general equation to fit the average delay is

$$T_{D_{PSBiCMOS}} = A + BC_L \tag{7}$$

$$A = [A_1(C_{in})^{-1} + A_2(C_{in})^{-1/2}] \tag{8}$$

$$B = \left[\frac{A_3 + A_4(C_{in})^{-1/2}}{C_{in} + A_5(C_{in})^{-1/2} + A_6}\right] \tag{9}$$

where A_1–A_6 are empirical fitting parameters that could be obtained from eight simulations of four different C_{in} and two different C_L values for each C_{in}. These parameters may, in turn, be expressed as functions of explicit circuit parameters in the delay expression such that the circuit could be optimized under different conditions. For example, to study the effect of the emitter area A_e on the circuit performance, the parameters become

$$A_1 = a_1\sqrt{A_e}$$

$$A_2 = a_2 A_e$$

$$A_3 = \text{constant}$$

$$A_4 = a_4\sqrt{A_e}$$

$$A_5 = a_5\sqrt{A_e}$$

$$A_6 = a_6 A_e.$$

Similar treatments could be done such that the delay expression contains other circuit parameters, such as the widths of MOS devices, supply voltage, etc.

The above delay expressions of the CMOS chain and the PSBiCMOS buffer were used to generate the crossover capacitance as a function of C_{in} for the 1-μm, V_{DD} = 5-V BiCMOS technology. The results in Fig. 5 show good agreement between the calculated results and HSPICE simulations. As will be shown later, the ability to generate the contours of equal delay or at different speedup factors is very important in the decision making during BiCMOS buffer design.

E. Design Guidelines

The above results could be used to generate the following design guides. For a given C_{in} and C_L, and the following requirements on delay T_D and area A, the following design guidelines can be used.

1) Minimum T_D and no area constraints: use BiCMOS if $C_L > C_{xov}$ and CMOS otherwise.

2) Minimum A and no delay constraint: use a single-stage CMOS buffer.

3) Minimum T_D and minimum A: use BiCMOS if $C_L > C_{Axov}$ and CMOS if $C_L < C_{xov}$. If $C_{Axov} > C_L > C_{xov}$ then this design requirement cannot be met.

4) Minimum T_D and $A \leq A_{max}$: the following procedure can be used:

a) If $A_{BiCMOS} > A_{max}$, then use CMOS. CMOS design under such an area constraint is given by steps 4e)–4i). If $A_{BiCMOS} < A_{max}$, then continue this procedure.

b) Calculate the optimum number of stages and tapering factor (N_o, f_o) of the CMOS chain and then find its area from (1). If $A_{CMOS} \leq A_{max}$, then use CMOS if $C_L < C_{xov}$, and BiCMOS if $C_L > C_{xov}$. If $A_{CMOS} > A_{max}$, then continue.

c) Design a CMOS chain under the area constraint using steps 4e)–4i).

d) Calculate the delay of the BiCMOS buffer and that of the CMOS buffer chain from (4) through (9) and use the one which has the minimum delay.

The procedure of CMOS buffer chain design under area constraint A_{max}:

e) Assume that the number of stages $N = N_o - 1$.

f) Calculate the CMOS chain area from (1).

g) If $A_{CMOS} > A_{max}$, then $N = N - 1$ and go to step 4f). Otherwise continue.

h) N_o is set equal to $N + 1$ instead of N if the following condition holds:

$$(N+1)[g + (C_L/C_{in})^{1/(N+1)}] \leq N[g + (C_L/C_{in})^{1/N}]. \tag{10}$$

i) The tapering factor is adjusted using (1) such that $A_{CMOS} = A_{max}$.

Unlike what one might think, this is not a lengthy process, since the optimum number of stages N_o is usually small. For example, for a ratio C_L/C_{in} of 1000, N_o is 5.

