Chapter 2
MOS Fabrication Technology

Abstract This chapter is concerned with the fabrication of metal–oxide–semiconductor (MOS) technology. Various processes such as wafer fabrication, oxidation, mask generation, photolithography, diffusion, deposition, etc. involved in the fabrication of MOS devices are introduced. Various steps used in the n-type MOS (nMOS) and complementary MOS (CMOS) fabrication are highlighted. The latch-up problem, an inherent problem of CMOS circuits, is introduced and appropriate techniques to overcome this problem are explained. Various short-channel effects arising out of the shrinking size of MOS devices are discussed. Some emerging MOS technologies such as high-K and FinFET to overcome short channel and other drawbacks are introduced.

Keywords Wafer fabrication · Oxidation · Mask generation · Photolithography · Diffusion · Ion implantation · Deposition · Fabrication steps · p-Well process · n-Well process · Twin-tub process · Silicon on insulator · Mask generation · Latch-up problem · Guard ring · Short-channel effect · High-K dielectric · Lightly doped drain structure · FinFET

2.1 Introduction

Metal–oxide–semiconductor (MOS) fabrication is the process used to create the integrated circuits (ICs) that are presently used to realize electronic circuits. It involves multiple steps of photolithographic and chemical processing steps during which electronic circuits are gradually created on a wafer made of pure semiconducting material. Silicon is almost always used, but various compound semiconductors such as gallium–arsenide are used for specialized applications. There are a large number and variety of basic fabrication steps used in the production of modern MOS ICs. The same process could be used for the fabrication of n-type MOS (nMOS), p-type MOS (pMOS), or complementary MOS (CMOS) devices. The gate material could be either metal or poly-silicon. The most commonly used substrate is either bulk silicon or silicon on insulator (SOI). In order to avoid the presence of parasitic transistors, variations are brought in the techniques that are used to isolate the devices in the wafer. This chapter introduces various technologies that are used to fabricate MOS devices. Section 2.2 provides various processes used in the fabrication of MOS devices. Section 2.3 introduces fabrication of nMOS devices. Steps for

A. Pal, *Low-Power VLSI Circuits and Systems*,
DOI 10.1007/978-81-322-1937-8_2, © Springer India 2015

the fabrication of CMOS devices are presented in Sect. 2.4 Latch-up problem and various techniques to prevent it are highlighted in Sect. 2.5. Short-channel effects (SCEs) have been considered in Sect. 2.6 and emerging technologies for low power have been considered in Sect. 2.7.

2.2 Basic Fabrication Processes [1, 2]

Present day very-large-scale integration (VLSI) technology is based on silicon, which has bulk electrical resistance between that of a conductor and an insulator. That is why it is known as a semiconductor material. Its conductivity can be changed by several orders of magnitude by adding impurity atoms into the silicon crystal lattice. These impurity materials supply either free electrons or holes. The donor elements provide electrons and acceptor elements provide holes. Silicon having a majority of donors is known as n-type. On the other hand, silicon having a majority of acceptors is known as p-type. When n-type and p-type materials are put together, a junction is formed where the silicon changes from one type to the other type. Various semiconductor devices such as diode and transistors are constructed by arranging these junctions in certain physical structures and combining them with other types of physical structures, as we shall discuss in the subsequent sections.

2.2.1 Wafer Fabrication

The MOS fabrication process starts with a thin wafer of silicon. The raw material used for obtaining silicon wafer is sand or silicon dioxide. Sand is a cheap material and it is available in abundance on earth. However, it has to be purified to a high level by reacting with carbon and then crystallized by an epitaxial growth process. The purified silicon is held in molten state at about 1500 °C, and a seed crystal is slowly withdrawn after bringing in contact with the molten silicon. The atoms of the molten silicon attached to the seed cool down and take the crystalline structure of the seed. While forming this crystalline structure, the silicon is lightly doped by inserting controlled quantities of a suitable doping material into the crucible. The set up is for wafer fabrication to produce nMOS devices is shown in Fig. 2.1a. Here, boron may be used to produce p-type impurity concentration of 10^{15} cm^3 to 10^{16} per cm^3. It gives resistivity in the range of 25–2 Ω cm. After the withdrawal of the seed, an "ingot" of several centimeters length and about 8–10 cm diameter as shown in Fig. 2.1b is obtained. The ingot is cut into slices of 0.3–0.4 mm thickness to obtain wafer for IC fabrication.

