COE 360 Principles of VLSI Design
Dr. Aiman El-Maleh

CMOS Processing Technology

CMOS PROCESSING TECHNOLOGY

3

The purpose of this chapter is to introduce the CMOS designer to the technology that is responsible for the semiconductor devices that might be designed. This is of importance in understanding the potential and limitations of a given technology. It also gives some background for the geometric design rules that are the interface medium between designer and fabricator.

The basics of semiconductor manufacturing are first introduced. Following this, a basic n-well CMOS process is described showing the process steps and how they relate to the design description passed from the designer to the fabrication engineer. Following this, a number of enhancements to the basic CMOS technology are described. Many of these are now required by mainstream CMOS logic and memory designers. The next section introduces the reader to layout design rules that prescribe how to manufacture the CMOS chip. The nature of CMOS latchup and the solutions to this problem are then covered. Finally, some CAD issues as they relate to process technology are covered. An appendix, Section 3.9, outlines the actual steps used in a CMOS process for those who want to get down to that level of detail.

3.1 Silicon Semiconductor Technology: An Overview

Silicon in its pure or *intrinsic* state is a semiconductor, having a bulk electrical resistance somewhere between that of a conductor and an insulator. The conductivity of silicon can be varied over several orders of magnitude by

introducing *impurity* atoms into the silicon crystal lattice. These *dopants* may either supply free electrons or holes. Impurity elements that use electrons are referred to as *acceptors* since they accept some of the electrons already in the silicon, leaving vacancies or holes. Similarly, *donor* elements provide electrons. Silicon that contains a majority of donors is known as *n-type* and that which contains a majority of acceptors is known as *p-type.* When n-type and p-type materials are brought together, the region where the silicon changes from n-type to p-type is called a *junction.* By arranging junctions in certain physical structures and combining these with other physical structures, various semiconductor devices may be constructed. Over the years, silicon semiconductor processing has evolved sophisticated techniques for building these junctions and other structures having special properties.

3.1.1 Wafer Processing

The basic raw material used in modern semiconductor plants is a *wafer* or disk of silicon, which varies from 75 mm to 230 mm in diameter and is less than 1 mm thick. Wafers are cut from ingots of single-crystal silicon that have been pulled from a crucible melt of pure molten polycrystalline silicon. This is known as the 'Czochralski,' method (Fig. 3.1) and is currently the most common method for producing single-crystal material. Controlled amounts of impurities are added to the melt to provide the crystal with the

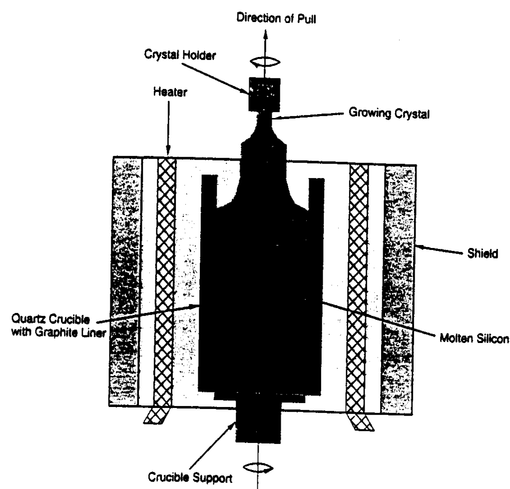

FIGURE 3.1 Czochralski method for manufacturing silicon ingots

required electrical properties. The crystal orientation is determined by a seed crystal that is dipped into the melt to initiate single-crystal growth. The melt is contained in a quartz crucible, which is surrounded by a graphite radiator. The graphite is heated by radio frequency induction and the temperature is maintained a few degrees above the melting point of silicon (≈1425°C). The atmosphere above the melt is typically helium or argon.

After the seed is dipped into the melt, the seed is gradually withdrawn vertically from the melt while simultaneously being rotated. The molten polycrystalline silicon melts the tip of the seed, and as it is withdrawn, refreezing occurs. As the melt freezes, it assumes the single crystal form of the seed. This process is continued until the melt is consumed. The diameter of the ingot is determined by the seed withdrawal rate and the seed rotation rate. Growth rates range from 30 to 180 mm/hour.

Slicing into wafers is usually carried out using internal cutting-edge diamond blades. Wafers are usually between 0.25 mm and 1.0 mm thick, depending on their diameter. Following this operation, at least one face is polished to a flat, scratch-free mirror finish.

