VLSI DESIGN
1999, Vol. 9, No. 2, pp. 203–211
Reprints available directly from the publisher
Photocopying permitted by license only

Influence of BJT Transit Frequency Limit Relation to MOSFET Parameters on the Switching Speed of BiCMOS Digital Circuits

A. SRIVASTAVA*

Department of Electrical and Computer Engineering, Louisiana State University, Baton Rouge, LA 70803-5901

(Received 25 March 1997; In final form 4 December 1997)

The use is made of the BJT transit frequency limit (f_{TL}) dependence on the MOSFET parameters (L, V_{th}) to design BiCMOS digital circuits. The f_{TL} relation is used in conjunction with the established BiCMOS gate delay models. It is shown that the minimum delay BiCMOS circuits driving the large capacitive load, can be designed at the transit frequency limit with the reduced BJT AREA factor. The time delay calculations are presented for a typical BiCMOS circuit and comparison is made with the results simulated using SPICE.

Keywords: BiCMOS, digital circuit design, VLSI design, gate delays, analysis, simulation

INTRODUCTION

BiCMOS is the technology combining the low power of CMOS with the high-speed and drive capability of bipolar for realizing high performance digital circuits [1–3]. Over the years BiCMOS integrated circuits have been well characterized for high-speed digital logic applications [4–16]. The circuit delay time model of the BiCMOS has been extensively studied and as a performance measure of the BiCMOS technology [12]. Many closed-form analytical expressions for the gate delay have been derived through the physical and electrical modeling and are available in the literature [5, 9, 10–13].

Greeneich and McLaughlin [5] have obtained closed-form analytical expressions for the gate delay and have shown its dependence on device and circuit parameters. The analysis does not include high-level injection effects in the bipolar transistor model. For the devices used, high-level injection effects are not a major consideration in transient response studies. Rosseel and Dutton [6] have studied the influence of device parameters on

* Tel.: (225) 388-5622; Fax: (225) 388-5200; e-mail: ashok@ee.lsu.edu

the switching speed of a BiCMOS buffer circuit. The analysis examines the high-level injection effects through the parameters I_{kF}, the upper knee current in bipolar transistor, and shows to strongly influence the performance. The analysis also examines the influence of different emitter sizes on the delay time and it is shown that for a given area, the delay is minimum for one optimal size ratio for the MOS and bipolar transistors. Rofail and Elmasry [11] have derived analytical and numerical BiCMOS gate delay models to characterize BiCMOS structures using long and short channel devices. Their numerical model include high-current effects and other second order effects. Raje *et al.* [12] have provided a piecewise delay expression of the BiCMOS gate delay model which takes into account high-current effects in bipolar transistor, short-channel effects in MOSFET and parasitic capacitances at the base and output. Fang *et al.* [10] have proposed a BiCMOS overall delay-optimization scheme. Under this scheme, the delay is minimized when the maximum collector current is equal to the onset current of high current effects of bipolar transistors. This indicates that the bipolar transistor in BiCMOS circuits operates at a collector current density below the high-current region for speed-optimized BiCMOS circuits. It is also shown that BiCMOS circuits can keep the speed advantage over CMOS circuits down to submicron dimensions under constant load conditions.

A comparison of these papers shows that different analytical solutions have been obtained through the use of different assumptions. None of these work seems to include and provide a proper correlation between the MOS and bipolar transistors dimensions in gate delay modeling and delay minimization in BiCMOS circuits. Rothermel and Hosticka [7] have shown that there is a transit frequency below which the bipolar transistor does not contribute to the speed improvement, and hence the digital BiCMOS design techniques do not offer any speed advantage over CMOS. In this work, we have further explored the usefulness of the relationship between the transit frequency limit of the bipolar transistor and MOSFET parameters on BiCMOS gate delay models.

THEORY

In the following, we will describe the relation between BJT transit frequency limit and MOSFET parameters, and BiCMOS gate delay models.

BJT Transit Frequency Relation to MOSFET Parameters

In order to develop a realistic BiCMOS speed improvement model, consider a simple BiCMOS buffer circuit which includes an internal logic as shown in Figure 1. The buffer circuit in the Figure 1 drives a capacitive load, C_L. Figure 1 provides the necessary parasitics which are to be included in developing a relation between the bipolar and MOS transistors parameters. Figure 2 shows the equivalent circuit for the buffer part of the Figure 1 where the transistor acts as a digital switching device. In Figure 2, the MOS transistor provides the input current source and bipolar transistor simulates the output stage of the buffer. We now assume that bipolar or MOS transistor charge and discharge the load capacitance, C_L, by the same output current (I_{OUT}). The input current is also assumed to be constant during switching. Thus, the time needed to build an active charge in base or channel region of bipolar and MOS transistors can be calculated from the input charge. Now for the same charging and discharging current, both transistors will have the same speed under identical input conditions, and BiCMOS drivers constructed with them will have the same delay times.

