APPENDIX A

# VLSI FABRICATION TECHNOLOGY

## Introduction

Since the first edition of this text, we have witnessed a fantastic evolution in **VLSI** (very-large-scale integrated circuits) technology. In the late 1970s, non-self-aligned metal gate MOSFETs with gate lengths in the order of 10 μm were the norm. Current VLSI fabrication technology is at gate lengths below 10 nm. This represents a 1000x reduction in device size, along with an even more impressive increase in the number of devices per VLSI chip. Ongoing development in VLSI technology relies upon new device concepts and new materials, taking quantum effects into account. While this is a very exciting time for researchers to explore new technology, we can also be assured that “traditional” **CMOS** and **BiCMOS** (bipolar CMOS) fabrication technologies will continue to be the workhorses of the microelectronic industry for many more years to come.

The purpose of this appendix is to familiarize the reader with VLSI fabrication technology. Brief explanations of standard VLSI processing steps are given. The variety of devices available in CMOS and BiCMOS fabrication technologies are also presented. In particular, the differences between components in the **IC** (integrated circuit) environment and those available for discrete circuit design will be discussed. In order to enjoy the economics of integrated circuits, designers have to overcome some serious device limitations (such as poor device tolerances) but may benefit from certain advantages (such as good component matching). An understanding of device characteristics is therefore essential in designing high-performance custom VLSIs.

This appendix will consider only silicon-based (Si) technologies. Although other compound semiconductors combining materials in groups III through V, such as gallium arsenide (GaAs) and aluminum gallium nitride (AlGaN), are also used to implement ICs, silicon is still the most popular material, with excellent cost–performance trade-off. The development of SiGe and strained-silicon technologies has further strengthened the position of Si-based fabrication processes in the microelectronic industry for many more years to come.

Silicon is an abundant element and occurs naturally in the form of sand. It can be refined using well-established purification and crystal growth techniques. It also exhibits suitable physical properties for fabricating active devices with good electrical characteristics. In addition, silicon can be easily oxidized to form an excellent insulator, $SiO_2$ (glass). This native oxide is useful for constructing capacitors and MOSFETs. It also serves as a diffusion barrier that can mask against unwanted impurities from diffusing into the high-purity silicon material. This masking property allows the electrical properties of the silicon to be altered in predefined areas. Therefore, active and passive elements can be built on the same piece of material (substrate). The components can then be interconnected using metal layers (similar to those used in printed-circuit boards) to form a monolithic IC.

# A.1 IC Fabrication Steps

The basic IC fabrication steps will be described in the following sections. Some of these steps may be carried out many times, in different combinations and/or processing conditions during a single complete fabrication run.

## A.1.1 Silicon Wafers

The starting material for modern integrated circuits is very-high-purity, single-crystal silicon. The material is initially grown as a single crystal ingot. It takes the shape of a steel-gray solid cylinder 10 cm to 30 cm in diameter and can be one to two meters in length. This crystal is then sawed (like a loaf of bread) to produce circular **wafers** that are 400 μm to 600 μm thick (a micrometer, or micron, μm, is a millionth of a meter). The surface of the wafer is then polished to a mirror finish using chemical and mechanical polishing (CMP) techniques. Semiconductor manufacturers usually purchase ready-made silicon wafers from a supplier and rarely start their fabrication process in ingot form.

Many basic electrical and mechanical properties of the wafer depend on the orientation of the crystalline structure, the impurity concentrations, and the type of impurities present. These variables are strictly controlled during crystal growth. A specific concentration of impurities can be added to the pure silicon in a process known as doping. This alters the electrical properties of the silicon, in particular its resistivity. Depending on the types of impurity, either holes (in ***p*-type** silicon) or electrons (in ***n*-type** silicon) can be responsible for electrical conduction. If a large number of impurity atoms is added, the silicon is said to be heavily doped (e.g., concentration $\gtrsim 10^{18}$ atoms/cm$^{-3}$). The relatively high concentration of free carries results in a correspondingly low resistivity. When designating the relative doping concentrations in semiconductor material, it is common to use the + and – symbols. A heavily doped (low-resistivity) $n$-type silicon wafer is referred to as $n+$ material, while a lightly doped material (e.g., concentration $< \sim 10^{16}$ atoms/cm$^{-3}$) is referred to as $n-$. Similarly, $p+$ and $p-$ designations refer to the heavily doped and lightly doped $p$-type regions, respectively. The ability to control the type of impurities and the doping concentration in the silicon permits the formation of diodes, transistors, and resistors in integrated circuits.

## A.1.2 Oxidation

In **oxidation**, silicon reacts with oxygen to form silicon dioxide ($SiO_2$). To speed up this chemical reaction, it is necessary to carry out the oxidation at high temperatures (e.g., 1000–1200°C) and inside ultraclean furnaces. To avoid the introduction of even small quantities of contaminants (which could significantly alter the electrical properties of the silicon), it is necessary to operate in a **clean room** . Particle filters are used to ensure that the airflow in the processing area is free from dust. All personnel must protect the clean-room environment by wearing special lint-free clothing that covers a person from head to toe.

The oxygen used in the reaction can be introduced either as a high-purity gas (referred to as a "**dry oxidation**") or as steam (forming a "**wet oxidation**"). In general, wet oxidation has a faster growth rate, but dry oxidation gives better electrical characteristics. The thermally grown oxide layer has excellent electrical insulation properties. The dielectric strength for $\mathbf{SiO_2}$ is approximately $10^7$ V/cm. It has a dielectric constant of about 3.9, and it can be used to form excellent MOS capacitors. Silicon dioxide can also serve as an effective mask against many impurities, allowing the introduction of dopants into the silicon only in regions that are not covered with oxide.

**IRIDESCENT WATERS**

Silicon dioxide is a transparent film, and the silicon surface is highly reflective. If white light is shone on an oxidized wafer, constructive and destructive interference will cause certain colors to be reflected. The wavelengths of the reflected light depend on the thickness of the oxide layer. In fact, by categorizing the color of the wafer surface, one can deduce the thickness of the oxide layer. The same principle is used by more sophisticated optical inferometers to measure film thickness. On a processed wafer, there will be regions with different oxide thicknesses. The colors can be quite vivid and are immediately obvious when a finished wafer is viewed with the naked eye.

## A.1.3 Photolithography

Mass production with economy of scale is the primary reason for the tremendous impact VLSI has had on our society. The surface patterns of the various integrated-circuit components can be defined repeatedly using photolithography. The sequence of photolithographic steps is as illustrated in Fig. A.1.

The wafer surface is coated with a photosensitive layer called photoresist, using a spin-on technique. After this, a photographic plate with drawn patterns (e.g., a quartz plate with chromium layer for patterning) will be used to selectively expose the photoresist to deep ultraviolet illumination (UV). The exposed areas become either softened (for positive photoresist), or hardened (for negative photoresist). The exposed or unexposed regions are then removed using a chemical developer, causing the mask pattern to be duplicated on the wafer. Very fine surface geometries can be reproduced accurately by this technique. Furthermore, the patterns can be projected directly onto the wafer, or by using a separate photomask produced by a 10x "step and repeat" reduction technique as shown in Fig. A.2.