5) Minimum A and $T_D \leq T_{max}$:

a) If $T_{D_{BiCMOS}} > T_{max}$ use CMOS if $C_L \leq C_{xov}$. The design procedure of a minimum area CMOS chain under a delay constraint is given below.

b) Starting with $N = 1$, increment N until a value is reached where the CMOS chain delay $\leq T_{max}$. If, at any time, N exceeded N_o or the CMOS chain area exceeded that of the BiCMOS, which is constant, then BiCMOS should be used.

c) The CMOS chain tapering factor is recalculated as

$$f = \frac{T_D}{\alpha N} - g \tag{11}$$

instead of $(C_L/C_{in})^{1/N}$, to give the minimum area at the required delay.

d) Calculate A_{CMOS} using (1) and use the BiCMOS buffer if $A_{BiCMOS} \leq A_{CMOS}$ and the CMOS chain otherwise.

F. Multistage BiCMOS Buffer Chains

Unlike CMOS, multistage BiCMOS buffers are seldom used. This is because the advantage of such a practice is not very obvious. However, as the above results indicate, the BiCMOS speed increases with increasing C_{in} until it reaches a certain limit. This means that there could be some situations in which using multilevel BiCMOS buff-

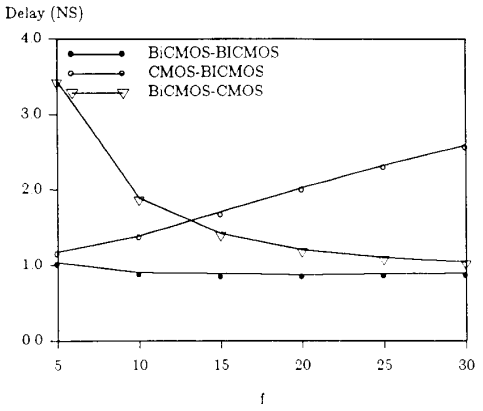


Fig. 7. The delays of the different multistage buffers versus the ratio from the input capacitance of the second stage to that of the first stage. C_{in} = 20 fF and C_L = 5 pF.

TABLE III
THE PERFORMANCE OF THE DIFFERENT CMOS/BiCMOS BUFFER CHAINS AND THE SINGLE-STAGE BiCMOS BUFFER

	CMOS Chain	Single BiCMOS	BiCMOS-BiCMOS	CMOS-BiCMOS	BiCMOS-CMOS
Delay (ns)	1.465	1.270	0.877	1.176	1.050
Normalized Area*	58.1	3.6	29.3	9.2	33.6

*Normalized to the area of the first stage of the CMOS buffer chain.

ers is advantageous. Such a situation arises when the required C_{in} is very small and C_L is very large. In this case, two BiCMOS buffer stages could be used, one with the required C_{in} and the second with a larger input capacitance, such that the delay is minimum. The value of the input capacitance of the second stage could be calculated using the delay expressions given by (4)–(9) to give the required delay under any constraint. Another option, as seen from the speedup contours, is to use a CMOS stage(s) followed by a BiCMOS stage. This option is useful under a severe area constraint or if the optimum input capacitance of the second stage is $< C_{xo1}$ for the given C_{in}. Another option is to use a BiCMOS buffer in the first stage followed by a CMOS chain with high input capacitance. Again, this option could be useful under area constraints or scaling, where BiCMOS is faster than CMOS over a very limited range. Fig. 7 shows a plot of the delay of the BiCMOS chain, the CMOS–BiCMOS chain, and the BiCMOS–CMOS chain versus f, where f is the ratio between the input capacitances of the first and second stages, C_{in} = 20 fF, and C_L = 5 pF. The number of stages in those buffers was limited to two stages only since there is no need for more stages. This can be depicted from the results above.

The simulations show that the minimum delay of the BiCMOS–BiCMOS chain occur at $f \approx 20$, while the optimum f as calculated from the delay expression was about 25. The delay of the CMOS–BiCMOS chain with $f = 5$ is larger than that of the BiCMOS–BiCMOS with $f = 20$ by about 34%. Its area, however, is about 38% of that of the BiCMOS–BiCMOS chain. The BiCMOS–CMOS chain with $f = 30$ has a delay that is 37% larger than the BiCMOS–BiCMOS minimum delay and an area which is smaller by 23%. Table III shows the optimum delay of each type depicted in Fig. 7 along with that of the CMOS chain and a single BiCMOS buffer. The areas of each type, normalized to the area of the first stage of the CMOS chain, are also shown in Table III. This table shows the trade-offs involved in using the different kinds of buffer chains. At one extreme, the BiCMOS–BiCMOS offers the highest speed. At the other extreme, the CMOS–BiCMOS offers the smallest area at a reasonable speed. Also, as Fig. 7 shows, the BiCMOS–BiCMOS chain speed does not change drastically with f. Therefore, a smaller f, hence a smaller area, could be used without much loss of speed.

