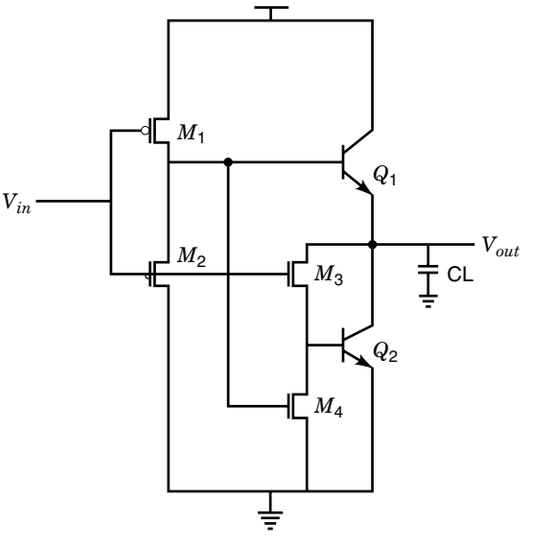


Figure 2. BiCMOS inverter gate.

BiCMOS LOGIC CIRCUITS

CMOS and bipolar techniques have their weak and strong points. CMOS offers an inverter with near-perfect characteristics such as high, symmetrical noise margins, high input and low output impedance, high packaging density, and low power dissipation. Speed is limited by the capacitive load and therefore is the only restricting factor, especially when capacitors must be driven. In contrast, bipolar digital circuits like ECL gates have a high current drive per unit area, high switching speed, and low I/O noise, but are power consuming. There is a performance gap between CMOS and ECL as shown in Fig. 1. The existence of this gap implies that neither CMOS nor ECL have the flexibility required to cover the full delay-power space. This can only be achieved by a technology such as BiCMOS which combines CMOS transistors and bipolar devices in a single process at a reasonable cost. The objective of the synergy of bipolar and CMOS is to exploit the advantages of both at the circuit and system levels.

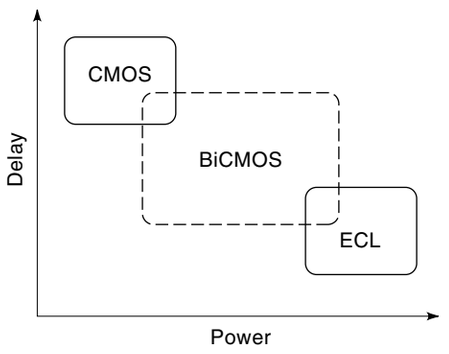


Figure 1. CMOS, BiCMOS, ECL delay and power.

The development of high-performance BiCMOS technology has received much attention in recent years. This technology opens a wealth of new opportunities because it is now possible to combine the high-density integration of CMOS logic with the current-driving capabilities of bipolar devices. A variety of digital BiCMOS circuit structures have been developed. An example of such circuits is a BiCMOS totem-pole gate which was originally developed by (1). This structure is currently the most commonly used BiCMOS gate (see Fig. 2). A BiCMOS inverter, which achieves just that mentioned above, is discussed in the following section. We first discuss the gate in general and then provide a more detailed discussion of the steady-state and transient characteristics.

BiCMOS INVERTER

Figure 2 shows the conventional BiCMOS gate. The primary motivation for BiCMOS is the possibility of using the best features of both MOSFETs and BJTs in a single integrated circuit. This has been done in a wide variety of applications such as memories, gate arrays, and processors. In the design of BiCMOS circuits, the MOSFETs are used to implement the logic function, and bipolar transistors are used to provide a fast, high-current output driver stage.

For example, to achieve high-speed BiCMOS adders, one requires a modest number of high-performance bipolar transistors for high-speed drivers on a critical path. As a result one can concentrate on bipolar devices optimized for high speed because a small number of these devices will not dominate the overall power consumption of the adder circuit. Furthermore, because the number of bipolar transistors used in the adder design is typically small, there is little need to make the packing density of the bipolar transistors compara-

ble to that of the CMOS part. This approach combines the area-efficient and low-power characteristics of CMOS layouts with the high-current drive capabilities of bipolar transistors. A BiCMOS inverter forms the basic circuit for the development of basic cells.

In order to achieve an engineering insight into the performance of BiCMOS cells, a transient analysis is performed on the driver cell shown in Fig. 2, using a step voltage excitation. The inverter gate is comprised of MOSFETs M_1 and M_2 which provide signals to drive the n–p–n bipolar transistors Q_1 and Q_2. NMOS transistors M_2 and M_4 provide a current discharge path for turning off the bipolar transistors. When the input falls, M_1 turns on and provides current to Q_1 which then charges up the load. When the input rises, M_2 turns on and provides current to Q_2 which then discharges the load capacitance. The BiCMOS gate has the features of

1. High input impedance provided by the MOSFET gate
2. Low output impedance provided by bipolar transistors
3. Superior driving capability of on- and off-chip loads
4. Lower delay sensitivity to loading
5. Transient drive with no dc power component

This gate represents a basic building block for digital logic circuits in BiCMOS. The inverter shown in Fig. 2 is easily extended to create multiple input basic tiling cells in a manner identical to pure CMOS basic cells.

Dc Characteristics

The output logic swing is constrained from going to supply voltage V_{dd} by the n–p–n base-emitter junction providing

$$V_{OH} = V_{dd} - V_{be1} \tag{1}$$

where V_{be1} is the base-emitter voltage of transistor Q_1. Similarly

$$V_{OL} = V_{be2} \tag{2}$$

is obtained when the lower n–p–n (Q_2) is biased on. The resulting logic swing of the BiCMOS gate is only a few tenths of a volt away from the supply voltages due to the low collector currents when the load capacitance is near full-charge or discharge.

Gate Transient Analysis

A basic understanding of the switching behavior of a digital BiCMOS gate is crucial to the circuit design. While accurate values for gate delay can be obtained simply by circuit simulations using simulators such as SPICE, a physical insight into the circuit and device parameters affecting gate delay can only be obtained from a detailed delay analysis.

A conventional BiCMOS gate is selected for concreteness, and a delay model is developed for a PMOS, driving the n–p–n emitter follower. It is observed that all digital BiCMOS gates have an MOS-BJT combination in common and the gate delay is primarily dependent on the switching properties of this combination. The analysis can be easily extended to cover all other subcells.

Consider the circuit for Fig. 3(a), which shows the principal circuit elements affecting the rise-time response, and a delay expression is derived in detail for this transient response. C_1 in Fig. 3(b) accounts for parasitic capacitance at the base:

$$C_1 = C_{bd1} + C_{bd2} + C_{g4} \tag{3}$$

The first two terms account for drain junction capacitances of M_1 and M_2, and C_{g4} is the gate oxide capacitance due to M_4. The base-emitter capacitance C_E and collector junction capacitance C_C are also included in the model.

For proper operation of the BiCMOS gate, the collector resistance R_C must be low enough to prevent foward biasing of the base-collector junction. With the equivalent circuit set up as described above, the equations governing the gate transient are derived when the input falls to its lowest level, at $t = 0$; M_1 turns ON and operates initially in the saturation region. Its drain current charges the base-emitter capacitance of Q_1 until $V_{be,Q1} = V_{be(on)}$ when Q_1 turns ON. The emitter current of Q_1 increases sharply, pulling up the base voltage of Q_1 and the output node. Referring to Fig. 2, as the base voltage of Q_1 exceeds the threshold voltage of the NMOS transistor, M_4 turns ON and discharges the base charge of Q_2. Thus, Q_2 is OFF, and as the base voltage of Q_1 reaches $V_{dd} - V_{DS(sat)}$, M_1 enters the triode region and its drain current drops gradually. Consequently the collector current of Q_1 starts decreasing. As the output voltage approaches $V_{dd} - V_{be(on)}$, transistor Q_1 gradually turns OFF.

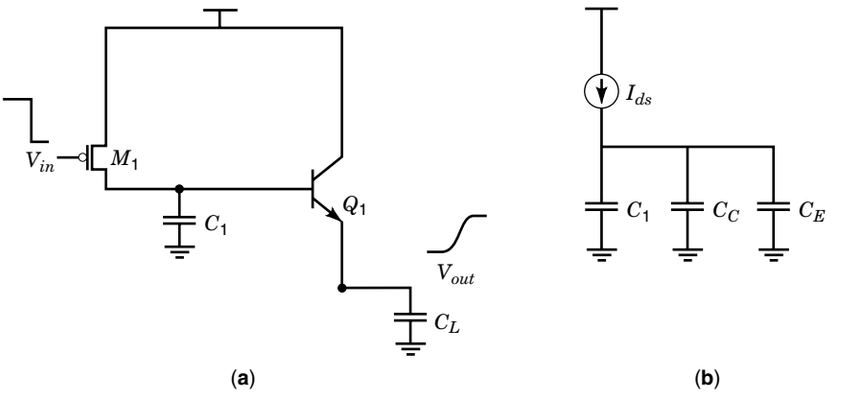


Figure 3. (a) Pull-up section of a BiCMOS gate. (b) Transient equivalent circuit when Q_1 is OFF.