3.1.2 Oxidation

Many of the structures and manufacturing techniques used to make silicon integrated circuits rely on the properties of the oxide of silicon, namely, silicon dioxide (SiO_2). Therefore the reliable manufacture of SiO_2 is extremely important.

Oxidation of silicon is achieved by heating silicon wafers in an oxidizing atmosphere such as oxygen or water vapor. The two common approaches are:

- Wet oxidation: when the oxidizing atmosphere contains water vapor. The temperature is usually between 900°C and 1000°C. This is a rapid process.
- Dry oxidation: when the oxidizing atmosphere is pure oxygen. Temperatures are in the region of 1200°C, to achieve an acceptable growth rate.

The oxidation process consumes silicon. Since SiO_2 has approximately twice the volume of silicon, the SiO_2 layer grows almost equally in both vertical directions. This effect is shown in Fig. 3.2 for an n-channel MOS device in which the SiO_2 (field oxide) projects above and below the unoxidized silicon surface.

3.1.3 Epitaxy, Deposition, Ion-Implantation, and Diffusion

To build various semiconductor devices, silicon containing varying proportions of donor or acceptor impurities is required. This may be achieved using epitaxy, deposition, or implantation. Epitaxy involves growing a single-crys-

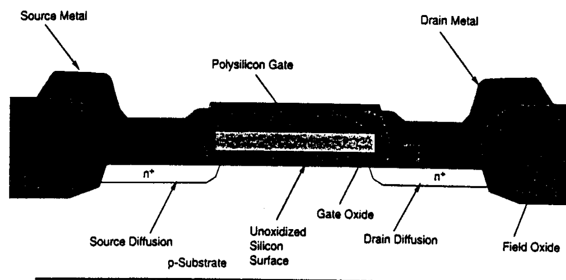

FIGURE 3.2 An nMOS transistor showing the growth of field oxide below the silicon surface

tal film on the silicon surface (which is already a single crystal) by subjecting the silicon wafer surface to elevated temperature and a source of dopant material. Deposition might involve evaporating dopant material onto the silicon surface followed by a thermal cycle, which is used to drive the impurities from the surface of the silicon into the bulk. Ion implantation involves subjecting the silicon substrate to highly energized donor or acceptor atoms. When these atoms impinge on the silicon surface, they travel below the surface of the silicon, forming regions with varying doping concentrations. At any elevated temperature (> 800°C) diffusion will occur between any silicon that has differing densities of impurities, with impurities tending to diffuse from areas of high concentration to areas of low concentration. Hence it is important once the doped areas have been put in place to keep the remaining process steps at as low a temperature as possible.

Construction of transistors and other structures of interest depends on the ability to control where and how many and what type of impurities are introduced into the silicon surface. What type of impurities are introduced is controlled by the dopant source. Boron is frequently used for creating acceptor silicon, while arsenic and phosphorous are commonly used to create donor silicon. How much is used is determined by the energy and time of the ion-implantation or the time and temperature of the deposition and diffusion step. Where it is used is determined by using special materials as masks. In places covered by the mask ion implantation does not occur or the dopant does not contact the silicon surface. In areas where the mask is absent the implantation occurs, or the predeposited material is allowed to diffuse into the silicon. The common materials used as masks include

- photoresist.
- polysilicon (polycrystalline silicon).
- silicon dioxide (SiO_2).
- silicon nitride (SiN).

The ability of these materials to act as a barrier against doping impurities is a vital factor in this process, called *selective diffusion*. Thus selective diffusion entails

- patterning *windows* in a mask material on the surface of the wafer.
- subjecting exposed areas to a dopant source.
- removing any unrequired mask material.