III. BiCMOS Complex Logic Gates

Two types of complex logic gates are considered. The first type is an AND-OR-INVERT gate (Fig. 7(a)) and the second type is a multilevel chain of NAND gates (Fig. 7(b)). The design and simulation results for both types are presented next.

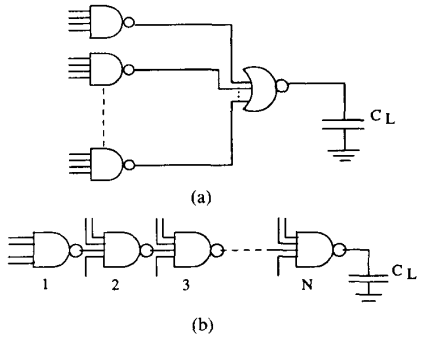


Fig. 8. The two types of complex logic gates considered: (a) a complex logic gate, and (b) a multilevel logic (MLL).

A. The AND-OR-INVERT Gate

The AND-OR-INVERT is a classical example of a complex logic function that is easy and straightforward to implement in CMOS. For this study, the fan_{in} was changed in multiples of eight, namely 8, 16, and 24, where every four inputs correspond to an AND gate of Fig. 8(a). The AND-OR-INVERT gate considered has been implemented using CMOS and FSBiCMOS, for the three generations of the BiCMOS technology. PSBiCMOS implementation was not considered because the complex gate chosen, due to the unscaled threshold voltage, will either not operate or will operate very poorly as the supply voltage is scaled down. Also the PSBiCMOS will not offer significant savings in area over the FSBiCMOS since the area of both implementations is dominated by the CMOS part.

The speedup factor of the BiCMOS implementation over the CMOS counter part is given in Fig. 9 versus the input capacitance per a single AND gate (C_{in}/AND), for different values of fan_{in}. Fig. 10 shows the speedup factor verus C_{in}/AND for the different BiCMOS technologies at $fan_{in} = 16$. The following can be observed.

1) The speedup factor decreases with increasing C_{in}/AND, for low values of fan_{in} (Fig. 9), which is the same result obtained in Fig. 2.

2) For high fan_{in}, the speedup factor decreases with C_{in} and then increases. This behavior is attributed to the fact that as fan_{in} increases, the sizing of the devices in the BiCMOS gate deviates from the optimum, hence the decrease in the speedup factor with C_{in} is faster. However, as C_{in} increases it becomes possible to size the MOS devices in the BiCMOS gate to give an optimum delay, and hence the increase in the speedup factor. Also another factor that could contribute to the reduction of the speedup factor is that the drivability of the CMOS circuit in the BiCMOS gate is reduced as fan_{in} increases, possibly forcing the BJT's to operate in the very low-current regime with large gain degradation.

3) The speedup factor, in general, decreases with scaling (Fig. 10), as was the case with the buffers. However, it increases with C_{in} for the third generation. This means that, unlike the buffers, for small values of C_{in}, the speedup factor might increase as C_{in} increases or fan_{in} decreases, specially for scaled-down technologies, where the BJT's gain degradation starts at a larger value of V_{BE}.

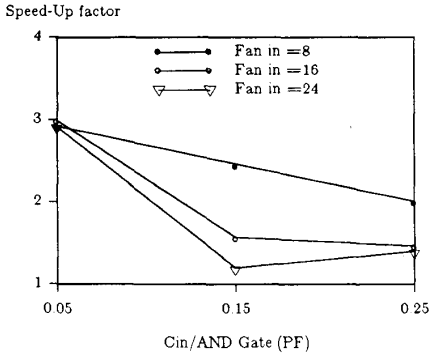


Fig. 9. The speedup factor of the FSBiCMOS complex logic gate over the CMOS counterpart versus the input capacitance of an AND gate for different values of fan-in and $C_L = 0.5$ pF.