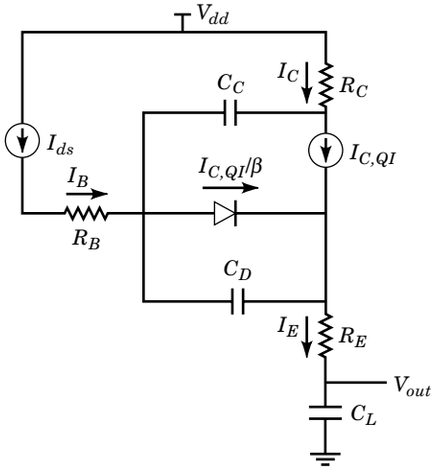


Figure 4. Transient equivalent circuit of the pull-up section when Q_1 turns ON.

The time period from $t = 0$ to the time when output reaches $V_{dd}/2$ at $t = T_d$ can be divided into three parts:

$$T_d = T_1 + T_2 + T_3 \tag{4}$$

where T_1 is the time needed for the drain current of M_1 to charge the net capacitance at the base until the base-emitter voltage V_{be} reaches the forward diode drop. During this period, diffusion capacitance C_D is absent from the circuit and M_1 operates in saturation. During T_2, M_1 still operates in saturation and the output also rises during this time, especially when C_L is small compared to base-emitter capacitance C_E, and during T_3, M_1 enters into the triode region until the output voltage reaches the switching voltage $V_{dd}/2$, which represents 50% of the rise time.

Writing the current equation at the base node of Q_1 in Fig. 3(a) yields the following differential equation:

$$I_{ds} = (C_C + C_E + C_1)\frac{dV_{be}}{dt} \tag{5}$$

Solving the above differential equation, we get

$$T_1 = \frac{(C_C + C_E + C_1)V_{be(on)}}{I_{ds}} \tag{6}$$

where

$$I_{ds} = \frac{C_{on}\mu(V_{dd} - |V_T|)^2 W}{2L} \tag{7}$$

During the second interval T_2, Q_1 is ON, and M_1 is in saturation. The equivalent circuit is shown in Fig. 4, where I_C is the collector current and $I_{C,Q1}/\beta$ is the time-varying current through the forward base diode. C_D represents the diffusion capacitance of the BJT due to the forward stored charge whose instantaneous value is $qI_{C,Q1}\tau_f/kT$. It is the inclusion of C_D that accounts for the role of the base transit time in gate switching. Applying Kirchhoff's current law at the base node,

$$I_{ds} = C_D\frac{dV_{be}}{dt} + \frac{I_{C,Q1}}{\beta} + C_C\frac{d(I_C R_C + V_{be} - V_{dd} + V_o)}{dt} \tag{8}$$

where

$$I_{C,Q1} = I_S e^{qV_{be}/kT} \tag{9}$$

V_{be} is relatively constant when compared to V_o and can thus be neglected in the last term of the equation. If $C_L > C_C$, then $I_C \approx I_{C,Q1}$. Additionally, if $\beta_f \gg 1$, then $I_E \approx I_C$. Hence

$$I_{C,Q1} = I_S e^{qV_{be}/kT} = C_L\frac{dV_o}{dt} \tag{10}$$

and Eq. (8) becomes

$$\begin{aligned} I_{ds} &= C_D\frac{dV_{be}}{dt} + \frac{I_{C,Q1}}{\beta} + C_C\frac{dV_o}{dt} + R_C C_C\frac{dI_{C,Q1}}{dt} \\ &= \frac{qI_{C,Q1}\tau_f}{kT}\frac{dV_{be}}{dt} + \frac{I_{C,Q1}\tau_f}{\beta} + \frac{C_C}{C_L}I_{C,Q1} + R_C C_C\frac{dI_{C,Q1}}{dt} \end{aligned} \tag{11}$$

since

$$\frac{dI_{C,Q1}}{dt} = \frac{q}{kT}I_{C,Q1}\frac{dV_{be}}{dt} = C_L\frac{d^2V_o}{dt^2} \tag{12}$$

Equation (11) becomes

$$I_{ds} = \frac{1}{\beta^*}I_{C,Q1} + \tau_f^*\frac{dI_{C,Q1}}{dt} \tag{13}$$

where

$$\frac{1}{\beta^*} = \left(\frac{1}{\beta} + \frac{C_C}{C_L}\right) \quad \text{and} \quad \tau_f^* = \tau_f + R_C C_C$$

The solution of Eq. (13) is

$$I_{C,Q1} = \beta^* I_{ds}(1 - e^{t/\beta^*\tau_f^*}) \tag{14}$$

Equation (10) is then solved for V_o using Eq. (14):

$$\frac{dV_o}{dt} - \frac{\beta^*}{C_L}I_{ds}(1 - e^{t/\beta^*\tau_f^*}) \tag{15}$$

which upon integrating yields

$$V_o(t) = \frac{I_{ds}(t^2)}{2\tau_f^* C_L} \tag{16}$$

At $t = T_2$,

$$|V_T| - V_{be(on)} = \frac{I_{ds}(T_2^2)}{2\tau_f^* C_L} \tag{17}$$

Solving

$$T_2 = \sqrt{\frac{2[|V_T| - V_{be(on)}]\tau_f^* C_L}{I_{ds}}} \tag{18}$$

During T_3, V_{be} is assumed to remain constant at 0.7V. I_C begins to rise, resulting in a rise in the output. At $t = T_3$ the output reaches $V_{dd}/2$, and the resulting equivalent circuit is shown in Fig. 5. M_1 is in the linear region and is modeled by an equivalent channel resistance.

Using Kirchhoff's current law at the base

$$\frac{V_{dd}-V_{be}-V_o}{R_{CH}+R_B}=\frac{I_{C,Q1}}{\beta}+C_D\frac{dV_{be}}{dt}+C_C\frac{d(V_{be}+I_CR_C+V_o-V_{dd})}{dt} \tag{19}$$

Since V_{be} is relatively small compared to V_o, it can be neglected in the last term. So

$$\begin{aligned}\frac{V_{dd}-V_{be}}{R_{CH}+R_B}=&\frac{V_o}{R_{CH}+R_B}+\frac{I_{C,Q1}}{\beta}+\frac{qI_{C,Q1}\tau_f}{kT}\frac{dV_{be}}{dt}\\&+C_CR_C\frac{qI_{C,Q1}}{kT}\frac{dV_{be}}{dt}+C_C\frac{dV_o}{dt}\end{aligned} \tag{20}$$

Using Eqs. (10) and (12) in Eq. (20),

$$\begin{aligned}\frac{V_{dd}-V_{be}}{R_{CH}+R_B}=&\frac{V_o}{R_{CH}+R_B}+\frac{C_L}{\beta}\frac{dV_o}{dt}+\tau_fC_L\frac{d^2V_o}{dt^2}\\&+C_CR_CC_L\frac{d^2V_o}{dt^2}+C_C\frac{dV_o}{dt}\end{aligned} \tag{21}$$

Equation (21) becomes

$$\frac{d^2V_o}{dt^2}+\frac{1}{\beta^*\tau_f^*}\frac{dV_o}{dt}+\frac{V_o}{T_0^2}=\frac{V_{dd}-V_{be}}{T_0^2} \tag{22}$$

where

$$T_0=\sqrt{(R_{CH}+R_B)C_L\tau_f^*} \tag{23}$$

Solving Eq. (22), the expression of the output voltage is

$$V(t)=(V_{dd}-V_{be})+[Ae^{(m_1t)}+Be^{(m_2t)}]e^{-t/2\beta^*\tau_f^*} \tag{24}$$

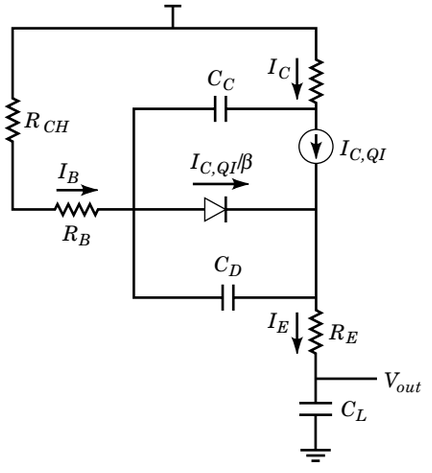


Figure 5. The transient equivalent circuit of the pull-up section when M_1 enters the triode region.

where

$$m_1,m_2=\pm\frac{1}{T_0}\sqrt{\left(\frac{T_0}{2\beta^*\tau_f^*}\right)^2-1} \tag{25}$$

In the above expression, the term $T_0/2\beta^*\tau_f^*$ is less than 1 for typical device and circuit parameters, forcing the roots m_1 and m_2 to be imaginary numbers. Consequently Eq. (24) can be expressed as

$$V(t)=(V_{dd}-V_{be})+\left[C\sin\left(\frac{t}{T}\right)+D\cos\left(\frac{t}{T}\right)\right]e^{-t/2\beta^*\tau_f^*} \tag{26}$$

where

$$T=\frac{T_0}{\sqrt{1-(T_0/2\beta^*\tau_f^*)^2}} \tag{27}$$

At time $t = T_2$,

$$V_o=|V_T|-V_{be(on)}=D+V_{dd}-V_{be(on)} \tag{28}$$

Since the threshold voltage of the MOSFET is so low, it is very close to the base-emitter turn-on voltage,

$$D=-(V_{dd}-V_{be}) \tag{29}$$

Also

$$I_C(t=0)=C_L\frac{dV_o}{dt}=0\Rightarrow C=\frac{-T(V_{dd}-V_{be})}{2\beta^*\tau_f^*} \tag{30}$$

Thus the expression for the output voltage is

$$V(t)=(V_{dd}-V_{be})\left[1-\left(\cos\left(\frac{t}{T}\right)+\frac{T}{2\beta^*\tau_f^*}\sin\left(\frac{t}{T}\right)\right)e^{-t/2\beta^*\tau_f^*}\right] \tag{31}$$

Then, since $V(t) = (V_{dd} - V_{be})/2$ at time $t = T_3$, the delay component for this interval is

$$T_3=\frac{\pi}{3}T \tag{32}$$

Hence the full 50% of the rise-time delay is given by

$$\begin{aligned}T_d=T_1+T_2+T_3=&\frac{(C_C+C_E+C_1)V_{be(on)}}{I_{ds}}\\&+\sqrt{\frac{2[|V_T|-V_{be(on)}]C_L\tau_f^*}{I_{ds}}}+\frac{\pi}{3}\frac{T_0}{\sqrt{1-(T_0/2\beta^*\tau_f^*)^2}}\end{aligned} \tag{33}$$

For typical device and circuit parameters the delay is predominantly determined by T_0.

