COURSE READER FOR EE143

MICROFABRICATION TECHNOLOGY

SPRING SEMESTER, 2006

Instructor
Professor Nathan Cheung
EECS
# TABLE OF CONTENTS

<table>
<thead>
<tr>
<th>Week</th>
<th>Resource</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Bank, “Introduction to Microengineering”</td>
</tr>
<tr>
<td>2</td>
<td>Overview of processing steps</td>
</tr>
<tr>
<td></td>
<td>Introduction to IC Devices, <a href="http://www.icknowledge.com">www.icknowledge.com</a></td>
</tr>
<tr>
<td></td>
<td>Streetman, Chap 3 “Energy Band and Charge carriers in Semiconductors”</td>
</tr>
<tr>
<td>3</td>
<td>Mayer-Lau, Chapter 9 Thermal Oxidation</td>
</tr>
<tr>
<td>4</td>
<td>Mayer-Lau, Chapter 8 Ion Implantation Notes on Ion Implant profiles</td>
</tr>
<tr>
<td>5</td>
<td>Wolf, Chap 8 Diffusion in Si</td>
</tr>
<tr>
<td>6</td>
<td>Sze, Chap 11 Lithography</td>
</tr>
<tr>
<td>7</td>
<td>Campbell -Chapter 13 Chemical Vapor Deposition</td>
</tr>
<tr>
<td></td>
<td>Reprints - What is a Vacuum, Sputtering</td>
</tr>
<tr>
<td></td>
<td>Deposition, Vacuum Systems, Vapor Pressure</td>
</tr>
<tr>
<td></td>
<td>Tables</td>
</tr>
<tr>
<td>8</td>
<td>Anan, Chap 10 Etching</td>
</tr>
<tr>
<td>9</td>
<td>Campbell, Chap 15 Isolation and Metallization</td>
</tr>
<tr>
<td>10</td>
<td>Uyemura, Chap 3 Layout of CMOS</td>
</tr>
<tr>
<td></td>
<td>Chen, Chap 5 CMOS Integration</td>
</tr>
<tr>
<td>11</td>
<td>Spring Recess</td>
</tr>
<tr>
<td>12</td>
<td>Streetman, Chap 8.3 MOS Principle</td>
</tr>
<tr>
<td>13</td>
<td>Streetman, Chap 8.3 MOS Principle</td>
</tr>
<tr>
<td>14,15</td>
<td>Kovac, Chapter 3 Mechanical Transducers</td>
</tr>
<tr>
<td>16</td>
<td>Nanofabrication Reprint Electrophoresis, 22 (2001)</td>
</tr>
</tbody>
</table>
Week 1

Introduction
INTRODUCTION TO MICROENGINEERING
by Danny Bank

What is Microengineering?

Microengineering refers to the technologies and practice of making three dimensional structures and devices with dimensions in the order of micrometers.

The two constructional technologies of microengineering are microelectronics and micromachining. Microelectronics, producing electronic circuitry on silicon chips, is a very well developed technology. Micromachining is the name for the techniques used to produce the structures and moving parts of microengineered devices.

One of the main goals of Microengineering is to be able to integrate microelectronic circuitry into micromachined structures, to produce completely integrated systems (microsystems). Such systems could have the same advantages of low cost, reliability and small size as silicon chips produced in the microelectronics industry.

When considering such small devices, a number of physical effects have different significance on the micrometer scale compared to macroscopic scales. Interest in microengineering has spawned or renewed interest in a number of areas dealing with the study of these effects on microscopic scales. This includes such topics as micromechanics, which deals with the moving parts of microengineered devices, and microfluidics, etc.

The remainder of this document introduces three of the micromachining techniques that are in use / under development.

Silicon micromachining is given most prominence, since this is one of the better developed micromachining techniques. Silicon is the primary substrate material used in the production microelectronic circuitry (ie, better silicon chips), and so is the most suitable candidate for the eventual production of microsystems.

The Excimer laser is an ultraviolet laser which can be used to micromachine a number of materials without heating them, unlike many other lasers which remove material by burning or vaporising it. The Excimer laser lends itself particularly to the machining of organic materials (polymers, etc).

LIGA is a technique that can be used to produce moulds for the fabrication of micromachined components. Microengineered components can be made from a variety of materials using this technique, however it does suffer the disadvantage that currently the technique requires X-rays from a synchrotron source.

A quick introduction to mask design is provided following discussion of techniques and structures, rather than directly following the photolithography section. This is so that the reader is able to become acquainted with the concept of creating structures by sequential photolithography and machining steps first, which hopefully makes it easier to understand what mask design software is trying to achieve.

Photolithography

Photolithography is the basic technique used to define the shape of micromachined structures in the three techniques outlined below. The technique is essentially the same as that used in the microelectronics industry, which will be described here. The differences in the photolithographic
techniques for Excimer laser micromachining and LIGA will be outlined in the relevant sections.

Figure 1a shows a thin film of some material (e.g., silicon dioxide) on a substrate of some other material (e.g., a silicon wafer). It is desired that some of the silicon dioxide (oxide) is selectively removed so that it only remains in particular areas on the silicon wafer (figure 1f).

Firstly a mask is produced. This will typically be a chromium pattern on a glass plate. The wafer is then coated with a polymer which is sensitive to ultraviolet light (figure 1b), called a photoresist. Ultraviolet light is then shone through the mask onto the photoresist (figure 1c). The photoresist is then developed which transfers the pattern on the mask to the photoresist layer (figure 1d).

There are two types of photoresist, termed positive and negative. Where the ultraviolet light strikes the positive resist it weakens the polymer, so that when the image is developed the resist is washed away where the light struck it - transferring a positive image of the mask to the resist layer. The opposite occurs with negative resist. Where the ultraviolet light strikes negative resist it strengthens the polymer, so when developed the resist that was not exposed to ultraviolet light is washed away - a negative image of the mask is transferred to the resist.

A chemical (or some other method) is then used to remove the oxide where it is exposed through the openings in the resist (figure 1e). Finally the resist is removed leaving the patterned oxide (figure 1f).

Figure 1.

Silicon micromachining

There are a number of basic techniques that can be used to pattern thin films that have been deposited on a silicon wafer, and to shape the wafer itself, to form a set of basic microstructures (bulk silicon micromachining). The techniques for depositing and patterning thin films can be used to produce quite complex microstructures on the surface of silicon wafer (surface silicon micromachining). Electrochemical etching techniques are being investigated to extend the set of basic silicon micromachining techniques. Silicon bonding techniques can also be utilised to extend the structures produced by silicon micromachining techniques into multilayer structures.
Basic techniques

There are three basic techniques associated with silicon micromachining. These are the deposition of thin films of materials, the removal of material ( patterning) by wet chemical etchants, and the removal of material by dry etching techniques. Another technique that is utilised is the introduction of impurities into the silicon to change its properties (ie, doping). Doping is only mentioned very briefly below.

It should be noted that all the techniques are outlined very briefly. Behind each outline hides a number of different methods, each with different advantages and disadvantages.

Thin films

There are a number of different techniques that facilitate the deposition or formation of very thin films (of the order of micrometers, or less) of different materials on a silicon wafer (or other suitable substrate). These films can then be patterned using photolithographic techniques and suitable etching techniques. Common materials include silicon dioxide (oxide), silicon nitride (nitride), polycrystalline silicon (polysilicon or poly), and aluminium.

A number of other materials can be deposited as thin films, including noble metals such as gold. Noble metals will contaminate microelectronic circuitry causing it to fail, so any silicon wafers with noble metals on them have to be processed using equipment specially set aside for the purpose. Noble metal films are often patterned by a method known as "lift off", rather than wet or dry etching.

Often, photoresist is not tough enough to withstand the etching required. In such cases a thin film of a tougher material (eg, oxide or nitride) is deposited and patterned using photolithography. The oxide / nitride then acts as an etch mask during the etching of the underlying material. When the underlying material has been fully etched the masking layer is stripped away.

Wet etching

Wet etching is a blanket name that covers the removal of material by immersing the wafer in a liquid bath of the chemical etchant. Wet etchants fall into two broad categories; isotropic etchants and anisotropic etchants.

Isotropic etchants attack the material being etched at the same rate in all directions. Anisotropic etchants attack the silicon wafer at different rates in different directions, and so there is more control of the shapes produced. Some etchants attack silicon at different rates depending on the concentration of the impurities in the silicon (concentration dependent etching).

Isotropic etchants are available for oxide, nitride, aluminium, polysilicon, gold, and silicon. Since isotropic etchants attack the material at the same rate in all directions, they remove material horizontally under the etch mask (undercutting) at the same rate as they etch through the material. This is illustrated for a thin film of oxide on a silicon wafer in figure 2, using an etchant that etches the oxide faster than the underlying silicon (eg, hydrofluoric acid).

![Figure 2](image.png)

This illustrates the isotropic wet etching of a thin film of material.
The photoresist is black, and the substrate yellow. The film is etched through, and the etching continues to further under-cut the mask.

Anisotropic etchants are available which etch different crystal planes in silicon at different rates. The most popular anisotropic etchant is potassium hydroxide (KOH), since it is the safest to use.

Silicon wafers are slices that have been cut from a large ingot of silicon that was grown from a single seed crystal. The silicon atoms are all arranged in a crystalline structure, so the wafer is monocrystalline silicon (as opposed to polycrystalline silicon mentioned above). When purchasing silicon wafers it is possible to specify that they have been sliced with the surface parallel to a particular crystal plane.

The simplest structures that can be formed using KOH to etch a silicon wafer with the most common crystal orientation (100) are shown in figure 3. These are V shaped grooves, or pits with right angled corners and sloping side walls. Using wafers with different crystal orientations can produce grooves or pits with vertical walls.

![Figure 3.](image)

Both oxide and nitride etch slowly in KOH. Oxide can be used as an etch mask for short periods in the KOH etch bath (ie, for shallow grooves and pits). For long periods, nitride is a better etch mask as it etches more slowly in the KOH.

Concentration Dependent Etching. High levels of boron in silicon will reduce the rate at which it is etched in KOH by several orders of magnitude, effectively stopping the etching of the boron rich silicon.

The boron impurities are usually introduced into the silicon by a process known as diffusion. A thick oxide mask is formed over the silicon wafer and patterned to expose the surface of the silicon wafer where the boron is to be introduced (figure 4a). The wafer is then placed in a furnace in contact with a boron diffusion source. Over a period of time boron atoms migrate into the silicon wafer. Once the boron diffusion is completed, the oxide mask is stripped off (figure 4b).

A second mask may then be deposited and patterned (figure 4c) before the wafer is immersed in the KOH etch bath. The KOH etches the silicon that is not protected by the mask, and etches around the boron doped silicon (figure 4d).
Boron can be driven into the silicon as far as 20µm over periods of 15 to 20 hours, however it is desirable to keep the time in the furnace as short as possible. With complex designs, etching the wafer from the front in KOH may cause problems where slow etching crystal planes prevent it from etching beneath the boron doped silicon. In such cases the wafer can be etched from the back, however this is not without disadvantages (longer etching times, more expensive wafers, etc). The high concentration of boron required means that microelectronic circuitry cannot be fabricated directly on the boron doped structure.

**Dry etching**

The most common form of dry etching for micromachining applications is reactive ion etching (RIE). Ions are accelerated towards the material to be etched, and the etching reaction is enhanced in the direction of travel of the ion. RIE is an anisotropic etching technique. Deep trenches and pits (up to ten or a few tens of microns) of arbitrary shape and with vertical walls can be etched in a variety of materials including silicon, oxide and nitride. Unlike anisotropic wet etching, RIE is not limited by the crystal planes in the silicon.

**Lift off**

Lift off is a stencilling technique often used to pattern noble metal films. There are a number of different techniques, the one outlined here is an assisted lift off method.

A thin film of the assisting material (eg. oxide) is deposited. A layer of resist is put over this and patterned, as for photolithography, to expose the oxide in the pattern desired for the metal (figure 5a). The oxide is then wet etched so as to undercut the resist (figure 5b). The metal is then deposited on the wafer, typically by a process known as evaporation (figure 5c). The metal pattern is effectively stencilled through the gaps in the resist, which is then removed lifting off the unwanted metal with it (figure 5d). The assisting layer is then stripped off too, leaving the metal pattern alone (figure 5e).
There are lift off techniques in which only photoresist is used as the stencil. In the assisted lift off, an intermediate layer assists in the process to ensure a clean lift off and well defined metal pattern. When noble metals are used it is desirable to deposit a thin layer of a more active metal (e.g., chrome) first, to ensure good adhesion of the noble metal.

**Basic structures**

**Bulk silicon micromachining**

One of the simplest possible, and most obvious, structures is the patterning of insulated electrical conductors. One possible application of this could be to use electric fields to manipulate individual cells.

Anisotropic etching with KOH can easily form V shaped grooves, or cut pits with tapered walls into silicon (figure 3 and figure 6).

KOH can also be used to produce mesa structures (figure 7a). When etching mesa structures the
corners can become bevelled (figure 7b), rather than right angle corners. This has to be compensated for in some way. Typically the etch mask is designed to include additional structures on the corners. These compensation structures are designed so that they are etched away entirely when the mesa is formed to leave 90 degree corners. One problem with using compensation structures to from right angled mesa corners is that they put a limit on the minimum spacing between the mesas.

![Figure 7](image)

Silicon diaphragms from about 50µm thick upwards can be made by etching through an entire wafer with KOH (figure 8). The thickness is controlled by timing the etch, and so is subject to errors.

![Figure 8](image)

Thinner diaphragms, of up to about 20µm thick, can be produced using boron to stop the KOH etch (figure 9) - Concentration dependent etching. The thickness of the diaphragm is dependent on the depth to which the boron is diffused into the silicon, which can be controlled more accurately than the simple timed KOH etch.

![Figure 9](image)

The silicon diaphragm is the basic structure used in microengineered pressure sensors, for example. It can also be adapted for use as an acceleration sensor.

Concentration dependent etching can also be used to produce narrow bridges, or cantilever beams. Figure 10a shows a bridge, defined by a boron diffusion, spanning a pit that was etched from the front of the wafer in KOH. A cantilever beam (a bridge with one end free) produced by the same method is shown in figure 10b.
The bridge and beam in figure 10 project across the diagonal of the pit to ensure that they will be etched free by the KOH. More complex structures are possible using this technique, but care must be taken to ensure that they will be etched free by the KOH.

If it is desired to make beams or bridges of a different orientation, the wafer can be etched through from the back in KOH (figure 11). This will ensure that the structure is released from the silicon. During such etching, it is necessary to ensure that the front of the wafer is adequately protected from the long KOH etch. Another alternative could be to produce a diaphragm, and etch the desired bridge or beam shape using a reactive ion etcher (dry etching).

One of the applications for these beams and bridges is as resonant sensors. The structure can be set vibrating at its fundamental frequency. Anything causing a change in the mass, length, etc., of the structure will register as a change frequency. Care has to be taken to ensure that only the quantity to be measured causes a significant change in frequency.

A combination of dry etching and isotropic wet etching can be used to form very sharp points. First a column with vertical sides is etched away using an RIE (figure 12a). A wet etch is then used, which undercuts the etch mask leaving a very fine point (figure 12b), the etch mask is then removed.

Very fine points like this can be fabricated on the end of cantilever beams as probes for use in atomic force microscopy. The technique can also be used to produce sharp, small blades.
Surface micromachining

The anisotropic wet etching, and concentration dependent etching techniques discussed in the "Basic Techniques" section are generally called bulk silicon micromachining techniques. This is because the microstructures are formed by etching away the bulk of the silicon wafer to achieve the desired result. Surface micromachining techniques build up the structure in layers of thin films on the surface of the silicon wafer (or any other suitable substrate).