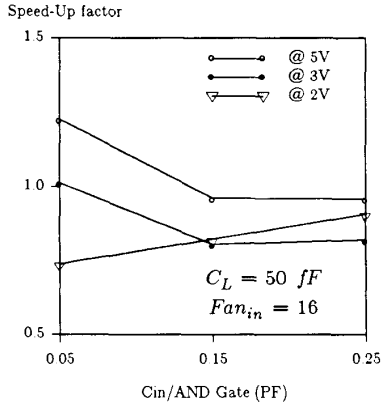


Fig. 10. The speedup factor of the FSBiCMOS complex logic gate versus the input capacitance of an AND gate for the three technology generations. $C_L = 50$ fF and $fan_{in} = 16$.

The above observations suggest that the BiCMOS implementations of complex gates tend to be faster than CMOS at low C_{in} and fan_{in}, and the speedup factor increases with C_L. However, CMOS becomes better with scaling. A similar concept to the speedup contours developed for the buffers in Section II could be applied to complex logic gates by obtaining an equivalent circuit of the complex logic. This is done by substituting the series MOS devices by a single device as in [9] and relating its saturation current to an effective value of C_{in}. By applying the analytical expressions obtained in the previous section the speedup contours could be obtained.

B. Multilevel Logic Gates

The options available in a BiCMOS technology to implement multilevel logic (MLL) are considered next. These options are: CMOS, CMOS with a FSBiCMOS driver, and CML with front/end conversion such that the inputs and outputs are CMOS levels. The input capacitance for all implementations is the same. The fan_{out} for each level of logic is one except for the last level where the fan_{out} and the wire capacitance are represented by a load capacitance C_L.

TABLE IV
THE AVERAGE STATIC POWER OF THE INTERNAL MCSL GATES AND THE END CML/CMOS CONVERSION STAGE FOR THE THREE GENERATIONS OF BiCMOS TECHNOLOGY

BiCMOS Tech.	Int. Gate	Conv. Stage
1.00 μm, 5 V	4.15 mW	14.20 mW
0.56 μm, 3 V	2.40 mW	7.15 mW
0.20 μm, 2 V	1.96 mW	4.74 mW

In the second option, only the last stage of the MLL was implemented in BiCMOS since all other levels have very small output capacitance. In the CML implementation, the basic cell used is the MCSL [11] where the conversion from CMOS to CML and the logic operation are done at the same time, with mixed CMOS and CML levels inputs. In [11] a 16-b adder implemented in MCSL was reported to achieve a speedup factor of about 5 over a similar adder implemented in CMOS. However, there was no mention of the input capacitance of the two implementations and the CML/CMOS conversion stage delay was not included in the total delay. In this work, it will be shown that, for the same input capacitance and including the CML/CMOS conversion stage delay, it is possible to achieve a speedup factor of 4 for six levels of logic at a total power of less than 35 mW. Table IV contains the average static power dissipation per an internal gate and the last conversion stage for the three BiCMOS technologies. The first gate in the MLL implementation using MCSL is shown in Fig. 11(a). An internal MCSL gate is shown in Fig. 11(b). The CML/CMOS conversion circuit at the end of the MLL, Fig. 11(c), is similar to the one in [12] except that a MCSL is used instead of the CML part.

The speedup factors of the CML and the CMOS + FSBiCMOS implementations over the CMOS for different load capacitances, as functions of the number of logic levels, are shown in Fig. 12. Fig. 13 shows the effect of scaling on the speedup factors for low C_L. There are two sets of results for the CML implementation: one where the end CML/CMOS conversion circuit had a fixed size FSBiCMOS buffer as a driver, and the other where the output driver was optimized as in Section II to show the effect of optimizing the output buffer. Also the areas of CMOS + BiCMOS and CML implementations (with optimized output), normalized to that of the CMOS implementation, are shown in Fig. 14.

From the above results, the following can be concluded.

1) For high C_L, the CMOS + BiCMOS implementation offers a good speed-up factor, especially for a low number of logic levels (Fig. 12), at a power level almost equal to that of pure CMOS. CML will offer a larger speedup, which increases as the number of logic levels increases (Fig. 12), at the expense of having a relatively large power dissipation.