The patterned photoresist layer can be used as an effective masking layer to protect materials below from wet chemical **etching** or **reactive ion etching** (RIE). Silicon dioxide, silicon nitride, polysilicon, and metal layers can be selectively removed using the appropriate

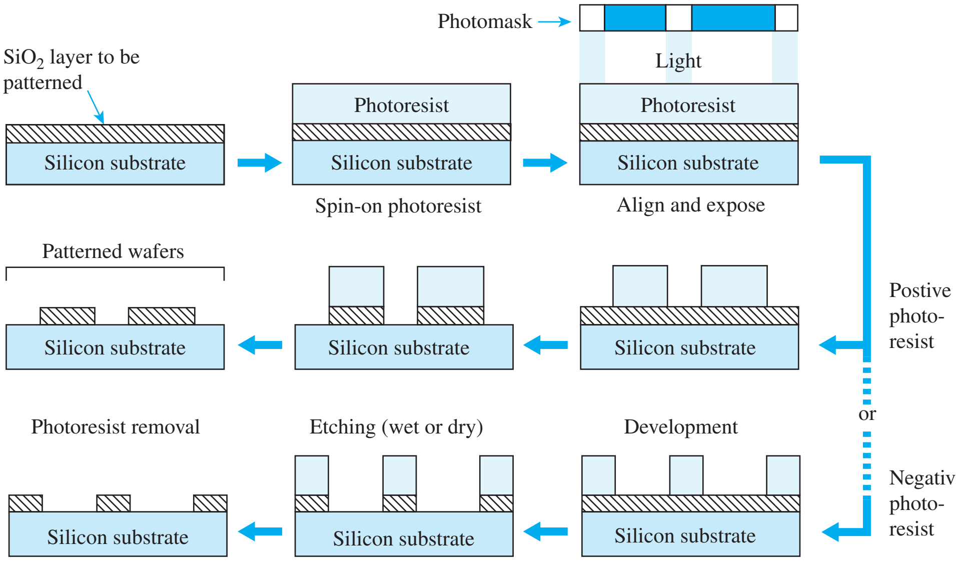

**Figure A.1** Photolithography using positive or negative photoresist.

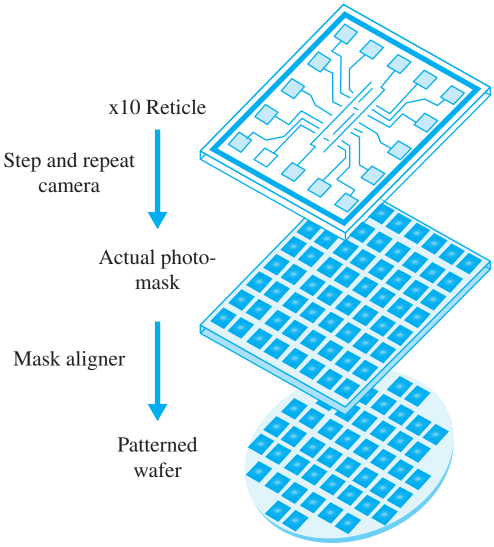

Figure A.2 Conceptual illustration of a step-and-repeat reduction technique to facilitate the mass production of integrated circuits.

etching methods (see next section). After the etching step(s), the photoresist is stripped away, leaving behind a permanent pattern of the photomask on the wafer surface.

To make this process even more challenging, multiple masking layers (which can number more than 20 in advanced VLSI fabrication processes) must be aligned precisely on top of previously etched patterns. This must be done with even finer precision than the minimum geometry size of the masking patterns. This requirement imposes very critical mechanical and optical constraints on the photolithography equipment.

## A.1.4 Etching

To permanently imprint the photographic patterns onto the wafer, chemical (**wet**) **etching** or RIE **dry etching** procedures can be used. Different chemical solutions can be used to remove different layers. For example, hydrofluoric (HF) acid can be used to etch $SiO_2$, potassium hydroxide (KOH) for silicon, phosphoric acid for aluminum, and so on. In wet etching, the chemical usually attacks the exposed regions that are not protected by the photoresist layer in all directions (**isotropic etching**). Depending on the thickness of the layer to be etched, a certain amount of undercut will occur whereby some material under the edges of the photoresist are removed. Therefore, the dimension of the actual pattern will differ slightly from the original pattern. If exact dimensions are critical, RIE **dry etching** can be used. This method is essentially a directional bombardment of the exposed surface using a corrosive gas (or ions). The cross section of the etched layer is usually highly directional (**anisotropic etching**) and has the same dimension as the photoresist pattern. A comparison between isotropic and anisotropic etching is given in Fig. A.3.

## A.1.5 Diffusion

**Diffusion** is a process by which atoms move from a high-concentration region to a low-concentration region. This is very much like a drop of ink dispersing through a glass of water except that it occurs much more slowly in solids. In VLSI fabrication, this is a

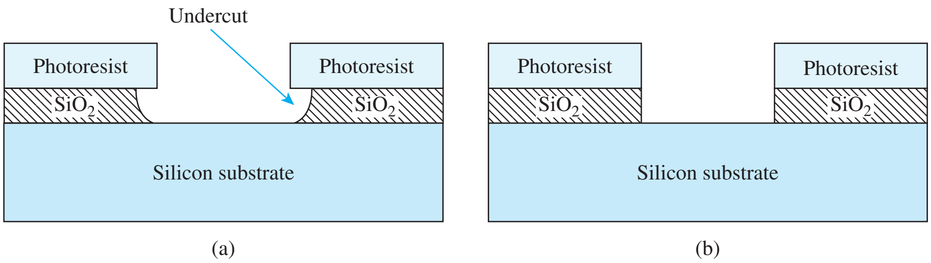

**Figure A.3** (**a**) Cross-sectional view of an isotropic oxide etch with severe undercut beneath the photoresist layer. (**b**) Anisotropic etching, which usually produces a cross section with no undercut.

method to introduce impurity atoms (dopants) into silicon, creating p- and n-type regions, diodes, transistors, and other devices. The rate at which dopants diffuse in silicon is a strong function of temperature. Diffusion of impurities is usually carried out at high temperatures (1000–1200°C) to obtain the desired doping profile. When the wafer is cooled to room temperature, the impurities are essentially "frozen" in position. The diffusion process is performed in furnaces similar to those used for oxidation. The depth to which the impurities diffuse depends on both the temperature and the processing time.

The most common impurities used as **dopants** are boron, phosphorus, and arsenic. Boron is a *p*-type dopant, while phosphorus and arsenic are *n*-type dopants. These dopants can be effectively masked by thin silicon dioxide layers. Heavy dopant concentrations can overwhelm previously introduced light dopant concentrations of the opposite type. For example, by diffusing boron into an *n*-type substrate, a *pn* junction (diode) is formed. If the doping concentration is heavy, the diffused layer can also be used as a conducting layer with very low resistivity.

## A.1.6 Ion Implantation

**Ion implantation** is another method used to introduce impurities into the semiconductor crystal. An ion implanter produces ions of the desired dopant, accelerates them by an electric field, and allows them to strike the semiconductor surface. The ions become embedded in the crystal lattice. The depth of penetration is related to the energy of the ion beam, which can be controlled by the accelerating-field voltage. The quantity of ions implanted can be controlled by varying the beam current (flow of ions). Since both voltage and current can be accurately measured and controlled, ion implantation results in impurity profiles that are much more accurate and reproducible than can be obtained by diffusion. In addition, ion implantation can be performed at room temperature. Ion implantation normally is used when accurate control of the doping profile is essential for device operation.