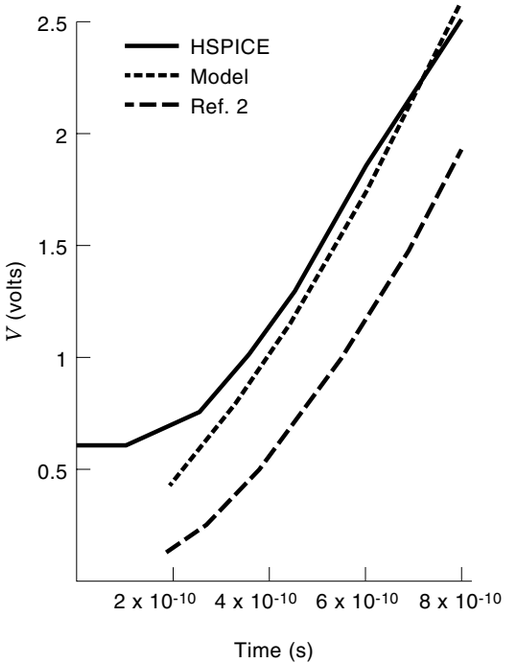


Figure 6. Comparison of analytical and HSPICE simulations.

Figure 6 demonstrates the reasonable agreement between the delay evaluation from the above derived analytical model and those extracted from HSPICE simulations. The figure also illustrates the plot referringg to the delay analysis reported in (2). The device parameters for the BiCMOS technology are listed in Table 1.

Gate Comparison

The logic function is restricted to a simple inverter with the justification that the inverter is the basic logic gate and that the performance of the inverter can be extrapolated to the performance of the other subcells. The CMOS and BiCMOS inverter delay is plotted in Fig. 7 as a function of fanout. The simulations obtained from HSPICE are shown for CMOS one-, two-, and three-stage drivers. Note that one-stage gate delay has a lower delay than the two- or three-stage gates up to a fanout of 5. It is also comparable to BiCMOS delay when the fanout is low. The two-stage gates crossover with the three-stage gates occurs at a relatively large fanout of 15 as shown in the figure. Hence for the smaller fanout the BiCMOS gate must be compared with the one-stage gate, and for the larger fanout with the two- and three-stage gates. Figure 7 shows that for fanouts up to 5, the BiCMOS advantage over the CMOS (single stage) improves. For high-fanout gate sites, a BiCMOS gate is superior to the CMOS drivers. The analysis can be easily extended to the other subcells of the BiCMOS gates.

This section introduced the basic cells implemented in BiCMOS technology, and a full transient analysis of the BiCMOS gate during switching is performed. Although the details are provided for the output pull-up transition of the conventional gate, the analysis is easily extended to other subcells. The performance comparison of BiCMOS and CMOS subcells is provided as a function of fanout and output load. In the following sections, we will first provide a systematic method for constructing an area-time optimal CMOS parallel adder. The approach is based on Ladner and Fischer's parallel prefix computation (3), and is essentially a lookahead addition. The basic tiling cells in CMOS used to implement the circuit blocks of a parallel adder are introduced. Following that, we will present improvements that can be achieved with the introduction of BiCMOS cells in the carry-generation circuit of CMOS adders. The carry-propagation delay due to large fanout and interconnect capacitances is a major factor determining the performance of parallel adders. Besides, with the view of driving capability of bipolar transistors, the BiCMOS cells are adopted to drive large fanout and heavy capacitive loads on the critical path of the fast carry-generation circuit. HSPICE simulation results, and for different

Table 1. Bipolar Device Parameters

Parameter	Unit	Value
I_S	A	1.4391×10^{-16}
β_f	—	100
τ_f	s	12p
I_{KF}	A	1.6939×10^{-2}
β_r	—	1.0
τ_r	s	0.0
I_{KR}	A	0.5
R_E	Ω	10
R_B	Ω	450
R_C	Ω	100
C_{JE}	F	12×10^{-15}
C_{JC}	F	10×10^{-15}
C_{JS}	F	25×10^{-15}

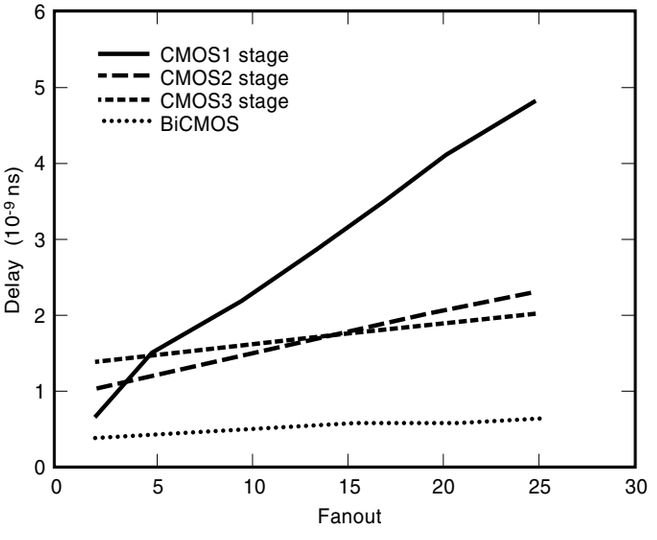


Figure 7. BiCMOS and CMOS delay versus fanout.

data width, parallel adders are presented to show the effectiveness of the mixed CMOS/BiCMOS parallel adder design.

CMOS PARALLEL ADDERS

Parallel Adders

Much attention has been paid to the tradeoff between time and number of gates, but little attention has been paid to the problem of connecting the gates in an economical and regular way to minimize the chip area and optimize the critical path delay.

The adder is the major component in an arithmetic logic unit (ALU), and the ALU is often the workhorse of a computational circuit. There are many kinds of adders available for conventional number systems, some implementations of adders are classified as follows, and the synonym and asymptotic time complexity of adders for n-bit data width are also given:

1. Basic ripple-carry adders : $O(n)$
2. Carry lookahead adders : $O(\log n)$
3. Carry-skip (bypass) adders : $O(n^{1/l})$, where l is the number of skip layers
4. Carry-select adders : $O(\log n)$

It is well known that the delay time of a standard ripple-carry adder can be dramatically decreased by employing the scheme of the carry lookahead addition which makes the slow signals arrive earlier. The carry-skip adders intend to improve the performance of the basic ripple-carry adder by making early signals more available in trading the available time against resources. In the carry-select adder, early signals are duplicated at the expense of additional resources to reduce the number of levels in the adder.

Carry Lookahead Adders

Variable carry lookahead adders have been investigated by many researchers (4,5,6,7). The resulting adder circuitry has constant delay time but contains certain gates whose fanin is unbounded and certain gates whose fanout is unbounded. Carry lookahead adders result from expanding the recurrence equation that describes the set of carries generated by the adder circuitry. From the equations describing the ith carry and sum bits,

$$c_i = (a_i + b_i)c_{i-1} + a_i b_i \quad \text{and} \quad s_i = a_i \oplus b_i \oplus c_i \tag{34}$$

the generate and propagate variables can be defined as

$$g_i = a_i b_i \quad \text{and} \quad p_i = a_i + b_i \tag{35}$$

Then the problem of computing the carries can be described by the simple first-order linear recurrence

$$c_i = p_i c_{i-1} + g_i \tag{36}$$

The relation above corresponds to the fact that the carry c_i is either generated by a_i and b_i or propagated from the previous carry c_{i-1}. Expanding Eq. (36), we obtain

$$c_i = g_i + \sum_{j=0}^{i-1} \left(\prod_{k=j+1}^{i} p_k \right) g_j \tag{37}$$

The fundamental carry operation, o, introduced by Brent and Kung (5) is used:

$$(g_l, p_l) o (g_r, p_r) = (g_l + p_l g_r, p_l p_r) \tag{38}$$

p_l denotes that a carry will propagate across bit position l, and g_l denotes that a carry is generated at bit position l. The term $p_l p_r$ denotes that a carry will propagate from bit r to bit l. Similarly $(g_l + p_l g_r)$ denotes that a carry is generated in at least one of the bit positions from r to l inclusive and propagated to bit position l.