2.2.2 Oxidation

Silicon dioxide layers are used as an insulating separator between different conducting layers. It also acts as mask or protective layer against diffusion and high-energy

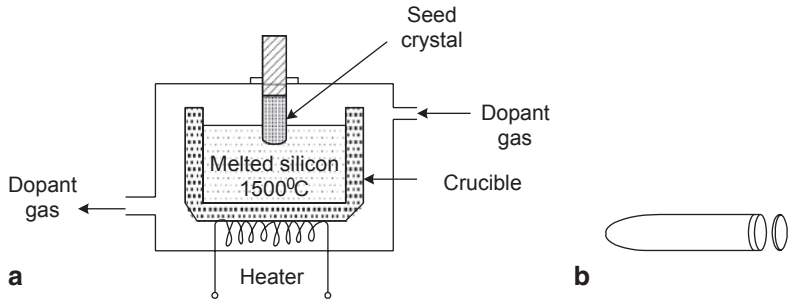


Fig. 2.1 **a** Set up for forming silicon ingot. **b** An ingot

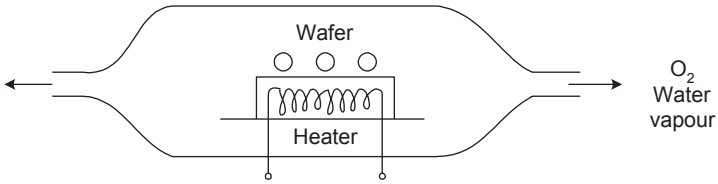


Fig. 2.2 Furnace used for oxidation

ion implantation. The process of growing oxide layers is known as oxidation because it is performed by a chemical reaction between oxygen (dry oxidation), or oxygen and water vapor (wet oxidation) and the silicon slice surface in a high-temperature furnace at about 1000 °C as shown in Fig. 2.2. To grow an oxide layer of thickness t_{ox}, the amount of silicon consumed is approximately $0.5t_{ox}$. Dry oxidation performed in O_2 with a few percent of hydrochloric acid added to produce thin, but robust oxide layers is used to form the gate structure. These layers are known as *gate oxide* layers. The wet oxidation produces a thicker and slightly porous layer. This layer is known as *field oxide* layer. The oxide thickness is limited by the diffusion rate of the oxidizing agent through the already grown layer and is about 1 μm at one atmospheric pressure, but can be doubled by using higher pressure, say approximately 20 atm. Another advantage of a high-pressure system is the possibility to grow thicker oxides in less time at high temperature.

2.2.3 Mask Generation

To create patterned layers of different materials on the wafer, masks are used at different stages. Masks are made of either inexpensive green glass or costly low-expansion glass plates with opaque and transparent regions created using photographic emulsion, which is cheap but easily damaged. Other alternative materials used for creating masks are iron oxide or chromium, both of which are more durable and give better line resolution, but are more expensive.

A mask can be generated either optically or with the help of an electron beam. In the optical process, a reticle, which is a photographic plate of exactly ten times the actual size of the mask, is produced as master copy of the mask. Transparent and opaque regions are created with the help of a pattern generator by projecting an image of the master onto the reticle. Special masking features such as parity masks and fiducials are used on the reticle to identify, align, and orient the mask. Master plates are generated from reticles in a step-and-repeat process by projecting an image of the reticle ten times reduced onto the photosensitized plate to create an array of geometrical shapes in one over the entire plate. Fiducials are used to control the separation between exposures and align the reticle images relative to one another. This process has the disadvantage that if there is a defect on the reticle, it is reproduced on all the chips. The step-and-repeat process not only is slow but also suffers from alignment problems and defect propagation due to dust specks. The electron beam mask generation technique overcomes these problems.

In the electron beam masking process, the masking plate is generated in one step. It is based on the raster scan approach where all the geometrical data are converted into a bit map of 1's and 0's. While scanning the masking plate in a raster scan manner, squares containing 1's are exposed and those containing 0's are not. Exposures are made by blanking and un-blanking the beam controlled by the bit map. Using this technique, several different chip types can be imprinted on the same set of masks. The main disadvantage of this approach is that it is a sequential technique. A better alternative is to use the soft X-ray photolithographic technique in which the entire chip can be eradicated simultaneously. This technique also gives higher resolution.