## A.1.7 Chemical Vapor Deposition

**Chemical vapor deposition** (CVD) is a process by which gases or vapors are chemically reacted, leading to the formation of solids on a substrate. CVD can be used to deposit various materials on a silicon substrate including $SiO_2$, $Si_3N_4$, polysilicon, and so on. For instance, if silane gas and oxygen are allowed to react above a silicon substrate, the end product, silicon dioxide, will be deposited as a solid film on the silicon wafer surface. The properties of the CVD oxide layer are not as good as those of a thermally grown oxide, but they are sufficient to allow the layer to act as an electrical insulator. The advantage of a CVD layer is that the oxide deposits at a faster rate and a lower temperature (below 500°C).

If silane gas alone is used, then a silicon layer will be deposited on the wafer. If the reaction temperature is high enough (above 1000°C), the layer deposited will be a crystalline layer (assuming that there is an exposed crystalline silicon substrate). Such a layer is called an **epitaxial** layer, and the deposition process is referred to as **epitaxy** instead of CVD. At lower temperatures, or if the substrate surface is not single-crystal silicon, the atoms will not be able to align along the same crystal orientation. Such a layer is called polycrystalline silicon (**poly Si**), since it consists of many small crystals of silicon aligned in random fashion. Polysilicon layers are normally doped very heavily to form highly conductive regions that can be used for electrical interconnections and MOSFET gates.

### A.1.8 Metallization

Metallization serves as wires to interconnect the various components (transistors, capacitors, etc.) that form the desired integrated circuit. Metallization involves the deposition of a metal over the entire surface of the silicon. The required interconnection pattern is then selectively etched. The metal layer is normally deposited via a sputtering process. A pure metal disk (e.g., 99.99% aluminum target) is placed under an Ar (argon) ion gun inside a vacuum chamber. The wafers are also mounted inside the chamber above the target. The Ar ions will not react with the metal, since argon is a noble gas. However, the ions are made to physically bombard the target and literally knock metal atoms out of the target. These metal atoms will then coat all the surface inside the chamber, including the wafers. The thickness of the metal film can be controlled by the length of the sputtering time, which is normally in the range of 1 to 2 minutes. The metal interconnects can then be defined using photolithography and etching steps. Contacts are needed to convey current between the semiconductor and the metal interconnect above. These are patterned prior to the metal by an additional mask step that creates openings in the $SiO_2$ layer. A conductive material such as tungsten is sputtered into the openings, providing a contact between the semiconductor and the metal layer that follows. Note that a high dopant concentration (either n- or p-type) is required under the contact to ensure low resistance.

### A.1.9 Packaging

A finished silicon wafer may contain several hundreds or thousands of finished circuits or chips. A chip may contain from 10 to more than $10^9$ transistors; each chip is rectangular and can be up to tens of millimeters on a side. The circuits are first tested electrically (while still in wafer form) using an automatic probing station. Bad circuits are marked for later identification. The circuits are then separated from each other (by a process called dicing), and the good circuits (dies) are mounted in packages (headers). Examples of such IC packages are given in Fig. A.4. Fine gold wires and/or balls of solder are normally used to interconnect the pins of the package to the metallization pattern on the die. Finally, the package is sealed using plastic or epoxy under vacuum or in an inert atmosphere.

## A.2 VLSI Processes

Integrated-circuit fabrication technology was originally dominated by bipolar technology. By the late 1970s, metal oxide semiconductor (MOS) technology became more promising for VLSI implementation with higher packing density and lower power consumption. Since the early 1980s, complementary MOS (CMOS) technology has almost completely dominated

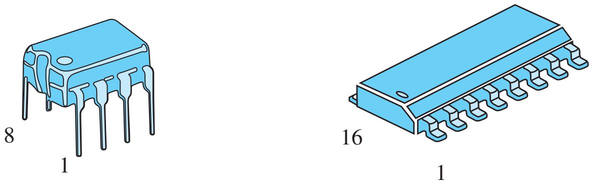

**Figure A.4** Examples of an 8-pin plastic dual-in-line IC package and a 16-pin surface-mount package.

the VLSI scene, leaving bipolar technology to fill specialized functions such as high-speed analog and RF circuits. CMOS technologies continue to evolve, and in the late 1980s, the incorporation of bipolar devices led to the emergence of high-performance bipolar-CMOS (BiCMOS) fabrication processes that provided the best of both technologies. However, high-quality bipolar devices offer no benefit for the digital logic that dominates most large integrated circuits, and they require additional processing steps during fabrication, hence additional cost.

The performance of CMOS and BiCMOS processes continues to improve with finer lithography resolution. However, fundamental limitations on processing techniques and semiconductor properties have prompted the need to explore alternative materials. Newly emerged SiGe and strained-Si technologies are good compromises to improve performance while maintaining manufacturing compatibility (hence low cost) with existing silicon-based CMOS fabrication equipment.

In the subsection that follows, we will examine a typical CMOS process flow, the performance of the available components, and the inclusion of bipolar devices to form a BiCMOS process.

## A.2.1 Twin-Well CMOS Process

Depending on the choice of starting material (substrate), CMOS processes can be identified as ***n*-well**, ***p*-well**, or **twin-well** processes. The latter is the most complicated but most flexible in the optimization of both the *n*- and *p*-channel MOSFETs. In addition, many advanced CMOS processes may make use of trench isolation and silicon-on-insulator (SOI) technology to reduce parasitic capacitance (hence increase speed) and to improve packing density.

A modern twin-well CMOS process flow is shown in Fig. A.5. An exemplar process with 10 masking layers is described here. In practice, most CMOS processes will require many additional layers such as *n*- and *p*-guards for better latchup immunity, and up to 10 or more layers of metalization for high-density interconnections. The inclusion of these layers would increase the total number of masking layers to 15–20 or more.

The starting material for the twin-well CMOS process is a *p*-type substrate. The process begins with the formation of the *p*-well and the *n*-well (Fig. A.5a). The *n*-well is required wherever *p*-channel MOSFETs are to be placed, while the *p*-well is used to house the *n*-channel MOSFETs. The well-formation procedures are similar. A thick photoresist layer is etched to expose the regions for *n*-well diffusion. The unexposed regions will be protected from the *n*-type phosphorus impurity. Phosphorus implantation is usually used for deep diffusions, since it has a large diffusion coefficient and can diffuse faster than arsenic into the substrate.

The second step is to define the active regions where transistors are to be placed using a technique called **shallow trench isolation** (STI). To reduce the chance of unwanted