The process would typically employ films of two different materials, a structural material (commonly polysilicon) and a sacrificial material (oxide). These are deposited and dry etched in sequence. Finally the sacrificial material is wet etched away to release the structure. The more layers, the more complex the structure, and the more difficult it becomes to fabricate.

A simple surface micromachined cantilever beam is shown in figure 13. A sacrificial layer of oxide is deposited on the surface of the wafer. A layer of polysilicon is then deposited, and patterned using RIE techniques to a beam with an anchor pad (figure 13a). The wafer is then wet etched to remove the oxide layer under the beam, freeing it (figure 13b). The anchor pad has been under etched, however the wafer was removed from the etch bath before all the oxide was removed from under the pad leaving the beam attached to the wafer.

Figure 13.

A variety of different chambers can be fabricated on the surface of silicon wafers using surface micromachining techniques. In figure 14, the chamber is defined by a volume of sacrificial oxide (figure 14a). A layer of polysilicon is then deposited over the surface of the wafer (figure 14b). A window is dry etched (RIE) through the polysilicon, and the wafer is then immersed in a wet etch that removes the oxide, leaving a windowed chamber (figure 14c).

Figure 14.

Surface micromachining can potentially produce quite complicated structures; such as microengineered tweezers, and gear trains.

Electrochemical etching of silicon

A variety of electrochemical silicon etching techniques are under development. One of these is the electrochemical passivation technique.

A wafer with a particular impurity concentration is used, and different impurities are diffused (or
implanted) into the wafer. This is done to form a diode junction at the boundary between the
differently doped areas of silicon. The junction will delineate the structure to be produced. An
electrical potential is then applied across the diode junction, and the wafer is immersed in a suitable
wet etch (KOH). This is done in such a way that when the etch reaches the junction an oxide layer
(passivation layer) is formed which protects the silicon from further etching.

This is another bulk silicon micromachining technique, and is essentially similar to the boron etch
stop technique (concentration dependent etching). The structures that can be produced are similar
to those produced by the boron etch stop technique. The main advantage of the electrochemical
method is that much lower concentrations of impurities are required, so the resulting structure is
more compatible with the fabrication of microelectronic circuitry.

**Wafer bonding**

There are a number of different methods available for bonding micromachined silicon wafers
together, or to other substrates, to form larger more complex devices.

A method of bonding silicon to glass that appears to be gaining in popularity is anodic bonding
(electrostatic bonding). The silicon wafer and glass substrate are brought together and heated to a
high temperature. A large electric field is applied across the join, which causes an extremely strong
bond to form between the two materials. Figure 15 shows a glass plate bonded over a channel
etched into a silicon wafer (RIE), forming a pipe through which fluid can flow.

![Figure 15](image)

It is also possible to bond silicon wafers directly together using gentle pressure, under water (direct
silicon bonding).

Other bonding methods include using an adhesive layer, such as a glass, or photoresist. Whilst
anodic bonding and direct silicon bonding form very strong joins they suffer from some
disadvantages, including the requirement that the surfaces to be joined are very flat and clean.

Wafer bonding techniques can potentially be combined with some of the basic micromachined
structures to form the valves, pumps, etc, of a microfluid handling system.

**Excimer laser micromachining**

Excimer lasers produce relatively wide beams of ultraviolet laser light. One interesting application of
these lasers is their use in micromachining organic materials (plastics, polymers, etc). This is because
the excimer laser doesn't remove material by burning or vaporising it, unlike other types of laser, so
the material adjacent to the area machined is not melted or distorted by heating effects.

When machining organic materials the laser is pulsed on and off, removing material with each pulse.
The amount of material removed is dependent on the material itself, the length of the pulse, and the
intensity (fluence) of the laser light. Below a certain threshold fluence, dependent on the material, the
laser light has no effect. As the fluence is increased above the threshold, the depth of material
removed per pulse is also increased. It is possible to accurately control the depth of the cut by counting the number of pulses. Quite deep cuts (hundreds of microns) can be made using the excimer laser.

The shape of the structures produced is controlled by using a chrome on quartz mask, like the masks produced for photolithography. In the simplest system the mask is placed in contact with the material being machined, and the laser light is shone through it (figure 16a). A more sophisticated and versatile method involves projecting the image of the mask onto the material (figure 16b). Material is selectively removed where the laser light strikes it.

![Figure 16](image16.png)

Structures with vertical sides can be created. By adjusting the optics it is possible to produce structures with tapered sidewalls (figure 17).

![Figure 17](image17.png)

Excimer lasers have a number of applications beyond those mentioned here. One area of application is in machining the cornea of the eye to change its optical properties; correcting for short sight.

**LIGA**

The acronym LIGA comes from the German name for the process (Lithographie, Galvanoformung, Abformung). LIGA uses lithography, electroplating, and moulding processes to produce microstructures. It is capable of creating very finely defined microstructures of up to 1000µm high.

In the process as originally developed, a special kind of photolithography using X-rays (X-ray lithography) is used to produce patterns in very thick layers of photoresist. The X-rays from a synchrotron source are shone through a special mask onto a thick photoresist layer (sensitive to X-rays) which covers a conductive substrate (figure 18a). This resist is then developed (figure 18b).

The pattern formed is then electroplated with metal (figure 18c). The metal structures produced can be the final product, however it is common to produce a metal mould (figure 18d). This mould can then be filled with a suitable material, such as a plastic (figure 18e), to produce the finished product.
in that material (figure 18f).

Figure 18.

As the synchrotron source makes LIGA expensive, alternatives are being developed. These include high voltage electron beam lithography which can be used to produce structures of the order of 100µm high, and excimer lasers capable of producing structures of up to several hundred microns high.

Electroplating is not limited to use with the LIGA process, but may be combined with other processes and more conventional photolithography to produce microstructures.

General LIGA Process Information

LIGA Process Flow

The overall process for the multi-project LIGA runs is shown in Figures 1-4 and has the following structure. Please note some changes since the last LIGA run.

Starting material: 4 inch silicon substrate
Oxidize to 5,000 Angstroms; unpatterned
Sputter plating base; unpatterned - 300 Angstroms of titanium followed by 5000 Angstroms of copper followed by another 300 Angstroms of titanium. The titanium and copper layers also act as sacrificial layers. The titanium and copper layers can be etched simultaneously to release the plated structures if desired. Apply thick photoresist (PMMA) and mill to thickness. Expose and develop the photoresist using the nickel level mask. Electroplate with pure nickel. For the thick nickel process, the metal is overplated and then milled to achieve a height of 200 +/- 5 microns. The photoresist is then removed to produce nickel structures with a height of 200 +/- 5 microns. For the thin nickel process, the metal is plated to 30 microns with no subsequent milling. The surface roughness will be that of the as-plated nickel. Estimated height control is +/- 5 microns and is pattern dependent. Released nickel structures can be produced by etching away the sacrificial titanium layer in a fresh solution of 1 part...
Ammonium Hydroxide, 1 part Hydrogen Peroxide, and 6 parts of Deionized Water, at room temperature. This release step is not necessary if only fixed structures are desired. Note: The structures on a die are either all fixed or all released as the sacrificial layer is not patterned. For released structures, the final release step must be performed by the user at the customer site.

**Hints for Good Mask Design**

Mechanical milling generates shear stresses on the photoresist and metal. This affects both photoresist and metal adhesion. Yield improves if retained photoresist areas are large. Yield decreases when the developed photoresist pattern contains mechanically weak structures. In the thick nickel process, the nickel and photoresist are milled after plating. Yield is better on larger nickel areas than on smaller ones. However, during milling the photoresist itself can also support milling stress and certain geometries are better than others. For example, a hexagonal nickel post can be milled easier than a circular one which will tend to rotate since it has no PMMA support. Uniformity in metal plating is pattern dependent. A good layout is one in which 50% of the chip area is covered with metal.
Fig. 1 Fabrication is done on a (100) silicon wafer with a 0.5μm oxide layer. A plating base is formed by sputtering 300nm of Ti and 5000A of Cu with a top layer of 300A Ti. The Ti and Cu also act as a release layer. Thick photoresist is applied and exposed using x-rays from a synchotron and developed with a solvent.

Fig. 2 The desired metal, in this case nickel, is electroplated onto the substrate, filling the voids in the PMMA.

Fig. 3 The metal and PMMA are milled back to produce a uniform top surface.

Fig. 4 Finally, the PMMA is removed. If desired, the customer can release the structures from the substrate by etching away the plating base in an NH$_4$O$_2$/H$_2$O$_2$ solution.
Week 2

Overview and Semiconductor Basics
Electronic Materials used in Si IC Technology

<table>
<thead>
<tr>
<th>Material</th>
<th>Preparation method</th>
<th>Substrate temperature</th>
<th>Material failure temperature</th>
<th>Comments</th>
</tr>
</thead>
<tbody>
<tr>
<td>Si Substrate</td>
<td>Crystal growth</td>
<td>~ 1400°C</td>
<td>~ 1400°C</td>
<td>Thermal stressing at &gt; 1300 °C can create crystalline defects such as dislocations and slipping planes</td>
</tr>
<tr>
<td>Poly-Si</td>
<td>CVD</td>
<td>~ 600°C</td>
<td>~1400 °C</td>
<td>Heavily doped poly-Si will have growth when T &gt; 1000 °C</td>
</tr>
<tr>
<td>Thermal SiO₂</td>
<td>Thermal oxidation</td>
<td>800-1100 °C</td>
<td>&gt; 1200 °C</td>
<td></td>
</tr>
<tr>
<td>CVD SiO₂ (SiO₂, PSG, BSG, BPSG)</td>
<td>CVD</td>
<td>300-600 °C</td>
<td>&gt; 1200 °C</td>
<td>The PSG, BSG, BPSG glasses will soften (“reflow”) when T &gt; 1000 °C</td>
</tr>
<tr>
<td>Spin-on Glass</td>
<td>spin-on</td>
<td>ambient</td>
<td>&gt;1200 °C</td>
<td></td>
</tr>
<tr>
<td>Si₃N₄</td>
<td>CVD</td>
<td>200-500 °C</td>
<td>&gt; 1200 °C</td>
<td></td>
</tr>
<tr>
<td>Al</td>
<td>Sputtering or evaporation deposition</td>
<td>ambient</td>
<td>560 °C</td>
<td>Melting point of Al/Si alloys is at 560 °C but process temperature is restricted to &lt; 450 °C to prevent junction spiking</td>
</tr>
<tr>
<td>Photoresist</td>
<td>spin-on</td>
<td>ambient</td>
<td>150 °C</td>
<td>Photoresist will reflow around 150 °C. Higher temperature will carbonize it.</td>
</tr>
<tr>
<td>Refractory metals</td>
<td>sputtering or CVD</td>
<td>&lt; 400 °C</td>
<td>&gt; 1000 °C</td>
<td>Used as interconnects and diffusion barriers</td>
</tr>
<tr>
<td>Polyimide</td>
<td>spin-on</td>
<td>ambient</td>
<td>450 °C</td>
<td>Used an insulating dielectric between metal layers</td>
</tr>
</tbody>
</table>

* The modern trend for IC processing is to minimize high-temperature cycles to avoid:

1) Thermal stress build-up in IC device structures
2) Excessive dopant redistribution by thermal diffusion.

[Comments]

* After Al deposition, high-temperature processing steps (e.g. post-implantation annealing at 900 °C) are not allowed.

* The photoresist layer has to be removed prior to any processing steps with substrate temperature > 150 °C. Usually, the only processing steps which allows photoresist to stay on the wafer are: ion implantation, sputtering or evaporation deposition.
Photolithography
Thermal Oxidation

- O$_2$ (or H$_2$O) diffuses through SiO$_2$ and reacts with Si at the interface to form more SiO$_2$.

- 1$\mu$m of SiO$_2$ formed consumes 0.44 $\mu$m of Si substrate.

- Thin oxide (e.g. gate oxide) - dry ox

- Thick oxide (e.g. field oxide) - wet ox

Oxide (X$_{ox}$) thickness

$\sqrt{t}$

Time (t)
Local Oxidation (LOCOS)

Window Oxidation
- Precise dose control and good uniformity.
Diffusion

\[ D = D_0 \exp \left( \frac{Q}{kT} \right) \]

\( D \) = Diffusion Constant

\( Q \) = Activation Energy

\( T \) = Temp in °K (D↑ as T↑)
**Predeposition**

- Usually replaced by low-energy implant these days.

- Si surface has dopant concentration = Cs (solid-solubility)

- Dose of dopant incorporation = \( \frac{Cs \cdot 2\sqrt{Dt}}{\sqrt{\pi}} \)

**Predeposition + Drive-in**

- Half-gaussian profile after long drive-in

- Dopant dose conserved during drive-in.

- Diffusion distance. \( \equiv \sqrt{Dt} \)
Chemical Vapor Deposition (CVD)

- Conformed Coverage
- Deposition Rate can be quite high.
Physical Vapor Deposition (PVD)

- Evaporation or Sputtering
- Line-of-sight deposition process
- Step Coverage Problem: (i) geometrical shadowing
  (ii) self-shadowing

```
---                  |
SiO₂               |
---                  |
 Si                   
```

```
---                  |
Al                  |
---                  |
 SiO₂               |
---                  |
 Si                   
```

```
---                  |
Photoresist            |
---                  |
  Si                  |
---                  |
  Si                  |
```

"Lift-off" process
Epitaxial Growth

(a) Single-crystal silicon

(b) Epitaxial layer

(c) Epitaxial layer

- Needs atomically clean Si surface and high growth temperature (> 1000°C)

- Applications include: Bipolar Subcollector and CMOS latch-up prevention.
2.0. Introduction to integrated circuit devices

2.1. Introduction

In this chapter the basics of the devices that make up integrated circuits will be presented. The objective of this chapter is to give the reader enough of an understanding of the operation of devices fabricated on integrated circuits (ICs) to be able to understand the process trade-offs and challenges discussed later in the book. For readers who already have a good working familiarity with device physics, this chapter may be skipped.

2.2. Basic electronic concepts

Electronic circuits regulate and control the flow of electric current. Electric current is the flow of electrons, the tiny subatomic particles that surround the nucleus of atoms. Electrons carry a fixed negative electric charge and the movement of electrons carries charge from one location to another - the flow of electrons is referred to as electric current. Electric current is driven by a difference in potential from one location to another location measured in volts. Electric current flows easily through materials that are conductors, and is blocked by materials that are insulators. The amount of resistance that a material presents to the flow of electric current is logically called resistance. Conductors have low resistance to the flow of electric current and insulators have extremely high resistance (essentially infinite until the voltage is so high that the material breaks down). For a given voltage, the higher the resistance the lower the current that will flow and the lower the resistance the higher the current that will flow. Conversely, for a given resistance, the higher the voltage the higher the current that will flow and the lower the voltage the lower the current that will flow.

2.3. Electronic circuit elements

Electronic circuits are made up of a number of elements used to control current flow. There are a wide variety of different circuit elements but for the purpose of this discussion the circuit elements will be restricted to the four most commonly used in ICs; resistors, capacitors, diodes and transistors. Resistors, provide a fixed amount of resistance to current flow. Capacitors, store electronic charge until discharged, somewhat similar to a battery. Diodes, allow current to flow in one direction but not in the opposite direction, a one way valve. Transistors, provides two major modes of action, one, a switch turning current flow on and off, or two, act as an amplifier whereby an input current produces a larger output current.

An IC is nothing more than a number of these components connected together as a circuit all formed on the same substrate.

2.4. Atomic structure and band theory

Atoms are made up of positively charged - protons, neutral - neutrons, and negatively charged electrons. Protons and neutrons are relatively heavy particles with similar masses and electrons are relatively light particles with much lower mass. The nucleus of an atom is made up of protons and neutrons with electrons arrayed around it. The electrons occupy specific allowed energy levels with each level accommodating two electrons. Outer levels may also contain sub levels each capable of holding two electrons.