In the case of an oxide mask, the process used for selectively removing the oxide involves covering the surface of the oxide with an acid resistant coating, except where oxide windows are needed. The SiO_2 is removed using an etching technique. The acid resistant coating is normally a photosensitive organic material called *photoresist* (PR), which can be polymerized by ultraviolet (UV) light. If the UV light is passed through a mask containing the desired pattern, the coating can be polymerized where the pattern is to appear. The polymerized areas may be removed with an organic solvent. Etching of exposed SiO_2 then may proceed. This is called a positive resist. There are also negative resists where the unexposed PR is dissolved by the solvent. This process is illustrated in Fig. 3.3. In established processes using PRs in conjunction with UV light sources, diffraction around the edges of the mask patterns and alignment tolerances limit line widths to around 0.8 μm. During recent years, electron beam lithography (EBL) has emerged as a contender for pattern generation and imaging where line widths of the order of 0.5 μm with good definition are achievable. The main advantages of EBL pattern generation are as follows:

- Patterns are derived directly from digital data.
- There are no intermediate hardware images such as recticles or masks; that is, the process can be direct.
- Different patterns may be accommodated in different sections of the wafer without difficulty.
- Changes to patterns can be implemented quickly.

The main disadvantage that has precluded the use of this technique in commercial fabrication lines is the cost of the equipment and the large amount of time required to access all points on the wafer.

3.1.4 The Silicon Gate Process

So far we have touched on the single-crystal form of silicon used in the manufacture of wafers and the oxide used in the manufacture and operation of circuits. Silicon may also be formed in a *polycrystalline* form (not having a single-crystalline structure) called *polysilicon*. This is used as an intercon-

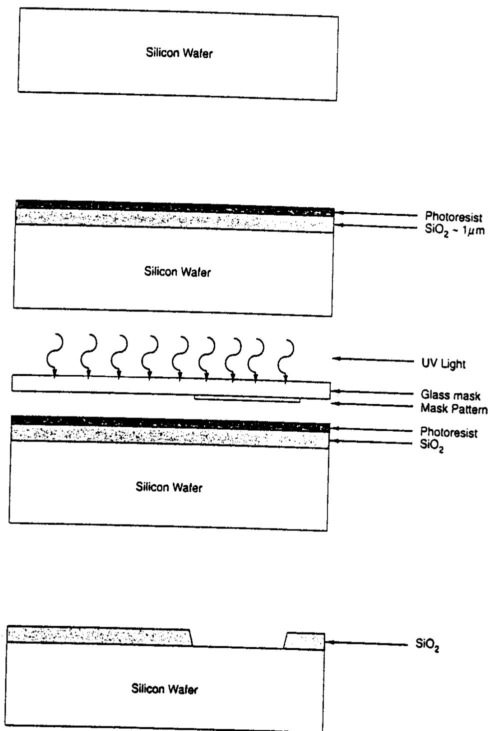

FIGURE 3.3 Simplified steps involved in the patterning of SiO_2: (a) Bare silicon wafer; (b) Wafer with SiO_2 and resist; (c) Exposing resist to UV light; (d) Final etched SiO_2

nect in silicon ICs and as the gate electrode on MOS transistors. The most significant aspect of using polysilicon as the gate electrode is its ability to be used as a further mask to allow precise definition of source and drain electrodes. This is achieved with minimum gate-to-source/drain overlap, which, we will learn, improves circuit performance. Polysilicon is formed when silicon is deposited on SiO_2 or other surfaces. In the case of an MOS transistor gate electrode, undoped polysilicon is deposited on the gate insulator. Polysilicon and source/drain regions are then normally doped at the same time. Undoped polysilicon has high resistivity. This characteristic is used to provide high-value resistors in static memories. The resistivity of polysilicon may be reduced by combining it with a refractory metal (see Section 3.2.4).

The steps involved in a typical silicon gate process entail photomasking and oxide etching, which are repeated a number of times during the processing sequence. Figure 3.4 shows the processing steps after the initial pattern-

