

(thus reducing the amount of time the wafers must be exposed to the high temperature).⁶⁶

3. The use of high-pressure oxidation to reduce the thermal cycle that must be employed to grow the field oxide.^{12,26,55} This also reduces the boron channel-stop dopant loss to the growing field oxide, as well as the interdiffusion of the wells (As a result, smaller isolation spacings between *n*- and *p*-channel devices can be used).

6.6.2.3 Definition of Active and Field Regions. In a conventional advanced CMOS process, a standard LOCOS process (described in detail in chaps. 2 and 5) is used to define the active and field regions (*Mask #2*) and to grow a thick thermal oxide over the field regions. Enhanced LOCOS approaches, which overcome some of the limitations of conventional LOCOS, can be used as alternative processes to form isolation structures in the field regions. Some that have been reported as having been integrated into submicron CMOS processes include the SILO process⁶⁷ and a poly-buffered LOCOS process²⁶ (see chap. 2 for details on both of these). A high-pressure oxidation process for growing the field oxide, was reported to have been used in both of these processes.

6.6.2.4 Gate-Oxide Growth and Threshold-Voltage Adjustment. The gate-oxide growth process is essentially the same as that used in the NMOS process described in chapter 5. The threshold-voltage implant process can be accomplished with either a single implant step, using boron (see section 6.3.1), or two separate implants (in which case additional masking steps are required). A separate selective punchthrough-prevention implant for the NMOS devices is usually needed, especially for submicron devices; this step also requires the use of an extra mask. The V_T -adjust and punchthrough-prevention implants can also be done either before or after the gate-oxidation step. To keep the boron V_T implant shallow (i.e., to improve the PMOS punchthrough characteristics, as described in section 6.3.2), implanting is sometimes done through the gate oxide. To protect the gate oxide from contamination (as well as to keep the implant shallow), part of the polysilicon can be deposited prior to the implant step,^{26,67} with the remainder deposited afterward.

Alternative gate-oxide materials are being studied as replacements for SiO_2 . Gate oxides of less than 10 nm will not be practical unless the gate voltage is also reduced, for two reasons. First, in order to prevent tunneling, the minimum gate oxide thickness must be $2V_{DD}/E_{ox}(\text{max})$, where $E_{ox}(\text{max})$ is 6 MV/cm. For $V_{DD} = 3$ V, the minimum thickness is thus 10 nm. The second reason has to do with the reliability and burn-in techniques used in accelerated lifetime testing. It becomes very difficult to screen out weak devices from good ones when the breakdown-voltage criterion defining a weak oxide falls within the range of the distribution exhibited by the good devices. This situation is encountered at oxide thicknesses below 10 nm.⁶⁸

If a material with a larger dielectric constant than that of SiO_2 (3.9) could be used, the gate dielectric could be thicker, while the same capacitance per unit area could be maintained. Extensive work has been conducted on such materials, including nitrided

oxides, thermal nitrides (dielectric constant = 8), and tantalum pentoxide, Ta_2O_5 (22). Rapid thermal processing has been explored as a means for forming the nitrided oxides, since the high temperatures needed for their formation can be produced by rapidly heating and cooling the wafers.

The Ta_2O_5 films have been prepared through thermal oxidation of a tantalum layer,⁶⁹ by reactive sputtering,⁷⁰ and CVD deposition.⁷¹ Several reports indicate that these films exhibit excellent characteristics, and thus show promise for potential applications in advanced integrated circuits.^{72,75}

In addition, a novel technique for forming ultrathin (15 nm), low-defect SiO_2 gate layers has been reported.⁷⁶ In this approach, a 5-nm thermal oxide is first grown. A 5-nm CVD oxide is then deposited, and finally another 5-nm oxide is thermally grown, *under* the CVD layer (Fig. 6-37). The layer is able to grow beneath the CVD oxide because the oxidizing species diffuse through the first two layers to the Si-SiO₂ interface (see Vol. 1, chap. 7). During the second thermal oxidation, the top CVD oxide is also densified. This combination layer exhibits a low defect density because the growth-induced micropores (which are the major factor contributing to the defect density in gate dielectrics), are misaligned. If they were to extend completely through the oxide film, these micropores would be potential paths for rapid diffusional mass transport, as well as for current leakage.

6.6.2.5 Polysilicon Deposition and Patterning. The process of forming the n^+ polysilicon gate structure involves the deposition of a polysilicon layer and patterning with *Mask #3*. The steps followed are essentially the same as those in the basic CMOS process. In most advanced CMOS processes the polysilicon film is overlaid with a refractory metal silicide to form a *polycide* structure (see Vol. 1, chap. 11). In some cases, a *salicide* (self-aligned silicided gate and source/drain regions – see Vol. 1, chap. 11) process is employed to reduce the parasitic resistance of the source and drain regions, as well as that of the gate material. In most such cases, the salicide

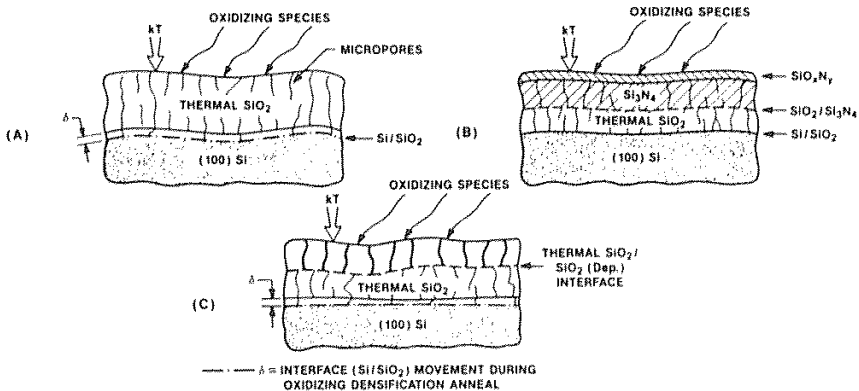


Fig. 6-37 Effects of oxidizing densification anneal: (a) SiO_2 films; (b) SiO_2 - Si_3N_4 dual dielectric; (c) thermal SiO_2 - deposited SiO_2 multilayered stacked films.⁷⁶ (© 1988 IEEE).

formation is carried out after the source and drain have been implanted.

The polysilicon can be doped either after deposition by diffusion or ion implantation, or *in situ* during deposition. The former approaches were used in most MOS processing prior to submicron device generations. For submicron devices, in which a reduced thermal budget becomes more imperative, *in situ* doping becomes more attractive, (since the long, high-temperature drive-in step can be eliminated).¹⁰³ The redistribution of channel and source/drain implants is thus minimized, allowing improved control of threshold voltage, punchthrough, and L_{eff} .

As was described in sections 6.3.3 and 6.3.4, alternative materials to n^+ polysilicon (or to polycides with n^+ polysilicon as the underlayer) have been investigated as gate materials (e.g., p^+ and n^+ polysilicon, or refractory metal gates). Each alternative process introduces some changes into the process flow; nevertheless it is predicted that as device dimensions shrink, such measures will need to be used to maintain adequate device performance.

Etching of the gate structure can be a difficult step, especially when gate lengths are in the submicron range. First, the dimensions must be accurately and uniformly produced across the wafer, as I_D is strongly dependent on the gate dimensions. Second, the sidewalls must be vertical in order to reduce asymmetric I_D characteristics in devices built with LDD structures (see section 5.6.5.2). Third, the formation of vertical sidewall gate structures implies the need for a completely anisotropic etch step (see Vol. 1, chaps. 15 and 16). In turn, such a step requires that the etch process be highly selective against etching the thin underlying gate oxide layer.

6.6.2.6 Formation of Source/Drain Regions. In advanced-CMOS processes, the gate lengths are short enough that LDD structures must be used to minimize hot-electron effects, especially in NMOS devices. Therefore, the procedures outlined in section 5.6.5 are used to fabricate such LDD structures. If these structures are used only for the NMOS transistors, *Masks #4* and *#5* are used to allow the sources and drains of the two transistor types to be selectively implanted.* Note that if LDD structures are also used for the PMOS devices, two additional masking layers may be needed (6-38a).

A removable-spacer LDD process for both NMOS and PMOS devices has been reported (Fig. 6-38b) that does not require the use of any other masks than the two needed to selectively form the sources and drains of the two transistor types.^{67,73} In the removable spacer process, the heavily doped source/drain implant is performed first, with the spacers in place. After the spacers have been removed, the implant that forms the lightly doped drain regions is carried out (Fig. 6-38b). This process can be used to provide LDD structures for one or both transistor types, as desired. Polysilicon is used as the material of the removable spacers in one approach,^{67,99} while a low-temperature

* A procedure that requires only one masking step for creating both types of source and drain regions has also been reported. In this approach, the heavy boron implant is carried out nonselectively. The mask is then used to cover the PMOS device active regions. An n -type implant heavy enough to overcompensate for the boron implant in the active regions of the n -channel devices is used to form the sources and drains of the NMOS transistors.

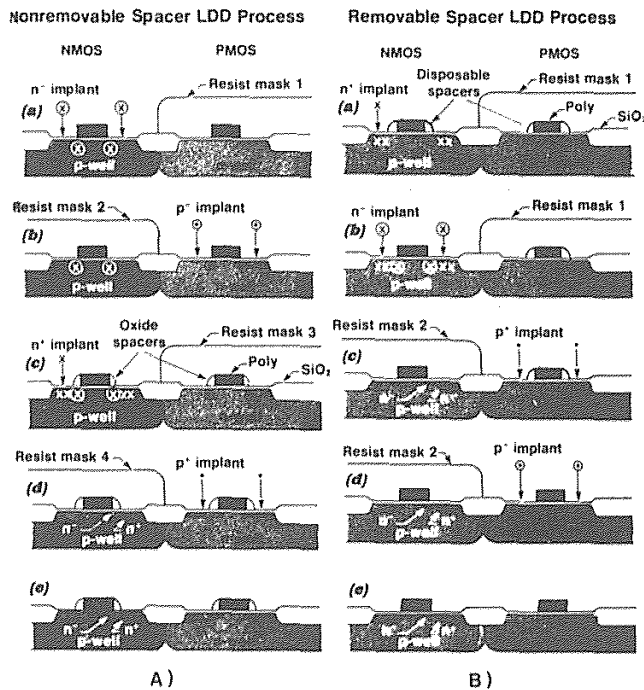


Fig. 6-38 (A) Non-removable oxide spacer LDD process: (a) n^+ implant, (b) p^+ implant, (c) after LDD spacer etch and n^+ implant, (d) after n^+ and p^+ implant, (e) after p^+ anneal. (B) Removable spacer LDD process: (a) after spacer etch and n^+ implant, (b) after NMOS spacer removal and n^+ implant, (c) p^+ implant, (d) after PMOS spacer removal and p^+ implant, (e) after p^+ anneal.⁶⁷ (© 1986 IEEE).

CVD oxide (over a thin polysilicon film) is used in the other.⁷³

Various techniques have been developed for forming the shallow source/drain junctions needed for submicron CMOS devices. First, As is implanted for the n -channel devices and BF_2^+ for the p -channel devices, since both species have very shallow projected ranges at implant energies of 30-50 keV (see Vol. 1, chap. 9). These implants are usually performed through a screen oxide to protect the source/drain regions from contamination during the implant procedure. Second, preamorphization of the silicon by implantation with Si or Ge reduces channeling and helps produce shallow junctions. It is necessary, however, to diffuse the implanted species past the layer of implant damage that cannot be annealed out, in order to prevent junction leakage (see Vol. 1, chap. 9). RTP techniques have been explored as a means of carrying out these anneal and diffusion thermal cycles. Shallow p^+n junction formation through the use of

diffusion (i.e., by utilizing RTP and either a solid diffusion source⁸² or a spin-on diffusion source⁸³) has also been reported. The use of an antimony implant to obtain shallow n^+p junctions with n^+ layers of lower resistivity than is possible with arsenic has also been described.⁸¹

Note that the screen oxide is usually stripped off following the source/drain implant, since it is damaged (and may be contaminated) by the heavy implants. A new oxide is grown over the source/drain regions and on the sidewall of the etched polysilicon electrode (and this step is referred to as *poly reoxidation*). During this thermal cycle, some (but not all) of the implantation damage is annealed out. The oxide formed on the poly sidewall shifts the remaining damage away from the gate edge to maintain the integrity of the thin gate oxide (Fig. 6-39).^{104,110} At the same time, the sidewall oxidation produces a *gate bird's beak*, or GBB, under the polysilicon edge. This reduces the gate-to-drain overlap capacitance and relieves the electric-field intensity at the corner of the gate structure. However, because the GBB encroachment can degrade transconductance of submicron MOSFETs and impact their subthreshold swing and threshold voltage,¹⁰⁹ this poly reoxidation process must be optimized for fabrication of submicron MOSFETs.¹¹¹

An approach to forming shallow junctions that uses CoSi_2 source/drain junctions has been reported.²⁹ In this approach, the CoSi_2 is formed before the source and drain junctions are created by means of a heavy ion implantation step. The implant is then performed so that the damaged layer, which would have occurred in the silicon crystal, is kept within the CoSi_2 layer. As a result, it is necessary to drive the implanted impurities just far enough so that they enter the silicon region to form the required pn junction. If a salicide process is used, the device has a cross-section, as shown in Fig. 6-40.

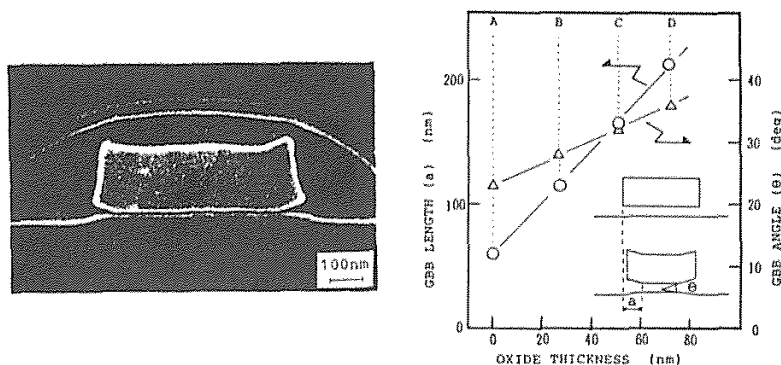


Fig. 6-39 (a) SEM micrograph of a gate structure with a GBB. (b) Profiles of GBB versus oxide thickness grown during re-oxidation.¹⁰⁴ This paper was originally presented at the Spring 1989 Meeting of The Electrochemical Society, Inc. held in Los Angeles, CA.

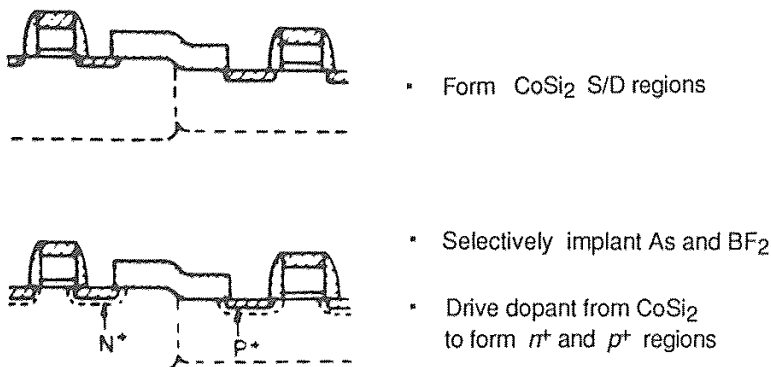


Fig. 6-40 Cross section of CMOS structure in which CoSi₂ is first formed on the source and drain regions, and then As and BF₂⁺ is selectively implanted into the CoSi₂. These dopants are then driven out of the CoSi₂ and into the Si substrate to form shallow p⁺ and n⁺ source/drain regions.²⁹ (© 1986 IEEE).

Another approach is to create so-called *elevated source-drains*. A thin (e.g., 200-nm) epitaxial layer of silicon can be selectively deposited onto the exposed source-drain areas of the MOS transistor, following the implantation of the lightly doped region of the LDD structure and formation of the spacers (Fig. 6-41a). This method has been used for the NMOS devices of a 1 Mbit SRAM.⁷⁹ In this case, a heavy BF₂⁺ implant is done such that the SEG film becomes heavily doped; the boron penetrates to the substrate during an RTP anneal that follows the implant. In this way, elevated, heavily doped, shallow source/drain regions are formed. The source-drain junction depths in the substrate are less than 0.2 μm deep (Fig. 6-41b). As noted earlier, the gate oxide that covers the source and drain regions is usually etched away and regrown following the implant step.

6.6.2.7 CVD Oxide Deposition and Contact Formation. Following the formation of the source and drain regions (and the salicide layers), a doped oxide is deposited by CVD, with procedures very much like those of the basic NMOS and CMOS processes (see section 5.4.1.6).

Contact windows are opened in this CVD layer to allow electrical connections to be made between the Metal 1 layer and the source/drain, gate, and substrate and well contact regions (*Mask #6*). Again, the details are similar to those of the basic NMOS and CMOS processes.

Advances in contact technology for CMOS include the use of barrier layers to prevent spiking through the shallow junctions, and the filling of contact holes by such materials as CVD W, polysilicon, or even SEG. In addition, as device dimensions decrease, the parasitic resistances of the source and drain regions become more significant. More details on all of these topics are provided in chapter 3.

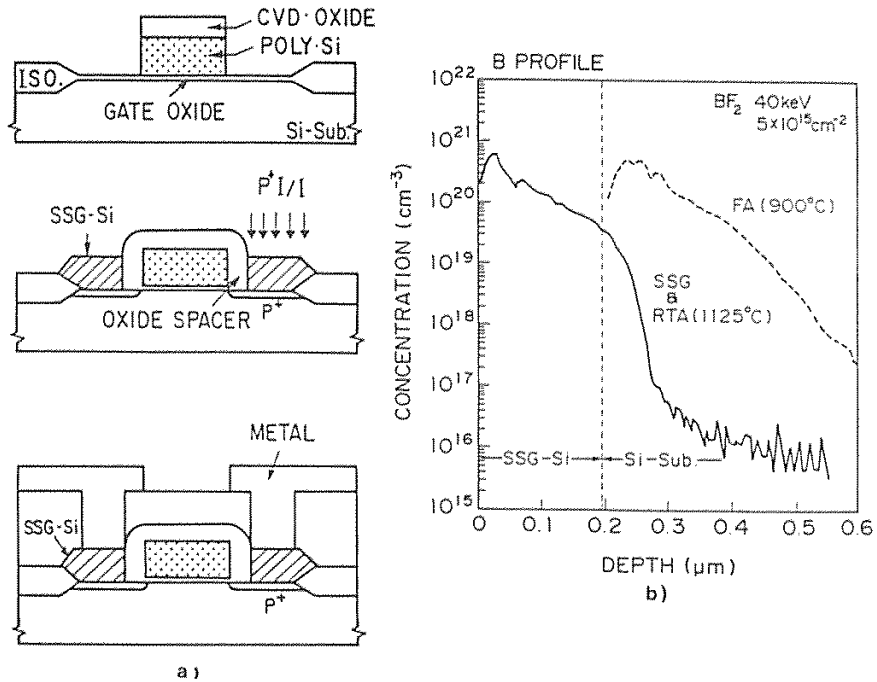


Fig. 6-41 Elevated source/drains. (a) Key process steps in forming elevated source/drains using SEG. (b) Comparison of the doping profile in source/drain regions of a PMOS device after a BF_2^+ implant directly into the regions (and a conventional furnace anneal) versus a BF_2^+ implant into the SEG regions and an RTP anneal.⁷⁹

6.6.2.8 Metal 1 Deposition and Patterning. The issues involved in the deposition and patterning of the Metal 1 layer are not significantly different from those described in chapter 5 for the NMOS process. *Mask #7* is used to pattern the Metal 1 layer. (For more information on the metal-layer deposition and patterning processes, see chapter 4 of this volume, and Vol. 1, chap. 9.)

Despite its higher resistivity compared to Al, tungsten's superior electromigration properties make it more appropriate for certain applications. As a result, CVD W has been selected as the Metal 1 material for a variety of circuit designs.

6.6.2.9 Intermetal Dielectric Deposition/Planarization and Via Patterning. Following the Metal 1 patterning, an intermetal dielectric must be deposited to electrically isolate Metal 1 from the Metal 2 layer. A variety of materials and deposition processes, have been utilized (see chap. 4).

However, deposition of this layer may make the wafer topography too severe to allow the Metal 2 film to be deposited with adequate step coverage. A number of *planarization* processes have been developed to overcome this problem (see chap. 4). If such techniques are successfully implemented, the wafer topography will be relatively planar, and any steps on its surface will be gently sloped rather than severely vertical.

It is necessary to open vias in the intermetal dielectric layer so that an electrical connection can be established between Metal 2 and Metal 1 at desired locations (*Mask #8*). Metal 2 must be deposited with adequate step coverage into the vias.

A number of techniques have been studied to improve coverage of Metal 2 in the vias, including the following (see chap. 4):

- Sloping the via walls by means of the via-etch process.
- Filling the vias with a blanket W or polysilicon deposition, and then etching back to provide a planar surface.
- Filling the vias with a selective deposition of W or Al.
- Increasing the step coverage into the vias through bias sputtering and heating of the substrate during deposition.
- Laser melting the Metal 2 film to increase step coverage.

6.6.2.10 Metal 2 Deposition and Patterning. The processing issues of depositing and patterning (*Mask #9*) the Metal 2 layer are discussed in chapter 4.

6.6.2.11 Passivation Layer Deposition and Patterning. The passivation-layer deposition and patterning (*Mask #10*) issues are the same those of the basic CMOS or NMOS processes.

6.7 MISCELLANEOUS CMOS TOPICS

6.7.1 Electrostatic-Discharge Protection

The input signals to an MOS IC are fed to the gates of MOS transistors. If the voltage applied to the gate insulator becomes excessive, the gate oxide can break down. The dielectric breakdown strength of SiO_2 is approximately 8×10^6 V/cm; thus, a 15-nm gate oxide will not tolerate voltages greater than 12 V without breaking down. Although this is well in excess of the normal operating voltages of 5-V integrated circuits, voltages higher than this may be impressed upon the inputs to the circuits during either human-operator or mechanical handling operations.

The main source of such voltages is triboelectricity (electricity caused when two materials are rubbed together). A person can develop very high static voltage (i.e., a few hundred to a few thousand volts) simply by walking across a room or by removing an integrated circuit from its plastic package, even when careful handling procedures are

followed. If such a high voltage is accidentally applied to the pins of an IC package, its discharge (referred to as *electrostatic discharge*, or ESD) can cause breakdown of the gate oxide of the devices to which it is applied. The breakdown event may cause sufficient damage to produce immediate destruction of the device, or it may weaken the oxide enough that it will fail early in the operating life of the device (and thereby cause device failure). A more detailed description of ESD failures in semiconductor devices can be found in reference 84.

All pins of MOS ICs must be provided with protective circuits to prevent such voltages from damaging the MOS gates. The need for such circuits is also mandated by the increasing use of VLSI devices in such high-noise environments as personal computers, automobiles, and manufacturing control systems. These protective circuits, normally placed between the input and output pads on a chip and the transistor gates to which the pads are connected, are designed to begin conducting or to undergo breakdown, thereby providing an electrical path to ground (or to the power-supply rail). Since the breakdown mechanism is designed to be nondestructive, the circuits provide a normally open path that closes only when a high voltage appears at the input or output terminals, harmlessly discharging the node to which it is connected.

Four types of circuits are used to provide protection against ESD damage, as follows:

1. diode breakdown
2. node-to-node punchthrough
3. gate-field-induced breakdown
4. parasitic *pnpn* diode latchup.

Often, a combination of protection methods is used, with a breakdown diode and one of the other protection devices connected in parallel with the gate being protected.

6.7.1.1 Diode Protection. Protection is obtained by using the diode-breakdown phenomenon to provide an electrical path in the silicon substrate that consists of a diffused resistor region (of a doping type opposite to that of the substrate). This diffused region is connected between the input pad and the gate (Fig. 6-42a). If a reverse-bias voltage greater than the breakdown voltage of the resultant *pn* junction is applied, the diffusion region (which otherwise works as a resistor), operates as a diode and undergoes breakdown. Furthermore, this diffused region will also clamp a negative-going transition at the chip input to one diode drop below the substrate voltage. In CMOS technologies, an additional protection diode can be added by utilizing the *pn* junction that exists between a p^+ node and the body region of the PMOS device (an *n*-type region, that is connected to V_{DD}). This diode is utilized as a protection device when a connection is made between the pad and a p^+ region. (Note that this second diode will clamp positive-going transitions to one diode drop *above* V_{DD} .)

6.7.1.2 Node-to-Node Punchthrough. As defined in section 5.5.2, *punchthrough* is the phenomenon by which the depletion region surrounding the drain of an MOS device extends along the channel of the device and contacts the depletion region of the source of the device. While the current that results from punchthrough is ordinar-

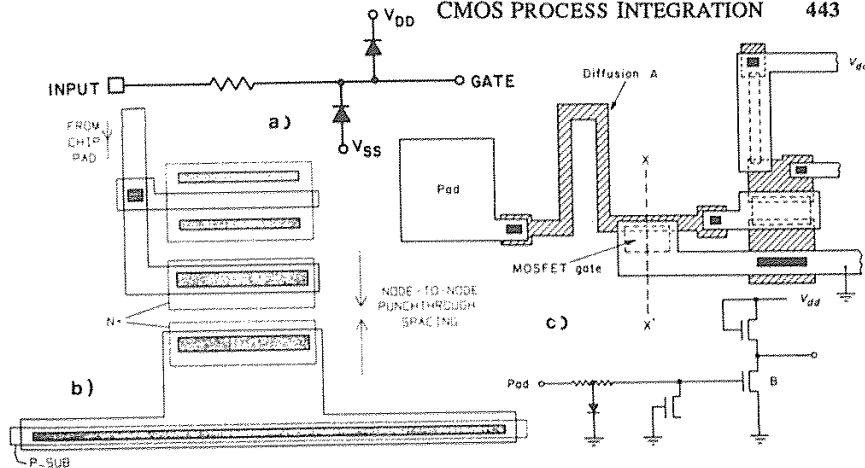


Fig. 6-42 MOS protective circuits: (a) Diode; (b) Punchthrough device; (c) Layout of an input protection structure containing both devices.¹²² From T. E. Dillinger, *VLSI Engineering*. Copyright Prentice-Hall, 1988. Reprinted with permission.

arily an unwanted effect, it is harnessed in ESD protection circuits for beneficial purposes. The source and drain regions of an ESD punchthrough protection structure are placed close enough together that punchthrough occurs at a voltage lower than that of the gate-oxide breakdown (Fig. 6-42b). Such a structure is laid out and tested to meet a specific ESD transient reliability measure. Once it has been verified as functioning according to the design specifications, it can be connected to each chip signal I/O pad.

6.7.1.3 Gate-Controlled Breakdown Structure. The electric field near the corner of an MOS device's drain node is strongest at the surface, since the depletion region is narrowest at this point⁸⁵ and the entire voltage across the depletion region must be dropped over a very short distance. The fact that the gate voltage can increase the strength of the electric field at the corner of the drain is the principle used in designing the gate-controlled breakdown structure (Fig. 6-43).

As the reverse-bias voltage applied to the drain is increased, the gate-to-drain voltage also increases (note that the gate is tied to ground). Because the presence of the gate reduces the breakdown voltage of the junction near the corner of the drain, a relatively small voltage can cause the junction to break down at this point. At breakdown, the junction conducts a large current; in this device, the current flows from the drain to the substrate. The actual input voltage at which breakdown occurs can be controlled by altering the oxide thickness. This produces a large voltage drop across the relatively high resistance of the diffusion area, reducing the voltage applied to the gate of the MOSFET.

Such structures were the workhorses of ESD protection devices. It has been found, however, that processing enhancements that increase the device performance of small-dimension devices sometimes have a negative impact on the failure resistance of the

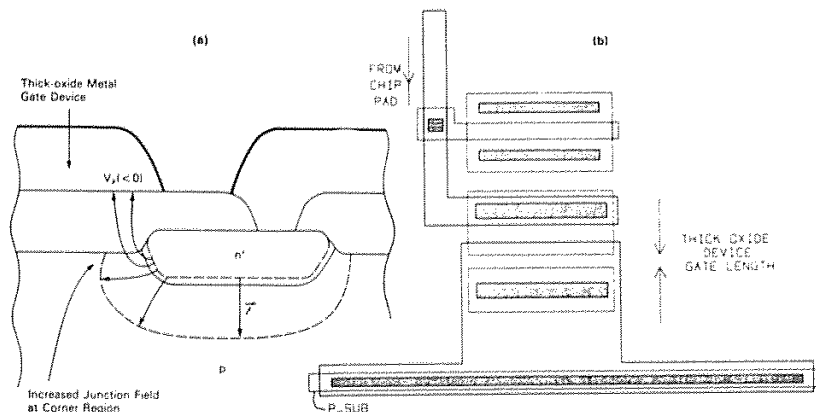


Fig. 6-43 (a) Cross-section, and (b) layout of a thick-oxide metal-gate device for node-to-substrate breakdown and input-pad protection.¹²² From T. E. Dillinger, *VLSI Engineering*. Copyright Prentice-Hall, 1988. Reprinted with permission.

ESD protective devices. Specifically, silicided source/drain regions and LDD structures can degrade the ESD performance of the gate-controlled breakdown structures so that the voltage at which *these devices themselves fail* is reduced from 4 kV to 1 kV.^{86,87} As a result, ESD protection structures can be rendered largely ineffective by the silicided regions, or else the circuits can be caused to fail by ESD in new failure modes as a result of the addition of performance-enhancement features.^{86,88}

6.7.1.4 *pnpn*-Diode ESD Protection for Advanced CMOS Circuits.