The binary o operation provides an interesting analogy between placing parentheses in an equation and different adder configurations. For example, the carry combination equation for a four-bit ripple adder is

$$((((g_0, p_0) o (g_1, p_1)) o (g_2, p_2)) o (g_3, p_3)) \tag{39}$$

Equation (39) indicates that the propagate and generate signals for the least significant groups (g_0, p_0) and (g_1, p_1) are combined first; then that result is combined with the next group, and so on, in a linear fashion. To combine n groups, $n - 1$ carry operations are performed sequentially,

$$(((g_0, p_0) o (g_1, p_1)) o (g_2, p_2)) o (g_3, p_3))) \tag{40}$$

Equation (40) indicates that the two lower and upper groups are combined simultaneously, and then the two results are combined. With this approach, $\log n$ sets of o operations are performed.

There are many problems that arise in attempts to implement a carry look-ahead adder in VLSI using Eq. (37). First of all, there are many multi-input gates contained in the resulting circuitry. For CMOS technology, the delay time may be proportional to the number of inputs to the gate. To solve the problem, each multi-input gate needs to be replaced with a balanced tree structure that has bounded fanin to each gate; modified circuitry then has a logarithmic delay time.

The other problem with the resulting circuitry for a carry look-ahead adder is the fanout effects; a large fanout represents large load capacitance and time delays. These fanouts can be traded for shorter interconnects and a smaller area, which may result in a faster circuit. Another problem in implementing a carry look-ahead adder is the area required to lay it out. Each carry requires a total of $1 + 2 + 3 + \ldots + i = i(i + 1)/2$ inputs to its gates, so that area $O(i^2)$ is required to realize it. The silicon area required to realize all the carries is thus $O(i^3)$, which ignores the interconnection complexity. The reason for the large amount of area computation required is that each carry directly generates all the subcomputations that it requires, so that much duplicate work is performed among all the carries. The key to reducing adder area is to avoid the duplicate work implemented in the adder.

One way to improve the speed of a carry look-ahead adder is to use BiCMOS technology which offers advantages enhancing the performance of VLSI circuits (8,9,10,11,12,13).

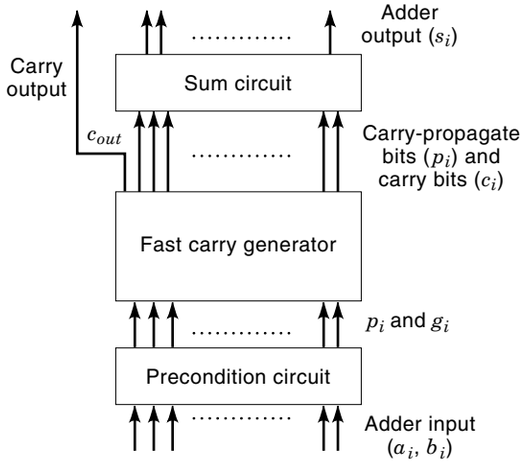


Figure 8. Three functional blocks of a parallel adder.

BiCMOS precharge circuits have been used in carry lookahead adder circuits (8) to improve speed performance. However, only the delay associated with the output was shortened. The propagation delay associated with the internal carry bits, which is important for a large-scale carry lookahead adder was still not improved. Many researchers have implemented the adder designs in CMOS and BiCMOS technology; however, the design of mixed CMOS/BiCMOS technology has not been explored.

The primary motivation of BiCMOS is the possibility of using the best features of both MOSFETs and BJTs in a single integrated circuit. BiCMOS circuits are utilized in combination with CMOS structures for the design of critical paths that would render an optimum system performance in terms of speed and area. A general comparison between optimized CMOS and BiCMOS design adders is carried out and the speedup factor of BiCMOS over CMOS is reported. From the results of the comparison, we can generalize that the BiCMOS adders can achieve significant speedup over CMOS as the data width increases.

CMOS Area-Time Optimal Carry Lookahead Adders

Basic Cells. The complete block diagram of a parallel adder is shown in Fig. 8. It consists of three functional blocks: the precondition circuit, the fast carry generator, and the sum circuit. The precondition circuit gates in the adder inputs a_i and b_i to generate the initial carry-propagate term p_i and carry-generate term g_i for each bit i. The computed p_i and g_i terms of each bit i are fed into the fast carry generator. This work is focused on the accelerated carry computation and the delay obtained in the carry generator. The carry bit c_i obtained from the fast carry generator is combined in the sum circuit with the carry-propagate bit p_i from the precondition circuit to generate the sum bit s_i,

$$s_i = p_i \oplus c_{i-1} \quad \text{for } i = 1, \ldots, n \tag{41}$$

To implement the design of a fast carry computation circuit, three basic types of tiling cells are required: black cells, white cells, and driver cells, as shown in Fig. 9. The terms "black" and "white" cells come from (5). Note that some of the inputs to the black and white cells "pass through" the cells. Specifically the (g_r, p_r) inputs of the black cells are available as outputs. This convention simplifies cascading the subcells and wiring diagrams.

The black cell is first implemented in static CMOS to perform the binary o operation: $(g_l, p_l)o(g_r, p_r) = (g_l + p_l g_r, p_l p_r)$, which are of two types, the black ba cell and the black bb cell as shown in Fig. 10. The ba cell of Fig. 10(a) gates in the positive-true signals and generates the complemented outputs, as the bb cell of Fig. 10(b) gates in the complemented inputs and outputs positive-true signals. Each of the cells shown in Figure 10(a) and 10(b) is composed of P and G subcells, which produce p_{out} and g_{out} signals, respectively. For equal drive capability, the widths of MOS devices are varied while keeping the lengths constant at 2λ. Minimum-length transistors are used for the pull-down network of each subcell implementing the black cell. PMOS transistors that form the pull-up circuit are ratioed in such a way that the maximum pull-up and pull-down channel resistances are made equal.

To maintain proper signal polarity, while implementing CMOS technology which features inverting logic, it is necessary to introduce inverters in the circuit. This is achieved by using the white cells shown in Fig. 11. To reduce wiring diagrams, white cells are of two types, wa and wb cells. The wb cell is a modified white cell that provides a turning corner for input signals.

The speed performance of parallel adders is mainly determined by the propagation delay involved in the critical paths

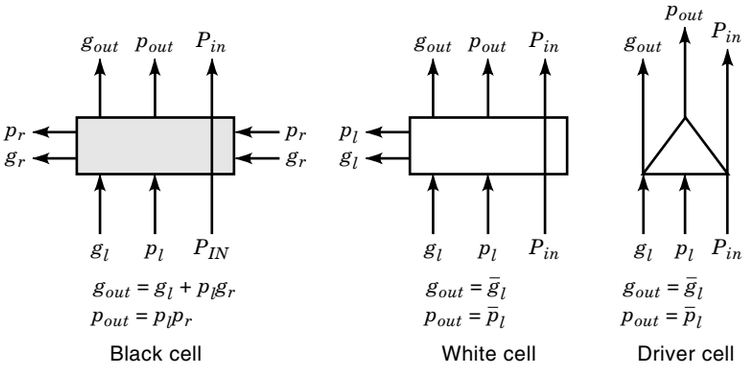


Figure 9. Basic types of tiling cells.

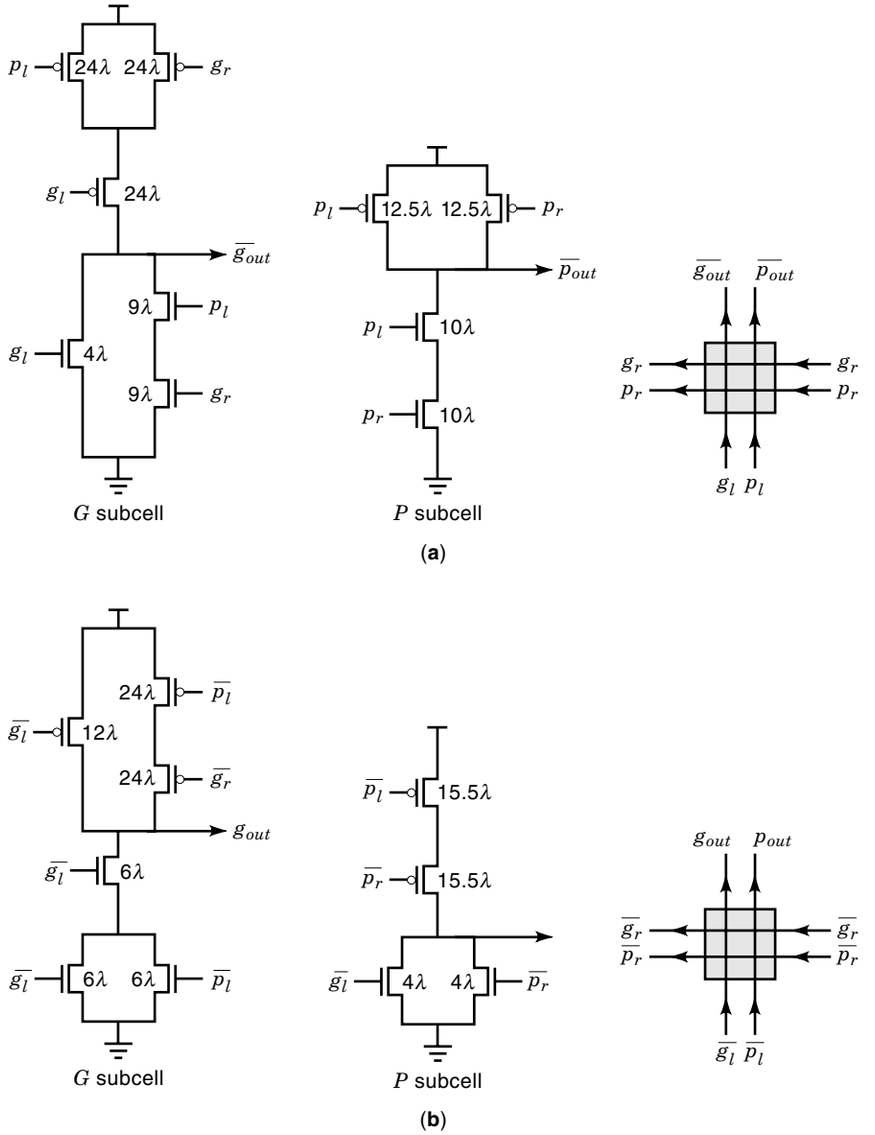


Figure 10. Black cell implementation in static CMOS. (a) The black *ba* cell (b) The black *bb* cell.