These master plates are usually not used for mask fabrication. Working plates made from the masters by contact printing are used for fabrication. To reduce turnaround time, specially made master plates can be used for wafer fabrication.

2.2.4 Photolithography

The photolithographic technique is used to create patterned layers of different materials on the wafer with the help of mask plates. It involves several steps. The first step is to put a coating of photosensitive emulsion called photo-resist on the wafer surface. After applying the emulsion on the surface, the wafer is spun at high speed (3000 rpm) to get a very thin (0.5–1 μm) and uniform layer of the photo-resist. Then the masking plate is placed in contact with the wafer in a precise position and exposed to the UV light. The mask plate, with its transparent and opaque regions, defines different areas. With negative photo-resist, the areas of the wafer exposed to UV light are polymerized (or hardened), while with positive photo-resist, the exposed areas are softened and removed.

The removal of the unwanted photo-resist regions is done by a process known as development. Unexposed (negative) or exposed (positive) portions of the photo-resist are chemically dissolved at the time of development. A low-temperature baking process hardens the subsequently remaining portion.

To create the desired pattern, actual removal of the material is done by the *etching* process. The wafer is immersed in a suitable etching solution, which eats out the exposed material leaving the material beneath the protective photo-resist intact. The etching solution depends on the material to be etched out. Hydrofluoric acid (HF) is used for SiO_2 and poly-silicon, whereas phosphoric acid is used for nitride and metal.

Another alternative to this wet chemical etching process is the plasma etching or ion etching. In this dry process, a stream of ions or electrons is used to blast the material away. Ions created by glow discharge at low pressure are directed to the target. Ions can typically penetrate about 800 Å of oxide or photo-resist layers, and thick layers of these materials are used as a mask of some area, whereas the exposed material is being sputtered away. This plasma technique can produce vertical etching with little undercutting. As a consequence, it is commonly used for producing fine lines and small geometries associated with high-density VLSI circuits.

Finally, the photo-resist material is removed by a chemical reaction of this material with fuming nitric acid or exposure to atomic oxygen which oxides away the photo-resist. Patterned layers of different materials in engraved form are left at the end of this process.

2.2.5 Diffusion

After masking some parts of the silicon surface, selective *diffusion* can be done in the exposed regions. There are two basic steps: pre-deposition and drive-in. In the pre-deposition step, the wafer is heated in a furnace at 1000 °C, and dopant atoms such as phosphorous or boron mixed with an inert gas, say nitrogen, are introduced into it. Diffusion of these atoms takes place onto the surface of the silicon, forming a saturated solution of the dopant atoms and solid. The impurity concentration goes up with a temperature up to 1300 °C and then drops. The depth of penetration depends on the duration for which the process is carried out. In the drive-in step, the wafer is heated in an inert atmosphere for few hours to distribute the atoms more uniformly and to a higher depth.

Another alternative method for diffusion is *ion* implantation. Dopant gas is first ionized with the help of an ionizer and ionized atoms are accelerated between two electrodes with a voltage difference of 150 kV. The accelerated gas is passed through a strong magnetic field, which separates the stream of dopant ions on the basis of molecular weights, as it happens in mass spectroscopy. The stream of these dopant ions is deflected by the magnetic field to hit the wafer. The ions strike the silicon surface at high velocity and penetrate the silicon layer to a certain depth as determined by the concentration of ions and accelerating field. This process is also followed by drive-in step to achieve uniform distribution of the ions and increase the depth of penetration.

Different materials, such as thick oxide, photo-resist, or metal can serve as mask for the ion implantation process. But implantation can be achieved through thin oxide layers. This is frequently used to control the threshold voltage of MOS

transistor. This control was not possible using other techniques, and ion implantation is now widely used not only for controlling the threshold voltage but also for all doping stages in MOS fabrication.