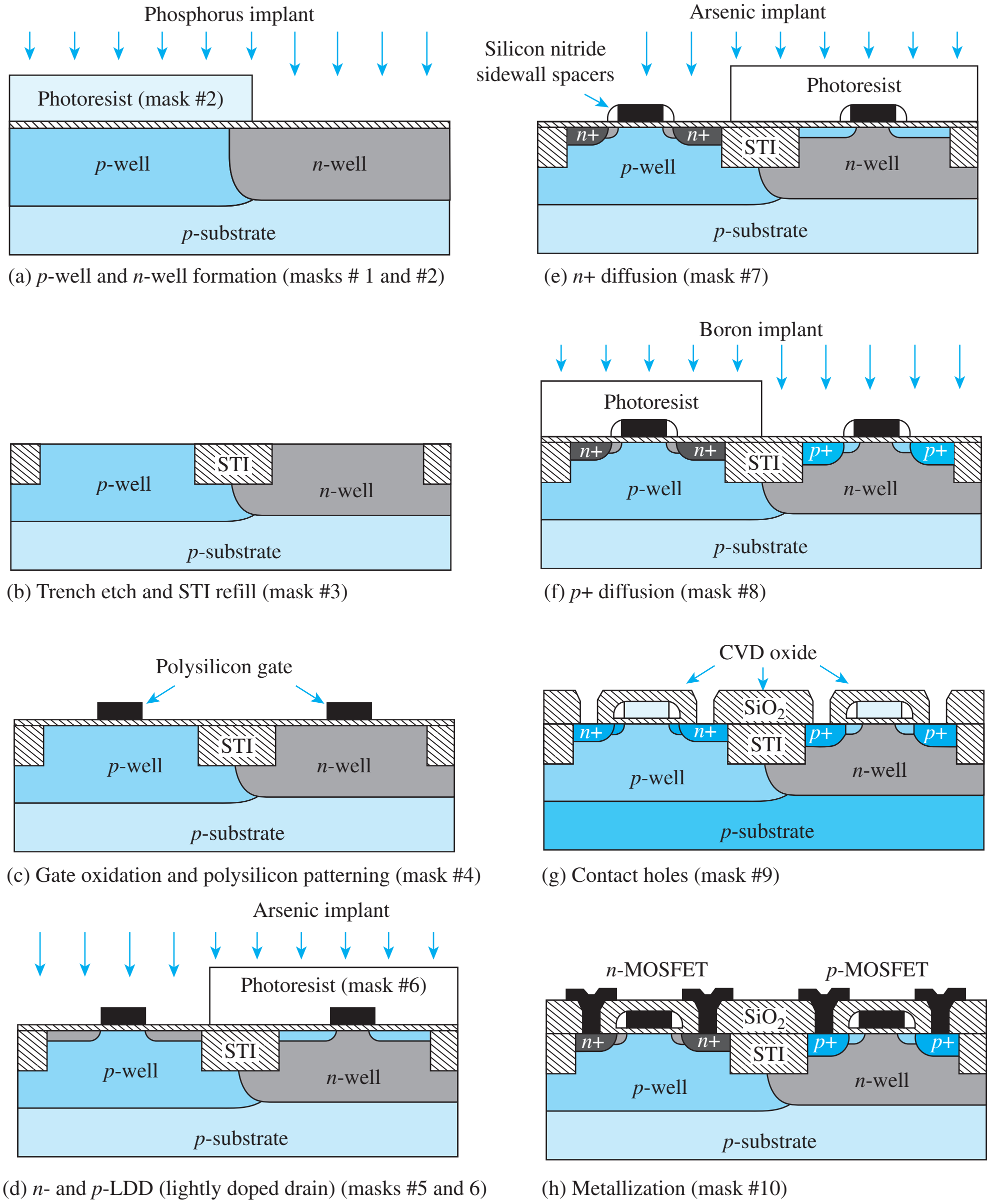

**Figure A.5** A modern twin-well CMOS process flow with shallow trench isolation (STI).

latchup (a serious issue in CMOS technology), dry etching is used to produce trenches approximately 0.3 μm deep on the silicon surface. These trenches are then refilled using CVD oxide, followed by a planarization procedure to ensure a flat surface topology (Fig. A.5b). An alternate isolation technique is called **local oxidation of silicon** (LOCOS). This older technology uses silicon nitride ($Si_3N_4$) patterns to protect selective regions of the wafer surface from oxidization. After a long wet-oxidation step, thick field oxide will appear in exposed regions between transistors. This produces an effect similar to that obtained in the STI process, but the isolation oxide occupies more area.

The next step is the formation of the polysilicon gate (Fig. A.5c). This is one of the most critical steps in the CMOS process. The thin oxide layer in the active region is first removed using wet etching followed by the growth of a high-quality thin gate oxide. Current deep-submicron CMOS processes routinely make used of oxide thicknesses as thin as 20 Å to 50 Å (1 angstrom = $10^{-8}$ cm) or even less. A polysilicon layer, usually arsenic doped (*n*-type), is then deposited and patterned. Some of the polysilicon traces formed in this step will subsequently serve as masks to define n- or p-type regions on either side, thus creating n- or p-type MOSFETs. In this way, the source and drain are automatically aligned to the polysilicon gate. The development of this self-aligned process allowed for much smaller more reliable MOSEFTs than was otherwise possible. The photolithography is most demanding in this step since the finest resolution is required to produce the shortest possible MOS channel length.

The formation of **lightly doped drain** (LDD) regions for MOSFETs of both types follows. Light doping prevents the generation of **hot electrons** that might affect the reliability of the transistors. A noncritical mask, together with the polysilicon gates, is used to form the self-aligned LDD regions (Fig. A.5d). The resistivity of the lightly doped regions is too high, so higher concentrations are next introduced throughout much of the source and drain regions.

Prior to the $n+$ and $p+$ drain region implant, a sidewall spacer step is performed. A thick layer of silicon nitride is deposited uniformly on the wafer. Due to the conformal nature of the deposition, the thickness of the silicon nitride layer at all layer edges (i.e., at both ends of the polysilicon gate electrode) will be thicker than those deposited over a flat surface. After a timed RIE dry etch to remove all the silicon nitride layer, pockets of silicon nitride will remain at the edge of the polysilicon gate electrode (Fig. A.5e). Such pockets of silicon nitride are called sidewall spacers. They are used to block subsequent $n+$ or $p+$ source/drain implants, protecting the LDD regions.

A heavy arsenic implant can be used to form the $n+$ source and drain regions of the *n*-MOSFETs. The polysilicon gate also acts as a barrier for this implant to protect the channel region. A layer of photoresist can be used to block the regions where *p*-MOSFETs are to be formed (Fig. A.5e). The thick field oxide stops the implant and prevents $n+$ regions from forming outside the active regions. A reversed photolithography step can be used to protect the *n*-MOSFETs during the $p+$ boron source and drain implant for the *p*-MOSFETs (Fig. A.5f). Note that in both cases the separation between the source and drain diffusions—channel length—is defined by the polysilicon gate mask alone, hence the self-aligned property.

Before contact holes are opened, a thick layer of CVD oxide is deposited over the entire wafer. A photomask is used to define the contact window opening (Fig. A.5g), followed by a wet or dry oxide etch. A thin conductive layer is then evaporated or sputtered onto the wafer. A final masking and etching step is used to pattern the interconnection (Fig. A.5h).

Not shown in the process flow is the final passivation step prior to packaging and wire bonding. A thick CVD oxide or pyrox glass is usually deposited on the wafer to serve as a protective layer.

## A.2.2 Integrated Devices

Besides the obvious *n*- and *p*-channel MOSFETs, other devices can be obtained by appropriate masking patterns. These include *pn* junction diodes, MOS capacitors, and resistors.

