

BiCMOS MEMORY CIRCUITS

Many applications, in particular computer caches, require memories that combine both high capacity and high speed. Traditionally, metal–oxide–semiconductor (MOS) memories have provided suitable capacity at low speeds, while bipolar memories have provided high speed but at lesser capacities and usually significantly higher power (see BIPOLAR MEMORY CIRCUITS). By adding a high-performance bipolar transistor to a complementary metal–oxide–semiconductor (CMOS) process, the bipolar complementary metal–oxide–semiconductor (BiCMOS) static random-access memories (SRAMs) were developed to have the density of CMOS SRAMs, yet the high speed of bipolar SRAMs.

BiCMOS memories achieve CMOS memory capacities by using CMOS memory cells, which have superior stability, lower susceptibility to alpha-particle-induced soft errors, much lower static power dissipation, and significantly smaller size than bipolar cells (see SRAM CHIPS). Bipolar or bipolar/CMOS periphery circuitry provides the high speed of bipolar SRAMs by borrowing and expanding on their high-speed circuit techniques. Finally, input/output (I/O) in high-speed stand-alone nonregistered SRAMs can account for nearly half of their access time. Bipolar transistors and emitter-coupled logic (ECL) voltage levels can reduce this I/O penalty significantly in BiCMOS SRAMs (see EMITTER-COUPLED LOGIC and TRANSISTOR-TRANSISTOR LOGIC).

Unfortunately, the advantages of BiCMOS SRAMs must be weighed against their costs. The additional complexity of BiCMOS processes over either CMOS or bipolar processes increases final parts costs both directly and indirectly, because of reduced yield. Furthermore, since BiCMOS is a niche market, development costs are higher than for CMOS, which, for example, can amortize fabricating equipment costs over a much larger market. In addition, power consumption in BiCMOS SRAMs often grows faster than performance in-

creases, hence the speed advantages must often be throttled back to meet system power requirements. Consequently, clever power-saving circuit techniques become mandatory to reap the speed improvements of BiCMOS.

Scaling, from lithography and processing improvements, fuels the semiconductor industry by increasing functionality, capacity, and performance with every succeeding generation, while simultaneously lowering costs. CMOS transistors, having a lateral structure, benefit more from scaling than the vertical bipolar transistor, and they are rapidly approaching the raw speed of bipolar transistors. Optimal performance scaling also dictates voltage reduction, which is problematic in BiCMOS circuit design because of the fixed V_{be} turn-on voltage of the bipolar transistor. (Note that the V_{be} voltage actually increases slightly as bipolar transistors scale, reaching 0.8 V for submicrometer devices.) Though less obvious, the bipolar transistor saturation voltage ($V_{ce,sat} \approx 0.4$ V) also limits BiCMOS voltage scaling because of the higher low-level output of the CMOS BiCMOS buffer and the minimum ECL current-source voltage drop. Supply voltages for CMOS circuits have reduced from 5 V to 3.3 V to 2.5 V in the quarter-micron generation, and they will further reduce to 1.8 V and smaller in future generations. Since power dissipation in CMOS scales as the square of the supply voltage, CMOS performance improves and power dissipation decreases with scaling. Bipolar transistor scaling, however, is slowed because of higher voltage requirements and thus little power dissipation reduction is possible from process improvements.

The combination of bipolar and CMOS transistors in BiCMOS processes permits a large variety of circuit techniques as shown in Fig. 1. CMOS transistors, or field-effect transistors (FETs), whose current drive is proportional to the square of the input voltage, rely on full-rail voltage swings for optimal performance. Most BiCMOS circuit families, such as BiCMOS, merged (mBiCMOS), and BiNMOS, are constructed by replacing or complementing one or more of the final drive FETs with a bipolar transistor (see BiCMOS LOGIC CIRCUITS). The BiCMOS variations attempt to improve on the basic BiCMOS gate, whose reduced output swing limits its performance as supply voltages scale (1). For supply voltages greater than 3 V and when heavily loaded, these BiCMOS gates can be much faster than CMOS gates because the bipolar transistor has a larger transconductance (g_m) than a FET with a similar input capacitance, i.e., a larger current-drive capability. Collectively we refer to these BiCMOS circuit families as “CMOS BiCMOS” because of the similarity to CMOS of the gates and voltage levels.

In contrast to FETs, bipolar transistors, whose current drive is exponential with input voltage, are better suited to current steering logic families such as ECL. These gates rely on the superior sensing capability of bipolar transistors, both better transconductance and better matching, than CMOS transistors. ECL gates have small voltage swings of typically around 550 mV single-ended, and only require a single type of bipolar transistor, the $n-p-n$ (see CURRENT-MODE LOGIC and EMITTER-COUPLED LOGIC). ECL circuits are typically faster than CMOS or CMOS BiCMOS circuits because of the reduced voltage swing they must drive and the minimal delay penalty for complex gates. BiCMOS designs with ECL are aptly referred to as “ECL BiCMOS.”

ECL requires a minimum supply voltage, V_{supply} , that is independent of process scaling. For conventional three-level se-

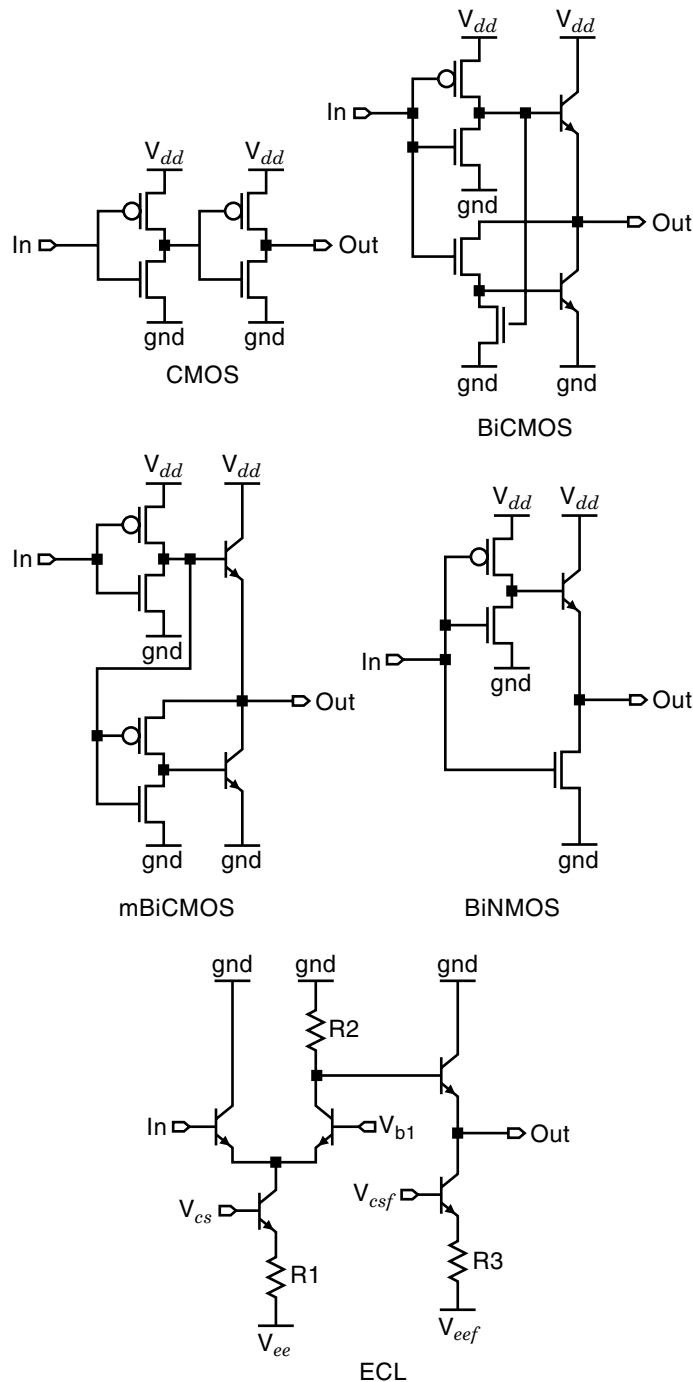


Figure 1. Buffers in five different BiCMOS circuit families. These circuit families or variations thereof make up the periphery circuits of BiCMOS memories.

ries-gated ECL,

$$V_{\text{supply}} > 4V_{\text{be}} + 2V_{\text{swing}} + V_{\text{ce,sat}} \approx 4.7 \text{ V} \quad (1)$$

for a typical voltage swing of 550 mV, and assuming $V_{\text{be}} = 0.8 \text{ V}$ and $V_{\text{ce,sat}} = 0.4 \text{ V}$. Two-level series gating with a reduced current-source resistor voltage can reduce the supply requirement to around 3.5 V or less if room/cold temperature operation is not required. Thus, with 5 V CMOS, conventional

three-level series-gated ECL and CMOS circuits can directly share supplies. However, the ECL-to-CMOS level conversion is slow because of the large voltage gain required. To share supplies in the 3.3 V CMOS generation, the ECL circuits must be limited to a modified two-level series-gated approach. Alternatively, a split supply with 5 V ECL and 3.3 V CMOS is possible, with the lower voltage either supplied externally or generated internally. Past the 3.3 V CMOS generation, mixed-supply circuits become necessary. Typically for ECL BiCMOS, the CMOS is referenced to the lower ECL supply (2, 3) because the $n-p-n$ transistor cannot pull up all the way to the upper supply; however, referencing to the upper supply also has some advantages (4).

BiCMOS PROCESS CONSIDERATIONS

Optimized CMOS BiCMOS circuits require different $n-p-n$ characteristics than ECL BiCMOS circuits, and hence different process flows. $N-p-n$ transistors in CMOS BiCMOS circuits need to deliver large current transients, which are limited by $n-p-n$ saturation from the collector resistance and beta roll-off from high-level injection. By increasing the collector doping, both of these effects are reduced at the cost of increased collector capacitance (5). In contrast, minimum power ECL BiCMOS gate delays are limited by the $n-p-n$ collector capacitance, which should be minimized for optimum ECL performance. Note that the $n-p-n$ peak, f_T , is only of secondary importance for high-density ECL circuits because of power constraints. Good ECL performance demands minimal $n-p-n$ collector-substrate capacitance (C_{js}), which results from reducing the collector area using trench isolation or other advanced isolation schemes. Small collector-base capacitance (C_{CB}) and base resistance (R_B) are also important and often require a self-aligned emitter double-polysilicon process. All of these schemes to increase ECL performance add complexity, reduce yield, and hence increase cost.

Since CMOS BiCMOS processes do not require the complex $n-p-n$ transistor formation of ECL processes, they are often simpler and can be developed as an add-on module to an existing CMOS process (6). ECL BiCMOS processes, on the other hand, are usually derived from good bipolar ECL processes, where the complicated $n-p-n$ formation has already been accomplished (7). A CMOS process that does not adversely affect the $n-p-n$ transistor performance is then incorporated. An alternative approach to ECL BiCMOS process optimization relies on aggressive scaling to provide smaller geometry $n-p-n$'s with reduced parasitics but without the complex process flows and their advanced isolation and emitter formation (8).

SRAM BASICS

Most fast SRAMs are operated in either register flow-through mode, or register-register mode (also known as fully pipelined). Because the I/O and SRAM access times are nearly comparable, fully pipelined stand-alone SRAMs have three times the throughput of nonpipelined parts with little latency penalty. Furthermore, the added system complexity in using a pipelined SRAM has become more acceptable as circuit integration, and hence functionality, has increased dramatically over time.