In BiCMOS circuits, if the bipolar transistor has to switch faster than the MOS transistor, charge build-up in the base of bipolar transistor should exceed the charge build-up in the channel region of the MOS transistor. This sets a lower limit on the bipolar transistor to perform better than the MOS transistor. The limit is described by the transit frequency limit, f_{TL} relation and is of the following

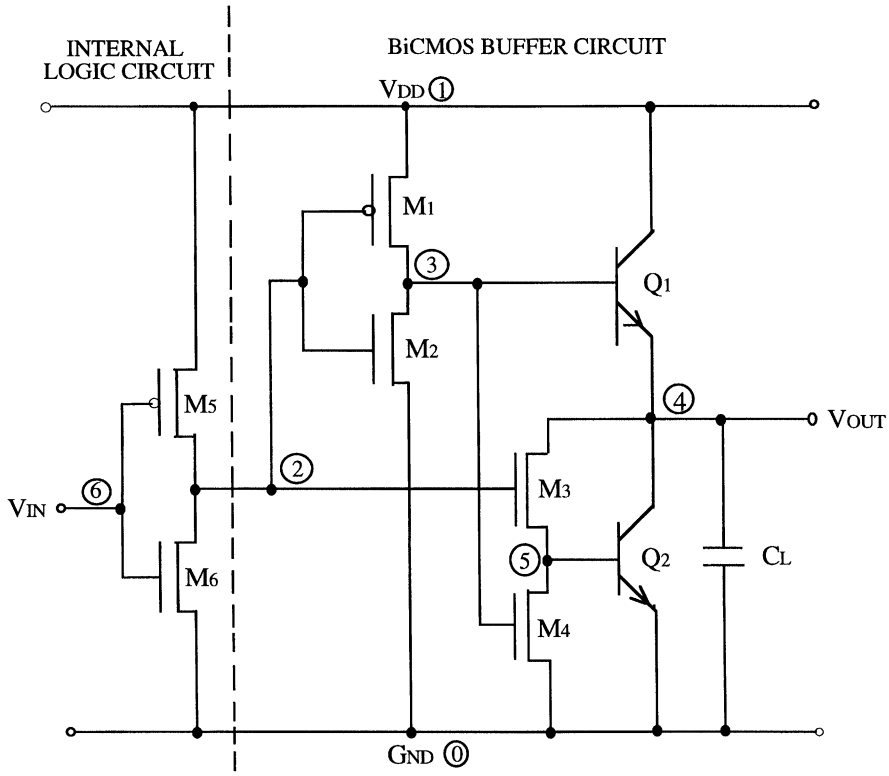

FIGURE 1 BiCMOS buffer with internal circuit.

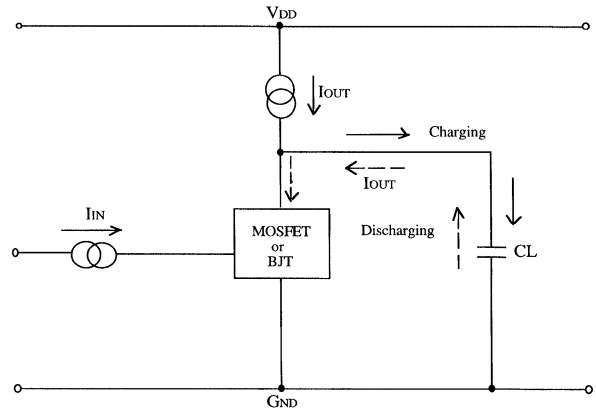

FIGURE 2 Transistor as a digital switch.

form [7]

$$f_{TL} = \frac{\mu_{ON}(V_{DD} - V_{th})}{\frac{32}{3}\pi L_{EFF}(L_{EFF} + 6L_{OV}) \cdot \left(1 + \frac{V_{DD}}{L_{EFF}E_{cm}}\right)} \quad (1)$$

where μ_{ON} is the mobility in low electrical field, L_{OV} the overlap length, V_{th} the threshold voltage, E_{cm} the critical electric field, and V_{DD} the power supply voltage. The transit frequency limit, f_{TL} is related to MOS transistor parameters (L_{EFF}, L_{OV} and V_{th}) and supply voltage V_{DD}. For a BiCMOS

circuit to perform better over the CMOS circuit, the transit frequency f_T of the BJT should be equal to or greater than the transit frequency limit, f_{TL}. Equation (1) for f_{TL} can be used to design BJT and MOSFET more realistically in a BiCMOS circuit for the optimum delay.