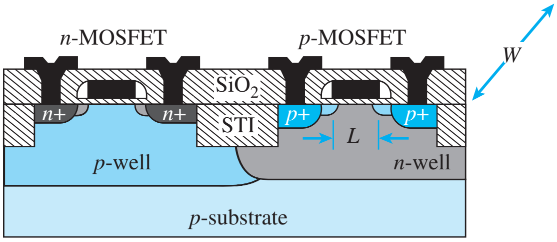

**Figure A.6** Cross-sectional diagram of $n$- and $p$-MOSFETs.

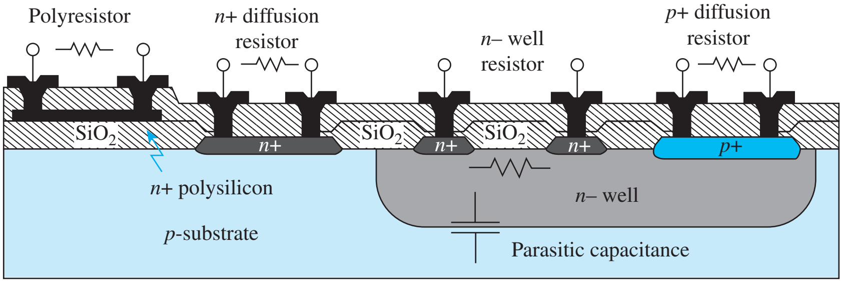

**Figure A.7** Cross sections of various resistor types available from a typical $n$-well CMOS process.

## A.2.3 MOSFETs

The $n$-channel MOSFET is often preferred in comparison to the $p$-MOSFET (Fig. A.6) because electron surface mobility is two to three times higher than that for holes. Therefore, with the same device size ($W$ and $L$), the $n$-MOSFET offers higher current drive (or lower on- resistance) and higher transconductance

In an integrated-circuit design environment, MOSFETs are characterized by their threshold voltage and by their device sizes. Usually the $n$- and $p$-channel MOSFETs are designed to have threshold voltages of similar magnitude for a particular process. The transconductance can be adjusted by changing the device surface dimensions ($W$ and $L$). This feature is not available for bipolar transistor, making the design of integrated MOSFET circuits much more flexible.

## A.2.4 Resistors

Resistors in integrated form are not very precise. They can be made from various diffusion regions as shown in Fig. A.7. Different diffusion regions have different resistivity. The $n$ well is usually used for medium-value resistors, while the $n+$ and $p+$ diffusions are useful for low-value resistors. The actual resistance value can be defined by changing the length and width of diffused regions. The tolerance of the resistor value is very poor (20–50%), but the matching of two similar resistor values is quite good (5%). Thus circuit designers should design circuits that exploit resistor matching and should avoid designs that require a specific resistor value.

All diffused resistors are self-isolated by the reverse-biased $pn$ junctions. A serious drawback for these resistors is the fact that they are accompanied by a substantial parasitic junction capacitance, making them not very useful for high-frequency applications. The

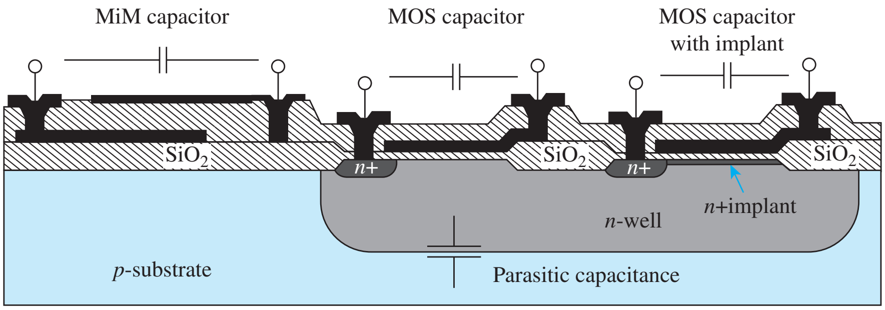

**Figure A.8** MIM and MOS capacitors in an $n$-well CMOS process.

reverse-biased *pn* junctions also exhibit a JFET effect, leading to a variation in the resistance value as the supply voltage is changed (a large voltage coefficient is undesirable). Since the mobilities of carriers vary with temperature, diffused resistors also exhibit a significant temperature coefficient.

A more useful resistor can be fabricated using the polysilicon layer that is placed on top of the thick field oxide. The thin polysilicon layer provides better surface area matching and hence more accurate resistor ratios. Furthermore, the polyresistor is physically separated from the substrate, resulting in a much lower parasitic capacitance and voltage coefficient.

## A.2.5 Capacitors

Two types of capacitor structure are available in CMOS processes: MOS and metal–insulator–metal (MiM) capacitors. The cross sections of these structures are as shown in Fig. A.8. The MOS gate capacitance, depicted by the center structure, is basically the gate-to-source capacitance of a MOSFET. The capacitance value is dependent on the gate area. The oxide thickness is the same as the gate oxide thickness in the MOSFETs. This capacitor exhibits a large voltage dependence. To eliminate this problem, an addition $n$+ implant is required to form the bottom plate of the capacitors, as shown in the structure on the right. Both these MOS capacitors are physically in contact with the substrate, resulting in a large parasitic *pn* junction capacitance between the bottom plate and substrate.

The MiM capacitor exhibits near-ideal characteristics but at the expense of less capacitance per unit area. This downside is ameliorated in advanced CMOS processes where as many as 10 metal layers can be alternated to realize more capacitance. Since this capacitor is placed on top of the thick field oxide, parasitic effects are kept to a minimum.

A third and less often used capacitor is the junction capacitor. Any *pn* junction under reversed bias produces a depletion region that acts as a dielectric between the $p$ and the $n$ regions. The capacitance is determined by geometry and doping levels and has a large voltage coefficient. This type of capacitor is often used as a varactor (variable capacitor) for tuning circuits. However, this capacitor works only with reverse-bias voltages.

For the MiM and MOS capacitors, the capacitance values can be controlled to within 5%. Practical capacitance values range from 10 fF to a few tens of picofarads. The matching between capacitors of similar size can be within 0.1%. This property is extremely useful for designing precision analog CMOS circuits.

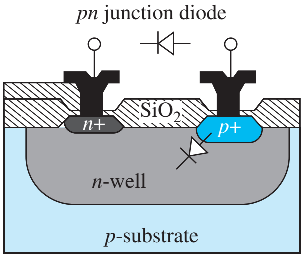

**Figure A.9** A *pn* junction diode in an *n*-well CMOS process.

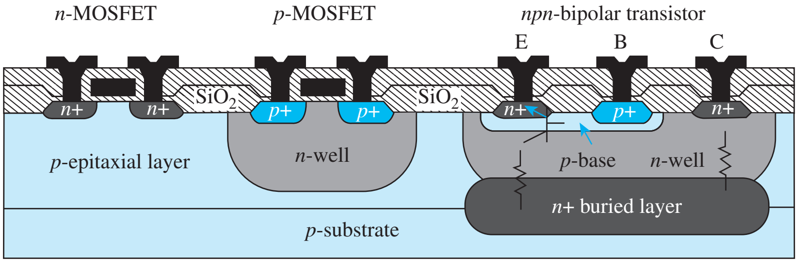

**Figure A.10** Cross-sectional diagram of a BiCMOS process.

## A.2.6 *pn* Junction Diodes

Whenever *n*-type and *p*-type diffusion regions are placed next to each other, a *pn* junction diode results. A useful structure is the *n*-well diode shown in Fig. A.9. The diode fabricated in an *n* well can provide a high breakdown voltage. This diode is essential for the input clamping circuits that protect against electrostatic discharge. The diode is also very useful as an on-chip temperature sensor by monitoring the variation of its forward voltage drop.

## A.2.7 BiCMOS Process

An *npn* vertical bipolar transistor can be integrated into the *n*-well CMOS process with the addition of a *p*-base diffusion region (Fig. A.10). The characteristics of this device depend on the base width and the emitter area. The base width is determined by the difference in junction depth between the *n*+ and the *p*-base diffusions. The emitter area is determined by the junction area of the *n*+ diffusion at the emitter. The *n*-well serves as the collector for the *npn* transistor. Typically, the *npn* transistor has a $\beta$ in the range of 50 to 100 and a cutoff frequency of greater than tens of gigahertz.