To read a basic SRAM, registered addresses are predecoded and combined with the registered control bits and then buffered and driven up the SRAM spine (see Fig. 2). These predecoded signals activate a single decoder that buffers up and drives a CMOS-level word line across the entire SRAM. Each activated SRAM cell then pulls current out of one of two differential bit lines, depending on the state of the cell. The bit-line differential is then sensed at the bottom of the SRAM, often both before and after multiple columns are multiplexed together. In large-capacity SRAMs, the large bit-line capacitance and small cell current combine to limit the bit-line slew rate and impact performance, since the slew rate $dv/dt = I/C$, where I is the cell current and C is the bit-line capacitance in this case. Increasing the cell current normally requires increasing the cell area and hence bit-line capacitance, thus offsetting the benefits, not to mention increasing the die size and SRAM cost.

To improve access times, the sense amplifiers (amps) must sense the smallest possible bit-line swings without impacting the SRAM reliability. Also, the SRAM should be reset once the output data have been latched, a technique termed “post-

discharge” in CMOS parlance. After the predecode outputs are disabled, the active word line falls and the bit-line voltages are restored or “equilibrated”. Finally, the sense amp is equilibrated if necessary. By resetting the SRAM, access-time critical-path signals travel in only one direction, which eliminates Miller capacitance effects. Furthermore, the designer can often speed up the critical path by tuning the gates to travel faster in one direction. Large SRAMs can also be split into multiple banks to improve performance.

To write the SRAM, the bit lines are overdriven by data buffers to reflect the data inputs, as shown in the left side of Fig. 2. The bit lines must be driven to full CMOS values to write the cell, whereas to read the cell, the bit lines only need to separate adequately for reliable sensing. Thus, bit-line equilibration after writing often limits the cycle time of the SRAM. Figure 3 illustrates how $n-p-n$ transistors can be used to speed up the write-cycle bit-line equilibration. During a write cycle, $n-p-n$ transistors Q1 and Q2 are turned off by driving the signal *write_L* low, and the bit lines are overdriven. To equilibrate the bit lines, signal *write_L* is driven high, which quickly restores the low bit line to one V_{be} below

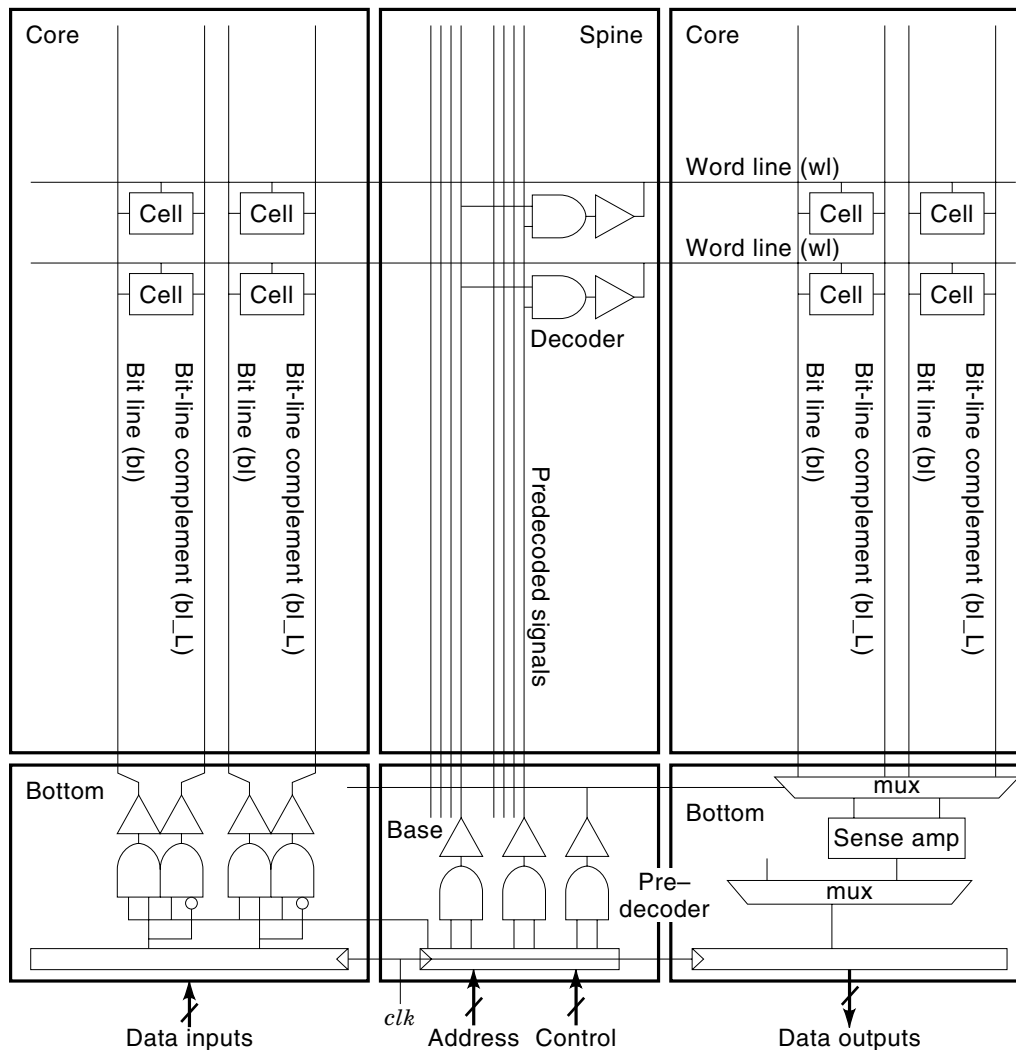


Figure 2. Basic SRAM floor plan (not to scale) with simplified write circuitry displayed on the left and read circuitry on the right, for illustration purposes only. The cells furthest from the base are in the critical speed path of the SRAM.

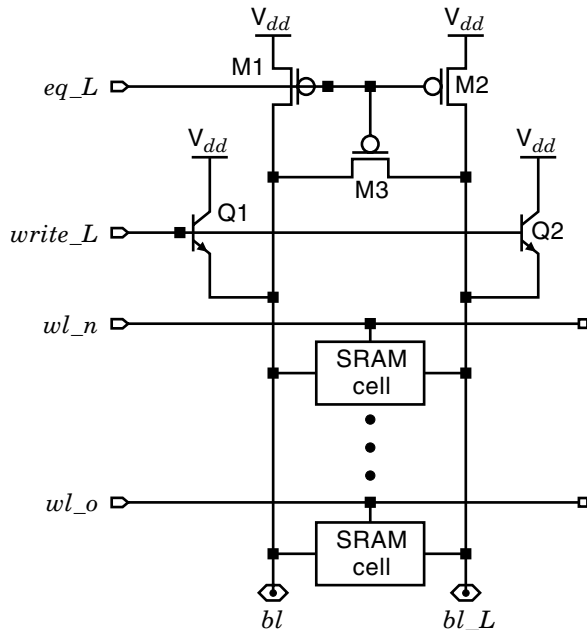


Figure 3. A bit slice of a BiCMOS SRAM illustrating the write-recovery n - p - n transistors and the equilibrate p -FET transistors. Bipolar transistors Q1 and Q2 speed up the bit-line write recovery and clamp the bit lines from excessive voltage swings during reading.

the supply. The bit lines are then equilibrated together and to the positive supply by the p -FETs M1, M2, and M3 when the signal eq_L is driven low, which is the same for both read and write cycles (9, 10). Note that Q1 and Q2 also clamp the bit lines from excessive voltage excursions during reading.

THE SRAM CELL

Most SRAMs are based on either six-transistor (6T) or four-transistor, two-resistor (4T-2R, or 4T for short) SRAM cells (see Fig. 4). 4T cells have been preferred for stand-alone SRAMs because of their smaller size, which is typically one-third to one-half the size of 6T cells in similar process generations. The 6T cell area cannot be shrunk as much as the 4T cell because a minimum p -FET-to- n -FET spacing is required to prevent latchup. The smaller size of the 4T cell, however, comes at a cost. 4T cells require extra process steps to form the high-impedance load resistor, which is normally in a second polysilicon layer. More importantly, the 4T cell stability is eroded as supply voltages scale, becoming unacceptable at supply voltages below 3.3 V.

In very large 4T SRAMs, power dissipation in the cell array can become a significant problem, due to the small but finite static current through the high-impedance resistors. To minimize this static power dissipation, the resistor size is designed only to be small enough to overcome the leakage of the n -FETs and preserve the high voltage in the cell (typically $>10\text{ G}\Omega$). When the cell is written with a full-swing word-line voltage, the n -FET access transistor pulls up the internal cell voltage to within a threshold voltage ($V_T \approx 0.8\text{ V}$) of the power supply. The high-value resistor must then restore the cell voltage to the full power supply, which typically requires many cycles to occur. Until the 4T cell can regain a full inter-

nal cell voltage, the noise margin and cell current are both significantly reduced, which impacts the read-access time. The smaller noise margins also increase the soft error rate (SER) (5). By boosting the word-line voltage to a threshold above the power supply, the cell voltage could be written directly to its full value (11). However, submicrometer processes, which could be of significant benefit, forbid boosting signals above the power supply (because of gate reliability concerns) unless they are running at a reduced supply voltage to minimize power consumption.

In contrast to the 4T cell, the word line's high voltage level is not critical in writing a 6T cell since the p -FETs quickly restore the full cell voltage after writing. A reduced word-line high voltage may even improve SRAM access times in primarily ECL BiCMOS SRAM designs because of the faster ECL-to-CMOS level conversion (2). Reducing the ratio of the pull-down to access n -FET size (i.e., the cell's β ratio) in the 6T cell can compensate for the lower word-line high voltage by increasing the cell current without eroding cell stability versus a typical full-swing word line (3).

The generic BiCMOS gate cannot be used to drive the word lines directly because of its reduced voltage swing. The diode-connected n - p - n pull-down generates a low voltage that is one V_{be} above ground, which would cause excessive leakage

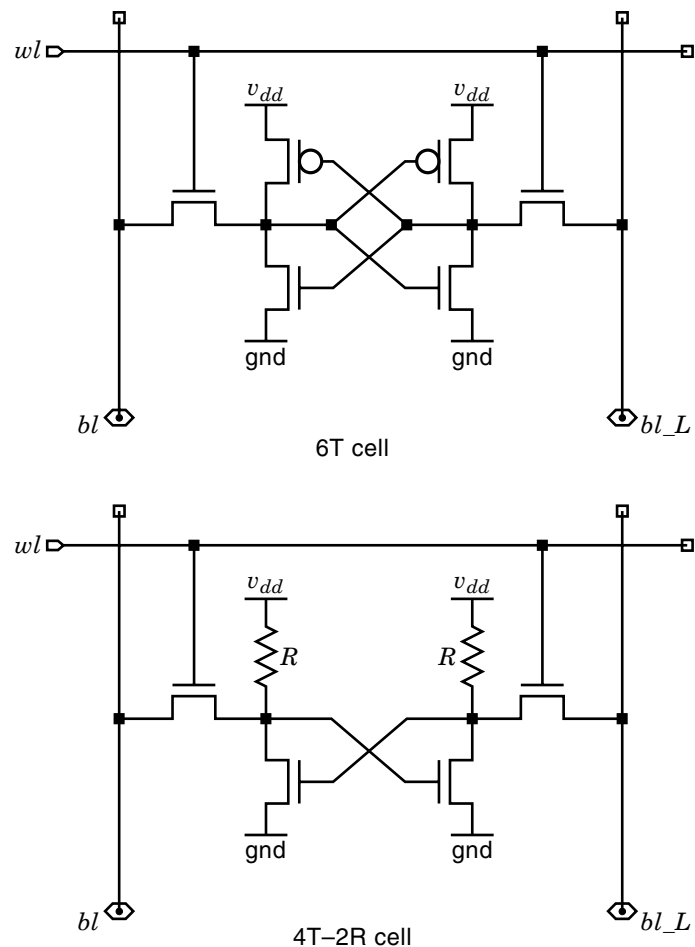


Figure 4. Two common CMOS SRAM cell schematics. While the 6-transistor (6T) cell is larger, it requires fewer process steps and is more stable, especially at lower power-supply voltages.

current in an SRAM array consisting of either 4T or 6T cells. While a BiNMOS gate can pull to ground, it cannot be used to drive a 4T cell because its reduced high-level swing would not drive a large enough voltage into the SRAM cell. By adding a parallel p -FET to the n - p - n pull-up in a BiNMOS gate (10), a full high level will be produced, albeit slowly. Alternatively, a BiCMOS gate may be followed by a CMOS inverter to restore the full voltage levels (12).

