SRAM CHIPS

Semiconductor memories play a vital role in today's electronics for storage of software programs instruction sets for microprocessor operation. They are used as stand-alone memory at the system level or as embedded memory for increased microprocessor speed. Memory devices are classified as volatile or nonvolatile. Volatile memories require power to retain the information while nonvolatile memories do not. One type of volatile memory is the dynamic random access memory (DRAM), which consists of a capacitor to store charge and of a transistor to control access to the capacitor. The other type of volatile memory is the static random access memory (SRAM), which consists of four transistors plus two load elements (either resistor or transistor) configured to remain in a fixed state until externally changed. SRAM lags DRAM in density per chip by roughly a factor of 4 because of the larger number of elements per cell. SRAM generally has superior data access time (less than half) and lower power dissipation (less than half) compared to DRAM. For example, in 1996 commercially available memory for DRAM was in the 4 megabit to 16 megabit (Mb) array size with read/write times around 70 ns while SRAM was available in 1 Mb to 4 Mb sizes with read/write times around 20 ns (1). SRAM finds specific applications for embedded memory in ASICs and microprocessors to increase speed (since interface circuits and package leads are eliminated) or as the main memory for very low power applications. SRAM is favored over DRAM when high-speed and/or low-power RAM is required for applications such as first-level cache memories. Cache memories are circuits that hold selected data from the larger main memory, allowing higher microprocessor performance due to the faster memory access time. The concept is similar to a person's library. The book-case containing most of the books is equivalent to the main memory while the books that are more readily accessible on the desk would be equivalent to the function of first-level cache memory. This article focuses on the SRAM chip, discussing the configuration, operation, comparison of various cell types, past and current trends in SRAM cells, circuit techniques used to increase SRAM performance, and failure and wear-out mechanisms. Viable technologies being developed to manufacture cost-effective, high-performance SRAM into the next century are also discussed.

THE BASICS OF MEMORY OPERATION

The basic architecture for a RAM memory chip is shown in Fig. 1 and is composed of the memory array, address predecoders and decoders, input and output buffers, sense amplifiers, and read/write circuitry. Specific operation of each of these circuit blocks are discussed in detail later, but a brief summary of the key functions is offered here by way of introduction.

The memory array is made up of individual storage elements or cells and can be configured in a square to maximize cell density over a given area as shown in the 8 × 8 (= 64 cell) array of Fig. 1, where each square represents a memory cell. In stand-alone memory chips the array is typically well over 50% of the entire area of the chip and is a key circuit block for die area reduction. While the DRAM cell is made up of two elements, the larger SRAM cell is made up of six elements. Both have the same memory function, which is to be

J. Webster (ed.), *Wiley Encyclopedia of Electrical and Electronics Engineering*. Copyright © 1999 John Wiley & Sons, Inc.

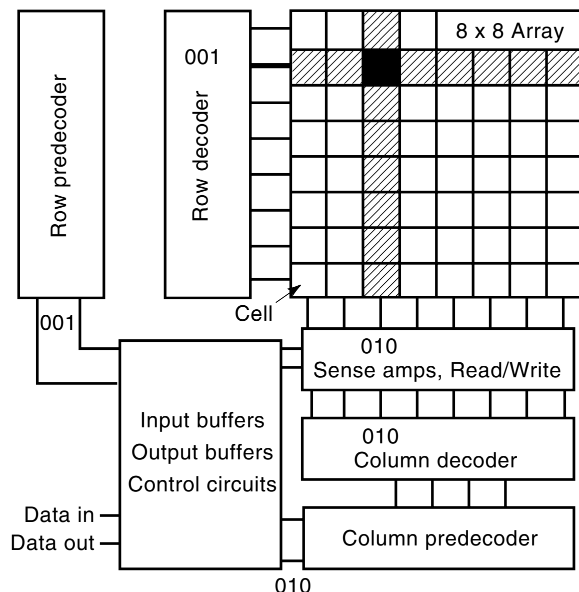


Figure 1. SRAM chip architecture in circuit block form.

in either one of two well-defined electrical states. These two states represent binary digits (or bits). In DRAM, the electrical state is determined by whether the capacitor element is holding charge or not. The electrical state of the SRAM cell is defined by the output voltage of the cell, which will be either high or low. Because of the dynamic nature of DRAM, its cells must be continually refreshed to keep the bit from losing its current logic state due to leakage from the capacitor. SRAMs do not require constant refreshing but maintain their latched logic state until forced into another by the write circuitry. However, both SRAM and DRAM require dc power for each of the memory cells to remain in its logic state.

The horizontal line connected to all cells in a given row is called the *word line*. In like manner, the line connected to the cells in a column is called the *bit line*. The cell state is read from or written to through the bit lines while the word line provides access to the cell. A specific cell of the array is accessed through the row and column decoders to allow bit line connection for reading from or writing to a specific cell. When the row decoder selects the desired address all of the cell access transistors in the selected row are turned on and any of the memory cells in this row are now accessible for read/write operation. The column decoder selects the bit lines of the specific cell to be read from or written to. For a specific row and column address, only one cell from the array will have both the access gate on and the bit lines accessed. Any bit in the array can be randomly accessed in this fashion, leading to the term *random access memory*. Figure 1 shows an example of address selection in the 8×8 array. The row decoder takes the binary input number of 001 ($= 2^0$) for the word-line address and selects the corresponding row 1. In the same manner, the column decoder takes the binary input number (010 in this example) and selects the desired column 2. The memory array is divided into $2^l \times 2^m$ number of bits, where l and m are the number of rows and columns, respectively, in the array. An 8-bit decoder can address 2^8 or 256 rows or columns. Semiconductor memories are typically offered in sizes of 2^n. Thus a 16 Mb memory is not 16,000,000 bits but 2^{24} or 16,777,216 bits.

The address location is kept in the input buffer until the decoder is ready to receive it. This occurs when the access to the location from the previous address is completed. Prior to the decoder circuit is the predecoder, which is used to simplify the circuit and to reduce die size. The predecoder sends the address into the decoder in smaller blocks, reducing the number of inputs to the decoder.

Once the desired bit is accessed, logic in the read/write circuitry dictates whether the state of the cell will be read or written to. A sense amplifier is contained in this circuit block to amplify the signal from the bit lines to allow for accurate and fast reading of the cell. This amplifier is necessary for high-speed operation because of the capacitive loads along the bit lines of the column of the selected cell, especially as the array gets larger. The data outputs are then fed into the output buffer for access from the external systems.

Memory designs can be categorized as synchronous, asynchronous, or static load. Synchronous memory requires a clock edge to enable memory operation while asynchronous memory does not. Asynchronous memory is designed to determine address changes and outputs the data following a change. Static load memories also require a clock. The relative complexity of a given design depends on which type of memory is used. Synchronous memory is faster since all of the inputs are clocked into the memory, but it does require a more complex design compared to asynchronous, which is simpler in design but suffers from internal time delays (1a,1b).

THE FUNDAMENTAL COMPONENTS OF THE SRAM CELL

The Inverter

This section addresses the fundamental components and the basic operation of the SRAM memory cell. The main functional component of the SRAM is the latch which is a bistable circuit made up of two inverters connected in a positive feedback loop. The inverter is the basic SRAM building block and incorporates a driver and a load tied in series as shown in the inset of Fig. 2. The driver functions as a voltage-controlled switch to invert an incoming voltage from a low voltage to a high voltage or vice versa. When the input voltage V_i moves high, the output voltage V_o is connected to ground. Conversely, when V_i goes low, V_o is connected to the power supply voltage V_{dd} through the load. The voltage-controlled switch is typically a bipolar transistor or a Metal-Oxide-Semiconductor Field Effect Transistor (MOSFET) and is often called the driver or pull-down transistor because it pulls the output to ground when it is on. The load is typically a transistor or resistor and is called the pull-up element because it pulls V_o up to V_{dd} when the driver is off and the load transistor is on. The ideal and typical voltage transfer characteristics (V_o as a function of V_i) for a typical inverter are shown in Fig. 2 as denoted by the dashed and solid lines, respectively. When V_i is lower than a specified low input voltage V_{il}, V_o is at the high output voltage V_{oh}. Conversely, when V_i is greater than a specified high input voltage V_{ih}, V_o is at the low output voltage V_{ol}. V_{il} defines the maximum V_i necessary to keep $V_o = V_{oh}$ whereas V_{il} defines the minimum V_i necessary to keep $V_o = V_{ol}$. Both are defined at the point on the voltage transfer curve where the slope $= -1$. Maximum inverter performance is obtained for the ideal case when $V_{il} = V_{ih} = V_{dd}/2$, $V_{oh} = V_{dd}$, and

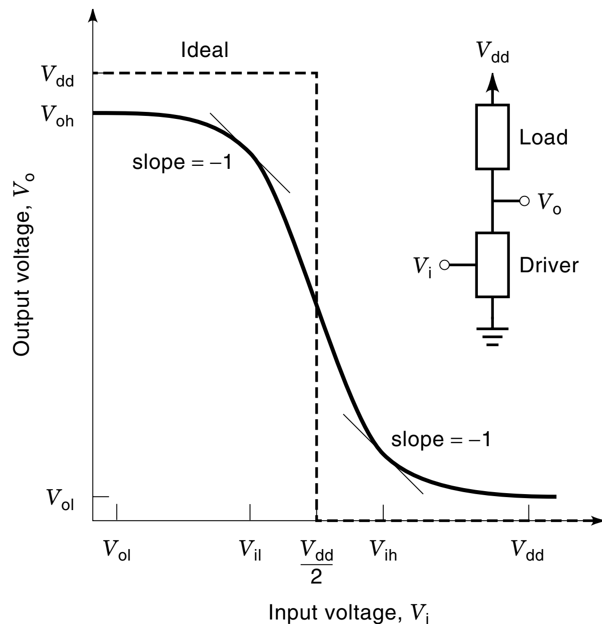


Figure 2. The ideal (dashed line) and typical (solid line) voltage transfer curve for the generic inverter as shown in the inset.

V_{ol} = 0. The choice of load element directly affects each of these values.

Three measures of inverter performance are the noise margin, the propagation delay, and the power dissipation. The noise margin NM defines the maximum amplitude allowed at the input without changing the output and thus quantifies the inverter's deviation from the ideal case. The NM helps specify the noise allowed on V_{in} of the gate such that the output is not affected. The high noise margin NM_h and low noise margin NM_l are defined as

$$NM_h = V_{oh} - V_{ih} \tag{1}$$

$$NM_1 = V_{il} - V_{ol} \tag{2}$$

In the ideal case $NM_h = NM_l = V_{dd}/2$. The propagation delay is defined as the average of the 50% points of the leading and trailing edges when the inverter is switching from low to high and from high to low. This is shown in Fig. 3 where τ_{phl} (or τ_{plh}) refers to the time difference between the 50% point on the rising (or falling) edge of V_i and the 50% point on the falling (or rising) edge of V_o. Then the propagation delay is defined as

$$\tau_p \equiv \frac{1}{2}(\tau_{phl} + \tau_{plh}) \tag{3}$$

The average power P_{av} dissipated in the inverter depends on whether the inverter is operating in the static (no switching) or dynamic (during switching) mode. The static power for an inverter with a MOSFET driver and resistor load R_L is given as $V_{dd}^2/2R_L$. The dynamic power is $C_L V_{dd}^2 f$ where f is the operating frequency and C_L is the load capacitance. These values depend on the driver and load used and are derived for various inverters in the next section. The power delay product $\tau_p P_{av}$ is a figure of merit often used to quantify the performance of the inverter.

The choice of load and voltage-controlled elements depends on the need of the application and directly affect array size, cost, switching speed, and power dissipation. For an inverter using a bipolar transistor as the driver, the load element is a low impedance resistor. When a MOSFET is used (typically n-type), the load element is either a poly resistor or NMOS (n-type MOSFET) or PMOS (p-type MOSFETs) device.

The Latch

Two inverters connected in a positive feedback configuration as shown in the inset of Fig. 4 form a circuit known as a latch. The solid line on the voltage transfer curve (VTC) represents the case when the node A is disconnected from node D and connected to an external power supply whereas the dashed line represents the case when the voltage V_A at node A is equal to that at node D, V_D. Figure 4 shows three possible operating points for the latch which are denoted by points i, ii, and iii. Point ii on the transfer curve is unstable because

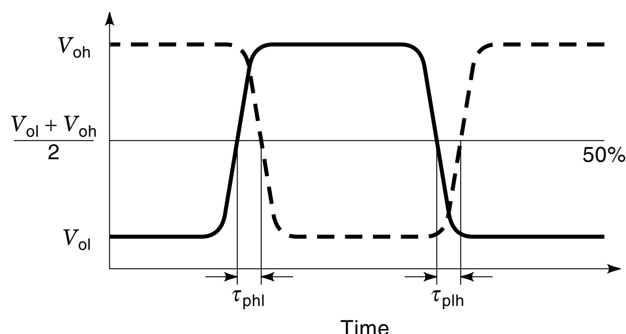


Figure 3. Inverter input and output voltage waveforms which show the definition of propagation delay for the high to low (τ_{phl}) and low to high (τ_{plh}) transistions. The output and input voltages are denoted by the dashed and solid curves, respectively.

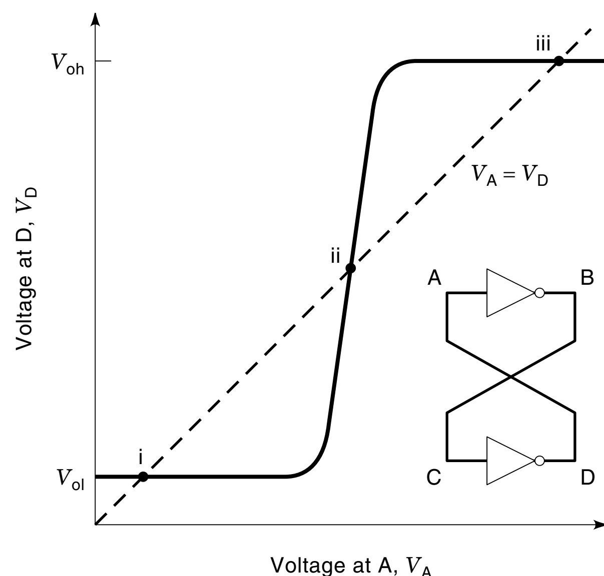


Figure 4. Voltage transfer curve (the solid line) for the basic inverter latch shown in the inset. The dashed line represents the case when $V_A = V_D$. The three possible operating points are denoted by i, ii, and iii.

any small fluctuation in the voltage amplifies and shifts V_D along the curve because of the feedback gain of the configuration. However, there is no gain at points i or iii and thus any incremental change in the voltage at V_A is not amplified as long as that change occurs at a voltage above V_{ih} or below V_{il}. Thus, the latch functions as a memory device because it remains in either of its two stable operating points which are represented as a logic 1 or 0.