Normally, an *n*+ buried layer is used to reduce the series resistance of the collector, since the *n* well has a very high resistivity. However, this further complicates the process by introducing *p*-type epitaxy and one more masking step. Other variations on the bipolar transistor include poly-emitter and self-aligned base contacts to minimize parasitic effects.

## A.2.8 Lateral *pnp* Transistor

The fact that most BiCMOS processes do not have optimized *pnp* transistors makes circuit design somewhat difficult. However, in noncritical situations, a parasitic lateral *pnp* transistor can be used (Fig. A.11).

In this case, the $n$ well serves as the $n$-base region, with the $p+$ diffusions as the emitter and the collector. The base width is determined by the separation between the two $p+$ diffusions. Since the doping profile is not optimized for the base–collector junctions and because the base width is limited by the minimum photolithographic resolution, the performance of this device is not very good: typically, $\beta$ is around 10, and the cutoff frequency is low.

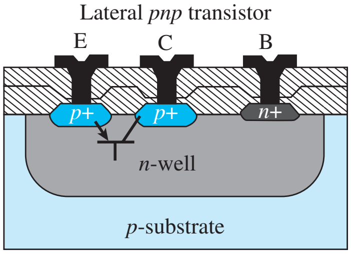

**Figure A.11** Lateral *pnp* transistor.

## A.2.9 *p*-Base and Pinched-Base Resistors

With the additional $p$-base diffusion in the BiCMOS process, two additional resistor structures are available. The $p$-base diffusion can be used to form a straightforward $p$-base resistor as shown in Fig. A.12. Since the base region is usually of a relatively low doping level and has a moderate junction depth, it is suitable for medium-value resistors (a few kilohms). If a large resistor value is required, the pinched-base resistor can be used. In this structure, the $p$-base region is encroached by the $n+$ diffusion, restricting the conduction path. Resistor values in the range of 10 kΩ to 100 kΩ can be obtained. As with the diffusion resistors discussed earlier, these resistors exhibit poor tolerance and temperature coefficients but relatively good matching.

## A.2.10 SiGe BiCMOS Process

With the burgeoning of wireless communication applications, the demand for high-performance, high-frequency RF integrated circuits is tremendous. Owing to the fundamental limitations of physical material properties, silicon-based technology cannot offer some transistor properties achievable with compounds from groups III through IV, such as GaAs. For example, by incorporating a controlled amount (typically no more than 15–20% mole

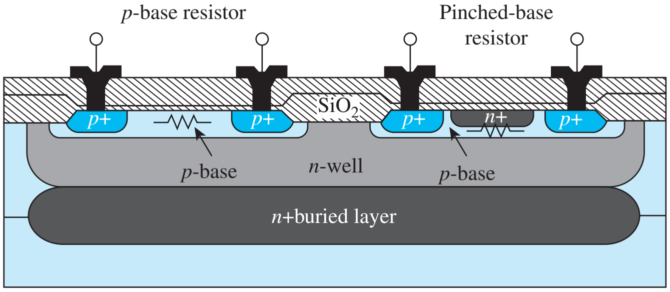

**Figure A.12** *p*-base and pinched *p*-base resistors.

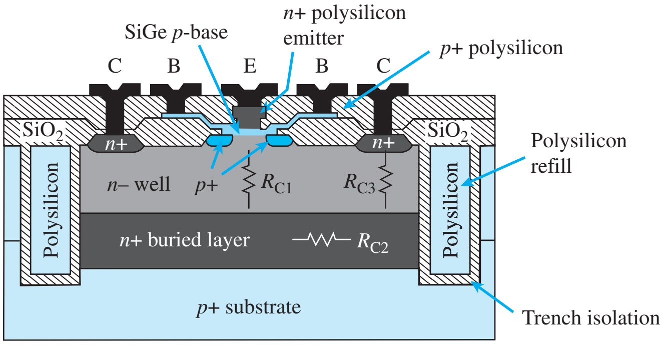

**Figure A.13** Cross-sectional diagram of a symmetric self-aligned SiGe heterojunction bipolar transistor, or HBT.

fraction) of germanium (Ge) into crystal silicon (Si) in the BJT's base region, the energy bandgap can be altered. The specific concentration profile of the Ge can be engineered in such a way that the energy bandgap can be gradually reduced from the pure Si region to a lower value in the SiGe region. This energy bandgap reduction produces a built-in electric field that can assist the movement of carriers, hence resulting in faster operating speed. Therefore, SiGe bipolar transistors can achieve significantly higher cutoff frequency (e.g., in the 100–200 GHz range). Moreover SiGe processing is compatible with existing Si-based fabrication technology, ensuring a very favorable combination of cost and performance.

To take advantage of the SiGe material characteristics, the basic bipolar transistor structure must also be modified to further reduce parasitic capacitance (for higher speed) and to improve the injection efficiency (for higher gain). A symmetric bipolar device structure is shown in Fig. A.13. The device made use of trench isolation to reduce the collector sidewall capacitance between the $n$-well/$n+$ buried layer and the $p$ substrate. The emitter size and the $p+$ base contact size are defined by a self-aligned process to minimize the base–collector junction (Miller) capacitance. This type of device is called a heterojunction bipolar transistor (HBT) since the emitter–base junction is formed from two different types of material, polysilicon emitter and SiGe base. The injection efficiency is significantly better than a homojunction device (as in a conventional BJT). This advantage, coupled with the fact that base width is typically only around 50 nm, makes it easy to achieve current gain of more than 100. In addition, not shown in Fig. A.13, is the possible use of multiple layers of metallization to further reduce the device size and interconnect resistance. All these device features are necessary to complement the high-speed performance of SiGe material.

## A.3 VLSI Layout

The designed circuit schematic must be transformed into a layout that consists of the geometric representation of the circuit components and interconnections. Today, computer-aided design tools allow many of the conversion steps, from schematic to layout, to be carried out semi- or fully automatically. However, any good mixed-signal IC designer must have

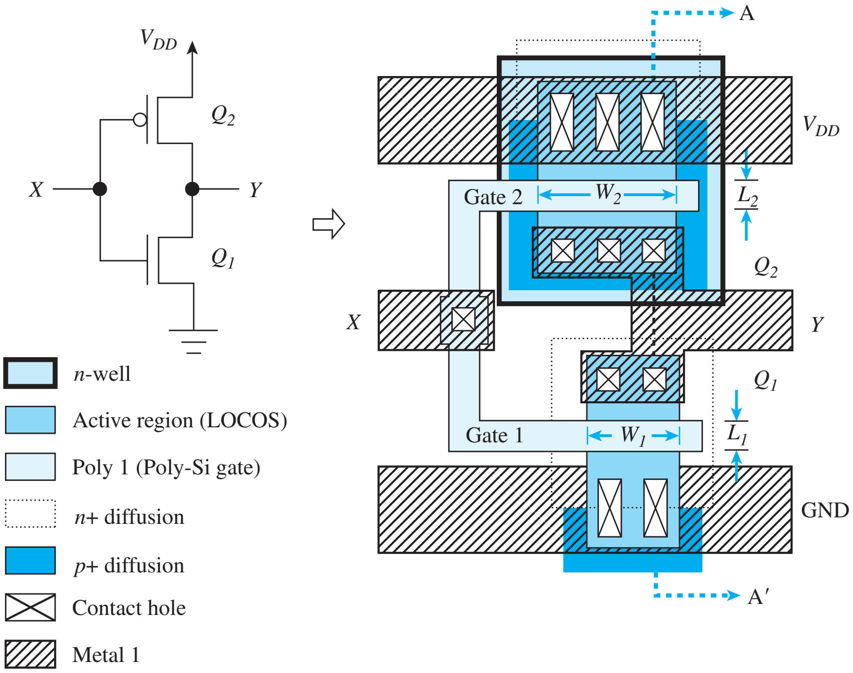

**Figure A.14** A CMOS inverter schematic and its layout.

practiced full custom layout at one point or another. An example of a CMOS inverter can be used to illustrate this procedure (Fig. A.14).