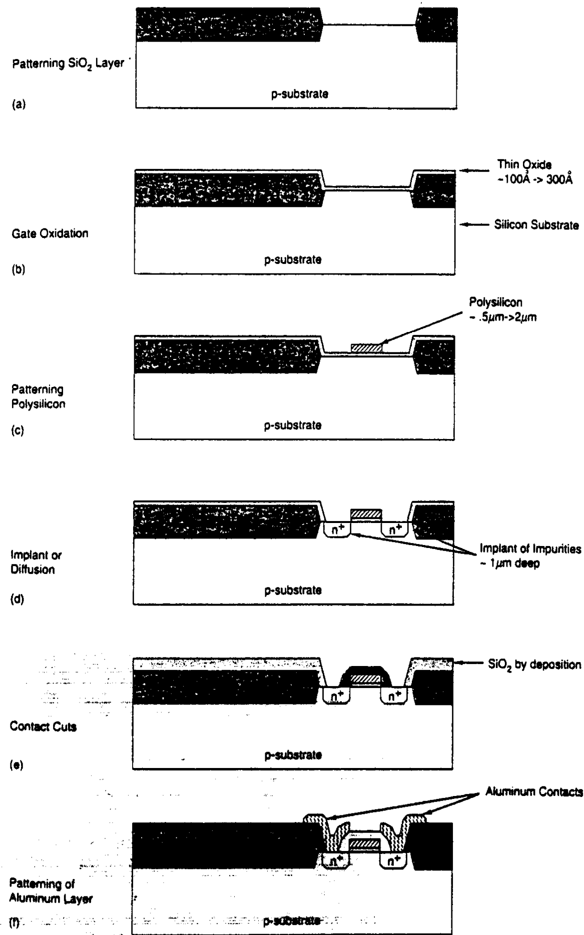

Figure 3.4 Fabrication steps for a silicon gate nMOS transistor

ing of the SiO_2, which was shown in Fig. 3.3. The wafer is covered with SiO_2 with at least two different thicknesses (Fig. 3.4b). A thin, highly controlled layer of SiO_2 is required where active transistors are desired. This is called the gate-oxide or thinox. A thick layer of SiO_2 is required elsewhere to isolate the individual transistors. This is normally called the field oxide. We will examine a variety of methods of achieving these two oxide thicknesses in Section 3.2.1.

Polysilicon is then deposited over the wafer surface and etched to form interconnections and transistor gates. Figure 3.4(c) shows the result of an etched polysilicon gate. The exposed gate oxide (not covered by polysilicon) is then etched away. The complete wafer is then exposed to a dopant source or is ion-implanted, resulting in two actions (Fig. 3.4d). Diffusion junctions are formed in the substrate and the polysilicon is doped with the particular type of dopant. This also reduces the resistivity of the polysilicon. Note that the diffusion junctions form the drain and source of the MOS transistor. They are formed only in regions where the polysilicon gate does not shadow the underlying substrate. This is referred to as a *self-aligned* process because the source and drain do not extend under the gate. Finally, the complete structure is covered with SiO_2 and contact holes are etched to make contact with underlying layers (Fig. 3.4e). Aluminum or other metallic interconnect is evaporated and etched to complete the final connection of elements (Fig. 3.4f). Further oxide layers, contact holes and metallization layers are normally added for extra interconnect.

Note that parasitic MOS transistors exist between unrelated transistors, as shown in Fig. 3.5. Here the source and drain of the parasitic transistor are existing source/drains and the gate is a metal or polysilicon interconnect overlapping the two source/drain regions. The "gate-oxide" is in fact the thick field oxide. The threshold voltage of this transistor is much higher than that of a regular transistor (this device is commonly called a field device) (Eq. 2.1). The high threshold voltage is usually ensured by making the field oxide thick enough and introducing a "channel-stop" diffusion, which raises

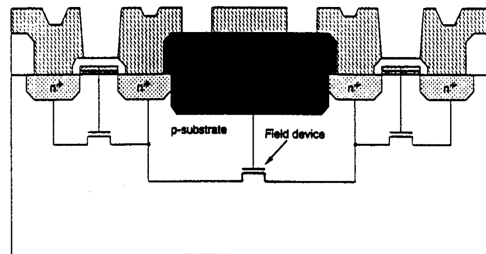

FIGURE 3.5 A parasitic MOS transistor or field device

the impurity concentration in the substrate in areas where transistors are not required, thus further increasing the threshold voltage (Section 2.1.3.1). These devices do have some useful purposes where the fact that they turn on at voltages higher than normal operating voltages may be used to protect other circuitry.

3.2 Basic CMOS Technology

CMOS (Complementary Metal Oxide Silicon) technology is recognized as the leading VLSI systems technology. CMOS provides an inherently low power static circuit technology that has the capability of providing a lower power-delay product than comparable design-rule bipolar, nMOS, or GaAs technologies. In this section we provide an overview of four dominant CMOS technologies, with a simplified treatment of the process steps. This is included primarily as a guide for better appreciation of the layout styles that may be used to implement CMOS gates.

The four main CMOS technologies are:

- n-well process.
- p-well process.
- twin-tub process.
- silicon on insulator.