SRAM PERIPHERY CIRCUITRY

High-speed BiCMOS SRAMs require high-speed periphery circuits for the predecoder/driver, the word-line decoder/driver, and the sense amps and data lines. For maximum speed, the periphery logic should be entirely ECL instead of CMOS or CMOS BiCMOS. The ECL-to-CMOS level conversion, which is required to drive the CMOS word line, is then postponed until the last possible moment. Because of power constraints, ECL-only periphery circuitry can be prohibitive. Even when low power is not required, the large number of rows, and hence word lines, in modern, high-capacity SRAMs requires the use of a certain amount of unconventional current-sharing techniques or active pull-down circuits. Low-power or very high capacity BiCMOS SRAMs often use CMOS BiCMOS circuits for the SRAM front end. With this approach, the performance improvements from the faster n - p - n sense amps and ECL data lines on the back end will justify the added complexity of the BiCMOS process over CMOS. Note also that this latter approach has the advantage of requiring a simpler CMOS BiCMOS n - p - n transistor.

Predecoder Circuits

While address- and control-bit predecoding is not required, predecoding can save power by reducing the number of wires driven per cycle up the spine and across the bottom of an SRAM and also improve performance. Furthermore, predecoding reduces the fan-in of the word-line decoder. Two-bit predecoding takes two signals and fully decodes them onto 2^2 , or 4, wires, which is the same number of wires needed for two unpredecoded signals driven both true and complement. Three-bit predecoding takes three signals and fully decodes them onto 2^3 , or 8, wires versus 6 wires for unpredecoded. While higher-order predecoding consumes more wires per signal, only one of those wires is active at any given time. Furthermore, the higher order the predecoding, the fewer the loads each wire has to drive. This combination of effects reduces the dynamic power in the signals and improves their speed. However, unlike CMOS or CMOS BiCMOS circuits, which consume very little static power, conventional ECL circuits consume almost entirely static power. Therefore, ECL circuits without current-sharing or active pull-down circuits actually consume more power with predecoding since each wire increases the static power consumption of the circuit while lowering the dynamic power consumption.

The generic ECL NOR gate, as shown in Fig. 5(a), predecodes signals in negative logic (i.e., a low input voltage signifies that the signal is active). Alternatively, an ECL stacked NAND gate can be used (13). For minimum power designs, the RC time constant of the resistor R_{swing} and the capacitance on node V_{swing} [see Fig. 5(a)] limit the gate speed. This capacitance is primarily due to the collector capacitance of the par-

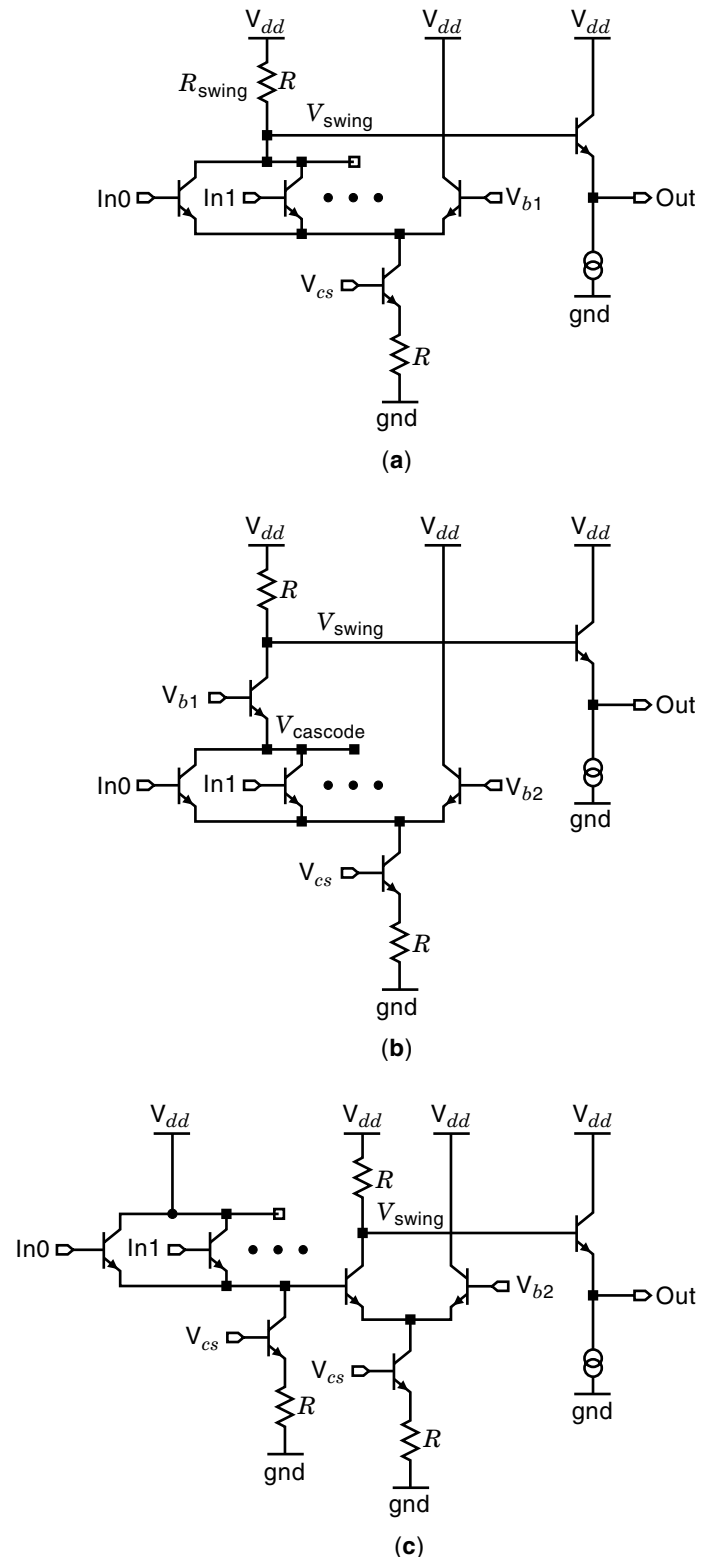


Figure 5. Three ECL NOR predecode circuits: (a) generic, (b) cascoded, (c) wire-or input. While higher power than CMOS variants, they are much faster.

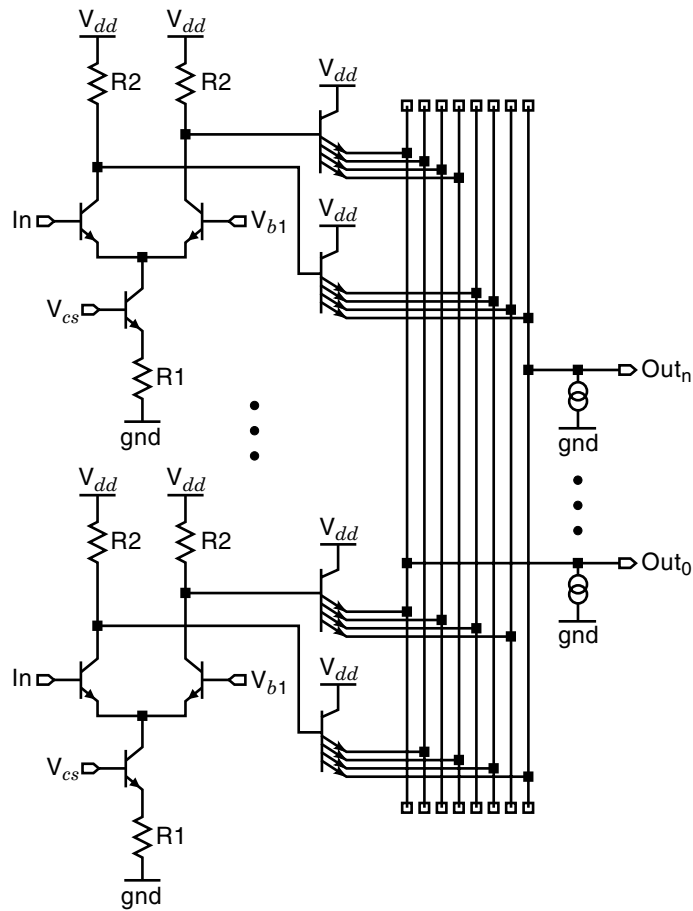


Figure 6. Wire-or predecode circuit. The wire-or eliminates one gate delay when merged into the previous logic.

allel $n-p-n$ transistors and increases linearly with the number of inputs. By cascoding the V_{swing} node, as illustrated in Fig. 5(b), the effective capacitance on node V_{swing} is reduced, which speeds up the gate. Note that the input signals must also be level shifted. Unfortunately, the cascode transistor also amplifies noise on node $V_{cascode}$, which is already very noisy because of the C_{CB} coupling of the many inputs that can transition simultaneously. A better circuit choice is shown in Fig. 5(c), in which the inputs are first wire-or'ed together. This circuit has the performance of the cascoded circuit without the noise sensitivity.

An alternative to the ECL NOR predecoder is the wire-or predecoder (9,14), which is shown in Fig. 6. The single-ended to differential buffer on the input can be merged into the previous logic such as the address or control register. Therefore, the only delay from the predecoder comes from the wire-or, which is very fast. However, because the wire-or output is one-low, unlike the ECL NOR decoder, the output currents cannot be shared (see later in this article).