BiCMOS Delay Models

In Figure 1, transistors M_1 and M_3 provide base current to transistors Q_1 and Q_2, respectively. Transistors M_2 and M_4 remove the base charge from the bipolar transistor during discharge process in switching transient. The buffer design provides identical pull-up and pull-down response. The delay time model can be developed by applying a step voltage at the input of the buffer circuit. It can be followed from Figure 1 that larger the transistor M_1, the higher is the base and emitter currents of transistor Q_1 which results in shorter gate delay. This condition may lead to BJT entering into the high-current region where Kirk effect [17, 18] takes place. In practical situation, the buffer circuit is driven by an internal CMOS logic circuit as shown in the Figure 1. Now, if the transistor M_1 is made larger in size by increasing its width, the gate capacitance of M_1 will increase. The gate capacitance of M_1 acts as a load to the internal logic circuit (inverter in the present design) and will reduce its speed. Thus, in order to achieve the shortest delay of the BiGMOS circuit of Figure 1, it may not be necessary to drive BJT(Q_1) in the high-current region during the pull-up transient.

BiCMOS gate delay is composed of pull-up delay due to charging of load capacitance C_L through the combination of M_1-Q_1 transistors, and pull-down delay due to discharging of C_L through the combination of M_3-Q_2 transistors. It has been shown in Ref. [10] that pull-up and pull-down transient response of the BiCMOS buffer circuit can be made equal by making BJTs Q_1 and Q_2 identical and keeping widths of M_1 and M_3 same. Thus, one can describe the switching behavior of a BiCMOS circuit either through pull-up or pull-down delay models. In the present work, we will study the switching response through the pull-up delay models.

It was stated earlier that BiCMOS buffer circuit operates below the current level (maximum output current) at which high-current effects set-in when integrated with the internal CMOS logic. The maximum output current can be approximated to be equal to upper knee current I_{kF} of the BJT as done by earlier workers [10] for an optimum delay calculation. It was also stated in this work that the BiCMOS circuit can be made to perform better over the CMOS circuit if transistors are designed to operate at the transit frequency limit f_{TL} described by the Eq. (1). The transit frequency limit f_{TL} can also be assumed to be the frequency at which high-current effects set-in. Thus, optimum gate delay for the BiCMOS circuit can be obtained at the f_{TL} at which it can be approximated that the maximum output current is equal to the knee current I_{kF}. Earlier delay models [5, 10] can be suitably modified to include the transit frequency relation of Eq. (1) and its dependence on MOSFET parameters. The operation of the BiCMOS circuit can be divided into three time intervals as shown in the Figure 3.

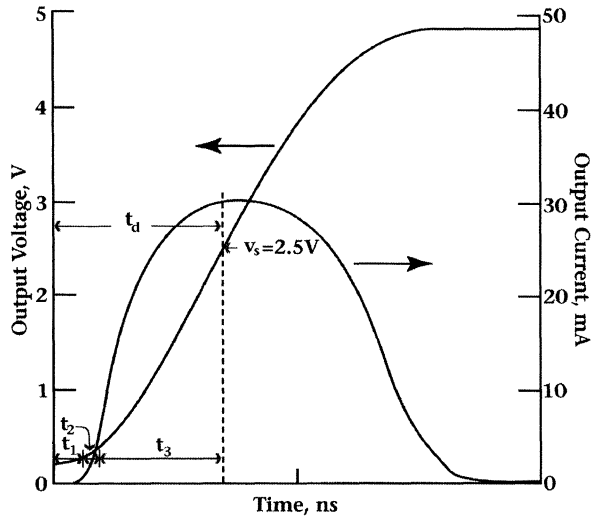

FIGURE 3 Three time delay intervals in a BiCMOS circuit shown on output voltage waveforms. Note: Output current waveform is also included to show the maximum collector current.