For silicon, the element of most interest to this publication, the outermost energy level has 3 sub levels with a total capacity of 6 electrons - two per sub...
level. The actual number of electrons occupying the outer most levels is 2, so there are 4 unfilled levels. In solid silicon, the outer most levels are filled by the four nearest silicon neighbor atoms sharing electrons, see figure 7.

![Silicon covalent bonds](image)

**Figure 7. Silicon covalent bonds**

When atoms exist in a solid, the proximity of electrons on neighboring atoms leads to a further splitting of allowed energy levels. The outer most energy levels split into two bands of closely spaced energy levels known as the conduction and the valence bands. The two bands are separated from each other by a gap in the allowed energy levels. The width of the energy gap determines the electrical conduction of the material. A large gap of several electron volts (eV) results in an insulating material, a gap of a few tenths of an eV to a few eV is a semiconductor, and a material with overlapping bands is a conductor.

The reason that the energy gap determines the conduction properties of the material is as follows. At zero degrees kelvin (minus 273 centigrade), the valence band is completely filled with electrons and the conduction band is empty. At room temperature the electrons in the material have gained thermal energy. If the energy gap is large, very few if any electrons will have sufficient thermal energy to jump to the conduction band where the electrons are free to move and participate in conduction. If the gap is small or non-existent, many electrons will jump to the conduction band and be free to move.

### 2.5. Intrinsic semiconductors

Materials with energy gaps of a few tenths of an eV to a few eV at room temperature will contain electrons that have jumped from the valence band to the conduction band due to thermal energy. The electrons in the conduction band will be free to move about in the material participating in conduction. In addition, when the electrons jump from the valence band they leave behind a missing electron or hole in the valence band. Electrons from neighboring atoms may move into the hole left behind by the escaping electron. The hole will then effectively have moved to a neighboring atom where it may move once again by another electron from a neighboring atom filling it. Since atoms with an associated hole in the valence band have one more proton than electron, the hole has an effective positive charge equal and opposite to the charge on an electron (atoms without holes have equal numbers of protons and electrons, i.e., they are electrically neutral). Each electron thermally excited into the conduction band results in a mobile negative electron in the conduction band and a mobile positive hole in the valence band. Although holes are less mobile than electrons, they may be thought of as positively charged particles free to move in their own right. Thermally generated holes and electrons are called intrinsic carriers and materials with significant numbers of thermally generated carriers at room temperature are called intrinsic semiconductors.

---

**Energy gap effect on conduction**

- **Insulators** - energy gaps of several eV.
- **Semiconductors** - energy gaps of a few tenths of an eV to a few eV.
- **Conductors** - overlapping energy bands with no energy gap.

**Intrinsic semiconductors**

- **Holes** - missing electrons in the valence band that “move” by capturing electrons from neighboring atoms thereby switching the hole to the neighboring atom.
- **Intrinsic carriers** - electron-hole pairs created by thermal energy.
- **Intrinsic semiconductor** - a material with a significant number of intrinsic carriers at room temperature.
2.6. **Extrinsic semiconductors**

Free electrons and holes may be created in a semiconductor material by introducing foreign elements known as dopants. Silicon is a group IV element in the periodic table. Group IV elements each have 4 electrons in an outer most energy level and are all semiconductors. Group III elements each have 3 electrons in the outer most energy level and group V elements each have 5 electrons in the outer most energy level. If a group III element is introduced into silicon, the group III element can bond with the four nearest silicon atoms leaving one missing electron or hole. If a group V element is introduced to silicon it can bond with the four nearest silicon atoms introducing an extra electron. Figure 8 illustrates a section of the periodic table containing group III, IV and V elements of interest.

![Figure 8. Periodic table section.](image)

Figure 9a illustrates a group III element boron, introduced into silicon. Boron introduces an unfilled energy level into the energy gap near the valence band that can “accept” an electron creating a hole. Figure 9b illustrates a group V element, phosphorus introduced into silicon. Phosphorus introduces an electron into the energy gap near the conduction band. The electron in the energy gap may easily be “donated” to the conduction band creating a free electron.

![Figure 9. Dopants in silicon.](image)

Group III dopants that accept electrons creating a hole are called p-type dopants for positive dopant. Group V elements that donate an electron are called n-type.

**Extrinsic semiconductors**

- dopant - an impurity that introduces holes or electrons to a semiconductor.
- p-type dopant - a dopant that accepts electrons creating holes, also called an acceptor.
- p-type - silicon doped to have more free holes than electrons.
- n-type dopant - a dopant that donates electrons, also called a donor.
- n-type - silicon doped to have more free electrons than holes.
- Majority carrier - the carrier type in the majority, holes for p-type silicon and electrons for n-type silicon.
- Minority carrier - the carrier type in the minority - electrons for p-type silicon and holes for n-type silicon.
n-type dopants. Throughout this text, p-type dopants will be indicated by green and n-type dopants will be indicated by purple.

If a sufficient number of dopants are introduced into a semiconductor to create more carriers of one type than the intrinsic carrier concentration, the material is said to be an extrinsic semiconductor. Extrinsic semiconductors have more of one type of carrier than the opposite type of carrier. Materials doped with p-type dopants are p-type semiconductors and have more holes than free electrons. Materials doped with n-type dopants are said to be n-type semiconductors and have more free electrons than holes. The carrier type, hole or electron in the majority is called the majority carrier and the type of carrier in the minority is called the minority carrier.

### 2.7. PN junction

When p-type silicon and n-type silicon are in contact, a PN junction results. Initially electrons flow from the n-type side of the junction to the p-type side of the junction and holes flow from the p-type side of the junction to the n-type side of the junction (opposite charges attract). Eventually the flow of carriers creates a voltage across the junction that prevents further current flow. If a voltage is applied across the pn junction with the positive terminal attached to the p-type side and the negative terminal attached to the n-type side and the voltage is higher than the built in voltage, the junction will turn “on” and electric current will flow. This is referred to as a forward biased junction, see figure 10 top. If a voltage is applied with the positive terminal attached to the n-type side and the negative terminal attached to the p-type side, the negative terminal will attract holes away from the junction and the positive terminal will attract electrons away from the junction resulting in an area at the junction depleted of mobile carriers called the depletion region. As the applied voltage is increased the depletion region will grow and no current will flow until the junction “breaks down” and current begins to flow. This is referred to as a reverse biased junction, see figure 10 bottom.

![Forward Biased Junction](image)

- **Depletion region** - a region “depleted” of mobile carriers.
- **Breakdown** - when a sufficiently high reverse bias is applied to a junction the junction will breakdown and conduct current.
- **Impact ionization** - carriers with sufficient energy impact atoms in the semiconductor exciting electrons into the conduction band creating electron-hole pairs.
- **Avalanche breakdown** - a kind of runaway condition where impact ionization creates electron-hole pairs that create new electron-hole pairs by further impact ionization.

**PN junction**
- Forward biased junction - a pn junction with a positive voltage applied to the p-type side of the junction and a negative voltage applied to the n-type side of the junction. The junction turns on if the applied voltage is greater than the built in voltage.
- Reverse bias - a pn junction with a negative voltage applied to the p-type side of the junction and a positive voltage applied to the n-type side of the junction.

![Reverse Biased Junction](image)

Figure 10. PN junction behavior.
The net result of the differing behavior for forward and reverse biased pn junction, is that the junction acts like a one way valve only allowing current flow in the forward direction. Diodes are frequently used in electric circuits to “steer” the current flow.

Although it may not be obvious, pn junction breakdown is a very important phenomena in ICs. The maximum voltage an IC can handle may be determined by junction breakdown so it is important to understand what determines the breakdown voltage of a pn junction.

Breakdown occurs when electrons and or holes in the semiconductor obtain enough energy to create new electron-hole pairs by colliding with atoms in the semiconductor exciting electrons into the conduction band, this is referred to as impact ionization. At a high enough applied voltage the electron-hole pairs created by impact ionization will have enough energy to create new electron -hole pairs creating a cascading runaway of electron-hole creation in the semiconductor. The carriers (electrons and holes) created by impact ionization will flood the depletion region collapsing it and causing a rapidly increasing current to flow, this is referred to as avalanche breakdown.

Most pn junctions have unequal doping on the p-type and n-type side. When a pn junction is reverses biased, a depletion region forms to oppose the applied voltage. The width of the depletion region depends on the applied voltage and the doping concentration in the semiconductor. The following rules determine the width of the depletion region:
1. The higher the applied voltage the thicker the depletion region will be until breakdown is reached.
2. For a given applied voltage the lower the doping concentration the wider the depletion region will be.
3. For a pn junction the region with the lowest doping level will have the widest depletion region, the highest energy carriers and will breakdown first leading to avalanche for the junction.

### 2.8. Bipolar transistors

A bipolar transistor is a three layer device with two pn junctions and three terminals- see figure 11. In figure 11 the transistor on the top is comprised of an n-type layer sandwiched between two p-type layers - a PNP, and the bottom transistor is comprised of a p-type layer sandwiched between two n-type layers - an NPN.

On the left side of the top of figure 11 is a p-type area referred to as the emitter, in the center is an n-type area referred to as the base and on the right side is a p-type area referred to as the collector. There is an electrical connection made to each of the three areas, emitter, collector and base referred to as terminals. For reasons that will be discussed shortly, the emitter is heavily doped p-type, the base is lightly doped n-type and of relatively narrow width and the collector is relatively lightly doped p-type. This type of bipolar transistor is called a PNP because an n-type layer is sandwiched between to p-type layers. There is also NPN transistors where a p-type layer is sandwiched between two n-type layers - see the bottom of figure 11.
In the preceding section it was shown that a forward biased pn junction would conduct current. If a positive voltage is applied to the emitter and a negative voltage is applied to the base of the PNP illustrated in figure 11 - see figure 12, then holes will be injected from the p-type emitter into the n-type base and electrons will be injected from the n-type base into the p-type emitter and current will flow. Under these conditions the emitter-base junction is forward biased. If the emitter is heavily doped relative to the base, under forward bias conditions the emitter will inject many more holes into the base than the base will inject into the emitter.

If a positive voltage is applied to the base and a negative voltage is applied to the collector, the collector base junction is reverse biased and no current will flow - see figure 13.

If the forward biased emitter-base junction exists in the same transistor with the reverse biased collector base junction, then the behavior of the transistor becomes really interesting. If the base of the bipolar transistor is relatively narrow and lightly doped, then holes injected into the base from the emitter may transit the base without recombining with electrons in the base and reach the collector-base junction. At the collector-base junction the holes are “collected” by the collector and current flows from

**Recombination**
- Recombination - when a free electron is captured by a hole annihilating the electron-hole pair.
emitter through the base to the collector. The current flowing from emitter terminal to collector terminal may be controlled by the current at the base terminal and the “gain” of the transistor is given by:

\[
gain = \frac{I_C}{I_B}
\]  

(2)

where, \(I_C\) is the collector current and \(I_B\) is the base current.

\[ \text{E} \rightarrow \text{B} \rightarrow \text{C} \]

\[ \text{P Type} \rightarrow \text{N Type} \rightarrow \text{P Type} \]

Figure 13. Reverse biased collector-base junction.

The configuration illustrated in figure 14 shows how a small signal connected to a base terminal may be used to modulate the current flowing from emitter to base. If a transistor is properly biased a small signal at the base terminal will be amplified at the collector terminal.

\[ \text{E} \rightarrow \text{B} \rightarrow \text{C} \]

\[ \text{P Type} \rightarrow \text{N Type} \rightarrow \text{P Type} \]

Figure 14. Bipolar transistor conducting.

A bipolar transistor is called bipolar because both carrier types, electrons and holes participate in conduction. In the section 2.10 a unipolar device the MOSFET will be presented. A MOSFET is a device in which only majority carriers participate in conduction.

2.9. MOS Capacitor

The MOS capacitor is the simplest MOS structure. A basic capacitor is formed when two conductive layers are separated by an insulating dielectric layer. If a voltage is applied across a capacitor, opposite charges accumulate on the two conductive layers and are stored there because the insulating layer prevents current flow from conductive layer to conductive layer. Eventually capacitors lose their stored charge by leakage through the dielectric, but for some period of time the capacitor will store charge that can later be released from the capacitor. In this case a capacitor is acting like a battery. The measure of charge a capacitor can store is the capacitance and is given by:

Bipolar Transistor 2

- Gain - the ratio of the collector current to the base current. Gain is often referred to as Beta.
- Bipolar - a device in which both carrier types, holes and electrons take part in conduction.
where $C$ is the capacitance, $k$ is the dielectric constant, $A$ is the area of the capacitor conductors and $d$ is the thickness of the dielectric layer. The $k$ value of a dielectric is an intrinsic property of the dielectric. From equation 2 it can be seen that capacitance increases with increasing dielectric constant $k$, increasing area $A$, or decreasing dielectric thickness $d$.

For a “standard” capacitor with two conductive plates capacitance is a fixed value versus voltage, however, for an MOS capacitor capacitance varies with voltage. Figure 15 illustrates an MOS capacitor under different voltages and polarities.

MOS Capacitors

- Capacitance - the capacity of a structure to store electric charge.
- Accumulation - an increase in concentration of carriers in a region of a semiconductor brought about by an applied electric field.
- Depletion - a reduction in the concentration of free carriers in a region of a semiconductor brought about by an applied electric field.
- Depletion region or layer - a region of a semiconductor in which no free carriers are present.
- Inversion - when a region of a semiconductor switches type from the majority carrier type to the minority carrier type due to accumulation of minority carriers brought about by an applied electric field.

In figure 15, the MOS capacitor consists of a metal top conductor, a silicon dioxide dielectric and an n-type silicon bottom “conductor”. When a voltage is applied with the positive terminal attached to the top conductor - figure 15a, the majority n-type carriers are attracted to the silicon-silicon dioxide interface (unlike charges attract). Under these conditions, the positive charges on the metal conductor and the negative charges in the silicon are separated by the minimum distance and from equation 2 this condition leads to the maximum capacitance determined by the silicon dioxide thickness, dielectric constant and the area of the top metal conductor.
If a relatively low voltage is applied to the capacitor with the negative terminal attached to the top conductor - figure 15b, the negative carriers in the silicon are repelled by the negative charge on the conductor plate (like charges repel). The majority n-type carriers are repelled from the silicon-silicon dioxide interface leaving behind a layer depleted of charges - a depletion layer forms. The depletion layers acts like a dielectric (insulator) due to the lack of availability of charges to conduct current. The charge on the top metal conductor and the charge in the silicon are now separated by a greater distance than just the thickness of the oxide. Since capacitance goes down as the thickness of the dielectric increases (in this case the effective thickness, not the physical thickness) the capacitance decreases.

If a relatively higher voltage is applied to the capacitor with the negative terminal attached to the upper metal conductor - figure 15c, the few p-type minority carriers that exist in n-type silicon will be attracted to the silicon-silicon dioxide interface while the n-type carriers are repelled. With sufficient voltage and time, the surface of the n-type silicon will “invert” to p-type. Since the carriers are now once again separated only by the thickness of the dielectric layer, the capacitance will again reach a maximum value for the thickness of dielectric, dielectric constant and area of the capacitor. Figure 16 illustrates capacitance versus voltage for the capacitor illustrated in figure 15.

![Capacitance Voltage Plot](image)

**Figure 16. Capacitance voltage plot.**

A key concept from figure 16 is the threshold voltage indicated by $V_T$ on the voltage axis. The threshold voltage is the voltage required to invert the surface of the semiconductor. Threshold voltage is effected by the type of metal used for the conductive top plate, the dielectric thickness and dielectric constant and the amount of doping and therefore carrier concentration in the silicon. For a common system of aluminum - silicon dioxide - silicon, the thinner the silicon dioxide layer and the lower the silicon doping the lower the threshold voltage. Conversely, thicker silicon dioxide layers with heavier doping levels have higher threshold voltages. Threshold voltage is a critical concept for other MOS devices that will be discussed in the next section.