2.2.6 Deposition

In the MOS fabrication process, conducting layers such as poly-silicon and aluminium, and insulation and protection layers such as SiO_2 and Si_3N_4 are deposited onto the wafer surface by using the chemical vapor deposition (CVD) technique in a high-temperature chamber:

$$\text{Poly: } SiH_4 \xrightarrow{1000^\circ C} Si + 2H_2$$

$$SiO_2\text{: } SiH_4 + O_2 \xrightarrow{400-450^\circ C} SiO_2 + 2H_2$$

$$Si_3N_4\text{: } 3SiCl_2H_2 + 4NH_3 \xrightarrow{600-750^\circ C} Si_3N_4 + 6HCl + 6H_2$$

Poly-silicon is deposited simply by heating silane at about 1000 °C, which releases hydrogen gas from silane and deposits silicon. To deposit silicon dioxide, a mixture of nitrogen, silane, and oxygen is introduced at 400–450 °C. Silane reacts with oxygen to produce silicon dioxide, which is deposited on the wafer. To deposit silicon nitride, silane and ammonia are heated at about 700 °C to produce nitride and hydrogen. Aluminium is deposited by vaporizing aluminium from a heated filament in high vacuum.

2.3 nMOS Fabrication Steps [2, 3]

Using the basic processes mentioned in the previous section, typical processing steps of the poly-silicon gate self-aligning nMOS technology are given below. It can be better understood by considering the fabrication of a single enhancement-type transistor. Figure 2.3 shows the step-by-step production of the transistor.

Step 1 The first step is to grow a thick silicon dioxide (SiO_2) layer, typically of 1 μm thickness all over the wafer surface using the wet oxidation technique. This oxide layer will act as a barrier to dopants during subsequent processing and provide an insulting layer on which other patterned layers can be formed.

Step 2 In the SiO_2 layer formed in the previous step, some regions are defined where transistors are to be formed. This is done by the photolithographic process discussed in the previous section with the help of a mask (MASK 1). At the end of this step, the wafer surface is exposed in those areas where diffusion regions along with a channel are to be formed to create a transistor.

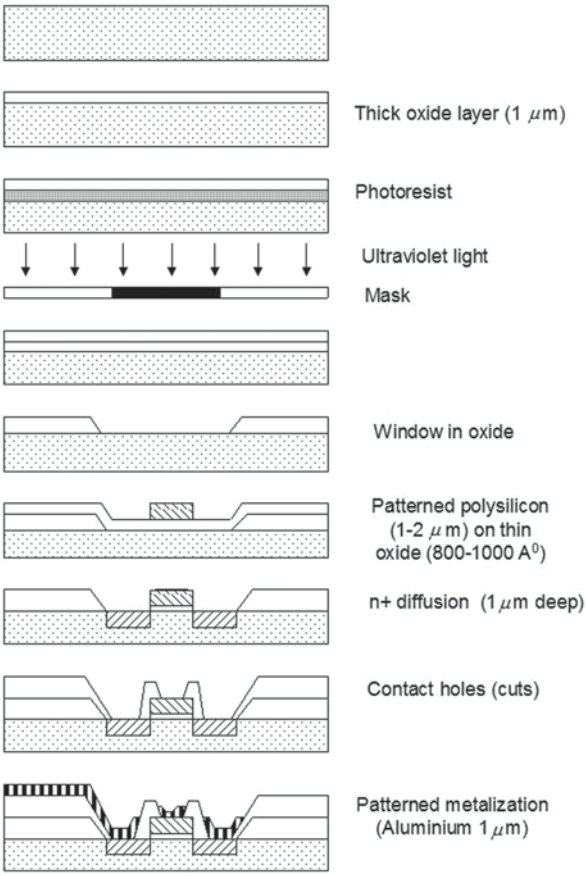


Fig. 2.3 nMOS fabrication steps

Step 3 A thin layer of SiO_2, typically of 0.1 μm thickness, is grown all over the entire wafer surface and on top of this poly-silicon layer is deposited. The poly-silicon layer, of 1.5 μm thickness, which consists of heavily doped poly-silicon is deposited using the CVD technique. In this step, precise control of thickness, impurity concentration, and resistivity is necessary.

Step 4 Again by using another mask (MASK 2) and photographic process, the poly-silicon is patterned. By this process, poly-gate structures and interconnections by poly layers are formed.