(i) The time interval t_1 is defined as the initial delay time. The MOSFET (M_1) turns-on and BJT (Q_1) begins to conduct until its base-emitter voltage V_{BE} reaches about $V_{BE}(on) \simeq 0.7\,V$. At the end of the time interval t_1, the MOSFET (M_1) still remains in saturation and the BJT (Q_1) turns-on. The initial time delay can be obtained from an analysis of the equivalent circuit during the time interval t_1 and is given by [5]

$$t_1 = R_{ch}(C_E + C_C)\frac{V_{BE}(on)}{V_{DD} - |V_{th,P}|} - \frac{R_C C_C^2}{C_E + C_C} \tag{2}$$

where

$$R_{ch} = \frac{2L}{W_1 \mu_p C_{ox}(V_{DD} - |V_{th,P}|)} \tag{3}$$

R_{ch} is the equivalent dc channel resistance of the MOSFET (M_1). C_E, C_C, R_B and R_C are the emitter and collector capacitances, R_B and R_C are the base and collector resistances, respectively of BJT (Q_1). C_E, C_C, R_B and R_C are assumed to be constant. C_{CS} is the collector-substrate capacitance of Q_2 and is assumed to be constant. W_1 and L are the channel width and length of the transistor M_1. μ_P and $|V_{th,P}|$ are the hole carrier mobility and threshold voltage. In a typical BiCMOS circuit, $R_C < R_{ch}$ which makes the second term in Eq. (2) negligible.

(ii) The delay time t_2 is the time interval during which BJT (Q_1) is on and the MOSFET (M_1) is still in saturation. It is to be mentioned that M_1 remains in saturation provided its threshold voltage $|V_{th,P}| > V_{BE}(on) \cong 0.7\,V$ of Q_1. The time delay can be obtained from an analysis of the equivalent circuit during the time interval t_2 and is given by [5]

$$t_2 = \left[2R_{ch}C_L^*\tau_T^*\left\{\frac{|V_{th,P}| - V_{BE}(on)}{V_{DD} - |V_{th,P}|}\right\}\right]^{1/2} \tag{4}$$

where C_L^* is the equivalent load capacitance, τ_T^* is the equivalent forward transit time of the BiCMOS circuit. C_L^* and τ_T^* are given by

$$C_L^* = C_L + C_{CS} \tag{5}$$

$$\tau_T^* = \tau_T + R_C C_C \tag{6}$$

where $\tau_T = (1/(2\pi f_T))$. f_T is the cut-off frequency or transit frequency of the transistor. C_E is neglected since $V_{BE}(on)$ nearly remains constant and equals to 0.7 V, ($V_{BE} \cong 0.7\,V$).

(iii) The delay time t_3 begins when $V_{OUT} = |V_{th,P}| - V_{BE}(on)$ and M_1 enters the linear region, and ends at the point where the output reaches the switching threshold $V_s = 2.5\,V$. M_1 can be substituted by an equivalent channel resistance R_{ch} in the equivalent circuit during the time interval t_3 which is of the form [5]

$$t_3 \cong T\cos^{-1}\left\{\frac{V_{DD} - V_{BE}(on)}{2(V_{DD} - |V_{th,P}|)}\right\} \tag{7}$$

where

$$T = \frac{T_O}{\sqrt{1 - \left(\frac{T_O}{2\beta_F^*\tau_T^*}\right)^2}} \tag{8}$$

and

$$T_\delta = \sqrt{(R_{ch} + R_B)C_L^*\tau_T^*} \tag{9}$$

$$\frac{1}{\beta_F^*} = \frac{1}{\beta_F} + \frac{C_C}{C_L^*} \tag{10}$$

It should be noted that for a typical device and circuit parameters $T_o \ll 2\beta_F^*\tau_T^*$. For $|V_{th,P}| = V_{BE}(on)$, Eq. 7 reduces to the form

$$t_3 \cong T\cos^{-1}(0.5) = T \tag{11}$$

The total delay time t_d is the sum of time delay intervals t_1, t_2 and t_3, described by Eqs. (2), (4), (7), or (11), respectively.