INVERTER ANALYSIS AND COMPARISON

The SRAM cell consists of a bistable latch connected to triggering circuitry to force the latch into either one of its stable operating points in which it remains as long as power is applied. The upper dashed box in the circuit shown in Fig. 5 encloses the standard six transistor (6T) SRAM cell made up of PMOS load elements (T7 and T8) and NMOS access transistors (T1 and T2) to each storage node (denoted as A and B). The lower dashed box shown in Fig. 5 is used for read operation and is discussed later.

Figure 6 shows two MOSFET inverters used for SRAM memory cells along with the respective driver and load current-voltage (I–V) characteristics. Figure 7 provides the voltage transfer curves for each of these inverters and is referred to later in the article.

Resistor Load NMOS

The simplest inverter to analyze and currently one of the more widely used for high-density SRAM is the NMOS driver with resistor load as shown in Fig. 6(a). With a high load resistance, the static power dissipation is reduced because it is equal to $V_{dd}^2/2R_L$. In early resistor load (or R-load) NMOS, the resistors were very large (because of the low poly sheet resistance), leading to large inverter areas. Advances in poly resistor processing have led to high-resistivity poly allowing for a significantly reduced length needed for high resistance loads. The poly resistor can be deposited over the top of the cell, resulting in a smaller cell area compared with the transistor load inverters. One disadvantage of the R-load inverter is the added masking step required to define the poly 2 load resistor (where poly 1 is used to define the gates of MOSFET

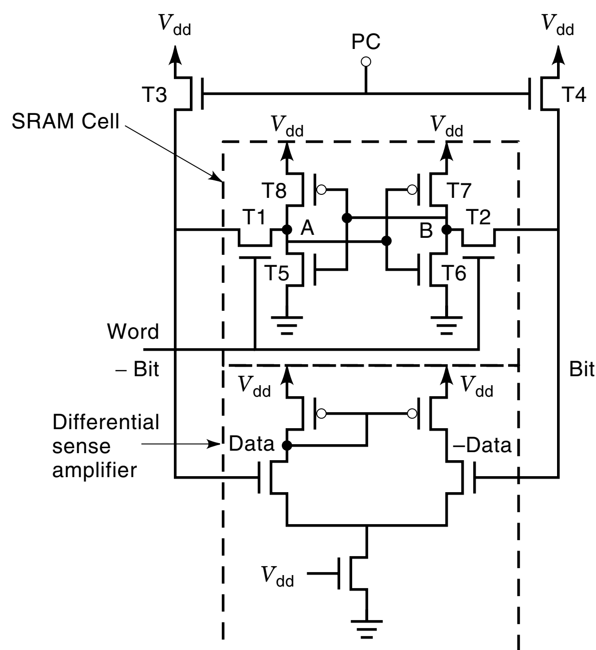
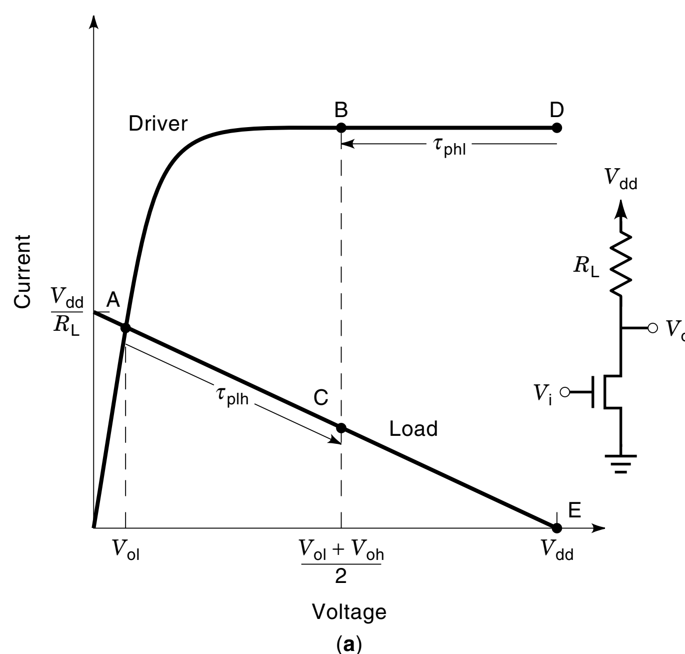


Figure 5. CMOS SRAM circuit configuration (contained within the dashed box), including the bit line pull up transistors, T3 and T4, and a simple sense amplifier used for read operation. Reprinted after Ref. 1b by permission of John Wiley & Sons, Inc. © 1991.



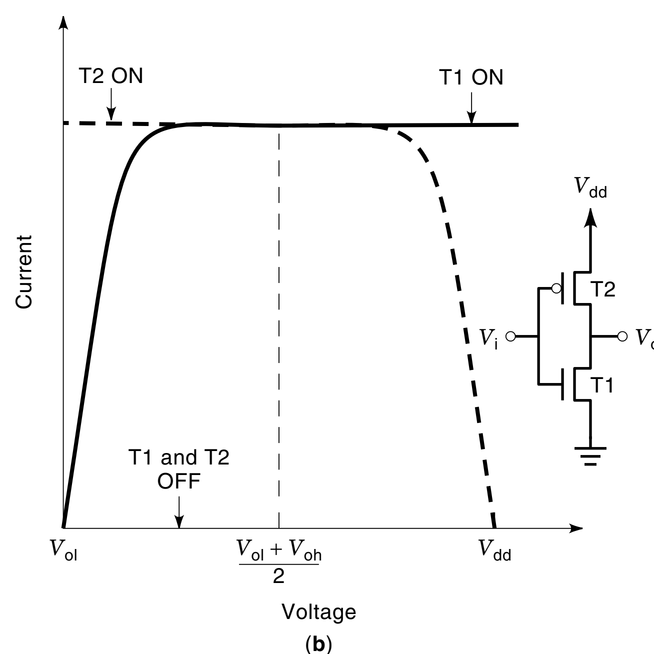


Figure 6. The driver transistor and load element I–V characteristics for the (a) resistor load inverter and (b) CMOS inverter. The propagation delays are noted on the appropriate I–V curve for each inverter as well as the minimum (V_{ol}), maximum (V_{oh}) and midpoint (V_{ol} + V_{oh})/2 output voltages.

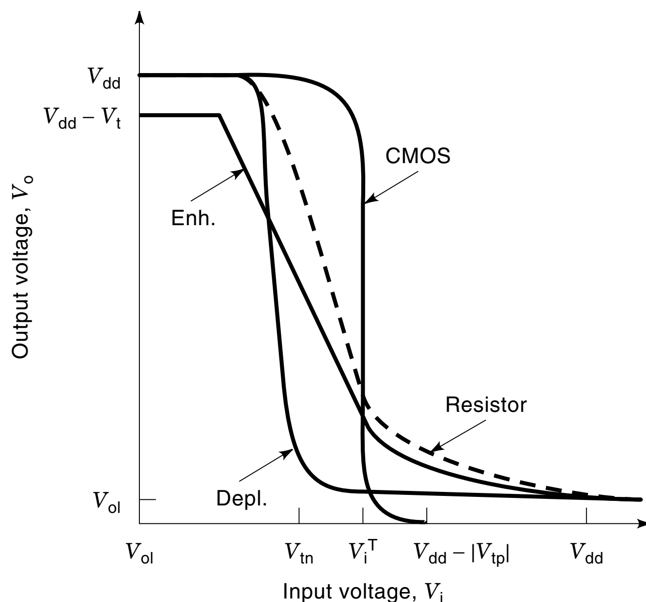


Figure 7. Comparison of voltage transfer curves for each of the inverters shown in Fig. 6, including the enhancement load and depletion load NMOS inverters. The key transition points for the CMOS inverter are indicated (V_{tn}, $V_{dd} - |V_{tp}|$, and V_i^T). Note how the CMOS curve approaches the ideal shown in Fig. 2.

driver and access transistors: T1, T2, T5, T6 in Fig. 5). In addition, strict process control is required to manufacture repeatable high-resistance poly in the range of 10 GΩ to 10 TΩ for today's SRAM.

In the R-load inverter, V_{oh} equals V_{dd} because there is not a measurable voltage drop across the resistive load when the NMOS driver is off. V_{ol} is obtained by equating the current through the NMOS transistor and the resistive load. As depicted in Fig. 2, V_{ol} occurs when $V_i > V_{ih}$. For $V_o = V_{ol}$ the gate-to-source voltage V_{gs} must be greater than V_{ih} and the device is on because V_{ih} is greater than the transistor turn-on or threshold voltage V_{tn} (a condition of the design is that $V_{il} < V_{th} < V_{ih}$). Because V_o is the drain-to-source voltage V_{ds}, the transistor is likely to be in the linear region of operation. The simple form of the drain current for a NMOS device in the linear region is given by

$$I_{dl} = \frac{W\mu_n C_{ox}}{2L}\left[2(V_{gs} - V_{tn})V_{ds} - V_{ds}^2\right] \tag{4}$$

where W and L are the gate width and length of the transistor, respectively. C_{ox} is the gate oxide capacitance and μ_n is the effective electron mobility because the NMOS transistor forms a channel of electrons for current flow. The current through the load resistor is given by

$$I_L = \frac{V_{dd} - V_{ol}}{R_L} \tag{5}$$

Setting Eq. (4) equal to Eq. (5) and solving for V_{ol} gives (2)

$$V_{ol} \approx \frac{V_{dd}}{1 + \dfrac{W\mu_n C_{ox}}{L} R_L (V_{dd} - V_{tn})} \tag{6}$$

Recall that V_{il} and V_{ih} are defined at the points where the slope $(dV_o/dV_i) = -1$ on the voltage transfer curve. The slope is found by equating the inverter to two resistors in series where V_o is the node between the resistors. The output resistance of the transistor driver is r_{ds} which gives

$$\frac{dV_o}{dV_i} = -\frac{dI_d}{dV_i}(R_L \| r_{ds}) = -1 \tag{7}$$

When $V_i = V_{il}$, V_o approaches V_{dd}. Therefore, the NMOS transistor is operating in the saturation region because $V_{ds} = V_o$. The drain current vs drain-to-source voltage curve ($I_d - V_{ds}$) for the MOSFET is nearly flat when the device is in saturation, and thus, r_{ds} is very high. In this case $R_L \| r_{ds}$ approaches R_L. The simplified equation of the current when the MOSFET is in saturation is

$$I_{ds} = \frac{W\mu_n C_{ox}}{2L}(V_{gs} - V_{tn})^2 \tag{8}$$

Inserting Eq. (8) into Eq. (7) (with $V_{gs} = V_{il}$) gives

$$\frac{W\mu_n C_{ox}}{L}(V_{il} - V_{tn})R_L = 1 \tag{9}$$

from which we can solve for V_{il}. V_{ih} is found similarly to V_{il}. In this case, dV_o/dV_i is differentiated as follows:

$$\frac{dV_o}{dV_i} = -\frac{dI_d}{dV_i}\frac{dV_o}{dI_d} = -1 \tag{10}$$

Using the expression in Eq. (4) for I_d in Eq. (10) gives

$$V_{ds} = \frac{V_{gs} - V_{tn}}{2} \tag{11}$$

When V_i (which is V_{gs}) is equal to V_{ih}, then V_o (which is V_{ds}) approaches V_{ol} and the NMOS transistor is in the linear operating region. Setting I_L equal to I_{dl} and using Eqs. (4) and (5), where V_{gs} equals V_{ih} and V_{ds} is given by Eq. (11), gives a quadratic equation for V_{ih}.

Once expressions for V_{oh}, V_{ol}, V_{il}, and V_{ih} are obtained, as outlined previously, the noise margins are calculated from Eqs. (1) and (2). These results show that V_{ih} and V_{il} increase and the slope in the transition region of the voltage transfer curve VTC decreases as R_L is decreased.

The propagation delay for the inverter consists of the time necessary to charge and discharge the capacitive loads connected to V_o which come from other transistors and parasitic capacitances in the circuit. The speed of charging and discharging depends on the current drive from the inverter (recall $I = dQ/dt$). Thus, a lower drain current takes longer to charge or discharge the load capacitance and, hence, takes more time to propagate a signal through the inverter or series of inverters. SPICE modeling is required for a more exact solution of the propagation delay because the NMOS transistor drain current is a nonlinear function of V_{gs} and V_{ds} (or V_i and V_o). However, a first-order estimate is obtained if we assume that a constant current charges or discharges the capacitive load. This current is an average of the current through the load device at the endpoint of the inverter transition. Given

$I_{av}dt = C_L dV$, τ_{plh} and τ_{phl} are expressed by (Ref. 2, p. 94)

$$\tau_{plh} = \frac{C_L(V_{oh} - V_{ol})/2}{I_{lh,avg}} \tag{12}$$

and

$$\tau_{phl} = \frac{C_L(V_{oh} - V_{ol})/2}{I_{hl,avg}} \tag{13}$$

Recall from Fig. 3 that for $I_{lh,avg}$, we are interested in the time it takes for V_o to charge from V_{ol} to the 50% point $(V_{oh} + V_{ol})/2$. Conversely, for $I_{hl,avg}$, we are interested in the time it takes for V_o to discharge from V_{oh} to the 50% point. Thus, $I_{lh,avg}$ and $I_{hl,avg}$ are determined from the key operating points on the load and driver I–V curves as shown in Fig. 6(a). In the transition from V_{ol} to V_{oh}, most of the current flows through the load resistor (from point A to point C) because the drive transistor is initially on (3):

$$I_{lh,avg} = \frac{[I_L(A) + I_L(C)]}{2} \tag{14}$$

In the transition from V_{oh} to V_{ol}, current flows through both the driver [from point D to B in Fig. 6(a)] and the load (from point E to C) which gives (3)

$$I_{hl,avg} = \frac{\{I_d(D) + [I_d(B) - I_L(C)]\}}{2} \tag{15}$$

where I_L and I_d are the load and NMOS currents, respectively. Equations (12) and (13) are solved using the results from Eqs. (14) and (15). Then the total propagation delay for the inverter is determined from Eq. (3). Equations (12)–(15) show that increasing the drive current reduces the propagation delay and is accomplished by reducing the gate length, gate oxide thickness, and/or threshold voltage, as is typically done in each technological shrink [see Eqs. (4) and (8)].