High-Speed Low-Power Buffers. Finding an active pull-down emitter follower is the goal of much research in ECL circuit design. Instead of having a fixed pull-down current source, an active pull-down circuit switches the current on only when needed. Such a technique significantly lowers the emitter-follower static power dissipation while potentially improving

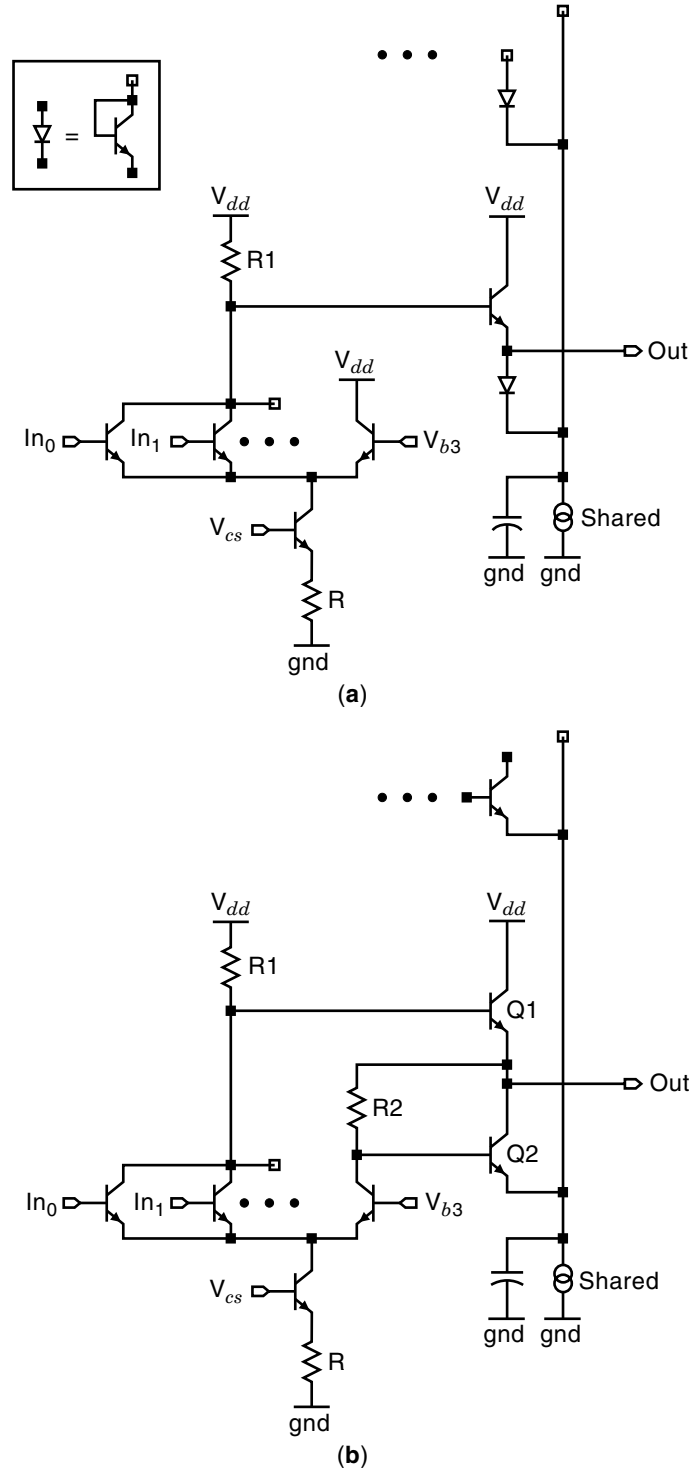


Figure 7. Two ECL NOR gates with emitter-follower current sharing: (a) diode, (b) active (3). The current sharing permits each decoder to be powered up for maximum speed without exceeding the overall memory's power budget. Reprinted with permission from "A Subnanosecond 64kb BiCMOS SRAM," Santoro, Tavrow, and Bewick, *Proc. BCTM*. © 1994 IEEE.

performance. Despite significant efforts by many groups (15), true active pull-down circuits remain fickle. They are either process intolerant or require special devices or supply voltages. Furthermore, they nullify one of the advantages of current-steering logic: low noise.

Fortunately for SRAMs, most highly loaded signals are fully decoded and hence the designer can share one pull-down current among many drivers (e.g., eight drivers can share the same pull-down current on a three-bit predecoded signal). In contrast, in an 8-bit data bus, there is no interrelationship between the bits and hence only a true active pull-down could be used. Figure 7(a) illustrates a diode-based current-sharing predecoder circuit. In the case of a three-bit predecoder, each gate has three inputs, and eight gates share the same emitter-follower current source. At any one time, only one of the outputs is high and the shared pull-down is steered through the diode associated with that output, giving it a fast pull-down. One disadvantage of this circuit is that the falling output voltage has a long tail, which results from the diode sharing back the current with the other parallel diodes. This effect becomes severe if another output simultaneously transitions high. Note also that by reducing the pull-down current in the nonactive gates, their outputs will rise slightly, which reduces the gates' static noise margins.

A current-sharing circuit that overcomes the limitations of the previous circuit is shown in Fig. 7(b) (3). In this circuit, the diode is replaced with an actively switched $n-p-n$ transistor. When the gate is deselected, the gate current is steered away from resistor R2 into resistor R1. Since no current flows through R2, there is no voltage drop across R2, and $n-p-n$ Q2 is essentially diode connected, much like in the previous circuit. When the gate becomes active, the gate current is steered away from R1 and into R2. Depending on the resistance ratio of R1 to R2, the designer can choose how much Q2 should be on (if at all) when the gate is active. Note that even if R2 is larger than R1, and hence transistor Q2 is fully off, the gate current still provides a pull-down current to Q1 so that it does not float up. When the gate is deactivated, the voltage across R2 collapses and the diode-connected $n-p-n$ Q2 steals the entire shared emitter-follower current-source current. Because Q2 is only active during the pull-down transient, the question of whether another gate is simultaneously switching high is of no consequence. Furthermore, the voltage tail can be eliminated with an f_T doubler (see later), which has a slower turn-off transient (3).

Note that the capacitance on the emitter-follower shared pull-down current source actually helps the current-sharing circuit in Fig. 7(b). When $n-p-n$ Q2 steals this current, it also receives the current required to charge up this capacitance.

While a minimum amount of capacitance on this wire is unavoidable, because the wire is distributed across many gates, an additional deliberately placed capacitance can be added. To a certain extent, a larger capacitance can be played off against a smaller current source to reduce power while maintaining the same level of performance.

Increasing Bipolar Drive Capability with Darlington. SRAMs require large buffering internally, especially following the address predecoder and the word-line decoder. Furthermore, to reduce power as much as possible, the predecoder and decoder gates should have a near-minimum current. Even though typical $n-p-n$ transistors have current gains, β , of around 100, the overall emitter-follower current gain over the current in the ECL current-steering gate is considerably less. Consider a 4:1 emitter-follower-current to gate-current ratio and a 550 mV voltage swing; the high static noise margin is already reduced by 24 mV from the emitter-follower in this configuration for a β of 100. Darlington configurations provide larger current gains and hence the possibility of faster buffering with less power consumption.

Some typical Darlington configurations are shown in Fig. 8 and, with the exception of the f_T doublers, behave as a single $n-p-n$ with a current gain of β^2 . Note that they also have a forward voltage drop of $2V_{be}$. The basic Darlington in Fig. 8(a) has no means of discharging the second $n-p-n$'s base and hence is not practical. Figure 8(b) and (c) provide two alternatives, the second of which is preferable because it requires no additional current source. In Fig. 8(d), the so-called f_T doubler, the diode provides the discharge path but can also be augmented with a current source to improve the turn-off speed, as shown in Fig. 8(e). Because the diode is an active device, it also affects the overall current gain. If the diode D2, which is a diode-connected $n-p-n$ transistor, is the same size as the second $n-p-n$ Q2, then the current gain is approximately 2β , since the same current must flow through D2 as through Q2.

A more severe problem than discharging the base is overshoot and ringing, which occurs to some degree in all emitter followers and is particularly severe in Darlington configurations because of their larger current gain. There are many physical explanations for overshoot in emitter followers, including the inductor-like behavior of the emitter resistor and the delay from discharging C_{CB} during $n-p-n$ turn-off. Alternatively, consider that β rolls off at both high frequency and high current. During a pull-up transient, the effective current gain is reduced from the ideal static value. When the emitter-follower output approaches its final output voltage, the pull-up current is quickly reduced and the high-frequency components of the

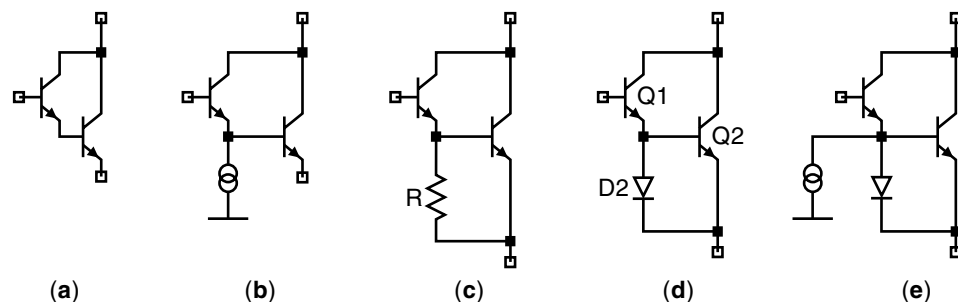


Figure 8. Darlington configurations: (a) basic, (b) current source, (c) resistor, (d) diode, or f_T doubler, and (e) f_T doubler and current source. These $n-p-n$ variations increase current drive (β) at a cost of $2V_{be}$ voltage drops and greater overshoot tendencies.

turn-on transient die off. Therefore, the $n-p-n$ current gain increases, returning to its nominal value, which injects more current into the output node and causes the circuit to overshoot.

Consider how the f_T doubler compares to other Darlington configurations for stability. The effective current gain of the f_T doubler is

$$\beta_{\text{eff}} = \beta \times \frac{A_{D2} + A_{Q2}}{A_{D2} + A_{Q2}/\beta} \approx \beta \times \left(1 + \frac{A_{Q2}}{A_{D2}}\right) \quad (2)$$

where A_{Q2} and A_{D2} are the emitter areas of $n-p-n$ Q2 and diode D2, respectively. Assuming that β is around 100 and that the emitter area of Q2 is less than 10 times D2, then Eq. (2) is linearly proportional to β . Consequently, the f_T doubler does not increase the overshoot problem over a simple emitter follower. While the current gain of the f_T doubler is less than for a Darlington, it can be set to a known and process-independent value, i.e., the ratio of layout areas, which reduces circuit variations due to changes in process and operational conditions.

Word-Line Decoder/Driver

Because of the large number of rows in a modern SRAM, the word-line decoder must be fast yet low-power. For a BiCMOS

SRAM with an ECL front end, the word-line decoder integrates an ECL-CMOS converter and a high-current-gain driver. CMOS BiCMOS front ends are inherently simpler (e.g., they contain no ECL-CMOS converter) and are directly based on a given BiCMOS circuit family, such as the ones shown in Fig. 1 (12,9). Since the bit lines must be driven to CMOS levels to write the SRAM, the write decoder/driver circuitry uses the same circuit techniques as the word-line decoder/driver.

Diode Decoder. An ECL NOR decoder can be used for both the predecoder and word-line decoder. An alternative approach replaces the ECL NOR decoder with a diode decoder (16,13), because it eliminates one gate delay from the access path. A diode decoder that uses an ECL NOR predecoder is illustrated in Fig. 9. Each diode decoder is attached to a unique combination of the one-high predecoder outputs. If all of these signals are high, then all of the diodes are off and the resistor R3 pulls up the internal node V_{int} to activate the gate. As with the simple NOR predecoder, the diode-decoder performance is limited by the RC time constant of resistor R3 and the collector capacitances of the diode-connected $n-p-n$ transistors. Since at most one decoder can be active at any one time, enough current must be sunk by the predecoder emitter-follower current sources to pull down all of the decoder resis-