It should be mentioned that the delay time interval t_3 will create a discontinuity in the

collector current at the boundary between t_2 and t_3. This is due to a constant term (C) which is set to 0 for mathematical simplicity in the derivation of Eq. (7) [5]. However, it was suggested in Ref. [10] that for high-speed BiCMOS circuits, $|V_{th,P}|$ can be approximated to be equal to $V_{BE}(on) = 0.7\,V$. Thus, it can be assumed that the MOSFET (M_1) operates in the linear region during the time interval t_2 without causing a significant error. It implies that $t_2 = 0$ and the time interval t_3 can then be described in the following form [10]

$$t_3 = T\tan^{-1}\left(\frac{2\beta_F^*\tau_T^*}{T}\right) \tag{12}$$

Equation (12) can be further simplified to the form

$$t_3 \cong \frac{\pi}{2}\cdot T = \frac{\pi}{2}\frac{T_O}{\sqrt{1-\left(\frac{T_O}{2\beta_F^*\tau_T^*}\right)^2}} \tag{13}$$

for $2\beta_F^*\tau_T^* \gg T$ for typical devices.

The delay time t_d is now the sum of time intervals t_1 and t_3 since $t_2 = 0$, described by Eqs. (1) and (13), respectively. The maximum output current $I_C(max)$ of Q_1 at the end of the time interval t_3 can then be described by the equation of the form [10]

$$I_C(max) = C_L^*\cdot\left\{\frac{V_{DD}-V_{BE}(on)}{T}\right\}\sin\left(\frac{\pi}{2}\right)\cdot\exp\left(-\frac{\pi T}{4\beta_F^*\tau_T^*}\right), \quad \text{for } 2\beta_F^*\tau_T^* \gg T \tag{14}$$

The maximum output current described by the Eq. (14) can be assumed to be approximately equal to the knee current I_{kF} at which high-current effects start. Equation (14) for $I_C(max)$ in combination with the Eq. (1) for f_{TL} can be used to design optimum delay BiCMOS circuits. It should be noted that f_{TL} influences time delays through the time delay interval equations for t_2 and t_3.

RESULTS AND DISCUSSION

Figure 4(a) shows the dependence of BJT transit time on the MOSFET channel length, obtained from Eq. (1) for a BiCMOS circuit shown in Figure 1. The typical MOSFET parameters used in the calculation are $\mu_{ON} = 5\times10^2/\,V-sec$, $V_{th,N} = 0.8\,V$, $L_{OV}\cong 0.13L_{EFF}$, $E_{cm} = 7\times10^4 V/cm$ and $V_{DD} = 5\,V$. The corresponding transit frequency limit dependence on the channel length is shown in the Figure 4(b). Figure 4 can be used to obtain dimensions of BJTs and MOSFETs in a BiCMOS circuit designed to operate at the minimum delay. Figure 5 shows the simulated output voltage

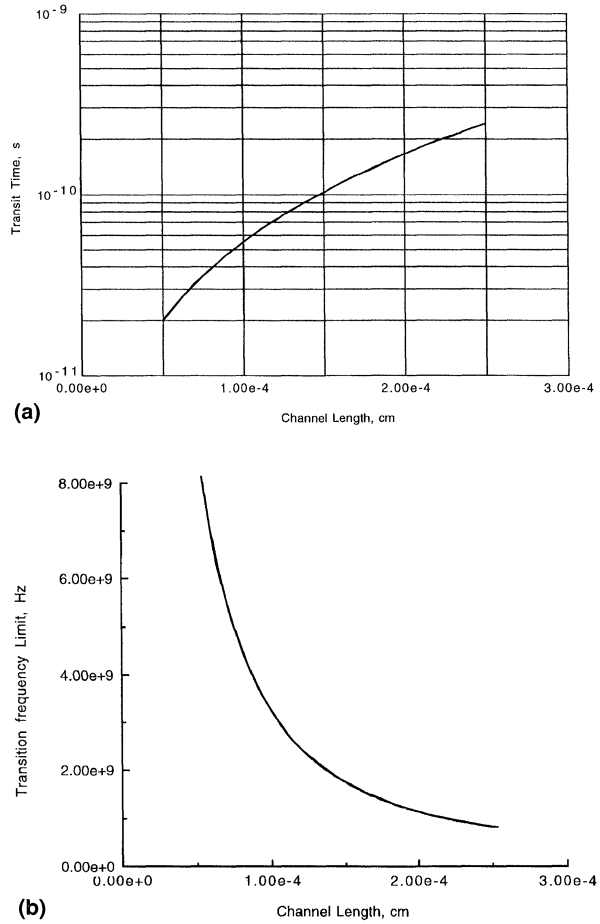

FIGURE 4 (a): BJT transit time dependence on MOSFET channel length; (b): BJT transit frequency limit dependence on channel length.