Even as devices are made smaller, a fixed minimum volume of silicon must always be used to dissipate the power associated with the current flowing through the protective devices to ground. If the power is dissipated in too small a volume of silicon, the silicon can be heated beyond its melting point; this can destroy the device, even if the circuit-protection device is operating properly. As a result, a minimum unscalable area (Fig. 6-44) will be required in order for the same degree of protection to be maintained through each generation of scaled technology.^{68,89}

Effective protection of VLSI must be provided for both automated and human handling. ESD from a machine is typically modeled as a high-voltage (4-kV), short-duration pulse (i.e., the *EIAJ machine model [MM] test method*), while ESD from a human operator is usually modeled as a longer, lower-voltage (300-V) but higher-current pulse (e.g., the *Mil. Std. human body model [HBM] test method*). Thus, if the gate-controlled breakdown structure itself fails at 1 kV, it ceases to provide protection against HBM ESD.⁸⁷

To solve this problem on the inputs of a CMOS chip, it has been proposed that the parasitic lateral *pnpn* structure inherent in all CMOS circuits be exploited (Fig. 6-45a)

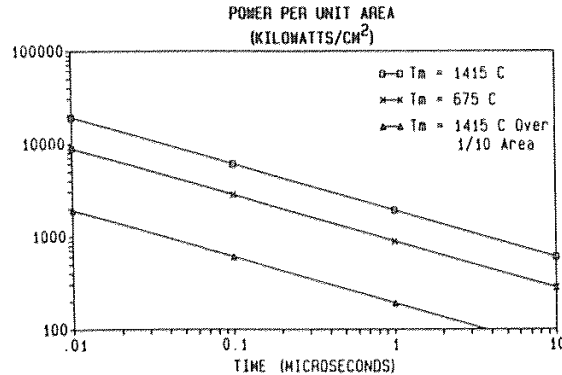


Fig. 6-44 Wunsch and Bell power curves for silicon melt, aluminum melt, and silicon melt for nonuniform current density.⁸⁹ (© 1968 IEEE).

as an ESD structure (see section 6.4). Such a structure – which is deliberately configured to latch up at a voltage lower than that required to damage the input MOS gate oxide – is built into each input circuit on the chip (D1 in Fig. 6-45b). Thus, this parasitic *pnpn* structure serves as the primary ESD structure of the input circuit by

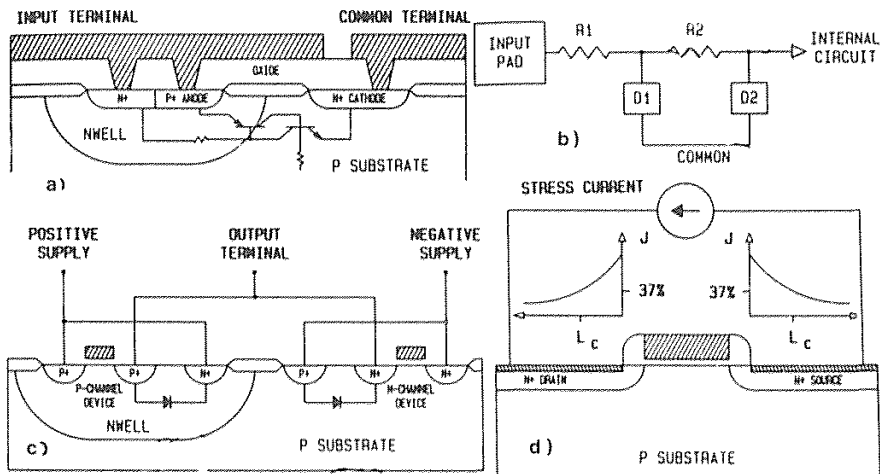


Fig. 6-45 (a) Lateral *pnpn* input protection structure showing parasitic and increased area at *n*-well and *p*-well junction resulting in lower power density. (b) Input-protection circuit showing primary protection elements R1 and D1 and secondary protection elements R2 and D2. (c) Representation of dual-diode output circuit cross section. Effectiveness depends on parasitic diode impedance as well as capacitance between supplies. (d) Cross section of MOS thin-oxide device showing current density at silicide and silicon interface.⁸⁷ (© 1979 IEEE).

providing a low-impedance path when it is turned on. It also offers increased area for power dissipation. Secondary ESD structures, however, must also be used on each input circuit (e.g., D2 and R2 in Fig. 6-45b) to provide effective protection until D1 achieves its low-impedance conducting state.

On the outputs of the chip, the parasitic diodes that exist in the CMOS output circuit are used to provide ESD protection, much as described in section 6.7.1.1 (Fig. 6-45c).⁸⁷ Such dual-diode CMOS outputs have proved to be effective when the resistance of the parasitic diode in the *p*-channel device is minimized and the failure threshold of the *n*-channel device is maximized. As shown in Fig. 6-45d, the stress current density in the diode of the silicided *n*-channel device is very nonuniform due to the silicided source/drain regions. As a result, the diode can fail if too high a current density passes through only a small fraction of the area of the silicided contact.

6.7.2 Power-Supply Voltage Levels for Future CMOS

As MOS has evolved, device scaling has been practiced to increase the speed and packing density of MOS ICs. Although scaling of the electric field to maintain a constant electric field (CE) would have promoted both high performance and high reliability, in practice a constant voltage (CV) has been maintained in order to maintain TTL compatibility and produce higher circuit performance. Such nonscaling of the voltage leads to reliability problems, caused by hot-electron effects and oxide breakdown. Scaling of the power supply voltage is thus desirable. M. Kakamu has calculated the optimum power-supply voltage as a function of shrinking line width, and has shown that this voltage can be scaled down in proportion to the square root of the design-rule shrinkage without sacrificing circuit performance.⁹⁰ The relationships given between the supply voltage and the design rule are

$$V_{DD} = 6.1 \times (L/2)^{1/2} \text{ (for conventional drain structures)} \quad (6-9)$$

and

$$V_{DD} = 8.2 \times (L/2)^{1/2} \text{ (for LDD structures)} \quad (6-10)$$

where L (μm) is the design rule and V_{DD} is the power-supply voltage (volts). These equations predict that at dimensions of $0.6 \mu\text{m}$, the power supply voltage will have to be reduced to 3.3 V in order for reliable device operation and circuit performance to be maintained if conventional drain structures are used. On the other hand, if LDD structures are used, it is predicted that $V_{DD} = 5 \text{ V}$ will still be feasible for $0.7\text{-}\mu\text{m}$ dimensions. A recent report indicates that these equations may be slightly pessimistic, in that MOSFETs with gate lengths of $0.6 \mu\text{m}$ that operate from a 5-V supply have been successfully fabricated.¹⁰² Through optimization of the LDD structures of these devices, a lifetime of more than 10 years in the face of hot-carrier degradation effects, is predicted.

6.7.3 Low-Temperature CMOS

Since it is possible to build MOS devices with minimum dimensions of less than 0.5 μm , low-temperature operation of CMOS (e.g., at 77°K) has been actively investigated. At low temperatures, the MOS devices exhibit lower subthreshold leakage, higher carrier mobility (which yields improved speed performance), and a steeper logarithmic current-voltage slope. It has been found that the normalized propagation delay in CMOS logic gates is reduced by a factor of nearly 2 or 3 when the devices are operated at 77°K.^{91,92,93,94} The major reason for the slow commercial introduction of such systems has been the difficulty associated with liquid-nitrogen refrigeration.

6.7.4 Three-Dimensional CMOS

Since the limits of two-dimensional CMOS may soon be approached, a logical direction for continued advances in device scaling is three-dimensional devices. CMOS is well-suited for 3-D integration, since low-power circuits will be mandatory in order for heat dissipation problems to be avoided. The 3-D approach will offer the benefits of increased packing density and higher speeds (due to shorter interconnects).

The inability to fabricate single-crystal silicon over insulating layers, however, has historically been an obstacle to the implementation of 3-D ICs. Recent advances in SOI technologies show promise of being able to surmount this obstacle and much work is currently being done to develop 3-D CMOS.⁹⁵

REFERENCES

1. J. Y. Chen, *CMOS Devices and Technology for VLSI*, Englewood Cliffs N.J., Prentice-Hall, 1989.
2. F. M. Wanlass and C. T. Sah, *IEEE Int. Solid-State Circ. Conf.*, February 1963.
3. R. D. Davies, "The Case for CMOS", *IEEE Spectrum*, October 1983, p. 26.
4. J. Y. Chen, "CMOS-The Emerging Technology," *IEEE Circuits and Systems Magazine*, March 1986, p. 16.
5. D. M. Brown, M. Ghezzi, and J. M. Pimbley, "Trends in Advanced CMOS Process Technology," *Proceedings of the IEEE*, December 1986, p. 1678.
6. W. C. Holton and R. K. Calvin, "A Perspective on CMOS Technology Trends," *Proceedings of the IEEE*, December 1986, p. 1646.
7. S. Prussin, private communication.
8. R. Chwang and K. Yu, "CHMOS-An n -Well Bulk CMOS Technology for VLSI," *VLSI Design*, Fourth Quarter, 1981, p. 42.
9. W. Wijaranakula et al., *J. Electrochem. Soc.*, December, 1988, p. 3113.
10. H. Tsuya et al., *Japanese J. Appl. Phys.*, 22, L16, (1983).
11. T. Yamaguchi et al., *IEEE Trans. Electron Dev.*, ED-31, 205 (1984).
12. L. C. Parrillo, "CMOS Active and Field Device Fabrication," *Semicond. Internat.*, April 1988, p. 64.

13. A. G. Lewis et al., "Vertical Isolation in Shallow n -Well CMOS Circuits," *IEEE Electron Dev. Letts.*, March 1987, p. 197.
14. L. C. Parrillo et al., "Twin-Tub CMOS," *IEDM Tech. Dig.*, 1980, p. 752.
15. J. Manoliu and J. O. Borland, "A Submicron Buried-Layer Twin-Well CMOS SEG Process," *IEDM Tech. Dig.*, 1987, p. 20.
16. R. D. Rung, C. J. Dell' Oca, and L. G. Walker, "A Retrograde p -Well for High-Density CMOS," *IEEE Trans. Electron Dev.*, **ED-28**, p. 1115, 1981.
17. R. A. Martin and J. Y. Chen, "Optimized Retrograde n -Well for 1 μm CMOS," *Proc. Custom Integ. Circ. Conf.*, 1985, p. 199.
18. S. R. Combs "Scalable Retrograde p -Well CMOS Technology," *IEEE Trans. Electron Dev.*, **ED-28**, p. 346, 1981.
19. Y. Taur et al., *J. Solid-State Circuits*, **SC-20**, p. 123, 1985.
20. S. M. Sze, Ed., *VLSI Technology*, 2nd. Ed., New York, McGraw-Hill, 1988, p. 491.
21. T. Ohzone et al., *IEEE Trans. Electron Dev.*, **ED-32**, 1789, (1980).
22. K. M. Cham, S.-Y. Oh, D. Chin, and J. L. Moll, *Computer-Aided Design and VLSI Device Development*, Boston, Kluwer Academic Publishers, 1986 p. 182.
23. G. J. Hu et al., *IEDM Tech. Dig.*, 1982, p. 710.
24. J. Zhu et al., *IEEE Trans. Electron Dev.*, February 1988, p. 145.
25. A. Schmitz and J. Y. Chen, "Design, Modelling, and Fabrication of Submicron CMOS Transistors," *IEEE Trans. Electron Dev.*, **ED-33**, p. 148, (1986).
26. R. A. Chapman et al., "A 0.8 μm CMOS Technology," *Tech. Dig. IEDM*, 1987, p. 362.
27. L. C. Parrillo et al., *Tech Dig. IEDM*, 1984, p. 418.
28. K. M. Cham et al., *IEEE Electron Dev. Lett.*, January 1986, p. 49.
29. S. J. Hillenius et al., "A Symmetric Submicron CMOS Technology," *Tech. Dig. IEDM*, 1986, p. 252.
30. Y. Pauleau, "Interconnect Materials for VLSI Circuits, Part-1," *Solid State Technology*, February 1987, p. 61.
31. V. Schwabe, F. Neppl, and E. P. Jacobs, *IEEE Trans. Electron Dev.*, **ED-31**, 1984, p. 988.
32. S. Iwata et al., *IEEE Trans Electron Dev.*, **ED-31**, 1984, p. 1174.
33. R. F. Kwasnick et al., *IEEE Trans. Electron. Dev.*, September 1988, p. 1432.
34. H. Oikawa and T. Amazawa, *Proc. 3rd. Intl. VLSI Symp., ECS Meeting*, May, 1985, p. 131.
35. D. M. Brown et al., *IEEE Trans. Electron Dev.*, **ED-18**, 1971, p. 931.
36. R. R. Troutman, *Latchup in CMOS Technology*, Kluwer, Boston, MA., 1986.
37. G. J. Hu, "A Better Understanding of CMOS Latchup," *IEEE Trans. Electron Dev.*, **ED-31**, p. 62, 1984.
38. R. R. Troutman, "Latchup in CMOS Technologies," *IEEE Circuits & Sys. Mag.*, May 1987, p. 15.
39. A. G. Lewis et al., *Tech. Dig. IEDM*, 1986, p. 248.
40. R. D. Rung and H. Momose, *IEEE Trans. Electron Dev.*, **ED-30**, 1983, p. 1647.
41. W. R. Dawes and G.F. Derbenwick, "Prevention of CMOS Latchup by Gold-Doping," *IEEE Trans. Nucl. Sci.*, **NS-23**, 1976, p. 2027.

42. J. R. Adams and R. J. Sokel, Presented at the 1979 Nuclear and Space Radiation Effects Conf., Santa Cruz, CA., July, 19, 1979.
43. J. O. Borland and T. Deacon, "Advanced CMOS Epitaxial Processing for Latchup Hardening," *Solid-State Technology*, p. 123, August 1984.
44. S. Ratanphanyarat et al., *Tech. Dig. IEDM*, 1987, p. 744.
45. F. S. Lai et al., *IEDM Tech. Dig. 1985*, p. 513.
46. M. L. Chen et al., *Tech. Dig. IEDM*, 1986, p. 256.
47. S. Swirhun et al., "Latchup-Free CMOS Using Guarded Schottky Barrier PMOS," *Tech. Dig. IEDM*, 1984, p. 402.
48. R. A. Martin et al., *IEDM Tech. Dig.*, 1985, p. 403.
49. A. G. Lewis, *IEEE Trans. Electron Dev.*, Oct. 1984, p. 1472.
50. D. Takacs et al., *Tech. Dig. IEDM*, 1983, p. 159.
51. K. W. Terrill et al., *Tech. Dig. IEDM*, 1984, p. 406.
52. R. Menozzi et al., "Layout Dependence of CMOS Latchup," *IEEE Trans. Electron Dev.*, Nov. 1988, p. 1892.
53. J. Y. Chen and D. E. Snyder, "Modeling Device Isolation in High-Density CMOS," *IEEE Electron Dev. Letts.*, February 1986, p. 64.
54. L. Herman, "Controlling CMOS Latchup," *VLSI Design*, April, 1985, p. 100.
55. M. -L. Chen et al., *Tech. Dig. IEDM*, 1986, p. 256.
56. K. M. Cham and S.-Y. Chiang, "A Study of Trench Surface Inversion Problem for the Trench CMOS Technology," *IEEE Electron Dev. Lett.*, Sept., 1983, p. 303.
57. Y. Nitsu et al., "Latchup Free CMOS Structure Using Shallow Trench Isolation," *Tech. Dig. IEDM*, 1985, p. 509.
58. R. D. Rung, H. Momose, and Y. Nagakubo, "Deep Trench Isolated CMOS Devices," *IEDM Tech. Dig.*, 1982, p. 6.
59. T. I. Kamins and S. Shiang, *Electron Dev. Letts.*, December 1985, p. 617.
60. N. Kasai et al., *IEEE Trans. Electron Dev.*, June 1985, p. 1331.
61. A. Stivers et al., *10th Internat. Conf. on CVD, ECS*, 87-8, p. 389.
62. R. D. Rung et al., *IEEE Trans. Electron Dev.*, ED-28, 1981, p. 1115.
63. L. C. Parrillo et al., "Twin-Tub CMOS II," *IEDM Tech. Dig.*, 1982, p. 706.
64. S. J. Hillenius and L. C. Parrillo, U.S. Patent No. 4,554,726, Nov. 25, 1985.
65. J. Pfister and J. Alvis, "Improved CMOS Field Isolation Using Ge/B Implantation," *IEEE Elect. Dev. Lett.*, August 1988, p. 391.
66. J. Pfister, *Electron Dev. Letts.*, November 1988, p. 561.
67. L. C. Parrillo et al., *Tech. Dig. IEDM*, 1986, p. 244.
68. M. H. Woods, "MOS VLSI Reliability and Yield Trends," *Proceedings of IEEE*, Dec. 1986, p. 1715.
69. G. S. Ohrlein, *J. Appl. Physics*, 59, 1587 (1986).
70. B. W. Shen et al., *Tech. Dig. IEDM*, 1987, p. 582.
71. M. Saitoh, T. Mori, and H. Tamura, *Tech. Dig. IEDM*, 1986, p. 680.
72. K. Ogiue et al., "Technology Improvement for High Speed ECL RAMs," *Tech. Dig. IEDM*, 1986, p. 468.
73. J. Pfister, *IEEE Electron Dev. Letters*, April, 1988, p. 189.
74. R. de Werdt et al., *Tech. Dig. IEDM*, 1987, p. 532.

75. S. G. Byeon and Y. Tzeng, *Tech. Dig. IEDM*, 1988, p. 722.
76. P. K. Roy et al., *Tech. Dig. IEDM*, 1988, p. 714.
77. N. Kasai, N. Endo, and A. Ishitani, *Tech. Dig. IEDM*, 1988, p. 242.
78. L. Manchanda, *Intl. Reliability Symp.*, 1986, p. 183.
79. K. Hashimoto, Presented at the Third Annual Applied Materials Inc., "Innovations in Epitaxial Technology for Advanced Device Structures" Seminar, December 15, 1988, Palo Alto, CA.
80. G. Sai-Halasaz and H. B. Harrison, *Electron Dev. Letts.*, September 1986, p. 534.
81. M. L. Chen et al., "Constraints in P-Channel Device Engineering for Submicron CMOS Technologies," *Tech. Dig. IEDM*, 1988, p. 390.
82. K. -T. Kim and C. -K. Kim, *IEEE Electron Dev. Letts.*, December 1987, p. 569.
83. E. Ling et al., *IEEE Electron Dev. Letts.*, March, 1987, p. 96.
84. B. A. Unger, *Intl. Reliability Physics Symp.*, 1981, p. 193.
85. A. S. Grove, *Physics and Technology of Semiconductor Devices*, Wiley, New York, 1967, Section 10.5.
86. K. - L. Chen et al., *Tech. Dig. IEDM*, 1986, p. 484.
87. R. N. Rountree, *Tech. Dig. IEDM*, 1988, p. 580.
88. C. Duvvury et al., "ESD Protection in 1 μ m CMOS," *Intl. Reliability Physics Symp.* 1986.
89. D. C. Wunsch and R. R. Bell, "Determination of Threshold Failure Levels of Semiconductor Diodes and Transistors due to Pulse Voltages," *IEEE Trans. Nucl. Sci.*, NS-15, p. 244, Dec. 1968.
90. M. Kakumu et al., *Tech. Dig. IEDM*, 1986, p. 399.
91. J. D. Plummer, "Low Temperature CMOS Devices and Technology," *Tech. Dig. IEDM*, 1986, p. 378.
92. J. Watt and J. D. Plummer, *Tech. Dig. IEDM*, 1987, p. 393.
93. J. Y-C. Sun et al., *Tech. Dig. IEDM*, 1986, p. 236.
94. J. S. T. Huang and J. W. Schrankler, *IEEE Electron Dev.*, January 1987, p. 101.
95. Y. Akasaka, "Three-Dimensional IC Trends," *Proc. IEEE*, December 1986, p. 1703.
96. M. H. El-Diwany et al., *Tech. Dig. IEDM*, 1987, p. 917.
97. R. H. Krambeck, C. M. Lee, H. F. S. Law, *IEEE J. Solid-State Circuits*, SC-17, (1982), p. 614.
98. M. Wong and K. Saraswat, *IEEE Electron Dev. Letts.*, November 1988, p. 579.
99. C. -Y. Yang et al., *Ext. Abs. of the Electrochemical Soc. Meeting*, Spring, 1988, Abs. No. 138, p. 207.
100. C. Y. Wong et al., *Tech. Dig. IEDM*, 1988, p. 238.
101. G. J. Hu and R. H. Bruce, *IEEE Electron Dev. Letts.*, EDL-5, p. 211, 1984.
102. H.-M. Mulhoff, K. H. Kusters, and H. Melzner, *Ext. Abs. of the Electrochemical Soc. Meeting*, Spring, 1989, Abs. No. 133, p. 186.
103. S. M. Kugelmass and J. P. Krusius, *Ext. Abs. of the Electrochemical Soc. Meeting*, Spring, 1989, Abs. No. 135, p. 188.
104. M. Kishimoto et al., *Ext. Abs. of the Electrochemical Soc. Meeting*, Spring, 1989, Abs. No. 137, p. 191.
105. P.-H. Pan, J. G. Ryan, and M. A. Lavoie, *Ext. Abs. of the Electrochemical Soc. Meeting*, Spring, 1989, Abs. No. 138, p. 193.
106. A. G. Lewis et al., *IEEE Trans. Electron Dev.*, June 1987, p. 1337.
107. Y. Sato, K. Ehara, and K. Saito, *J. Electrochem. Soc.*, June, 1989, p. 1777.

108. C. Mead and L. Conway, *Introduction to VLSI Systems*, Addison-Wesley, Reading, MA, 1980.
109. J. R. Pfeister et al., *IEEE Electron Dev. Letts.*, August 1989, p. 367.
110. C. Y. Wong et al., "Process Induced Degradation of Thin Oxides," in *Proc. 1st Int. Symp. Ultra Large Scale Integration Sci. Technol.*, S. Broydo and C. M. Osburn, Eds., Electrochem. Soc., 1987, p. 155.
111. C. Y. Wong et al., *IEEE Electron Dev. Letts.*, September, 1989, p. 420.
112. J. Y. Chen, *IEEE Trans. Electron Devices*, ED-31, p. 910 (1984).
113. K. Tanno, F. Shimura, and T. Kawamura, *J. Electrochem. Soc.*, 128, 395, (1981).
114. W. C. Till and J. T. Luxton, *Integrated Circuits: materials, devices, and fabrication*, Prentice-Hall, Englewood Cliffs, N. J., 1982.
115. B. Hoefflinger and G. Zimmer, in "J. Carroll, ed., *Solid-State Devices 1980, Institute of Physics Conf. Ser. 57*, from the 10th European Solid-State Device Research Conf., Sept. 1980.
116. A. G. Lewis et al., *IEEE Trans. Electron Dev.*, ED-34, 2156, (1987).
117. J. Y. Chen, *VLSI Design*, July 1984, p. 78.
118. H. Momose et al., *Tech. Dig. IEDM*, 1984, p. 706.
119. L. C. Parrillo, G. W. Reutlinger, and L. K. Wang, U.S. Patent No. 4,435,895, March 31, 1984.
120. K. M. Cham et al., *Tech. Dig. IEDM*, 1986, p. 296.
121. M. L. Chen, *Semiconductor International*, April 1988, p. 78.
122. T. E. Dillinger, *VLSI Engineering*. Englewood Cliffs, N.J., Prentice-Hall, 1988.
123. F. K. Baker et al., *Tech. Dig. IEDM*, 1989, p. 443.
124. J. M. Sung et al., *Tech. Dig. IEDM*, 1989, p. 447.

PROBLEMS

6.1 An n -well CMOS process starts with a substrate doping of $3 \times 10^{15} \text{ cm}^{-3}$. The well doping g near the surface is approximately constant at a level of $3 \times 10^{16} \text{ cm}^{-3}$. The gate oxide thicknesses are both 40 nm. (a) Calculate the threshold voltages of the n - and p -channel transistors. (b) Calculate the boron doses needed to shift the NMOS threshold to +1 V and the PMOS threshold to -1 V. Assume that the threshold shifts are achieved through shallow ion implantations. Neglect oxide charge.

6.2 In a p -well CMOS technology, the n -substrate has a doping concentration of $1 \times 10^{15} \text{ cm}^{-3}$, and the p -well concentration is $5 \times 10^{16} \text{ cm}^{-3}$. Calculate the source/drain junction capacitances and body-effect coefficients for both n - and p -channel MOSFETs assuming the gate oxide is 20 nm thick in both devices.

6.3 In an n -well CMOS process, specify the necessary masking sequence in correct order and also state the number of masks needed a process that uses both a double-poly and double-level metal process.

6.4 A CMOS inverter is driven with a square wave voltage source with a period of 1 μs . The power supply voltage, $V_{DD} = +12 \text{ V}$ and the load capacitance C_L is 20 pF. Find the power dissipated by the inverter.

452 SILICON PROCESSING FOR THE VLSI ERA – VOLUME II

6.5 If an LDD structure is needed for both n - and p -channel MOSFETs in a submicron CMOS process, list the CMOS masking sequence and associated major processing steps. If the doping concentration of the p^+ LDD is an order of magnitude larger than that of the n^- LDD, can any of these steps be omitted?

6.6 Comparing a retrograde well CMOS process to a conventional well CMOS process, what is the most important difference in the two process sequences and in the completed device structures? How do these differences impact the device isolation?

6.7 In a (p^+ diffusion)-(n -well)-(p -on- p^+ substrate) structure, the n -well is uniformly doped at $1 \times 10^{16} \text{ cm}^{-3}$ and is $1 \mu\text{m}$ deep, the p -epi thickness is $4 \mu\text{m}$ and is uniformly doped at $1 \times 10^{15} \text{ cm}^{-3}$, and the p^+ diffusion is $0.4 \mu\text{m}$ deep. If a substrate bias generator is needed and the p^+ diffusion and n -well are both biased at 5 V, what is the maximum (in absolute value) substrate voltage that can be applied before punchthrough occurs?

6.8 Explain why using a lightly doped epitaxial layer on a heavily doped substrate is one of the most effective ways to increase latchup resistance. Why is a thinner epi even more helpful and what limits the continuous scaling of epi thickness?

6.9 For an n -well CMOS structure, draw a lumped equivalent circuit model of the parasitic device structures that give rise to latchup and derive the conditions for latchup to occur.

6.10 Explain why latchup cannot be sustained if the holding voltage is greater than the supply voltage. What is the most effective technique for increasing the holding voltage?

6.11 If the n^+ -to- n -well spacing is $3 \mu\text{m}$, the n -well depth is $1.3 \mu\text{m}$, and the p^+ diffusion depth is $0.3 \mu\text{m}$, estimate the minimum pulse width needed to trigger latchup.

6.12 In an n -well CMOS structure the well depth is $1 \mu\text{m}$ and the well concentration $1 \times 10^{16} \text{ cm}^{-3}$. The p^+ diffusion is $0.3 \mu\text{m}$ deep and it is doped at $1 \times 10^{19} \text{ cm}^{-3}$. Estimate the vertical pnp transistor common-emitter current gain (β) using one-dimensional analysis. (*Hint:* You will need to calculate the minority-carrier diffusion length in the base of the transistor, and compare this with the base width in order to estimate α , which can then be used to find β .)

CHAPTER 8

SEMICONDUCTOR MEMORY

PROCESS INTEGRATION

Memories store digital information (or data) in terms of *bits*, or binary digits (ones or zeros). Modern digital systems use *memory devices* to store and retrieve large quantities of digital data at electronic speeds. Early digital computers used magnetic-cores as the devices in fast-access memories. With the introduction of semiconductor memory chips in late 1960s, however, magnetic cores began to be replaced by integrated circuits (which implemented a much higher-density digital-memory function). This not only increased the performance capabilities of the memory, but also drastically decreased its cost. By the end of the 1970s, magnetic-core memories had been completely displaced as high-speed memory devices.

8.1 TERMINOLOGY OF SEMICONDUCTOR MEMORIES

Memory capacities in digital systems are usually expressed in terms of bits, since a separate storage device or circuit is used to store each bit of data. Each storage element is referred to as a *cell*. Memory capacities are also sometimes stated in terms of *bytes* (8 or 9 bits) or *words* (32 – 80 bits). Each byte typically represents an alphanumeric character. Every bit, byte or word is stored in a particular location, identified by a unique numeric address, and only a single bit, byte, or word is stored or retrieved during each cycle of memory operation.

Memory-storage capability is expressed in units of kilobits and megabits (or kilobytes and megabytes). Since memory addressing is based on binary codes, capacities that are integral powers of 2 are typically used. As a result, a memory device with a 1-kbit capacity can actually store 1,024 bits, and a 64-kbit device can store 65,536 bits.

In digital computers, the number of memory bits is usually 100 to 1000 times greater than the number of logic gates, which implies that the memory cost per bit must be kept very low. In addition, it is desirable for the memory devices to be as small as possible (since this will allow the highest density of cells on a chip), to operate at a high speed, to have a small power consumption, and to operate reliably.

Memory cells could be designed to possess a set of characteristics close to those of an ideal digital-logic-element. Such an ideal cell would be able to

1. perform the desired logic function;
2. robustly quantize the signal levels of the stored data;
3. exhibit a high degree of input-output isolation and fan-out; and
4. regenerate the stored logic-levels.

However, to enable each memory cell to possess all of these attributes would require the use of a complex circuit to implement each cell. Memory-cell design therefore involves trading off most of the desired properties of digital-logic devices in order to achieve a cell that is as simple and compact as possible. Consequently, the cell itself is not capable of outputting digital data in an electrical form compatible with the requirements of the remainder of the system. To restore the electrical characteristics of the cell's outputted data to adequate values, properly designed peripheral circuits (e.g., sense amplifiers, memory registers, and output drivers) are necessary. These circuits are designed to be shared by many memory cells. The trade-off thus made is that of a less-robust output signal from the cell, in exchange for a simple, compact memory cell design (consisting of only 1 to 6 transistors).

8.1.1 Random-Access and Read-Only Memories (RAMs and ROMs)

The most flexible digital memories are those that allow for data storage (or *writing*) as well as data retrieval (or *reading*). Memories in which both of these functions can be rapidly and easily performed, and whose cells can be accessed in random order (independent of their physical locations), are referred to as *random-access memories* (RAMs). *Read-only memories* (ROMs) are those in which only the read operation can be performed rapidly (although ROMs are generally configured so that their cells are also randomly accessible, and data can be entered into them). Entering data into a ROM, however, is referred to as *programming* the ROM, to emphasize that this operation is much slower than the writing operation used in RAMs.