The entire preceding discussion has centered around n-type silicon. P-type silicon exhibits the same behavior except that p-type silicon accumulates when a negative voltage is applied to the top metal conductor and inverts when a positive voltage is applied to the top conductor, the exact opposite of n-type silicon behavior versus polarity.

---

**Capacitance voltage plot**

- Capacitance voltage plot - a plot of capacitance versus voltage for an MOS capacitor. Also called a CV plot.
- Threshold voltage - the voltage at which inversion begins.
2.10. The MOSFET
The Metal Oxide Semiconductor Field Effect Transistor or MOSFET is the basic device building block underlying the vast majority of ICs produced in the world today - see figure 17.

**Figure 17. Relative market share of different technologies.**

In a MOSFET, an MOS capacitor is used to apply an electric field that controls the transistor switching. The structure of a MOSFET comes in two basic types, NMOS and PMOS - see figure 18.

**Figure 18. NMOS and PMOS MOSFETs.**

---

### The MOSFET
- MOSFET - Metal Oxide Semiconductor Field Effect Transistor
- NMOS - a MOSFET that uses n-type carriers for conduction. NMOS is made up of an n-type source region and an n-type drain region separated by a p-type region.
- PMOS - a MOSFET that uses p-type carriers for conduction. PMOS is made up of a p-type source region and a p-type drain region separated by an
In an NMOS device - figure 18 top, two highly doped n-type areas are separated by a lightly doped p-type with a conductive “gate” over an insulating oxide over the p-type region. In a PMOS device - figure 18 bottom, two highly doped p-type areas are separated by a lightly doped n-type area.

2.10.1. MOSFET off-state
If a positive voltage is applied to the drain of an NMOS device and the gate and source are tied to ground, then the drain to body pn junction is reverse biased and no current flows - see figure 19.

The doping of the body of the MOSFET is very important to a number of operating parameters of the device. The doping right under the gate oxide determines the threshold voltage of the device. The doping needs to be light if the device is to have a low threshold voltage - a very desirable quality. On the other hand, low doping in the body of the device means the depletion region spreads further for a given applied voltage. If the depletion region spreads all the way to the source, the device “punches through” and breaks down. For very small MOSFETs, punch-through breakdown limits the operating voltage of the device to a value lower than the expected avalanche breakdown voltage. A great deal of device engineering that will be discussed in the later chapters is dedicated to punch-through control.

2.10.2. MOSFET on-state
If a positive voltage is applied to the drain of an NMOS device, the source is grounded and the gate also has a positive voltage applied greater than the threshold voltage, the surface under the gate will invert connecting the source and drain and allowing current to flow - see figure 20. In figure 20 the n-type source and drain areas become connected by an n-type inversion layer, so only n-type carriers participate in conduction making a MOSFET a unipolar device.

2.11. MOSFET scaling
In figure 17 it was shown that MOSFETs are the dominant type of IC in production today. The reason that MOSFETs have taken over from Bipolar devices is simply put, MOSFETs scale and Bipolar transistors don’t.
Chapter 2 - Introduction to integrated circuit devices

At the 1974 International Electron Devices Meeting Dennard, et.al., presented the classic paper on scaling “Design of Ion-Implanted MOSFET's with Very Small Physical Dimensions”. Dennard's group had discovered that if a MOSFET's physical dimensions were scaled down while maintaining a constant electric field every other MOSFET performance parameter improved. For example, if a 250nm gate length MOSFET is scaled by 1.4, the new MOSFET has a gate length of 250/1.4 ~ 180nm. The operating voltage of 1.8 volts is also scaled by 1.4 to achieve a constant electric field and the new MOSFET operates at 1.8/1.4 ~ 1.3 volts. The resulting MOSFET is 1/2 the size, 1.4 times as fast and consumes 1/2 the power.

2.12. Conclusion
In this chapter the four basic device type commonly used in integrated circuits have been presented. In the next chapter we will discuss how NMOS and PMOS devices are used together to make CMOS (Complementary MOS), the structures required for CMOS and the issues that arise as CMOS is scaled to submicron geometries.

Scaling
- Gate length - the distance between on side of the gate electrode and the other side in the direction defined by a line drawn from drain to source of the MOSFET.
- Electric field - voltage per unit length. For example, the voltage applied between drain and source of a MOSFET divided by the distance between the drain and source of the MOSFET.
Notes on IC Device Structures

1) Si Contacts

We need to transfer charge carriers from device to device in integrated circuits through interconnects. The interface between the interconnect material (e.g., Al, poly-Si, or other metals) and the Si substrate can form either rectifying contacts or "ohmic" contacts. The term “ohmic” refers to an interface which has negligible resistance compared with the parasitic resistance of the Si device. You can consult Chapter 3 of Muller and Kamins for the physical principles of these contacts. The following contact combinations are often used in IC structures:

a) Tunneling Ohmic Contact

This is the most common ohmic contact structure used in ICs. The heavily-doped (> $10^{20}$/cm$^3$) p+ and n+ regions are specially fabricated to facilitate a thin energy barrier between the metal and the Si substrate. The majority carriers can tunnel easily through this thin barrier with low impedance. Note that the p+/p or n+/n interface is also highly conductive.

b) Schottky Contacts

Schottky contacts can be formed by metals deposited on lightly doped Si. The thermionic emission energy barrier for majority carrier conduction is called the Schottky barrier height. With the same metal, the sum of the Schottky barrier heights for n-substrate and p-substrate is equal to the energy gap (i.e., $\phi_{Bn} + \phi_{Bp} = E_g$). For example, Al/n-Si contact has $\phi_{Bn} \approx 0.75$eV which will form a rectifying contact. Thus, Al/p-Si will have $\phi_{Bp} \approx 0.37$ eV which will form a Schottky “ohmic” contact. The later is called “ohmic” because the barrier height is so low that the conducting impedance is small compared with that of the Si device.

c) Poly-Si Contacts

Heavily-doped poly-Si (either n+ or p+ with doping concentrations > $10^{20}$/cm$^3$) is used as the gate material and also as an interconnect material in MOS integrated circuits. It is also used a dopant source to diffuse dopants into the underlying Si substrate if a high-temperature diffusion cycle is used in the process. The n+ poly-Si/n-Si, p+ poly-Si/p-Si and Al/ n+ (or p+) poly-Si contacts are all ohmic.
2) Device Isolation Structures

There major techniques can be used to isolate current conduction between device through the substrate:

a) PN Junction isolation

Most IC devices requires this junction isolation technique: diffuse resistors, BJT, CMOS etc.. To ensure the pn junctions are reverse biased, the substrate has to be connected to the most negative potential of the IC for p-substrate and most positive potential in the case of n-substrate. The state-of-the-art junction isolation leakage current density is $< 1\text{nA/cm}^2$ at a reverse bias of -5V.

[Note] MOSFETs using inversion layer to control channel conduction has build-in junction isolation which prevents the inversion charges to flow to the substrate.

b) Oxide Isolation

The lateral conduction between devices is isolated by SiO$_2$.

[Note] MOS Channel Stop Isolation
MOSFET are usually having lateral isolation by thick field oxides (FOX). However, we may have metal interconnects carrying high voltages lying on top of the field oxide region. To avoid charge inversion through the parasitic MOS structure (i.e., metal/FOX/substrate), the region below the field oxide is doped to a higher concentration to raise the threshold voltage much higher than the supply voltage.

c) Silicon-on-Insulator

This is the ideal device isolation technique. High cost of fabricating SOI substrates is the limitation.
NPN Bipolar Transistor
with n+ subcollector

N-well CMOS Inverter

THE BIPOLAR PROCESS

Figure 5-9: Layout (a) and cross-section of the complete nnp transistor (b).

N-WELL CMOS TECHNOLOGY

Figure 5-12: CMOS Inverter. Composite layout (a), cross-section (b), and electrical diagram (c).
Polysilicon gate $V_G$

Source electrode $V_S$

Channel stop

$p^+$ Source

$n^+$ Drain

Bulk electrode $V_B$

Structural perspective layout (a)

Field oxide (FOX)

Drain electrode $V_D$

$n^+$ Source

$p^+$

$p^+$

$n^+$

Channel stop

$p^+$

$n^+$

Ion-implanted n-type channel layer

Bulk electrode $V_B$

Basic D-mode MOSFET structure (a)

Poly gate

$n^+$ (Source)

Source electrode (Metal)

$n^+$ (Drain)

Drain electrode (Metal)

Top view (b)

$n$-channel enhancement-mode MOSFET.

Circuit symbol (c)

$G + V_{GS}$

$B + V_{DS}$

$W/L$

$D + V_{DS}$

$I_D$

$I_N$

$V_{DS}$

$V_{GS}$

$V_{SS}$

Depletion-mode MOSFET.
Simple process to make an ohmic contact (Al/n⁺-Si)

1. Starting substrate
2. Oxide growth, resist application and soft-bake
3. Ultraviolet light exposure
4. Oxide etch
5. Stripping resist, dope through window
6. Drive-in or implant anneal with oxide growth
7. Oxide contact cut
8. Evaporation of Al, application and patterning of resist
9. Aluminium etch
10. Final structure

* The lithographic steps are shown here
Process Flow of NMOS Inverter, with both Enhancement and Depletion MOSFETs.
Double Poly Process Flow

Initial active area

(a)

Poly 1

Gate oxide

FOX

p`

p, p`

After gate oxide growth and first-level poly (poly 1) deposition/patterning

(b)

Oxide etch

(c)

Oxide coating

Gate oxide (for poly II)

FOX

p`

p, p`

Oxidation for second poly MOSFET gate oxide

(d)

Poly II deposition/patterning

(e)

Oxide etch, n* implant

CMOS Process Flow

For pMOS bulk connection

n-tub formation

(a)

After LOCOS

(b)

After poly deposition/patterning and drain/source implantation

(c)

Final structure

(d)

CMOS process flow.
In this chapter we begin to discuss the specific mechanisms by which current flows in a solid. In examining these mechanisms we shall learn why some materials are good conductors of electric current, whereas others are poor conductors. We shall see how the conductivity of a semiconductor can be varied by changing the temperature or the number of impurities. These fundamental concepts of charge transport form the basis for later discussions of solid state device behavior.

In Chapter 2 we found that electrons are restricted to sets of discrete energy levels within atoms. Large gaps exist in the energy scale in which no energy states are available. In a similar fashion, electrons in solids are restricted to certain energies and are not allowed at other energies. The basic difference between the case of an electron in a solid and that of an electron in an isolated atom is that in the solid the electron has a range, or band, of available energies. The discrete energy levels of the isolated atom spread into bands of energies in the solid because in the solid the wave functions of electrons in neighboring atoms overlap, and an electron is not necessarily localized at a particular atom. Thus, for example, an electron in the outer orbit of one atom feels the influence of neighboring atoms, and its overall wave function is altered. Naturally, this influence affects the potential energy term and the boundary conditions in the Schrodinger equation, and we would expect to obtain different energies in the solution. Usually, the influence of neighboring atoms...
on the energy levels of a particular atom can be treated as a small perturbation, giving rise to shifting and splitting of energy states into energy bands.

### 3.1.1 Bonding Forces in Solids

The interaction of electrons in neighboring atoms of a solid serves the very important function of holding the crystal together. For example, alkali halides such as NaCl are typified by ionic bonding. In the NaCl lattice, each Na atom is surrounded by six nearest neighbor Cl atoms, and vice versa. Four of the nearest neighbors are evident in the two-dimensional representation shown in Fig. 3-1. The electronic structure of Na (Z = 11) is [Ne]3s^1, and Cl (Z = 17) has the structure [Ne]3s^23p^5. In the lattice each Na atom gives up its outer 3s electron to a Cl atom, so that the crystal is made up of ions with the electronic structures of the inert atoms Ne and Ar (Ar has the electronic structure [Ne]3s^23p^6). However, the ions have net electric charges after the electron exchange. The Na^+ ion has a net positive charge, having lost one electron, and the Cl^- ion has a net negative charge, having gained an electron.

Each Na^+ ion exerts an electrostatic attractive force upon its six Cl^- neighbors, and vice versa. These coulombic forces pull the lattice together until a balance is reached with repulsive forces. A reasonably accurate calculation of the atomic spacing can be made by considering the ions as hard spheres being attracted together (Prob. 3.1).

An important observation in the NaCl structure is that all electrons are tightly bound to atoms. Once the exchange has been made between the Na and Cl atoms to form the Na^+ and Cl^- ions, the outer orbits of all atoms are completely filled. Since the ions have the closed-shell configurations of the inert atoms Ne and Ar, there are no loosely bound electrons to participate in current flow, as a result, NaCl is a good insulator.

![Figure 3-1](image)

**Figure 3-1** The NaCl lattice: an example of ionic bonding.

In a metal atom the outer electronic shell is only partially filled, usually by no more than three electrons. We have already noted that the alkali metals (e.g., Na) have only one electron in the outer orbit. This electron is loosely bound and is given up easily in ion formation. This accounts for the great chemical activity of the alkali metals, as well as for their high electrical conductivity. In the metal the outer electron of each alkali atom is contributed to the crystal as a whole, so that the solid is made up of ions with closed shells immersed in a sea of free electrons. The forces holding the lattice together arise from an interaction between the positive ion cores and the surrounding free electrons. This is one type of metallic bonding. Obviously, there are complicated differences in the bonding forces for various metals, as evidenced by the wide range of melting temperatures (234 K for Hg, 3643 K for W). However, the metals have the sea of electrons in common, and these electrons are free to move about the crystal under the influence of an electric field.

A third type of bonding is exhibited by the diamond lattice semiconductors. We recall that each atom in the Ge, Si, or C diamond lattice is surrounded by four nearest neighbors, each with four electrons in the outer orbit. In these crystals each atom shares its valence electrons with its four neighbors (Fig. 3-2). Bonding between nearest neighbor atoms is illustrated in the diamond lattice diagram of Fig. 1-9. The bonding forces arise from a quantum mechanical interaction between the shared electrons. This is known as covalent bonding; each electron pair constitutes a covalent bond. In the sharing process it is no longer relevant to ask which electron belongs to a particular atom—both belong to the bond. The two electrons are indistinguishable, except that they must have opposite spin to satisfy the Pauli exclusion principle. Covalent bonding is also found in certain molecules, such as H₂.

As in the case of the ionic crystals, no free electrons are available to the lattice in the covalent diamond structure of Fig. 3-2. By this reasoning Ge and Si should also be insulators. However, we have pictured an idealized lattice at 0 K in this figure. As we shall see in subsequent sections, an electron can be thermally or optically excited out of a covalent bond and thereby become free to participate in conduction. This is an important feature of semiconductors.

![Figure 3-2](image)

**Figure 3-2** Covalent bonding in the Si crystal, viewed along a (100) direction (see also Figs. 1-8 and 1-9).
Compound semiconductors such as GaAs have mixed bonding, in which both ionic and covalent bonding forces participate. Some ionic bonding is to be expected in a crystal such as GaAs because of the difference in placement of the Ga and As atoms in the periodic table. The ionic character of the bonding becomes more important as the atoms of the compound become further separated in the periodic table, as in the II–VI compounds.

3.1.2 Energy Bands

As isolated atoms are brought together to form a solid, various interactions occur between neighboring atoms, including those described in the preceding section. The forces of attraction and repulsion between atoms will find a balance at the proper interatomic spacing for the crystal. In the process, important changes occur in the electron energy level configurations, and these changes result in the varied electrical properties of solids.

Qualitatively, we can see that as atoms are brought together, the application of the Pauli exclusion principle becomes important. When two atoms are completely isolated from each other so that there is no interaction of electron wave functions between them, they can have identical electronic structures. As the spacing between the two atoms becomes smaller, however, electron wave functions begin to overlap. The exclusion principle dictates that no two electrons in a given interacting system may have the same quantum state, thus there must be a splitting of the discrete energy levels of the isolated atoms into new levels belonging to the pair rather than to individual atoms.