The average power P_{av} is $V_{dd}I_{dd}(\max)/2$ for the NMOS inverter considered here, where $I_{dd}(\max)$ is the maximum power supply current. The average currents are proportional to $I_{dd}(\max)$ and hence the power delay product is proportional to (Ref. 2, p. 96)

$$\frac{C_L(V_{oh} - V_{ol})}{I_{dd}(\max)} \frac{V_{dd}I_{dd}(\max)}{2} = \frac{C_L(V_{oh} - V_{ol})V_{dd}}{2} \tag{16}$$

Thus, to minimize power dissipation per logic decision in an NMOS inverter the logic swing, power supply voltage, and/or capacitive loading should be reduced. The load capacitance depends on the gate oxide thickness of the MOSFETs and parasitic capacitances, from metal to substrate, metal to poly, metal 1 to metal 1, and metal 1 to metal 2. A simple expression is (4)

$$C_L = C_g \times \text{F.O.} + C_j + C_m \tag{17}$$

where C_g is the gate capacitance, C_j is the drain junction capacitance, C_m is the wiring load capacitance and F.O. is the fan-out or number of load gates on the output. The gain and drain capacitances are more or less dictated by the device requirements of the technology, but the process can be optimized to minimize the metal capacitive loads.

CMOS Inverter

The other inverter shown in Fig. 6(b) is the complementary MOS (CMOS) inverter which uses a PMOS for the pull-up transistor (whose well is tied to V_{dd}). Current flows in this inverter only during switching because either the pull-down or pull-up transistor is off during standby. Thus, the static power dissipation comes only from leakage current and is essentially zero. As indicated by the load line of the CMOS inverter in Fig. 6(b), the maximum amount of current is available across the entire range of output voltages, which leads to high switching speeds. In addition, the noise margin for the CMOS inverter is maximum because $V_{oh} = V_{dd}$ and $V_{ol} = 0$ V. A disadvantage of the CMOS inverter is the added process complexity and cost because a simple nine-mask R-load NMOS process is increased to 12 masks for the CMOS process, an increase of 33% in both cycle time and cost. In addition, larger diffusion spacing design rules (e.g., N+ to P+) are required to avoid latch-up conditions in the array. To maximize switching speed, the PMOS gate size must be nearly 2.5 times that of NMOS for the same drive current because of the lower carrier mobility μ_p in PMOS. Both of these factors lead to increased cell size for CMOS. Figure 6(b) shows the case when the drive current of the PMOS is the same as that of the NMOS. Despite all of these disadvantages, the CMOS inverter is used extensively for SRAM because of the lower power, greater noise immunity, and better operation at low voltage. Thus, CMOS is particularly useful for low voltage battery applications and embedded memory in today's high-performance microprocessors.

Figure 7 shows how the VTC of the CMOS inverter approaches ideal operation. When $V_i < V_{tn}$, the PMOS device is on, the NMOS device is off, and $V_o = V_{oh} = V_{dd}$. When $V_i = V_{tn}$, the NMOS pull-down transistor begins to conduct, and V_o drops. As the input voltage is further increased to $V_i = V_{dd} - |V_{tp}|$, the PMOS pull-up transistor turns off, making $V_o = V_{ol} = 0$ V. V_{il} and V_{ih} are solved for similarly, as discussed previously, by equating the drain currents and differentiating with respect to V_i. For $V_i = V_{il}$, the NMOS is in the linear region and the PMOS device is in saturation whereas the opposite is true when $V_i = V_{ih}$. With this information and the relationship $dV_o/dV_i = -1$ (Ref. 2, p. 100),

$$V_{ih} = \frac{2V_o + V_{tn} + \left[\dfrac{(W\mu_p/L)_p}{(W\mu_n/L)_n}\right](V_{dd} - |V_{tp}|)}{1 + \left[\dfrac{(W\mu_p/L)_p}{(W\mu_n/L)_n}\right]} \tag{18}$$

and

$$V_{il} = \frac{2V_o - V_{dd} - |V_{tp}| + \left[\dfrac{(W\mu_n/L)_n}{(W\mu_p/L)_p}\right]V_{tn}}{1 + \left[\dfrac{(W\mu_n/L)_n}{(W\mu_p/L)_p}\right]} \tag{19}$$

The transition voltage V_i^T from PMOS to NMOS conduction occurs when V_i is between V_{tn} and $V_{dd} - |V_{tp}|$. From V_{tn} to V_i^T, the NMOS device is in saturation, and the PMOS device is in linear operation. Between V_i^T and $V_{dd} - |V_{tp}|$, the PMOS device is in saturation, and the NMOS device is in linear operation. At V_i^T, both devices are in saturation, and their currents are

equal ($I_{dsat,p} = I_{dsat,n}$). Using Eq. (8) (5),

$$\left(\frac{W\mu_n C_{ox}}{L}\right)_n (V_i^T - V_{tn})^2 = \left(\frac{W\mu_p C_{ox}}{L}\right)_p (V_{dd} - V_i^T - |V_{tp}|)^2 \tag{20}$$

and

$$V_i^T = \frac{\left[V_{dd} + V_{tn}\sqrt{\frac{(W\mu_n/L)_n}{(W\mu_p/L)_p}} - |V_{tp}|\right]}{\left[1 + \sqrt{\frac{(W\mu_n/L)_n}{(W\mu_p/L)_p}}\right]} \tag{21}$$

When $V_{tn} = |V_{tp}|$ and $(W\mu_n/L)_n = (W\mu_p/L)_p$, we get the ideal VTC with $V_i^T = V_{dd}/2$ and equally fast rise and fall times. Recalling that $Idt = C_L dV$, the time needed to discharge the capacitor (high to low transition) while the NMOS is in saturation is given by (Ref. 3, p. 875)

$$\tau_{phl,sat} = \frac{C_L[V_{dd} - (V_{dd} - V_{tn})]}{\left(\frac{W\mu_n}{L}\right)_n (V_{dd} - V_{tn})^2} \tag{22}$$

Integrating $C_L dV$ from $V_{dd} - V_{tn}$ to $V_{dd}/2$ gives (Ref. 3, p. 876)

$$\tau_{phl,lin} = \frac{C_L}{2\left(\frac{W\mu_n}{L}\right)_n (V_{dd} - V_{tn})} \ln\left(\frac{3V_{dd} - 4V_{tn}}{V_{dd}}\right) \tag{23}$$

The total delay time τ_{phl} is equal to the sum of Eqs. (22) and (23). The analysis for the low to high transition is the same where the PMOS device is in operation. Equations (22) and (23) indicate that the propagation delay is minimized by minimizing the load capacitance and maximizing the drive current, as was the case for the *R*-load NMOS inverter. It is important to note that the power supply voltage continues to decrease as geometric design rules shrink to avoid degradation of device performance because of hot carrier effects. Equations (22) and (23) indicate that this leads to an increase in the propagation delay as demonstrated by Norishima et al. who found that the delay per stage nearly doubles when V_{dd} is lowered from 5 V to 2 V (4).

The CMOS inverter has the best noise margin compared to its NMOS counterparts as shown in Fig. 7. The enhancement and depletion load NMOS inverters were used in early SRAM chips but are seldom used today. The reasons are that the enhancement load NMOS inverter has a reduced NM ($V_{oh} = V_{dd} - V_t$) and the depletion load NMOS inverter has high-power dissipation since the load is always on. In CMOS static power is nearly eliminated while the dynamic power is the same as for NMOS ($C_L V_{dd}^2 f$). Thus, the overall CMOS power consumption is reduced compared with the NMOS inverter which dissipates considerable amounts of static power.

Reducing the size of the PMOS device relative to the NMOS results in an asymmetric voltage transfer curve and slower propagative times because of the reduced drive current. To reduce the cell size while maintaining the advantages of CMOS performance an alternative approach was demonstrated (6,7) by using a PMOS thin-film transistor (pTFT). The TFT is a poly Si transistor built on top of the NMOS bulk driver. The cross section of the pTFT load cell is given in Fig. 8 (8). The disadvantage of this cell is that is has a lower current drive compared with bulk PMOS loads and it adds a masking step. Improving the pTFT load cell performance is addressed below. The key parameter is the ratio I_{on}/I_{off}, which is the pTFT on current to off current.

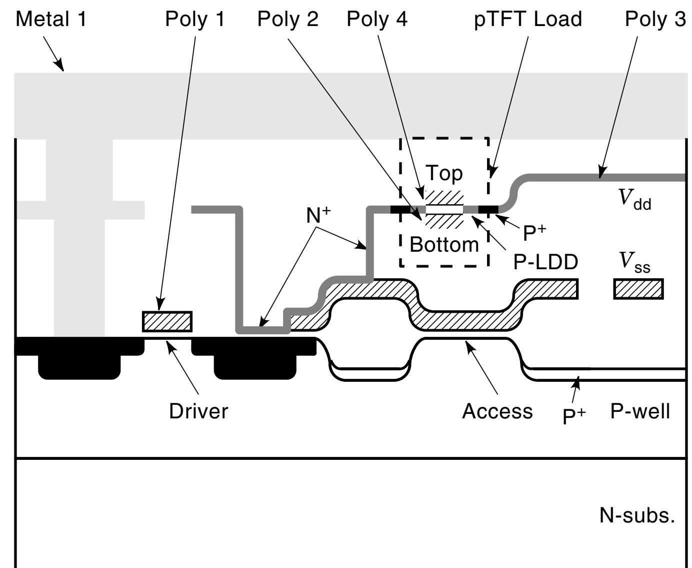


Figure 8. Cross section of pTFT load SRAM cell. The pTFT load shown within the dashed box is a double-gated device with a P-LDD region. The driver and access transistors are noted. Reprinted after Ref. 8 by permission of IEEE (© 1992 IEEE).

Other Inverter Technologies

In many cases SRAM designs have been tailored to take advantage of the benefits of the different technologies available. For example, one application is to use depletion load NMOS in the SRAM array and CMOS technology in the external circuitry to maximize speed and array density while lowering the overall power consumption compared with an all NMOS design.

When speed is the most critical objective for a given SRAM application (e.g., cache memory in high-speed computers), the bipolar transistor can be implemented for the driver because of its much faster switching speed relative to MOSFET devices. A common bipolar SRAM configuration utilizes emitter-coupled logic (ECL). Access times of less than 1 ns are achieved with bipolar SRAM technology (9). However, much higher power dissipation results because of the need for lower impedance resistors (high current) and a much larger cell area.

Bipolar CMOS (BiCMOS) technology has been developed to take advantage of the low-power CMOS for the SRAM cell and logic circuitry while using bipolar devices for the circuits needing high speed and high gain. These circuits include high capacitive nodes in the decoders, word-line drivers, output buffers, and the sense amps, which require high gain and need high input sensitivity for fast sensing of small differential bit-line swings (10). Application of bipolar devices decreases the access times and thus improves overall SRAM chip performance. One design implements bipolar, pTFT, and CMOS technologies to optimize performance, cell area, and static power dissipation (11).

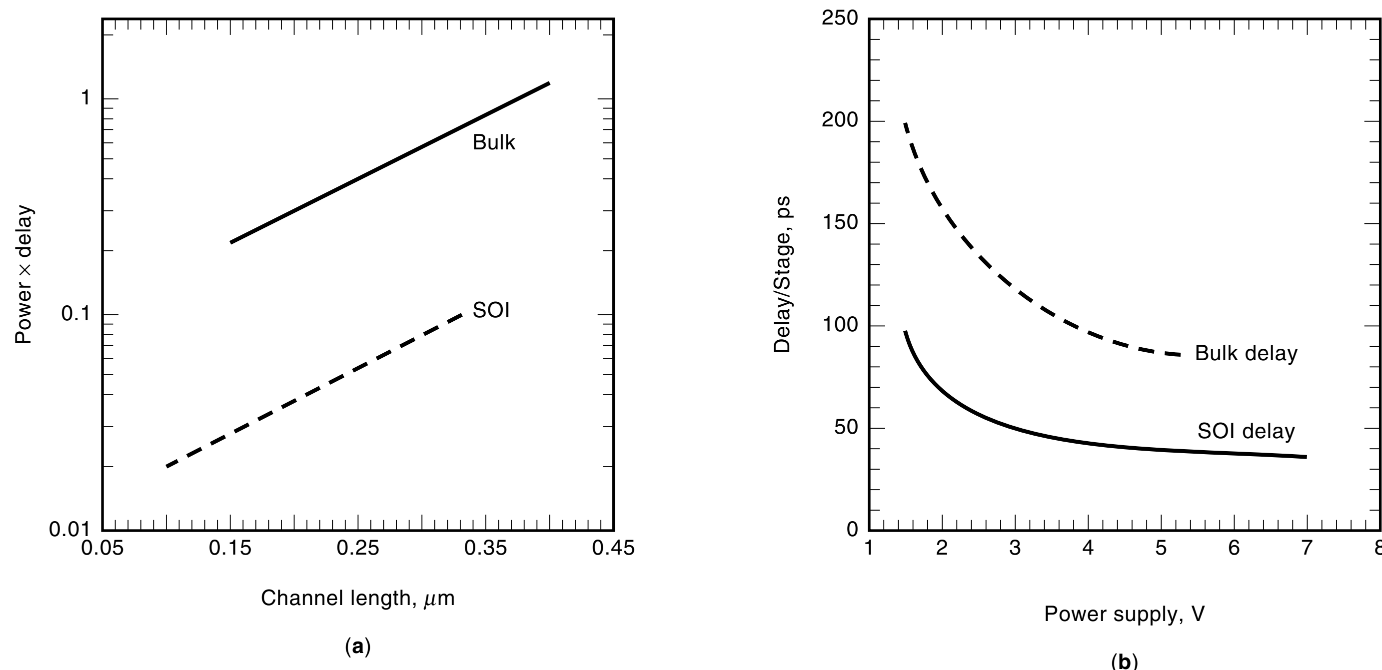


Figure 9. Comparison of SOI and bulk CMOS SRAM performance for (a) power-delay product vs channel length and (b) delay per stage vs power supply voltage. Reprinted after Ref. 12 and 63 by permission of IEEE (© 1993, 1989 IEEE).

Another technology being developed is semiconductor on insulator (SOI) where the active Si used for the SRAM is isolated from the substrate by a thick silicon dioxide layer. The use of SOI significantly reduces the parasitic capacitance associated with the substrate in standard MOSFET technologies. Much lower propagative delay times are obtained, particularly as the channel length and power supply voltages decrease as shown in Fig. 9 (12,63). In fact, Shahidi et al. from IBM demonstrated a 3.5 ns access time at 1 V using 0.1 μm CMOS (12). Additional advantages of SOI include significantly reduced latch-up and body effect and fewer soft errors. However, the higher cost and defectivity of SOI is limiting its current use. As gate lengths decrease below the 0.25 μm range, SOI may be an acceptable tradeoff to achieve the better performance.