8.1.2 Semiconductor-Memory Architecture

The organization of large semiconductor memories is shown in simplified form in Fig. 8-1.¹ The storage cells of the memory are arranged in an array consisting of horizontal rows and vertical columns. Each cell shares electrical connections with all the other cells in its row, and column. The horizontal lines connected to all the cells in the row are called *word lines*, and the vertical lines (along which data flows into and out of the cells) are called *bit lines*. Each cell therefore has a unique memory location, or address, which can be accessed at random through selection of the appropriate word and bit line. (Some memories are designed so that four or eight cells are accessed simultaneously.) Thus, in semiconductor memories such as that shown in Fig. 8-1, any cell can be accessed in random order, at a fixed rate, for the purpose of either reading

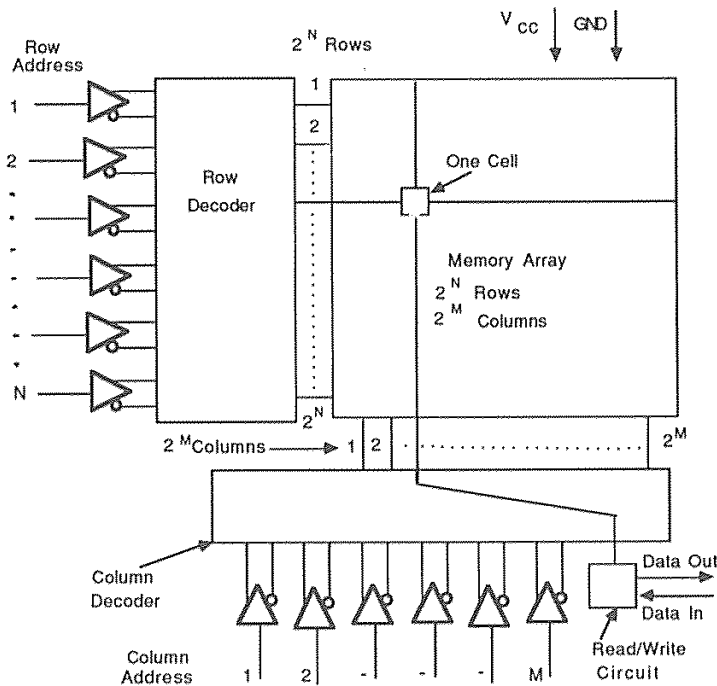


Fig. 8-1 Organization of a random access memory (RAM).

or writing data. The array configuration of semiconductor memories lends itself well to the regular, structured designs favored in VLSI.

There are also a number of important circuits at the periphery of the array. The first such peripheral circuit is the *address decoder*. Two of these are used on each chip – one for the word lines, the other for the bit lines. These circuits allow a large number of word and bit lines to be accessed with the fewest number of address lines. Address decoders for this purpose have 2^n output lines, with a different one selected for each different n -bit input code. In later generations of memory circuits, address multiplexing was integrated on some memory chips to reduce the number of address pins by half. (Note that the address decoder was the first peripheral logic-circuit to be built on the memory chip).

The read/write control circuitry shown in Fig. 8-1 determines whether data is to be written into or read from the memory. Because such circuits also amplify and buffer the data signals retrieved from the cells, one of the important circuits in this subsystem is the *sense amplifier*. In dynamic memories that need periodic data refreshing, refresh circuitry may also be provided.

Note that most RAMs have only one *input-data lead* and one *output-data lead* (or only one combined *input/output lead*). Writing into and reading from such RAMs is done one bit at a time. Other RAMs have a number of input- and output-data leads, with the number determined by the word length of the system's data bus. ROMs, on the other hand, are typically organized so that the number of output-data leads (usually eight) is the same as the number of lines on the data bus. ROMs are programmed word by word (i.e., eight bits, or one byte, at a time) and are read from in the same manner.

8.1.3 Semiconductor-Memory Types

In semiconductor RAMs, information is stored on each cell either through the charging of a capacitor or the setting of the state of a bistable flip-flop circuit. With either method, the information on the cell is destroyed if the power is interrupted. Such memories are therefore referred to as *volatile memories*. When charge on a capacitor is used to store data in a semiconductor-RAM cell, the charge needs to be periodically refreshed, since leakage currents will remove it in a few milliseconds. Hence, volatile memories based on this storage mechanism are known as *dynamic RAMs*, or *DRAMs*.

If the data is stored (i.e., written into the memory) by setting the state of a flip-flop, it will be retained as long as power is connected to the cell (and no other write signals are received). RAMs fabricated with such cells are known as *static RAMs*, or *SRAMs*. Volatile RAMs can be treated as nonvolatile if they are provided with battery backup. Some DRAM and SRAM chips are even packaged together with a battery to facilitate implementation of this approach.

It is often desirable to use memory devices that will retain information even when the power is temporarily interrupted (or when the device is left without applied power for indefinite periods). Magnetic media offer such nonvolatile-memory storage. In addition, a variety of semiconductor memories have been developed with this characteristic. At present, virtually all such nonvolatile memories are ROMs. While data *can* be entered into these memories, the programming procedure varies from one type of ROM to the other (and none of them can be considered to be RAMs).

The first group of nonvolatile memories consists of those ROMs in which data is entered during manufacturing, and cannot subsequently be altered by the user. These devices are known as *masked ROMs* (or simply *ROMs*). The next category consists of memories whose data can be entered by the user (*user-programmable ROMs*). In the first example of this type, known as a *programmable ROM*, or *PROM*, data can be entered into the device only *once*.

In the remaining ROM types, data can be erased as well as entered. In one class of erasable ROMs, the cells must be exposed to a strong ultraviolet light in order for stored data to be erased. These ROMs are called *erasable-programmable ROMs*, or *EPROMs*. In the final type, data can be *electrically erased* as well as entered into the device; these are referred to as *EEPROMs*. The time needed to enter data into both EPROMs and EEPROMs is much longer than the time required for the *write* operation in a RAM. As a result, none of the ROM types can at present be classified as fully functional RAM devices.

A few nonvolatile RAMs have been developed, but these have not yet reached the state of development at which they can be considered an important class of semiconductor memory.

8.1.4 Read-Access and Cycle Times in Memories

The two principal time-dependent performance characteristics of a memory are the *read-access time* and the *cycle time*. The first is the propagation delay from the time when the address is presented to the memory chip until data stored at that address is available at the memory output. The cycle time is the minimum time that must be allowed after the initiation of a read operation (or a write operation, in a RAM) before another read operation can be initiated. The minimum cycle times for reading and writing in a RAM are not necessarily equal, but for simplicity of design most systems employ a single minimum cycle time. For semiconductor RAMs, the read-access time is typically 50 – 90% of the read-cycle time.

8.1.5 Recently Introduced On-Chip Peripheral Circuits

Additional peripheral circuits have recently been added to the basic memory-organization structure shown in Fig. 8-1. These circuits serve mainly to improve the manufacturability and testability of the chips. Those designed to increase manufacturability include redundancy circuits and error-correction circuits. Redundancy circuits allow some defective chips to be salvaged, while self-testing circuits reduce testing time.

Redundancy allows a defective row or column of cells to be replaced with a spare. Replacement techniques include the use of electrically or laser-blown fuses, or of one-time-programmable memory cells (which control on-chip multiplexers that switch in spare rows or columns). Redundancy measures typically improve manufacturing yields by factors of between 1.5 and 5.

Error-detection and correction techniques involve the addition of parity bits to allow the system to detect bad data, as well as circuitry to accomplish parity checking and error correction. This imposes an area penalty: for example, if one parity bit is added to each byte, the size of a 32k x 8k chip would be increased to 32k x 9k. If a failure is detected during programming, the parity bits can be sacrificed to act as a spare bit field for byte-wide applications. In DRAM designs, the use of error-correction coding (ECC) requires another 27% of memory cells. Because the ECC approach corrects soft errors as well as hard errors, the problem of soft errors can be reduced for the life of the product.

8.1.6 Logic-Memory Circuits

Special-purpose circuits that combine logic and memory on the same chip have recently been introduced. These fall into two categories: (1) memories that have some additional logic capabilities (*logic-in-memory circuits*, or *special-application RAMs*); and (2) logic circuits that contain some memory capability (*memory-in-logic circuits*).

A number of different special-application RAMs have been developed, including *video RAMs* (VRAMs) and *multiport SRAMs*.⁴⁵ Video RAMs are DRAMs designed to support the high-capacity requirements of frame buffers and display memories found in graphics terminals and systems. They have two input/output ports – one for random access (as in conventional RAMs), and one for serial-access. The serial port accesses the memory sequentially and performs the various serialization tasks necessary to drive cathode-ray tubes or other serial-data devices. The fast serial-readout-rate also enables quick refreshing of the graphics screen. The random-access port is driven by a graphics processor and is used to build the screens of displayed data (i.e., the random-access capability allows real-time updating of pictures).

Multiport SRAMs are being offered for use in multiprocessor systems. For example, dual-port SRAMs are becoming widely used to allow two independent logic circuits to simultaneously access one memory in a read and write mode. The two circuits can thus communicate with each other by passing data through the common memory. (Two processor-containing components of a digital system might be a CPU and a disk controller, or two processors working on two related but different tasks.) The use of the dual-port memory would eliminate the need for any special data-communication hardware.

Many applications for memory-in-logic circuits have also been envisioned, but these have only begun to be implemented. Some manufacturers of *application-specific circuits* (ASICs) do offer large memory blocks as part of prefabricated gate arrays, as well as increasingly larger memory blocks in their standard-cell libraries. Other, more advanced circuits are still being developed. For instance, more extensive use of memory will be needed in *image processing* and *coder-decoder* (CODEC) ULSI circuits. Image processing systems will be greatly enhanced when a processor is available with a memory of at least 2 Mbits that can store a frame of a picture. Similarly, CODECs used to move pictures could use such an image processor with an on-chip memory for data compaction. Finally, microprocessor chips are being built with on-board memories. (For example, the Intel 486 μ P contains *cache memory* on the chip, which consumes ~40% of the chip area.) Even more memory integration might allow for an on-chip memory hierarchy. Products that could use such circuits include *point-of-sale terminals*, *smart cards*, and *telecommunications circuits*.

8.2 STATIC RANDOM-ACCESS MEMORIES (SRAMS)

SRAMs, the first type of semiconductor memory to be implemented, are referred to as *static memories* because they do not require periodic refresh signals in order to retain their stored data. The bit state in an SRAM is stored in a pair of cross-coupled inverters, which form a circuit known as a flip-flop. The voltage on each of the two outputs of a flip-flop circuit is stable at only one of two possible voltage levels, because the operation of the circuit forces one output to a high potential, and the other to a low potential. The memory logic state of the cell is determined by whichever of the two inverter outputs is high. Flip-flops maintain a given state for as long as the circuit

receives power, but they can be made to undergo a change in state (i.e., to flip), through the application of a trigger voltage of sufficient magnitude and duration to the appropriate input. Once the circuit has settled into its new stable state, the trigger voltage can be removed. SRAM cells can be implemented in NMOS (Fig. 8-2a), CMOS (Fig. 8-2b), bipolar (Fig. 8-2c), or BiCMOS technologies.

The chief disadvantage of an SRAM cell is that it consists of at least six devices, as compared to only two for the dynamic-memory cell. Thus, even when the same set of design rules is used, an SRAM chip cannot be built with as many cells as a DRAM chip (as is illustrated in Figs. 8-3a and 8-3b).

On the other hand, SRAMs are the fastest semiconductor memories. Their speed is derived from the self-restoring nature of the flip-flop and the static peripheral circuits of the memory chip. Bipolar SRAMs are the fastest of all, and MOS SRAMs are the fastest among MOS memories. Bipolar SRAMs, however, dissipate much more power than CMOS SRAMs (e.g., 0.1-to-1.0 mW/bit versus $\sim 25 \mu\text{W/bit}$).

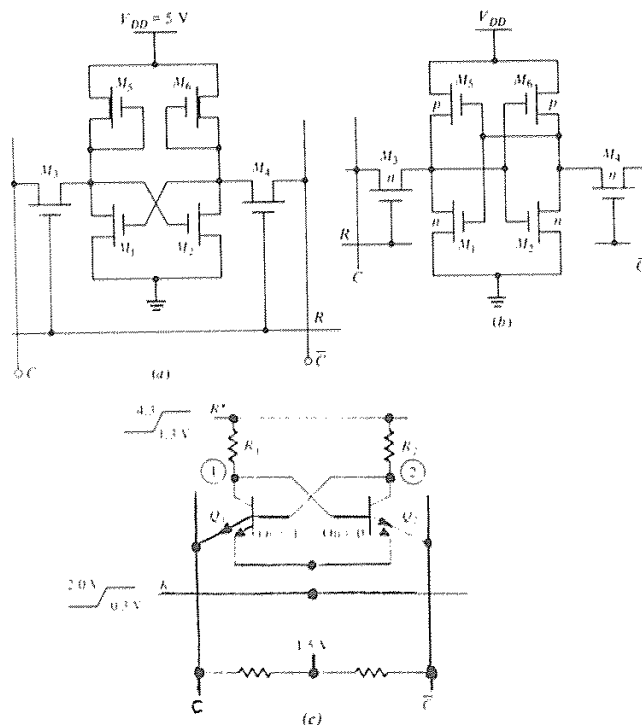


Fig. 8-2 Circuit schematic of (a) NMOS SRAM cell, (b) CMOS SRAM cell, (c) Bipolar SRAM cell.¹ From D. A. Hodges and H. G. Jackson, *Analysis and Design of Digital Integrated Circuits*, Copyright, 1983 McGraw-Hill Book Co. Reprinted with permission.

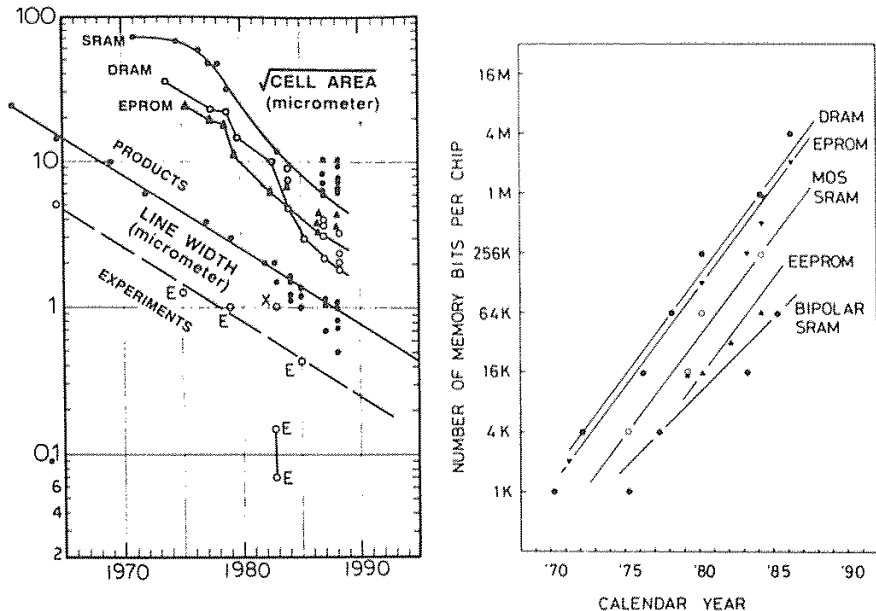


Fig. 8-3 (a) Trends of the laboratory and the production lithographic line width (two lower straight lines) and of the cell size (the square root of cell area) of production SRAM, DRAM, and EPROM cells.³ (© 1988 IEEE). (b) Level of integration of semiconductor memories which have been presented at the *IEEE International Solid-State Circuits Conference* versus calendar year.² (© 1986 IEEE).

SRAMs are also characterized by their input/output (I/O) capability: one group has a TTL I/O capability, while the other has an ECL I/O capability (see chap. 7, section 7.2.2 for a comparison of TTL and ECL I/O signals). For the past several generations, high-speed CMOS technologies have been used to build TTL I/O SRAMs, while ECL I/O SRAMs have been implemented using bipolar technology. This is changing, since BiCMOS technology has begun to be used to fabricate both TTL and ECL I/O SRAMs. For example, 256-kbit ECL I/O SRAMs with 15-ns access times were commercially introduced in 1989, and 1-Mbit devices are expected shortly. TTL-I/O BiCMOS SRAMs are challenging incumbent CMOS and bipolar devices and already offer higher performance than all but the best of the CMOS SRAMs.

As SRAMs have evolved, they have undergone an increase in density. Most of this has been due to the use of smaller line widths (e.g., the 4-kbit MOS SRAM used 5- μm lines, while the 16-kbit, 64-kbit, 256-kbit, and 1-Mbit SRAMs were built with 3.0- μm , 2.0- μm , 1.2- μm , and 0.8- μm lines, respectively). The remainder of the density increase has been due to improvements in process technology, novel cell designs, and circuit innovations.²

Figure 8-2 indicates that one word line and two bit lines are connected to each SRAM cell; consequently two access transistors are also provided in each cell. In principle, it should be possible to achieve all memory functions using only one column line, one bit line, and one access transistor. In practice, however, normal variations in device parameters and operating conditions, make it difficult (or impossible) to obtain reliable operation at maximum speed using a single access-line to flip-flop cells. Therefore, the symmetrical bit lines (bit 0 and bit 1), are necessary. In a matrix of memory cells, each pair of bit-0 and bit-1 lines is shared by all memory cells in each column, and each word line is shared by all memory cells in each row.

8.2.1 MOS SRAMs

MOS SRAMs can be fabricated in either NMOS and CMOS, with early MOS SRAMs implemented in the former (Fig. 8-2a depicts a schematic diagram of an NMOS-SRAM cell). The load devices, M_5 and M_6 , are depletion-mode NMOS transistors; the driver transistors, M_1 and M_2 , and the access transistors, M_3 and M_4 , are enhancement-mode NMOS transistors. The fully static CMOS SRAM (*full-CMOS*) cell shown in Fig. 8-2b was developed next. In this cell, the load devices are PMOS enhancement-mode transistors, while the other four transistors are enhancement-mode NMOS devices. Finally, a four-transistor SRAM cell was developed, with high-valued polysilicon resistors used as the load devices (*poly-load cell*). Since all the transistors in this cell are NMOS devices, the cells can be built in CMOS p -wells. When an array of such poly-load cells was fabricated in a p -well and combined with full-CMOS peripheral circuits, the resulting SRAMs demonstrated a significant decrease in power consumption compared to NMOS SRAMs (since most of the power in an SRAM is dissipated in the peripheral circuits, such as the off-chip drivers). These SRAMs also displayed a higher packing density than did SRAMs built with fully static CMOS cells.⁴

The full-CMOS cells dissipate less power than do the other types of MOS SRAM cells when in the standby mode of operation (i.e., when the cell is not being written into or read from). In the other types of cells, one inverter is always *ON*, and hence significant current is drawn from V_{DD} (much more, of course, in the six-transistor NMOS cell). In the full-CMOS cell, however, one transistor in each of the coupled inverters is *OFF*; thus, only junction-leakage current is drawn from V_{DD} . This current is approximately three orders of magnitude less than that drawn by the poly-load cell,⁵ demonstrating that very little power is dissipated in the full-CMOS cell. In addition, the stability of the full-CMOS cell is high, since higher alpha-particle immunity and smaller junction leakage sensitivity is exhibited. The insensitivity to leakage allows operation at higher temperatures. Fully static CMOS SRAMs have been exploited as low-power, battery-backed memory devices in battery-operated consumer goods and portable office equipment.

The isolation of n -channel from p -channel devices means that the full-CMOS cell generally requires a larger area than does the poly-load cell. Furthermore, in order to establish contacts from the drain regions of the p -channel devices to those of the

n -channel devices (as well as to the n^+ polysilicon-gate materials), the metal layer must be used. In four-transistor/poly cells, there are no p -channel devices; so no n -channel-to- p -channel isolation is needed in the memory array of the chip. Buried contacts which take up less space can also be used to connect the drains of the access transistors and the gate of driver-transistors. Since it is also more costly to manufacture SRAMs with full-CMOS cells, four-transistor/poly-load cells have been used in the design of most high-density CMOS SRAMs. By 1989, 1-Mbit CMOS SRAMs were being offered commercially.

Alternative cell structures have also been investigated as a way to increase the density of SRAMs. One such approach is to stack the transistors on top of one another. For example, the full-CMOS cell can be built with the active p -channel transistor load stacked above the n -channel devices. The second layer of transistors can be fabricated on recrystallized silicon¹² or built using a hydrogen-passivated polysilicon transistor.¹³ Although the processing techniques needed to fabricate such three-dimensional stacked structures are complex and difficult to control, these devices will become more attractive as the need to form higher-density structures becomes greater.

A novel SRAM cell based on the reverse base current of a bipolar transistor and consisting of only one bipolar transistor and one MOS transistor was described by Sakui et al.¹⁴ Such a compact cell would allow SRAMs to be built with the same densities as DRAMs.

8.2.1.1 Circuit Operation of MOS SRAM Cells. NMOS and CMOS SRAM cells all exhibit the same basic circuit behavior (Figs. 8-2a and b). When writing or reading data in such a cell is desired, the word line of the cell (which is held low in the standby state) is raised to V_{DD} (e.g., +5 V). This causes the enhancement-mode NMOS access transistors M_3 and M_4 to be turned *ON*. *Writing* is performed by forcing one of the bit lines low (e.g., close to 0 V), while maintaining the other at its standby value (about 3 V). For example, to write a *1*, the bit-0 line must be forced low.*

When this occurs, M_1 turns *OFF* and its drain voltage rises due to the currents flowing through M_5 and M_3 . When M_2 has been turned *ON*, the bit line can be returned to its standby level, leaving the cell in the state of storing a *1*. (The operation of writing a *0* is complementary to that just described).

For *reading* a *1*, the bit lines must both be biased at about 3 V. When the cell is selected, current flows through M_4 and M_2 to ground and through M_5 and M_3 to the bit *1* line. The gate voltage of M_2 does not fall below 3 V, so M_2 remains *ON*. The voltage of the bit-0 line is thus reduced to less than 3 V, while the voltage of the bit-1 line is pulled up above 3 V, since M_1 is *OFF* but M_5 is *ON*. As a result, a differential output signal exists between the bit-0 and bit-1 lines. This signal is fed to the sense amplifier, which in SRAMs is a differential amplifier capable of providing rapid sen-

* The cell must be designed so that the conductance of the access transistor is several times larger than that of the load transistor (i.e., comparing M_4 to M_6), in order for the drain of M_2 and the gate of M_1 to be brought below V_T .)

sing. Consequently, one of the bit lines needs to be only slightly discharged in order to generate a differential input signal large enough to drive the sense amplifier. To avoid a change in the state of the cell during reading, however, it is necessary for the conductance of M_2 to be around three times as large as that of M_4 so that the drain voltage does not rise above V_T . (The operation of reading a 0 is complementary to the one just described).

Among the most important factors limiting the maximum speed of MOS SRAMs are the delay associated with signal propagation through address buffers and decoders (which gets longer as the number of inputs and outputs increases), and the delay associated with the charging and discharging of the word and bit lines (which increases as the RC product of the word- and bit-line structures increases).

Bit-lines are typically formed in metal (Al), and hence their resistance is not a significant limitation. Word-lines, however, are normally implemented with polysilicon or polycide, so their higher resistance is considerably larger than that of the bit-lines. This then becomes one factor that limits SRAM speed. The parasitic capacitance of the word- and bit-line structures themselves, combined with the many paralleled access transistors (which are connected to each word and bit line), results in a large equivalent lumped capacitance on each of these lines.

Finally, there is a delay associated with the signal propagation through the sense-amplifier and data-output circuits. Considerable effort has been expended to develop high-speed sense amplifiers for SRAMs. In addition, a circuit technique known as *address-transition detection* (ATD) has also been used to speed up sensing in MOS SRAMs. In this technique, the bit lines are equilibrated upon detection of a change in the address input. (Sense-amplifier design, and the details of ATD are discussed in texts on VLSI circuit design, and so will not be further described here.

8.2.1.2 SRAM Cell Layout and Processing Issues. For maximum density to be achieved in a memory device, the cells must be laid out in as small a size as possible. The size is determined by the cell's topology and by the design rules of the IC fabrication technology. A completed layout design represents the outcome of years of development, and a great deal of design experience. Figures 8-4a, 4b, and 4c are examples of SRAM-cell designs for a six-transistor NMOS SRAM cell,⁶ a poly-load cell,⁸ and an advanced full-CMOS cell,⁷ respectively. These cell layouts also reflect some of the process enhancements that have made possible the improvements in SRAM performance, speed, and density.

Table 8-1 shows the evolution of the MOS SRAM; shrinking line widths and a variety of process enhancements can be seen to have been primarily responsible for the density and performance improvements. Figure 8-5 presents the same information graphically. The process enhancements are summarized in the following paragraphs describing the cells used in 1-Mbit SRAMs.

Several 1-Mbit CMOS SRAMs based on the poly-load cell were described in detail in the *IEEE Journal of Solid-State Circuits* (October, 1988), and a 4-Mbit SRAM (October, 1989).¹⁰⁹ The access time of these devices ranges from 7.5 to 18 ns (although the access times of commercially available 1-Mbit SRAMs in 1989 were

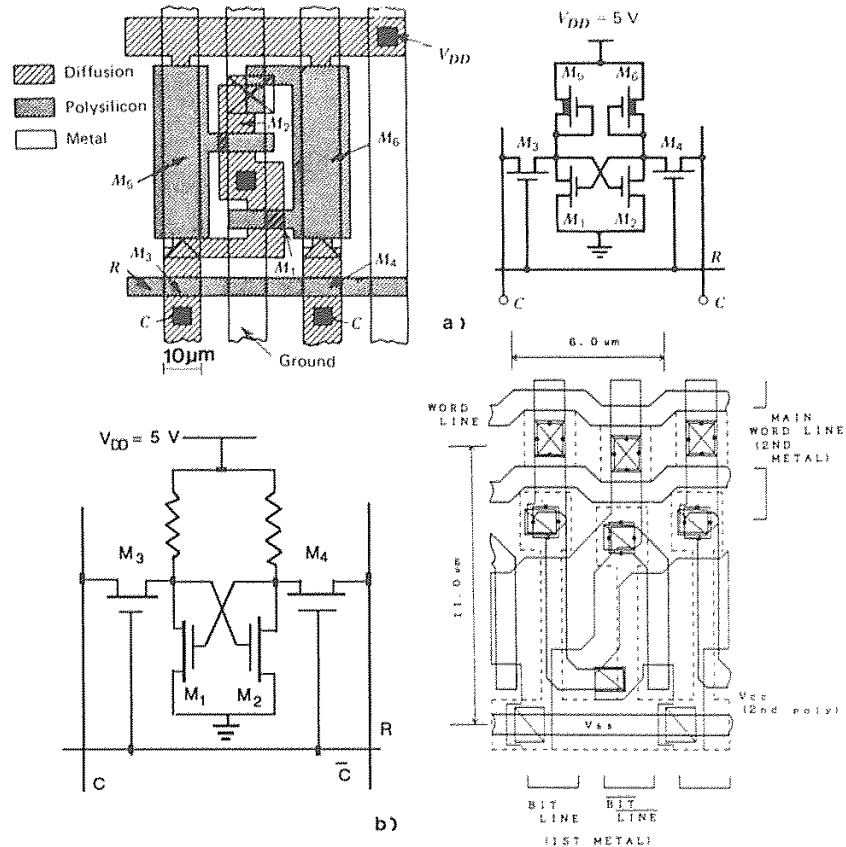


Fig. 8-4 (a) Schematic and layout of NMOS SRAM cell.⁶ After R. Hunt, "Memory Design and Technology," in M. J. Howes and D. V. Morgan, eds., *Large Scale Integration*. Copyright 1981, John Wiley & Sons. Reprinted with permission. (b) Schematic and layout of poly load SRAM cell.⁸ (© 1988 IEEE).

being given as 25-120 ns).³² From the layout of the 7.5-ns CMOS SRAM cell (Fig. 8-4b), it can be seen that a double-polysilicon, double-level metal process is used, and that the cell area is $66 \mu\text{m}^2$ ($6.0 \times 11 \mu\text{m}$). The first level of polysilicon is a polycide structure, which is used for the V_{SS} power line in the memory array as well as for the gates of the MOS transistors. The second poly layer is used to form both the high-valued load resistors and the low-resistance V_{DD} lines. The bit lines are formed in Metal 1, and the word lines in Metal 2. A $0.8\text{-}\mu\text{m}$ twin well-CMOS process is used in which the channel lengths of the n - and p -channel transistors are 0.8 and $1.0 \mu\text{m}$, respectively.

Other advanced poly-load-based CMOS SRAM designs include such process enhancements as trench-isolation structures, triple-level polysilicon, self-aligned con-

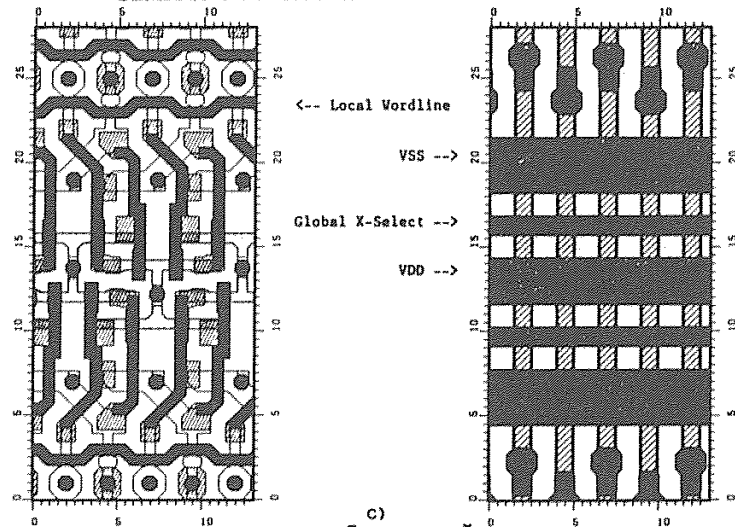


Fig. 8-4 (c) Layout of a full-CMOS SRAM cell.⁷ (© 1988 IEEE).

tacts, buried contacts, and spare rows and columns for redundancy.