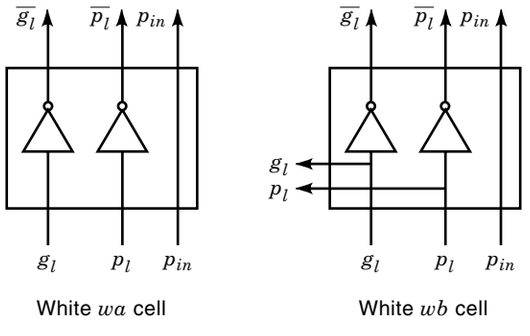


Figure 11. White cells implemented in CMOS: *wa* and *wb* cells.

of the fast carry-generator circuit, which drive large fanouts and interconnect capacitances. Driver cells are used in case of long wire interconnects or large fanouts. A specially ratioed inverter in single stage or in cascaded stages is the singular subject used as the driver cell shown in Fig. 12. It is clear at this stage that the black cells are used for computation, the white cells are used for electrical requirements, and the driver cells are used for performance improvements.

Design Architecture. To construct a fast adder, the signal delay associated with each type of the subcell is analyzed. For the CMOS design, the cell resistance and capacitances are estimated in order to compute the associated signal delay.

For equal drive capability, the width of the MOS devices are varied while keeping the lengths to a minimum. PMOS transistors, which form the pull-up circuit, are ratioed in a way that the maximum pull-up and pull-down channel resis-

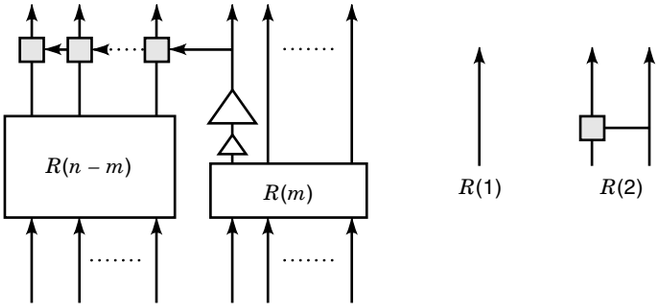


Figure 12. Drivers used in recursive construction of the fast carry generator.

tances R_c are made equal. In integrated systems, capacitances of circuit nodes are due not only to the capacitance of gates connected to the nodes but also to capacitances to ground of signal paths connected to the nodes and to other stray capacitances. The total interconnect capacitance C_i and resistance R_i due to the signal paths are not negligible.

In a static CMOS design, a pair of PMOS pull-up and an NMOS pull-down transistors constitutes a basic inverting unit. The input signal drives both the pull-up and pull-down transistors. Let C_g be the total gate capacitance of the unit, then the approximate generation time of the output signal is given by

$$t_{out} = t_{in} + (R_c + R_i f)(C_i + C_g) f \tag{42}$$

where t_{in} is the input ready time and f is the fanout factor associated with the cell.

As can be seen, the delay per inverting stage is multiplied by a fanout factor. The overall performance of the circuit may be seriously degraded if it contains any large fanouts. In such cases the drivers that are an integral part of the circuit layout are modeled explicitly.

In metal interconnects, if the channel resistance R_c is much greater than $R_i f$, then t_{out} becomes

$$t_{out} = t_{in} + R_c(C_i + C_g) f \tag{43}$$

Let τ be the primary time constant in calculating the delay through elementary inverting logic stages, defined as

$$\tau = R_c(C_i + C_g) \tag{44}$$

then

$$t_{out} = t_{in} + \tau f \tag{45}$$

which is a simple, conventional timing model (14).

In the case of the adder design, the fanout f of a subcell varies depending on the type of the subcell and on the type and number of its succeeding cells. This can be illustrated by considering the layout of the 32-bit adder shown in Fig. 13 where each cell is identified by a pair of height and bit coordinates.

For example, in Fig. 13, consider the black cell at (4, 8) which refers to the fourth cell on the vertical path of bit 8. Recall that the black cell is implemented to perform the binary o operation: $(g_l, p_l) o (g_r, p_r) = (g_l + p_l g_r, p_l p_r)$. Therefore the left operand of cell (4, 8), namely (g_l, p_l), comes from cell (3, 8) which is just vertically below cell (4, 8). The outputs p_{out} and g_{out} of cell (3, 8) are the inputs p_l and g_l of cell (4, 8). The fanout of p_{out} of cell (3, 8) is 2, since it drives both P and G subcells of cell (4, 8), whereas the fanout of g_{out} of cell (3, 8) is 1, since it drives only the G subcell of cell (4, 8). The same analysis extends to all the cells in the circuit.

The right operand of cell (4, 8), namely (p_r, g_r), comes from driver cell (3, 3) whose output signals make a turn in wb cell (4, 3) and supply the right operand to each of black cells (4, 4), (4, 5), (4, 6), (4, 7) and (4, 8). Thus the fanout of g_{out}(or p_{out}) of driver cell (3, 3) is 6, since it drives each subcell of (4, 3), (4, 4), (4, 5), (4, 6), (4, 7), and (4, 8). All the cells driven by the driver cells in the horizontal path are indicated by a bold line.

Since the delay through a cascaded driver depends on the drives fanout f_d, the ratio r between the successive stages, and the number s of cascaded stages (14), the minimum delay is obtained by taking the driver ratio,

$$r = f_d^{1/(s+1)} \tag{46}$$

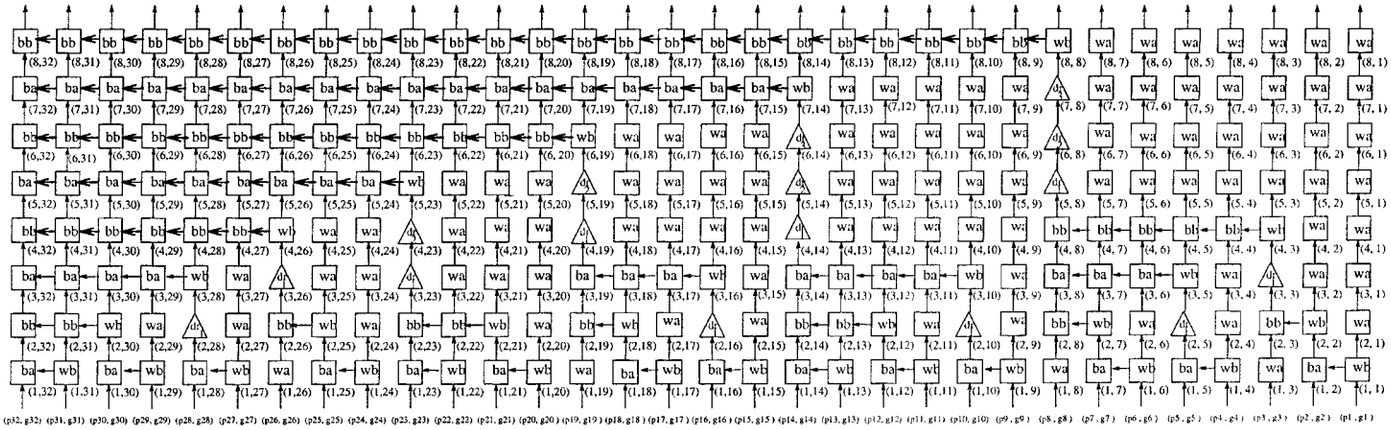


Figure 13. An optimal CMOS 32-bit fast carry generator.

The corresponding minimum propagation delay of the cascaded driver is

$$\text{delay}(s, f_d) = (s+1)(f_d^{1/(s+1)})\tau \tag{47}$$

Thus, for an s-stage driver of fanout f_d,

$$t_{dout} = t_{din} + \text{delay}(s, f_d) \tag{48}$$

Note that if the number of stages is zero, that is, in the case of a single inverter, then Eq. (48) is the same as Eq. (45).