In addition, by adding bipolar transistors a range of BiCMOS processes are possible.

During the discussion of CMOS technologies, process cross-sections and layouts will be presented. Figure 3.6 summarizes the drawing conventions.

3.2.1 A Basic n-well CMOS Process

A common approach to n-well CMOS fabrication has been to start with a lightly doped p-type substrate (wafer), create the n-type well for the p-channel devices, and build the n-channel transistor in the *native* p-substrate. Although the processing steps are somewhat complex and depend on the fabrication line, Fig. 3.7 illustrates the major steps involved in a typical n-well CMOS process. The mask that is used in each process step is shown in addition to a sample cross-section through an n-device and a p-device. Although we have shown a polysilicon gate process, it is of historical significance to note that CMOS was originally implemented with metal (aluminum) gates. This technology (in p-well form) formed the basis for the majority of low

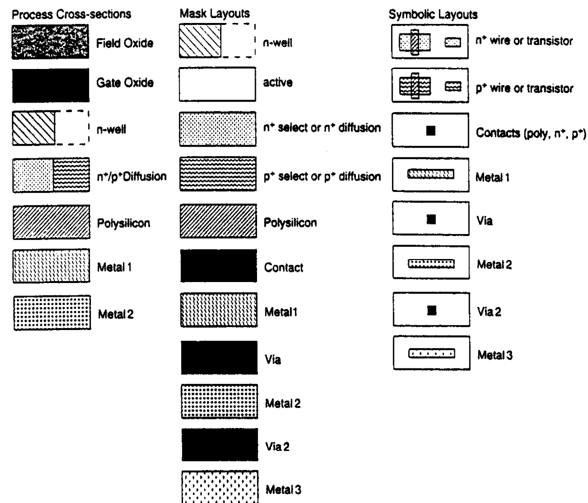

FIGURE 3.6 CMOS process and layout drawing conventions

power CMOS circuits implemented in the 1970s. The technology is robust and still in use. As can be seen from Fig. 3.7, the mask levels are not organized by component function. Rather they reflect the processing steps.

- The first mask defines the n-well (or n-tub); p-channel transistors will be fabricated in this well. Ion implantation or deposition and diffusion is used to produce the n-well (Fig. 3.7a). The former tends to produce shallower wells which are compatible with fine dimension processes. As the diffusion process occurs in all directions, the deeper a diffusion is the more it spreads laterally. This lateral spread affects how near to other structures wells can be placed. Hence, for closely spaced structures a shallow well is required. From a patterned well shape, the final well will extend outside the patterned dimension by the lateral diffusion.
- The next mask is called the "active" mask, because it defines where areas of thin oxide are needed to implement transistor gates and allow implantation to form p- or n-type diffusions for transistor source/drain regions (Fig. 3.7b). Other terms for this mask include *thinox, thinoxide, island,* and *mesa.* A thin layer of SiO_2 is grown and covered with SiN. This is used as a masking layer for the following two steps.

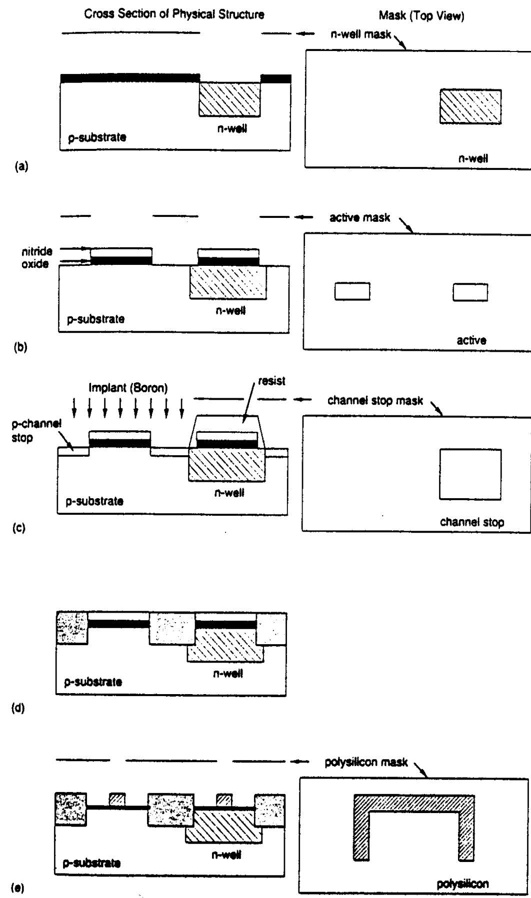

FIGURE 3.7 A typical n-well CMOS process

- The channel-stop implant is usually then completed. This uses the p-well mask (the complement of the n-well mask). It dopes the p-substrate in areas where there are no n-transistors p^+ using a photoresist mask (Fig. 3.7c). This, in conjunction with the thick field oxide that will cover these areas, aids in preventing conduction between unrelated transistor source/drains.