A 1-Mbit CMOS SRAM based on a full-CMOS cell was reported in the same issue of the *IEEE Journal* mentioned above (Fig. 8-4b).⁷ It has a longer access time (25 ns) but consumes only 1 μ W of power in the standby mode. Its memory-cell size is $5 \times 12 \mu\text{m}$, which implies that clever cell design and advanced processing techniques can produce fully static cells with sizes comparable to those of poly-load cells. (Note that the major reduction of cell area in such full-CMOS cells has been attributed to the use of a local interconnect layer¹⁰⁸ — chap. 3, section 3.11.2. The local interconnect layer allows the two inverters of the cell to be cross coupled with only two contacts per cell

Table 8-1. Evolution of MOS SRAM Technology

Introduction Date	Size (bits)	Access Time	Minimum Feature Size	Process Enhancements
1969	PMOS 256 bit			Silicon Gate, CVD Oxide
1972	NMOS 1k		8 μm	Depletion-Mode Load
1975	NMOS 4k	4 ns (1988)	5 μm	Ion-Implant V_T Adjust
1978	NMOS 16k		3 μm	Plasma Etching /Wafer Stepper
1982	CMOS/NMOS 64k	15 ns	2 μm	Double-Poly
1985	CMOS/NMOS 256k	25 ns (1988)	1.2 μm	Polycide/Poly, LDD Structures
1988	CMOS/NMOS 1M	25 ns (1988)	0.8 μm	(Polycide/Poly, Double-Metal, Twin-Well, LDD Structures)
	Full CMOS 1M	25 ns (1988)		
1989	CMOS/NMOS 1M	10 ns		
	CMOS/NMOS 4M	25 ns	0.5 μm	3.3 V, Retrograde p -Well,
	BicMOS 1M	8 ns	0.8 μm	25 Mask Levels, Twin-Well

– compared to nine contacts in a cell implemented with double-level metal, but with no local interconnect level – and thus the cell size can be made significantly smaller.)

This SRAM is fabricated using a 14-mask process, and it employs a single level of polysilicon and two layers of metal. The poly layer is selectively doped *p*-type when it acts as the gate for PMOS devices, and *n*-type when it is a gate for NMOS devices. A silicide strap is also used to connect poly lines to diffused regions. Spare rows and columns are included for redundancy.

Another full-CMOS SRAM (256 kbits in size) has also been described.⁹ This 35-ns access-time part uses TiN as a local-interconnect structure between gates and diffused regions, and 0.8- μm MOS devices. The 100-nm-thick TiN layer has a sheet resistance slightly lower than that of a 500-nm doped-poly layer (14 Ω/sq vs. 20 Ω/sq). The TiN makes contact to the TiSi_2 layer that is formed on the surfaces of both the diffusion and gate regions. Since TiN is also an effective diffusion barrier, it prevents the phosphorus dopant in the n^+ polycide structure from diffusing and counterdoping the diffused drain regions of the PMOS devices when a connection is formed between them. (The formation and properties of TiN as an interconnect and barrier material is discussed in greater detail in chap. 3).

Several 4-Mbit CMOS SRAMs are described in the October, 1989 issue of the *IEEE Journal of Solid-State Circuits*. The decrease in SRAM cell size as a function of minimum feature size is shown in Fig. 8-5.

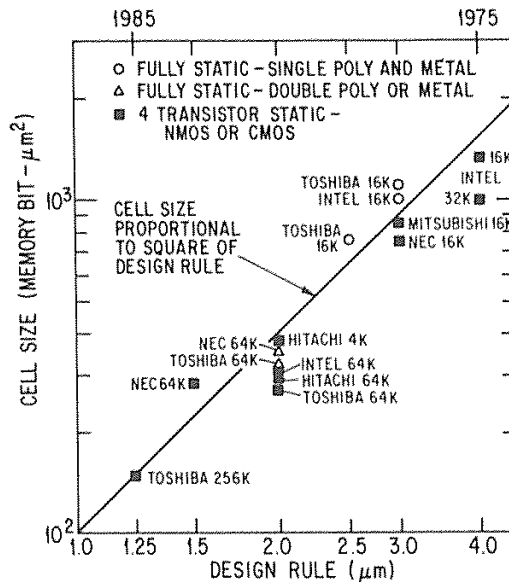


Fig. 8-5 MOS SRAM cell size versus design rule history.¹¹⁶ (© 1986 IEEE).

8.2.1.3. High-Valued Polysilicon Load Resistors for MOS SRAMs

High-valued resistors are used as the load devices in the poly-load SRAM cell (Fig. 8-2b). In order to minimize power consumption and yet maintain an optimum soft-error rate, the load current of the cell is set to about 31 pA.¹⁵ Very high-valued load resistors must be used to obtain such small load currents. For example, it has been calculated that 164 G Ω resistors must be used for 64-kbit and 256-kbit SRAMs, and 97 G Ω resistors are needed for 1-Mbit and 4-Mbit SRAMs (Fig. 8-6a).¹⁵ Films made of materials with very high sheet resistances must be used to fabricate these load resistors to avoid the consumption of excessive area.

Undoped polysilicon films exhibit high sheet-resistivity values, making them good candidates for fabricating such structures (Fig. 8-6b). When undoped polysilicon films are implanted with arsenic in doses from $\sim 1 \times 10^{13}/\text{cm}^2$ to $1 \times 10^{15}/\text{cm}^2$, the sheet resistivity can be controlled from $10^4 \Omega/\text{sq}$ up to about $10^{12} \Omega/\text{sq}$ (Fig. 8-6c). Hence, to fabricate a high valued resistor (for example, a 97-G Ω resistor for a 1-Mbit SRAM cell), a polysilicon film with a sheet resistance of 26 G Ω/sq can be used. This sheet resistance can be obtained with an As implant dose of $\sim 3 \times 10^{13}/\text{cm}^2$. The length of a 97-G Ω resistor fabricated in such a 50-nm-thick, 1.2- μm -wide line of polysilicon would be 4.0 μm .

Undoped polysilicon exhibits such high resistivity because some of the impurities in the films segregate to the grain boundaries and do not effectively produce free carriers. In addition, the grain-boundary regions trap some of the free carriers that are produced (see Vol. 1, chap. 6).

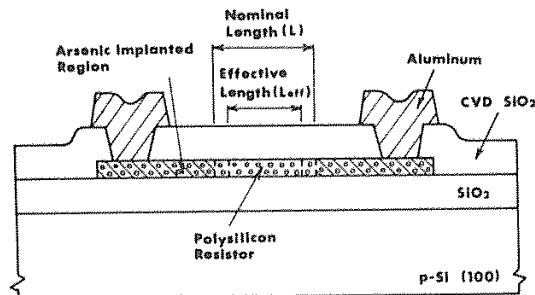
The I-V characteristics of the high-valued polysilicon resistors predicted by the trapping model of Lu et al. fit the experimental data fairly well if the resistor length is not too short.¹⁶ On the other hand, lateral diffusion from adjacent higher-doped regions in the poly can significantly alter the resistance value if such diffusion takes place over a large enough fraction of the resistor length. Because the potential energy barrier to diffusion along the grain boundaries is lower than that in the bulk (see Vol. 1, chap. 8), the rapid diffusion of impurities along grain boundaries can bring impurities to the lightly doped poly regions, even at relatively low temperatures.

The effect just described can be important in the design of polysilicon-load SRAM cells. The resistors are normally formed in the second polysilicon layer, and the remainder of this layer is implanted with a much higher dose (so that it can serve as a low-resistance interconnect path). During this implant, the high-resistivity poly regions are covered with a mask to avoid impurity doping. The minimum size of the mask is limited by the effect of the lateral diffusion of impurities from the highly doped regions during the activation anneal of the polysilicon following ion implantation (e.g., 950°C for 30 min). Hence, a lower limit of about 3 μm was initially predicted for the length of such resistors.

A technique for reducing the extent of the lateral diffusion by implanting the polysilicon with a very heavy dose of oxygen ($\sim 1 \times 10^{22}/\text{cm}^3$) has been reported.¹⁷ High-valued resistors can be fabricated with lengths as small as 0.8 μm (Fig. 8-6d). The oxygen apparently segregates to the grain boundaries, retarding the diffusion of the

Memory size	Feature Size	Power Supply Voltage	Load Current per Bit	Typical Memory Standby Current	Load Resistance	Memory Cell Size	L/W of Load Resistor	Sheet Resistance	Thickness of Poly-silicon Resistor	Chip Size	
(bits)	(μm)	(V)	(pA)	(μA)	(G Ω)	(μm^2)	(μm)	(G $\Omega/\square$)	(nm)	(mm ²)	
[14] 64K	2.0	5.0	31	2	164	16 $\times$ 19 (304)	7.0/2.0	47	100	5.44 $\times$ 5.80	
Estimated Value	256K	1.2	5.0	31	8	164	10 $\times$ 11 (110)	4.0/1.2	47	70	$\sim$ 6.5 $\times$ 7.5
	1M	0.8	3.0	31	33	97	7.0 $\times$ 7.5 (52.5)	3.0/0.8	26	50	$\sim$ 8.5 $\times$ 9.5
	4M	0.5	3.0	31	130	97	3.4 $\times$ 4.2 (14.3)	2.0/0.5	24	30	$\sim$ 8.5 $\times$ 10
	16M	0.25	1.5	31	520	48	1.7 $\times$ 2.1 (3.6)	1.0/0.25	12	30	$\sim$ 8.5 $\times$ 10

a)



b)

Fig. 8-6 (a) Comparison of parameters about the load resistors in SRAM cells. (b) Schematic cross section of a polysilicon resistor.

arsenic (and perhaps also increasing the potential barrier height by forming silicon-oxygen bonds).

The high sheet resistance of polysilicon-load resistors is also reduced by hydrogen diffusion into the polysilicon from plasma-deposited nitride passivation films. This was found to be controllable by sandwiching the polysilicon film with an LPCVD silicon-nitride film (which contains much less hydrogen than does plasma-deposited nitride; see Vol. 1, chap. 6).¹⁹

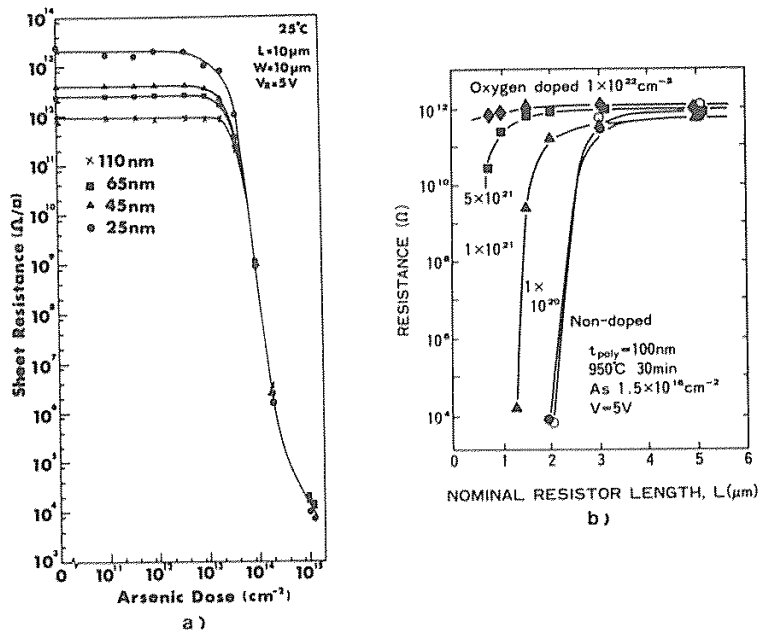


Fig. 8-6 (c) Sheet resistance versus arsenic dose for ion-implanted polysilicon resistors.¹⁵ (© 1985 IEEE). (d) Resistance versus nominal resistor length for oxygen-doped and undoped polysilicon resistors. Resistance is normalized to $1\text{-}\mu\text{m}$ width.¹⁷ (© 1987 IEEE).

8.2.1.4 Soft Errors in SRAMs. SRAMs offer better resistance than DRAMs to both transient and total-dose radiation, making them better suited for some military and space applications. Until recently, the soft-error rates of SRAMs (see section 8.3.4) were negligible compared to those of DRAMs. However, as geometries have been scaled down to produce circuits of greater density, alpha-particle-induced soft-error rates have also become a concern in SRAMs.³³ Although *p*-well CMOS in itself raises the threshold against soft-error failures, the use of extra buried *p*-layers has also been explored as a way to reduce such errors by an additional three orders of magnitude.¹⁰ In addition, the full-CMOS cells exhibit less susceptibility than poly-load cells to single-event upsets and soft errors.

CMOS/SOS technology provides inherently harder parts than does bulk CMOS, and SRAMs have thus been built in CMOS/SOS for such applications. A report detailing the causes of the increase in soft-error rates in densely packed SRAM cells is given in reference 11. Another report on the modeling of alpha-particle sensitivities of SRAMs indicates that state-of-the-art CMOS SRAMs from 64 kbits to 1 Mbit can be made to exhibit sufficient insensitivity to alpha particles.¹⁸

8.2.2 Bipolar and BiCMOS SRAMS

Although MOS SRAMs have achieved much higher device densities on a chip, as well as lower cost and lower power per bit, bipolar SRAMs using emitter-coupled logic (ECL) technology are still faster. Hence, they are chiefly used in applications where highest-speed operation is required (e.g., in the cache memory of high-speed computers). ECL SRAMs are classified into two groups: *high speed* (7-15 ns access times), and *ultra-high speed* (<7 ns access time).²⁴ The fastest 16-kbit bipolar SRAMs have access times of <4 ns,²⁰ and a subnanosecond 5-kbit bipolar SRAM has been reported.^{21,31} As noted earlier, such bipolar SRAMs exhibit an ECL I/O capability.

The emitter-coupled cell shown in Fig. 8-2c is the most widely used bipolar SRAM cell. The load devices affect both the current in the standby mode and the saturation conditions of the driver transistors. In addition, since these devices also determine the read/write current, they have a significant impact on the access time. In early bipolar SRAMs the load for the flip-flop was simply a resistor, which caused the driver transistors to saturate when accessed. In more recent high-speed ECL SRAMs, another cell (as shown in Fig. 8-7a), has been used that utilizes a *pnp* transistor as the load. The advantage of this is that it allows for a smaller cell size, since the transistors are fabricated in parasitic elements. However, because the loads still allow the driver transistors to saturate, SRAMs which use such cells cannot achieve ultra-high speeds. Hence, such cells are used in medium-speed, high-density bipolar SRAMs.²³

In ultra-high-speed ECL cells, saturation is avoided through the use of a Schottky diode in combination with resistors (the so-called *Schottky-barrier-diode, [SBD] switched-load-resistor cell*; Fig. 8-7b).²² The load-resistance value in this cell is high during standby and low when active, making the current during standby about one-thousandth of that drawn by the cell during reading and writing.

Although this cell was invented quite some time ago, it did not gain wide acceptance because of its low density. Recent cell designs that incorporate trench isolation (to allow closer spacing of the transistors) and two levels of polysilicon (one for the interconnections and the other for the resistors) have allowed cells about half the size of conventional SBD cells to be realized.

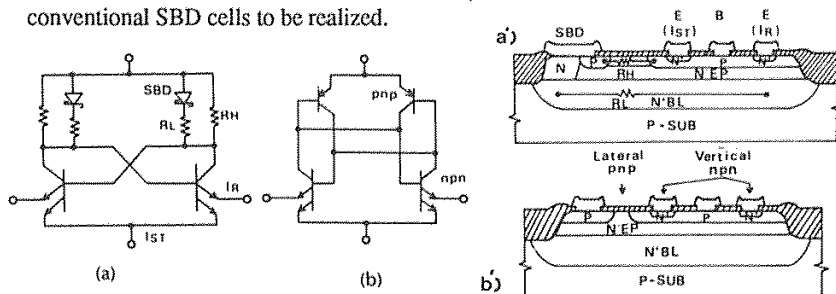


Fig. 8-7 Equivalent-circuit diagrams and cross-sectional views of (a) switched-load memory cell, and (b) cross-coupled *pnp* load cell.² (© 1986 IEEE).

In ultra-high-speed SRAMs, the Schottky diode speeds the switching response and increases the soft-error immunity by sustaining most of the stored charge on the SBD capacitance. A large SBD capacitance is therefore needed in order for high reliability to be maintained. The SBD capacitance, however, is quite small per unit area ($2.8 \text{ fF}/\mu\text{m}^2$). An alternative cell that incorporates a separate Ta_2O_5 capacitor has been reported.^{24,27} The use of this capacitor, which has a capacitance of $8.5 \text{ fF}/\mu\text{m}^2$, makes it possible to reduce the cell size by 30% compared to a conventional SBD cell.

Another ECL-SRAM cell using polysilicon diodes as the load elements has been developed.¹¹³ The advantages of this approach include: compact cell size, very low standby current, very low parasitics, and small active junction area. Access times of 1.5 ns for a 1-kbit SRAM have been demonstrated.

Process innovations that have been incorporated to allow faster, higher-density SRAMs include the use of electromigration-resistant aluminum alloys (to permit higher current densities in metal interconnect stripes) and U-groove isolation (see chap. 2) which reduces the isolation width by a factor of three compared to fully recessed LOCOS isolation (see Fig. 2-35a).

Circuit techniques have also been used to enhance the performance of bipolar SRAMs. For example, read/write current has been concentrated in the active region of device operation, and the word delay has been reduced through the use of Darlington drivers. Table 8-2 summarizes the evolution of bipolar SRAM technology.

Table 8.2 Bipolar SRAM Evolution

Introduction Date	Access Size	Access Time	Load Device	Process Enhancements	Circuit Enhancements
1975	1 k	1.5 ns	Resistor	Al-Cu	Non-Saturated Read/Write Current
1978	4 k	2.2 ns	Schottky Diode		Darlington Drivers
1982	16 k	3.0 ns	<i>pnp</i> Transistor	U-Groove Isolation	
1986	64 k	5.0 ns			

8.2.2.1 BICMOS SRAMs. Although high-speed ECL SRAMs up to 64 kbits in size have been fabricated, such large bipolar SRAMs have power dissipation and yield problems. Power dissipation increases because each cell draws a minimum standby current of about $2 \mu\text{A}$ to maintain sufficient noise margins and immunity from alpha particle soft errors. Defects in the narrow base region make it difficult for high-yielding circuits containing 262,144 narrow bases to be produced (i.e., each cell of a 64-kbit ECL SRAM designed with *pnp* loads contains four transistors).

On the other hand, CMOS alone cannot be used to build such high-performance, higher-density SRAMs because the driving capability of CMOS is inferior to bipolar, and it is practically impossible to design an input and output buffer circuit in CMOS that has an ECL I/O capability.

SRAMs have been developed which combine both bipolar and CMOS devices on the same chip. A comparison of 64-kbit SRAMs built using ECL and BiCMOS techno-

Table 8-3 Comparison of 64-kbit Bipolar and BiCMOS SRAMs²⁴

	1.3- μm BiCMOS	2.0- μm BiCMOS	1.2- μm Bipolar	Unit
Organization	64 k x 1	16 k x 1	64 k x 1	word x bit
Address Access Time	7	13	10	ns
Write Pulse Width	4	7	11	ns
Operating Power	350	500	1300	mW
Memory-Cell Size	97	230	524	μm^2
Die Size	20	30	55.4	mm^2

logics is given in Table 8-3.²⁴ More recently, 256-kbit BiCMOS SRAMs have been announced, with access times of 8-12 ns.^{25,26} As noted earlier, BiCMOS SRAMs can be designed with TTL as well as ECL I/O capabilities.

In early BiCMOS SRAMs, moderate-speed bipolar transistors were integrated into what was essentially a CMOS technology in order to provide faster output buffers and sense amplifiers. In more recent designs, the CMOS devices are being fabricated on a basically high-speed bipolar chip. In an early 16-kbit BiCMOS SRAM, the cells of the memory array were poly-load cells, the peripheral circuits were CMOS, and the I/O buffers and sense amplifiers were bipolar circuits. In a more recent 256-kbit BiCMOS SRAM design, the memory array was implemented with full-CMOS cells, and bipolar sense amplifiers and ECL output buffers were used. This memory had a reported access time of 8 ns and could be operated with battery backup (i.e., since it draws only 1 μA during standby, a battery with minimal power can provide long-term backup).

Figure 8-8 shows the speed-versus-density characteristics of bipolar, CMOS, and BiCMOS SRAMs.³⁴ These curves indicate that bipolar technologies no longer offer the fastest performance at such densities, and that ECL I/O BiCMOS offers higher

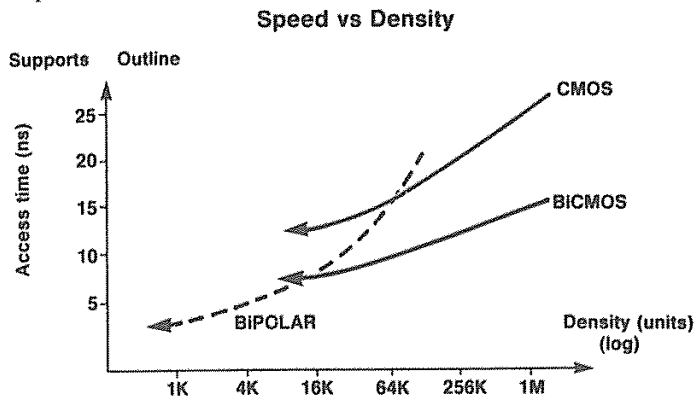


Fig. 8-8 Comparison of speed and density of CMOS, Bipolar and BiCMOS SRAMs.³⁴ Reprinted with permission of Semiconductor International.

speeds than bipolar ECL at middle and higher densities (because of the severe speed, power, and density trade-off in bipolar ECL). In addition, BiCMOS SRAMs are somewhat faster than CMOS SRAMs at equivalent power and density for the same geometry, and TTL I/O BiCMOS SRAMs are a little slower than ECL I/O SRAMs because of their larger output voltage swing. Finally, BiCMOS offers a range of speed and power trade-offs (from very fast, low-density BiCMOS at high power, to slightly slower, mid-density memory circuits at moderate power).

Recently, an 8 ns 1-Mbit BiCMOS SRAM¹¹⁴ and a 3.5 ns 16-kbit ECL BiCMOS SRAM¹¹⁵ have been reported.

8.3 DYNAMIC RANDOM ACCESS MEMORIES (DRAMs)

As noted earlier, *dynamic random access memories* (DRAMs) are so named because their cells can retain information only temporarily (on the order of milliseconds); even with power continuously applied. The cells must therefore be read and refreshed at periodic intervals. While the storage time may at first appear to be very short, it is actually long enough to allow for many memory operations between refresh cycles.

Despite of this apparently complex operating mode, the advantages of cost per bit, device density, and flexibility of use (i.e., both read and write operations are possible) have made DRAMs the most widely used form of semiconductor memory to date.

The earliest DRAMs used three-transistor cells and were fabricated using PMOS technology. Nevertheless, their introduction represented an immediate, dramatic decrease in the minimum semiconductor memory-cell size, since they could replace SRAMs based on a six-transistor cell. As a result, more cells per chip could be implemented. However, DRAM cells consisting of only one transistor and one capacitor were quickly implemented,²⁸ and such cells have been used in DRAMs ever since.

8.3.1 Evolution of DRAM Technology

The earliest MOS combinational logic networks, referred to as *static logic circuits*, operated without any need for periodic clock signals. However, it was recognized that clock signals could be used to advantage in combinatorial and sequential logic circuits. By introducing clock signals at arbitrary circuit nodes, it was possible to achieve faster operation, greater circuit density, and reduced power dissipation. Such logic circuits became known as *dynamic logic circuits*.

Data in these circuits was temporarily stored in dynamic registers (in the form of charge on the gate of an MOS transistor), rather than in static registers (which store data as the state of a flip-flop circuit). Thus, dynamic shift registers could be built with fewer transistors and, consequently, on much smaller areas of silicon than were needed for static shift registers. This allowed a dramatic increase in logic-circuit density.

A question arose, however, as to how long the gate of a MOS transistor could store charge before that charge would be lost through leakage currents. It turned out that at

near room temperature, the charge could be stored for more than 10 milliseconds. If a clock signal were to arrive at intervals significantly shorter than this, a large fraction of the initially stored charge would still remain on the MOS transistor gates. Therefore, if a clock signal were to be applied at least this frequently, dynamic registers could serve as efficient charge-storage nodes.

It was also quickly realized that if the dynamic stored-charge approach was practical in dynamic logic circuitry, it might also work for semiconductor-memory designs. In the case of a dynamic memory, however, a *refresh signal* had to be applied to each charged cell node at sufficiently frequent intervals (typically, every 4-8 ms), to allow the temporarily stored data on each cell to be retained indefinitely.

The concept of the DRAM was patented by Dennard of IBM in 1968, and the first commercial DRAM was introduced by Intel in 1970. The latter was built using a three-transistor cell (Fig. 8-9) in PMOS silicon-gate technology, while Dennard's patent used a one-transistor cell. In the three transistor cell, the charge is stored on the parasitic capacitance of the gate of transistor M_1 . (Although the capacitance in this cell is a parasitic effect, it is drawn explicitly in the circuit schematic of Fig. 8-9 because it is essential for normal memory-cell operation.) The leakage current of the reverse-bias junction of the drain region of transistor M_3 discharges this capacitance over a period of several milliseconds or more. Hence, a periodic signal must arrive at the node in order for the charge stored on the capacitor to be maintained. Since one-transistor DRAM cells quickly replaced the three-transistor cells, the rest of our discussion will be restricted to them.

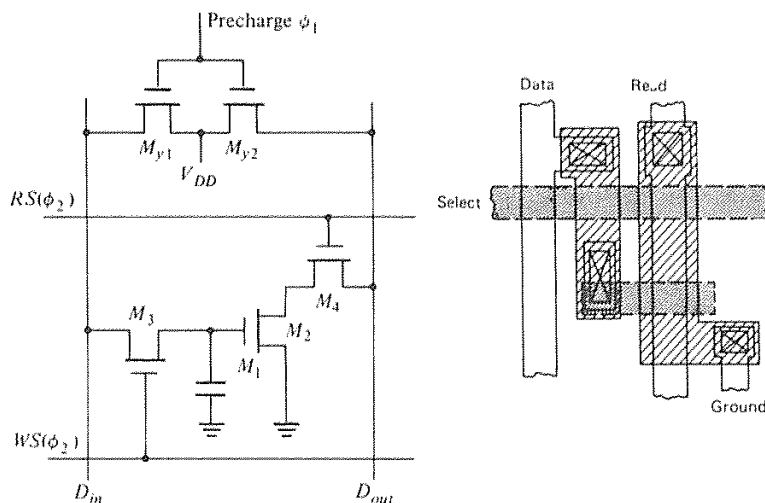


Fig. 8-9 Layout and circuit for a 3-transistor DRAM cell.

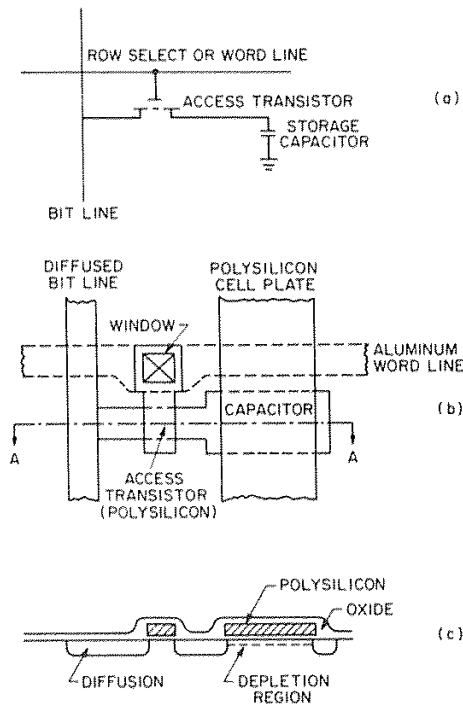


Fig. 8-10 Single-transistor DRAM cell and storage capacitor. (a) Circuit schematic. (b) Cell layout. (c) Cross section through A - A.⁶ After R. Hunt, "Memory Design and Technology," in M.J. Howes and D.V. Morgan, eds., *Large Scale Integration*. Copyright 1981, John Wiley & Sons. Reprinted with permission.

8.3.1.1 One-Transistor DRAM Cell Design. The influence of Dennard's one-transistor cell²⁸ (which actually consists of one transistor and one capacitor) is considered to be comparable to that of the invention of the transistor itself.³ The design of this cell has been rendered in many versions since its invention, and we will begin the description of its evolution (which, by the way, is far from over), with a description of a simple cell that utilizes a polysilicon layer as one plate of the cell capacitor (Fig. 8-10). The first such cell was fabricated with $8\text{-}\mu\text{m}$ features, used $1280\text{ }\mu\text{m}^2$ of silicon area, and was employed in the design of the 4-kbit NMOS DRAM. There have since been many variations of even this simple cell, with single or double layers of polysilicon, different methods of capacitor formation, and different materials used for the word and bit lines. These are reviewed in detail in reference 30.