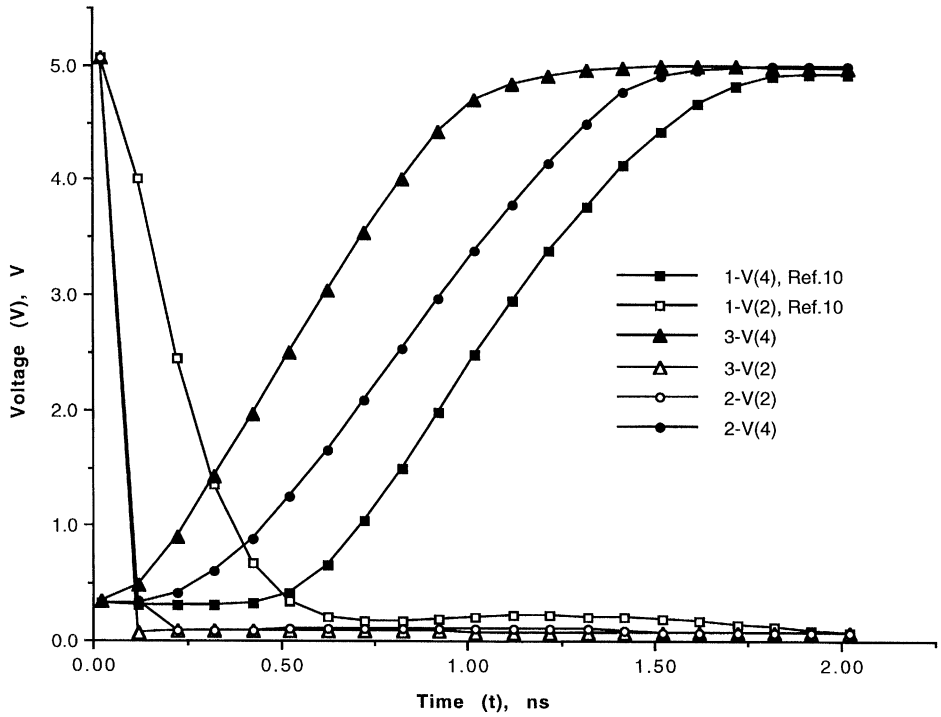

FIGURE 5 Voltage response at nodes 4 and 2, respectively obtained from SPICE3.

response of the BiCMOS buffer with the internal circuit. The step function input voltage which switches from 0 to 5 V is applied at a node 6. The circuit is simulated using SPICE3 (LEVEL 2) MOSFET model parameters. BJT and MOSFET model parameters are taken from the Ref. [10]. The load capacitance, C_L is 5 pf. In Figure 5, $V(4)$ and $V(2)$ are the voltage responses at nodes (4) and (2), respectively. In a group of $V(4)$ plots 1-$V(4)$, Ref. [10] plot shows the output voltage response corresponding to device dimensions of Ref. [10] for the gate delay. The corresponding input voltage response at a node 2 is shown by 1-$V(2)$, Ref. [10] plot. The MOSFET channel length is 2 μm and BJT transit time is 20 ps for this set of curves. The plots 2-$V(4)$ and 3-$V(4)$ show the output voltage response for $L = 1.0$ and 0.5 μm, respectively. Transit time (τ_T) parameters are obtained from the Figure 4(a). The corresponding input voltage response at a node 2 is shown by 2-$V(2)$ and 3-$V(2)$ curves. The plots 1-$V(4)$, Ref. [10], 2-$V(4)$ and 3-$V(4)$ are obtained using the AREA factor – 14, 14 and 8 and τ_T = 20, 55 and 20 ps, respectively for the BJT design. Figure 5 shows that the BiCMOS gate delay is further reduced from the value that is obtained for the optimized design described in Ref. [10].

Figure 6 compares the output current response corresponding to $L = 2$ and 0.5 μm, respectively. The current response corresponding to $L = 1$ μm is not included in this figure. The plot 1-I_c, Ref. [10] shows the output (collector) current variation for the BiCMOS circuit design parameters of the Ref. [10]. The output current peaks at approximately 30 mA and the AREA factor is 14. This plot is compared with the plot for 2-I_c for $L = 0.5$ μm, and τ_T = 20 ps obtained from the Figure 4(a). The AREA factor is 8 for this curve which provides nearly the same peak current as in the 1-I_c, Ref. [10] curve. Figure 6 also shows that the output