THE FUNDAMENTAL COMPONENTS OF THE BASIC SRAM ARCHITECTURE

As noted above, there are a number of factors that limit the operating speed of the SRAM chip. In addition to the memory array itself, the access time of a SRAM is also influenced by the address buffer, decoders, sense amplifier, and output buffer circuitry (47) due to the delay with signal propagation through these circuits. The delay increases as the number of inputs and outputs increases. This section discusses the important details of these peripheral circuits and the methods used to improve their performance.

The Row Decoder

The decoder circuitry is divided into the predecoder and decoder. The function of the decoder is to select the desired row or column in the array based on the address that is sent to it from the input buffer. A conventional two input/four bit NAND/inverter-based decoder is provided as an example in Fig. 10(a) while Fig. 10(b) shows the transistor schematic for the NAND gate. The truth table for the NAND gate is given in Table 1. Only when both of the inputs (A and B in Fig. 10) are at a logic state "1" (or "high") does a logic state "0" (or low) get passed as the output of the NAND gate. An example of the decoder function is shown in Fig. 10(a), where the ad-

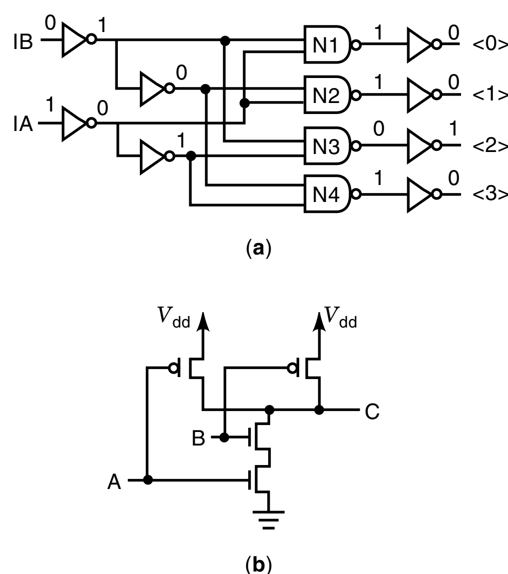


Figure 10. (a) Schematic of decoder circuit using NAND gates and inverters. (b) Transistor schematic of NAND gate.

Table 1. Truth Table for NAND Gate Shown in Fig. 10

Input A	Input B	Output C
0	0	1
0	1	1
1	0	1
1	1	0

dress of row 2 (10) is presented at the two inputs of the decoder. In this case, the NAND gate, N1, has two inputs, 1 and 0, and thus gives an output of 1. The same inputs and outputs occur at NAND gates N2 and N4. However, both inputs at N3 are 1, and, thus, N3 has an output of 0. The outputs from the NAND gates feed into the inputs of the inverters as shown in Fig. 10(a). As a result, the only line that is set high (logic state 1) is the line corresponding to row 2 which was the address selected. All other rows are at a logic state of 0 (or off). Table 2 summarizes the resulting output for each possible input address for the 4 bit decoder in Fig. 10.

The same principle of operation exists for larger decoders, which have a greater number of inputs and outputs and, thus, take up a larger area and operate at lower speeds. Each address line in the array would require a NAND gate output, resulting in extensive area overhead for larger memory arrays. Most have adopted the use of a predecoder to minimize the increase in area and improve overall decoder speed. The predecoder takes on the functionality of the decoder and is thus the same layout shown for the decoder example in Fig. 10(a), with the exception that the inverters at each NAND gate input are eliminated. Each output from the predecoder feeds into one of the corresponding NAND gate inputs in the decoder. The other input to each NAND gate in the decoder is connected to an enable circuit, which is triggered when the decoder is intended to be in operation. The output of each NAND gate in the decoder is then fed through an inverter into the corresponding word line in the array.

One can appreciate that as the memory array size increases, the number of devices required for the predecoder and decoder circuits can increase significantly, leading to speed and area penalties. The conventional two-input NAND gate decoder architecture results in a large total gate capacitance and large layout area which limits fast decoding operation (39,50). A simple estimate of the delay time associated with the word line can be obtained from (Ref. 12a, p. 831)

$$t_d = t_{90\%} - t_{10\%} = 2.303\tau - 0.105\tau \quad (24)$$

$$\tau = R_{wl}C_T \quad (25)$$

$$C_T = C_{wl}(A_{wl} - A_g) + C_gA_g \quad (26)$$

where C_{wl} and A_{wl} are the capacitance and area associated with the word line (typically a poly or polycide layer). C_g and A_g are the capacitance and area associated with the gates along the word line. This is a pessimistic estimate but gives values on the order of the actual circuit. A distributed lumped-parameter model would give a more accurate estimate (12a). Equations (24) and (25) are given to indicate the key components for load capacitance along the word line. Simulation software can be developed to address the minimization of these circuit loads (51).

More advanced schemes have been developed to improve speed and reduce area in the row decoder. One such improvement was the implementation of the predecoder circuit as discussed above. Another is the use of a divided word line (Refs. 1b, p. 446; 12b), which decreases the large capacitive load on the global word line by splitting it into a global, subglobal, and local word-line scheme. High-performance SRAM chips typically incorporate bipolar junction transistors (BJT) in a BiCMOS process because of their much larger gain compared to the MOSFET. The BJTs are effective at high-capacitive nodes such as exist in decoders, word lines, and output buffers (Ref. 1b, p. 190; 10). The penalty for using the BJTs is a higher power dissipation.

Table 2. State of Output Rows for Each Possible Input to the 4 Bit Decoder in Fig. 10

Address	Input IA	Input IB	Row 0	Row 1	Row 2	Row 3
00	0	0	1	0	0	0
01	0	1	0	1	0	0
10	1	0	0	0	1	0
11	1	1	0	0	0	1

The Column Decoder

The column decoder is typically smaller in design, utilizing pass gates for accessing the bit lines and for the transfer of data. A single pass gate approach is shown for two-column access in Fig. 11. The PC node is the precharge node used for increasing read speed and will be discussed in the Read/Write Operation section below. The incoming address lines, ⟨0⟩ and ⟨1⟩, come from the column predecoder which receives the address from the input buffer. When the pass gates associated with that column address are set high, data are allowed to transfer into or out of the cell being accessed (recall, the row decoder has selected the word line of the cell to be accessed at the same time), depending on whether the operation is read or write. The delay times associated with the pass gate column decoder are not as severe as for the NAND gate decoder due to the fewer number of transistors used. The column predecoders will be the same as or very similar to the row predecoders and, thus, require high fanout capability. BiCMOS SRAM designs typically utilize bipolar devices in the

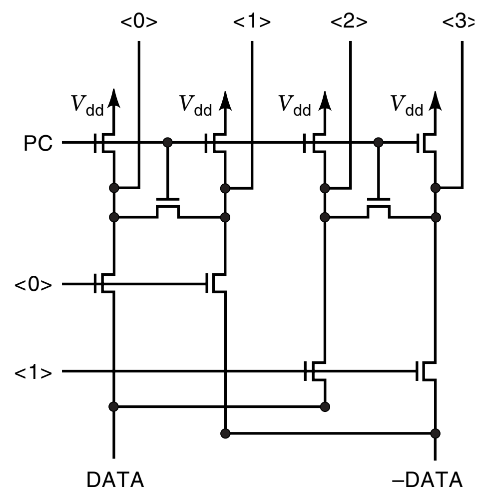


Figure 11. Pass gate column decoder circuit.

row predecoders to decrease access time while CMOS logic is often adequate for the pass gate approach of the column decoder shown in Fig. 11 (Ref. 10, p. 21).

SRAM Read/Write Operation

This section briefly discusses the read/write operation of SRAM. Two key measures of SRAM speed are the "read access time" and the "cycle time" (Ref. 10, p. 15). The read access time is the propagation delay from the time when the address is presented at the input of the memory chip until the data are available at the output. The cycle time is the minimum time that must be allowed after the initiation of the read/write operation before another read/write is initiated. Writing to any cell or falsely reading an incorrect state or cell in the array during read operation must be avoided. In addition, writing to an incorrect cell or disturbing the logic state of another cell during write operation must not occur.

The read operation is understood by examining the SRAM in Fig. 5 which is a CMOS SRAM cell with the addition of the bit-line pull-up transistors (T3 and T4) and a simple differential sense amplifier (1b). Two bit lines are generally needed to ensure maximum operating speeds (Ref. 2, p. 378). Read occurs by pulling the bit lines high (set the precharge node PC to V_{dd}) and then turning on the access transistors (T1 and T2) by applying V_{dd} to the word line. The read operation is designed this way because a single NMOS device is poor at passing a one, and the PMOS devices are generally small (Ref. 1a, p. 567). The logic state of the cell means that either node A or node B (see Fig. 5) will be low, and thus one of the bit lines is pulled low. Sense amp circuitry is connected to the bit lines to compare the voltages on the bit and −bit lines and thus determine the voltage at nodes A and B. When the access gates of T1 and T2 are turned on (word-line level is high), the bit that is pulled down by the logic state of the cell during read falls to a value, which is function of the size of T1, T2, T3, T4, T5, and T6. A typical waveform, which plots the various key cell node voltages as a function of time, is shown in Fig. 12 for the circuit in Fig. 5. Before setting the word line high, the bit lines are near 4 V in this example (V_{dd} = 5 V), and the DATA and −DATA nodes are just above 2 V. After the word line is set to V_{dd}, the bit line and the DATA and −DATA voltages diverge. DATA goes above 3 V, and −DATA goes below 1 V. The larger the pull-down transistors relative to the pull-up transistor, then the larger the difference between DATA and −DATA and, hence, the faster the sense amp is able to differentiate a signal. However, the size of the pull-down transistors is limited to keep the RAM cell size small, and thus there is a tradeoff between speed and differential voltage (Ref. 1a, p. 570). In addition, the conductance of the driver (T6 in Fig. 4) must be much larger than that of access device T2 so that the drain voltage of T6 does not rise above its V_{tn} and result in a change in the state of the cell during reading (Ref. 2, p. 380). This is an issue of cell stability which is discussed in detail later.

The read operation discussed above where PC is set to V_{dd} during the entire read, as shown in Fig. 12 is called a static read. To minimize power loss and pull-up time, a dynamic precharge design is used. Its configuration is the same as Fig. 5, except that PC is not tied to V_{dd} during the entire read and the sense amp is replaced by an inverter whose output is the data. In this case, the PC node is given a short pulse followed by setting word line to V_{dd}. Then the data on the output of an inverter connected to one of the bit lines is read. A key design issue with the precharge approach is the timing of the precharge pulse and the activation of the word line. If the word line is set high before the precharge is off, SRAM cells along the word line in the array flip the state of the cell (Ref. 1a, p. 569).

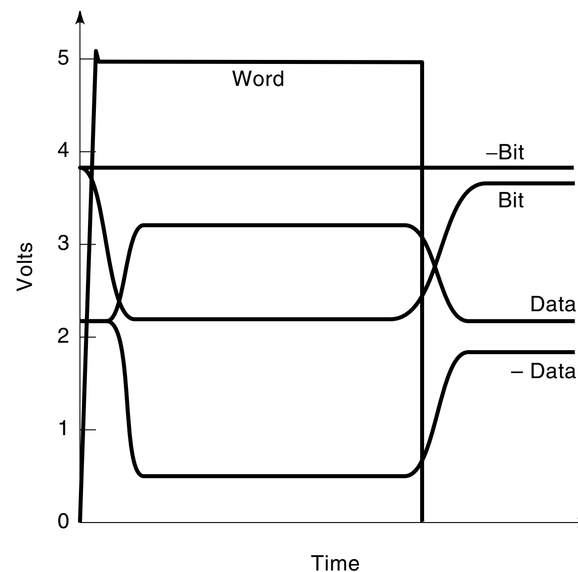


Figure 12. Simple SRAM read waveforms of the circuit in Fig. 5. The word and bit lines and the key sense amp nodes (DATA and −DATA) waveforms are given. Note the duration of the word-line voltage over the entire read operation for the static read configuration.

Write operation occurs by pulling one bit line low and leaving the complementary bit line at its high level. This forces the state of the cell when the access transistors T1 and T2 (see Fig. 5) are turned on. Figure 13 shows a basic circuit used to write data where the write access transistors T9 and T10 are turned on to set the bit lines to the desired values. Then the word line is set high to turn on the access transistors T1 and T2. To write a one to the cell the −bit line is pulled low whereas the bit line is left at $V_{dd} - V_{t,T4}$ which sets node A low and node B high.

SRAMs can be operated in the common asynchronous mode where no external clock is required, and thus the circuit design is simplified. For faster SRAM operation the synchronous or clocked mode can be implemented in the design at the expense of more complex circuitry. This can be done by adding latches to the input. Address transition detection (ATD) circuits are used to provide the initial pulse so that asynchro-

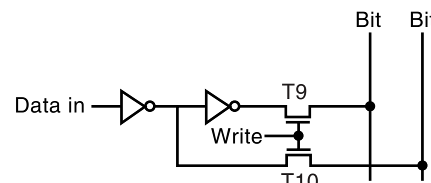


Figure 13. Basic write configuration circuit model. Reprinted after Ref. 1b by permission of John Wiley & Sons Inc. © 1991.

nous SRAMs can be operated as if synchronous (Ref. 1b, p. 394). This pulse is generated when one or more of the inputs (such as addresses or chip selects) have changed. It acts as an original clock for subsequent internal clocks. Use of ATD methods results in higher speed and lower power in asynchronous SRAM (12c).

Another method of improving access time is known as pipelining (12d,12e). Essentially, the circuit is divided into pipe segments, which are input-triggered and self-resetting circuit blocks. The read or write operation cycles through each block. After it has gone through the first block, it enters the second block while at the same time the next read or write operation can begin in the first block. In this manner, the access time was nearly cut in half, as demonstrated by Chappell et al. (12e).

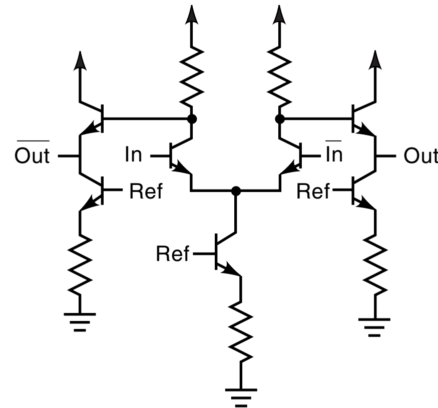


Figure 14. Differential amplifier utilizing bipolar junction transistors.