In a solid, many atoms are brought together, so that the split energy levels form essentially continuous bands of energies. As an example, Fig. 3-3 illustrates the imaginary formation of a diamond crystal from isolated carbon atoms. Each isolated carbon atom has an electronic structure $1s^22s^22p^2$ in the ground state. Each atom has available two $2s$ states, six $2p$ states, and higher states (see Tables 2-1 and 2-2). If we consider $N$ atoms, there will be $2N, 2N_1$ states of type $1s, 2s$, and $2p$, respectively. As the interatomic spacing decreases, these energy levels split into bands, beginning with the outer $(n = 2)$ shell. As the “2s” and “2p” bands grow, they merge into a single band composed of a mixture of energy levels. This band of “2s–2p” levels contains $8N$ available states. As the distance between atoms approaches the equilibrium interatomic spacing of diamond, this band splits into two bands separated by an energy gap $E_g$. The upper band (called the conduction band) contains $4N$ states, as does the lower (valence) band. Thus, apart from the lowly and tightly bound “1s” levels, the diamond crystal has two bands of available energy levels separated by an energy gap $E_g$ wide, which contains no allowed energy levels for electrons to occupy. This gap is sometimes called the “forbidden band,” since in a perfect crystal it contains no electron energy states.

We should pause at this point and count electrons. The lower “1s” band is filled with the $2N$ electrons which originally resided in the collective 1s states of the isolated atoms. However, there were 4N electrons in the original isolated $n = 2$ shells ($2N$ in $2s$ states and $2N$ in $2p$ states). These $4N$ electrons must occupy states in the valence band or the conduction band in the crystal. At 0 K the electrons will occupy the lowest energy states available to them. In the case of the diamond crystal, there are exactly $4N$ states in the valence band available to the $4N$ electrons. Thus at 0 K, every state in the valence band will be filled, while the conduction band will be completely empty of electrons. As we shall see, this arrangement of completely filled and empty energy bands has an important effect on the electrical conductivity of the solid.

3.1.3 Metals, Semiconductors, and Insulators

Every solid has its own characteristic energy band structure. This variation in band structure is responsible for the wide range of electrical characteristics observed in various materials. The diamond band structure of Fig. 3-3, for example, can give a good picture of why carbon in the diamond lattice is a good insulator. To reach such a conclusion, we must consider the properties of completely filled and completely empty energy bands in the current conduction process.

Before discussing the mechanisms of current flow in solids further, we can observe here that for electrons to experience acceleration in an applied electric field, they must be able to move into new energy states. This implies there must be empty states (allowed energy states which are not already occupied by electrons) available to the electrons. For example, if relatively few electrons reside in an otherwise empty band, ample unoccupied states are available into
which the electrons can move. On the other hand, the diamond structure is such that the valence band is completely filled with electrons at 0 K and the conduction band is empty. There can be no charge transport within the valence band, since no empty states are available into which electrons can move. There are no electrons in the conduction band, so no charge transport can take place there either. Thus carbon in the diamond structure has a high resistivity typical of insulators.

Semiconductor materials at 0 K have basically the same structure as insulators — a filled valence band separated from an empty conduction band by a band gap containing no allowed energy states (Fig. 3-4). The difference lies in the size of the band gap $E_g$, which is much smaller in semiconductors than in insulators. For example, the semiconductor Si has a band gap of about 1.1 eV compared with 5 eV for diamond. The relatively small band gaps of semiconductors (Appendix III) allow for excitation of electrons from the lower (valence) band to the upper (conduction) band by reasonable amounts of thermal or optical energy. For example, at room temperature a semiconductor with a 1-eV band gap will have a significant number of electrons excited thermally across the energy gap into the conduction band, whereas an insulator with $E_g = 10$ eV will have a negligible number of such excitations. Thus an important difference between semiconductors and insulators is that the number of electrons available for conduction can be increased greatly in semiconductors by thermal or optical energy.

In metals the bands either overlap or are only partially filled. Thus electrons and empty energy states are intermixed within the bands so that electrons can move freely under the influence of an electric field. As expected from the metallic band structures of Fig. 3-4, metals have a high electrical conductivity.

3.1.4 Direct and Indirect Semiconductors

The "thought experiment" of Section 3.1.2, in which isolated atoms were brought together to form a solid, is useful in pointing out the existence of energy bands and some of their properties. Other techniques are generally used, however, when quantitative calculations are made of band structures. In a typical calculation, a single electron is assumed to travel through a perfectly periodic lattice. The wave function of the electron is assumed to be in the form of a plane wave moving, for example, in the $x$-direction with propagation constant $k$, also called a wave vector. The space-dependent wave function for the electron is

$$\psi(x) = U(k_x, x)e^{ik_xx}$$

where the function $U(k_x, x)$ modulates the wave function according to the periodicity of the lattice.

In such a calculation, allowed values of energy can be plotted vs. the propagation constant $k$. Since the periodicity of most lattices differs in various directions, the $(E, k)$ diagram must be plotted for the various crystal directions, and the full relationship between $E$ and $k$ is a complex surface which should be visualized in three dimensions.

The band structure of GaAs has a minimum in the conduction band and a maximum in the valence band for the same $k$ value ($k = 0$). On the other hand, Si has its valence band maximum at a different value of $k$ than its conduction band minimum. Thus an electron making a smallest-energy transition from the conduction band to the valence band in GaAs can do so without a change in $k$ value, on the other hand, a transition from the maximum point in the Si conduction band to the maximum point of the valence band requires some change in $k$. Thus there are two classes of semiconductor energy bands: direct and indirect (Fig. 3-5). We can show that an indirect transition, involving a change in $k$, requires a change of momentum for the electron.

---

*Discussion of plane waves are available in most sophomore physics texts or in introductory electromagnetics texts.*
EXAMPLE 3.1
Assuming that \( U \) is constant in Eq. (3-1) for an essentially free electron, show that the \( x \)-component of the electron momentum in the crystal is given by \( \langle p_x \rangle = \hbar k \).

SOLUTION
From Eq. (3-1),
\[
\psi_h(x) = U e^{ikx}.
\]
Using Eq. (2-21b) and the momentum operator,
\[
\langle p_x \rangle = \int_{-\infty}^{\infty} U e^{-ikx} \frac{\partial}{\partial x} e^{ikx} dx = \int_{-\infty}^{\infty} U dx \cdot \frac{\partial}{\partial x} \int_{-\infty}^{\infty} U dx = \hbar k.
\]
This result implies that \((E, k)\) diagrams such as shown in Fig. 3-5 can be considered plots of electron energy vs momentum, with a scaling factor \( \hbar \).

The direct and indirect semiconductors are identified in Appendix III. In a direct semiconductor such as GaAs, an electron in the conduction band can fall to an empty state in the valence band, giving off the energy difference \( E_v \) as a photon of light. On the other hand, an electron in the conduction band minimum of an indirect semiconductor such as Si cannot fall directly to the valence band maximum but must undergo a momentum change as well as changing its energy. For example, it may go through some defect state \((E_v)\) within the band gap. We shall discuss such defect states in Sections 4.21 and 4.32. In an indirect transition which involves a change in \( k \), the energy is generally given up as heat to the lattice rather than as an emitted photon. This difference between direct and indirect band structures is very important for deciding which semiconductors can be used in devices requiring light output. For example, semiconductor light emitters (Section 6.4) and lasers (Chapter 10) generally must be made of materials capable of direct band-to-band transitions or of indirect materials with vertical transitions between defect states.

Band diagrams such as those shown in Fig. 3-5 are cumbersome to draw in analyzing devices, and do not provide a view of the variation of electron energy with distance in the sample. Therefore, in most discussions we shall use simple band pictures such as those shown in Fig. 3-4, remembering that electron transitions across the band gap may be direct or indirect.

3.1.5 Variation of Energy Bands with Alloy Composition
As III–V ternary and quaternary alloys are varied over their composition ranges (see Sections 1.2.4 and 1.4.1), their band structures change. For example, Fig. 3-6 illustrates the band structure of GaAs and AlAs, and the way in which the bands change with composition \( x \) in the ternary compound Al\(_x\)Ga\(_{1-x}\)As. The binary compound GaAs is a direct material, with a band gap of 1.43 eV at room temperature. For reference, we call the direct \((k = 0)\) conduction band minimum \( \Gamma \). There are also two higher-lying indirect minima in the GaAs conduction band, but these are sufficiently far above \( \Gamma \) that few electrons reside there (we discuss an important exception in Chapter 12 in which high-field excitation of electrons into the indirect minima leads to the Gunn effect). We call the lowest-lying GaAs indirect minimum \( L \) and the other \( X \). In AlAs the direct \( \Gamma \) minimum is much higher than the indirect \( X \) minimum, and this material is therefore indirect with a band gap of 2.16 eV at room temperature.

In the ternary alloy Al\(_x\)Ga\(_{1-x}\)As all of these conduction band minima move up relative to the valence band as the composition \( x \) varies from 0 (GaAs) to 1 (AlAs). However, the indirect minimum \( X \) moves up less than the others, and for compositions above about 38 percent Al this indirect minimum becomes the lowest-lying conduction band. Therefore, the ternary alloy AlGaAs is a direct semiconductor for Al compositions on the column III sublattice up to about 38 percent, and is an indirect semiconductor for higher Al mole fractions. The band gap energy \( E_v \) is shown in color on Fig. 3-6(c).

The variation of energy bands for the ternary alloy GaAs\(_{1-x}\)P, is generally similar to that of AlGaAs shown in Fig. 3-6. GaAsP is a direct semiconductor from GaAs to about GaAs\(_{0.5}\)P, and is indirect from this composition to GaP (see Fig. 6-19). This material is often used in visible LEDs.

Figure 3-6
Variation of direct and indirect conduction bands in AlGaAs as a function of composition.
(a) The \((E, k)\) diagram for GaAs, showing three minima in the conduction band,
(b) AlAs band diagram.
The mechanism of current conduction is relatively easy to visualize in the case of a metal, the metal atoms are imbedded in a "sea" of relatively free electrons, and these electrons can move as a group under the influence of an electric field. This free electron view is oversimplified, but many important conduction properties of metals can be derived from just such a model. However, we cannot account for all of the electrical properties of semiconductors in this way. Since the semiconductor has a filled valence band and an empty conduction band at 0 K, we must consider the increase in conduction band electrons by thermal excitations across the band gap as the temperature is raised. In addition, after electrons are excited to the conduction band, the empty states left in the valence band can contribute to the conduction process. The introduction of impurities has an important effect on the energy band structure and on the availability of charge carriers. Thus there is considerable flexibility in controlling the electrical properties of semiconductors.

3.2 Electrons and Holes

As the temperature of a semiconductor is raised from 0 K, some electrons in the valence band receive enough thermal energy to be excited across the band gap to the conduction band. The result is a material with some electrons in an otherwise empty conduction band and some unoccupied states in an otherwise filled valence band (Fig. 3.7). For convenience, an empty state in the valence band is referred to as a hole. If the conduction band electron and the hole are created by the excitation of a valence band electron to the conduction band, they are called an electron–hole pair (abbreviated EHP).

After excitation to the conduction band, an electron is surrounded by a large number of unoccupied energy states. For example, the equilibrium number of electron–hole pairs in pure Si at room temperature is only about $10^{10}$ EHP/cm$^3$, compared to the Si atom density of more than $10^{25}$ atoms/cm$^3$. Thus the

Figure 3.7
Electron–hole pairs in a semiconductor.

---

1In Fig. 3.7 and in subsequent discussions, we refer to the bottom of the conduction band as $E_c$ and the top of the valence band as $E_v$. 
few electrons in the conduction band are free to move about via the many available empty states.

The corresponding problem of charge transport in the valence band is somewhat more complicated. However, it is possible to show that the effects of current in a valence band containing holes can be accounted for by simply keeping track of the holes themselves.

In a filled band, all available energy states are occupied. For every electron moving with a given velocity, there is an equal and opposite electron motion elsewhere in the band. If we apply an electric field, the net current is zero because for every electron $j$ moving with velocity $v_j$, there is a corresponding electron $j'$ with velocity $-v_j$. Figure 3-8 illustrates this effect in terms of the electron energy vs wave vector plot for the valence band. Since $k$ is proportional to electron momentum, it is clear the two electrons have oppositely directed velocities. With $N$ electrons/cm$^3$ in the band we express the current density using a sum over all of the electron velocities, and including the charge $-q$ on each electron. In a unit volume,

$$J = (-q) \sum_j v_j = 0 \quad \text{(filled band)} \quad (3-2a)$$

Now if we create a hole by removing the $j$th electron, the net current density in the valence band involves the sum over all velocities, minus the contribution of the electron we have removed.

$$J = (-q) \sum_j v_j - (-q)v_j \quad \text{(jth electron missing)} \quad (3-2b)$$

But the first term is zero, from Eq. (3-2a). Thus the net current is $+qv_j$. In other words, the current contribution of the hole is equivalent to that of a positively charged particle with velocity $v_j$, that of the missing electron. Of course, the charge transport is actually due to the motion of the now uncompensated electron ($j'$). Its current contribution $(-q)(-v_j)$ is equivalent to that of a positively charged particle with velocity $+v_j$. For simplicity, it is customary to treat empty states in the valence band as charge carriers with positive charge and positive mass.

In all the following discussions we shall concentrate on the electrons in the conduction band and on the holes in the valence band. We can account for the current flow in a semiconductor by the motion of these two types of charge carriers. We draw valence and conduction bands on an electron energy scale $E$, as in Fig. 3-8. However, we should remember that in the valence band, hole energy increases oppositely to electron energy, because the two carriers have opposite charge. Thus hole energy increases downward in Fig. 3-8 and holes, seeking the lowest energy state available, are generally found at the top of the valence band. In contrast, conduction band electrons are found at the bottom of the conduction band.

### 3.2.2 Effective Mass

The electrons in a crystal are not completely free, but instead interact with the periodic potential of the lattice. As a result, their "wave-particle" motion cannot be expected to be the same as for electrons in free space. Thus, in applying the usual equations of electrodynamics to charge carriers in a solid, we must use altered values of particle mass. In doing so, we account for most of the influences of the lattice, so that the electrons and holes can be treated as "almost free" carriers in most computations. The calculation of effective mass must take into account the shape of the energy bands in three-dimensional $k$-space, taking appropriate averages over the various energy bands.

Find the $(E, k)$ relationship for a free electron and relate it to the electron mass.

**EXAMPLE 3-2**

![Figure 3-8](image)

A valence band with all states filled, including states $j$ and $j'$, marked for discussion. The jth electron with wave vector $k_j$ is matched by an electron at $j'$ with the opposite wave vector $-k_j$. There is no net current in the band unless an electron is removed. For example, if the jth electron is removed, the motion of the electron at $j'$ is no longer compensated.

From Example 3-1, the electron momentum is $p = mv = \hbar k$. Then

$$E = \frac{1}{2}mv^2 = \frac{1}{2} \frac{p^2}{m} = \frac{\hbar^2}{2m} k^2$$

**SOLUTION**
Thus the electron energy is parabolic with wave vector \( k \). The electron mass is inversely related to the curvature (second derivative) of the \((E, k)\) relationship, since

\[
\frac{d^2E}{dk^2} = \frac{\hbar^2}{m}
\]

Although electrons in solids are not free, most energy bands are close to parabolic at their minima (for conduction bands) or maxima (for valence bands). We can also approximate effective mass near those band extrema from the curvature of the band.

The effective mass of an electron in a band with a given \((E, k)\) relationship is given by

\[
m^* = \frac{\hbar^2}{d^2E/dk^2}
\]

(3-3)

Thus the curvature of the band determines the electron effective mass. For example, in Fig. 3-6a it is clear that the electron effective mass in GaAs is much smaller in the direct \( \Gamma \) conduction band (strong curvature) than in the \( L \) or \( X \) minima (weaker curvature, smaller value in the denominator of the \( m^* \) expression).