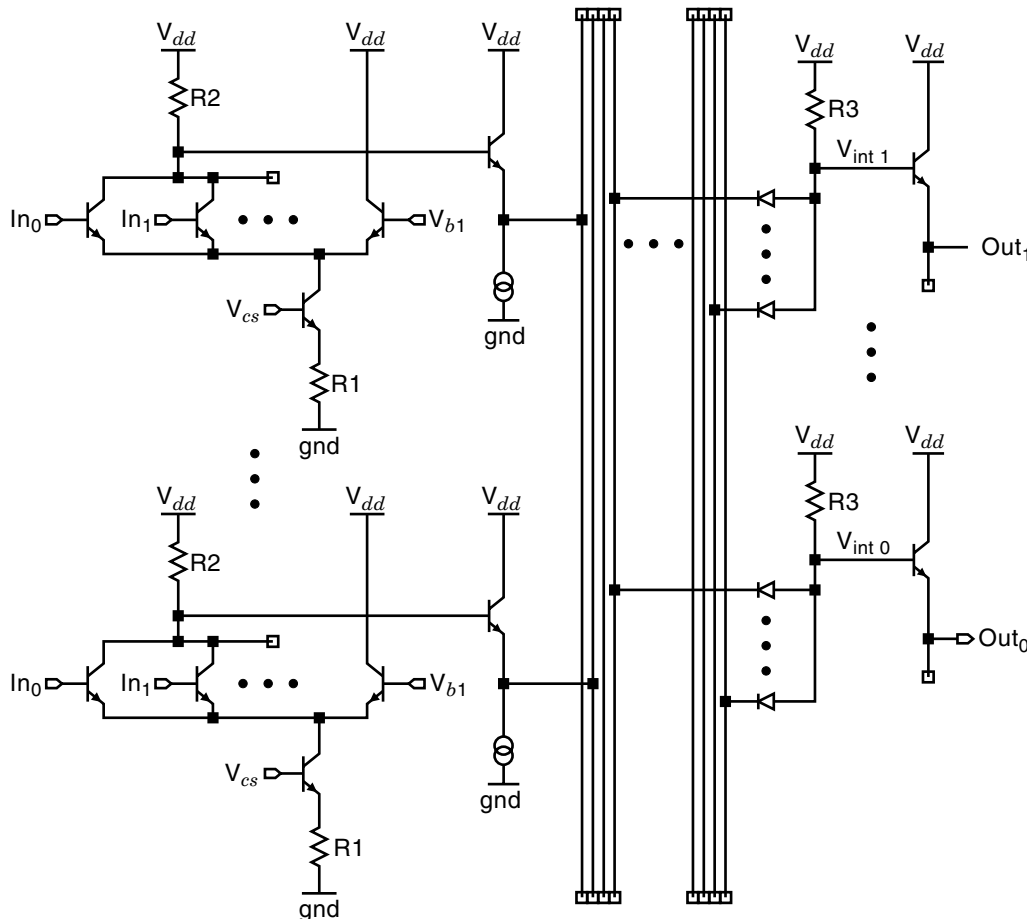


Figure 9. A diode-decoder circuit with ECL NOR predecoders. Only the output stage that is connected to all high inputs will be high.

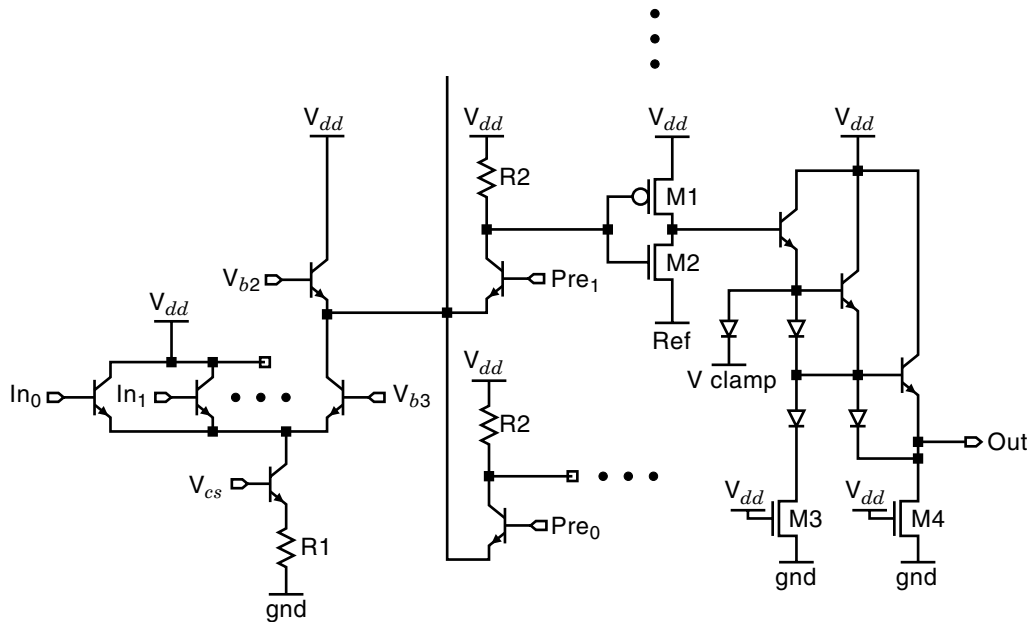


Figure 10. A word-line driver circuit with a current-sharing ECL OR decoder, CMOS inverter, and triple-Darlington driver (3). This circuit decodes, ECL-CMOS converts, and buffers in only two gate delays. Reprinted with permission from “A Subnanosecond 64 kb BiCMOS SRAM,” Santoro, Tavrow, and Bewick, *Proc. BCTM*. © 1994 IEEE.

tances. Therefore, current sharing is not possible and resistor R3 must be increased to reduce power, which also directly decreases performance. Alternatively, resistor R3 can be dynamically varied (e.g., with a p -FET) to reduce power in inactive banks, without impacting performance (13).

When the predecoder emitter followers are low, most of the pull-down current is sourced by the diodes in the diode decoder instead of the emitter-follower n - p - n , which erodes the lower noise margin of the diode decoder. To counteract this problem, resistor R2 must be made larger than R3. In addition, by increasing the ratio of resistors R2 and R3 to R1, the output voltage swing can be increased (up to the point that the predecoder input n - p - n transistors become saturated), which helps in the succeeding ECL-CMOS conversion. However, in both cases, the predecoder gate slows down unless the power can also be increased, or dynamically switched.

NOR Decoder. An ECL NOR decoder with current sharing and an integrated ECL-CMOS converter is illustrated in Fig. 10 (3). By inverting the output signal with a CMOS inverter (FETs M1 and M2), the ECL NOR gate can be replaced by an OR gate. Because only the selected OR gate requires current to pull the output low, many OR gates can share the same current source (e.g., 8 or 16). In the circuit in Fig. 10, two levels of decoding are used: the lower level selects whether any output in the group is active and the one-high predecoded inputs select the single active output. In addition, the resistor ratio of R2 to R1 is chosen to be 4, which increases the ECL voltage swing enough to directly drive the CMOS inverter. Note that the widths of M1 and M2 are adjusted to skew the inverter trip point toward V_{dd} . The larger R2 resistor, and hence larger RC time constant on the OR output, can be offset by increasing the gate current. The larger gate current also offsets the additional delay of the wire and emitter capaci-

tance of sharing more outputs; therefore, the voltage gain does not directly slow down the circuit or increase its power. In contrast, current sharing is not possible in a conventional ECL NOR decoder, and the voltage gain on the output further reduces the performance (2).

The circuit in Fig. 10 also uses a triple Darlington with an overshoot clamp and two f_T doublers to buffer up the CMOS inverter output current and drive a reduced-level word-line voltage. Note that the clamp voltage is shared among all of the word-line drivers; thus the clamp voltage generation consumes very little overall power. Instead of attempting to switch off the pull-down n -FETs M3 and M4, they are left on and they consume static power when the word line is high. However, since only one word line can be high at any given time, this current is minimal. In addition to the large current gain of the triple Darlington, the three V_{be} drops partially temperature-compensate the CMOS cells. As the temperature increases, the V_{be} voltage drops at the rate of 1.5 mV/°C (17). Therefore, at higher temperatures, the word-line swing is increased, which compensates for the CMOS cell current reduction due to the higher temperature.

ECL-CMOS Converters. A variety of stand-alone static ECL-CMOS converters is displayed in Fig. 11. The two circuit variations in Fig. 11(a) and (b) rely on an n -FET current mirror to provide an opposing pull-down current to the p -FET input pull-up. The current path through the current mirror is much slower than the direct p -FET pull-up, and thus these circuits produce asymmetric rise and fall times. The ECL-CMOS converter in Fig. 11(c), though drawn differently, is similar to the previous two. The ECL gate, which consists of the differential pair and emitter-follower n - p - n transistors, would normally be merged into the preceding gate, leaving two p -FETs and an n -FET current mirror. In this circuit, the

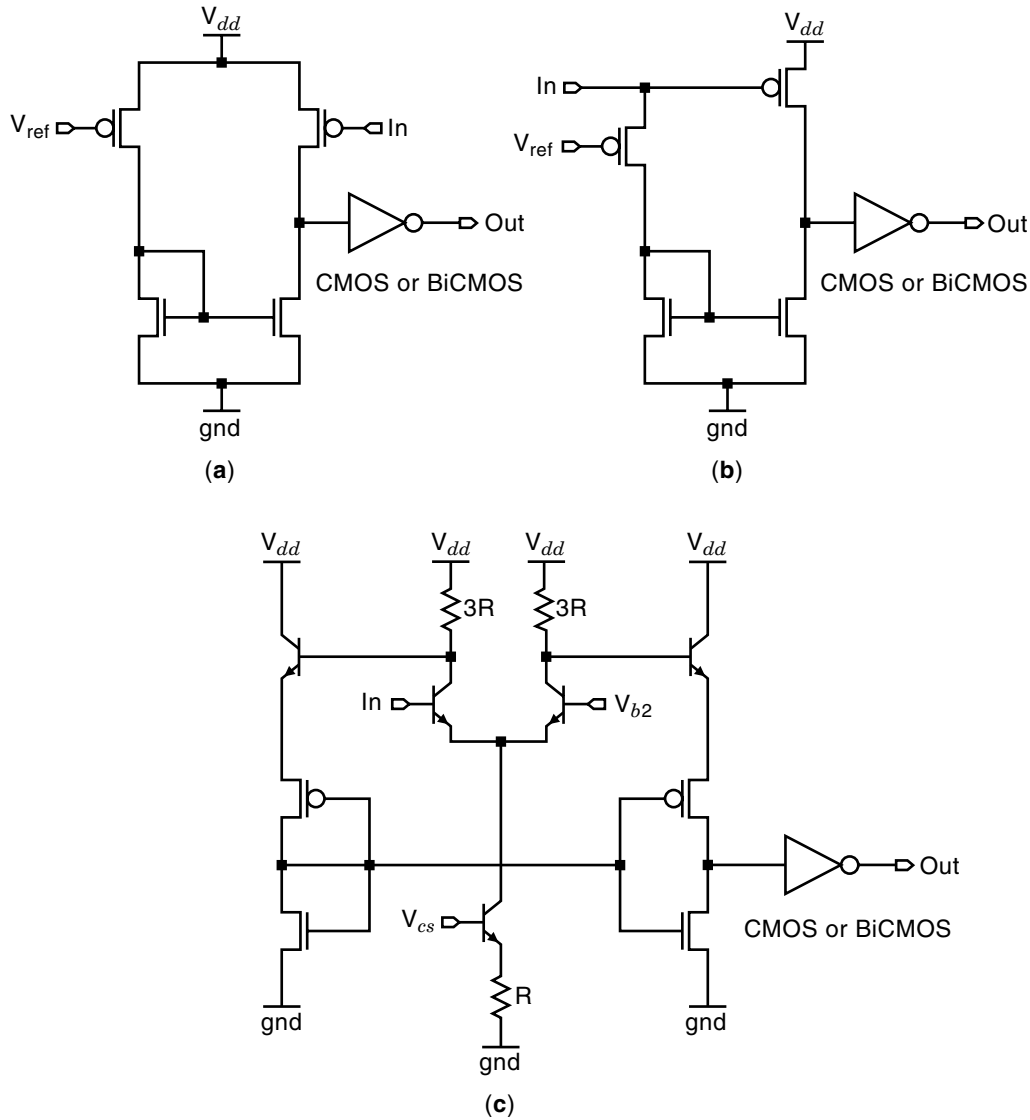


Figure 11. Three ECL-to-CMOS level converter circuits. Sense amp (c) is very fast but requires mixed voltages and consumes static power.

inverter on the left biases the inverter on the right to the high gain region of its transfer function; therefore, small voltage changes on the p -FET source terminal induce large output voltage excursions. Note that this circuit also has unequal rise and fall times.