The Sense Amplifier

The sense amplifier circuitry noted in Fig. 1 is critical in achieving fast access times for high-performance SRAMs. As stated above, its function is to amplify the difference between the signals on the two complementary bit lines during a read operation while the memory cell drives the capacitance of the bit lines. Thus, the gain of the amplifier is a key metric for fast sensing. A simple CMOS inverter such as the one shown in Fig. 6(b) can be used if low power is required at the expense of speed due to low gain (Ref. 1a, p. 579; Ref. 1b, p. 162). A more common CMOS sense amp is the differential sense amplifier such as the one shown in Fig. 5. The common mode rejection ratio (CMRR) is used as a performance metric and is defined as *differential gain / common mode gain*. Common mode gain is the voltage gain when the inputs to both NMOS devices (Bit and −Bit in Fig. 5). With an ideal current source the gain would be unity. The differential gain is the voltage gain of the amplifier for different input voltages. The larger that the CMRR is then the better the ability of the differential amplifier to resolve the differential mode signal over the common mode signal. If one of the NMOS drivers and the two PMOS active loads have matched transconductances then (Ref. 12a, p. 444)

$$\text{CMRR} \approx \frac{1 + 2g_{\text{md}}r_{\text{dss}}}{2} \tag{27}$$

where g_{md} is the transconductance of the other NMOS driver and r_{dss} is the drain-to-source resistance of the current source device.

Many SRAM manufacturers have implemented a BiCMOS technology in order to take advantage of the higher gain bipolar junction transistor (BJT) for the peripheral circuitry requiring high gain and high fanout (e.g., decoders and sense amps) at the expense of higher power dissipation (45,46). A simple BJT differential amplifier utilizes emitter coupled logic (ECL) and would have the same layout as the CMOS differential amplifier, except that the NMOS transistors would be NPN BJTs. The PMOS active loads would be resistors or active load PNP BJTs. For an active load bipolar differential amplifier of the ECL design (Ref. 12a, p. 449)

$$\text{CMRR} \approx \frac{1 + 2g_{\text{md}}r_{\text{os}}}{4} \tag{28}$$

where g_{md} is the transconductance of the other *npn* driver and r_{os} is the output resistance of the current source device. For the ideal case where the common mode gain is unity, CMRR would reduce to $V_A/V_{gs}–V_t$ and $V_A/2V_T$ for the CMOS and BJT differential amplifiers, respectively (Ref. 3, pp. 519, 532). V_A is the early voltage and is on the order of 30 V to 200 V for CMOS and 50 V to 100 V for BJTs. V_{gs} is the gate-to-source voltage, which would be at or near V_{dd}, and V_t is the threshold voltage of the driver (assumed to be the same as the PMOS active load). V_T is kT/q, which is 25 mV at room temperature. Typical values for V_A, V_{dd}, and V_t for the CMOS amplifier would give CMRR values of 10 to 100 in this very simplified estimation. In contrast, the range of CMRR values for the bipolar amplifier would be 1000 to 2000, which is more than an order of magnitude better than the CMOS. An example of a resistive load BJT differential sense amplifier is shown in Fig. 14.

An alternate family of sense amps are the current sense amps that are designed to amplify the change in current between the two bit lines during the read operation (39,48,49). These sense amps are low-impedance circuits and thus the *RC* delay in driving the bit lines may be decreased (Ref. 1a, p. 572). A conventional current-mirror, as shown in Fig. 15

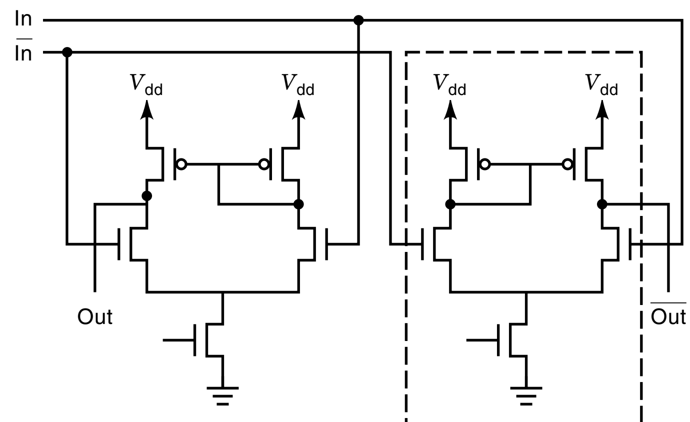


Figure 15. Conventional current sense amp mirror. The simplified CMOS sense amp is shown in the dashed box. Reprinted after Ref. 50 by permission of IEEE (© 1990 IEEE).

(50), has a reasonable gain over a wide input voltage range. A 20% increase in the gain of the current sense mirror was achieved by adding additional PMOS devices in parallel with each PMOS active load of the convention current mirror (50). Another high-performance CMOS sense amp is the latched sense amp, which has been shown to be capable of identifying voltage swings as low as 10 mV (47), resulting in bit-line delay times of 0.3 ns from selection of the word line.

IMPROVING TODAY'S SRAM CELL COST AND PERFORMANCE

To develop faster and larger memory arrays, MOS IC minimum feature sizes continue to shrink, including transistor gate lengths as seen in the trends in Fig. 16 (13,18,43). However, the downside of shrinking gate lengths is that lower power supply voltages are required to minimize hot carrier degradation of the gate oxide. This reduces the noise margin, making cell stability more critical. In addition, the lower V_{dd} increases the propagation delay as shown in Eqs. (22) and (23) and in Fig. 9 (4). This section addresses the various process and design problems confronting high-performance and cost-effective SRAM cell as IC dimensions shrink.

The earliest memories utilized bipolar devices (Ref. 10, p. 17) which delivered high-speed memory but had high power dissipation and low memory density. The first MOS SRAM was built from PMOS (Ref. 9, p. 579). Depletion load NMOS became popular once enhancement mode NMOS could be reliably built. The transition from depletion load NMOS to resistive load NMOS occurred between the 1 kilobit (kb) to 4 kb array size because of better noise margin and the reduced standby power which could be obtained with high resistive loads. Resistive load CMOS became more popular between the 16 kb and 64 kb memory size because of its lower standby power. Resistive loads were used primarily for high-density arrays whereas full CMOS was used for embedded memory in logic applications. However, poly load resistors require high resistance for low standby current but not too high in order to maintain a minimum current to keep the storage node charged. Production of BiCMOS began around the 256 kb to 1 Mb array size for high-speed applications while minimizing static power loss (Ref. 9, p. 579).

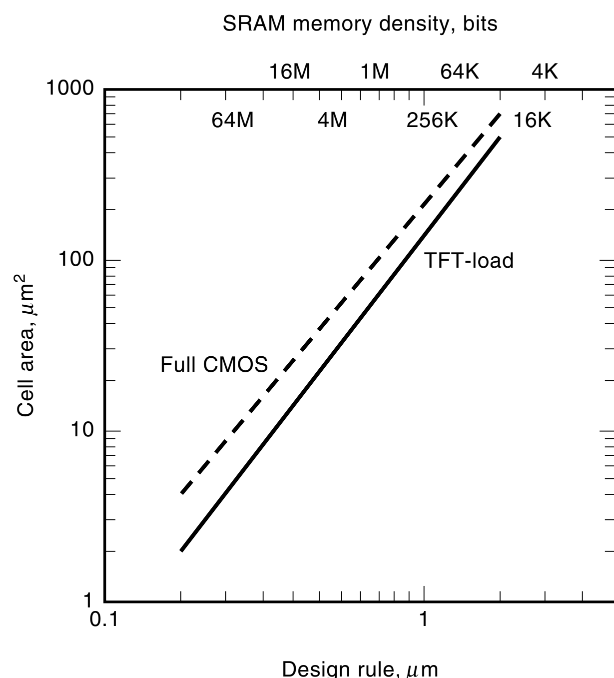


Figure 16. Cell area comparison between pTFT and bulk PMOS load inverters as a function of minimum design rule. Reprinted after Ref. 18 and 43 by permission of IEEE (© 1996, 1993 IEEE).

The R-load cells become more problematic as the power supply drops which has led to a wider use of pTFTs (6,7). The pTFT cell size is reduced by more than a factor of 2 compared with the standard bulk PMOS technology, as shown in Fig. 16. TFTs must achieve a high I_{on} to keep the storage node (node B in Fig. 5) charged while maintaining a low I_{off} for low standby power. High-performance pTFT technologies with a high I_{on}/I_{off} are required and must be able to operate at lower power supply voltages as the cell areas continue to decrease. Design modifications, such as split word lines are being introduced to minimize cell size and increase cell stability.

High cell stability, low power consumption, low process complexity (for low cost), small cell size, and high memory speed are the key objectives to consider as the device size continues to shrink. The competing technologies against which these objectives are weighed are the advanced six transistor (6T) and the four transistor (4T) cells. The 6T cells are composed of bulk Si PMOS loads and are mainly used for on-chip microprocessor or other logic circuits because of the lower power dissipation. The much smaller 4T cells are composed of NMOS drivers with poly resistor or pTFT loads and dominate the stand-alone market because of their smaller cell size. The 4T cell is less susceptible to latch-up because the memory array is made up only of bulk NMOS devices. The 4T cell suffers, however, from poorer cell stability, higher soft error rates (SER), and higher cost.

4T vs 6T: Cell Stability

Two disadvantages of the 4T cell are the cost (due to process complexity) and the inferior cell stability (particularly at lower V_{dd}). An analysis by C. Lage et al. (14) indicates that the 4T cell has a higher process complexity and cost [$\geq$22 mask steps compared to $\approx$18 for full CMOS (15)]. The larger the array, however, and the higher the yield, then the more cost-effective the complex processes become. They also point out, however, that the more complex the process is, then the less compatible it is with more standard CMOS logic processes that a company may be running on the same manufacturing line (14). The advanced 6T cell incorporates a number of technological improvements compared with the simple 6T cell, including trench isolation (16,17), self aligned contacts (14,18), optical proximity effect correction (18), and local interconnect (14).

Data stability is of primary concern (especially as V_{dd} is reduced) and depends on the cell ratio, data leakage, and soft errors (13,19). If the cell is overly sensitive to switching or noise, then data is lost during reads (19). Memory cell stability is measured by the static noise margin (SNM) which is understood by considering the circuit diagram of the SRAM cell during read operation (refer to Fig. 5). The SNM is defined as the critical value of an equivalent static noise margin voltage source V_{snm} (placed between the input of each inverter

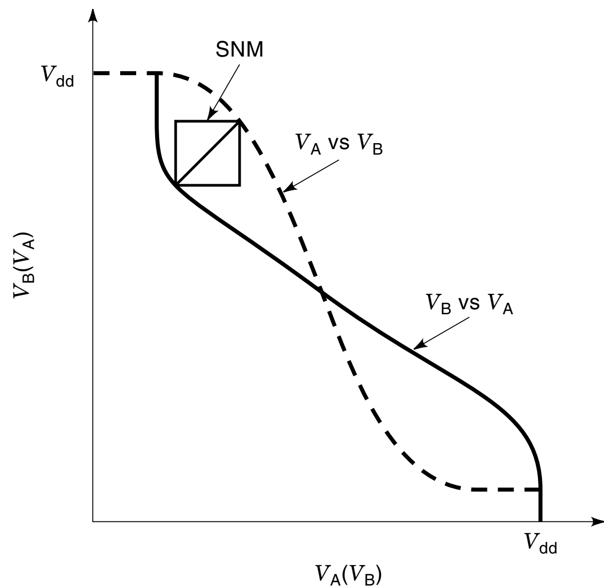


Figure 17. Extraction of SNM from the voltage transfer curves at the SRAM storage nodes A and B in Fig. 5. V_A as a function of V_B is denoted by the dashed line whereas V_B vs V_A is denoted by the solid line.

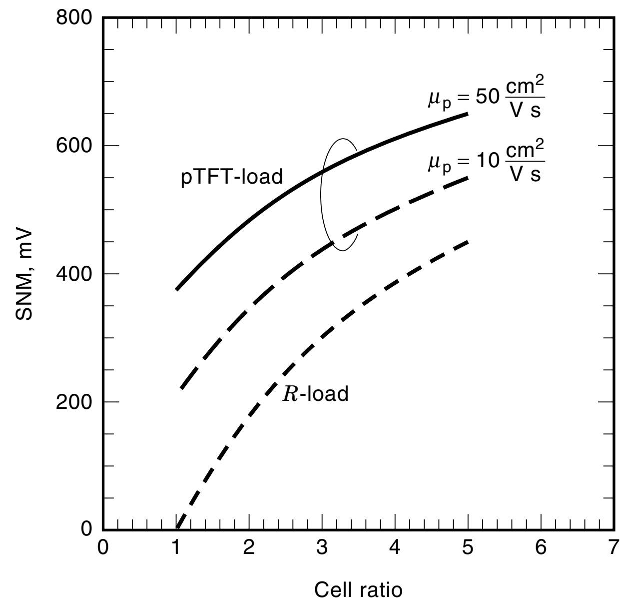


Figure 18. Calculated SNMs as a function of cell ratio for pTFT load and R-load inverters. The pTFT load shows results for two different hole mobilities which reflects drive current capability. Reprinted after Ref. 8 by permission of IEEE (© 1992 IEEE).

and node A or B) above which the two stable states of Node A and Node B change. Figure 17 shows SNM is obtained from the voltage difference between the transfer curves of the storage nodes (A and B in Fig. 5) of the SRAM cell.

The cell ratio significantly affects the SNM and is defined as the ratio of the size of the driver transistor (T5 or T6 in Fig. 5) to the size of the neighboring access transistor (T1 or T2) or

$$\beta = \frac{(W/L)_D}{(W/L)_A} \tag{29}$$

Calculated SNMs from the poly resistor load and pTFT load cells are compared for various cell ratios in Fig. 18 for V_{dd} = 3 V and gate lengths of 2 μm (8). This figure indicates that the SNM improves with increased cell ratio and is better for pTFT loads compared with resistive loads because of the higher load currents. Figure 18 also shows that the SNM increases as the drive current I_{on} of the pTFT load device increases [with increasing carrier mobility μ_p, see Eq. (8)] (8,15,20). The relationship between the SNM and the minimum power supply voltage $V_{dd,min}$ necessary to maintain SRAM operation (21,22) is provided in Fig. 19. This result combined with that of Fig. 18 indicates that, to achieve the smaller $V_{dd,min}$ necessary for smaller geometries, the cell ratio must increase, causing an increase in cell area. Simulated results given by Yuzuriha et al. in Fig. 19 (22) show that a cell ratio of about 3 is required for 3 V operation. In an ideal SRAM cell (23),

$$V_{dd,min} = (1 + \gamma_a)V_{td} + V_{ta} \tag{30}$$

where V_{td} and V_{ta} are the driver and access transistor threshold voltages, respectively, and γ_a is the access transistor body effect coefficient (because its source is floating). Thus, $V_{dd,min}$ increases as the V_t's increase. A comparison of simulated values of SNM for 6T and 4T cells versus supply voltage and cell ratio shows that the 6T cells have superior performance at lower voltages, particularly below 3 V, as shown in Fig. 20 (14). In fact, a recent paper demonstrated an SNM > 500 mV at V_{dd} = 2.5 V and SRAM functionality down to 0.6 V for a full CMOS 6T cell (18). These results emphasize the need for

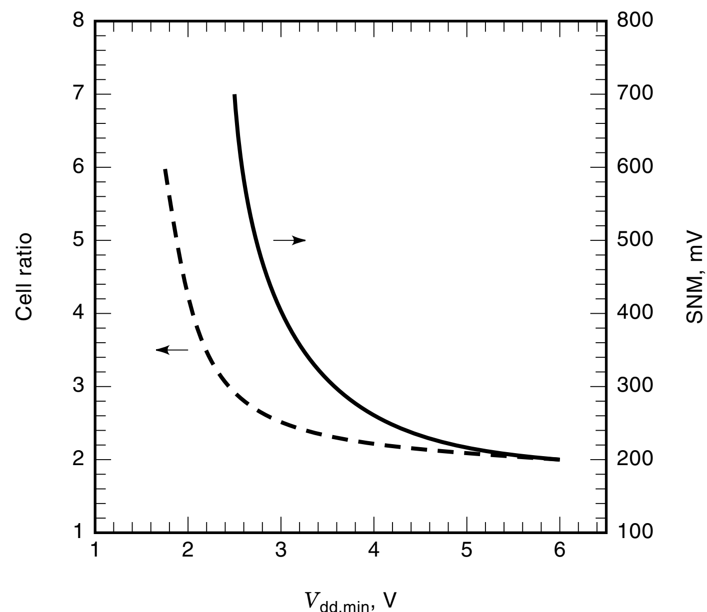


Figure 19. Relationship between Cell ratio and SNM with the minimum power supply voltage necessary for SRAM operation. The solid line reflects the SNM curve whereas the dashed line denotes the cell ratio curve. Reprinted after Refs. 21 and 22 by permission of IEEE (© 1993, 1991 IEEE).