In the cell shown in Fig. 8-11a, the capacitor stores the charge on the cell (storage capacitor), and the NMOS transistor allows the bit line to access the charge-storage capacitor), and the NMOS transistor allows the bit line to access the charge-storage

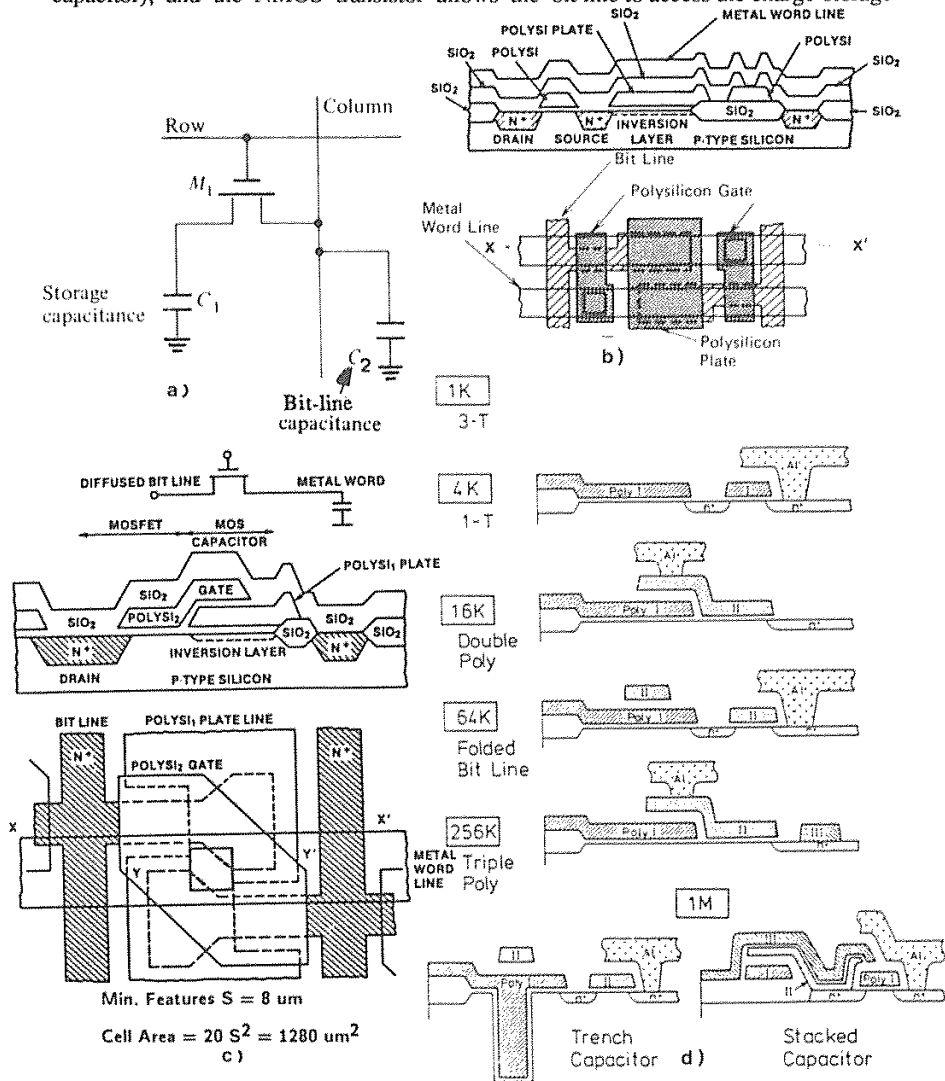


Fig. 8-11 (a) DRAM cell connections to word and bit lines. (b) and (c) Cross-sectional and layout views of (b) the single-poly-word-line/diffused-bit-line 4-kbit DRAM cell, and (c) the double-poly diffused bit line merged DRAM cell.³ (After C. N. Berglund) (d) Structural innovations of DRAMs.⁴¹ (© 1985 IEEE).

region of the capacitor. The storage capacitor consists of a polysilicon plate over a thin oxide film (which is the capacitor dielectric), with the semiconductor region under the oxide serving as the other capacitor plate. An n^+ diffused region in the semiconductor substrate serves as the bit line and an aluminum stripe as the word line. As can be seen in the cross-section of the cell (Fig. 8-11b), the bit-line diffused region makes contact with the n^+ -diffused source region of the access transistor. A contact between the word line and the polysilicon gate of the access transistor is made, as also shown in this figure.

One widely implemented enhancement of this basic cell design is shown in Fig. 8-11c. In this modified cell, the floating drain region of the access transistor is eliminated and a second layer of polysilicon transfers the charge from the bit line to the storage capacitor. This not only allows the cell to be reduced in size, but also increases its storage capacity.³⁵ The disadvantage is that a double-polysilicon process must be used. As has often been the case in the evolution of VLSI, increased packing density and better performance are achieved at the price of somewhat greater process complexity. Figure 8-11d summarizes the structural innovations used as DRAMs have evolved.

8.3.1.2 Operation of the One-Transistor DRAM Cell. To study the operation of the cell in Fig. 8-11c, assume that the substrate is grounded and that 5 V are applied to the polysilicon top plate of the storage capacitor (which we'll refer to as the *plate electrode* of the capacitor). The semiconductor region under the polysilicon plate serves as the other capacitor electrode, and in an NMOS cell this p -type region is normally inverted by the 5-V bias. As a result, a layer of electrons is formed at the surface of the semiconductor, and a depleted region is created below the surface. (The electrode on which the charge is stored will be referred to as the *storage electrode*.)

To write a *one* into the cell, 5 V are applied to the bit line, and a 5-V pulse is simultaneously applied to the word line. The access transistor is turned *ON* by this pulse, since its V_T is about 1 V. The source of the access transistor is biased to 5 V, since it is connected to the bit line. However, the electrostatic potential of the channel beneath both the access-transistor gate and the polysilicon plate of the storage capacitor is less than 5 V, because some of the applied voltage is dropped across the gate oxide. As a result, any electrons present in the inversion layer of the storage capacitor will flow to the lower potential region of the source, causing the storage electrode to become a depletion region that is emptied of any inversion-layer charge. When the word-line pulse returns to 0 V, an *empty potential well* remains under the storage gate. This empty well represents a *binary one*, and it is shown as the deep-depletion space-charge region in Fig. 8-12.

For writing a *zero*, the bit-line voltage is returned to 0 V, and the word line is again pulsed to 5 V. With the access transistor turned *ON*, electrons from the n^+ source region (whose potential has been returned to 0 V) have access to the empty potential well (whose potential is now lower than that of the source region). Hence, the electrons from the source move to fill it, thus restoring the inversion layer beneath the poly plate. When the word-line voltage is returned to zero, the inversion-layer charge present

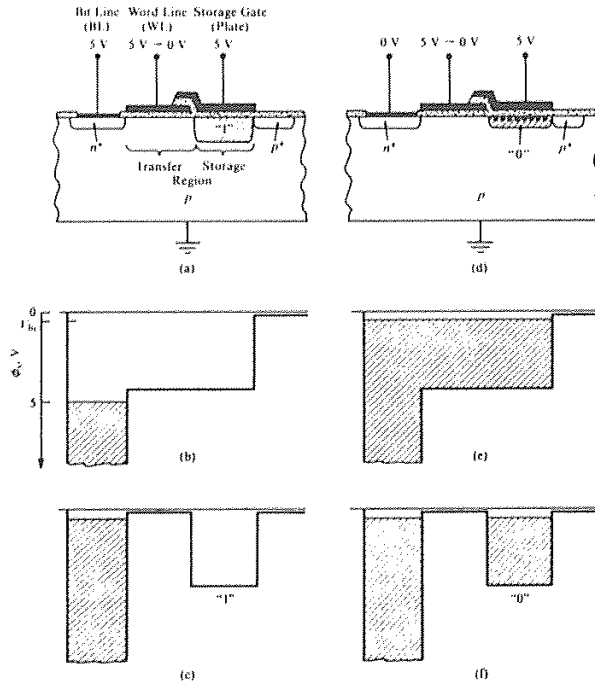


Fig. 8-12 A basic dynamic RAM cell, showing (a) a stored one and (d) a stored zero. The writing of a one is shown in (b) and (c), and the writing of a zero is shown in (e) and (f). From D. K. Schroder, *Advanced MOS Devices*. Copyright 1988, Addison-Wesley. Reprinted with permission.

on the storage capacitor is isolated beneath the storage gate. This condition represents a stored *binary zero*.

Note that when a one is stored, an empty well exists. This is not an equilibrium condition, since electrons that should be present in the inversion layer have been removed. As electrons are thermally generated within and nearby the depletion region surrounding the well, they will move to re-create the inversion layer. A stored one thus gradually becomes a stored zero as electrons refill the empty well. The nature of the planar one-transistor DRAM cells is that ones become zeros, and zeros remain zeros.

To prevent the ones from being lost, each cell must be periodically refreshed by the memory so that the correct data remains stored at each bit location. The time interval between refresh cycles is called the *refresh time*. The total leakage current of the cell must be low enough that the cell does not discharge and lose its memory state between refreshes. A typical guideline has allowed a 20% degradation of the charged state of a cell during the refresh time.¹⁰⁵ For example, if this interval is 8 ms and the charge stored on a cell is 10^6 electrons, the maximum allowed leakage current at maximum

operating chip temperatures and worst-case voltages and timing conditions is ~ 4 pA. This specification can be achieved fairly easily at the average cell level. However, since the leakage currents in a large number of cells exhibit a distribution, the high side of the distribution must be monitored in the course of technology development – and manufacturing – to ensure that the maximum allowable current is not exceeded in any of the memory chip cells.

It is predicted that as memory sizes continue to increase, the maximum allowable leakage current will be reduced. Since alternative capacitor dielectric materials may be used in advanced DRAMs, the leakage currents across such dielectrics will be an important material characteristic. In addition, the access transistor of the cells will have to be turned off very "hard" when the memory cell is not being accessed, since any source-drain leakage can also discharge the storage capacitor.

8.3.1.3 Writing, Reading, and Refreshing DRAM Cells. It's useful to consider the reading and writing of DRAM cells from a circuit perspective as well as from the device physics perspective just discussed. Figure 8-13 provides a simplified schematic of this operation.

Switches S1 and S2 are used to select the appropriate bit and word lines that are connected to the cell in question. Such switch closure is the response to an address

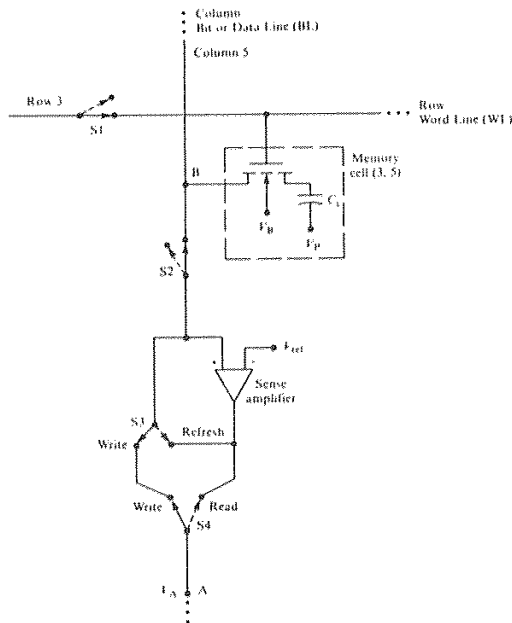


Fig. 8-13 Schematic diagram showing the writing, reading, and refreshing operations of a DRAM. After ref. [117], reprinted by permission of Scientific American, Inc.

signal fed to the bit (column) and word (row) decoders. The closure of S1 causes a "high" voltage to be applied to that particular word (row) line (in turn causing the access transistors controlled by this line to be turned ON).

If it is desired to write data into the cell at that point, an appropriate voltage must also exist on the correct bit line. In addition, switches S3 and S4 must be set to the write positions. This allows the voltage at point A on the bit line, V_A , to be applied to the source of the cell being addressed (point B). As long as Switch S1 is closed and the access transistor is turned on, the capacitor of the cell can be charged to $V_A - V_P$ (where V_P is the *storage-electrode potential*). When the write operation is completed, switches S1 and S2 are opened, and another cell can be written.

If it is desired to read data from a particular cell, the appropriate switches S1 and S2 are once again closed (it is assumed that all of the cells already contain the stored information). Switch S4, however, is set to the read position. The storage capacitor of the cell being read is now connected to one input of the sense amplifier. The sense amplifier is a *comparator circuit*, with its other input being connected to a reference voltage, V_{ref} . Therefore, if the cell-capacitor voltage is larger than V_{ref} , a logic 1 is read; if it is smaller, a logic 0 is read.

A logic 1 corresponds to the condition in which the storage electrode is depleted of its inversion layer charge. If a logic 1 is read immediately after the cell has been written, the signal will be strong. As time passes, thermal electron-hole pair generation will cause refilling of the empty potential well, thereby degrading the amplitude of the logic 1 signal. If too long a time elapses between the writing and reading, the inversion charge will be reestablished, and a logic 0 will be produced when the cell is read. Once a logic 0 is stored on a cell, however, it will continue to be read for as long as power remains applied to the capacitor plate electrode.

This indicates that the logic 1 must be periodically refreshed to allow it to be retained on the cell for indefinite time periods. Because it is not known what logic level is stored on each cell at any instant of time (especially since the cells are randomly read and written), it is mandatory that the entire memory be refreshed at periodic intervals (usually every few milliseconds). Furthermore, since reading a cell changes the charge on its capacitor, the cell must also be refreshed immediately following each read operation.

The refresh procedure is accomplished by switching S3 to the refresh position after the sense amplifier output has been set by the read operation. The output voltage of the sense amplifier will then write the appropriate information back onto the cell capacitor. If it is desired to refresh the entire memory, each cell can be read and refreshed. It is apparent, however, that data cannot be written while reading or refreshing is in progress.

One sense amplifier must be available for each bit line. Note that the sense amplifier is an extremely sensitive comparator (basically of the cross-coupled flip-flop type), and its design is critically important to the success of DRAM manufacture. Although the details of the design task are not our subject here, some aspects of sense-amplifier performance should be mentioned. When a cell is read, the charge stored on the cell capacitor is shared with the 10 to 20 times larger capacitance of the bit line (which, as we saw earlier, is a long conductor line connected to the sources of all of the cells in the

column). After the time interval between refresh pulses has elapsed (e.g., 8 ms) the difference in stored voltage between a 1 and a 0 may be as small as 2 V. As a result, there may only be a 100-200 mV difference between the 1 and 0 signals applied to the sense-amplifier input.

8.3.1.4 DRAM-Cell Charge Storage and Capacitance. A *one* must be clearly distinguishable from a *zero* when the read operation is performed. The zero is represented by the inversion charge present when the potential well is full. This quantity in an MOS capacitor, Q_s , is given by

$$Q_s = (V_{G'} - 2\phi_f) C_{ox} \approx V_{G'} C_{ox} \quad (8-1)$$

where $V_{G'}$ is the voltage applied to the gate, ϕ_f is the difference in potential between the intrinsic Fermi level (E_i) and the Fermi level (E_F), and C_{ox} is the capacitance of the capacitor oxide.* In order to pack a great many cells onto a DRAM chip, the cell size is made as small as possible. This implies that it is also desirable to make the area of the storage capacitor as small as possible. On the other hand, Q_s of the storage capacitor must be large enough to send a sufficiently strong signal to the sense circuitry and to provide sufficient immunity from soft errors (see section 8.3.5). Novel cell designs have been developed in an attempt to satisfy these apparently contradictory requirements (such designs will be discussed later).

EXAMPLE 8-1: Calculate the charge stored in the inversion layer when a *zero* is stored (both in units of *coulombs*, and in terms of the number of electrons present) when 5 V are applied to an MOS capacitor whose dimensions are $4 \times 4 \mu\text{m}$, and which has an SiO_2 dielectric that is 15 nm thick. Also, find its capacitance, C_s .

SOLUTION: $Q_s(0) \approx V_{G'} C_s = (\epsilon \times A \times V_{G'}) / t_{ox}$

$$= (3.9 \times 8.85 \times 10^{-14} \text{ F/cm} \times 16 \times 10^{-8} \text{ cm}^2 \times 5 \text{ V}) / 1.5 \times 10^{-6} \text{ cm}$$

$$= 18.6 \times 10^{-14} \text{ C} = 186 \text{ fC, or since } q = 1.6 \times 10^{-19} \text{ C/electron}$$

$$Q_s = 1.15 \times 10^6 \text{ electrons; and}$$

$$C_s = Q_s / V_A = 37 \text{ fF.}$$

The above shows that the most important parameters involved in increasing the charge stored on the capacitor are the dielectric constant and thickness of the insulator, and the area of the capacitor.

The capacitance of the DRAM cell is also important. As described earlier, when the

* The approximation used in Eq. 8-1 is valid when $V_{G'} > 2\phi_f$, which is the case for $V_{G'} = 5 \text{ V}$ and $\phi_f \approx 0.4 \text{ V}$.

contents of the cell are sensed, the charge stored on it is "dumped" into the bit line connected to the sense amplifier. Because the bit-line capacitance (C_B) is typically 7 to 15 times larger than the cell capacitance, the capacitively-divided voltage difference applied to the sense amplifier (ΔV_{sa}) is substantially smaller than that existing in the cell alone. ΔV_{sa} is given approximately by

$$\Delta V_{sa} = (1/2) V_A (C_{ox}) / (C_{ox} + C_B) \quad (8-2)$$

The minimum detectable voltage difference that the sense amplifiers of 1-Mbit DRAMs can detect (i.e., their *sensitivity*) is in the neighborhood of 150-200 mV. (Note that this sensitivity must be maintained under worst-case operating conditions of voltage, temperature, and noise, as well as worst-case variations of processing conditions.) It is predicted that the sensitivity of sense amplifiers will have to be significantly increased in order for higher-density DRAMs to be fabricated.

8.3.1.5 High-Capacity (HI-C) DRAM Cells. Another technique for increasing the cell's charge-storage capacity without increasing its size was suggested independently by Sodini and Kamins,³⁷ and Tasch, et al.¹⁰⁶ This novel technique involves multiple ion implantations to increase the substrate doping in the local vicinity of the storage node.

A deep implantation of *p*-type impurities (boron) is first performed under the storage-plate area. This increases the substrate doping, which in turn increases the depletion-region capacitance of the storage capacitor. However, this single implant alone does not increase the charge-storage capacity of the cell, since the extra *p*-doping also reduces the difference in the surface potential between an empty and a full potential well.

To restore surface potential to its previous value it (without compromising the increased capacitance), a very shallow layer of *n*-type dopant (arsenic) is implanted under the storage plate area (Fig. 8-14). The implanted *donor* atoms, which are very close to the Si/SiO₂ interface, behave like a fixed positive oxide charge (and the presence of such oxide charge acts to increase the surface potential in NMOS structures). By increasing the depletion-region capacitance without simultaneously increasing the surface potential, the charge-storage capacity of these so-called *high capacity cells* (or *Hi-C cells*) is enhanced by about 50 percent compared to conventional cells.^{35,104} Although, this technique increases the storage capacities of planar capacitor structures in these types of one-transistor cells, the cell size that must be used in order for adequate charge-storage capacity to be obtained eventually becomes too large for this type of cell to be used in advanced DRAMs (i.e., >1 Mbit). New cell structures have thus been developed for larger DRAMs.

8.3.1.6 CMOS DRAMs. With the introduction of the 256-kbit DRAM, the design of DRAM circuits began to change from NMOS to a mixed NMOS/CMOS technology. The cells of the mixed-technology memory array are all built in a common well of a CMOS wafer. The access transistor and storage capacitor of each cell are usually still fabricated using NMOS technology, while the peripheral circuits are

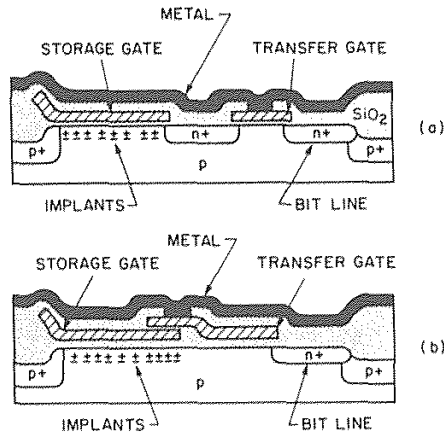


Fig. 8-14 High-capacity (Hi-C) dynamic RAM structure with arsenic (+) and deeper boron (-) implants. (a) One-transistor cell with single-level polysilicon. (b) Double-level polysilicon cell.³⁵ (© 1978 IEEE).

designed and fabricated in CMOS technology. The majority of 1-Mbit DRAM designs were executed in NMOS/CMOS technology, and this trend is expected to continue.

The advantages of NMOS/CMOS over NMOS DRAMs include lower power dissipation (i.e., by a factor of around 3) and smaller soft-error rates (see section 8.3.5). In addition, as power-supply voltages are reduced to allow smaller MOS transistors to be built, NMOS device design must become much more complex to allow the cells to function properly. CMOS on the other hand, can easily operate at such lower voltages. Finally, a circuit technique known as *static column decoding*, which significantly reduces the memory access time, can be successfully implemented in CMOS but not in NMOS (i.e., since to do so in NMOS would require a circuit dissipating excessively large standby power).³⁸

In late 1989 1-Mbit BiCMOS DRAMs were introduced. Their access times of 40 ns placed them between the high-speed 1-Mbit SRAMs (access times of 20 ns), and the slower 1-Mbit CMOS DRAMs (access times of 60-80 ns).

8.3.2 Design and Economic Constraints on Advanced DRAM Cells

As the DRAM cell has been scaled down in size, the minimum amount of stored charge needed to maintain reliable memory operation has remained the same. This constant charge-storage value has had to be maintained within fabrication-cost constraints. From the system point of view, a new generation of DRAM will be embraced if it allows a density increase of about fourfold at the circuit board level, provided that this is accompanied by a cost reduction. To allow such an increase to be realized, the new

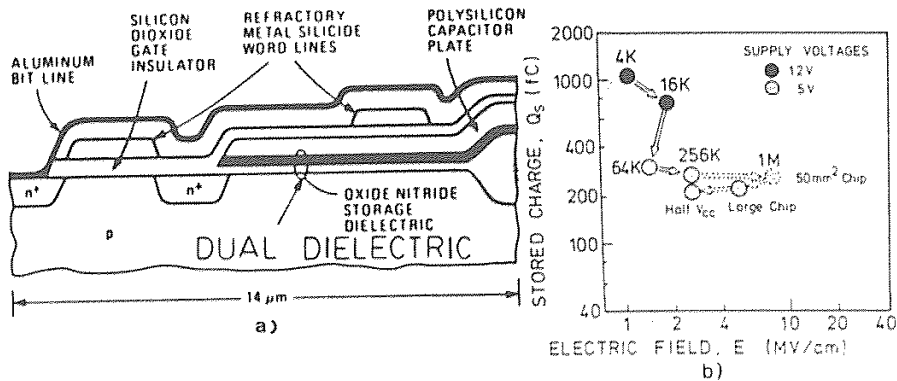
DRAM generation must be able to use the same size package as that used by the previous generation. This explains why the 300-mil package housed DRAMs for five generations. To squeeze enough cells onto a chip to allow this package to be retained implies that the cell size of the 1-Mbit DRAM cannot exceed $20 \mu\text{m}^2$.

For future DRAM generations, however, increased memory size will be achieved by both increasing the chip size and shrinking the cell size. Since the chip size is predicted to increase by a factor of 1.5 from generation to generation, the cell area will therefore need to shrink to 40% of the size of the previous generation (which can be done by shrinking the minimum line width to 70% of that used previously). For 4-Mbit DRAMs, the cell size must therefore be no larger than $9 \mu\text{m}^2$; for 16-Mbit DRAMs, no larger than $4 \mu\text{m}^2$, and so on. To meet the signal-to-noise ratio constraints and the soft-error rate requirements, a minimum of ~ 200 fC of charge ($\sim 10^6$ electrons) will have to be stored. For the fabrication process to be economically feasible, as few steps as possible should beyond those required to fabricate the transistors and interconnects should be added.

Recovery of the equipment and development costs of the 256-kbit, 1-Mbit, and 4-Mbit DRAMs has become the major factor dictating the three-year product-introduction and delivery cycle so that a profit can be generated from each DRAM generation. In a typical three-year cycle, fewer than 1 million chips would be shipped for sampling during the introduction year. In the first production year, 5 million chips would be shipped; production would increase to 50 million chips in the second year, and 500 million in the third. (Note that in the second year the market would be mainframe computers, and in the third year it would be personal computers.)³ The three-year model predicts that 16-Mbit DRAMs will be introduced in 1992, and 64-Mbit DRAMs in 1995.

It was noted in section 8.3.1.3 that the cell's storage capacity could be increased by making the capacitor dielectric thinner, by using an insulator with a larger dielectric constant, or by increasing the area of the capacitor. The first two options are not currently viable, since capacitor dielectrics thinner than those now being used in DRAM cells (10 nm) will suffer leakage due to Fowler-Nordheim tunneling, and dielectrics with significantly larger dielectric constants than of SiO_2 have not yet been accepted for DRAM-cell application (although research work is under way to develop such higher-dielectric-constant materials).³⁹ One recent report described a plasma-CVD process for depositing high-quality Ta_2O_5 films.¹¹⁹ (Ta_2O_5 exhibits a much higher dielectric constant than SiO_2 [22 vs. 3.9], but normally also suffers from a much higher leakage current.) However, by reacting TaCl_5 with N_2O under optimized plasma-CVD conditions, Ta_2O_5 films with a thickness that yielded capacitances equivalent to those of a 30-Å SiO_2 film, demonstrated very low leakage currents for up to 10-year operation at 3.3 V.

It should also be noted that since the 256-kbit DRAM generation bilayer films (consisting of both silicon nitride and SiO_2), have been used as the capacitor dielectric to increase cell capacitance (Fig. 8-15a). The higher dielectric constant of Si_3N_4 (twice as large as that of SiO_2) was responsible for this increase.



Factors	256K	1M	4M
Type	NMOS	CMOS	CMOS- Trench or Stack
No. of Masks	9~10	17~19	20~25
Total Process Steps	200	350	450
Test Time (Sec.)	60	120	240
Chip Size (Ratio)	1	2	3
Logical No. of Chips/ Wafer (Ratio)	3	2	1
Good Dies/Lot (Ratio)	6	2	1
Clean Room Class	100	10	1
Cost Ratio (on a Mature Production)	1	4	10

c)

Fig. 8-15 Cross-sectional view of a planar-capacitor DRAM cell with a two-layer capacitor dielectric.³ (© 1988 IEEE). (b) Trends in storage charge and electric field across capacitor insulator.⁴¹ (© 1985 IEEE). (c) Comparison of DRAM production factors.⁴⁴

One technique that did allow thinner dielectric films to be used was the *half- V_{CC}* approach.⁴⁰ That is, the *plate electrode* is biased to $V_{CC}/2$ (i.e., to 2.5 V when $V_{CC} = 5$ V), and the *storage electrode* is allowed to swing between 0 V and 5 V. As a result, the same quantity of charge can be stored on the capacitor, but the value of the electric field acting on the dielectric is only half the value that exists when the voltage between the two plates equals V_{CC} . This technique has been implemented in the fabrication of 1-Mbit DRAMs (Fig. 8-15b).

The third option (increasing the capacitor area) can be effective if the area is increased by forming the storage capacitor in a trench etched in the substrate or by using a stacked capacitor structure. Both approaches have been implemented, and many variations of such three-dimensional capacitors have been reported.

The planar capacitor structure used in the one-transistor DRAM cell described in section 8.3.1.1 was predicted to be usable up to the 256-kbit DRAM generation. In this generation, the capacitor consumes 30 to 40% of the cell area. It was generally agreed that beyond this, a three-dimensional capacitor structure would be needed in order for sufficient charge storage to be obtained. It turned out, however, that virtually all of the DRAM manufacturers elected to squeeze everything they could from the planar capacitor, and continued to use it to manufacture 1-Mbit DRAMs. This decision was due largely to the difficulty in achieving a reliable capacitor dielectric in a trench cell at the time 1-Mbit DRAMs were introduced. The use of both larger chip sizes and the half- V_{CC} plate-electrode voltage technique permitted the planar capacitor to perform adequately for 1-Mbit DRAMs. Reference 42 presents the details of a 1-Mbit DRAM technology using a 38-fF planar-capacitor structure in which the cell size is $37 \mu\text{m}^2$.

As DRAM size increases, process complexity is expected to increase markedly as well. For example, a 1-Mbit DRAM is reported to require ~18 masks and 350 processing steps, all of which could be successfully carried out in a Class 10 cleanroom (Fig. 8-15c). In comparison, the 4-Mbit DRAM is expected to need 20-25 masks and in excess of 450 processing steps, and will thus require a Class 1 cleanroom processing facility.^{43,44} A detailed report on the technology issues that will need to be addressed in the design and fabrication of 64- and 256-Mbit DRAMs has recently been published.¹⁰⁵

In 1989 1-Mbit CMOS DRAMs with access times ranging from 6-100 ns were being commercially offered. (The fabrication of a high speed 22-ns CMOS DRAM was announced in late 1989, but it was not being offered for sale.)¹¹⁰ At that time, 4-Mbit DRAMs with access times of 80-120 ns were also being offered, and 16-Mbit CMOS DRAMs with access times as small as 45 ns were being reported.¹¹¹ Finally, 1-Mbit BiCMOS DRAMs with access times of 30 ns were being introduced.¹¹²

8.3.3 Trench-Capacitor DRAM Cells

8.3.3.1 Trench Capacitor Processing for DRAMs. Trench-capacitor structures have been developed as a way to achieve DRAM cells with larger capacitance values without increasing the area these cells occupy on the chip surface. (For example, the silicon-area reduction of a trench capacitor compared to a planar capacitor for the same specific capacitance is a factor of 18 or more. Specifically, a 4.0- μm -deep trench capacitor with surface dimensions of $0.87 \times 2.4 \mu\text{m}$ will occupy less than $3 \mu\text{m}^2$ of chip area but will have a capacitance of 40 fF.)⁶⁶ Many of the processing details involved in trench-capacitor fabrication are the same as those described in chapter 2, section 2.6.3, which deals with the process technology of trench-isolation structures. In this section we discuss those issues that are unique to the fabrication of trench capacitors used in DRAM cells.

There are several differences between the trench structures used for isolation and those used as DRAM capacitors. In the former, the dielectric film on the trench walls can be relatively thick, and the trench can be refilled with polysilicon or CVD SiO_2 . In the latter, the insulator formed on the trench walls serves as the capacitor dielectric, and it

must therefore be as thin as possible. Since the material that refills the trench serves as one plate of the capacitor, it must consist of highly doped polysilicon. Furthermore, in order for increased capacitance to be obtained through increases in trench depth (while all other parameters remain constant), the trench walls must be highly vertical. To allow for reliable refilling of the trenches, however, some trench sidewall slope must be allowed, and a compromise process that produces a nominal sidewall slope of 87° has been suggested.⁴⁶ Finally, to obtain such structures as Hi-C capacitors, the trench walls may need to be selectively doped.