The circuit must first be “flattened” and redrawn to eliminate any interconnection crossovers, similar to the requirement of a printed-circuit-board layout. Each process is made up of a specific set of masking layers. In this case, seven layers are used. Each layer is usually assigned a unique color and fill pattern for ease of identification on a computer screen. The layout begins with the placement of the transistors. For illustration purposes, the $p$ and $n$ MOSFETs are placed in an arrangement similar to that shown in the schematic. In practice, the designer is free to choose the most area-efficient layout. The MOSFETs are defined by the active areas overlapped by the “poly 1” layer. The MOS channel length and width are defined by the width of the “poly 1” strip and that of the active region, respectively. The $p$-MOSFET is enclosed in an $n$ well. For more complex circuits, multiple $n$ wells can be used for different groups of $p$-MOSFETs. The $n$-MOSFET is enclosed by the $n$+ diffusion mask to form the source and drain, while the $p$-MOSFET is enclosed by the $p$+ diffusion mask. Contact holes are placed in regions where connection to the metal layer is required. Finally, the “metal 1” layer completes the interconnections.

The corresponding cross-sectional diagram of the CMOS inverter along the AA′ plane is as shown in Fig. A.15. The poly-Si gates for both transistors are connected to form the input terminal, $X$. The drains of both transistors are tied together via “metal 1” to form the output terminal, $Y$. The sources of the $n$- and $p$-MOSFETs are connected to GND and $V_{DD}$, respectively. Note that butting contacts consist of side-by-side $n$+/$p$+ diffusions that

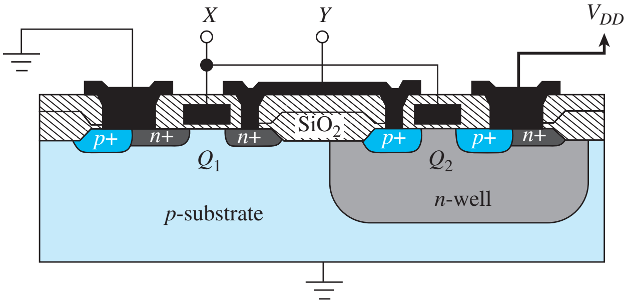

**Figure A.15** Cross section along the plane AA$'$ of a CMOS inverter. Note that this particular layout is good for illustration purposes, but is not necessarily appropriate for latchup prevention.

are used to tie the body potential of the *n*- and *p*-MOSFETs to the appropriate voltage levels.

When the layout is completed, the circuit must be verified using CAD tools such as the circuit extractor, the design rule checker (DRC), and the circuit simulator. Once these verifications have been satisfied, the design can be “taped out” to a mask-making facility. A pattern generator (PG) machine can then draw the geometries on a glass or quartz photoplate using electronically driven shutters. Layers are drawn one by one onto different photoplates. After these plates have been developed, clear and dark patterns resembling the geometries on the layout will result. A set of the photoplates for the CMOS inverter example is shown in Fig. A.16. Depending on whether the drawn geometries are meant to be opened as windows or kept as patterns, the plates can be **clear** or **dark field**. Note that each of these layers must be processed in sequence. The layers must be aligned within very fine tolerance to form the transistors and interconnections. Naturally, the greater the number of layers, the more difficult it is to maintain the alignment. This also requires better photolithography equipment and may result in lower yield. Hence, each additional mask will be reflected in an increase in the final cost of the IC chip.

## A.4 Beyond 20 nm Technology

The rapid advancement of VLSI fabrication technology has followed a prediction called Moore’s Law for more than four decades. In 1965, Gordon Moore, one of the cofounders of Intel, foresaw that the number of transistors that can be integrated onto a VLSI chip would roughly double every two years. In order to achieve this, the size of the transistor has to be reduced accordingly. Otherwise, the size of the VLSI chip would have grown to an unacceptable size, leading to low yield and high cost. Instead of redesigning a fabrication technology from scratch every time, a scaling procedure is normally carried out. The scaling process is not only an optical shrink of the device surface layout, it also requires the reduction in vertical dimensions such as gate oxide thickness, source and drain junction depths, etc. Ideally, all dimensions and the supply voltage are reduced proportionately so that the electrical field intensities remain constant. However, this approach to scaling has proven difficult to sustain at supply voltages around 1 V. A MOSFET threshold voltage below a few 100 mV is then required, resulting in unacceptably high drain-source leakage