2) For low C_L, CML still offers a good speedup factor, which again increases as the number of logic levels increases (Fig. 12).

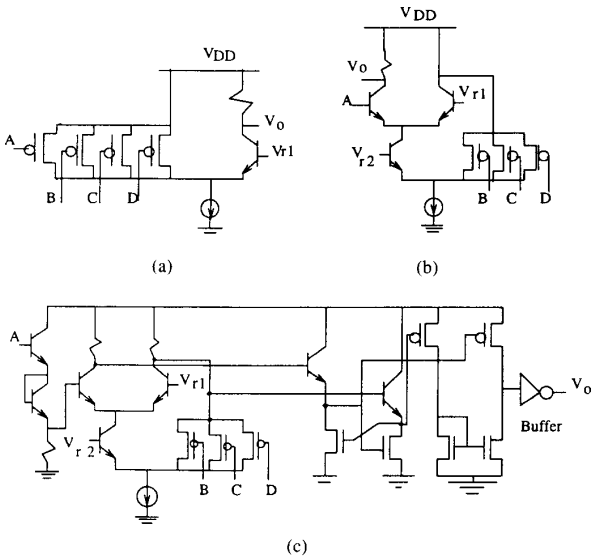


Fig. 11. The different MCSL building blocks used in the CML implementation of the MLL: (a) the first gate in the MLL, (b) an internal gate, and (c) the last gate in the MLL logic with CML/CMOS conversion. The buffer is selected according to the procedure outlined in Section II to give optimum performance.

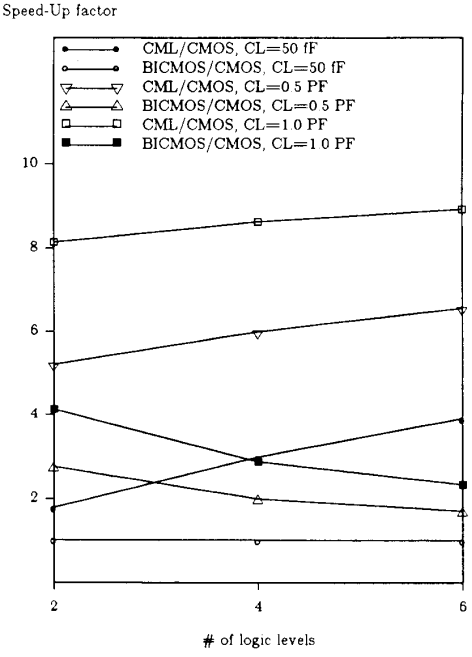


Fig. 12. The speedup factor of the CML and CMOS + BiCMOS implementations of the MLL over the CMOS counterpart versus the number of logic levels and for different C_L, at 5 V.

3) The optimization of the output buffer in the CML/CMOS conversion circuit increases the speedup significantly, especially for the scaled BiCMOS technologies (Fig. 13).

4) The CMOS + BiCMOS implementation does not offer any advantages as the technology scales (Fig. 13).

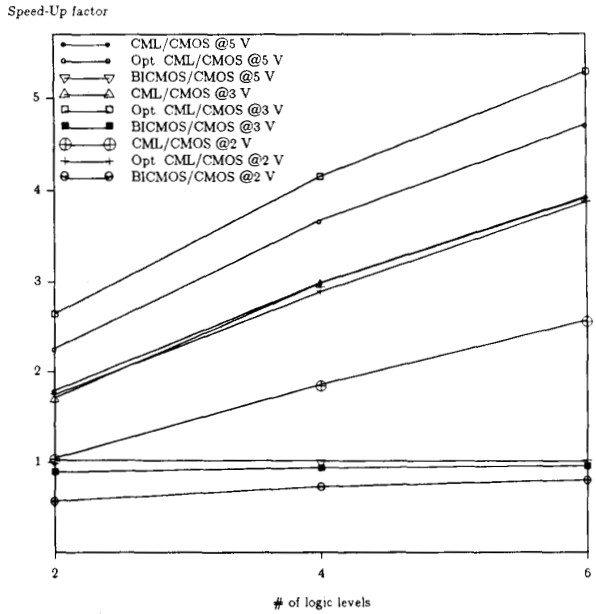


Fig. 13. The speedup factor of the CML and CMOS + BiCMOS implementations of the MLL versus the number of logic levels for the three technologies and at C_L = 50 fF.