Several techniques have been developed for achieving a dielectric capacitor film that is thin enough to provide both high capacitance and high reliability (that is, the dielectric must be able to provide the same equivalent breakdown voltage as the planar capacitor used in previous DRAM generations). First, composite dielectric films (e.g., thermally grown oxide and CVD nitride) are frequently used.^{47,62,68} Since the nitride has a higher dielectric constant than SiO_2 , a thicker composite film will yield the same capacitance as a thinner single SiO_2 layer. This thicker film prevents capacitor leakage due to dielectric breakdown or Fowler-Nordheim tunneling.

The growth of the thermal oxide film is also a key step. Unless preventative measures are taken, a thinner oxide will grow in the bottom corners (concave) and top corners (convex) of the trench. A higher electric field will exist across these regions, causing trench capacitors to exhibit higher leakage currents than planar capacitors.

This problem is avoided for the bottom corners by ensuring that the etch process produces a trench with rounded bottom corners (see chap. 2). In addition, an oxidation step for edge rounding and stripping is performed prior to the growing of the actual capacitor SiO_2 film. One report indicates that a 50-nm SiO_2 film is grown in this process and is then stripped in dilute HF (Fig. 8-16a).⁴⁸ In addition to smoothing out any sharp bottom corners, this step also removes any plasma damage from the trench walls.

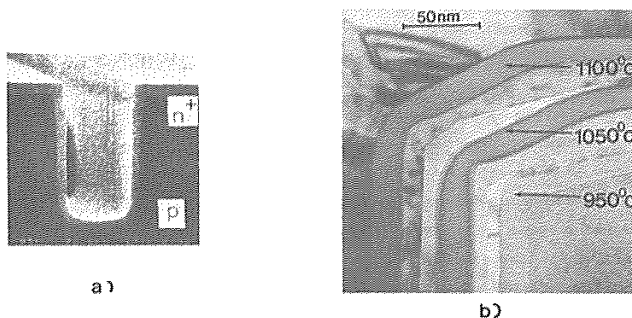


Fig. 8-16 (a) Rounding-off oxidation can produce trenches with smooth bottom corners.⁴⁸ (© 1985 IEEE). (b) Rounding-off oxidation can also reduce the severity of the sharp upper corner of the trench.⁵⁰ This paper was originally presented at the Spring 1989 Meeting of The Electrochemical Society, Inc. held in Los Angeles, CA.

The electric field is intensified at the top corners of the trench because they are normally quite sharp after etch, and this magnifies the effect of any oxide thinning that may occur. (The extent of the electric-field intensification is modeled in reference 49.) The edge-rounding oxidation step mentioned above also increases the curvature radius of the top corner of the trenched Si surface (Fig. 8-16b), thus helping to produce a trench capacitor with low leakage currents under high electric fields. However, because it is necessary to use a process that allows viscoelastic flow during oxide growth (to relieve the stresses that inhibit oxide growth at the convex corners), a higher temperature oxidation process (e.g. 1100°C) is usually involved.^{49,50} The use of rapid thermal processing (RTP) to grow the trench oxide has also been reported.⁵¹ Good leakage-current behavior is exhibited when RTP cycles of 1150°C for 25 sec in O₂ were used to grow the trench oxide.

The polysilicon that fills the trench must also be highly doped to prevent depletion effects. In situ doping of the poly is thus necessary. The conventional process for in situ doping of polysilicon employs gaseous phosphine as the dopant source. Unfortunately, this reduces the polysilicon deposition rate by a factor of about 25 (see Vol. 1, chap. 6). Specially designed LPCVD furnaces with caged boats are needed to improve the process.¹¹⁷ However, these furnaces have particulate problems and cannot be automated, making them incompatible with a high-volume fabrication environment. A recent report described the use of t-butylphosphine (TBP) as an alternative doping source.¹¹⁸ It can be used in standard automated 100-wafer LPCVD furnaces to produce in-situ doped polysilicon films. A higher deposition rate can be achieved (~20 Å/min), with adequate thickness uniformity. This material is also less toxic than phosphine.

8.3.3.2 First-Generation Trench-Capacitor-Based DRAM Cells.

Trench structures for storage capacitor application in DRAMs were first reported in 1982-83 (Fig. 8-17a).⁵² The processing technology that made these structures possible was anisotropic etching of Si by RIE. Earlier V-groove structures etched in Si by means of wet etching resulted in crystallographically produced sharp edges, which in turn degraded the gate-oxide integrity to the point where devices could not be reliably manufactured. One of the first tests that had to be met by RIE-etched trench capacitors was that of exhibiting breakdown characteristics equal to those of planar-type capacitors. As described in the previous section (and summarized in Fig. 8-17b), several reports showed that this could be achieved through the implementation of trench etching control measures, the use of edge-rounding procedures, or the use of combination films for the trench dielectric (e.g., thermal SiO₂ and CVD nitride).

In the first generation of trench-capacitor-based cells *the plate electrode of the storage capacitor is inside the trench, and the storage electrode is in the substrate*. The access transistor is a planar MOS transistor fabricated beside the trench capacitor, and the trenches are 3-4 μm deep. The cell size of the basic cells of this generation requires about 20 μm² of surface area, making the cells suitable for 1-Mbit DRAM designs. It was thought that with appropriate design-rule shrinkage, these cells would be appropriate for early 4-Mbit DRAM designs.⁵³

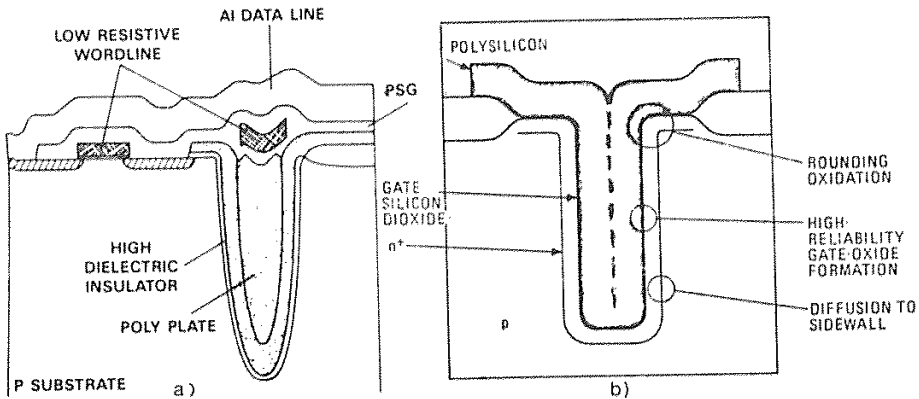


Fig. 8-17 (a) Basic DRAM trench capacitor structure.⁵² (© 1982 IEEE). (b) Processing techniques used to insure fabrication of high-quality trench structures.

In one variation of this cell type, the plate electrode is grounded, and the substrate is biased between 0 and 5 V (which improves device isolation between adjacent cells).⁵⁴ The walls of the storage electrode (i.e., those in the *p*-type substrate) are doped *n*-type, creating a Hi-C type cell.

These first generation cells exhibit some disadvantages for smaller-sized DRAM cells. Since the charge is stored in a potential well in the substrate, if the cells are too close together, high leakage currents arise between adjacent cells (due to punchthrough or surface conduction). This problem can be alleviated through increased doping of the region between the cells or through the use of deeper, narrower trenches, but at the cost of creating other problems. First, the required doping in the substrate will lead to avalanche breakdown of the reverse-biased junction of the access transistors at spacings $\leq 0.8 \mu\text{m}$. Second, deeper, narrower trenches are significantly more difficult to fabricate reliably and for practical trench dimensions the spacing limit is nearly reached for the cell sizes needed in 4-Mbit DRAMs. Further, since the storage node is in the substrate, there is no immunity to charge collection from alpha particles. Consequently, this type of trench capacitor is as vulnerable to alpha-particle-induced soft errors as cells made with planar storage capacitors. Several design modifications have been developed to increase capacitance without either making the trenches deeper or increasing cell size.

In the first modification, the plate electrode is folded around the sides of the storage electrode, creating a structure called the *folded-capacitor cell*, FCC (Figs. 8-18a and b).⁵⁵ A shallow trench is etched around most of the perimeter of the storage electrode. The plate electrode is deposited over this trench, much as a tablecloth is laid over a table top.⁵⁶ When both the sides and the planar area (tabletop) are covered, a capacitor with a larger area is obtained, and the capacitance is thereby increased (Fig. 8-18c).

Interestingly, the capacitor of this cell apparently utilizes *both* the planar- and trench-capacitor concepts. In addition, the cell's storage plate edges are electrically isolated

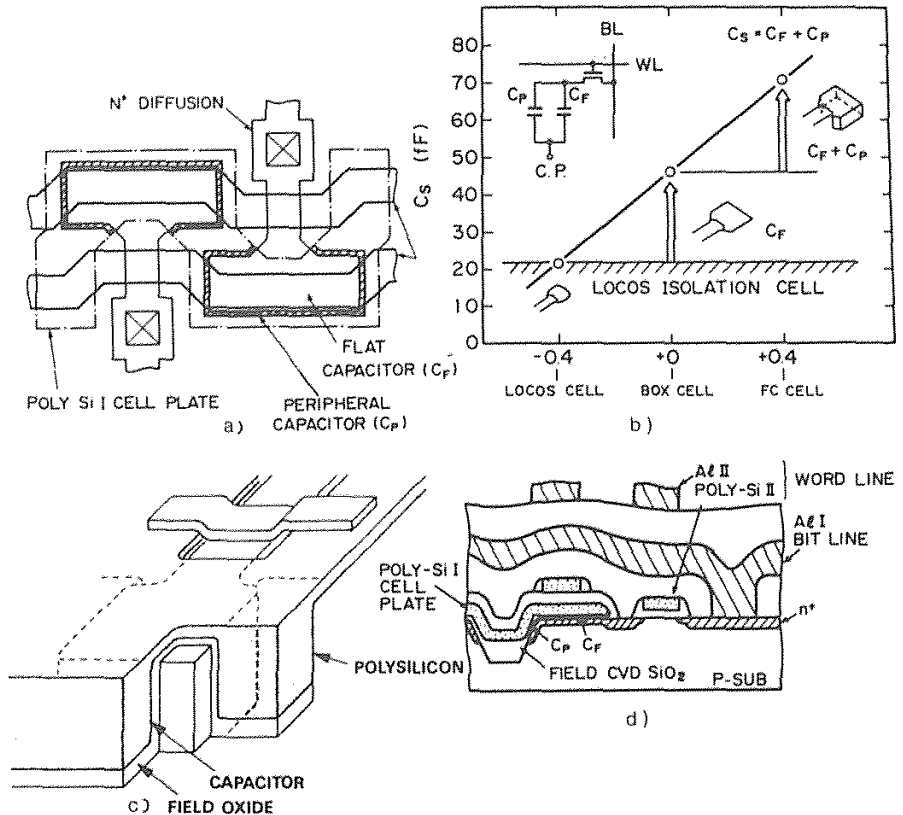


Fig. 8-18 (a) and (b) Top and perspective views of the folded capacitor cell (FCC).⁵⁵ (© 1984 IEEE). (c) Increase in capacitance by FCC. (d) Cross section of FCC showing CVD SiO₂ BOX-isolation structure.⁵⁶ (© 1986 IEEE).

from those of adjacent cells by means of a *BOX-type* isolation structure, rather than a *LOCOS* isolation structure (Fig. 8-18d). This increases the memory-array packing density (and in effect decreases the cell size), while also increasing the capacitance. An FCC cell size of 32 μm^2 with a 70-fF capacitor was reportedly used to fabricate 1-Mbit DRAMS. This cell appears to be scalable to 4-Mbit and 16-Mbit DRAM requirements.

In a second novel approach, the walls of the storage electrode were made to follow the outside edges of the cell perimeter, and the access transistor was placed inside (*Isolation Vertical Capacitor cell*, or IVEC, Fig. 8-19a).⁵⁷ A third invention folded the plate electrode around the storage electrode but used selective doping of certain trench walls to achieve isolation (i.e., the substrate trench walls that act as isolation structures were

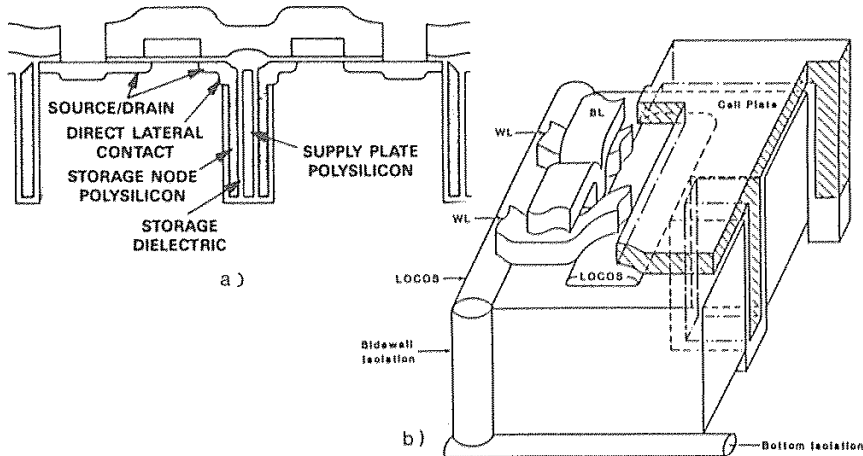


Fig. 8-19 (a) Isolation-merged VERTICAL Capacitor (IVEC) cell.⁵⁷ (© 1984 IEEE). (b) Perspective view of the FASIC cell.⁵⁸ (© 1987 IEEE).

selectively boron doped by means of oblique ion implantation) and storage (i.e., those walls used as the storage plate surfaces were arsenic doped by means of oblique ion implantation, creating a Hi-C storage capacitor). This latter cell was named the *folded bit-line adaptive sidewall isolated capacitor* (FASIC) cell (Fig. 8-19b).⁵⁸ FASIC cells can be made as small $10 \mu\text{m}^2$ and with capacitances as large as 50 fF, making them suitable for use in 4-Mbit DRAMs. They require trenches of only $2 \mu\text{m}$ in depth.

8.3.3.3 Trench-Capacitor Structures with the Storage Electrode Inside the Trench (Inverted Trench Cell). One set of trench-capacitor designs sought to reduce punchthrough and soft-error problems by placing the plate electrode on the *outside* of the trench, and the storage electrode *inside* (Fig. 8-20a). Since the charge is stored inside the trench (which is therefore completely oxide isolated except in the region of lateral contact to the access transistor), it can leak only through the capacitor oxide or the lateral diffused contact.

Four examples of early approaches using such cell designs are the *buried-storage-electrode cell* (BSE) (Fig. 8-20b),⁶⁰ the *substrate-plate-trench cell*, (SPT) (Fig. 8-20c),⁶¹ and the *stacked-transistor-capacitor cell*, (STT) (Fig. 8-20d).⁶² In the first two, the plate electrode is heavily *p*-doped and is connected to the power supply, while the inside storage plate is heavily *n*-doped. Since the substrate is maintained at essentially an equipotential, the punchthrough problem exists only around the region through which the charge is introduced into the trench. Note that for heavily-doped storage electrodes (e.g., $>2 \times 10^{19} \text{ cm}^{-3}$), inversion will not occur at 5 V or less. Instead, the bias applied to the capacitor causes both plates of the capacitor to deplete; together with the oxide capacitor, these two depletion regions make this type of trench

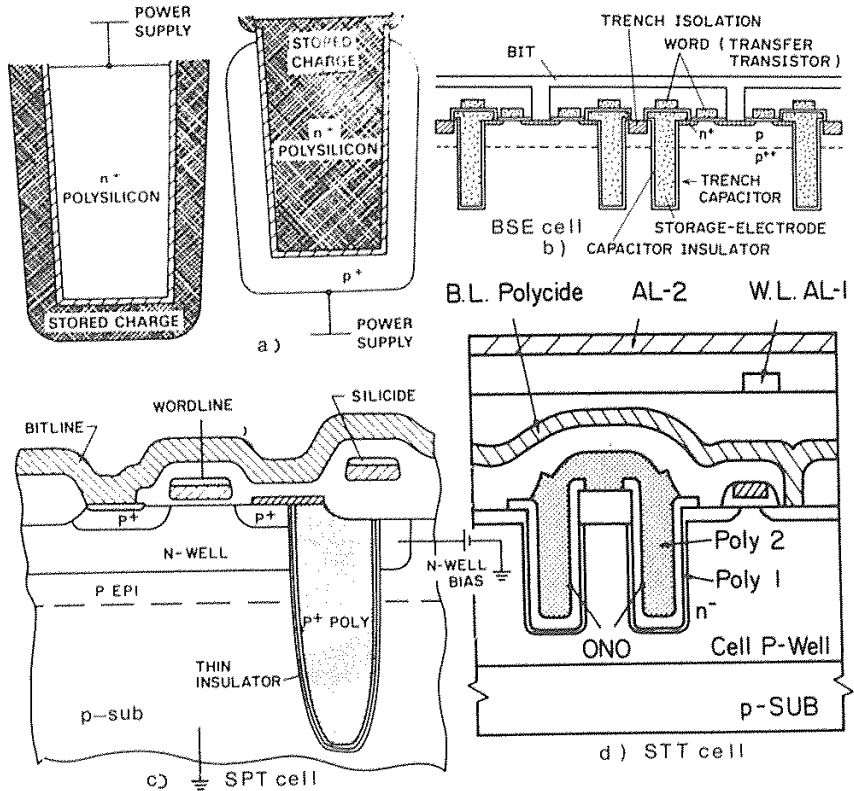


Fig. 8-20 (a) Mechanism of charge storage on outer and inner plates of the DRAM trench storage capacitor. (b) Cross section and process sequence of BSE cell.⁶⁰ (© 1985 IEEE). (c) Cross section of SPT cell.⁶¹ (© 1985 IEEE). (d) Cross section of STT cell.⁶² (© 1987 IEEE).

capacitor equivalent to three capacitor elements in series. Since the depletion regions grow with increasing voltage, the total trench capacitance decreases monotonically. The heavy doping of the plates therefore helps to maximize the cell capacitance. Finally, in such cells a 0 logic level is stored as 0 V and a 1 level as 5 V.

The problem with this type of cell is that the *gated-diode structure* shown in Fig. 8-21a can cause a significant leakage current to flow into the storage node, adversely affecting the cell's retention time. (The physics of the gated diode structure is treated in detail in reference 63.) An alternative cell (the IBM SPT cell) overcomes this problem by using PMOS access transistors and *p*-type doped inner-storage electrodes, and then creating the SPT cells in an *n*-well on a *p*-substrate (Fig. 8-21b).⁶⁴ As a result, the storage electrode gates the *n*-well-to-substrate junction, and the leakage current (as well

as the *band-to-band tunneling-induced leakage current* generated in the bulk silicon⁶⁵ is collected at the *n*-well contact instead of at the storage electrode. If such a cell is not built in a well (e.g., the BSE cell), the storage electrode will gate the junction formed by the storage electrode and the substrate, and the resulting leakage current will be collected by the storage electrode.

In the most advanced type of cell that does not use the substrate as the storage electrode, *both* the plate and storage electrodes are fabricated inside the trench opening, allowing both electrodes to be completely oxide-isolated. Lightly doped epitaxial layers on heavily doped substrates are not needed, and the cells will be free from punchthrough at arbitrarily small cell spacings. In addition, the soft-error rate will be reduced further than it is in the other inverted trench cells. However, these improvements are achieved through a substantial increase in process complexity.

Several such cells have been reported, including the *dielectrically encapsulated trench* (DIET) capacitor (Fig. 8-22a),⁶⁶ the *half- V_{CC} sheath-plate capacitor* (HPSC) (Fig. 8-22b),⁶⁷ and the *double-stacked capacitor* (DSP) (Fig. 8-22c).⁶⁸ The last has two polysilicon plates, one biased to V_{BB} and the other to $V_{CC}/2$. The capacitors formed by the lower poly plate and substrate (separated by the outer dielectric layer), and by the two poly layers (separated by the interpoly dielectric) act in parallel, almost doubling the cell's storage capacitance. A DSP cell of $6\ \mu\text{m}^2$ in size with trench depths of $4\ \mu\text{m}$ is reported to exhibit a capacitance of 50 fF.

8.3.3.4 Trench-Capacitor Cells with the Access Transistor Stacked above the Trench Capacitor. The access transistor occupies a significant fraction of the cell area in trench-transistor cell designs. When this transistor is a planar transistor and is placed alongside the trench capacitor, surface area must be devoted to both structures. Attempts to use short-channel lengths for the access

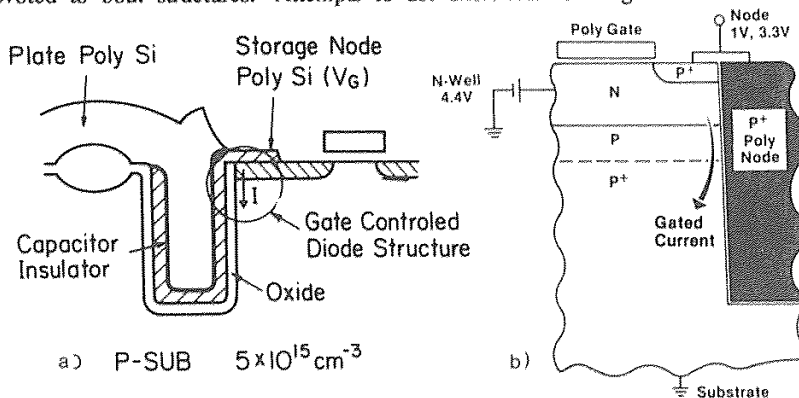


Fig. 8-21 (a) Cell structure with gate controlled diode. (b) Schematic representation of SPT cell bias conditions — *p*-substrate-to-*n*-well junction is gated by the polysilicon node.⁶⁴ (© 1987 IEEE).

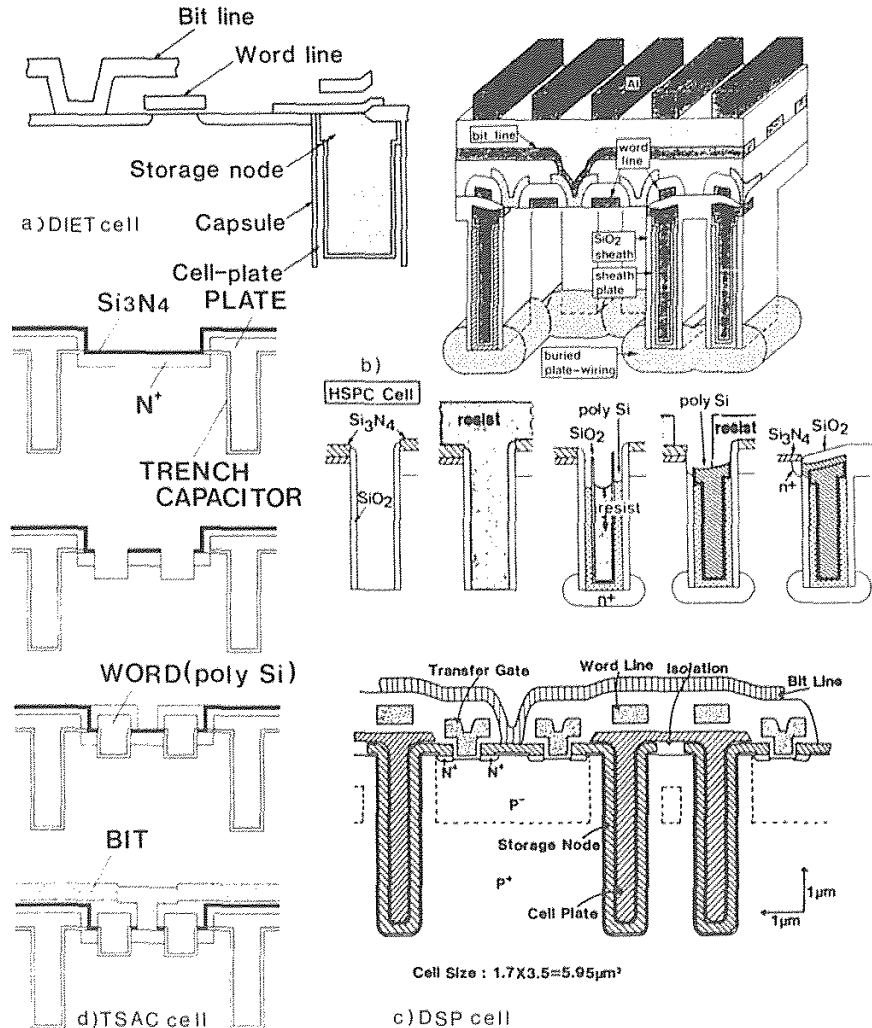


Fig. 8-22 (a) Cross section of DIET cell.⁶⁶ (© 1986 IEEE). (b) Perspective view and process sequence of the HPSCC cell.⁶⁷ (© 1987 IEEE). (c) Cross section of the DSP cell.⁶⁸ (© 1987 IEEE). (d) Fabrication process of the TSAC cell.⁶⁹ (© 1986 IEEE).

transistor have run up against the effects of drain-induced barrier lowering (see section 5.5.2).

One technique for overcoming this problem extends the gate length of the access transistor by forming a trench in the transistor channel (Fig. 8-22d). This reduces the

area of the planar access transistor without decreasing its channel length.⁶⁹ Using this technique with a self-aligned contact structure, a cell size of $9\ \mu\text{m}^2$ was realized; such a cell is making it suitable for a 4-Mbit DRAM.*

A more efficient use of space would be to stack the transistor above the trench capacitor (and, if possible, to form a vertical-access transistor). Two examples of such cells are the *trench-transistor cell* (Fig. 8-23) and the *self-aligned epitaxy over trench cell* (SEOT) (Fig. 8-24).⁷¹

In the trench-transistor cell, the vertical-access (or trench) transistor is built in the top $2\ \mu\text{m}$ of the trench. Its source is connected to the n^+ polysilicon storage electrode of the capacitor by a lateral contact, made by means of an oxide undercut etch and polysilicon refill. The drain, gate, and source of the trench transistor are formed by a diffused buried n^+ bit line, an n^+ polysilicon word line, and a lateral contact, respectively. The gate-oxide thickness is $\sim 25\ \text{nm}$ and the channel length is $1.5\ \mu\text{m}$. The transistor width is determined by the perimeter of the trench. The electrical behaviors of this trench transistor have been modeled, and the results are presented in reference 72. This cell has reportedly been used to build 4-Mbit DRAMs.

A *surrounding gate transistor* (SGT) cell that extends the trench-transistor cell approach has recently been reported (Fig. 7-23d).¹²³ This cell can be made smaller than the trench-transistor cell because it uses trench isolation for the bit-line isolation, rather than the LOCOS isolation used in the latter cell. The transistor and capacitor of this cell surround a silicon pillar, allowing the cell size to be shrunk to $1.2\ \mu\text{m}^2$ while still providing 30 fF storage capacitance. The SGT cell is being studied as a candidate for 64/256-Mbit DRAMs.

In the SEOT cell the storage electrode is first completely isolated from the substrate (Fig. 8-25a), and selective epitaxy is then grown. With the exposed Si area surrounding the trench acting as a seed, a single-crystal-silicon layer grows over the top of the trench (Fig. 8-25b). When the epitaxy growth is stopped before the lateral epitaxial film has grown completely over the trench, a self-aligned window is formed on top of the trench. The capping oxide on the top of trench surface is then etched, and a second epitaxial film is grown. A pyramidal window of polysilicon is formed on top of the exposed polysilicon in the trench; the material surrounding this pyramid is single-crystal silicon formed by means of lateral epitaxy. A planar surface is achieved after a specific minimum of epitaxial growth, and the isolation structure and MOS transistors are then fabricated. An $8\text{-}\mu\text{m}^2$ cell size has been achieved using $0.85\text{-}\mu\text{m}$ design rules, making this cell suitable for 4-Mbit DRAMs. With some process improvements and design modifications, the cell appears to be scalable to 64-Mbit DRAM dimensions.

8.3.4 Stacked Capacitor DRAM Cells

Another approach that allows the cell to shrink in size without a loss of its storage capacity is that of stacking the storage capacitor on top of the access transistor, as

* A report that studied the design methodology and size limitations of submicron access transistors for DRAM applications is published in reference 70.