Sense Amps and Data Lines

Bipolar transistors make nearly ideal sense amps because of their low mismatch (typically much less than 1 mV with proper layout) and exponential current–voltage relationship. For example, a differential pair with n - p - n transistors and a 10 mV differential input voltage steers 60% of the current away from the low side, and 88% at 50 mV. In contrast, a CMOS differential pair with a 10 mV differential steers only 51% of the current away from the low side, and 55% at 50 mV, assuming perfect matching. In addition, the CMOS transistors are likely to have between one and two orders of magnitude higher threshold-voltage mismatch than an n - p - n

transistor. Both of these factors contribute directly to sense amp speed. Note that because of the relatively poor performance of CMOS differential pairs, fast CMOS designs rely instead on positive-feedback clocked sense amps, which require special critical timing pulses for both equilibration and sensing.

Differential-Pair Sense Amp. Typical n - p - n sense amps are simply current-mode logic (CML) gates, as shown in Fig. 12(a) (i.e., ECL gates without emitter followers). Because of the large number of sense amps required in modern SRAMs and because CML consumes static power, having one sense amp per SRAM column can be prohibitive. In addition, because of the large footprint of the n - p - n transistor, especially in non-trench-isolated technologies, the column pitch may be too tight to lay out one sense amp per column. Two techniques have been developed to solve these problems: (1) powering the sense amps down after sensing by turning off the current

source and (2) multiplexing many sets of bit lines together to reduce the number of sense amps needed.

Figure 12(b) displays a CML gate modified to have a switchable current source. Because the simple n -FET current source cannot provide an accurate current that is stable over process, voltage, and temperature, diode clamps have been inserted on the outputs. Otherwise, under certain conditions the voltage swings can become large enough to saturate the n - p - n transistors. The output impedance of an n -FET current source, however, is worse than for an n - p - n /resistor current source. On the positive side, the n -FET current source can function at lower voltages than the n - p - n current source. Note that the current source control voltage requires a special timing signal and additional circuit complexity.

Multiplexing. Since most SRAMs have more columns than data outputs, data multiplexing is required. Instead of adding an extra gate with its associated gate delay, the multiplexor can be merged into the sense amp with nearly no additional delay. Furthermore, most multiplexor schemes actually reduce overall power consumption. Figure 13(a) shows an ECL mux/latch that combines the functions of sensing, multiplexing, and latching the bit-line data all in one gate with a single gate current. The sense inputs are one-high predecoded and select the active differential pair. This approach requires three-level series gating, however, by moving the clock input n - p - n to the second level and by swinging a single-ended clock higher than the sense inputs, a two-level realization is also possible.

In Fig. 13(b), an alternative muxing scheme is illustrated that uses an n -FET current source to select the appropriate differential pair. Furthermore, the differential pairs can be distributed, even among different banks, in which case they are called data lines. These low-swing differential data lines are typically faster and consume less power than full-swing CMOS. With the cascode n - p - n transistor as shown in Fig. 13(b), these circuits can be very fast, even with the potentially large capacitance on the distributed data lines (14,12,18). However, careful layout is required to make such a delicate circuit robust against noise.

A third approach, shown in Fig. 13(c), actually reduces the number of differential pairs and thus eases the problem of pitch-matching the sense amp to the SRAM cell column (19). Since the bit lines are typically precharged high, the CMOS transmission-gate n -FET is not required. Note that the $sense_L$ inputs are one-low predecoded signals in this case. This technique can be readily modified to have an n -FET current source and/or distributed data lines as in the circuit in Fig. 13(b).

Current Sense Amps. An alternative to the familiar voltage-based differential-pair sense amp is a fully-differential current sense amp, such as the one shown in Fig. 14 (20). The cross-coupled diodes in this circuit produce differential sensitivity similar to a differential pair and improve the common-mode rejection significantly over a simple cascoded sense amp (i.e., a sense amp similar to the one in Fig. 14, but without the cross-coupled diodes). When both of the bit lines fall together, for example, the diodes pull down the n - p - n bases in equal measure, leaving the outputs nearly unchanged. Differential bit-line signals, however, are amplified by the diodes since the lower bit line will tend to turn off the opposite n - p - n transistor through the cross-coupled diode. This sense amp also clamps the bit lines against large voltage swings. On the negative side, power supply spikes can disable this sense amp until the bit lines recover. For example, if V_{dd} falls during sensing, a common occurrence because of the large current sourced by the front-end periphery circuitry, the bit lines must be pulled down an equal measure before sensing. This requires relatively large pull-down currents in the current sources.

ALTERNATIVE APPROACHES

In addition to speeding up the SRAM periphery circuitry, the basic SRAM core performance must also be improved to achieve the fastest BiCMOS SRAM performance. Driving the word lines and waiting for the bit lines to slew adequately for reliable sensing adds significantly to the access times of most

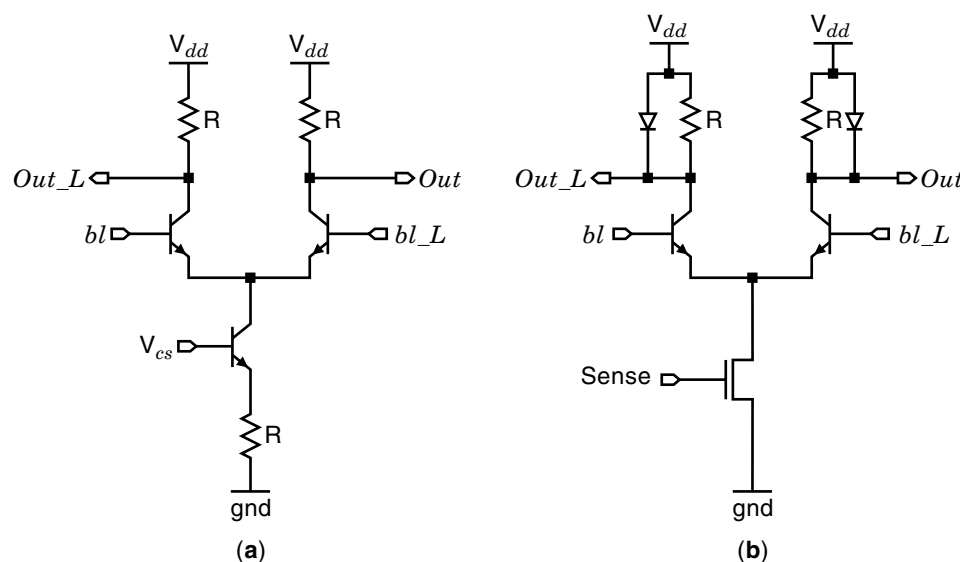


Figure 12. Two differential-pair sense amps: (a) n - p - n /resistor current source, (b) n -FET current source. The n -FET current source can be dynamically switched to save power when the sense amp is idle.

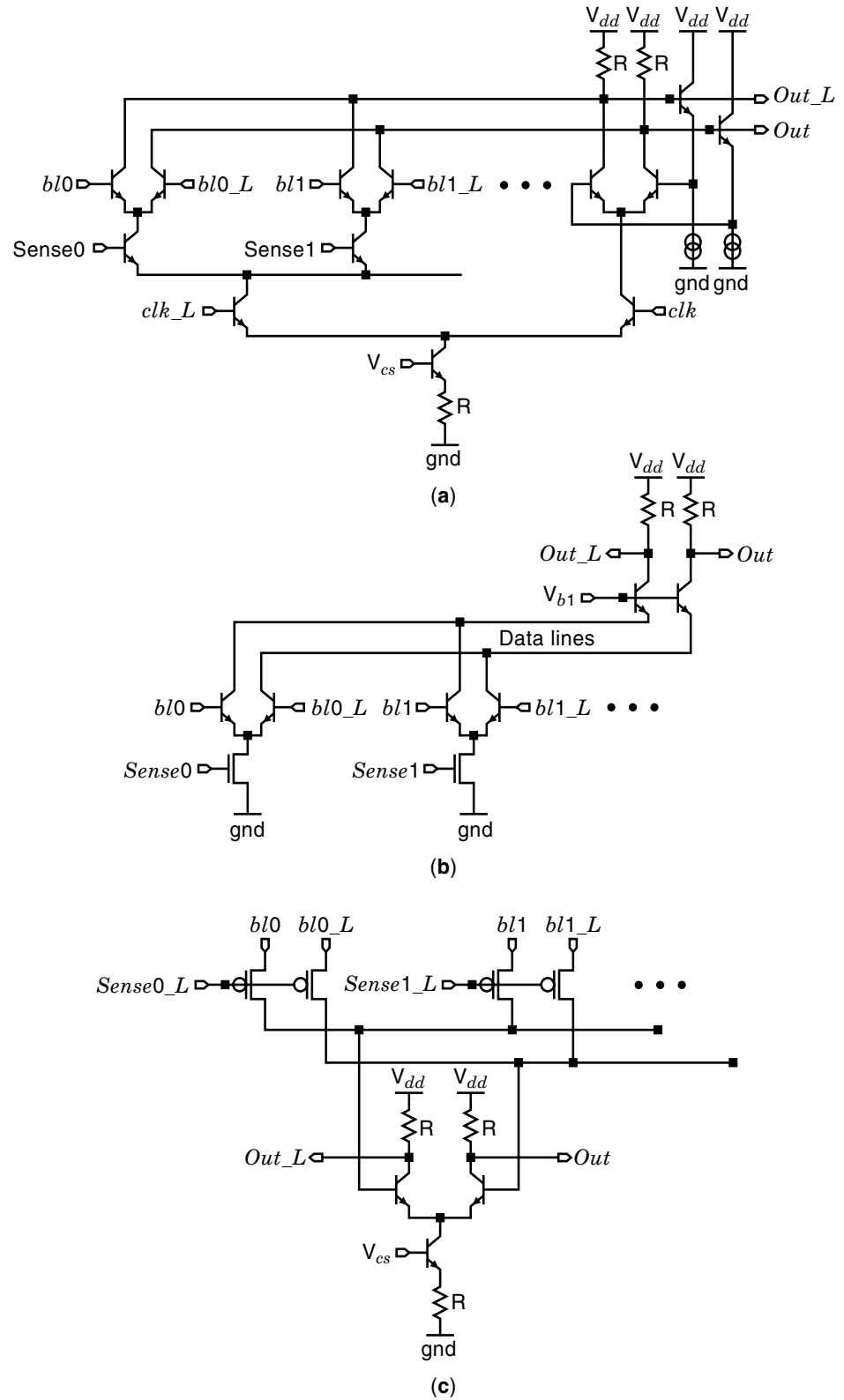


Figure 13. Three sense amps with integrated multiplexors: (a) three-level series-gated ECL mux/latch, (b) distributed differential pairs, (c) *p*-FET input mux. These sense amps reduce power, increase performance, and simplify layout.

SRAMs. With increasingly higher capacity SRAMs, even maintaining access times becomes a challenge as word-line and bit-line lengths, and thus capacitances, increase. At some point, the performance increase that results from splitting the design in half outweighs the delay associated with multi-

plexing and driving the address and data lines to and from different banks. However, “banking” only provides a modest performance increase and usually costs significantly in area overhead. Alternative approaches are thus sorely needed to increase SRAM performance.