Step 5 Then the thin oxide layer is removed to expose areas where n-diffusions are to take place to obtain source and drain. With the poly-silicon and underlying thin oxide layer as the protective mask, the diffusion process is performed. It may be noted that the process is self-aligning, i.e., source and drain are aligned automatically with respect to the gate structure.

Step 6 A thick oxide layer is grown all over again and holes are made at selected areas of the poly-silicon gate, drain, and source regions by using a mask (MASK 3) and the photolithographic process.

Step 7 A metal (aluminium) layer of 1 μm thickness is deposited on the entire surface by the CVD process. The metal layer is then patterned with the help of a mask (MASK 4) and the photolithographic process. Necessary interconnections are provided with the help of this metal layer.

Step 8 The entire wafer is again covered with a thick oxide layer—this is known as *over-glassing.* This oxide layer acts as a protective layer to protect different parts from the environment. Using a mask (MASK 5), holes are made on this layer to provide access to bonding pads for taking external connections and for testing the chip.

The above processing steps allow only the formation of nMOS enhancement-type transistors on a chip. However, if depletion-type transistors are also to be formed, one additional step is necessary for the formation of n-diffusions in the channel regions where depletion transistors are to be formed. It involves one additional step in between step 2 and step 3 and will require one additional mask to define channel regions following a diffusion process using the ion implantation technique.

2.4 CMOS Fabrication Steps [2, 3]

There are several approaches for CMOS fabrication, namely, p-well, n-well, twin-tub, triple-well, and SOI. The n-well approach is compatible with the nMOS process and can be easily retrofitted to it. However, the most popular approach is the p-well approach, which is similar to the n-well approach. The twin-tub and silicon on sapphire are more complex and costly approaches. These are used to produce superior quality devices to overcome the *latch-up problem,* which is predominant in CMOS devices.

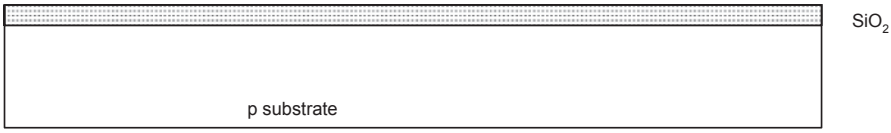
2.4.1 The n-Well Process

The most popular approach for the fabrication of n-well CMOS starts with a lightly doped p-type substrate and creates the n-type well for the fabrication of pMOS transistors. Major steps for n-well CMOS process are illustrated as follows:

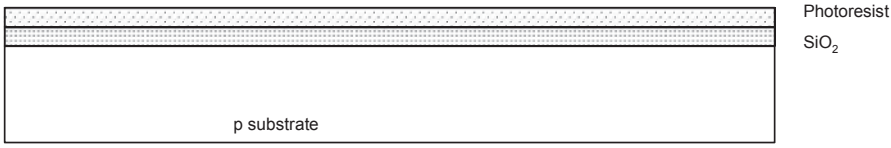
Step 1 The basic idea behind the n-well process is the formation of an n-well or tub in the p-type substrate and fabrication of p-transistors within this well. The formation of an n-well by ion implantation is followed by a drive-in step (1.8×102 p cm^{-2}, 80 kV with 1150 °C for 15 h of drive-in). This step requires a mask (MASK 1), which defines the deep n-well diffusions. The n-transistor is formed outside the well. The basic steps are mentioned below:

- Start with a blank wafer, commonly known as a substrate, which is lightly doped.

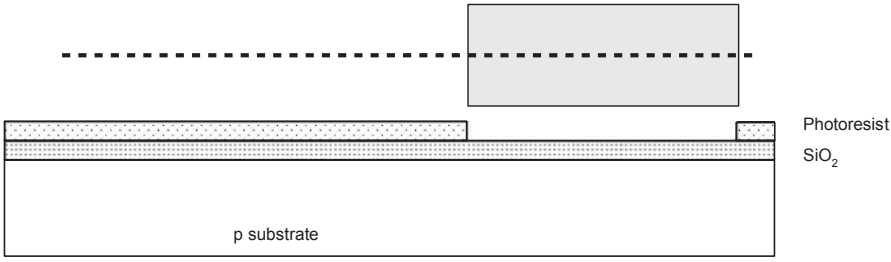
- Cover the wafer with a protective layer of SiO_2 (oxide) using the oxidation process at 900–1200 °C with H_2O (wet oxidation) or O_2 (dry oxidation) in the oxidation furnace.