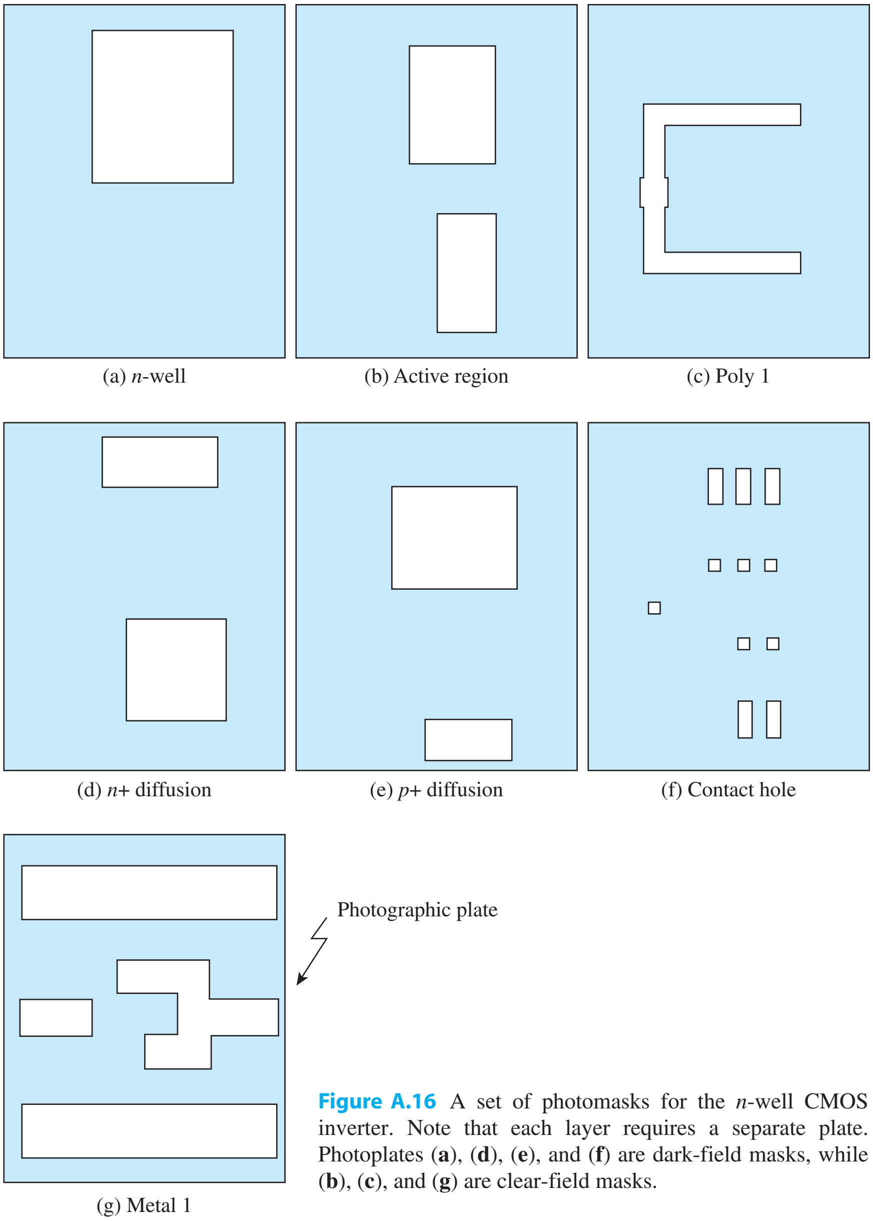

**Figure A.16** A set of photomasks for the *n*-well CMOS inverter. Note that each layer requires a separate plate. Photoplates (**a**), (**d**), (**e**), and (**f**) are dark-field masks, while (**b**), (**c**), and (**g**) are clear-field masks.

current when the transistor is off, especially with many millions of leaky transistors integrated on a single VLSI chip. The effect of MOSFET scaling is illustrated in Fig. A.17. VLSI fabrication technology is categorized by the minimum dimension that it can define. This is usually referred to as the channel length of the MOS gate. The reduction in device dimensions not only allows higher integration density, the shorter channel length and closer proximity of the devices also allow higher switch speed, hence better performance. Moreover, the smaller capacitances of the transistors and their interconnect mean that less charge and less energy is required to turn them on and off, thus reducing circuit power consumption. As a rule of thumb, scaling cannot be carried out with an aggressive factor. Normally, 50% reduction in

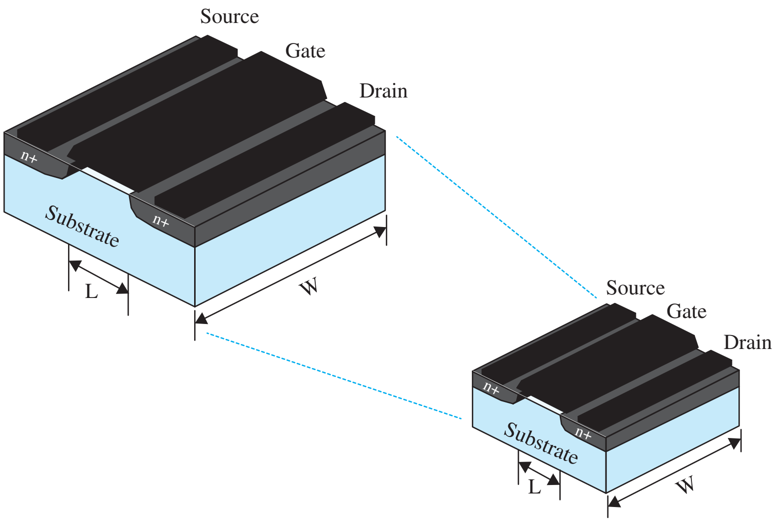

**Figure A.17** MOSFET scaling consists of the reduction of both the surface and vertical dimensions. In addition, modification of the doping profiles and choice of materials are also necessary.

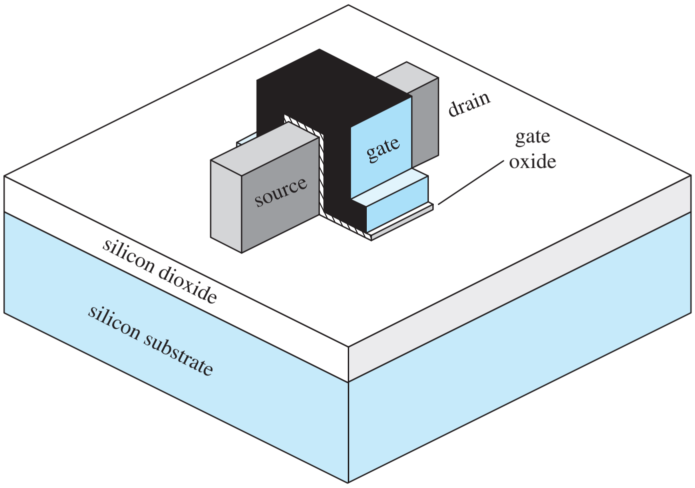

**Figure A.18** A perspective view of the FINFET showing a 3D gate warped around a very thin slab of silicon fin. The source and drain contact areas are actually larger than the intrinsic device.

dimensions is achieved every two generations. Therefore, a scaling factor of approximately 0.7 is normally used. This is why we have technology nodes such as 1 μm in 1990, to 0.7 μm, 0.5 μm, 0.35 μm, 0.25 μm, 0.18 μm, 0.13 μm, 90 nm and so on.

However, this scaling approach cannot continue forever. As we approach the 20 nm technology node, the ultrathin gate dielectric and ultrashort channel length lead to an unacceptable level of gate and drain to source leakage currents. Some of these problems can be addressed by the use of other materials. For example, instead of reducing the gate oxide thickness, similar benefits are obtained by insulating the gate with a higher dielectric constant material such as $HfO_2$. Introducing some germanium can strain the crystal lattice of a MOSFET and improve its carrier mobility. Circuit performance can also be improved by reducing the resistance of the interconnect, hence copper is now sometimes used in place of aluminium.

Patterning such fine features also presents fundamental challenges for the photolithography. Diffraction makes it difficult to accurately pattern features with dimensions far below the wavelength of light used. As a result, there is a trend towards the use of shorter ultraviolet wavelengths. It is also possible to go beyond the diffraction limit by combining multiple photolithographic patterning steps, however this increases the number of masks, processing steps, and cost.

In order to maintain the quest for an even higher level of integration, new device structures have been studied. One of the most promising technologies is the ultra-thin-body (UTB) device. In particular, the FINFET, as illustrated in Fig. A.18, has a three-dimensional gate wrapped around a very thin slab of silicon (the fin) that stands vertically from the surface of an SOI wafer. The thin silicon fin is fully depleted during off condition to suppress drain-source leakage current. In 2018, 7 nm FINFET technology is already being used for the production of high performance VLSI chips.

## Summary

- This appendix presents an overview of the various aspects of VLSI fabrication procedures. This includes component characteristics, process flows, and layouts. This is by no means a complete account of state-of-the-art VLSI technologies. Interested readers should consult other references on this subject for more detailed descriptions.

## Bibliography

S. A. Campbell, Fabrication Engineering at the Micro- and Nanoscale, 4th ed., Oxford University Press, 2014.

R. S. Muller, T.I. Kamins, and M. Chan, *Device Electronics for Integrated Circuits,* 3rd ed., Hoboken, NJ, John Wiley & Sons, 2003.

J. D. Plummer, M.D. Deal, and P.B. Griffin, *Silicon VLSI Technology*, Upper Saddle River, NJ, Prentice Hall, 2000.

S. Wolf, *Microchip Manufacturing*, Lattice Press (www.latticepress.com), 2004.