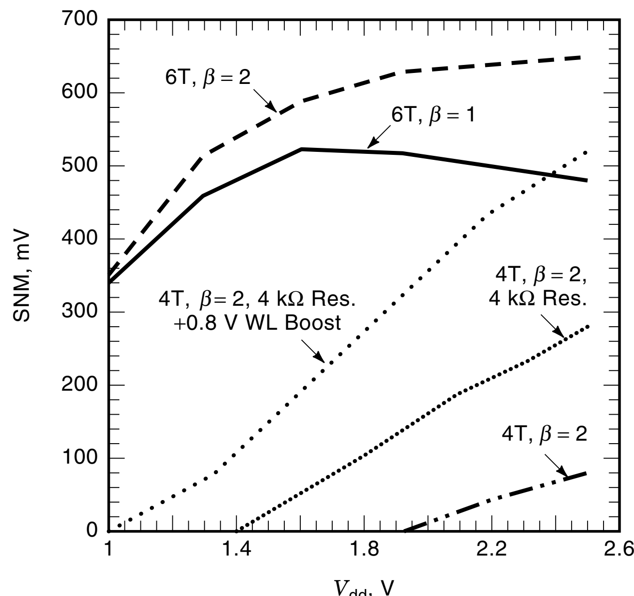


Figure 20. Simulation comparing various 6T and 4T SRAM cell SNM as a function of power supply voltage. The 6T cells shown have cell ratios of 1 and 2. All of the 4T cells have a cell ratio of 2 and include the enhancements of an N-resistor with and without a 0.8 V word-line boost. Reprinted after Ref. 14 by permission of IEEE (© 1996 IEEE).

high I_{on} in the load device. However, the pTFT size cannot be increased arbitrarily to obtain a larger I_{on} because this increases both cell area and I_{off}. A higher I_{off} increases standby power. In fact, the requirement of less than 1 μA for 16 Mb SRAM dictates that I_{off} must be less than 60 fA per pTFT and $I_{on}/I_{off} > 10^8$ (13,22).

A number of demonstrated improvements make the pTFT more robust with a larger I_{on} for improved cell stability. One method for increasing the drive current of the pTFT without taking up more cell area is putting the gate electrode on both the top and the bottom of the poly silicon TFT channel (24–27). The poly 4 layer shown in Fig. 8 is added to form a double-gated TFT (DGT). With this design, the effective area of the device is increased thereby increasing the current drive because $I \approx W/L$. An extension of the DGT is the gate all around TFT (GAT) which essentially is the DGT with sidewall transistor action as well (24). The gate poly Si surrounding the channel poly Si simultaneously, and thus no contact holes for the bottom gate are required, saving a masking step compared to the DGT. Maegawa et al. (20,24) demonstrated a reduction in I_{off} of nearly two orders of magnitude at a drain voltage of -3 V and an increase in I_{on} of a factor of 2 for the GAT compared with the single-gate TFT (SGT). The SNM was about 300 mV and $I_{on}/I_{off} > 10^7$ for a 0.4 um channel length pTFT.

One method for reducing I_{off} is using a resistor in the lightly doped drain (LDD) region (28), as highlighted in Fig. 8, where the lighter *p*-type poly is formed between the pTFT channel and the P+ source/drain. As the LDD dose is decreased, I_{off} decreases (28) but the maximum drive current is also reduced (28,29). AT&T demonstrated an $I_{on}/I_{off} = 4 \times 10^5$ at -3.3 V with a cell dimension of 0.35 μm × 0.35 μm (28). The I_{on}/I_{off} is improved by increasing the channel length (30,31) because I_{off} is significantly reduced, as demonstrated in Ref. 28, where $I_{on}/I_{off} > 10^8$ for a channel length of 0.7 um. Modifications have been made to improve the LDD TFT characteristics, such as forming an N^- offset resistor between the access transistor and the storage node of the cell by blocking the N^+ poly implant in this region. With a resistance of 10 kΩ, the SNM improved from 100 mV to 270 mV (at a V_{dd} of 3 V) and $V_{dd,min}$ improved from 2 V to 3.6 V (for a cell ratio of 3) (28). The additional resistance, however, decreases the ON current of the access transistor by 30% and thus increases the effective cell ratio of the SRAM cell (28). The 6T cell is the approach of choice for operating voltages ≤1.5 V on the basis of the cell stability requirements demonstrated in the simulations in Fig. 20 (18). This figure compares the 6T cell with cell ratios of 1 and 2 with 4T cells with the LDD resistor, a word-line boost, and a cell ratio of 2. The word-line boost incorporates pushing the word-line voltage above the power supply voltage to increase the voltage written into the cell.

Recently, two new 4T cell structures maintain 6T-like cell stability with the advantages of the smaller cell area afforded by the TFT load. One cell implements PMOS drivers and nTFT loads (31). The nTFTs show superior device characteristics compared with the pTFTs and do not require an LDD layer, saving a mask step. This cell operates at 1.2 V because of improved cell stability with the nTFTs and has an I_{on}/I_{off} of 5×10^5 compared with 1×10^4 for the a pTFT with LDD. Another 4T pTFT structure presented by Mitsubishi (29) replaces the dual bit-line bulk NMOS access transistors by a single bit line with a pTFT in parallel with a bulk NMOS device for the access transistors. The word line is connected to the gate of the bulk NMOS access transistor, and the complementary word line is connected to the gate of the pTFT access device. The GAT is used for improved pTFT performance. The overall cell size is reduced by 84% under conventional 0.3 um design rules compared to the SGT with two bulk access transistors and two bit lines.

4T vs 6T: Soft Error Rate

The soft error rate (SER) is another problem which must be addressed as SRAM densities continue to increase (32). A hard error is defined as a location in the memory array which always fails to output the data previously written to it and is often caused by physical defects which occur during processing. Conversely, soft errors are single nonrecurring errors in the array which are not caused by process defects. Rather, they are circuit induced by power supply noise, inadequate noise margin, or sense amplifier imbalance (Ref. 9, p. 616). May and Woods (33) identify a major source of soft errors in DRAM as alpha particles originating from the decay of trace uranium and thorium in IC packages. The SER results from electron-hole pair generation by the ionizing radiation which charges up the DRAM capacitors. The problem increases for smaller geometries because of the smaller capacitances and hence, less charge needed to cause an upset.

SRAM soft errors occur (32), when the voltage drop in the storage node of the cell that is induced by the impinging alpha particles is not compensated for in time by the current supply to the node (6). Thus, high-speed operation of the SRAM has a higher soft error rate. A detailed treatment of this issue shows that the more likely source of soft errors is cosmic ray events rather than alpha particles (34). The best

protection against soft errors is to maintain sufficient stored charge in the cell to compensate for cosmic rays (34). Because the SRAM cell area continues to decrease in size, the amount of stored charge in the cell continues to decrease (19), which leads to a higher SER. Figure 21 shows that the capacitance per cell and critical charge necessary to cause a soft error continue to get smaller for larger SRAM arrays (19,34). The issues becomes particularly acute as the operating voltages decrease because the charge on a capacitor is directly proportional to the voltage (recall $Q = CV$). A reduction in SER results when I_{on} of the load and the stored capacitance in the cell are increased.

The SER in resistor load cells is problematic because of the low currents. This led to the development of the higher current drive pTFTs (6,29,35,36). Thus, the entire discussion regarding the need to increase I_{on} in pTFTs applies directly to SER reduction because a higher drive current allows faster recharging of the storage node during high-speed operation (35,37). 6T CMOS SRAM cells have a much better SER compared with 4T cells for comparable geometry because of the higher I_{on} for bulk PMOS pull-up transistors and the higher cell capacitance of the larger cell area (13).

Another method for reducing SER in both 4T and 6T cells is to introduce additional cell capacitance without increasing cell size. A number of options have been developed to meet this objective. The bottom gate pTFT structure in Fig. 8 (6) has additional cell capacitance (~5 fF) from the cross-coupled capacitors from the bottom gate of the TFT (poly 2) to the channel poly (poly 3) of the TFT and the access gate poly (poly 1). Another cell design utilizes a top gate TFT and implements a V_{dd} plate stacked over the TFT to form a cell node capacitor (21). The "fin" capacitor typically used for DRAM is implemented so that each fin adds additional capacitance in the range of 15 fF to 20 fF (19). Motorola added an oxide-nitride-oxide (ONO) capacitor between the third poly layer (which functions as a resistor load) and an added fourth poly layer (34).

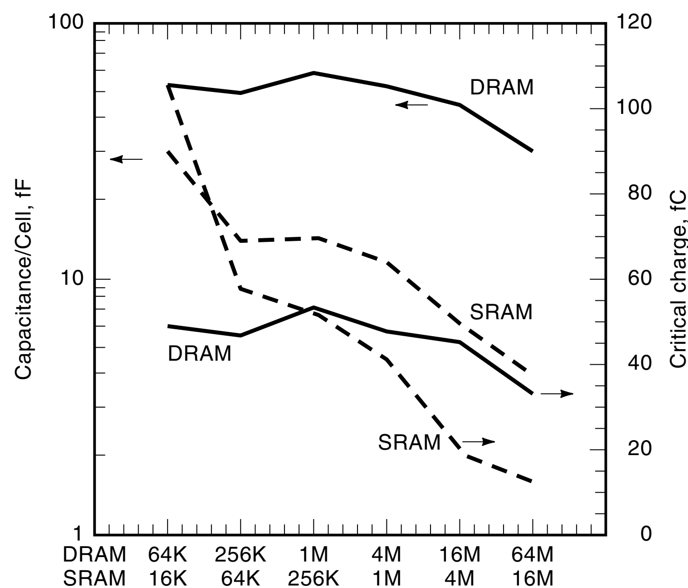


Figure 21. Comparison of capacitance per cell and critical charge necessary to avoid a soft error between SRAM and DRAM at various array densities. The DRAM and SRAM curves are denoted by the solid and dashed lines, respectively. Reprinted after Ref. 34 by permission of IEEE (© 1991 IEEE).

By comparison, SOI has a much lower SER compared with bulk Si because the radiation hardness is much better (37).

Split Word Line

Higher cell stability for a given cell ratio is realized by making the cell more symmetric. One method of accomplishing this is to use a split word line (19,21,30,38,39,43). A layout comparing the conventional and the split word-line (SWL) cells is given in Fig. 22 showing the active area and first poly Si layer defining the bulk driver and access transistors (19). The conventional cell has the access transistors on one side and the drivers on the other, which can negatively affect the SNM and $V_{dd,min}$ because of the lack of symmetry. Adding an additional word line and reorienting the location of the bulk transistors makes the cell symmetric and produces a well balanced flip-flop function (30). This symmetric layout improves cell stability over a wider range of process variation (38), relative to mask alignment between the active and the first poly Si, and reduces threshold voltage variation. The SWL cell keeps the effective transistor width stable by reducing lateral expansion of the field oxide bird's beak, minimizing variation in the driver transistor current (19,39). The stability is improved and cell size is reduced in a 16 Mb SRAM by stacking the split word lines over the pull-down transistors (21). Another implementation of cell symmetrization to improve cell stability was accomplished in a 64 Mb SRAM by designing the word line through the center of the cell. At 2.5 V this smaller cell has an I_{on}/I_{off} near 5×10^7 and a $V_{dd,min}$ of 1.7 V.

Summary

The choice of 4T vs 6T cell architecture, BiCMOS, bipolar, SOI, or the various design options mentioned previously depends on a number of engineering tradeoffs between cell area, operating speed, design complexity, chip area, process cost, cell stability, power, and SER. The end user/customer defines performance requirements that the SRAM manufacturer must meet as cost effectively as possible to be profitable. The decision on cell architecture depends on which of these items is most critical for the application and is the least expensive technology to meet customer objectives.