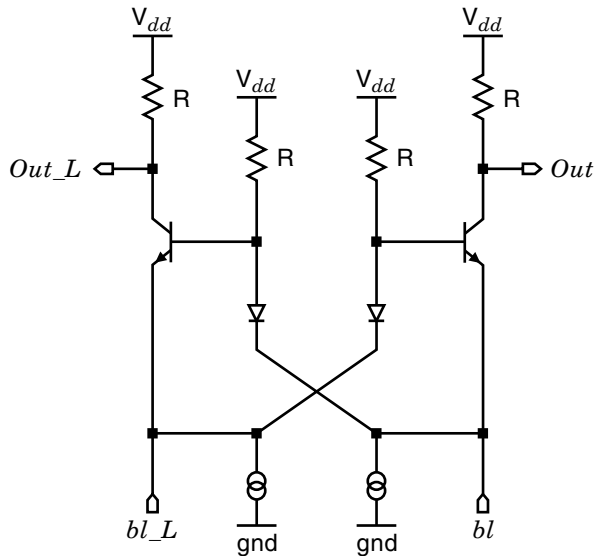


Figure 14. A current-sense amp with cross-coupled diodes to improve common-mode rejection (20). Current-sense amps theoretically can sense faster since they require smaller voltage differentials than voltage-sense amps. Reprinted with permission from “A 3.5 ns, 1 W, ECL register file,” Horowitz et al., *ISSCC Tech. Dig.* © 1990 IEEE.

The CSEA Cell

The CMOS-storage emitter access (CSEA) cell (13) (see Fig. 15) provides two advantages over a conventional all-CMOS cell: (1) an ECL-level word line and (2) an increased cell current. The CSEA cell consists of two cross-coupled CMOS inverters (M1–M4); a single n -FET for writing (M5), using a separate write path; and an n - p - n transistor (Q1) that is wire-or’ed together with the n - p - n transistors in all of the cells that share the same read bit line. To access the CSEA cell, the read word line is raised, which raises the base of Q1 only if p -FET M2 is on. Because of the n - p - n transistor in the cell, a large current is sourced onto the bit lines using only an ECL-level read-word-line swing, and thus the time-consuming ECL–CMOS conversion on the word line can be

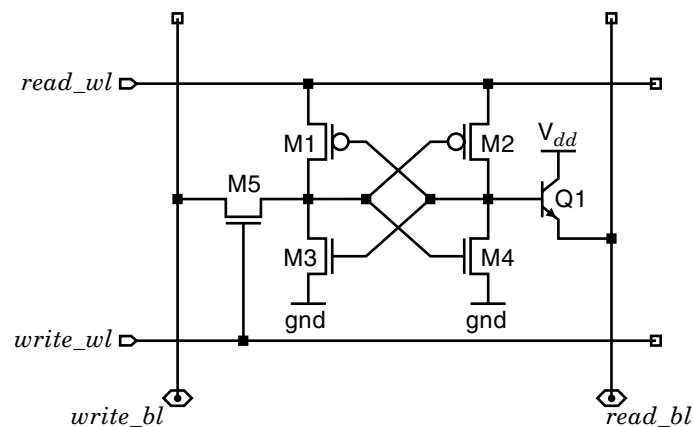


Figure 15. A CMOS-storage emitter access (CSEA) SRAM cell. The bipolar transistor in the cell converts a smaller word-line voltage swing into a larger bit-line current, at the cost of a larger cell and a single-ended sense amp.

avoided, while at the same time improving the bit-line slew rate. In addition, because the read and write paths are completely separate, a CSEA SRAM can be simultaneously read from and written to, and thus functions as a two-port SRAM.

On the negative side, the CSEA cell is much larger than a 6T cell (typically twice the size) because of the relatively large n - p - n transistor. Furthermore, n - p - n transistors typically have lower yields than CMOS transistors because of the complicated emitter formation, which further increases the SRAM cost. The larger cell current is offset by the single-ended bit-line swing, which is much more noise-sensitive than the conventional 4T or 6T cell’s differential bit lines. Even when a second n - p - n transistor is added to the CSEA cell, which greatly increases its size, the common-mode rejection on the bit lines is inferior to a conventional CMOS cell because word-line noise (even for deselected word lines) is coupled onto only one bit line, depending on each individual cell’s state. Writing a “1” into the CSEA cell is also difficult because of the single write transistor M5, unlike in a conventional

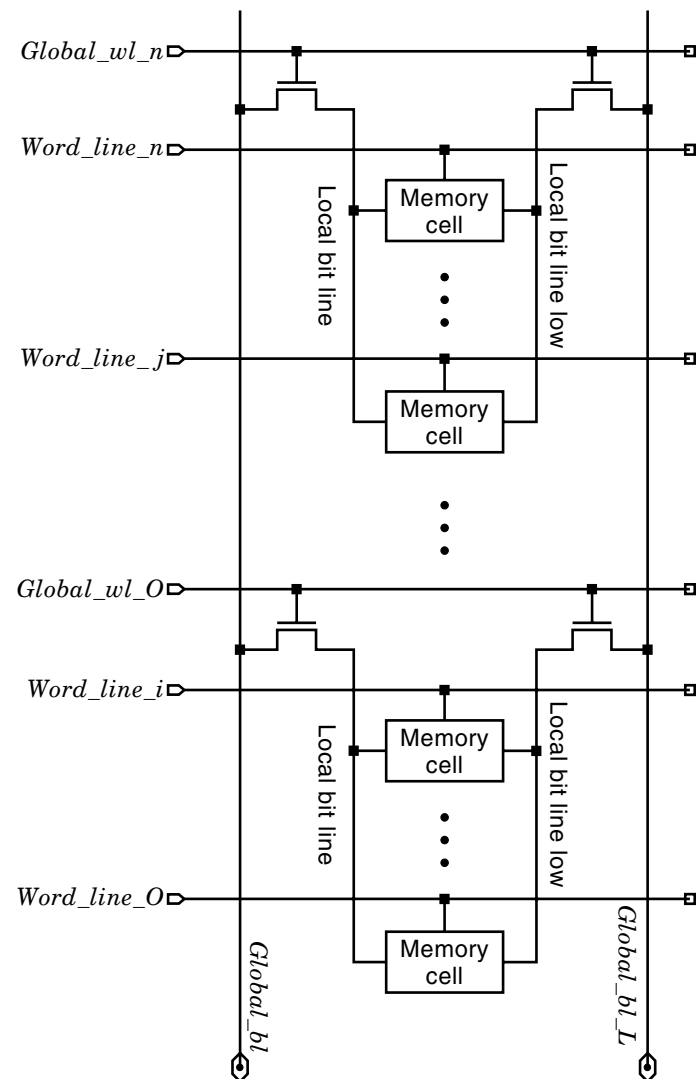


Figure 16. An SRAM bit slice with a divided bit line. The n -FET pass-gate connects or isolates the local bit lines to or from the global bit lines, which reduces the effective bit-line capacitance and improves the bit-line slew rate.

CMOS cell with two access/write transistors. While transistors M1–M5 can be sized to facilitate writing over all corners, adding another write bit line and n -FET write transistor is probably desirable. Finally, the CSEA cell has a reliability problem because of the large negative V_{be} voltage on the n - p - n when the cell contains a “0.” This problem can be partially alleviated by lowering the quiescent voltage of the read word line; however, the cell stability will suffer. All in all, the CSEA cell can be used very effectively for small, fast multiport SRAMs, such as register files.

Embedded Access Trees

Embedded access trees (EAT) improve SRAM performance by reducing the bit-line capacitance attached to a given SRAM cell (3). This is accomplished by embedding a tree structure into the SRAM read access and write paths, in which only the branch of the tree that is selected drives the next higher level branch through a switch. The simplest implementation of this technique is shown in Fig. 16 (21), and is called a divided or hierarchical bit line. If the switch has current gain (i.e., an

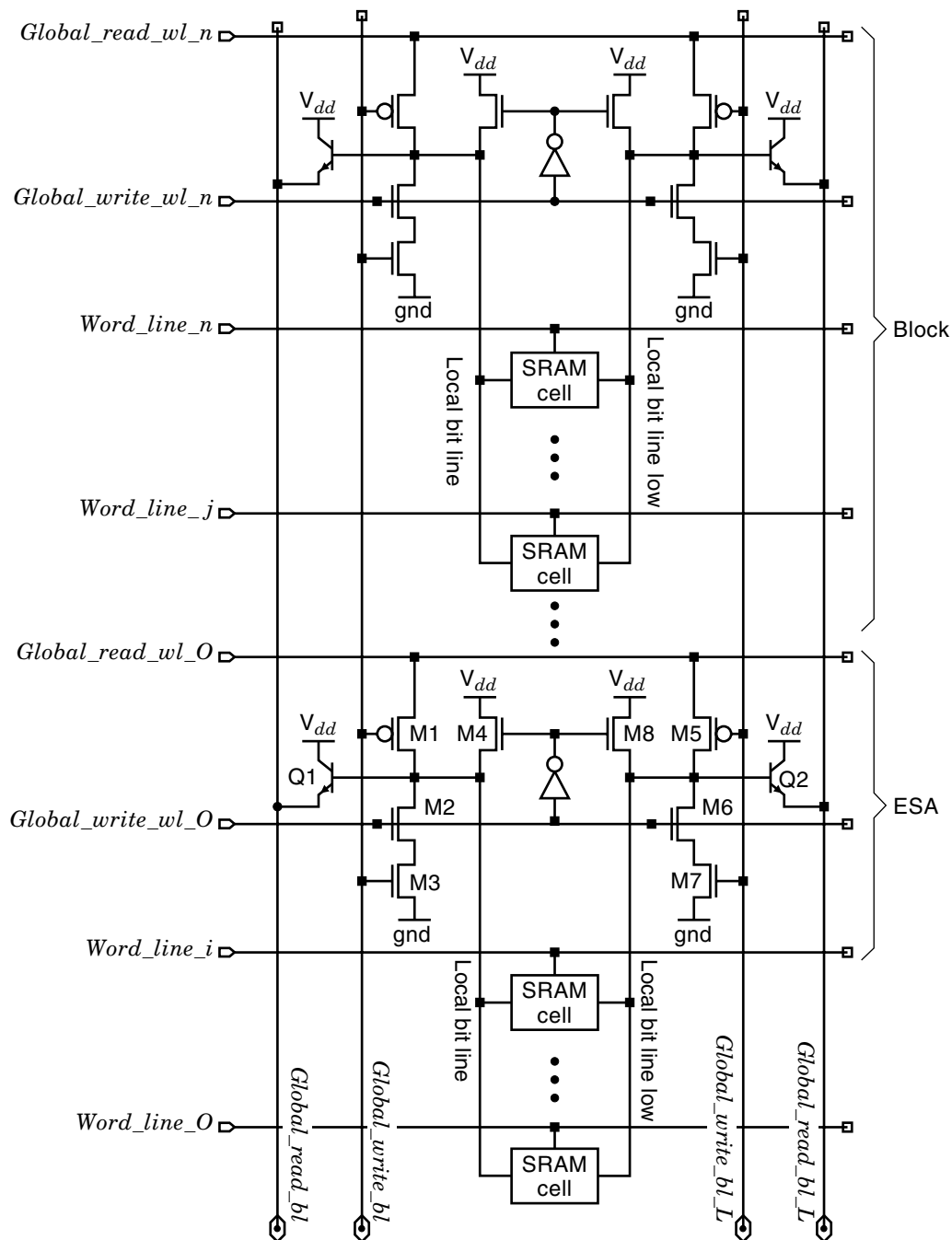


Figure 17. An embedded access tree (EAT) SRAM bit slice that details the embedded sense amp (ESA). The current gain in the ESA reduces the effective bit-line capacitance to be only the capacitance associated with the local bit lines.

embedded sense amp, or ESA), then a given memory cell only drives the loads from the small number of cells on its local branch and the switch load (3). A two-level tree can be readily built with one additional metal layer and the minimal added overhead of the ESA. Note that extra metal routing layers are typically available in embedded SRAM processes, but must normally be added to stand-alone SRAM processes.