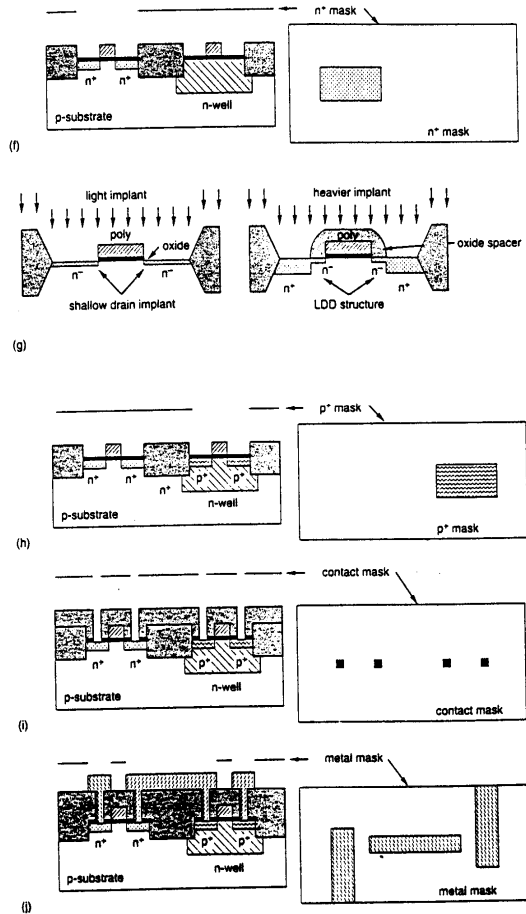

FIGURE 3.7 *(continued)*

- Following the channel-stop implant, the photoresist mask is stripped, leaving the previously masked SiO_2/SiN sandwich defining the active regions. The thick field oxide is then grown. This grows in areas where the SiN layer is absent. The oxide grows in both directions ver-

tically and also laterally under the SiO_2/SiN sandwich (Fig. 3.7d). This lateral movement results in what is called a "bird's beak" because of the shape of the oxide encroachment under the gate oxide mask. This general oxide construction technique is called LOCOS for LOcal Oxidation Of Silicon. The oxide encroachment results in an active area that is smaller than patterned. In particular, the width dimension of a transistor will be reduced from what might be expected from the photolithography. Other techniques such as SWAMI (SideWAll Masked Isolation)[1,2] have been developed to reduce the effect of the bird's beak. Of additional concern is the final planarity of the field oxide/gate oxide interface. If the difference in height is too great, the subsequent conductors may have "step coverage" problems in which a conductor thins and can even break as it crosses a thick to thin oxide boundary. To counter this, many planarization techniques have been developed. One such technique is to pre-etch the silicon in areas where the field oxide is to be grown by around half the final required field oxide thickness. The LOCOS oxide is then grown and the final field oxide/gate oxide interface is very planar.

- An n/p-transistor threshold voltage adjust step might then be performed using a p/n-well photoresist mask. In current fabrication processes the polysilicon is normally doped n^+. With normal doping concentrations suitable for small dimension processes, this results in threshold voltage for n-devices of around 0.5–0.7 volts. However, the p-device threshold is around −1.5 to −2.0 volts. Thus the p-device has to have its threshold voltage adjusted more than the n-device. This is done by introducing an additional negatively charged layer at the silicon/oxide interface. This moves the channel from the silicon/oxide interface further into the silicon, creating a "buried channel" device.[3] Following these two steps the gate oxide is grown.
- Polysilicon gate definition is then completed. This involves covering the surface with polysilicon and then etching the required pattern (in this case an inverted "U"). As noted previously, the "poly" gate regions lead to "self-aligned" source-drain regions (Fig. 3.7e).
- An n-plus (n^+) mask is then used to indicate those thin-oxide areas (and polysilicon) that are to be implanted n^+. Hence a thin-oxide area exposed by the n-plus mask will become an n^+ diffusion area (Fig. 3.7f). If the n-plus area is in the p-substrate, then an n-channel transistor or n-type wire may be constructed. If the n-plus area is in the n-well (not shown), then an *ohmic* contact to the n-well may be constructed. An ohmic contact is one which is only resistive in nature and is not rectifying (as in the case of a diode). In other words, there is no junction (n-type and p-type silicon abutting). Current can flow in both directions in an ohmic contact. This type of mask is sometimes