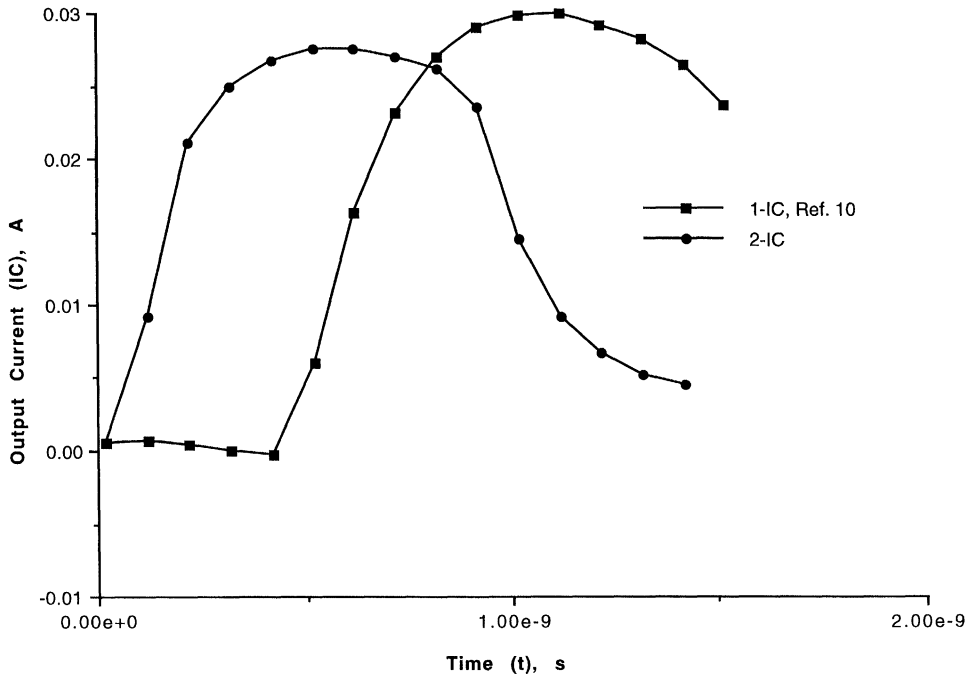

FIGURE 6 Output current (collector current) response.

TABLE I BiCMOS delay time

L μm	I_C(max) (mA)	t_1 (ns)	t_2 (ns)	t_3 (ns)	t_d (ns)	$t_d(t_{2=0})$ (ns)	$t_{d(SPICE3)}$ (ns)	BJT AREA factor	C_L (pF)
0.5	28	0.052	0.085	0.519	0.656	0.674	0.60	8	5
2.0	30	0.092	0.090	0.504	0.686	0.742	0.70	14	5

current peaks approximately at a half the time needed to reach the switching threshold $V_S = 2.5\,V$ in plot 1-I_C Ref. [10]. Thus, we can conclude from Figures 5 and 6 that the Eq. (1) gives a lower gate delay with the reduced BJT AREA factor under the same load condition. This is further shown in the Table I where the time intervals t_1, t_2, t_3 and the total delay time t_d for the BiCMOS buffer stage are summarized. In Table I, delay time results are included for $L = 0.5\,\mu m$ without using the Eq. (1) [10]. SPICE3 (LEVEL 2) simulation results are also shown for comparison. In the Table I, results obtained from the use of the transit frequency relation (Eq. 1) for $L = 0.5\,\mu m$ suggest that a minimum delay BiCMOS circuit can be designed even with the reduced BJT AREA factor. The presented design approach provides nearly the same maximum output current (Fig. 6, plot 2-I_c) to drive the same load ($C_L = 5\,pF$) as in $L = 2\,\mu m$ based design.

CONCLUSION

The BJT transit frequency limit dependence on MOSFET parameters and its direct relationship to BiCMOS delay time intervals give a minimum delay. The minimum delay is obtained with reduced BJT AREA factor and large maximum output current which is approximately to be the

upper knee current. The high-current effects which start at the knee current, occur at the end of time interval t_3 or the delay time t_d. Since the BiCMOS circuit operates below the knee current, high current effects such as degradation in current gain and base transit time are not included in the time delay analysis. This approach holds good for output voltage to rise to its switching threshold which is normally the case. The results and analysis of the present work suggest that the minimum delay BiCMOS digital circuits could be designed using BJT transit frequency limit criteria in relation to MOSFET parameters.

Acknowledgement

Author is grateful to the reviewers for their valuable comments and suggestions.