For a band centered at \( k = 0 \) (such as the \( \Gamma \) band in GaAs), the \((E, k)\) relationship near the minimum is usually parabolic

\[
E = \frac{k^2}{2m^*} + E_\Gamma
\]

(3-4)

Comparing this relation to Eq. (3-3) indicates that the effective mass \( m^* \) is constant in a parabolic band. On the other hand, many conduction bands have complex \((E, k)\) relationships that depend on the direction of electron transport with respect to the principal crystal directions. In this case, the effective mass is a tensor quantity. However, we can use approximate averages over such bands in most calculations.

A particularly interesting feature of Figs. 3-5 and 3-6 is that the curvature \( d^2E/dk^2 \) is positive at the conduction band minima, but is negative at the valence band maxima. Thus, the electrons near the top of the valence band have negative effective mass, according to Eq. (3-3). Valence band electrons with negative charge and negative mass move in an electric field in the same direction as holes with positive charge and positive mass. As discussed in Section 3.2.1, we can fully account for charge transport in the valence band by considering hole motion.

In any calculations involving the mass of the charge carriers, we must use effective mass values for the particular material involved. Table 3-1 lists the effective masses for Ge, Si, and GaAs appropriate for one type of calculation. In this table and in all subsequent discussions, the electron effective mass is denoted by \( m_e^* \) and the hole effective mass by \( m_h^* \). The \( n \) subscript indicates the electron as a negative charge carrier, and the \( p \) subscript indicates the hole as a positive charge carrier.

### Table 3-1

<table>
<thead>
<tr>
<th>Material</th>
<th>( m_e^* )</th>
<th>( m_h^* )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ge</td>
<td>0.059</td>
<td>0.070</td>
</tr>
<tr>
<td>Si</td>
<td>0.039</td>
<td>0.060</td>
</tr>
<tr>
<td>GaAs</td>
<td>0.067</td>
<td>0.045</td>
</tr>
</tbody>
</table>

### 3.2.3 Intrinsic Material

A perfect semiconductor crystal with no impurities or lattice defects is called an intrinsic semiconductor. In such material there are no charge carriers at 0 K, since the valence band is filled with electrons and the conduction band is empty. At higher temperatures electron–hole pairs are generated as valence band electrons are excited thermally across the band gap to the conduction band. These EHPs are the only charge carriers in intrinsic material.

The generation of EHPs can be visualized in a qualitative way by considering the breaking of covalent bonds in the crystal lattice (Fig. 3-9). If one of the Si valence electrons is broken away from its position in the bonding structure such that it becomes free to move about in the lattice, a conduction electron is created and a broken bond (hole) is left behind. The energy required to break a bond is called the band gap energy \( E_g \). This model helps in visualizing the physical mechanism of EHP creation, but the energy band model is more productive for purposes of quantitative calculation. One important difficulty in the "broken bond" model is that the free electron and the hole seem deceptively localized in the lattice. Actually, the positions of the free electron and the hole are spread out over several lattice spacings and should be considered quantum mechanically by probability distributions (see Section 2.4).

Since the electrons and holes are created in pairs, the conduction band electron concentration \( n \) (electrons per \( cm^3 \)) is equal to the concentration of holes in the valence band \( p \) (holes per \( cm^3 \)). Each of these intrinsic carrier concentrations is commonly referred to as \( n \). Thus, for intrinsic material

\[
n = p = n_i
\]

(3-5)

![Figure 3-9](image-url)

_Electron–hole pairs in the covalent bonding model of the Si crystal._
At a given temperature there is a certain concentration of electron–hole pairs \( n \). Obviously, if a steady state carrier concentration is maintained, there must be recombination of EHPs at the same rate at which they are generated. Recombination occurs when an electron in the conduction band makes a transition (direct or indirect) to an empty state (hole) in the valence band, thus annihilating the pair. If we denote the generation rate of EHPs as \( g \), (EHP/cm²·s) and the recombination rate as \( r \), equilibrium requires that

\[
r = g. \tag{3-6}
\]

Each of these rates is temperature dependent. For example, \( g(T) \) increases when the temperature is raised, and a new carrier concentration \( n \) is established such that the higher recombination rate \( r(T) \) just balances generation. At any temperature, we can predict that the rate of recombination of electrons and holes \( r \) is proportional to the equilibrium concentration of electrons \( n \) and the concentration of holes \( p \):

\[
r = \alpha n_p p_n = \alpha n^2, \tag{3-7}
\]

The factor \( \alpha \) is a constant of proportionality which depends on the particular mechanism by which recombination takes place. We shall discuss the calculation of \( n \) as a function of temperature in Section 3.3.3, recombination processes will be discussed in Chapter 4.

### 3.2.4 Extrinsic Material

In addition to the intrinsic carriers generated thermally, it is possible to create carriers in semiconductors by purposefully introducing impurities into the crystal. This process, called doping, is the most common technique for varying the conductivity of semiconductors. By doping, a crystal can be altered so that it has a predominance of either electrons or holes. Thus there are two types of doped semiconductors, n-type (mostly electrons) and p-type (mostly holes). When a crystal is doped such that the equilibrium carrier concentrations \( n \) and \( p \) are different from the intrinsic carrier concentration \( n \), the material is said to be extrinsic.

When impurities or lattice defects are introduced into an otherwise perfect crystal, additional levels are created in the energy band structure, usually within the band gap. For example, an impurity from column V of the periodic table (P, As, and Sb) introduces an energy level very near the conduction band in Ge or Si. This level is filled with electrons at 0 K, and very little thermal energy is required to excite these electrons to the conduction band (Fig. 3-10). Thus at about 50–100 K virtually all of the electrons in the impurity level are "donated" to the conduction band. Such an impurity level is called a donor level, and the column V impurities in Ge or Si are called donor impurities. From Fig. 3-10 we note that the material doped with donor impurities can have a considerable concentration of electrons in the conduction band, even when the temperature is too low for the intrinsic EHP concentration to be appreciable. Thus semiconductors doped with a significant number of donor atoms will have \( n \) near \( n_p \) at room temperature. This is n-type material.
into the bonding structure of the lattice and is therefore loosely bound to the Sb atom. A small amount of thermal energy enables this extra electron to overcome its coulombic binding to the impurity atom and be donated to the lattice as a whole. Thus it is free to participate in current conduction. This process is a qualitative model of the excitation of electrons out of a donor level and into the conduction band (Fig. 3-10). Similarly, the column III impurity Al has only three valence electrons to contribute to the covalent bonding (Fig. 3-12), thereby leaving one bond incomplete. With a small amount of thermal energy, this incomplete bond can be transferred to other atoms as the bonding electrons exchange positions. Again, the idea of an electron "hopping" from an adjacent bond into the incomplete bond at the Al site provides some physical insight into the behavior of an acceptor, but the model of Fig. 3-11 is preferable for most discussions.

We can calculate rather simply the approximate energy required to excite the fifth electron of a donor atom into the conduction band (the donor binding energy). Let us assume for rough calculations that the Sb atom of Fig. 3-12 has its four covalent bonding electrons rather tightly bound and the fifth "extra" electron loosely bound to the atom. We can approximate this situation by using the Bohr model results, considering the loosely bound electron as ranging about the tightly bound "core" electrons in a hydrogen-like orbit. From Eq. (2-15) the magnitude of the ground-state energy (n = 1) of such an electron is

\[ E = \frac{m_q^4}{2K^2 R^2} \]  

(3-8)

The value of \( K \) must be modified from the free-space value 4\( \pi \varepsilon_0 \), used in the hydrogen atom problem to

\[ K = 4\pi \varepsilon_0 \varepsilon_r, \]  

(3-9)

where \( \varepsilon_r \) is the relative dielectric constant of the semiconductor material. In addition, we must use the effective mass \( m^* \) typical of the semiconductor.

EXAMPLE 3-3 Calculate the approximate donor binding energy for Ge (\( \varepsilon_r = 16, m^* = 0.12m_0 \))

SOLUTION From Eq. (3-8) and Appendix II we have

\[ E = \frac{m_q^4}{8(\varepsilon_0\varepsilon_r)\hbar^3} = \frac{0.12(9.11 \times 10^{-34})(1.6 \times 10^{-19})^4}{8(8.85 \times 10^{-12}) \times (16)(6.63 \times 10^{-34})^3} \]

\[ = 1.02 \times 10^{-22} \text{ J} = 0.0064 \text{ eV} \]

The "density of states" effective mass listed in Table 3-1 is not appropriate for this calculation. The value used here is as an average of the effective mass in different crystallographic directions, called the "conductivity effective mass."
band gaps. For example, Fig. 3-13 shows the spatial variation in conduction and valence bands for a multilayer structure in which a very thin layer of GaAs is sandwiched between two layers of AlGaAs, which has a wider band gap than the GaAs. We will discuss the details of such heterojunctions (junctions between dissimilar materials) in Section 5.8. It is interesting to point out here, however, that a consequence of confining electrons and holes in a very thin layer is that these particles behave according to the particle in a potential well problem, with quantum states calculated in Section 2.4.3. Therefore, instead of having the continuum of states normally available in the conduction band, the conduction band electrons in the narrow-gap material are confined to discrete quantum states as described by Eq. (2-33), modified for effective mass and finite barrier height. Similarly, the states in the valence band available for holes are restricted to discrete levels in the quantum well. This is one of the clearest demonstrations of the quantum mechanical results discussed in Chapter 2. From a practical device point of view, the formation of discrete quantum states in the GaAs layer of Fig. 3-13 changes the energy at which photons can be emitted. As electron on one of the discrete conduction band states \( E_i \) in Fig. 3-13 can make a transition to an empty discrete valence band state in the GaAs quantum well (such as \( E_h \)), giving off a photon of energy \( E_i + E_h + E_\text{ph} \), greater than the GaAs band gap. Semiconductors lasers have been made in which such quantum wells are used to raise the energy of the transition from the infrared, typical of GaAs, to the red portion of the spectrum. We will see other examples of quantum wells in semiconductor devices in later chapters, including applications of tunneling of electrons through such quantum states.

In calculating semiconductor electrical properties and analyzing device behavior, it is often necessary to know the number of charge carriers per cm\(^2\) in the material. The majority carrier concentration is usually obvious in heavily doped material, since one majority carrier is obtained for each impurity atom (for the standard doping impurities). The concentration of minority carriers is not obvious, however, nor is the temperature dependence of the carrier concentrations.

To obtain equations for the carrier concentrations we must investigate the distribution of carriers over the available energy states. This type of distribution is not difficult to calculate, but the derivation requires some background in statistical methods. Since we are primarily concerned here with the application of these results to semiconductor materials and devices, we shall accept the distribution function as given for this discussion.

### 3.3.1 The Fermi Level

Electrons in solids obey Fermi-Dirac statistics. In the development of this type of statistics, one must consider the indistinguishability of the electrons, their wave nature, and the Pauli exclusion principle. The rather simple result of these statistical arguments is that the distribution of electrons over a range of allowed energy levels at thermal equilibrium is given by

\[
f(E) = \frac{1}{1 + e^{\frac{E - E_F}{kT}}} \quad (3-10)
\]

where \( k \) is Boltzmann's constant \( (k = 8.62 \times 10^{-5} \text{ eV}/\text{K} = 1.38 \times 10^{-23} \text{ J/K}) \). The function \( f(E) \), the Fermi-Dirac distribution function, gives the probability that an available energy state at \( E \) will be occupied by an electron at absolute temperature \( T \). The quantity \( E_F \) is called the Fermi level, and it represents an important quantity in the analysis of semiconductor behavior. We note here the number of states in the vicinity of the Fermi level.

\[
n_0 = \frac{e^{-E_F/kT}}{e^{E_F/kT}} = e^{-E_F/kT}
\]

assuming the two levels have an equal number of states. The exponential term \( e^{(-\Delta E/kT)} \) is commonly called the Boltzmann factor. It appears also in the denominator of the Fermi-Dirac distribution function. We shall return to the Boltzmann distribution in Chapter 10 in discussions of the properties of lasers.
For an energy $E$ equal to the Fermi level energy $E_F$, the occupation probability is

$$f(E_F) = \left[ 1 + e^{(E_F - E) \mu T} \right]^{-1} = \frac{1}{1 + 1} = \frac{1}{2} \quad (3-11)$$

Thus an energy state at the Fermi level has a probability of $\frac{1}{2}$ of being occupied by an electron.

A closer examination of $f(E)$ indicates that at 0 K, the distribution takes the simple rectangular form shown in Fig. 3-14. With $T = 0$ in the denominator of the exponent, $f(E)$ is $1/(1 + 0) = 1$ when the exponent is negative ($E < E_F$), and is $1/(1 + \infty) = 0$ when the exponent is positive ($E > E_F$). This rectangular distribution implies that at 0 K every available energy state up to $E_F$ is filled with electrons, and all states above $E_F$ are empty.

At temperatures higher than 0 K, some probability exists for states above the Fermi level to be filled. For example, at $T = T_1$ in Fig. 3-14 there is some probability $f(E)$ that states above $E_F$ are filled, and there is a corresponding probability $[1 - f(E)]$ that states below $E_F$ are empty. The Fermi function is symmetrical about $E_F$ for all temperatures (Prob. 3.4); that is, the probability $f(E_F + \Delta E)$ that a state $\Delta E$ above $E_F$ is filled is the same as the probability $[1 - f(E_F - \Delta E)]$ that a state $\Delta E$ below $E_F$ is empty. The symmetry of the distribution of empty and filled states about $E_F$ makes the Fermi level a natural reference point in calculations of electron and hole concentrations in semiconductors.

In applying the Fermi-Dirac distribution to semiconductors, we must recall that $f(E)$ is the probability of occupancy of an available state at $E$. Thus if there is no available state at $E$ (e.g., in the band gap of a semiconductor), there is no possibility of finding an electron there. We can best visualize the relation between $f(E)$ and the band structure by turning the $f(E)$ vs $E$ diagram on its side so that the $E$ scale corresponds to the energies of the band edges. For intrinsic material we know that the concentration of holes in the valence band is equal to the concentration of electrons in the conduction band. Therefore, the Fermi level $E_F$ must lie at the middle of the band gap in intrinsic material. Since $f(E)$ is symmetrical about $E_F$, the electron probability "tail" of $f(E)$ extending into the conduction band of Fig. 3-15a is symmetrical with the hole probability tail $[1 - f(E)]$ in the valence band. The distribution function has values within the band gap between $E_F$ and $E_v$, but there are no energy states available, and no electron occupancy results from $f(E)$ in this range.

The tails in $f(E)$ are exaggerated in Fig. 3-15 for illustrative purposes. Actually, the probability values at $E_F$ and $E_v$ are quite small for intrinsic material at reasonable temperatures. For example, in Si at 300 K, $n = p = 10^9 \text{ cm}^{-3}$, whereas the densities of available states at $E_F$ and $E_v$ are on the order of $10^{10} \text{ cm}^{-3}$. Thus the probability of occupancy $f(E)$ for an individual state in the conduction band and the hole probability $[1 - f(E)]$ for a state in the valence band are quite small. Because of the relatively large density of states in each band, small changes in $f(E)$ can result in significant changes in carrier concentration.

In n-type material, there is a high concentration of electrons in the conduction band compared with the hole concentration in the valence band (recall Fig. 3-10). Thus in n-type material the distribution function $f(E)$ must lie above its intrinsic position on the energy scale (Fig. 3-15b). Since $f(E)$ retains its shape for a particular temperature, the larger concentration of electrons at $E_F$...
in n-type material implies a correspondingly smaller hole concentration at \( E \). We notice that the value of \( f(E) \) for each energy level in the conduction band (and therefore the total electron concentration \( n_e \)) increases as \( E_c \) moves closer to \( E_F \). Thus the energy difference \((E_F - E_c)\) gives a measure of \( n_e \) we shall express this relation mathematically in the following section.