called the *select* mask because it *selects* those transistor regions that are to be n-type. In modern small dimension processes, to reduce hot carrier effects, considerable effort may go into what is termed "drain engineering."[4] Rather than using one single diffusion or implantation step and mask to produce the source/drain regions, quite complicated structures are constructed. Typical of these structures is the LDD or Lightly Doped Drain structure, which is illustrated in Fig. 3.7(g). This consists of a shallow n-LDD implant that covers the source/drain region where there is no poly (i.e., the normal source/drain region). A spacer oxide is then grown over the polysilicon gate. An n^+ implant is then used to produce n^+ implants that are spaced from the edge of the original poly gate edges. The spacer is then removed, resulting in a structure that is more resistant to hot-electron effects. Current 0.25μm processes revert to a simpler self-aligned structure presumably because of the complexity of the LDD structure.

- The next step usually uses the complement of the n-plus mask, although an extra mask is normally not needed. The "absence" of an n-plus region over a thin-oxide area indicates that the area will be a p^+ diffusion or p-active. P-active in the n-well defines possible p-transistors and wires (Fig. 3.7h). A p^+ diffusion in the p-substrate allows an ohmic contact to be made. Following this step, the surface of the chip is covered with a layer of SiO_2. The LDD step is not necessarily done for p-transistors because their hot-carrier susceptibility is much less than that of n-transistors. For this reason, the drawn length dimension of p-transistors might be larger than that of the n-transistors.
- Contact cuts are then defined. This involves etching any SiO_2 down to the surface to be contacted (Fig. 3.7i). These allow metal (next step) to contact diffusion regions or polysilicon regions.
- Metallization is then applied to the surface and selectively etched (Fig. 3.7j) to produce circuit interconnections.
- As a final step (not shown), the wafer is passivated and openings to the bond pads are etched to allow for wire bonding. Passivation protects the silicon surface against the ingress of contaminants that can modify circuit behavior in deleterious ways.

The cross-section of the finished n-well process is shown in Fig. 3.8(c). The layout of the n-well CMOS transistors corresponding to this cross-section is illustrated in Fig. 3.8(b). The corresponding schematic (for an inverter) is shown in Fig. 3.8(a). From Fig. 3.8 it is evident that the p-type substrate accommodates n-channel devices, whereas the n-well accommodates p-channel devices. (Figure 3.8 also appears in color as Plate 1.)

In an n-well process, the p-type substrate is normally connected to the negative supply (V_{SS}) through what are termed V_{SS} substrate contacts, while

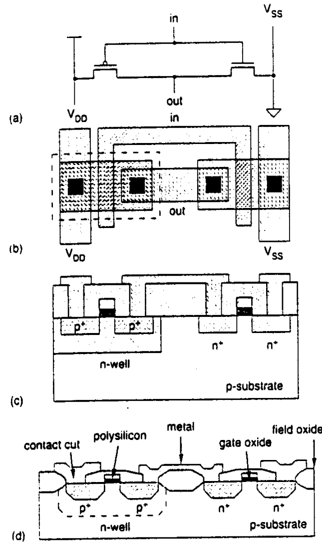

FIGURE 3.8 Cross section of a CMOS inverter in an n-well process

the well has to be connected to the positive supply (V_{DD}) through V_{DD} substrate (or well or tub) contacts. As the substrate is accessible at the top of the wafer and the bottom, connecting the substrate may be accomplished from the backside of the wafer. Topside connection is preferred because it reduces parasitic resistances that could cause latchup (see later). Substrate connections that are formed by placing n^+ regions in the n-well (V_{DD} contacts) and p^+ in the p-type substrate (V_{SS} contacts) are illustrated by Fig. 3.9(a). The corresponding layout is shown in Fig. 3.9(b). Other terminology for these contacts include "well contacts," "body ties," or "tub ties" for the V_{DD} substrate connection. We will use the term "substrate contact" for both V_{SS} and V_{DD} contacts, because this terminology can be commonly used for most bulk CMOS processes. It should be noted that these contacts are formed during the implants used for the p-channel and n-channel transistor formation.