APPLICATION-SPECIFIC SRAMSs

SRAM arrays with logic circuitry designed for a specific task are referred to as application-specific SRAMs (ASSRAM). A number of ASSRAMs are on the market as shown in the summary in Table 3 (Ref. 10, pp. 35 and 75). One interesting application is the nonvolatile SRAM (specifically "Shadow RAM" or NVSRAM) which combines SRAM with electrically erasable programmable read only memory (EEPROM). NVSRAM is useful for memory applications which require critical data storage that will not be lost if the power supply drops below the necessary operating voltage of the SRAM. This application combines the high speed of SRAM with the nonvolatile memory capability of the EEPROM. SRAM requires a minimum power supply to remain functional, but EEPROM does not. However, EEPROM technologies have a limited number

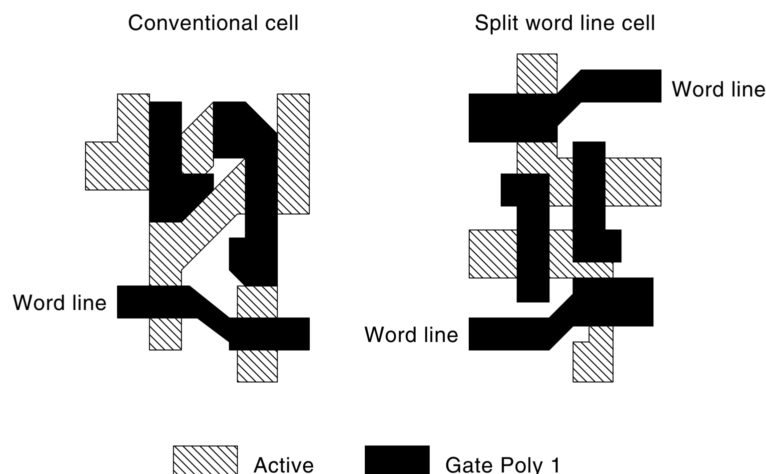


Figure 22. Active and poly 1 cell layouts for the conventional single word-line and the split word-line cells. The active and poly 1 areas are denoted by the dashed and solid regions, respectively. Reprinted after Ref. 19 by permission of IEEE (© 1991 IEEE).

of write/erase cycles, and they are very slow (on the order of milliseconds not nanoseconds as is the case for SRAM). The NVSRAM design consists of a corresponding "shadow" EEPROM cell for each SRAM cell. Logic circuitry is designed to transfer data from the SRAM to the EEPROM when the operating voltage drops below a predetermined threshold voltage set just above the operating voltage necessary for the SRAM to function. The SRAM array is again utilized once this threshold voltage is exceeded by reconnecting power or replacing the batteries. Sixty-four kb to 256 kb memory sizes are currently available (Ref. 10, p. 121).

Another ASSRAM on the market is the content-addressable memory (CAM), otherwise called an associative memory (Ref. 1b, p. 483). The function of this chip is to examine a word of data presented at its input and compare it to data stored internally in the CAM. If there is a match between the two, then a signal is passed to a RAM to enable a specific data word at the output (Ref. 1a, p. 589; Ref. 10, p. 39). Applications using CAMs include database management, disk caching (virtual memory), artificial intelligence, pattern and image recognition, and networks (e.g., ethernet) filtering (Ref. 1b, p. 484; Ref. 10, p. 39). The CAM is rarely used as a stand-alone memory because it requires additional input pins as well as having at least 9 transistors per cell (compared to 6 for the standard CMOS SRAM cell), resulting in a larger array.

Table 3. Listing of Commonly Used ASSRAMs[a]

Acronym	Description
SBSRAM	Synchronous burst static RAM
SRAM	Static RAM
SSRAM	Synchronous static RAM (clocked)
FIFO	First-in-first-out serially accessed memory (line buffers)
Dual-Port RAM	Two independent devices have simultaneous read-write access
NVSRAM	Nonvolatile SRAMs (EEPROM and SRAM combined)
CAM	Content-addressable memory (associates an address with data)
Cache TAG RAM	Like CAM, stores TAG
BRAM	Battery backed-up SRAM

[a] Reprinted after p. 75 of Ref. 10 by permission of IEEE (© 1997 IEEE); (Ref. 1b, pp. 478–483).

SRAM TESTING AND RELIABILITY

Testing and Redundancy

SRAM chips, like any other manufactured product, are tested following processing to ensure proper functionality before being shipped for packaging and sale. The key issues in developing an adequate test program is balancing complete test coverage with total test time. As SRAM density and circuit complexity increase, more test vectors are necessary to ensure proper operation in the application. With more tests and more SRAM cells to test, the test time can become a bottleneck for manufacturing cycle time. An example of this is given by Prince in comparing the test time for two different size arrays. A simple array test taking approximately 30 min on a 64k SRAM would take over 6 h using the same test on a 1M array (Ref. 1b, p. 703). The testing becomes more difficult for embedded memories, which require test algorithms that incorporate both the microprocessor and the SRAM.

A number of failure/fault models have been derived to assist in test development by directing tests to look for such faults. A detailed list of the fault modes with lengthy explanation of each type can be found in Chap. 4 of Ref. 10. A brief list will be given here by way of summary in order to provide a sample of the type of fault models that exist.

Stuck at fault (SAF): A cell or line always stuck at a "1" or "0" which can't be changed. For N distinct logic lines there are $2N$ possible single SAFs. These can be found by testing each cell

Bridging fault (BF): A short between two or more cells or lines. This type of fault can be more difficult to detect since it does not remain stuck

Stuck open fault (SOF): An open circuit word line

Data retention fault (DRF):	Due to loss in cell data over time caused by parasitics such as leakage
Pattern sensitive fault (PSF):	The contents of the cell are influenced by the contents of other cells in the array
Dynamic fault (DF):	An electrical fault that is time dependent and is internal to the chip. This type of fault is typically observed in sense amp or write recovery, retention faults, etc.

A March test can be used to detect SAFs, BFs, and CFs. This test marches from the lowest address bit to the highest. The simple March test writes and then reads a "0" and "1" in succession. The order of the march can be varied to determine the type of the fault. There are many different array tests in addition to the March test, as can be found in Chapters 4 and 5 of Ref. 10. The more thorough the test the longer the test time.

The typical SRAM test program can be divided into two categories: dc and ac testing (Ref. 10, p. 158). Dc tests include quiescent (static) and operating supply currents, output voltage tests (V_{oh}, V_{ol}) to measure the high and low voltages on outputs when driving a load, input current tests (I_{ih}, I_{il}) to measure the amount of current drawn under a specified high and low voltage, and input/output pin leakage (I_{ilk}, I_{olk}). Ac tests include voltage and current data retention, read cycle time, address access time, chip select/enable times, output hold/enable times, address setup and hold, write pulse width and other key timing related tests.

Many SRAM manufacturers implement built-in self-tests (BIST), which are tests that provide the capability of the chip or circuit to test itself. On-line concurrent BIST is performed simultaneously during normal functional operation, whereas on-line nonconcurrent BIST is performed while the chip is idle. Off-line BIST is a test mode performed when the chip is not in operation. Off-line BIST does not detect errors in real time as can be done with on-line BIST. There are two general circuit approaches in implementing BIST: random logic and microcoded ROM. Circuit complexity, speed, and chip area are some of the items to be considered during design. Error-correcting codes (ECC) in the chip are used to correct both hard and soft errors in the array by utilizing parity bits to detect bit errors in the array. The penalty for more extensive ECC techniques is chip area.

Redundant cells can be implemented in SRAM designs by adding extra rows and columns to replace those in the array which have bad bits found at wafer sort. Redundancy can be achieved by current blown fuses, laser blown fuses, or laser annealed resistor connections (Ref. 1b, p. 127). The use of redundancy is to assist in yield improvement in the early stages of manufacturing a new SRAM technology. As the technology matures, the yields will reach a point where it is more cost-effective to reduce the total chip area by eliminating the redundant cells.

Reliability

It is one thing to build a memory array with high yield and performance coming out of the manufacturing line but quite another to guarantee the same performance over a long period of use. There are a number of wear-out mechanisms which can lead to the failure of a part. This section addresses these failures and others and discusses process and/or design improvements implemented to prevent early failure. Generally, reliability is compromised by the physical integrity of the multilayer chip, the presence of contaminants which erode physical features or shift device characteristics or change in the electrical behavior of the circuit elements.

Aluminum (Al) is the predominant metal used in ICs to carry current and suffers from electromigration at higher current densities over time. Essentially, ionized atoms in the metal collide with the electron current and are scattered in the direction opposite to the current flow. As a result, voids form which increase the resistance, completely severing the metal line, or causing shorts to underlying or overlying metal layers. Alloying Cu in the Al lines and placing a maximum current limit during the design has greatly minimized this problem. Corrosion can result from in-line process contamination, air exposure, or from penetration of halide ions during the packaging process. Another metal-related reliability problem is junction "spiking" of the contacts which results from Si diffusing from the source or drain junction into the metal line at the contact. The voided Si is replaced by Al from the metal, resulting in extremely high junction leakages. The alloying of the Al with Si and the use of TiN or TiW for a metal barrier has minimized this problem. However, if care is not taken to develop a robust barrier process, the barrier wears out or develops cracks at high stress points through which junction spiking occurs. Metal lines are also sheared off by high film stresses of the dielectrics which sandwich the metals or are placed over the metals, such as a high compressively stressed Si_3N_4 film on metal to function as a passivation. In addition, thermal stresses result in metal line cracking or shearing which leads to functional failure over time if the cracking does not completely sever the line upon packaging.

The top passivation layer provides mechanical protection for the chip and also protects against penetration of moisture and other contaminants which degrade chip performance over time. Si_3N_4 is the most commonly used dielectric because of its very good barrier properties. However, if pinholes are formed at deposition, the die passes testing after packaging but fails over time as moisture or other contaminants diffuse into the chip. The passivation layer also delaminates from the underlying layers because of stress or contaminants, and takes some of the metal with it. Dielectric delamination also occurs in the interlayer dielectrics (poly to poly, poly to metal, and/or metal to metal). The resulting voids are stress points which may crack or shear neighboring metal lines, particularly if the die temperature increases, increasing the pressure in the void.

As mentioned previously, it is important to keep ion contaminants, particularly alkalis, out of the chip. As the temperature increases, the alkali ions (e.g., Na or K) diffuse to the gate oxide and shift the threshold voltage of the transistors or degrade the oxide over time, which leads to functional failures.

Gate oxide integrity has been given a great deal of attention because it shifts device characteristics over time or causes device failure. Thin gate oxide integrity is also compromised during processing because of the large number of process steps which utilize ion/electron plasmas. If regions of the

chip have a large ratio of conductor area (poly or metal) over field oxide relative to the thin gate oxide area, then charge buildup during plasma processing (e.g., etches or low temperature dielectric deposition) or ion implantation dissipates through the thin gate regions. As a result, the oxide is not so damaged that it does not work coming out of fabrication but could be damaged enough that additional stresses during normal chip operation result in oxide failure or shifting of transistor parametrics, leading to SRAM failure. Charge damage also results from high current implants, such as source and drain implants. Processes must be carefully monitored to minimize the extent of charge damage during processing.

Additional reliability hazards in the Si include latch-up, electrostatic discharge (ESD), and electrical overstress (EOS). Latch-up occurs when bias conditions on a CMOS chip are such that bipolar action occurs between the source and well of one device with that of another. A positive feedback loop forms and the current increases until the devices lock up. One way to avoid this is to increase the P+ to N+ spacing requirements and/or implement a guard ring around the corresponding well and diffusion which both increase SRAM array size. Another is to tailor the well profiles to minimize the gain of the parasitic bipolar device which can be done with retrograde wells that utilize high-energy ion implementation. Trench isolation or SOI between the diffusions both minimize latch-up. Latch-up is a concern for CMOS SRAM technologies because the P+ to N+ spacing continues to shrink with the technologies to maintain the shrink in array size. ESD and EOS occur when excessive voltage or charge connects to the chip pins. Various input protections on the pads have been designed to withstand normal voltages that the packaged chip encounters during handling and packaging.

Finally, hot carrier injection (HCI) also leads to SRAM failures in the field. HCI occurs as the gate length and gate oxide thickness are scaled, resulting in an increase in the lateral electric field from the source to the drain. This, in turn, leads to the generation of a significant number of electron-hole pairs caused by the impact of high-energy electrons accelerated through the channel. Some of the current generated is injected into the gate oxide, degrading the oxide integrity and shifting transistor characteristics. The problem with this threshold shift is that it results in mismatching of V_t for the driver and access transistor if they are under different bias conditions on their respective drain, gate, and source. Equation (30) indicates that the operating voltage minimum will increase, leading to SRAM failure. In addition, HCI degrades the drive current which also leads to poor cell stability over time. As the operating voltages continue to decrease for smaller device sizes, the allowed shift in SRAM cell device parameters will narrow, making HCI a key concern for each new technology. The implementation of the lightly doped drain (LDD) technology and lower V_{dd} reduce but do not eliminate HCI. Other process enhancements, such as gate nitridation, are being developed to address this issue.

NOVEL SRAM CELL CONFIGURATIONS FOR FUTURE HIGH-SPEED/HIGH-DENSITY APPLICATIONS

This final section presents a few brief examples of novel SRAM structures which are being developed to improve cell performance, increase memory density, and/or decrease costs. The focus of current research is on those devices which exhibit at least two stable operating points to mimic the SRAM cell bistability first shown in Fig. 4. Three-element SRAM cell operation has been demonstrated with structures utilizing quantum wells formed by delta-doped (δ-doped) layers (e.g., thicknesses <100 Å and doping concentrations $>10^{19}$ cm^{-3}) or from heterostructures (e.g., GaAs/AlGaAs) (52–56) as the storage node. An access gate and load are required to complete the three-element cell. These cells were made, however, with III–V materials which are not readily compatible with the more mature processing and lower cost of silicon. In addition, these devices have very narrow noise margins and large power dissipation from the lack of a well-defined "off" state. Room temperature Si-based multistate quantum devices are significantly inferior to date compared with GaAs-based materials. A multistate Si-based device has been developed (57) with a very narrow noise margin (<1 V) and a high standby current.

Novel Si structures utilizing bipolar technology are also under development (58,59). WSI used the latch configuration with NMOS drivers and access gates with bipolar loads (58). Toshiba developed a cell which uses the reverse base current in an n-p-n bipolar device as the storage element (59). One of the more promising Si-based approaches to date utilizes a bistable SiGe diode with closely spaced p-type and n-type δ-doped layers in a SiGe layer (60). Distinct bistability was obtained as shown in the diode I–V curve in Fig. 23 where the ratio of the resistance in both stable states is over 3×10^6. Details of the physical operation of the bistable diode are found elsewhere (60). SRAM operation was demonstrated (61) with a V_{oh} of 3.3 V and V_{ol} of 1.0 V under an operating voltage of 3.5 V. An all-Si bistable device was recently fabricated and exhibited characteristics very similar to those shown in Fig.

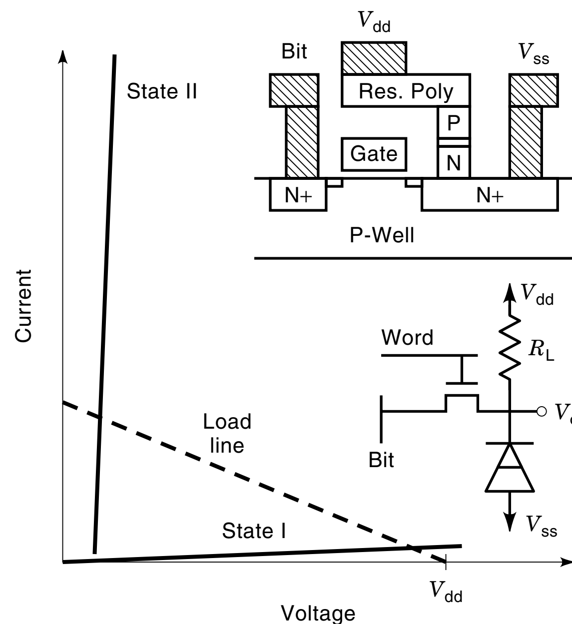


Figure 23. I–V curve for the bistable diode. The lower inset shows a proposed three-element SRAM cell utilizing the bistable diode as the storage element whose cross section is shown in the upper inset. The resistor load line is denoted by the dashed line. Reprinted after Ref. 61 by permission of IEEE (© 1995 IEEE).