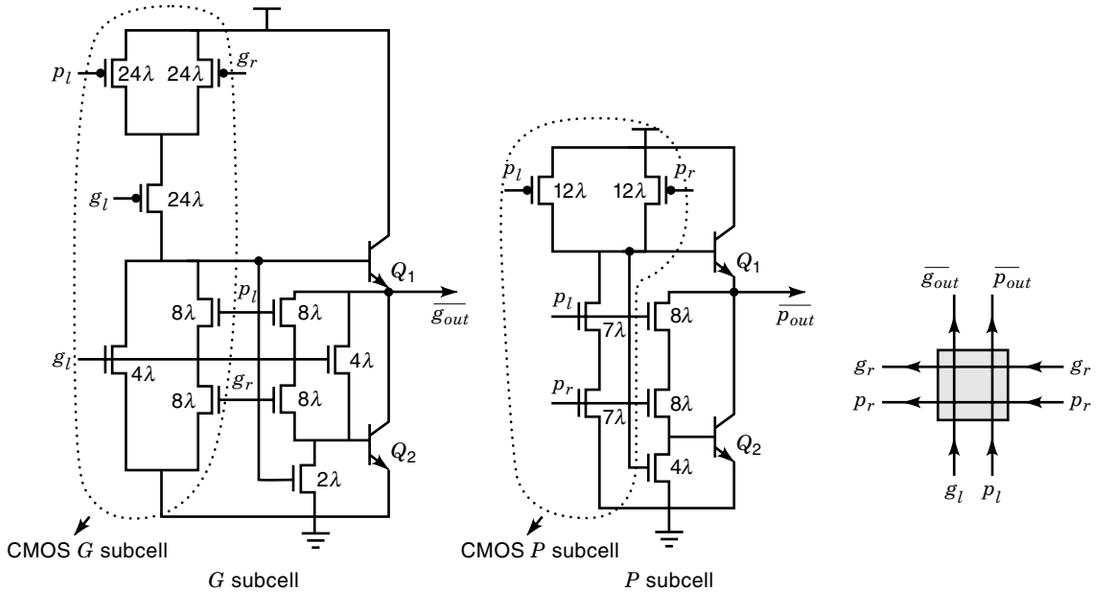
With the analysis above the generation time for each circuit signal as the sum of its input ready time and delay factor can be analyzed. Let t_{gout} be the time when signal g_{out} is ready, and let t_{gl} be the t_{gout} of the cell producing g_l and t_{gr} the t_{gout} of the cell producing g_r. Similarly t_{pl} and t_{pr} represent the t_{pout} of the cell producing p_l and p_r, respectively.

Consider the g subcell of the black cell; the input ready time t_{gin} can be formulated as

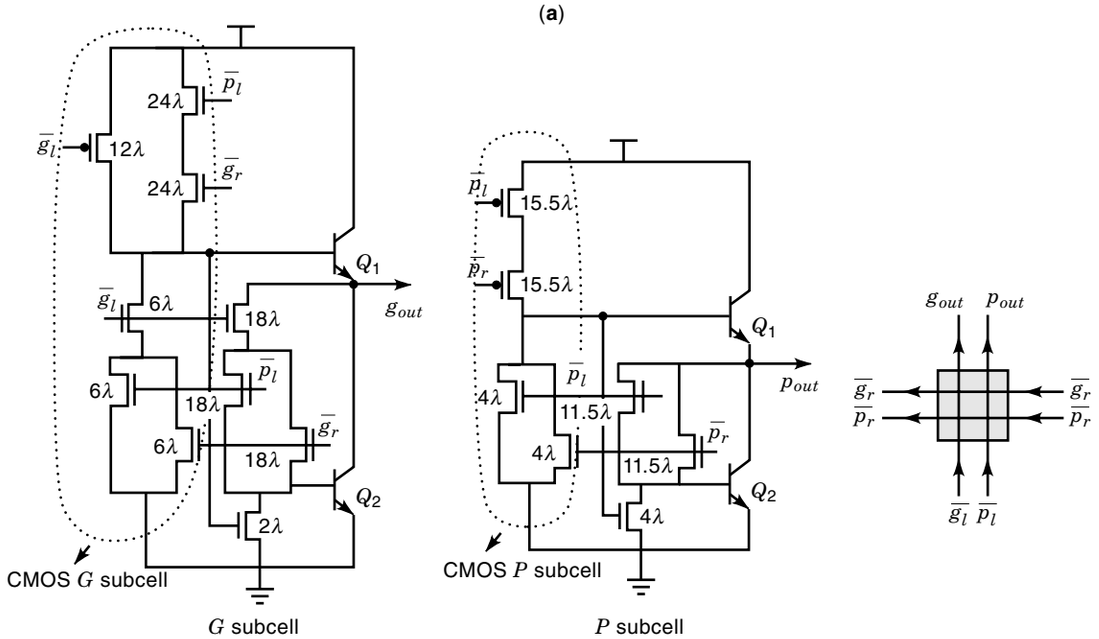
$$t_{gin} = \max\{t_{gl}, t_{pl}, t_{gr}\} \tag{49}$$

and t_{pin} for the p subcell of the black cell can be formulated as

$$t_{pin} = \max\{t_{pl}, t_{pr}\} \tag{50}$$



(a)



(b)

Figure 14. Black cell implementation in BiCMOS. (a) The black *ba* cell; (b) the black *bb* cell.

Let f_g and f_p be the fanout of the g subcell and p subcell under analysis, then

$$t_{gout} = t_{gin} + \text{delay}(s, f_g) \tag{51}$$

and

$$t_{pout} = t_{pin} + \text{delay}(s, f_p) \tag{52}$$

Equations (51) and (52) depend on the fanout of the cell under consideration, which is determined by the interconnection of modular cells.

With the evaluation of the timing behavior of the basic cells, consider the construction of the fast carry generator $R(n)$ based on the recursive construction of the basic cells $R(1)$ and $R(2)$ as shown in Fig. 12. The $R(n)$ circuit is composed of the subcircuits $R(m)$ and $R(n - m)$ which, in turn, are composed of circuits of even smaller sizes. The $R(n)$ circuit has a large fanout from the most significant bit of the right block, that is, bit m, broadcasting it to all bit positions of the left block. To shorten the critical delay due to this large fanout, a multistage driver is placed at the most significant bit of $R(m)$.

Another critical delay comes from the propagation of the signal through the leftmost vertical path. Since both critical paths converge at the leftmost top cell, it is necessary to decompose the n-bit adder into subcircuits $R(m)$ and $R(n - m)$ by choosing the best placed m. Thus, to evaluate the input ready time of the most significant bit of an n-bit adder, $t_{gin}(n, 1)$, consider the following recurrence:

$$t_{gin}(i, j) = \min_{j \le m < i} \{\max[t_{gin}(i, m+1) + \tau, t_{pin}(i, m+1) + 2\tau, t_{gin}(m, j) + f(i, m, j)\tau]\} \tag{53}$$

$$t_{pin}(i, j) = \min_{j \le m < i} \{\max[t_{pin}(i, m+1) + 2\tau, t_{pin}(m, j) + f(i, m, j)\tau]\} \tag{54}$$

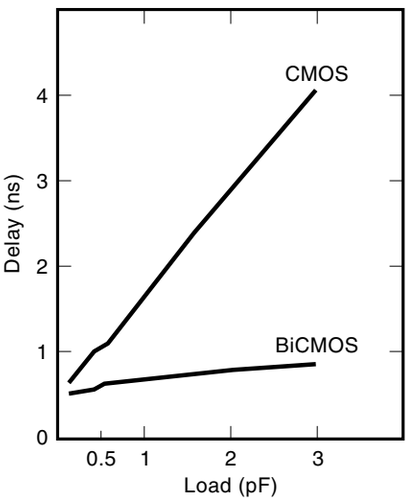


Figure 15. Delay performance of BiCMOS and CMOS of the P subcell of a ba cell.

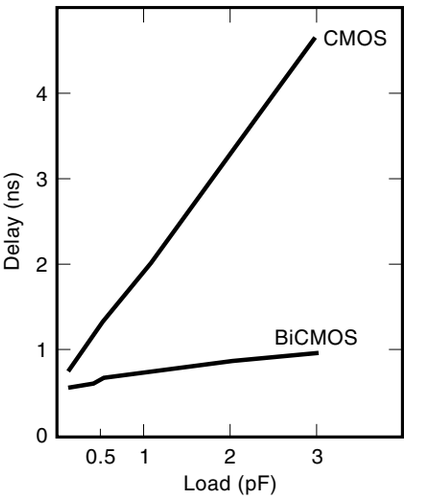


Figure 16. Delay performance of BiCMOS and CMOS of the G subcell of a ba cell.

where

$t_{gin}(i, j)$ is input ready time for the g term of the most significant bit of an adder block of size $i - j + 1$.
$t_{pin}(i, j)$ is input ready time for the p term of the most significant bit of an adder block of size $i - j + 1$.
$f(i, m, j)$ is load function of the block of size $m - j + 1$ driving that of size $i - m$.

The load function $f(i, m, j)$ is defined as

$$f(i, m, j) = \min_{\substack{0 \le s < u \\ \text{and } s \equiv u \bmod 2}} \{\text{delay}(s, i - m + 1)\} \tag{55}$$

where $u = \max\{0, \text{depth}(i, m + 1) - \text{depth}(m, j)\}$.

In Eq. (55), s is chosen to minimize the signal delay through the driver. The depth of the s-stage driver is limited to u, which is the depth difference between two component adder blocks, $R(i - m)$ and $R(m - j + 1)$. Since the optimal splitting m value for the p signal and the g signal is the same, one-dimensional dynamic programming can be used to calculate the optimal fast carry-generator configurations for n up to any desired datawidth (4).