References

[1] Alvarez, A. R. (Ed.) (1989). *BiCMOS Technology and Applications*, Kluwer Academic Publications, Boston, MA.
[2] Buchanan, J. E. (1991). *BiCMOS/CMOS Systems Design*, McGraw-Hill, NY.
[3] Kang, S. M. and Leblebici, Y. (1996). *CMOS Digital Integrated Circuits: Analysis and Design*, McGraw-Hill, NY.
[4] De Los Santos, H. J. and Hoefflinger, B. (1986). "Optimization and scaling of CMOS-bipolar drivers for VLSI interconnects," *IEEE Trans. Electron Devices*, **ED-33**(11), 1722–1730.
[5] Greeneich, E. W. and McLaughlin, K. L. (1988). "Analysis and characterization of BiCMOS for high-speed digital logic," *IEEE J. Solid-State Circuits*, **23**(2), 558–565.
[6] Rosseel, G. P. and Dutton, R. W. (1989). "Influence of device parameters on the switching speed of BiCMOS buffers," *IEEE J. Solid-State Circuits*, **24**(1), 90–99.
[7] Rothermel, A. and Hosticka, B. (1989). "Speed comparison of digital BiCMOS and CMOS circuits," *IEE Proc.*, Pt. G, **136**(2), 49–56.
[8] Srivastava, A. (1991). "Switching properties of polysilicon emitter transistor operating in saturation," *IEE Proc.-G*, **138**(3), 358–362.
[9] Diaz, C. H. and Leblebici, Y. (1991). "An accurate analytical delay model for BiCMOS driver circuits," *IEEE Trans. Computer-Aided Design*, **10**, 577–588.
[10] Fang, W., Brunnschweiler, A. and Ashburn, P. (1992). "An accurate analytical BiCMOS delay expression and its application to optimizing high-speed BiCMOS circuits," *IEEE J. Solid-State Circuits*, **27**(2), 191–202.
[11] Rofail, S. S. and Elmasry, M. I. (1992). "Analytical and numerical analysis of the delay time of BiCMOS structures," *IEEE J. Solid-State Circuits*, **27**(5), 834–839.
[12] Raje, P. A., Saraswat, K. C. and Cham, K. M. (1992). "Accurate delay models for digital BiCMOS," *IEEE Trans. Electron Devices*, **39**(6), 1456–1464.
[13] Lai, P. T. and Cheng, Y. C. (1994). "A closed-form delay expression for digital BiCMOS circuits with high-injection effects," *IEEE J. Solid-State Circuits*, **29**(5), 640–643.
[14] Rafail, S. S. (1994). "Low-voltage, low-power BiCMOS digital circuits," *IEEE J. Solid-State Circuits*, **29**(5), 572–579.
[15] Kuroda, T. and Matsuo, K. (1994). "Analysis and optimization of BiCMOS gate circuits," *IEEE J. Solid-State Circuits*, **29**(5), 564–571.
[16] Chiang, C. H. and Kuo, J. B. (1995). "An analytical transient model for a 1.5 V BiCMOS dynamic logic circuit for low-voltage deep submicrometer BiCMOS VLSI," *IEEE Trans. Electron Devices*, **42**(3), 549–554.
[17] Kirk, C. T. (1962). "A theory of transistor cutoff frequency (f_T) falloff at high current densities," *IRE Trans. Electron Devices*, **ED-9**, 164–174.
[18] Feng, W. (1990). "Accurate analytical delay expressions for ECL and CML circuits and their applications to optimizing high-speed bipolar circuits," *IEEE J. Solid-State Circuits*, **25**(2), 572–583.

Authors' Biography

A. Srivastava is currently an Associate Professor of Electrical and Computer Engineering at the Louisiana State University in Baton Rouge. He was previously on the faculty in Department of Electrical and Computer Engineering of the State University of New York, New Paltz; North Carolina State University, Raleigh, University of Cincinnati; and Birla Institute of Technology and Science, Pilani (India). During 1979–80, he was also a UNESCO FELLOW at the University of Cincinnati and University of Arizona, Tucson. In the past, he has also served as a Scientist at the Central Electronics Engineering Research Institute, Pilani (India). His research interests include Digital and Analog/Mixed-Mode VLSI Design, Modeling; Smart Sensors and CMOS–MEMS Integration; Rad-Hard GaAs and CMOS ICs; and Cryogenics. He is a Senior Member of IEEE.