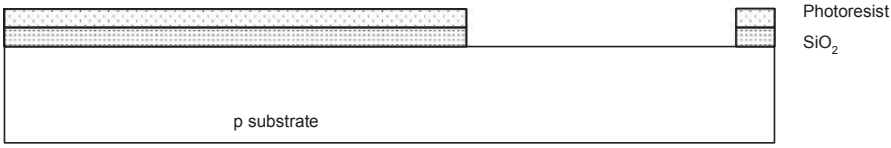
- Spin on photoresist, which is a light-sensitive organic polymer. It softens where exposed to light.



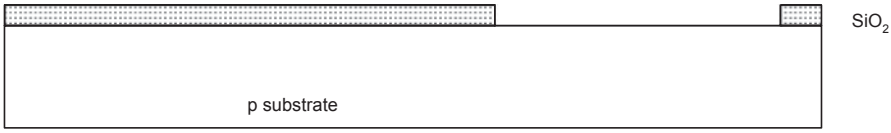
- Expose photoresist through the n-well mask and strip off the exposed photoresist using organic solvents. The n-well mask used to define the n-well in this step is shown below.



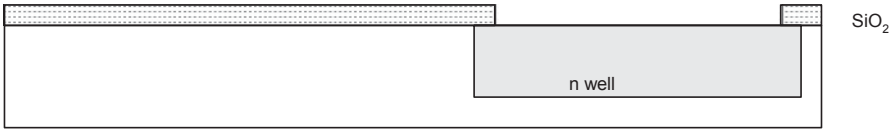
- Etch oxide with HF, which only attacks oxide where the resist has been exposed.



- Remove the photoresist, which exposes the wafer.

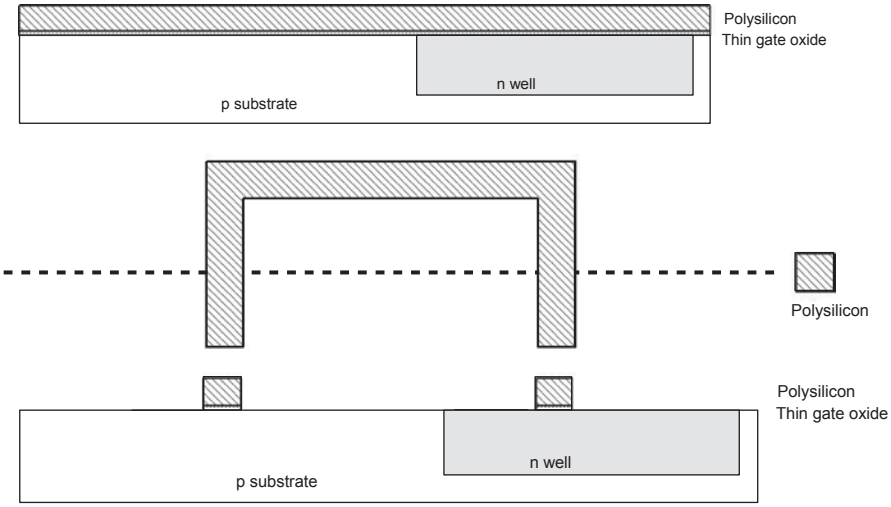


- Implant or diffuse *n* dopants into the exposed wafer using diffusion or ion implantation. The ion implantation process allows shallower wells suitable for the fabrication of devices of smaller dimensions. The diffusion process occurs in all directions and dipper the diffusion more it spreads laterally. This affects how closely two separate structures can be fabricated.



- Strip off SiO_2 leaving behind the p-substrate along with the n-well.

Step 2 The formation of thin oxide regions for the formation of p- and n–transistors requires MASK 2, which is also known as active mask because it defines the thin oxide regions where gates are formed.



http://www.springer.com/978-81-322-1936-1

Low-Power VLSI Circuits and Systems
Pal, A.
2015, XXXVII, 389 p. 303 illus., Hardcover
ISBN: 978-81-322-1936-1