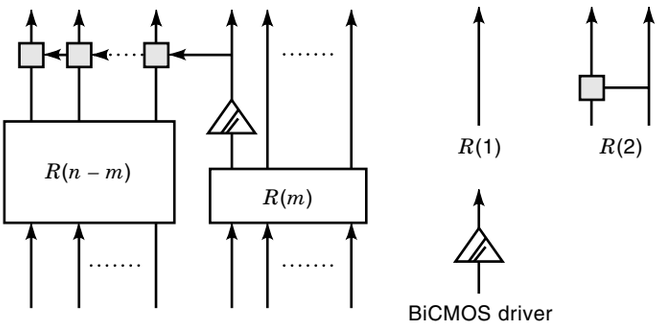


Figure 17. BiCMOS drivers used in recursive construction of the fast carry generator.

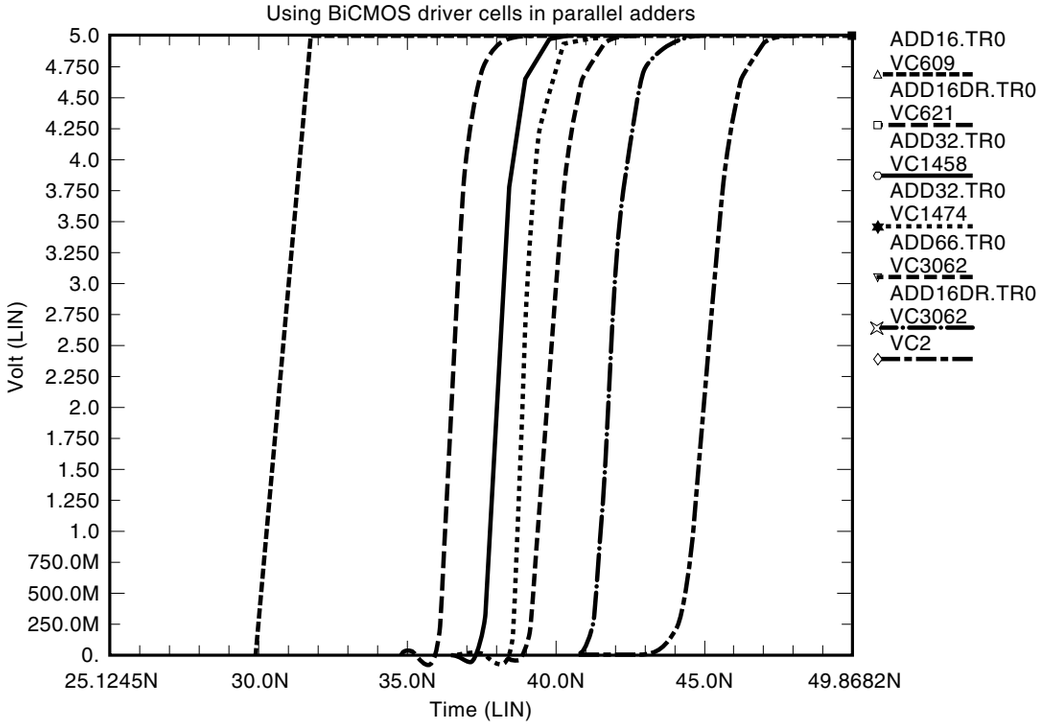


Figure 18. HSPICE simulation of the critical path delay of 16-bit, 32-bit, and 66-bit parallel adders.

In order to regularize and minimize the layout efforts, the number of driver stages is limited to 2 or 3 stages. In addition the driver ratio between successive stages is usually an integer 2 or 3. The performance of CMOS adders is discussed and compared with mixed CMOS/BiCMOS adders in the next section.

MIXED CMOS/BiCOMOS AREA-TIME OPTIMAL ADDERS

Basic Cells

Figure 14 shows the black *ba* and *bb* cells implemented in BiCMOS technology. The logic function is implemented using PMOS and NMOS blocks. The bipolar transistors Q_1 and Q_2 are used as current drivers. HSPICE-measured delay times of the *P* subcell and *G* subcell of black *ba* cell are shown in Figs. 15 and 16. The delay performance of the black *bb* cell is similar to the results of the *ba* cell.

BiCMOS gates are faster than the CMOS gates, especially as the load capacitance, due to high fanout, increases. Using only BiCMOS cells to design the parallel adder, the integration density due to CMOS-based parallel adder is not achieved, and for low-fanout sites, CMOS subcells are faster than BiCMOS ones. Therefore using CMOS *P* and *G* subcells to drive smaller loads, which are not in the critical path, and BiCMOS *P* and *G* subcells to drive the heavy ones, which are in the critical path, gives an optimal solution to the parallel adder design. The following section investigates the performance improvement in the design of optimal parallel adders by using the BiCMOS cells in CMOS adders.

Table 2. Comparison of HSPICE Simulations of 16-bit, 32-bit, and 66-bit Parallel Adders

Data Width	T_{dlh}	T_{dhl}	Avg T_d	Difference in T_d	% Decrease in T_d
CMOS16	7.23	7.9	7.565	—	—
BiCMOS16(dr)	5.67	6.1	5.885	1.68	28.54
BiCMOS16	5.12	6.3	5.71	1.855	32.486
CMOS32	10.96	9.39	10.175	—	—
BiCMOS32(dr)	8.12	7.33	7.725	2.45	31.715
BiCMOS32	6.11	6.74	6.425	3.75	58.365
CMOS66	14.17	11.49	12.83	—	—
BiCMOS66(dr)	8.95	8.68	8.815	4.015	45.547
BiCMOS66	6.81	7.45	7.13	5.7	79.944

Note: CMOS## ⇒ Adder implemented in CMOS. BiCMOS##(dr) ⇒ Adder implemented in CMOS with BiCMOS driver cells. BiCMOS## ⇒ Adder implemented in BiCMOS.

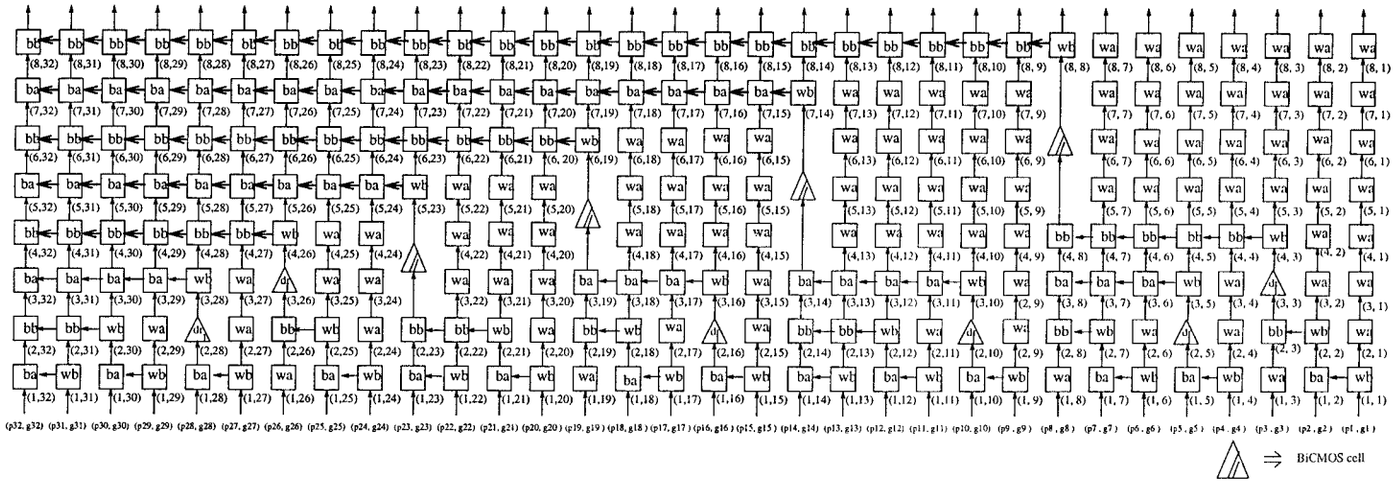


Figure 19. 32-bit fast carry generator using BiCMOS driver cells.

Implementation of BiCMOS Parallel Adder

To achieve a fast carry-generation circuit, the critical path delay must be shortened. The following section discusses the improvement achieved by introducing BiCMOS cells into a CMOS parallel adder design. Figure 17 shows two possible critical paths in the recursive construction of the fast carry-generation circuit. One is the delay through the leftmost bit of $R(m)$, which drives a large a fanout and interconnect capacitances, and the other is the leftmost vertical path of $R(n)$. Since the critical path converges at the leftmost top cell, C_{n-1} is obviously one of the slowest carry output bits.

Using BiCMOS Driver Cells in Parallel Adders

To accelerate the critical path through the leftmost bit $R(m)$, BiCMOS driver cells are introduced at the site of a large fanout to minimize the delay. Note that in the CMOS carry-generator circuit a multistage driver is needed but that it is now replaced by a single BiCMOS driver cell, without increasing the total chip size.