A bit-slice of an EAT SRAM, including the ESA circuitry, is shown in Fig. 17. All of the global bit lines run in the additional metal layer, which is typically Metal-3. To access a particular cell, its word line and the associated global read word line are raised. The global write word line and global write bit lines are low; therefore, p -FETs M1 and M5 are on and n -FETs M2–M4 and M6–M8 are off. Depending on the active cell's state, the current from one of the local bit lines creates a voltage drop across either M1 or M5. This voltage differential is then driven onto the global read bit lines by n - p - n transistors Q1 and Q2, which effectively increases (and inverts) the cell current. Note that the global read bit-line capacitance is also less than the bit-line capacitance in a conventional SRAM since it contains only a relatively small number of ESAs (e.g., 8 or 16) as compared to many hundreds of SRAM cells (or more). The global read bit lines are then sensed at the bottom of the SRAM using a conventional BiCMOS sense amp. The global read word line's ECL-level swing is translated into a true differential global read bit-line voltage, unlike in the case of the CSEA, because the bases of both Q1 and Q2 are raised above the bases of all of the other n - p - n transistors that are also wire-or'ed onto the global read bit lines. This approach also uses many fewer n - p - n transistors than a CSEA SRAM. Note that the timing relationship between the word line and global word line is not critical. Also, by adding more global word lines and muxing together adjacent ESAs onto the same global bit line, the number of global bit lines can be reduced to one per cell (3).

To write the EAT SRAM, a single word line and its associated global write word line are raised in addition to one of the global write bit lines, depending on the data to be written. Either n -FETs M2–M3 or n -FETs M6–M7 then pull down one of the local bit lines, which writes the cell. Once the global word line is lowered, either M4 or M8 quickly restores the low local-bit-line voltage to a threshold below the power supply, much like in Fig. 3, but using an n -FET instead of an n - p - n . Once the write bit lines are driven low, p -FETs M1 and M5 fully equilibrate the local bit lines. Because both the local and global write bit lines have lower capacitances than in a conventional SRAM, the write speed is also improved in the EAT SRAM.

BiCMOS CAMS

Content-addressable memories (CAMs) retrieve the address that matches a given data word, unlike RAMs, which retrieve the data word at a given address. CAMs are often used in translation-lookaside buffers (TLBs) in microprocessors and routing lookup tables in network routers. A typical CAM might need to match on a 32-bit or 64-bit word, which requires many large, and thus slow, comparators. By incorporating an n - p - n transistor into an otherwise CMOS CAM cell, a very fast match line can be realized using a low-swing wire-or function (22). BiCMOS circuits can also improve the

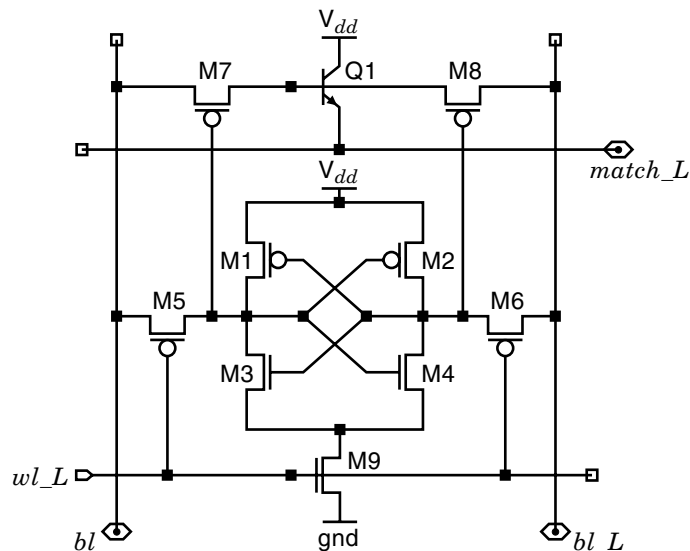


Figure 18. A BiCMOS content-addressable memory (CAM) cell (22). The normally slow ring function between adjacent cells to signify a match is significantly improved by the wire-or structure permitted by the addition of the bipolar transistor. Reprinted with permission from "A 4-ns BiCMOS translation-lookaside buffer," Tamura et al., *IEEE J. Solid-State Circuits*. © 1990 IEEE.

priority encoder circuits that handle potential multiple data matches that are unique to CAMs.

Figure 18 shows a BiCMOS CAM cell with nine CMOS transistors and one n - p - n transistor. During matching, the bit lines are driven with the input data. If the matching does not rely on this bit position, then both the bit lines are driven high. If the internal cell potential mismatches the input data, either p -FET M7 or M8 is turned on and the base of n - p - n transistor Q1 is pulled high, which pulls up the *match_L* line. If no cells on a given *match_L* line mismatch, the *match_L* line stays low, which indicates that a match has been found. To improve performance, the bit-line voltage swing is kept small [700 mV in (22)], which yields a similar *match_L* line voltage swing.

CAM cells must also be written and read (though typically only for diagnostic purposes). During writing, the *wl_L* signal is lowered, which turns off n -FET M9. p -FET access transistors M5 and M6 then drive the bit-line voltage onto the internal nodes of the cross-coupled inverter. Because M9 is off, the p -FETs do not have to overpower n -FETs M3 and M4 in order to write. Therefore, the CAM cell can be written with a small voltage swing on the bit lines. While the small bit-line voltage swing improves performance, it also limits the reverse base-emitter voltage on Q1, which would otherwise affect the n - p - n 's reliability. Note that in a CAM, all cells on a given *wl_L* line are always written together, which permits the use of M9, unlike in an SRAM, in which many data words typically share the same row and are multiplexed during access and writing.

FUTURE TRENDS

Despite the advantages of bipolar transistors over CMOS transistors and the inherent speed advantage of current

steering logic, BiCMOS SRAMs are likely to disappear in the future because of their higher power and poor voltage scaling. In the short term, lateral bipolar transistors on silicon-on-insulator (SOI) processes could provide another generation of BiCMOS SRAMs. Lateral bipolars can have significantly smaller parasitic collector capacitances than vertical bipolars, which produce similar ECL gate delays at lower gate currents (23). Furthermore, SOI processes provide many advantages over bulk processes for SRAMs, such as near-zero alpha-particle susceptibility, reduced n -FET-to- p -FET spacing and thus smaller 6T cell size, and lower parasitic capacitance.

In the longer term, new bipolar device structures with lower V_{be} drops may allow ECL-like circuits to operate at lower voltages. The bipolar-FET hybrid transistor is one example in which the CMOS transistor gate is tied to the floating body in a CMOS SOI process (24). This device structure has bipolar current-voltage characteristics and a V_{be} voltage that is about 0.3 V less than a conventional bipolar transistor.

BIBLIOGRAPHY

1. R. B. Ritts et al., Merged BiCMOS logic to extend the CMOS/BiCMOS performance crossover below 2.5 V supply, *IEEE J. Solid-State Circuits*, **26**: 1606–1613, 1991.
2. K. Yamaguchi et al., A 1.5-ns access time, 78- μm^2 memory-cell size, 64-kb ECL-CMOS SRAM, *IEEE J. Solid-State Circuits*, **27**: 162–173, 1992.
3. M. Santoro, L. Tavrow, and G. Bewick, A Subnanosecond 64kb BiCMOS SRAM. In *Proc. BCTM*, 1994, pp. 95–98.
4. T. Douseki et al., Fast-access BiCMOS SRAM architecture with a V_{ss} generator, *IEEE J. Solid-State Circuits*, **26**: 513–517, 1991.
5. A. R. Alvarez, (ed.), *BiCMOS Technology and Applications*, 2nd ed., Boston: Kluwer Academic Publishers, 1993.
6. J. Hayden et al., A high performance 0.5 μm BiCMOS technology for fast 4-Mb SRAM's *IEEE Trans. Electron Devices*, **39**: 1669–1677, 1992.
7. J. Kirchgessner et al., An advanced 0.4 μm BiCMOS technology for high performance ASIC applications. In *IEDM Tech. Dig.*, 1991, pp. 97–100.
8. A. A. Iranmanesh et al., A 0.8- μm advanced single poly BiCMOS technology for high-density and high-performance applications, *IEEE J. Solid-State Circuits*, **26**: 422–426, 1991.
9. R. A. Kertis, D. D. Smith, and T. L. Bowman, A 12-ns ECL I/O 256Kx1-bit SRAM Using a 1- μm BiCMOS Technology, *IEEE J. Solid-State Circuits*, **23**: 1048–1053, 1988.
10. T. Douseki and Y. Ohmori, BiCMOS circuit technology for a high-speed SRAM, *IEEE J. Solid-State Circuits*, **23**: 68–73, 1988.
11. H. Toyoshima et al., A 6-ns, 1.5-V, 4-Mb BiCMOS SRAM, *IEEE J. Solid-State Circuits*, **31**: 1610–1616, 1996.
12. M. Odaka et al., A 512kb/5ns BiCMOS RAM with 1KG/150ps logic gate array. In *ISSCC Tech. Dig.*, 1989, pp. 28–29.
13. D. E. Wingard, D. C. Stark, and M. A. Horowitz, Circuit techniques for large CSEA SRAM's, *IEEE J. Solid-State Circuits*, **27**: 908–919, 1992.
14. M. Suzuki et al., A 3.5-ns, 500-mW, 16-kbit BiCMOS ECL RAM, *IEEE J. Solid-State Circuits*, **24**: 1233–1237, 1989.
15. K.-Y. Toh et al., A 23-ps/2.1-mW ECL gate with an ac-coupled active pull-down emitter-follower stage, *IEEE J. Solid-State Circuits*, **24**: 1301–1306, 1989.
16. N. Homma et al., A 3.5-ns, 2-W, 200- μm^2 , 16-kbit ECL bipolar RAM, *IEEE J. Solid-State Circuits*, **SC-21**: 675–679, 1986.
17. W. R. Blood Jr., *MECL System Design Handbook*, 4th ed., Phoenix, AZ: Motorola Inc., 1988, p. 113.
18. K. Nakamura et al., A 6-ns ECL 100k I/O and 8-ns 3.3-V TTL I/O 4-Mb BiCMOS SRAM, *IEEE J. Solid-State Circuits*, **27**: 1504–1510, 1992.
19. T. Shiomi et al., A 5.8-ns 256-kb BiCMOS TTL SRAM with T-shaped bit line architecture, *IEEE J. Solid-State Circuits*, **28**: 1362–1369, 1993.
20. M. Horowitz et al., A 3.5ns, 1 Watt, ECL register file. In *ISSCC Tech. Dig.*, 1990, pp. 68–69.
21. R. Taylor and M. Johnson, 1Mb CMOS DRAM with a divided bitline matrix architecture, *ISSCC Tech. Dig.*, 1985, pp. 242–243.
22. L. R. Tamura et al., A 4-ns BiCMOS translation-lookaside buffer, *IEEE J. Solid-State Circuits*, **25**: 1093–1101, 1990.
23. W.-L. M. Huang et al., TFSOI complementary BiCMOS technology for low power applications, *IEEE Trans. Electron Devices*, **42**: 506–512, 1995.
24. S. A. Parke, C. Hu, and P. K. Ko, Bipolar-FET hybrid-mode operation of quarter-micrometer SOI MOSFET's, *IEEE Electron Device Lett.*, **14**: 234–236, 1993.

LEE S. TAVROW
Micro Magic, Inc.

BICMOS MEMORY CIRCUITS. See FIELD EFFECT TRANSISTOR MEMORY CIRCUITS.
BICONICAL ANTENNAS. See CONICAL ANTENNAS.