For p-type material the Fermi level lies near the valence band (Fig 3-15c) such that the \((1 - f(E))\) tail below \( E_c \) is larger than the \( f(E) \) tail above \( E_c \). The value of \((E_F - E_c)\) indicates how strongly p-type the material is.

It is usually inconvenient to draw \( f(E) \) vs. \( E \) on every energy band diagram to indicate the electron and hole distributions. Therefore, it is common practice merely to indicate the position of \( E_F \) in band diagrams. This is sufficient information, since for a particular temperature the position of \( E_F \) implies the distributions in Fig 3-15.

3.3.2 Electron and Hole Concentrations at Equilibrium

The Fermi distribution function can be used to calculate the concentrations of electrons and holes in a semiconductor, if the densities of available states in the valence and conduction bands are known. For example, the concentration of electrons in the conduction band is

\[
n_e = \int_{E_c}^{E_F} f(E) N(E) \, dE \tag{3-12}
\]

where \( N(E) \, dE \) is the density of states (cm\(^{-3}\)) in the energy range \( dE \). The subscript 0 used with the electron and hole concentration symbols \((n_e, p_h)\) indicates equilibrium conditions. The number of electrons per unit volume in the energy range \( dE \) is the product of the density of states and the probability of occupancy \( f(E) \). Thus the total electron concentration is the integral over the entire conduction band, as in Eq (3-12). The function \( N(E) \) can be calculated by using quantum mechanics and the Pauli exclusion principle (Appendix IV).

It is shown in Appendix IV that \( N(E) \) is proportional to \( E^{3/2} \), so the density of states in the conduction band increases with electron energy. On the other hand, the Fermi function becomes extremely small for large energies. The result is that the product \( f(E) N(E) \) decreases rapidly above \( E_c \), and very few electrons occupy energy states far above the conduction band edge. Similarly, the probability of finding an empty state (hole) in the valence band \((1 - f(E))\) decreases rapidly below \( E_c \), and most holes occupy states near the top of the valence band. This effect is demonstrated in Fig 3-16, which shows the density of available states, the Fermi function, and the resulting number of electrons and holes occupying available energy states in the conduction and valence bands at thermal equilibrium (i.e., with no excitations except thermal energy). For holes, increasing energy points down in Fig 3-16, since the \( E \) scale refers to electron energy.

Figure 3-16: Schematic band diagram, density of states, Fermi-Dirac distribution, and the carrier concentrations for (a) intrinsic, (b) n-type, and (c) p-type semiconductors at thermal equilibrium.

The result of the integration of Eq. (3-12) is the same as that obtained if we represent all of the distributed electron states in the conduction band by an effective density of states \( N_e \) located at the conduction band edge \( E_c \). Therefore, the conduction band electron concentration is simply the effective density of states at \( E_c \), times the probability of occupancy at \( E_c \)

\[
n_e = N_e f(E_c) \tag{3-13}
\]

The upper limit is actually improper in Eq (3-12), since the conduction band does not extend to infinite energy. This is unimportant in the calculation of \( n_e \), however, since \( f(E) \) becomes negligibly small for large values of \( E \). Most electrons occupy states near the bottom of the conduction band in equilibrium.
In this expression we assume the Fermi level \( E_F \) lies at least several \( kT \) below the conduction band. Then the exponential term is large compared with unity, and the Fermi function \( f(E_f) \) can be approximated as

\[
f(E_f) = \frac{1}{1 + e^{(E_f - E_c)/kT}} \approx e^{-(E_f - E_c)/kT}
\]  

(3-14)

Since \( kT \) at room temperature is only 0.026 eV, this is generally a good approximation. For this condition the concentration of electrons in the conduction band is

\[
N_e = N_L e^{-(E_f - E_c)/kT}
\]  

(3-15)

The effective density of states \( N_e \) is shown in Appendix IV to be

\[
N_e = \frac{2\pi m^*}{h^2} \left( \frac{2kT}{e} \right)^{3/2}
\]  

(3-16)

Since the quantities in Eq. (3-16) are known, values of \( N_e \) can be tabulated as a function of temperature. As Eq. (3-15) indicates, the electron concentration increases as \( E_F \) moves closer to the conduction band. This is the result we would predict from Fig. 3-15b.

By similar arguments, the concentration of holes in the valence band is

\[
p_n = N_L [1 - f(E_F)]
\]  

(3-17)

where \( N_L \) is the effective density of states in the valence band. The probability of finding an empty state at \( E_F \) is

\[
1 - f(E_F) = 1 - \frac{1}{1 + e^{-(E_F - E_V)/kT}} = e^{-(E_F - E_V)/kT}
\]  

(3-18)

for \( E_F \) larger than \( E_V \), by several \( kT \). From these equations, the concentration of holes in the valence band is

\[
p_n = N_L e^{-(E_F - E_V)/kT}
\]  

(3-19)

The effective density of states in the valence band reduced to the band edge is

\[
N_v = \frac{2\pi m^*}{h^2} \left( \frac{2kT}{e} \right)^{3/2}
\]  

(3-20)

As expected from Fig. 3-15c, Eq. (3-19) predicts that the hole concentration increases as \( E_F \) moves closer to the valence band.

The electron and hole concentrations predicted by Eqs. (3-15) and (3-19) are valid whether the material is intrinsic or doped, provided thermal equilibrium is maintained. Thus for intrinsic material, \( E_F \) lies at some intrinsic level \( E_i \) near the middle of the band gap (Fig. 3-15a), and the intrinsic electron and hole concentrations are

\[
n_i = N_L e^{-(E_i - E_c)/kT}, \quad p_i = N_L e^{-(E_V - E_i)/kT}
\]  

(3-21)

The product of \( n_0 \) and \( p_0 \) at equilibrium is a constant for a particular material and temperature, even if the doping is varied.

\[
n_0 p_0 = (N_L e^{-(E_f - E_c)/kT})(N_L e^{-(E_V - E_i)/kT}) = N_L^2 e^{-(E_f - E_c)/kT}
\]

(3-22a)

\[
N_L p_i = (N_L e^{-(E_V - E_i)/kT}) = N_L e^{-(E_f - E_c)/kT}
\]

(3-22b)

The intrinsic electron and hole concentrations are equal (since the carriers are created in pairs), \( n_i = p_i \), thus the intrinsic concentration is

\[
n_i = \sqrt{N_L e^{-(E_f - E_c)/kT}}
\]

(3-23)

The constant product of electron and hole concentrations in Eq. (3-22) can be written conveniently as

\[
n_0 p_0 = n_i^2
\]

(3-24)

This is an important relation, and we shall use it extensively in later calculations. The intrinsic concentration for Si at room temperature is approximately \( n_i = 1.5 \times 10^{10} \text{ cm}^{-3} \).

Comparing Eqs. (3-21) and (3-23), we note that the intrinsic level \( E_i \) is the middle of the band gap \( (E_c - E_V)/2) \), if the effective densities of states \( N_f \) and \( N_v \) are equal. There is usually some difference in effective mass for electrons and holes, however, and \( N_f \) and \( N_v \) are slightly different as Eqs. (3-16) and (3-20) indicate. The intrinsic level \( E_i \) is displaced from the middle of the band gap, more for GaAs than for Ge or Si (Table 3.1).

Another convenient way of writing Eqs. (3-15) and (3-19) is

\[
n_i = n_0 e^{-(E_f - E_i)/kT}
\]

(3-25a)

\[
p_i = p_0 e^{-(E_V - E_i)/kT}
\]

(3-25b)

obtained by the application of Eq. (3-21). This form of the equations indicates directly that the electron concentration is \( n_i \) when \( E_f \) is at the intrinsic level \( E_i \), and that \( n_i \) increases exponentially as the Fermi level moves away from \( E_i \) toward the conduction band. Similarly, the hole concentration \( p_i \) varies from \( n_i \) to larger values as \( E_V \) moves from \( E_i \) toward the valence band. Since these equations reveal the qualitative features of carrier concentration so directly, they are particularly convenient to remember.

A Si sample is doped with \( 10^{17} \) As atoms/cm\(^3\). What is the equilibrium hole concentration \( p_0 \) at 300 K? Where is \( E_f \) relative to \( E_i \)?

\[
\frac{n_0}{p_0} = \frac{2.25 \times 10^{10}}{3.0} = 2.25 \times 10^9 \text{ cm}^{-3}
\]

**EXAMPLE 3.4**

**SOLUTION**
From Eq (3-25a), we have

$$E_v - E_i = kT \ln \frac{n_i}{p_i} = 0.0259 \ln \frac{10^{10}}{1.5 \times 10^5} = 0.407 \text{ eV}$$

The resulting band diagram is

![Band Diagram](image)

3.3.3 Temperature Dependence of Carrier Concentrations

The variation of carrier concentration with temperature is indicated by Eq (3-25). Initially, the variation of $n_i$ and $p_i$ with $T$ seems relatively straightforward in these relations. The problem is complicated, however, by the fact that $n_i$ has a strong temperature dependence [Eq (3-23)] and that $E_v$ can also vary with temperature. Let us begin by examining the intrinsic carrier concentration. By combining Eqs (3-23), (3-16), and (3-20) we obtain

$$n_i(T) = 2\left(\frac{2\pi kT}{h^2}\right)^{3/2} \left(m^*_e m^*_h\right)^{1/2} e^{-E_v/2kT}$$

(3-26)

The exponential temperature dependence dominates $n_i(T)$, and a plot of $\ln n_i$ vs $10^2/T$ appears almost linear (Fig 3-17). In this figure we neglect variations due to the $T^2$ dependence of the density-of-states function and the fact that $E_v$ varies somewhat with temperature. The value of $n_i$ at any temperature is a definite number for a given semiconductor, and is known for most materials. Thus we can take $n_i$ as given in calculating $n_i$ or $p_i$ from Eq (3-25). With $n_i$ and $T$ given, the unknowns in Eq (3-25) are the carrier concentrations and the Fermi level position relative to $E_v$. One of these two quantities must be given if the other is to be found. If the carrier concentration is held at a certain value, as in heavily doped extrinsic material, $E_v$ can be obtained from $n_i$.

---

1. When plotting quantities such as carrier concentration, which involve a Boltzmann factor, it is common to use an inverse temperature scale. This allows terms which are exponential in $1/T$ to appear linear in the semilogarithmic plot. When reading such graphs, remember that temperature increases from right to left.

2. For Si the band gap $E_g$ varies from about 1.11 eV at 300 K to about 1.16 eV at 0 K.

3. Care must be taken to use consistent units in these calculations. For example, if an energy such as $E_v$ is expressed in electron volts (eV), it should be multiplied by $q$ (1.6 x 10^-19 C) to convert to joules if $k$ is in J/K. Alternatively, $E_v$ can be kept in eV and the value of $k$ in eV/K can be used. At 300 K we can use $kT = 0.0259$ eV and $E_v$ in eV.

---

Figure 3-17: Intrinsic Carrier Concentration for Ge, Si, and GaAs as a function of inverse temperature. The room temperature values are marked for reference.

Eq (3-25) The temperature dependence of electron concentration in a doped semiconductor can be visualized as shown in Fig 3-18. In this example, Si is doped n-type with a donor concentration $N_d$ of $10^{19}$ cm^{-3}. At very low temperatures (large $1/T$), negligible intrinsic EHPs exist, and the donor electrons are bound to the donor atoms. As the temperature is raised, these electrons are denoted to the conduction band, and at about 100 K $(1000/T = 10)$ all the donor atoms are ionized. This temperature range is called the ionization region. Once the donors are ionized, the conduction band electron concentration is $n_d = N_d = 10^{19}$ cm^{-3}, since one electron is obtained for each donor atom. When every available extrinsic electron has been transferred to the conduction band, $n_d$ is virtually constant with temperature until the concentration of intrinsic carriers becomes comparable to the extrinsic concentration $N_d$. Finally, at higher temperatures $n_d$ is much greater than $N_d$, and the intrinsic carriers dominate. In most devices it is desirable to control the carrier concentration by dop-
a hole concentration in the valence band commensurate with the acceptor concentration. In fact, the filling of the $E_a$ states occurs at the expense of the donated conduction band electrons. The mechanism can be visualized as follows. Assume an acceptor state is filled with a valence band electron as described in Fig. 3-11, with a hole resulting in the valence band. This hole is then filled by recombination with one of the conduction band electrons. Extending this logic to all the acceptor atoms, we expect the resultant concentration of electrons in the conduction band to be $N_e - N_a$ instead of the total $N_a$. This process is called compensation. By this process it is possible to begin with an n-type semiconductor and add acceptors until $N_e = N_a$ and no donated electrons remain in the conduction band. In such compensated material, $n_e = n_i = p_i$, and intrinsic conduction is obtained. With further acceptor doping the semiconductor becomes p-type with a hole concentration of essentially $N_d - N_a$.

The exact relationship among the electron, hole, donor, and acceptor concentrations can be obtained by considering the requirements for space charge neutrality. If the material is to remain electrostatically neutral, the sum of the positive charges (holes and ionized donor atoms) must balance the sum of the negative charges (electrons and ionized acceptor atoms):

$$p_i + N_d^+ = n_i + N_a^-$$  \quad (3-27)

Thus, in Fig 3-19 the net electron concentration in the conduction band is

$$n_0 = p_0 + (N_a^+ - N_d^-)$$  \quad (3-28)

If the material is doped n-type ($n_0 \gg p_0$) and all the impurities are ionized, we can approximate Eq. (3-28) by $n_0 = N_d - N_a$.

Since the intrinsic semiconductor itself is electrostatically neutral and the doping atoms add or subtract, the requirement of Eq. (3-27) must be maintained at equilibrium. The electron and hole concentrations and the Fermi level adjust such that Eqs. (3-27) and (3-25) are satisfied.

Knowledge of carrier concentrations in a solid is necessary for calculating current flow in the presence of electric or magnetic fields. In addition to the values of $n$ and $p$, we must be able to take into account the collisions of the charge carriers with the lattice and with the impurities. These processes will affect the ease with which electrons and holes can flow through the crystal, that is, their mobility within the solid. As should be expected, these collision and scattering processes depend on temperature, which affects the thermal motion of the lattice atoms and the velocity of the carriers.

1.4.1 Conductivity and Mobility

The charge carriers in a solid are in constant motion, even at thermal equilibrium. At room temperature, for example, the thermal motion of an individual electron may be visualized as random scattering from lattice atoms, impurities, other electrons, and defects (Fig. 3-20). Since the scattering is ran-
dom, there is no net motion of the group of $n$ electrons/cm$^3$ over any period of time. This is not true of an individual electron, of course. The probability of the electron in Fig. 3-20 returning to its starting point after some time $t$ is negligibly small. However, if a large number of electrons is considered (e.g., $10^6$ cm$^{-3}$ in an n-type semiconductor), there will be no preferred direction of motion for the group of electrons and no net current flow.

If an electric field $\vec{E}$ is applied in the $x$-direction, each electron experiences a net force $-q\vec{E}$ from the field. This force may be insufficient to alter appreciably the random path of an individual electron; the effect when averaged over all the electrons, however, is a net motion of the group in the $-x$-direction. If $p_x$ is the $x$-component of the total momentum of the group, the force of the field on the $n$ electrons/cm$^3$ is

$$-nP_x = \frac{dp_x}{dt}_{\text{avg}} \quad (3.29)$$

Initially, Eq. (3.29) seems to indicate a continuous acceleration of the electrons in the $-x$-direction. This is not the case, however, because the net acceleration of Eq. (3.29) is just balanced in steady state by the decelerations of the collision processes. Thus while the steady field $\vec{E}$ does produce a net momentum $p_x$, the net rate of change of momentum when collisions are included must be zero in the case of steady state current flow.