23 (62). The all-Si structure is readily integrated into existing CMOS process flows because the growth occurs at temperatures less than 650°C. The cross section shown in Fig. 23 is one potential application of the bistable diode in the three-element SRAM cell which can be built in a very simple double-poly, double-metal process. The bistable diode is grown over the source of the access device and is contacted directly to the poly load resistor. The cell size can be made as small as current DRAM cells because the diode is a vertical element which can be built to the minimum design rule. It is estimated that the switching speed of the diode is on the order of picoseconds (61) and thus SRAM speed is only limited to external circuitry and the load. If the standby power of this structure is reduced and the manufacturing technology to deposit the thin layers with acceptable thickness variation is developed, this novel cell is one idea which may prove useful for future high-performance, high-density SRAM.

BIBLIOGRAPHY

1. T. Makimoto, Market and technology trends in the nomadic age, *1996 Symp. VLSI Tech. Dig. Tech. Papers,* 1996, pp. 6–9.

1a. N. Weste and K. Eshraghian, *Principles of CMOS VLSI Design.* 2nd ed., Reading MA: Addison-Wesley, 1993, p. 564.

1b. B. Prince, *Semiconductor Memories,* 2nd ed., New York: Wiley, 1991, p. 166.

2. D. Hodges and H. Jackson, *Analysis and Design of Digital Integrated Circuits,* New York: McGraw-Hill, 1983, p. 71.

3. A. Sedra and K. Smith, *Microelectric Circuits.* 2nd ed., New York: Holt, Rinehart and Winston, 1987, p. 860.

4. M. Norishima et al., High-performance 0.5 μm CMOS technology for logic LSIs with embedded large capacity SRAMs, *IEDM Tech. Dig.,* 1991, pp. 489–492.

5. D. K. Schroder, *Modular Series on Solid State Devices: Advanced MOS Devices,* Reading, MA: Addison-Wesley, 1990, p. 173.

6. T. Yamanaka et al., A 25 μm^2 new poly-Si PMOS load (PPL) SRAM cell having excellent soft error immunity, *IEDM Tech. Dig.,* 1988, pp. 48–51.

7. M. Ando, A 0.1 A standby current bouncing-noise-immune 1 Mb SRAM, *1988 Symp. VLSI Tech. Dig. Tech. Papers,* 1988, p. 49.

8. Y. Uemoto et al., A stacked-CMOS cell technology for high-density SRAM's, *IEEE Trans. Electron Devices,* **39**: 2359–2363, 1992.

9. S. Wolf, *Silicon Processing for the VLSI Era, vol 2: Process Integration,* Sunset Beach, CA: Lattice Press, 1990, p. 584.

10. A. Sharma, *Semiconductor Memories: Technology, Testing and Reliability,* Piscataway, NJ: IEEE Press, 1997, p. 20.

11. M. Ishida et al., Cornerless active area cell and Bi-T-MOS process for sub-half microm SRAM's, *1994 Symp. on VLSI Tech. Dig. of Tech. Papers,* 1994, pp. 103–104.

12. G. Shahidi et al., SOI for a 1-volt CMOS technology and application to a 512 Kb SRAM with 3.5 ns access time, *IEDM Tech. Dig.,* 1993, pp. 813–816.

12a. R. Geiger, P. Allen, and N. Strader, *VLSI Design Techniques for Analog and Digital Circuits,* New York: McGraw-Hill, 1990, p. 831.

12b. T. Hirose, A 20-ns 4 Mb CMOS SRAM with hierarchical word decoding architecture, *ISSCC Proc.,* 1990, p. 132.

12c. S. T. Flannagan et al., Two 13-ns 64k CMOS SRAMs with very low active power and improved asynchronous circuit techniques, *IEEE J. Solid-State Circuits,* **SC-21**: 692–703, 1986.

12d. K. O'Conner, A prototype 2k × 8b pipelined static RAM, *ISSCC Proc.,* 1989.

12e. T. Chappell et al., A 2-ns cycle, 3.8-ns access 512-kb CMOS ECL SRAM with a fully pipelined architecture, *IEEE J. Solid-State Circuits,* **26** (11): 1577–1585, 1991.

13. S. Flannagan, Future technology trends for static RAMS, *IEDM Tech. Dig.,* 1988, pp. 40–43.

14. C. Lage, J. Hayden, and D. Subramanian, Advanced SRAM technology—the race between 4T and 6T cells, *IEDM Tech. Dig.,* 1996, pp. 271–274.

15. M. Helm et al., A low cost, microprocessor compatible, 18.4 μm^2, 6-T bulk cell technology for high speed SRAMs, *1993 Symp. VLSI Tech. Dig. Tech. Papers,* 1993, pp. 65–66.

16. K. Ishimaru et al., Trench isolation technology with 1 μm depth n- and p- wells for a full-CMOS SRAM cell with a 0.4 μm n+/p+ spacing, *1994 Symp. VLSI Tech. Dig. Tech. Papers,* 1994, pp. 97–98.

17. T. McNelly et al., High performance 0.25 mm SRAM technology with tungsten interpoly plug, *IEDM Tech. Dig.,* 1995, pp. 927–930.

18. M. Ueshima et al., A 5 μm^2 full-CMOS cell for high-speed SRAMs utilizing a optical-proximity-effect correction (OPC) technology, *1996 Symp. VLSI Tech. Dig. Tech, Papers,* 1996, pp. 146–147.

19. K. Itabashi et al., A split wordline cell for 16 Mb SRAM using polysilicon sidewall contacts, *IEDM Tech. Dig.,* 1991, pp. 477–480.

20. S. Maegawa et al., Impact of μA-on-current gate-all-around TFT (GAT) for static RAM of 16 Mb and beyond, *Jpn. J. Appl. Phys.,* **35**: 910–914, 1996.

21. S. Ikeda et al., A stacked split word-line (SSW) cell for low voltage operation, large capacity, high speed SRAMs, *IEDM Tech. Dig.,* 1993, pp. 809–812.

22. K. Yuzuriha et al., A large cell-ratio and low node leak 16 Mb SRAM cell using ring-gate transistors, *IEDM Tech. Dig.,* 1991, pp. 485–488.

23. I. Naiki et al., Center wordline cell: a new symmetric layout cell for 64 Mb SRAM, *IEDM Tech. Dig.,* 1993, pp. 817–820.

24. S. Maegawa et al., A 0.4 μm gate-all-around TFT (GAT) using a dummy nitride pattern for high-density memories, *Jpn. J. Appl. Phys.,* **34**: 895–899, 1995.

25. H. Kuriyama et al., A C-switch cell for low-voltage operation and high-density SRAMs, *IEDM Tech. Dig.,* 1996, pp. 279–282.

26. A. Adan et al., *1990 Proc. Symp. VLSI Tech.,* 1990, pp. 19–20.

27. T. Hashimoto et al., Ext. Abstr. *22nd 1990 Int. Conf. Solid State Devices Mater.,* Sendai (Business Center of Academic Societies Japan, Tokyo, 1990), p. 393.

28. C. Liu et al., High reliability and high performance 0.35 μm gate-inverted TFT's for 16 Mb SRAM applications using self-aligned LDD structures, *IEDM Tech. Dig.,* 1992, pp. 823–826.

29. F. Hayashi et al., A highly stable SRAM memory cell with top-grated P^- -N drain poly-Si TFT's for 1.5 V operation, *IEDM Tech. Dig.,* 1996, pp. 283–286.

30. H. Ohkubo et al., 16 Mb SRAM cell technologies for 2.0 V operation, *IEDM Tech. Dig.,* 1991, pp. 481–484.

31. C. Liu et al., Using n-channel TFT's without LDD structures for high stabilities of 1.2-V high-density SRAMs, *IEDM Tech. Dig.,* 1995, pp. 919–922.

32. J. S. Fu, Scaling studies of CMOS SRAM soft-error tolerances - from 16 K to 256 K, *IEDM Tech. Dig.,* 1987, pp. 540–543.

33. T. May and M. Woods, A new physical mechanism for soft errors in dynamic memories. *Proc. Rel. Phys. Symp.,* 1978, pp. 2–9.

34. C. Lage et al., Soft error rate and stored charge requirements in advanced high-density SRAMs, *IEDM Tech. Dig.,* 1993, pp. 821–824.

35. T. Yoshida et al., Crystallization technology for low voltage operated TFT, *IEDM Tech. Dig.,* 1991, pp. 843–846.

36. J. Hayden et al., A new toroidal TFT structure for future generation SRAMs, *IEDM Tech. Dig.,* 1993, pp. 825–828.

37. M. Hashimoto et al., Small geometry SOI technology for high density SRAMs, *IEDM Tech. Dig.,* 1991, pp. 973–975.

38. J. Hayden et al., A high-performance quadruple well, quadruple poly BiCMOS process for fast 16 Mb SRAMs, *IEDM Tech. Dig.,* 1992, pp. 819–822.

39. M. Matsumiya et al., 15-ns 16 Mb CMOS SRAM with interdigitated bit-line architecture. *IEEE J. Solid-State Circuits,* **27** (11): 1497–1502, 1992.

40. A. Kinoshita et al., A study of delay time on bit lines in megabit SRAM's, *IEICE Trans. Electron.,* **E75-C** (11): 1383–1386, 1992.

41. K. Sasaki et al., A 7-ns 140-mW 1-Mb CMOS SRAM with current sense, amplifier, *IEEE J. Solid-State Circuits,* **27** (11): 1511–1518, 1992.

42. T. Seki et al., A 6 ns 1 Mb CMOS SRAM with high-performance sense amplifier, *1992 Symp. VLSI Tech. Dig. Papers,* 1992, pp. 26–27.

43. K. Sasaki et al., A 16-Mb CMOS SRAM with a 2.3-μm^2 single-bit-line memory cell, *IEEE J. Solid-State Circuits,* **28** (11): 1117–1129, 1993.

44. M. Ukita et al., A single-bit-line cross-point cell activation (SCPA) architecture for ultra-low-power SRAM, *IEEE J. Solid-State Circuits,* **28** (11): 1114, 1993.

45. T. Kikuchi et al., A 0.35 μm ECL-CMOS process technology on SOI for 1 ns megabits SRAM's with 40 ps gate array, *IEDM Tech. Dig.,* 1995, pp. 923–926.

46. H. Takahashi et al., 250 MHz BiCMOS synchronous *SRAM, NEC Res. & Develop.* **34** (4): 453–460, 1993.

47. T. Seki et al., A 6-ns 1-Mb CMOS SRAM with latched sense amplifier, *IEICE Trans. Electron.,* **E76-C**: 818–822, 1993.

48. T. Blalock and R. Jaeger, A high-speed sensing scheme for 1 T dynamic RAM's utilizing the clamped bit-line sense amplifier, *IEEE J. Solid-State Circuits,* **27** (4): 618–625, 1992.

49. E. Seevinck, P. van Beers, and H. Ontrop, Current-mode techniques for high-speed VLSI circuits with application to current sense amplifier for CMOS SRAM's, *IEEE J. Solid-State Circuits,* **26** (4): 525–536, 1991.

50. S. Aizaki et al., A 15-ns 4 Mb CMOS SRAM, *IEEE J. Solid-State Circuits,* **25** (5): 1063–1067, 1990.

51. H. Goto et al., A 3.3-V 12-ns 16-Mb CMOS SRAM, *IEEE J. Solid-State Circuits,* **27** (11): 1490–1495, 1992.

52. J. Chen et al., Single transistor static memory cell: Circuit application of a new quantum transistor, *Appl. Phys. Lett.,* **62** (1): 96–98, 1993.

53. T. Hanyu, Y. Yabe, and M. Kameyama, Multiple-valued programmable logic array based on a resonant-tunneling diode model, *IEICE Trans. Electron.,* **E76-C**: 1126–1132, 1993.

54. A. Seabaugh, Y.-C. Kao, and H.-T. Yuan, Nine-state resonant tunneling diode memory, *IEEE Electron. Dev. Lett.,* **13** (9): 479–481, 1992.

55. T. Mori et al., A static random access memory cell using a double-emitter resonant-tunneling hot electron transistor for gigabit-plus memory applications, *Jpn. J. Appl. Phys.,* **33**: 790–793, 1994.

56. P. van der Waft, A. Seabaugh, and E. Beam, III, RTD/HFET low standby power SRAM gain cell, *IEDM Tech. Dig.,* 1996, pp. 425–428.

57. C. Liu et al., A novel amorphous silicon doping superlattice device with double switching characteristics for multiple-valued logic applications, *IEEE Electron Device Lett.,* **14** (8): 391–393, 1993.

58. C. Brown, SRAM research focuses on simplicity, *Electron. Eng. Times,* p. 35, Sept. 11, 1995.

59. K. Sakui et al., A new static memory cell based on reverse base current (RBC) effect of bipolar transistor, *IEDM Tech. Dig.,* 1988, pp. 44–47.

60. X. Zheng, T. Carns, and K. Wang, A GeSi/Si bistable diode exhibiting a large ON/OFF conductance ratio, *Appl. Phys. Lett.,* **66**: 2403, 1995.

61. T. Carns, X. Zheng, and K. Wang, A novel high speed, three element Si-based static random access memory (SRAM) cell, *IEEE Electron Dev. Lett.,* **16** (6): 256–258, 1995.

62. X. Zhu et al., A Si bistable diode utilizing interband tunneling junctions, *Appl. Phys. Lett.* **71** (15): 2190–2192, 1997.

63. P. H. Woerlee et al., Half-micron CMOS on ultra-thin silicon on insulator, *IEDM Tech. Dig.,* 1989, pp. 821–824.

TIMOTHY K. CARNS
ZILOG, Inc.
XINYU ZHENG
KANG L. WANG
University of California, Los Angeles

STABILITY, ABSOLUTE. See ABSOLUTE STABILITY.

STABILITY, CIRCUIT. See CIRCUIT STABILITY.

STABILITY IN FORCED FLOW CONDUCTORS. See SUPERCONDUCTORS, CRYOGENIC STABILIZATION.

STABILITY, NYQUIST. See NYQUIST CRITERION, DIAGRAMS, AND STABILITY.

STABILITY OF AN EQUILIBRIUM. See LYAPUNOV METHODS.

STABILITY, POWER SYSTEM. See POWER SYSTEM STABILITY.