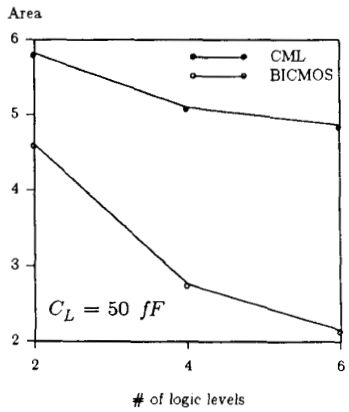


Fig. 14. The areas of the CML and CMOS + BiCMOS implementations of the MLL normalized to that of the CMOS implementation versus the number of logic levels and for C_L = 50 fF.

The CML, however, can offer the same speedup with much less power dissipation (Fig. 13 and Table IV).

5) The normalized areas of the CML and CMOS + BiCMOS implementations decrease with increased number of logic levels, and for low values of C_L the CML area is larger than that of the CMOS + BiCMOS (Fig. 14). The area ratio between CML and pure CMOS cannot decrease below a certain limit, and that is the ratio from a CML stage to a CMOS stage. However, this ratio decreases as C_{in} increases since the BJT area does not increase as C_{in} increases.

IV. Conclusion

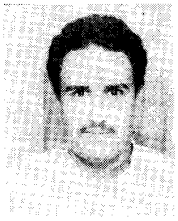
In this work, several buffering circuits available in a BiCMOS technology have been compared and speedup contours were obtained. These contours show that BiCMOS is faster than CMOS at small values of input capacitance, and the speedup factor increases with the load capacitance until a certain limit is reached after which it starts to decrease. However, at such a point, the area of the CMOS buffer chain is much larger than that of the BiCMOS buffer. General and easy-to-automate design guidelines for buffer chain design in a BiCMOS environment were obtained. The possibility of multistage mixed CMOS/BiCMOS buffers was investigated. It was found that a two-stage BiCMOS buffer can achieve a lower delay than all other buffers available in a BiCMOS technology. The CMOS–BiCMOS buffer chain could achieve a speed closer to that of the BiCMOS–BiCMOS chain at a much smaller area. It was shown that BiCMOS complex logic gates can achieve significant speedup over CMOS for low input capacitances, low fan_{in}, and high fan_{out}. BiCMOS complex gates are more difficult to optimize due to associated limitations on device sizes, which become more severe as the gate complexity increases. Finally, the different options for implementing multilevel logic gates in a BiCMOS technology were compared. It was found that replacing the last stage of a CMOS MLL with a BiCMOS gate offers a good speedup factor for high C_L with no increase in power. However, for scaled-down BiCMOS technologies this is not the case, and to achieve any significant speedup, CML has to be used.

Acknowledgment

The authors thank Dr. S. Rofail for his valuable comments.