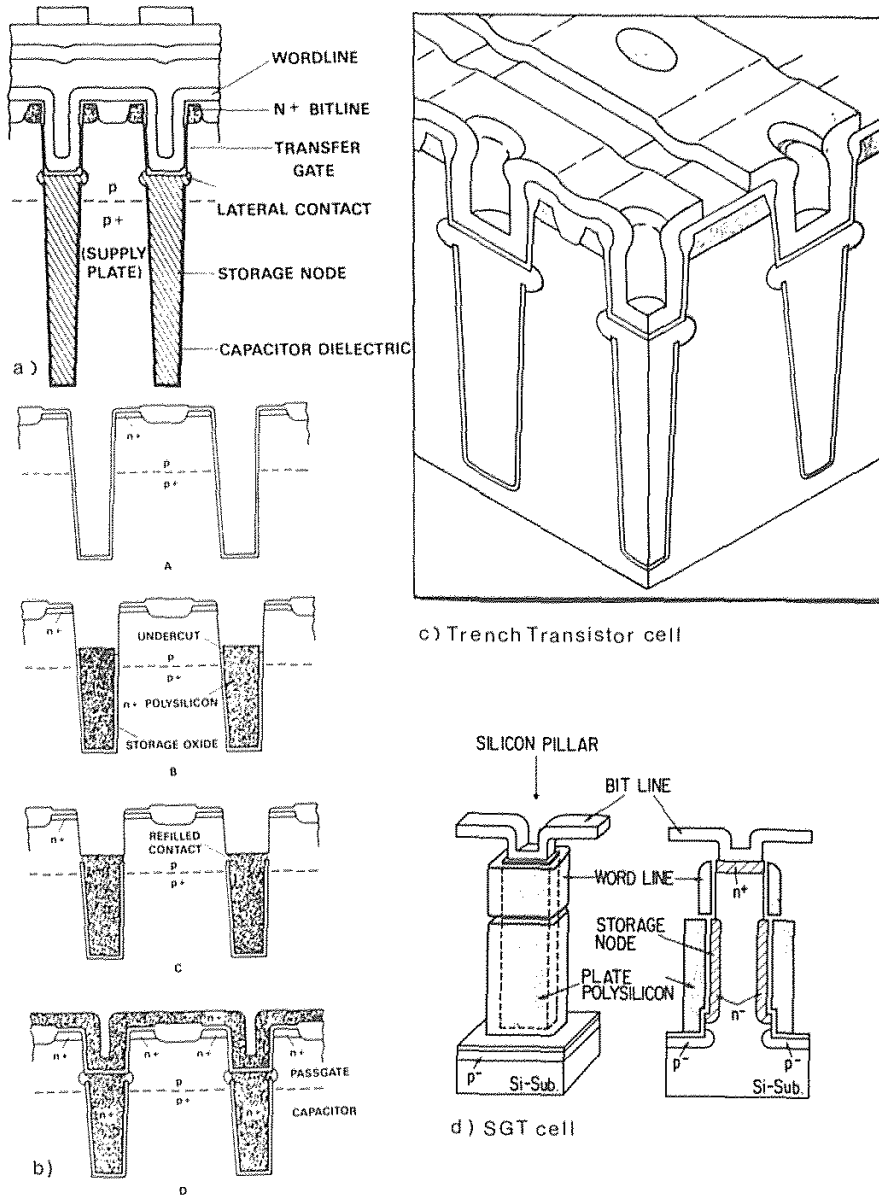


Fig. 8-23 (a) Cross section; (b) perspective view; and (c) fabrication sequence of the trench transistor cell;⁵⁹ (© 1985 IEEE). (d) Schematic view of SGT cell.¹²³ (© 1989 IEEE).

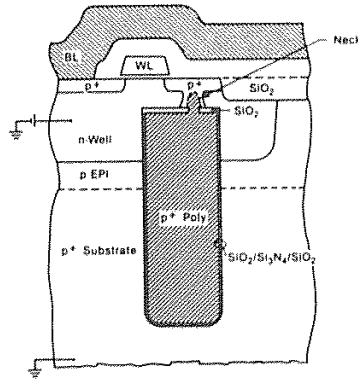


Fig. 8-24 Cross section of the SEOT cell.⁷¹ (© 1988 IEEE).

shown in Fig. 8-26.⁷³ The lower electrode of the stacked capacitor is in contact with the drain of the access transistor, and the bit line runs over the top of the stacked capacitor. Although some stacked capacitor (STC) type cells have been used to fabricate

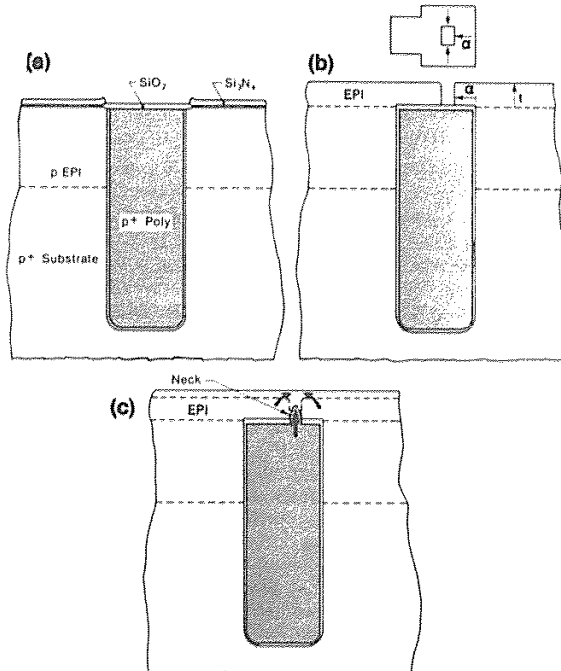


Fig. 8-25 Key processing steps of the SEOT technology.⁷¹ (© 1988 IEEE).

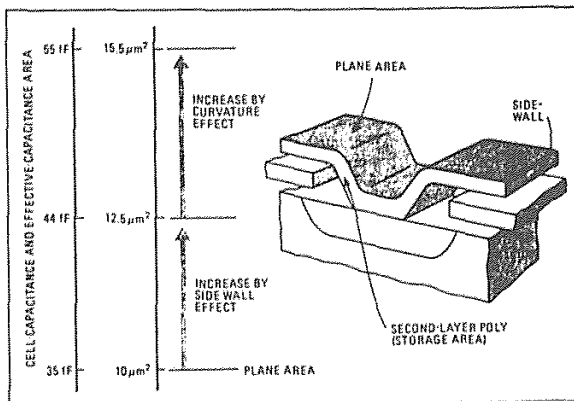


Fig. 8-26 Stacked capacitor (STC) cell structure.

256-kbit DRAMs, the minimum cell size required by conventional stacked-capacitor cells for adequate charge-storage capacity is too large for these cells to be used in larger DRAMs.

For STC cells to be made feasible for 1-Mbit DRAMs and beyond, an insulator with a larger dielectric constant than that of SiO_2 must be used, or novel cell structures must be developed. Although research is continuing into the use of such higher dielectric constant materials as tantalum pentoxide, currently acceptable insulators do not make *conventional* STC cells viable for 1-Mbit and larger DRAMs. However, several novel STC cell designs have been reported.

In the first of these, the contact hole used to connect the lower capacitor electrode to the drain of the access transistor is not filled up with polysilicon, as is the case in conventional STC cells (Figs. 8-27a and b).⁷⁴ Such filling up of the contact hole reduces the effective area of the capacitor, especially for holes with small dimensions. Instead, the contact hole is opened *after* the lower capacitor-electrode polysilicon film is deposited, and only a thin second polysilicon film is subsequently deposited into the hole (Fig. 8-27c). Good contact between the second film and the substrate is established by means of an ion-beam mixing implantation step following the thin-poly deposition (see section 3.4.2.5). This process produces a cell capacitance that is ~ 1.3 times as large as that obtained with a conventional STC. (Figure 8-27d shows an SEM photograph of the new cell.) A cell capacitance of ~ 35 fF with a cell size of $8.8 \mu\text{m}^2$ is achieved with this design.

As shown in Fig. 8-27d, by trenching into the Si substrate it is possible to produce even more capacitance for a cell size of the same area (or a comparable capacitance for a smaller cell size). The trenched STC cell of Fig. 8-27d is predicted to be able to provide 30 fF of capacitance in a cell area of $1.3 \mu\text{m}^2$, which would make it suitable for 64-Mbit DRAMs. A trench depth of $\sim 1 \mu\text{m}$ is needed to achieve this capacitance value.

In the second novel STC cell, the bit lines are formed before the stacked capacitor is fabricated. In addition, the capacitor is laid out on a diagonal with respect to the bit and

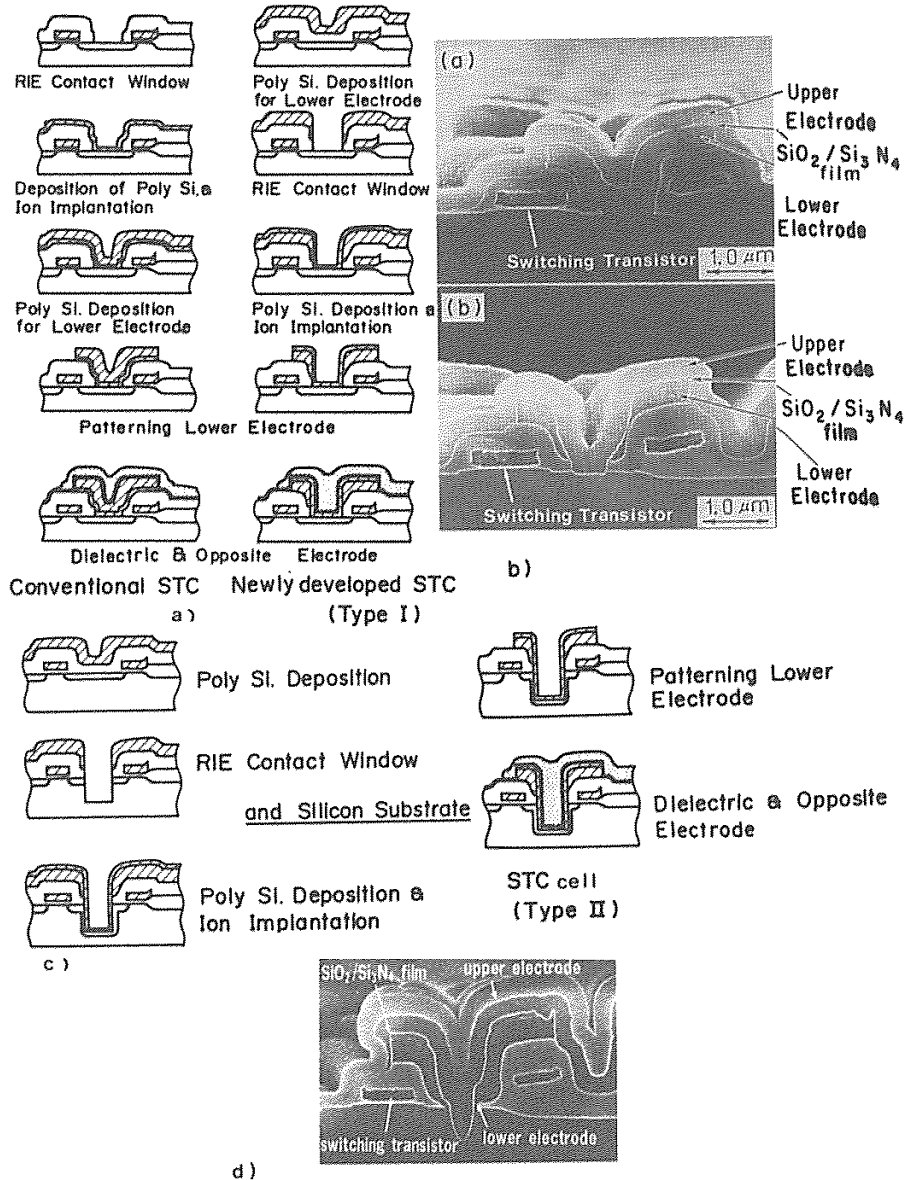


Fig. 8-27 (a) Process sequence of conventional and newly-developed type-I STC cells. (b) Cross-sectional SEM pictures of these two cells. (c) Process sequence of an advanced STC cell (type II). (d) SEM picture of this cell.⁷⁴ (© 1988 IEEE).

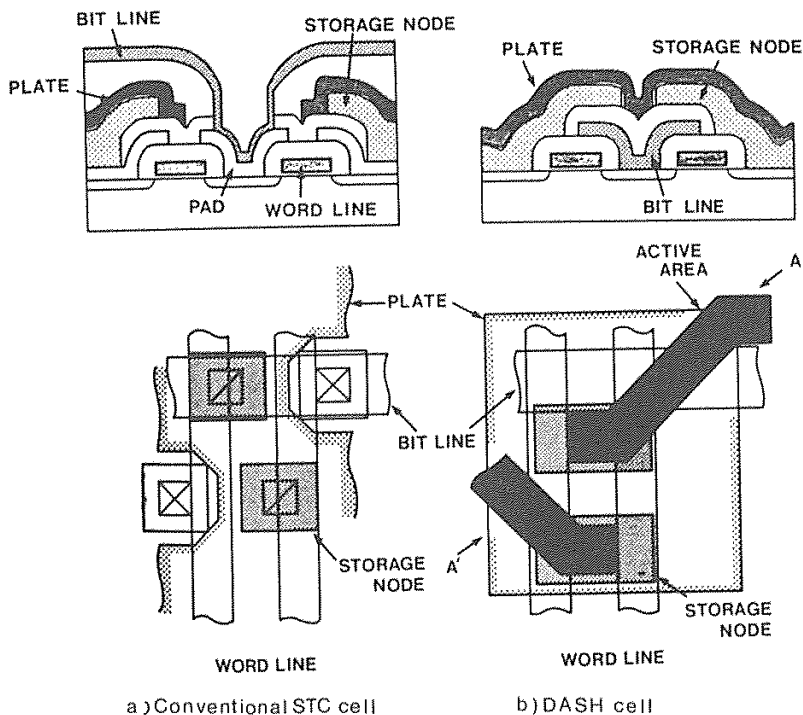


Fig. 8-28 Cross section and layout of: (a) conventional STC cell. (b) DASH cell.⁷⁵ (© 1988 IEEE).

word lines, which increases the cell area without increasing its size. This cell is thus called a *diagonal active-stacked-capacitor cell with a highly packed storage node (DASH)*.⁷⁵ Figure 8-28a shows the cross section of the DASH cell, and Fig. 8-28b compares the layouts of this type of cell with that of a conventional STC cell. The DASH cell yields a 35-fF capacitance for a cell size of $3.4 \mu\text{m}^2$, and hence could be used in 16-Mbit DRAMs.

In the third novel STC cell, a unique fin structure is used to fabricate a capacitor with a high capacitance in a small area. (Figure 8-29a shows the fabrication process sequence.)⁷⁶ This structure would allow cells with two fins to be used in a 16-Mbit DRAM. When the bit lines are formed prior to formation of the fin structure (Fig. 8-29b), a cell structure can be obtained that has sufficiently high capacitance in a small enough area to allow fabrication of a 64-Mbit DRAM.

A fourth novel STC cell, called a *spread stacked capacitor (SSC)* cell,¹²² the storage electrode is expanded into the neighboring 2nd memory cell area (Fig. 8-29c and d), and the storage electrode of the 2nd memory cell is expanded to the 1st memory-cell area.

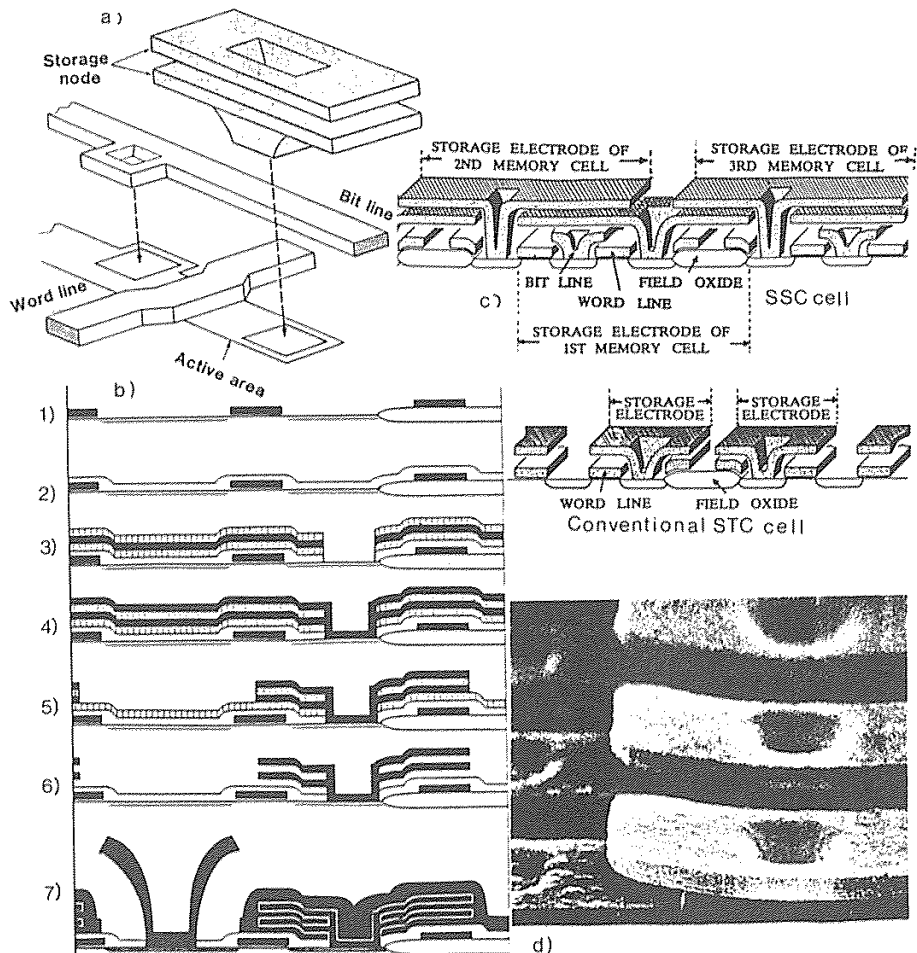


Fig. 8-29 (a) Schematic view of fin-STC structure. (b) Schematic view of fin structure and fabrication sequence.⁷⁶ (© 1988 IEEE). (c) Schematic cross sections of the SSC cell and a conventional-STC cell. (d) SEM photo of a spread-stacked capacitor.¹²² (© 1989 IEEE).

This allows the storage capacitance of the SSC cell to be $\sim 1.8\times$ as great as that of the conventional STC cell, making a possible candidate for 64-Mbit DRAMs.

8.3.5 Soft-Error Failures in DRAMs

When a DRAM is tested, each cell is operated to verify that it functions correctly. A *hard fail*, or *hard error*, indicates that a particular array location repeatedly fails to output

correct data values previously written into the location. Although there are a number of potential sources of hard fails, by and large they are caused by random physical defects. As noted earlier, various layout design approaches and on-chip error-detection and correction circuits (that replace failed bits with spares), have been implemented to eliminate faulty cells from a memory, allowing some defectively manufactured chips to be salvaged.

Soft errors, which are single-nonrecurring read errors on single bits of a memory array, are also a significant problem in DRAMs. A soft error is not a permanent error, in the sense that the cause is not a process defect. A write (then read) cycle in an array location that was previously in error carries no greater or lesser probability of error again than does a cycle in any other array location.

Although soft errors can be caused by such circuit related problems as supply-voltage noise, inadequate noise margins, and sense-amplifier imbalance, there is one specific *physical* failure mode that will cause soft errors, even when all circuit-related failure modes are eliminated. The cause of this failure mode was identified in 1979 by May and Woods⁷⁷ as the alpha particles originating from the decay of uranium and thorium atoms. These radioactive atoms are naturally occurring trace impurities in the materials used to make IC packages, and the alpha particles they emit have energies in the 8-9 MeV range.

Since a bit of information is stored on a cell by the presence or absence of charge in the potential well of the storage capacitor, the number of electrons that distinguishes an empty well from a full one (and hence which differentiates between a logical *one* and a *zero*) is known as the *critical-upset charge*. The generally quoted value for this charge is 45-50 fC (2.5×10^5 electrons),⁷⁸ which is about 25% of Q_s (see section 8.3.1.8).

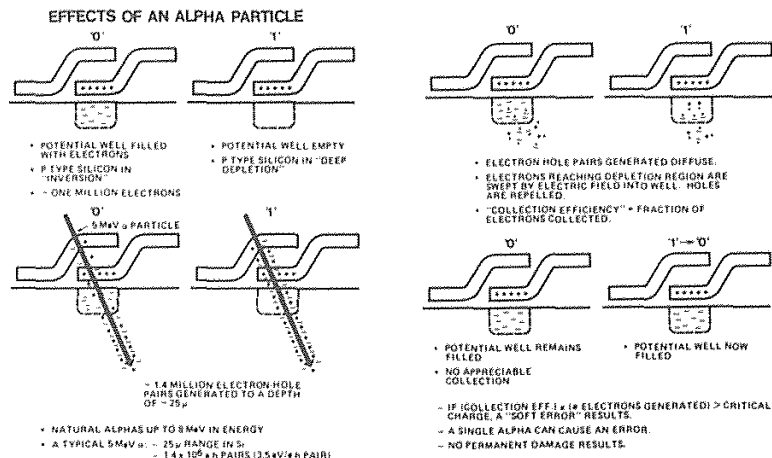


Fig. 8-30 Collection process of free carriers generated by an incident alpha particle. This shows that the cell is sensitive to the alpha particle hit when the storage node is depleted, which is when the cell is storing a *one* for the double-poly cell shown.⁷⁷ (© 1979 IEEE).

If the temperature is raised, or if light is incident on the memory, the generation rate of electron-hole pairs in the substrate can be significantly increased. If enough electrons are produced that an empty well fills up in the interval between refresh pulses, a soft error will occur. Electron-hole pairs are also produced by ionizing radiation incident on a memory chip – specifically, when energetic alpha particles strike a semiconductor substrate (Fig. 8-30).

The electron-hole pairs produced as an alpha particle first passes through the chip's passivation layers are not collected by the empty well, implying that some fraction of the initial alpha-particle energy is lost before it reaches the substrate. Even if only half of the initial energy is retained, there will still be enough energy (4 MeV) to produce roughly 10^6 electron-hole pairs along a trajectory $\sim 25\text{ }\mu\text{m}$ in length.⁷⁹ Any of the electrons generated within the potential well will be swept into the storage node by the depletion region electric field. Those electrons generated in the bulk that diffuse to the edge of the well will likewise be collected by the storage node. The remainder will recombine in the bulk. If a large enough fraction of the electrons generated by an alpha-particle strike is collected by the empty potential well, a soft error can result. The entire collection process occurs in a matter of microseconds.

8.3.5.1 Techniques Used to Reduce Soft-Error Rates in DRAMs.

The *soft-error rate* (SER), expressed as the number of errors per hour, describes the degree of susceptibility to soft-error phenomena. When the storage nodes on silicon devices were on the order of $25\text{ }\mu\text{m}$ in length, it was possible (but not likely) for all of the electrons from one strike to be collected on one node. However, since Q_s for these structures was also larger (e.g., 2×10^6 electrons), the SER was so low that the problem was not noticed. As devices decreased in size, SERs grew significantly larger.

Soft errors in DRAMs were first reported in 16-kbit DRAMs (1 soft-failure per 1000 hours of operation was typical), and they became an appreciable concern in the design and packaging of 64-kbit DRAMs and larger. When new DRAMs are designed, an attempt is made to ensure that the SER is comparable to the hard-error rate.

It was found that as long as a DRAM cell can store more than 6×10^5 electrons, the SER can be made comparable to the hard-error rate. For storage devices in DRAMs of up to 16 kbits, this was accomplished by scaling the oxide thickness to maintain adequate capacitance in the cell. The reliability trade-off in this approach was an increased hard failure rate due to increased defects in the thinner gate oxide. The Hi-C cell was invented as a means of increasing the storage capacity without increasing the capacitor size. An added benefit of this cell is that the doping gradient of the deep *p*-implant produces a potential energy barrier to the diffusing electrons, preventing some of them from reaching the depleted well of the storage capacitor. Hence, the SER of memories built with such cells is decreased. An additional advance was the implementation of the insulator of the storage capacitor with a dual dielectric film (SiO_2 and Si_3N_4), which increased the cell's capacitance (see section 8.3.1.5).

The same SER-reducing phenomenon is exploited when DRAMs are built in CMOS technology. That is, the entire array of NMOS DRAM cells is built in a *p*-well. The well-substrate junction acts as a reflecting barrier for diffusing minority carriers created

outside of the well, preventing most of them from reaching the unfilled storage nodes. Putting the memory on a heavily doped substrate with a lightly doped epitaxial layer on the surface is also beneficial. Since minority carriers generated in the heavily doped substrate have much shorter lifetimes than those in the lightly doped regions, they are prevented from reaching the epitaxial region in which the cells are built (see also section 6.8.4.2).

The use of trench-capacitor structures can also reduce the SER of DRAMs, increasing the number of stored electrons per cell without requiring the lateral device or chip to be made larger. A trench cell in which the storage plate is inside the trench provides better SER than one in which the storage plate is on the outer walls. (If there is little or no isolation between closely spaced trench cells and the track of an alpha particle intersects two of the cells, considerable charge can be transferred. The charge flow along the alpha track can cause transient forward-biasing of the cell junctions. This bipolar-like effect becomes more dominant as the cell-to-cell spacing decreases.)¹⁰⁵

A final method used to reduce SER is to apply a thick coating of a radioactive-contaminant-free polymer on top of the IC. For example, alpha particles with energies of up to 8 MeV are completely absorbed by a 50- μm -thick layer of polyimide. In addition, packaging materials are now manufactured with much lower concentrations of radioactive impurities. Purification of the materials used in wafer fabrication (particularly metallization) is also being pursued, and trace levels of radioactive impurities are now low enough that they can be eliminated them as a significant source of alpha particles.

Even with thick polymer coatings, chips are still vulnerable to strikes by high-energy cosmic rays (in the form of high-atomic-number nuclei). A small fraction of such very energetic particles can penetrate the earth's atmosphere, the package, and the polymer coating, entering the silicon. Under some conditions, these nuclei can also produce soft-errors through the generation of electron-hole pairs. At present, it is estimated that cosmic-ray events produce an SER about an order of magnitude lower than that due to alpha particles from currently available "clean" packaging materials.

8.3.6 The DRAM as a Technology Driver

The DRAM has also been used as a *technology driver* over a large part of its life, since it makes a good test vehicle for advancing silicon integrated-circuit process technology. The regular, repetitive architecture of the DRAM chip requires the least amount of engineering design time to create a circuit with hundreds of thousands, or millions, of devices that can be produced in full-scale production in the factory. Hence, a new MOS technology can be fully tested in the shortest time using the DRAM as the production test vehicle. After the flaws in the process and manufacturing procedures have been ironed out using the technology driver, the process can be transferred to the manufacture of other, more design-intensive circuits.

In the mid-1980s, U.S. semiconductor manufacturers increasingly turned to high-density MOS logic arrays as a replacement technology-driver circuit, since most of them had been forced to abandon the the DRAM market after the price erosion of the 64-kbit

DRAM. The American preference for the logic test chip was also based on the fact that considerable engineering and manufacturing expertise from the production of such dense logic chips as the 80386 and 68000 MPU families. By the late 1980s, however, the world wide manufacturing trend had swung to CMOS SRAMs, since these circuits are easier to design than DRAM circuits (which require intricate clock circuitry). In addition, the 4-Mbit and larger DRAMs use three-dimensional capacitor structures that are unique to the DRAM, and their yield statistics may thus not be as valuable for learning how to manufacture other circuits that do not utilize such structures.

In 1989, however, Osamu of Toshiba presented a case for why DRAMs could still serve as excellent technology drivers.⁸⁰ First, the repetitive structure allows failure categories to be easily identified so that efforts can be undertaken to correct processes that lead to the failures. Second, the factors that impact the RAM yield can be analyzed more completely and in a much shorter period of time than those in logic devices (i.e., the tasks of test-vector generation and design for testability are much less onerous for RAMs than for logic circuits). Finally, the volume production of DRAMs is so great that memory-fabrication processing experience is acquired much more rapidly. These lessons can be transferred to other volume IC production lines. For example, DRAM production can reach 10^6 parts per year at a single manufacturer (with 2×10^6 devices per chip in the 1-Mbit DRAM), versus ~50K microprocessor parts per year (with 1×10^6 devices per typical 32-bit microprocessor).

8.4 MASKED READ-ONLY MEMORIES (ROMS)

Masked (or *mask-programmed*) ROMs are nonvolatile memories into which information is permanently stored through the use of custom masks during fabrication. Users thus must provide the ROM manufacturer with the desired bit pattern of the memory. Subsequent changes of stored data are impossible, and only *read* operations can be performed. Since only a single customized mask is required to personalize these ROMs for a specific application, however, many designs can be economically implemented.

Such memory circuits have been implemented in bipolar, NMOS, and CMOS technologies. As of 1989, 4-Mbit CMOS ROMs were the largest available, but parts containing as many as 32 Mbits are envisioned. The fastest high-density MOS-based ROMs have access times of about 80 ns,⁸¹ while less dense (256-kbit) ROMs are faster (50 ns). The faster ROMs are most often used to simplify the interface to microprocessors by eliminating wait states, thereby permitting more rapid program execution. Bipolar ROMs are even faster (e.g., 10 ns access time for a 1-kbit ECL ROM), but they are also much less dense.

In addition to interfacing with microprocessors, ROMs have been used for a variety of applications, including these:

- *Look-Up Tables.* These are used for mathematical calculations in which evaluations of square roots and of trigonometric functions, logarithmic functions, and exponential functions are needed. The procedure would usually be much more time consuming if software subroutines were used to calculate the series

expansions of particular functions. Other look-up applications include spell checking and dictionary servicing.

▪ *Character Generators.* All digital systems rely on the display of alphanumeric characters of such input and output devices as CRTs and dot-matrix printers. In most of these applications, the patterns of pixels, segments, or dots used to display the set of characters are stored in a ROM. In Japan, large (1-Mbit and larger) ROMs are used as character generators for the complex *kanji* characters.

▪ *Microcontrol Store.* In microprogrammed digital systems, the execution of an encoded macroinstruction involves the generation of a sequence of signal vectors for gating or control purposes. ROMs are often chosen for such applications because their higher packing density and simpler fabrication procedure make them less costly than static or dynamic RAMs. In addition, ROMs are more stable than SRAMs and DRAMs. Parts of computer programs that are completely debugged and that do not need to be rewritten during computation are thus often stored in ROMs.

8.4.1 Masked-ROM Implementation

Each bit of information in a ROM is stored by the presence or absence of a data path from the word (access) to a bit (sense) line. The data path is eliminated simply by ensuring that no circuit element joins a word and bit line. Thus, when the word line of a ROM is activated, the presence of a signal on the bit line will mean that a 1 is stored, whereas the absence of a signal will indicate that the bit location is storing a 0. As shown in Fig. 8-31 (which uses an NMOS array as the example), there can be two basic forms of the ROM, with implementation by either the NOR function (Fig. 8-31a), or the NAND function (Fig. 8-31b). Note that programming of masked ROMs in bipolar technology is done by selectively omitting a contact, at the contact mask in the emitter-follower or Schottky-diode ROM arrays shown in Fig. 8-31d.

Although the speed of the ROM depends on the details of the MOS fabrication process, NOR arrays usually have faster access times; in addition, the stored bit pattern can be set by the metal-interconnection layer. Therefore, unprogrammed NOR ROMs can be manufactured up until the metal mask step and can then be stored in inventory. Such almost-completed wafers can be quickly completed (programmed) by using a custom mask that patterns the metal layer.