In the parallel adder architecture, only the driver cells with a large fanout are replaced by the BiCMOS driver cells, and all other basic cells are composed of only CMOS devices. The bipolar device works as a high-current output circuit. It

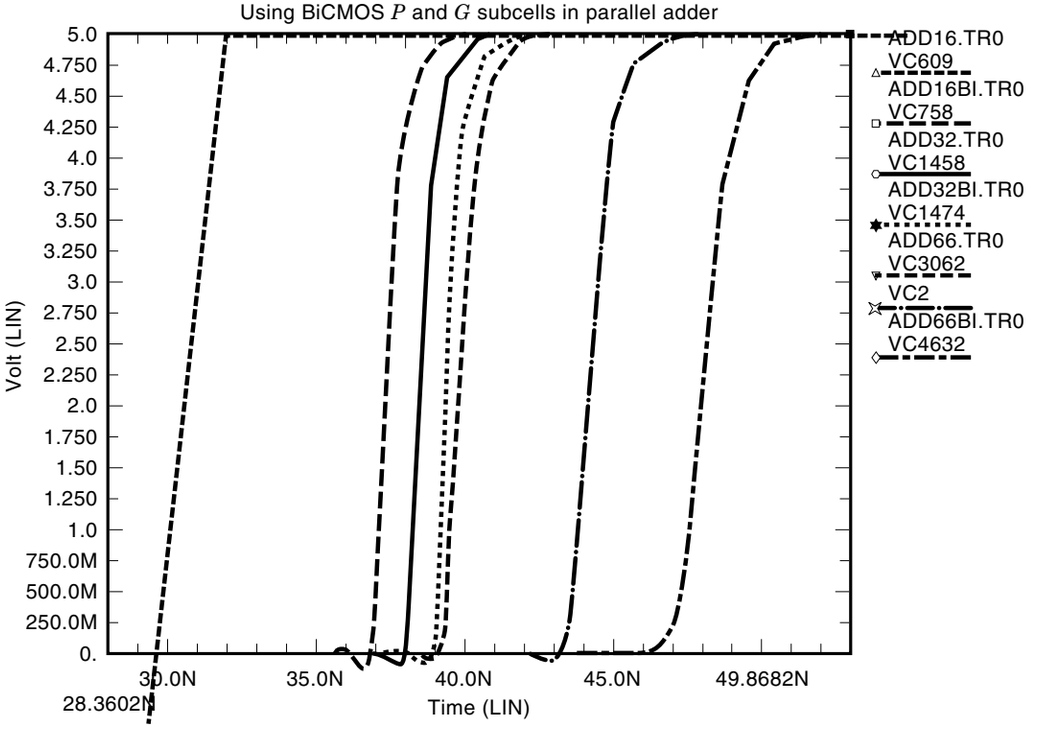


Figure 20. HSPICE simulation of the critical path delay of 16-bit, 32-bit, and 66-bit parallel adders.

accelerates the operation of the CMOS logic circuits; this is verified by evaluating the CMOS parallel adders with BiCMOS driver cells. From HSPICE simulations shown in Fig. 18, it is observed that by introducing BiCMOS driver cells alone, the critical path delay of the parallel adders is shortened by 28.54% in the case of the 16-bit adder, 31.715% in the case of the 32-bit adder, and 45.547% in the case of the 66-bit adder. The results are tabulated in Table 2.

Using BiCMOS *P* and *G* Subcells in Parallel Adders

Figure 19 shows the implementation of a 32-bit parallel adder that uses BiCMOS driver cells to shorten the critical path delay, and the propagation delay through the leftmost vertical path of the carry-propagation circuit is shortened by introducing BiCMOS subcells in the critical path as shown by the bold path of arrows.

From the mixed CMOS/BiCMOS parallel adder design, the speed of the arithmetic operation in the critical path of carry-generation circuit is highly increased, as verifed by evaluating the 16-bit, 32-bit, and 66-bit parallel adders. To compare the performances of the parallel adders for different data widths, HSPICE simulations were obtained for 16-bit, 32-bit, and 66-bit adders as shown in Fig. 20. HSPICE simulations compared with the results of CMOS adders show that the delay was shortened by 32.486% in the case of the 16-bit, 58.365% in the case of the 32-bit, and 79.944% in the case of the 66-bit adder as shown in Table 2.

CONCLUSIONS

This article presented mixed CMOS/BiCMOS parallel adders, which is an improvement over the high-speed area-time optimal adders that have been realized by CMOS static circuits. The CMOS parallel adders suffer from the speed penalty as a result of long wire interconnects and large fanout sites. In order to achieve higher speed in the carry computation, high-speed technology is needed.

BiCMOS technology suitable for high-speed circuits is chosen for the implementation of the basic cells in the carry-generation circuit to drive large-fanout and capacitive loads. A full transient analysis of BiCMOS gate switching has been carried out. The design of the BiCMOS gate is illustrated for the conventional basic inverter cell. The concept introduced to perform the design can be directly applied to the other subcells used in the implementation of the parallel adder. A comparison is made between the CMOS and BiCMOS digital gate performance based on the output capacitive and fanout loads.

From the analysis of the basic cells, CMOS cells are chosen to drive light loads and BiCMOS cells are chosen to drive heavy loads, resulting in a design of mixed CMOS/BiCMOS parallel adders. To demonstrate the performance of the mixed CMOS/BiCMOS parallel adders, design circuits of 16-bit, 32-bit, and 66-bit parallel adders are implemented. From the results shown in Table 2, it can be observed that as the bit number increases, the propagation delay of the mixed CMOS/BiCMOS adders is shortened almost linearly in comparison with the CMOS parallel adders.

Scaling is one of the issues that needs to be investigated. BiCMOS scaling for digital logic gates is different from conventional CMOS scaling because basic circuit behavior is different. BiCMOS gate delay needs to be analyzed at a reduced voltage supply. Voltage supply reduction is necessary only if the device feature sizes are scaled. Therefore an analysis of the effect of the voltage supply reduction must be carried out with device scaling.

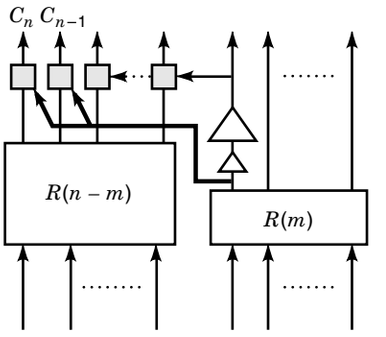


Figure 21. The P and G signals from the leftmost bit of $R(m)$ are directly fed to the most critical bits.

In the case of the high-speed arithmetic logic unit (ALU) used to perform 2's complement signed arithmetic and logic operations on a large data set of numbers, it is advantageous to know the result of the carryout bit and the overflow bit. An overflow output pin flags arithmetic operations that exceed the available 2's complement number range. This pin is logically the exclusive-OR of the carry-output pins C_n and C_{n-1} of an n-bit parallel adder. At the most significant end of the adder, this pin indicates that the result of an arithmetic 2's complement operation has overflowed into the sign bit, causing the sign bit to become erroneous.

If the information is retrieved earlier, the carry bits from the leftmost bit of $R(m)$ could be propagated to the most significant bits immediately, instead of propagating through the buffer cells. The other bit positions are fed through the buffer maintaining the adder architecture. This way the carryout and overflow bit will flag earlier than at the sum of the computation time. The concept of direct feeding to the most significant bits is illustrated in Fig. 21.

BIBLIOGRAPHY

1. H. C. Lin et al., CMOS-bipolar transistor structure, *IEEE Trans. Electron Devices,* **6** (11): 945–951, 1969.
2. W. Fang, A. Brunnschweiler, and P. Ashburn, An accurate analytical BiCMOS delay expression and its application in optimizing high-speed BiCMOS circuits, *IEEE J. Solid-State Circuits,* **27** 2, Feb. 1992.
3. R. E. Ladner and M. J. Fischer, Parallel prefix computation, *J. ACM,* **27** (4): 831–838, 1980.
4. B. W. Y. Wei and C. D. Thompson, Area-time optimal adder design, *IEEE Trans. Comput.,* **39** (5): 1990.
5. R. P. Brent and H. T. Kung, A regular layout for parallel adders, *IEEE Trans. Comput.,* **31** (3): 1982.
6. B. W. Y. Wei and Y.-F. Chen, QAC: A CMOS implementation of the 32-bit Q adder, *Proc. IEEE Int. Conf. Comput. Design: VLSI Comput.,* Port Chester, NY, October 1985.
7. T. F. Ngai, M. J. Irwin, and S. Rawat, Regular, area-time efficient carry-lookahead adders, *J. Parallel Distrib. Comput.,* **3**: 92–105, 1986.

8. J. B. Kuo, H. J. Liao, and H. P. Chen, A BiCMOS dynamic carry lookahead adder circuit for VLSI implementation of high-speed arithmetic unit, *IEEE J. Solid-State Circuits,* vol. 28, no. 3, March 1993.
9. T. Hotta, et al., CMOS/bipolar circuits for 60-MHz digital processing, *IEEE J. Solid-State Circuits,* **21** (5): 1986.
10. T. Hotta, et al., A 70-MHz 32-b microprocessor with 1.0-μm BiCMOS macrocell library, *IEEE J. Solid-State Circuits,* vol. 25, no. 3, June 1990.
11. J. D. Gallia, et al.; High-performance BiCMOS 100k-gate array, *IEEE J. Solid-State Circuits,* **25** (1): 1990.
12. S. H. K. Embabi, A. Bellaour, and M. I. Elmasry, *Digital BiCMOS Integrated Circuit Design,* Drodrecht, The Netherlands: Kulwer, 1993.
13. E. W. Greeneich and K. L. McLaughlin, Analysis and characterization of BiCMOS for high-speed digital logic, *IEEE J. Solid-State Circuits,* **23**: 558–565, 1988.
14. C. Mead and L. Conway, *Introduction to VLSI Systems,* Reading, MA: Addison-Wesley, 1980.

CHIEN-IN HENRY CHEN
Wright State University

ANUP KUMAR
Credence Systems Corporation