References

[1] M. I. Elmasry, A. Bellaouar, and S. H. Embabi, *BiCMOS IC Design*. Norwell, MA: Kluwer Academic, 1991.
[2] P. Raje *et al.*, "A new BiCMOS/CMOS gate comparison/design methodology and supply voltage scaling model," in *IEDM Tech. Dig.*, 1989, pp. 433–436.
[3] *HSPICE User's Manual*, Meta-Software, Inc., Campbell, CA, 1990.
[4] A. Bellaouar, S. Embabi, and M. I. Elmasry, "Scaling of BiCMOS digital circuit structures," in *IEDM Tech. Dig.*, 1989, pp. 437–440.
[5] A. Bellaouar, S. Embabi, and M. I. Elmasry, "Scaling of digital BiCMOS circuits," *IEEE J. Solid-State Circuits*, vol. 25, pp. 932–941, 1990.
[6] N. Hedenstierna and K. Jeppson, "CMOS circuit speed and buffer optimization," *IEEE Trans. Computer-Aided Design*, vol. CAD-6, pp. 270–281, 1987.
[7] G. Rosseel and R Dutton, "Delay analysis for BiCMOS drivers," in *BCTM Tech. Dig.*, 1988, pp. 220–222.
[8] S. H. K. Embabi, A. Bellaouar, and M. I. Elmasry, "Analysis and optimization of BiCMOS digital circuit structures," Univ. Waterloo, Waterloo, Ont., Canada, Int. Rep., 1990.
[9] T. Sakurai and A. Newton, "Delay analysis of series-connected MOSFET circuits," *IEEE J. Solid-State Circuits*, vol. 26, pp. 122–131, 1991.
[10] P. Raje *et al.*, "BiCMOS gate performance optimization using a unified delay model," in *Symp. VLSI Technology Tech. Dig.*, 1990, pp. 91–92.
[11] W. Heimsch *et al.*, "Merged CMOS/bipolar current switch logic (MCSL)," *IEEE J. Solid-State Circuits*, vol. 24, pp. 1307–1311, 1989.
[12] T. S. Yang *et al.*, "A 4-ns 4K × 1-bit two-port BiCMOS SRAM," *IEEE J. Solid-State Circuits*, vol. 23, pp. 1030–1040, 1988.
[13] S. H. K. Embabi, A. Bellaouar, and M. I. Elmasry, "Analysis and optimization of BiCMOS digital circuit structures," *IEEE J. Solid-State Circuits*, vol. 26, pp. 676–679, 1991.



Muhammad S. Elrabaa (S'90) was born in Khartoume, Sudan, on August 6, 1968. He received the B.Sc. degree from Kuwait University, Kuwait, and the M.A.Sc. degree from the University of Waterloo, Waterloo, Ont., Canada, all in electrical engineering, in 1989 and 1991, respectively. His research interests include ECL circuits and digital BiCMOS circuits. He is currently studying at the University of Waterloo toward the Ph.D. degree.

Mohamed I. Elmasry (S'69–M'73–SM'79–F'88) was born in Cairo, Egypt, on December 24, 1943. He received the B.Sc. degree from Cairo University, Cairo, Egypt, and the M.A.Sc. and Ph.D. degrees from the University of Ottawa, Ottawa, Ont., Canada, all in electrical engineering, in 1965, 1970, and 1974, respectively.

He has worked in the area of digital integrated circuits and system design for the last 25 years. He worked for Cairo University from 1965 to 1968 and for Bell-Northern Research (BNR), Ottawa, Canada, from 1972 to 1974. He has been with the Department of Electrical and Computer Engineering, University of Waterloo, Waterloo, Ont., Canada, since 1974, where he is a Professor and founding Director of the VLSI Research Group. He has a cross appointment with the Department of Computer Science where he is a Professor. He has held the NSERC/BNR Research Chair in VLSI design at the same University since 1986. He has served as a consultant to research laboratories in Canada and the United States, including AT&T Bell Labs, GE, CDC, Ford Microelectronics, Linear Technology, Xerox, and BNR, in the area of LSI/VLSI digital circuit/subsystem design. During sabbatical leaves from Waterloo he was at the Micro Components Organization, Burroughs Corporation (Unisys), San Diego, CA, Kuwait University, Kuwait, and Swiss Federal Institute of Technology, Lausanne, Switzerland. He has authored and co-authored over 150 papers on integrated circuit design and design automation. He has several patents to his credit. He is the editor of the IEEE Press books *Digital MOS Integrated Circuits* (1981), *Digital VLSI Systems* (1985), and *Digital MOS Integrated Circuits II* (1991). He is also author of the book *Digital Bipolar Integrated Circuits* (Wiley, 1983) and co-author of the book *Digital BiCMOS Integrated Circuits* (Kluwer, 1991).

Dr. Elmasry has served in many professional organizations in different positions including the Chairmanship of the Technical Advisory Committee of the Canadian Microelectronics Corporation. He is a founding member of the Canadian VLSI Conference, the International Conference on Microelectronics, and the founding President of Pico Electronics Inc. He is a member of the Association of Professional Engineers of Ontario and is a Fellow of the IEEE for his contributions to "digital integrated circuits."