The NAND-type ROMs, on the other hand, have a longer access time, and they must be programmed through the implantation of dopants into the channel of selected transistors wherever stored zeros are desired (Fig. 8-31c). Such a step must be performed earlier in the manufacturing sequence, which increases the TAT. The advantage of NAND ROMs over the NOR type is that they have a considerably higher density when fabricated using the same process and design rules.

The access time of ROMs is limited by the resistance and capacitance of the word and bit lines, as well as by the currents available to drive these lines. Because of their higher density (i.e., 1-Mbit CMOS ROMs were introduced in 1983), ROMs were the

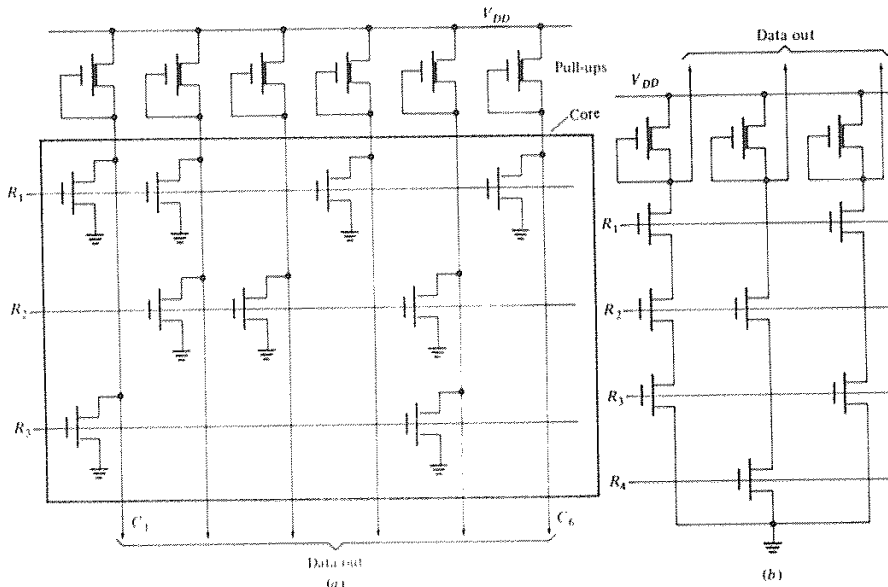


Fig. 8-31 (a) MOS ROM NOR array, (b) MOS ROM NAND array.¹

first semiconductor memories in which on-chip error-correction circuitry (ECC) was implemented.

8.5 PROGRAMMABLE ROMS (PROMS)

If only a small quantity of ROM circuits is needed for a specific application, custom fabrication of even a single mask layer may be too expensive and/or time consuming. In such cases, it is faster and cheaper for users to program each ROM chip individually. ROMs with such capabilities are referred to as *user or field-programmable* memories. Many types of these have been developed.

In this section we describe *programmable read-only memories* (PROMs), a type of ROM into which information can be *programmed only once* and then cannot be erased. Subsequent sections describe ROMs that allow data to be erased after entry.

In PROMs, a data path exists between *every* word and bit line at the completion of chip manufacture (corresponding to a stored *1* in every data position). Storage cells are selectively altered to store a *0* following manufacture by electrically *blowing open* the appropriate word-to-bit connection paths. Since the write operation is destructive, once a *0* has been programmed into a bit location it cannot be erased back to a *1* in a PROM. PROMs were originally implemented in bipolar technology, although MOS PROMs have recently become available as well.

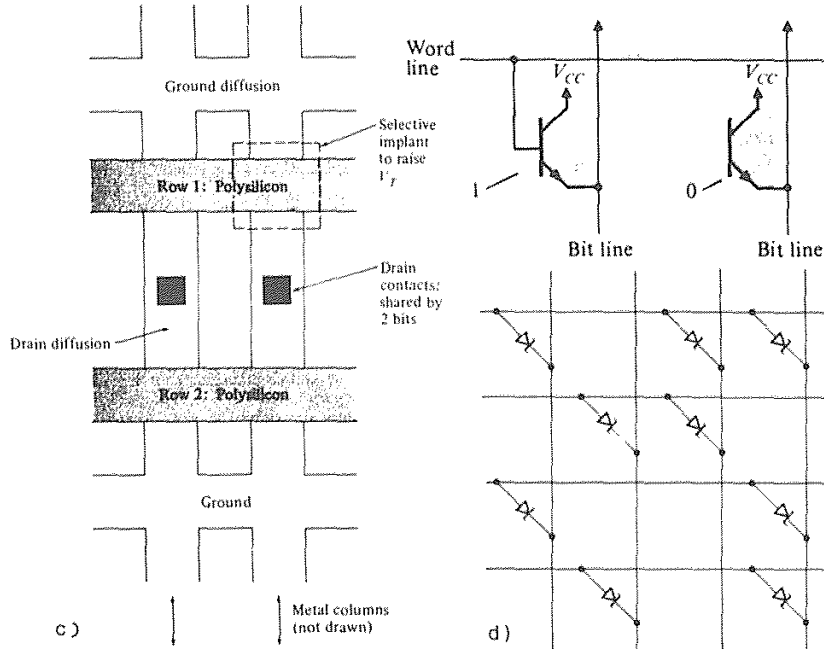


Fig. 8-31 (c) MOS NAND ROM - simplified layout, (d) BJT ROM cells.¹ From D. A. Hodges and H. G. Jackson, *Analysis and Design of Digital Integrated Circuits*, Copyright, 1983 McGraw-Hill Book Co. Reprinted with permission.

In bipolar PROMs, two techniques are used to eradicate the word-to-bit-line paths in desired bit locations. The first involves the use of a bipolar transistor in series with a *fuse*, as shown in Fig. 8-32. In such *emitter-follower* bipolar PROMs, a small fusible link is placed in series with each emitter. Early links were implemented with nichrome thin films, but these exhibited a *growback* problem (in which disconnected fuses became reconnected after some time). PROM fuses are now generally made of polysilicon. Fuse-link PROMs are designed to operate with a 5-V supply for reading data, but higher voltages (10-15 V) are needed to produce the 10-30 mA required to blow open the fuses. Such large voltages may be supplied by off-chip programming devices or by special electronic circuits available on the chip.

The second technique makes use of a *pn* diode that is short-circuited by an avalanching pulse. An example of a 64-kbit bipolar PROM using *pn* diode cells with an access time of 50 ns is described in reference 82.

MOS PROMs have also been introduced, and in 1985 they became available in 1-Mbit size. Such components are actually MOS erasable and programmable read-only memories (EPROMs) housed in inexpensive plastic packages (rather than in the costly, quartz-windowed ceramic packages needed by *erasable* MOS PROMs). Without a quartz

window these MOS PROMs can be field-programmed only once. As a result, they are commonly known as *one-time-programmable* (OTP) ROMs. The advantage of these over bipolar PROMs is that they can be fabricated in much higher densities. High-density CMOS OTP ROMs are now being built with access times close to those of bipolar PROMs, but with more bits per chip and much lower power dissipation. For example, a 256-kbit CMOS OTP ROM with an access times of 50 ns has recently been introduced; this approaches the access time (~ 40 ns) of large [64-kbit] bipolar PROMs.⁸⁷ In addition, OTP ROMs are no longer much more expensive than ROMs, and hence they are also expected to increasingly replace masked ROMs.

8.6 ERASABLE PROGRAMMABLE READ-ONLY MEMORIES (EPROMS)

Erasable PROMs depend on the long-term retention of electronic charge as the information-storage mechanism. The charge is stored on a *floating polysilicon gate* of an MOS device (the term *floating* refers to the fact that no electrical connection exists to this gate). The charge is transferred from the silicon substrate through an insulator.

Each of the various mechanisms implemented to transfer (and remove) charge from the floating gate has been the basis of a different erasable-PROM device type. This section describes the so-called *electrically programmable ROM* (EPROM), which also requires that the device be irradiated with ultraviolet (UV) light for removing (or *erasing*) the stored charge from the floating gate.

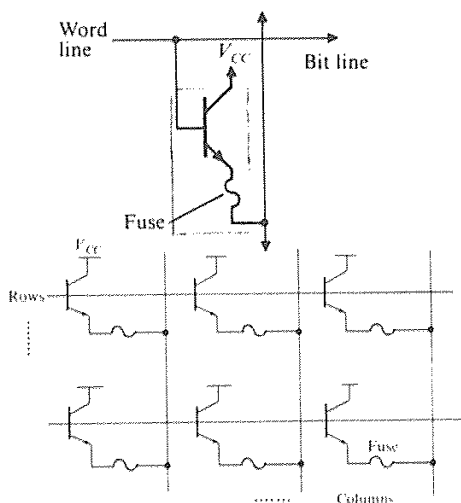


Fig. 8-32 (a) Emitter-follower bipolar PROM.¹ From D. A. Hodges and H. G. Jackson, *Analysis and Design of Digital Integrated Circuits*, Copyright, 1983 McGraw-Hill Book Co. Reprinted with permission.

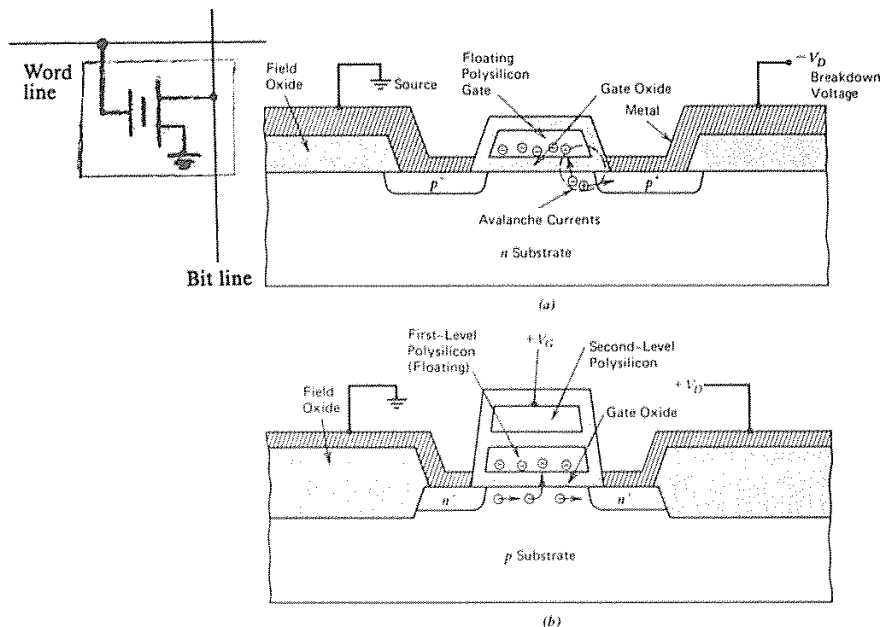


Fig. 8-33 (a) Circuit schematic and cross section showing the mechanism of charge injection into the gate by avalanche in a FAMOS memory element. (b) A FAMOS element made with two layers of polysilicon and suitable for n -channel MOS applications. From R. S. Muller and T. I. Kamins, *Device Electronics for Integrated Circuits*, 2nd Ed. Copyright 1986, John Wiley & Sons. Reprinted with permission.

Traditionally, such EPROMs have been used as prototyping vehicles to ensure that no glitches remained in the code. Once the programs were finalized, the code was usually fixed into ROM components. However, the cost of EPROMs is shrinking with advances in technology, and as a result, their use is growing at the expense of ROMs. Another factor in favor of EPROMs is their faster turn around time (which also plays a role in the choice of technology used to implement masked ROMs).

The charge-transfer mechanism is based on the injection of hot electrons into the floating polysilicon gate, which is completely encapsulated by SiO_2 . The original EPROM devices were fabricated in PMOS technology and consisted simply of a MOSFET with a floating gate (Fig. 8-33a). If a sufficiently high reverse-bias voltage is applied to the drain, the drain-substrate pn junction will experience avalanche breakdown, causing hot electrons to be generated. Some of these will have enough energy to pass over the oxide potential-energy barrier and charge the floating gate (see section 5.6.2). These EPROM devices were thus called *Floating-gate, Avalanche-injection MOS transistors* (FAMOS).

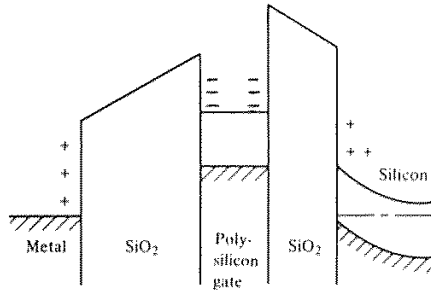


Fig. 8-34 Energy band diagram of a FAMOS device with charge stored in the silicon gate. From E. S. Yang, *Microelectronic Devices*, Copyright 1988, McGraw-Hill Book Company. Reprinted with permission.

Once electrons are transferred to the gate, they are trapped there, as illustrated by the energy-band diagram shown in Fig. 8-34. Since the potential-energy barrier at the oxide-silicon interface is greater than 3 eV, the rate of spontaneous emission of electrons from the oxide over this barrier is negligibly small. The electronic charge on the floating gate can thus be retained for many years.

If the floating gate is charged with a sufficient number of electrons, inversion of the channel under the gate will occur. A conducting channel then forms between the source and the drain, exactly as would occur if an external gate voltage were applied. The presence of a 1 or 0 in each bit location is therefore determined by the presence or absence of a conducting channel in a programmed device.

Subsequent advances in process technology (Fig. 8-33b) made it possible to implement EPROMs with 5 V, *n*-channel devices.^{83,84} In such EPROMs the cells can also be laid out in NOR or NAND arrays; we will use the NOR array configuration to describe the operation of these newer cells.

Two layers of polysilicon are used to form a double gate in the transistor, as shown in Fig. 8-33b. Gate #1 is the floating gate and is placed under Gate #2. Cell selection is controlled by Gate #2, which therefore plays the role of the single gate in conventional MOS transistors. Initially, Gate #1 is uncharged; thus, if the drain, source, and Gate #2 of the transistor are grounded, Gate #1 will also be at 0 V. If a voltage (V_2) is subsequently applied to Gate #2, the voltage on Gate #1 (V_1) will be given by:

$$V_1 = [C_2 / (C_1 + C_2)] V_2 \quad (8-3)$$

because the two gates represent a capacitive divider as shown in Fig. 8-35. From the electrical perspective of Gate #2, the transistor appears to have a larger V_T . In order to turn on this transistor, a larger gate voltage must be applied to Gate #2 (typically somewhat more than twice the normal V_T). For example, if a conventional NMOS

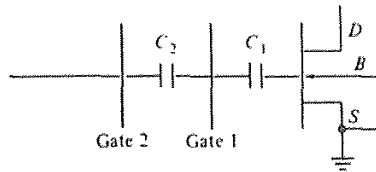


Fig. 8-35 Equivalent capacitive divider of an EPROM structure.

device with a $V_T = 1$ V was fabricated with the double gate of Fig. 8-33b, a voltage of 2 V would have to be applied to Gate #2 to turn it *ON*; a voltage of 5 V for reading the cell would also cause it to turn *ON* (Fig. 8-36). Such a turned-on device would cause a *positive logic stored zero* to appear at the output of the bit line if the device was used in a NOR array. As a result, the programming of the EPROM begins by discharging all of the floating gates through exposure to UV radiation, so that every cell initially stores a 0. A 1 is then selectively written into the desired cells.

For a 1 to be written into a cell, both Gate #2 and the drain are raised to about 12 V (for a few hundred microseconds), while the source and substrate are kept grounded (early EPROMs required 30 V programming voltages for several milliseconds). Hot electrons are created near the drain and are attracted to the floating gate (which, due to capacitive coupling, has a more positive potential than the drain). Some fraction of the electrons will traverse the oxide and charge the floating gate. When the voltages on Gate #2 and the drain are returned to zero, these charges remain trapped on Gate #1. The electrons trapped on Gate #1 cause its potential to be at about -5 V. Therefore, if a signal of only 5 V is applied to Gate #2 when the EPROM is being read, no channel will form in the transistor. Under this circumstance, a 1 is stored in the cell. The electron-trapping process is self-limiting, because once electrons are stored on the floating gate they begin to inhibit further electron injection.

In order for the cells to be erased, the stored charge must be removed from the floating gate. This is accomplished by flood exposure of the EPROM with strong ultraviolet light for approximately 20 minutes. The UV light creates electron-hole pairs in the SiO_2 , providing a discharge path for the charged floating gate.

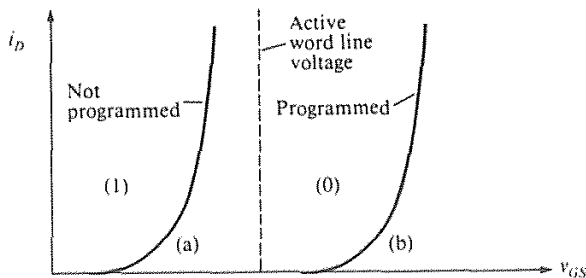


Fig. 8-36 Transfer characteristic of a floating-gate transistor.

One of the advantages of EPROMs is that the cells consist of only one transistor, allowing them to be fabricated with high densities (e.g., a 4-Mbit CMOS EPROM with an access time of 120 ns and 0.8- μm channel-length transistors has been reported).⁸⁵ In addition, they cost less to manufacture than electrically erasable PROMs (EEPROMs – see the next section).

A disadvantage of EPROMs is that they require UV light for erasing and must therefore be packaged in an expensive ceramic package with a UV-transparent quartz window. In addition, they must be removed from the circuit board and put into a special UV eraser. (Note that since sunlight and fluorescent lamps contain some UV, one week of sunlight or three years of room-level fluorescent lighting are likely to erase some of the cells. Therefore, except during erasure, the window should be covered at all times with an opaque label.) Another disadvantage is that the high voltage needed to program the EPROM is generally not available on the integrated circuit, so a special programming setup must also be provided. This limitation, combined with the fact that EPROM programming takes a relatively long time, means that these cells are used primarily for reading information and are only occasionally rewritten. (Note, however, that the programming time is decreasing dramatically. In the first 64-kbit EPROMs, it took about 50 ms to program each byte, adding up to almost seven minutes for the entire chip; in the 4-Mbit EPROM,⁸⁵ the program time has been reduced to 10 μs /byte, so that the entire chip can be programmed in only five seconds!)

OTP ROMs compete with high-density masked ROMs because they offer the benefit of a significantly shorter TAT (albeit at a somewhat higher cost). OTP ROMs are also less expensive than bipolar PROMs, and offer a PROM capability with much higher density. While bipolar PROMs are generally faster, a three-transistor EPROM cell was recently reported that would allow CMOS EPROMs to be built with the same speed and density as bipolar PROMs, but with much lower power dissipation and 100% testability.⁸⁶ Another one-transistor cell, split-gate 256-kbit CMOS EPROM with an access time of 50 ns has also been reported.⁸⁷

Some of the relevant process and circuit-design enhancements used in fabricating the large CMOS EPROMs include the following:⁸⁸

- Use of thin (20 nm) reliable interpoly dielectric materials, often consisting of composite films of $\text{SiO}_2/\text{Si}_3\text{N}_4/\text{SiO}_2$, for increased capacitive coupling between Gates #1 and #2.
- Self-aligned contacts to the control gate (as well as a self-aligned floating gate) to achieve the $3.1 \times 2.9 \mu\text{m}^2$ small cell size.
- Use of a low-resistance polycide gate for the word line to achieve high speed.
- Reduction of the programming voltage to 10.5 V, along with a reduction of the programming time to $\sim 10 \mu\text{s}$ /byte.
- On-chip test circuits.

A novel self-aligned planar-array EPROM cell has also been proposed.^{89,90} This

cell appears to make possible the fabrication of 4-Mbit EPROMs with 1- μm design rules because it uses buried n^+ bit lines that are self-aligned to the FAMOS transistor.

8.7 ELECTRICALLY ERASABLE PROMS (EEPROMS)

In some applications it is desirable to erase the contents of a ROM electrically, rather than to use a UV light source. In other circumstances it is useful to be able to change one byte at a time, without having to erase the entire IC. A variety of *electrically erasable PROMs* have been developed to serve these applications. Such EEPROMs are the most sophisticated of the ROM families in terms of the physical operating principles and process complexity. For example, EEPROMs must be fabricated with unique tunnel oxides, as well as with high-voltage transistors (for programming and erasing the devices).

Three technologies have been developed for EEPROM fabrication: (1) MNOS transistors; (2) Floating-gate Tunnel Oxide (FLOTOX) MOS transistors; and (3) textured-polysilicon floating-gate MOS transistors. Although MNOS transistor-based devices were among the first EEPROMs to be commercially manufactured, their technology limitations have made them less widely adopted than the others. Therefore, we will devote most of our attention to FLOTOX and textured-poly EEPROMs.

8.7.1 MNOS-Based EEPROMs

The MNOS EEPROM cell consists of a single MOS-like transistor that employs a composite gate-dielectric layer (Si_3N_4 , ~50 nm thick, on top of a very thin [~2 nm thick] SiO_2 layer), as shown in Fig. 8-37. (See ref. 118 for more details on MNOS devices.) Unlike in floating-gate MOS devices, the charge is stored in discrete traps in the nitride bulk. The charge transfers from the substrate to the nitride traps (and back, during erasure) by tunneling through the thin oxide layer. Programming is accomplished by applying a high voltage to the top gate; erasing is done by grounding the top gate and raising the well to a high potential. MNOS transistors are built within wells (akin

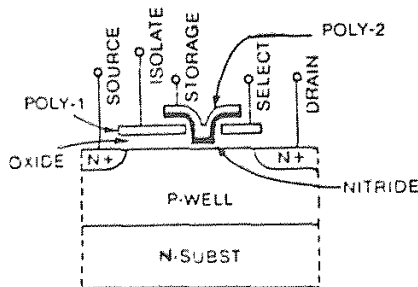


Fig. 8-37. Cross-sectional structure of an MNOS memory cell.¹⁰⁷ (© 1983 IEEE).

to those used in CMOS) so that their channel potentials can be controlled.

The manufacturing process of MNOS transistors involves the following modifications to the standard single polysilicon-gate MOS technology: thin oxide growth, nitride deposition, and post-nitride temperature cycles. Mastering the processes used to grow the ultra-thin oxide and deposit the nitride and to control their quality is a challenging task. Furthermore, while the basic transistor is very small and highly scalable, each cell of the memory array requires a select transistor. This requirement, coupled with the need to fabricate wells, produces a relatively large effective cell size. Finally, the charge stored on the nitride traps continually leaks away through the thin oxide by means of tunneling, even when no erase voltage is applied. The charge loss is thus time dependent, making charge retention the main reliability concern with MNOS devices. (With MNOS structures, as the switching speed is increased, the ability to retain stored charge is decreased. Thus, devices with a retention time of tens of years can be fabricated if a slow switching speed can be tolerated.)

Nevertheless, MNOS exhibits higher tolerance to ionizing radiation than do either of the other EEPROM technologies. Thus, MNOS EEPROMs currently find their main use in low-density military applications that need radiation-hardened EEPROMs; this appears to be the niche to which MNOS EEPROMs will be relegated in the future.⁹¹

8.7.2 FLOTOX EEPROMs

The *floating-gate tunneling oxide (FLOTOX)* transistor, shown in Fig. 8-38a, consists of an MOS transistor with two polysilicon gates. A thin (8-12 nm) gate oxide (or oxynitride) region is formed near the drain. The lower polysilicon layer is the floating-gate while the other is the control gate. The remainder of the floating-gate oxide is typically 50 nm thick, and the interpoly oxide is ~50 nm thick. Programming of this transistor is done by causing electrons to be transferred from the substrate to the floating

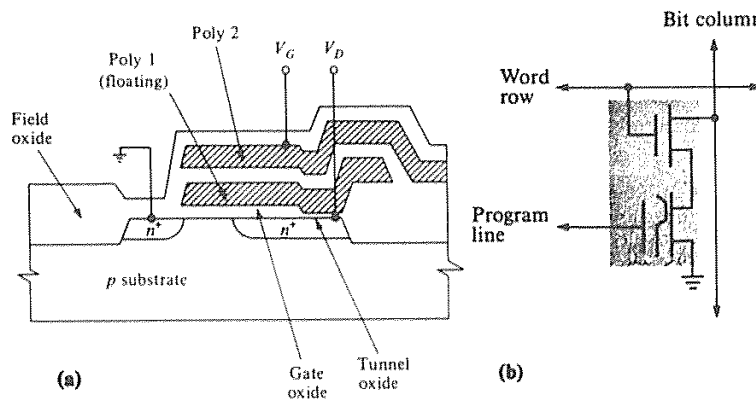


Fig. 8-38 (a) Cell structure of a Flotox transistor structure, (b) Connection in an EEPROM.⁹¹ (© 1986 IEEE).

gate through the thin oxide layer by means of Fowler-Nordheim tunneling.⁹²

The control-gate voltage is raised to a sufficiently high value so that tunneling ensues (e.g., 12 V in modern FLOTOX EEPROMs). As electronic charge builds up on the floating gate, the electric field is reduced, which decreases the electron flow. Since the tunneling process is reversible, the floating gate can be erased by grounding the control gate and raising the drain voltage, indicating that tunneling is used both to *program* and *erase* the FLOTOX transistor. Programming and erase times are on the order of 9 ms. Electron transfer by Fowler-Nordheim tunneling, however, requires a minimum electric-field strength of around 10 MV/cm. Thus, for oxides of 10 nm in thickness, such tunneling will be negligible when normal 5-V signals are applied. As a result, FLOTOX transistors can be expected to retain their charges for more than 10 years if the memory is subjected only to normal read cycles.

The FLOTOX transistor must be isolated by a select transistor. Otherwise, the high voltage applied to the drain of the selected cell during erasing would also appear on the drain of the other *unselected* cells in the same memory column. A FLOTOX EEPROM cell must therefore consist of two transistors (Fig. 8-38b). Although this limits the density of such EEPROMs in comparison to EPROMs and flash EEPROMs, it makes it possible to erase and re-program one byte of the memory without having to erase the entire IC. In addition, two cycles are needed to load the correct data into the memory. In the first, all the cells in a byte are programmed (i.e., the floating gates are charged); in the second, selected cells are erased, with the drain used for data control.

The fabrication of FLOTOX EEPROMs involves a modification of the polysilicon-gate MOS process. A double-polysilicon process is used, together with a thin tunnel-oxide growth process. The growth of a high-quality, thin tunneling oxide is, in fact, the critical manufacturing step in this technology. The tunneling dielectric reportedly can be successfully implemented with nitrided oxides, since the barrier between silicon nitride and silicon is lower than that between SiO₂ and Si. As a result, a higher tunneling current can be obtained for the same voltage.⁹³

Despite the fact that a reliable process for growing thin tunneling oxides must be developed, FLOTOX-based devices have become the most widely manufactured of the EEPROM types. They are still the easiest to learn to manufacture for companies that have already successfully developed an EPROM process. Since it is desirable to be able to program and erase the EEPROM while it remains in place on a PC board, considerable effort has also been expended to make this memory type fully operational with a 5 V power supply. (This type of operational capability is referred to as *5 V only*.)

FLOTOX-based EEPROMs are best suited for applications in which low-cost, low-density, nonvolatile memories are required – for example, in microcontrollers and programmable logic devices. Another potential application is for smart credit cards; several Japanese companies have announced 64-kbit FLOTOX-based EEPROMs for this market.

On the other hand, scaling and reliability considerations appear to limit the maximum size of FLOTOX EEPROMs to 256 kbits. The need for two transistors, and the relatively large size of the select transistor (due to the large voltages needed for

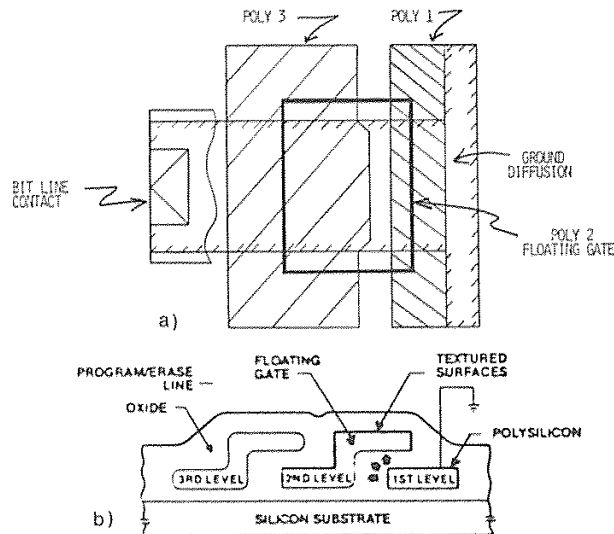


Fig. 8-39 Textured-polysilicon memory cell. (a) Top view. (b) Cross-sectional structure.⁹¹ (© 1986 IEEE).

programming and erasure) are contributing factors. The poly-to-poly area of the sense transistor must also be large, due to the high oxide capacitance of the thin tunnel oxide.⁴⁰ Furthermore, the FLOTOX EEPROMs exhibit high failure rates, caused by defect-related oxide-breakdown problems as memory size is increased.⁹⁴ Error-detection and correction codes (EDCC) can be used to overcome this limitation, but such solutions impose a penalty of increased die cost. Reference 95 gives an example of a 50-ns access time, 256-kbit FLOTOX-based CMOS EEPROM using a single-bit EDCC.

The reliability of FLOTOX-based EEPROMs compares favorably with that of the other two EEPROM types. As with the others, there is a very low failure rate during 5-V operation; reliability problems occur as a result of the high voltages that must be used during programming and erasing. Random single-bit failures occur in FLOTOX devices due to oxide defects, resulting in leaky oxides that lose charge over time. The number of cycles that most FLOTOX EEPROMs are specified to be able to endure before the thin oxide becomes too leaky to retain data sufficiently, is 10^3 – 10^4 cycles. However, a process for increasing this endurance level to 10^6 cycles has been reported.⁹⁶

8.7.3 Textured-Polysilicon EEPROMs

Textured-polysilicon EEPROMs, introduced in 1983 as an alternative to the tunneling oxide types of devices, are also based on the floating-gate MOS technology. The cell consists of three layers of polysilicon that partially overlap (Fig. 8-39) to create a cell