3.2.2. The p-well Process

N-well processes have emerged in popularity in recent years. Prior to this, p-well processes were one of the most commonly available forms of CMOS. Typical p-well fabrication steps are similar to an n-well process, except that a p-well is implanted rather than an n-well. The first masking step defines the p-well regions. This is followed by a low-dose boron implant driven in by a

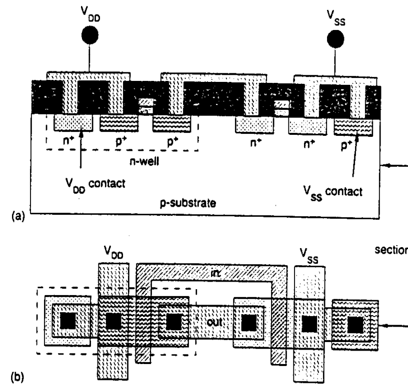

FIGURE 3.9 Substrate and well contacts in an n-well process

high-temperature step for the formation of the p-well. The well depth is optimized to ensure against n-substrate to n^+ diffusion breakdown, without compromising p-well to p^+ separation. The next steps are to define the devices and other diffusions; to grow field oxide; contact cuts; and metallization. A p-well mask is used to define p-well regions, as opposed to an n-well mask in an n-well process. A p-plus (p^+) mask may be used to define the p-channel transistors and V_{SS} contacts. Alternatively, we could use an n-plus mask to define the n-channel transistors, because the masks usually are the complement of each other.

P-well processes are preferred in circumstances where the characteristics of the n- and p-transistors are required to be more balanced than that achievable in an n-well process. Because the transistor that resides in the native substrate tends to have better characteristics, the p-well process has better p devices than an n-well process. Because p-devices inherently have lower gain than n devices, the n-well process exacerbates this difference while a p-well process moderates the difference.

3.2.3 Twin-Tub Processes

Twin-tub CMOS technology provides the basis for separate optimization of the p-type and n-type transistors, thus making it possible for threshold voltage, body effect, and the gain associated with n- and p-devices to be independently optimized.[5,6] Generally, the starting material is either an n^+ or p^+ substrate with a lightly doped *epitaxial* or *epi* layer, which is used for protection against latchup (see Section 3.5). The aim of *epitaxy* (which means "arranged upon") is to grow high-purity silicon layers of controlled thick-

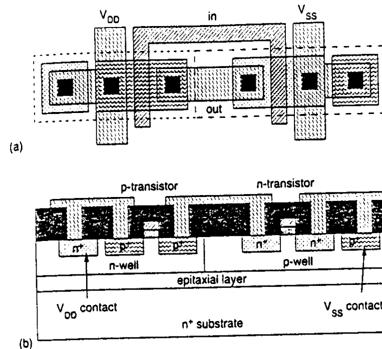

FIGURE 3.10 Twin-well CMOS process cross section

ness with accurately determined dopant concentrations distributed homogeneously throughout the layer. The electrical properties of this layer are determined by the dopant and its concentration in the silicon. The process sequence, which is similar to the n-well process apart from the tub formation where both p-well and n-well are utilized, entails the following steps:

- Tub formation.
- Thin-oxide construction.
- Source and drain implantations.
- Contact cut definition.
- Metallization.

Since this process provides separately optimized wells, balanced performance n-transistors and p-transistors may be constructed. Note that the use of threshold adjust steps is included in this process. These masks are derived from the active and n-plus masks. The cross-section of a typical twin-tub structure is shown in Fig. 3.10. The substrate contacts (both of which are required) are also included.

3.2.4 Silicon On Insulator

Rather than using silicon as the substrate, technologists have sought to use an insulating substrate to improve process characteristics such as latchup and speed. Hence the emergence of Silicon On Insulator (SOI) technologies. SOI CMOS processes have several potential advantages over the traditional CMOS technologies.[7] These include closer packing of p- and n-transistors, absence of latchup problems, and lower parasitic substrate capacitances. In