To find the total rate of momentum change from collisions, we must investigate the collision probabilities more closely. If the collisions are truly random, there will be a constant probability of collision at any time for each electron. Let us consider a group of $N_0$ electrons at time $t = 0$ and define $N(t)$ as the number of electrons that have not undergone a collision by time $t$. The rate of decrease in $N(t)$ at any time $t$ is proportional to the number left unscattered at $t$,

$$-\frac{dN(t)}{dt} = \frac{1}{\tau} N(t) \quad (3.30)$$

where $\tau^{-1}$ is a constant of proportionality

The solution to Eq. (3.30) is an exponential function

$$N(t) = N_0 e^{-t/\tau} \quad (3.31)$$

and $\tau$ represents the mean time between scattering events, \(1\), called the mean free time. The probability that any electron has a collision in the time interval $dt$ is

$$-\frac{dp_x}{dt} = \frac{p_x}{\tau} \quad (3.32)$$

The rate of change of $p_x$ due to the decelerating effect of collisions is

$$\frac{dN_x}{dt}_{\text{coll}} = -\frac{p_x}{\tau} \quad (3.33)$$

The sum of acceleration and deceleration effects must be zero for steady state.

Taking the sum of Eqs. (3.29) and (3.33), we have

$$-\frac{p_x}{\tau} - gq\vec{E}_x = 0 \quad (3.34)$$

The average momentum per electron is

$$\langle p_x \rangle = \frac{p_x}{n} = -gq\vec{E}_x \quad (3.35)$$

where the angular brackets indicate an average over the entire group of electrons. As expected for steady state, Eq. (3.35) indicates that the electrons have on the average a constant net velocity in the negative $x$-direction.

$$\langle v_x \rangle = \frac{p_x}{mv_x} = -gq\vec{E}_x \quad (3.36)$$

Actually, the individual electrons move in many directions by thermal motion during a given time period, but Eq. (3.36) tells us the net drift of an average electron in response to the electric field. The drift speed described by Eq. (3.36) is usually much smaller than the random speed due to thermal motion $v_0$.

The current density resulting from this net drift is just the number of electrons crossing a unit area per unit time ($n\langle v_x \rangle$) multiplied by the charge on the electron ($-q$).}

$$J_x = -q n\langle v_x \rangle \quad \begin{bmatrix} \text{ampere} \quad \text{coulomb} \quad \text{electrons} \quad \text{cm}^{-2} \quad \text{cm} \cdot \text{s}^{-1} \end{bmatrix} \quad (3.37)$$

\footnote{Equations (3.30) and (3.31) are typical of events dominated by random processes, and the forms of these equations occur often in many branches of physics and engineering. For example, in the radioactive decay of unstable nuclear isotopes, $N_0$ nuclei decay exponentially with a mean lifetime $\tau$. Other examples will be found in this text, including the absorption of light in a semiconductor and the recombination of excess EHPs.}
Using Eq. (3-36) for the average velocity, we obtain

$$J_s = \frac{neτ}{m^*} E$$  \hspace{1cm} (3-38)

Thus the current density is proportional to the electric field, as we expect from Ohm's law.

$$J_s = \sigma E$$, \hspace{1cm} where \( \sigma = \frac{neτ}{m^*} $$  \hspace{1cm} (3-39)

The conductivity \( \sigma (\Omega \cdot \text{cm})^{-1} \) can be written

$$\sigma = qne\mu_s$$, \hspace{1cm} where \( \mu_s = \frac{qY}{m^*} $$  \hspace{1cm} (3-40)

The quantity \( \mu_s \), called the electron mobility, describes the ease with which electrons drift in the material. Mobility is a very important quantity in characterizing semiconductor materials and in device development.

The mobility defined in Eq. (3-40) can be expressed as the average particle drift velocity per unit electric field. Comparing Eqs. (3-36) and (3-40), we have

$$\mu_s = -\frac{\langle v_s \rangle}{E}$$  \hspace{1cm} (3-41)

The units of mobility are \( \text{cm}^2/(\text{V} \cdot \text{s}) \), as Eq. (3-41) suggests. The minus sign in the definition results in a positive value of mobility, since electrons drift opposite to the field.

The current density can be written in terms of mobility as

$$J_s = qne\mu_s E$$  \hspace{1cm} (3-42)

This derivation has been based on the assumption that the current is carried primarily by electrons. For hole conduction we change \( n \) to \( p \), \(-q\) to \(+q\), and \( \mu_n \) to \( \mu_p \), where \( \mu_p = +\langle v_s \rangle / E \) is the mobility for holes. If both electrons and holes participate, we must modify Eq. (3-42) to

$$J_s = q(n\mu_n + p\mu_p)E = \sigma E$$  \hspace{1cm} (3-43)

Values of \( \mu_n \) and \( \mu_p \) are determined for many of the common semiconductor materials in Appendix III. According to Eq. (3-40), the parameters determining mobility are \( m^* \) and mean free time \( \tau \). Effective mass is a property of the material's band structure, as described by Eq. (3-3) Thus we expect \( m^* \) to be small in the strongly curved \( \Gamma \) minimum of the GaAs conduction band (Fig. 3-6), with the result that \( \mu_s \) is very high. In a more gradually curved band, a larger \( m^* \) in the denominator of Eq. (3-40) leads to a smaller value of mobility. It is reasonable to expect that lighter particles are more mobile than heavier particles (which is satisfying, since the common-sense value of effective mass is not always apparent). The other parameter determining mobility is the mean time between scattering events, \( \tau \). In Section 3.4.3 we shall see that this is determined primarily by temperature and impurity concentration in the semiconductor.

3.4.2 Drift and Resistance

Let us look more closely at the drift of electrons and holes. If the semiconductor bar of Fig. 3-21 contains both types of carrier, Eq. (3-43) gives the conductivity of the material. The resistance of the bar is then

$$R = \frac{\rho L}{wt} = \frac{L}{wt} \frac{1}{\sigma} $$  \hspace{1cm} (3-44)

where \( \rho \) is the resistivity (\( \Omega \cdot \text{cm} \)). The physical mechanism of carrier drift requires that the holes in the bar move as a group in the direction of the electric field and that the electrons move as a group in the opposite direction. Both the electron and the hole components of current are in the direction of the \( E \) field, since conventional current is positive in the direction of hole flow and opposite to the direction of electron flow. The drift current described by Eq. (3-43) is constant throughout the bar. A valid question arises, therefore, concerning the nature of the electron and hole flow at the contacts and in the external circuit. We shall see that the contacts to the bar of Fig. 3-21 are ohmic, meaning that they are perfect sources and sinks of both carrier types and have no special tendency to inject or collect either electrons or holes.

If we consider that current is carried around the external circuit by electrons, there is no problem in visualizing electrons flowing into the bar at one end and out at the other (always opposite to \( I \)). Thus for every electron leaving the left end (\( x = 0 \)) of the bar in Fig. 3-21, there is a corresponding electron entering at \( x = L \), so that the electron concentration in the bar remains constant at \( n \). But what happens to the holes at the contacts? As a hole reaches the ohmic contact at \( x = L \), it recombines with an electron, which must be supplied through the external circuit. As this hole disappears, a corresponding hole must appear at \( x = 0 \) to maintain space charge neutrality. It is reasonable to consider the source of this hole as the generation of an EHP at \( x = 0 \), with the hole flowing into the bar and the electron flowing into the external circuit.

![Figure 3-21 Drift of electrons and holes in a semiconductor bar](image)
EXAMPLE 3-5  Find the resistivity of intrinsic Ge at 300 K

SOLUTION  From Appendix III, \( \mu_e = 3900 \) and \( \mu_h = 1900 \text{ cm}^2/\text{V-s} \) for intrinsic Ge. Thus, since \( n_0 = p_0 = n_i \),

\[
\begin{align*}
\sigma_i &= q(\mu_e + \mu_h)n_i = 1.6 \times 10^{-19}(5800)(2.5 \times 10^{10}) \\
&= 2.32 \times 10^{-2} \text{ (}\Omega\cdot\text{cm})^{-1} \\
\rho_i &= \sigma_i^{-1} = 43 \text{ (}\Omega\cdot\text{cm)}
\end{align*}
\]

3.4.3 Effects of Temperature and Doping on Mobility

The two basic types of scattering mechanisms that influence electron and hole mobility are lattice scattering and impurity scattering. In lattice scattering, a carrier moving through the crystal is scattered by a vibration of the lattice, resulting from the temperature. The frequency of such scattering events increases as the temperature increases, since the thermal agitation of the lattice becomes greater. Therefore, we should expect the mobility to decrease as the sample is heated (Fig. 3-22). On the other hand, scattering from crystal defects such as ionized impurities becomes the dominant mechanism at low temperatures. Since the atoms of the cooler lattice are less agitated, lattice scattering is less important; however, the thermal motion of the carriers is also slower.

Since a slowly moving carrier is likely to be scattered more strongly by an interaction with a charged ion than by a carrier with greater momentum, impurity scattering events cause a decrease in mobility with decreasing temperature. As Fig. 3-22 indicates, the approximate temperature dependencies are \( T^{-2} \) for lattice scattering and \( T^{-3/2} \) for impurity scattering. Since the scattering probability of Eq. (3-32) is inversely proportional to the mean free time and therefore to mobility, the mobilities due to two or more scattering mechanisms add inversely:

\[
\frac{1}{\mu} = \frac{1}{\mu_1} + \frac{1}{\mu_2} + \cdots \quad (3-45)
\]

As a result, the mechanism causing the lowest mobility value dominates, as shown in Fig. 3-22.

![Figure 3-23](image1.png)

Variation of mobility with total doping impurity concentration \((N_x + N_e)\) for Ge, Si, and GaAs at 300 K

---

1Collective vibrations of atoms in the crystal are called phonons. Thus lattice scattering is also known as phonon scattering.
As the concentration of impurities increases, the effects of impurity scattering are felt at higher temperatures. For example, the electron mobility \( \mu_e \) of intrinsic silicon at 300 K is 1350 cm\(^2\)(V-s). With a donor doping concentration of \( 10^{18} \text{ cm}^{-3} \), however, \( \mu_e \) is 700 cm\(^2\)(V-s). Thus the presence of the \( 10^{11} \) ionized donors/cm\(^3\) introduces a significant amount of impurity scattering. This effect is illustrated in Fig. 3-23, which shows the variation of mobility with doping concentration at room temperature.

### 3.4.4 High-Field Effects

One assumption implied in the derivation of Eq. (3-39) was that Ohm's law is valid in the carrier drift processes. That is, it was assumed that the drift current is proportional to the electric field and that the proportionality constant (\( \sigma \)) is not a function of field \( E \). This assumption is valid over a wide range of \( E \). However, large electric fields (>10\(^5\) V/cm) can cause the drift velocity and therefore the current \( J = -q \sigma v_d \) to exhibit a sublinear dependence on the electric field. This dependence of \( \sigma \) upon \( E \) is an example of a hot carrier effect, which implies that the carrier drift velocity \( v_d \) is comparable to the thermal velocity \( v_t \).

In many cases an upper limit is reached for the carrier drift velocity in a high field (Fig. 3-24). This limit occurs near the mean thermal velocity (“10\(^6\) cm/s”) and represents the point at which added energy imparted by the field is transferred to the lattice rather than increasing the carrier velocity. The result of this scattering limited velocity is a fairly constant current at high field. This behavior is typical of Si, Ge, and some other semiconductors. However, there are other important effects in some materials; for example, in Chapter 12 we shall discuss a decrease in electron velocity at high fields for GaAs and certain other materials, which results in negative conductivity and current instabilities in the sample. Another important high-field effect is avalanche multiplication, which we shall discuss in Section 5.4.2.

#### 3.4.5 The Hall Effect

If a magnetic field is applied perpendicular to the direction in which holes drift in a p-type bar, the path of the holes tends to be deflected (Fig. 3-25). Using vector notation, the total force on a single hole due to the electric and magnetic fields is

\[
\mathbf{F} = q (\mathbf{E} + \mathbf{v} \times \mathbf{B})
\]

In the \( y \)-direction the force is

\[
F_y = q (E_y - v_y B_z)
\]

The important result of Eq. (3.47) is that unless an electric field \( E_y \) is established along the width of the bar, each hole will experience a net force (and therefore an acceleration) in the \(-y\)-direction due to the \( qv_y B_z \) product. Therefore, to maintain a steady state flow of holes down the length of the bar, the electric field \( E_y \) must just balance the product \( v_y B_z \),

\[
E_y = v_y B_z
\]

so that the net force \( F_y \) is zero. Physically, this electric field is set up when the magnetic field shifts the hole distribution slightly in the \(-y\)-direction. Once the electric field \( E_y \) becomes as large as \( v_y B_z \), no net lateral force is experienced by the holes as they drift along the bar. The establishment of the electric field \( E_y \) is known as the Hall effect, and the resulting voltage \( V_{sh} = E_y w \) is called

---

**Figure 1-24**
Saturation of electron drift velocity at high electric fields for Si.

**Figure 3-25**
The Hall effect.
the Hall voltage. If we use the expression derived in Eq. (3-37) for the drift velocity (using +q and p₀ for holes), the field Eₓ becomes

\[ Eₓ = \frac{J_x}{qµ_p} = R_n \frac{I_x}{qV_{ds}/w} \]

Thus the Hall field is proportional to the product of the current density and the magnetic flux density. The proportionality constant \( R_n = \frac{qµ_p}{2} \) is called the Hall coefficient. A measurement of the Hall voltage for a known current and magnetic field yields a value for the hole concentration \( p₀ \)

\[ p₀ = \frac{qR_n}{\mu_p} = \frac{I_x}{qV_{ds}/w} = \frac{I_x}{q(V_{ds}/w)} \quad (3-50) \]

Since all of the quantities in the right-hand side of Eq. (3-50) can be measured, the Hall effect can be used to give quite accurate values for carrier concentration.

If a measurement of resistance \( R \) is made, the sample resistivity \( ρ \) can be calculated:

\[ ρ(Ω·cm) = \frac{Rwl}{l} = \frac{V_{cm}/I_x}{I_x/w} \quad (3-51) \]

Since the conductivity \( σ = 1/ρ \) is given by \( qµ_pρ_0 \), the mobility is simply the ratio of the Hall coefficient and the resistivity

\[ υ_x = \frac{σ}{qρ} = \frac{1/ρ}{q(1/qR_n)} = \frac{R_n}{ρ} \quad (3-52) \]

Measurements of the Hall coefficient and the resistivity over a range of temperatures yield plots of majority carrier concentration and mobility vs. temperature. Such measurements are extremely useful in the analysis of semiconductor materials. Although the discussion here has been related to p-type material, similar results are obtained for n-type material. A negative value of \( q \) is used for electrons, and the Hall voltage \( V_{SH} \) and Hall coefficient \( R_n \) are negative. In fact, measurement of the signs of the Hall voltage is a common technique for determining if an unknown sample is p-type or n-type.

**EXAMPLE 3-6**

A sample of Si is doped with \( 10^{20} \) phosphorus atoms/cm². What would you expect to measure for its resistivity? What Hall voltage would you expect in a sample 100 µm thick if \( I_x = 1 \) mA and \( B_y = 1 \) kG = \( 10^{-2} \) Wb/cm²?

**SOLUTION**

From Eq. 3-23, the mobility is 700 cm²/(V·s). Thus the conductivity is

\[ σ = qµ_pρ = (1.6 \times 10^{-19})(700)(10^{20}) = 1128Ω^{-1} \]

since \( ρ_0 \) is negligible. The resistivity is

\[ ρ = σ^{-1} = 0.0893 Ω·cm \]

The Hall coefficient is

\[ R_n = -(qρ_0)^{-1} = -625 cm²/C \]

from Eq. 3-49, or we could use Eq. 3-52. The Hall voltage is

\[ V_{SH} = \frac{I_x}{1} R_n = \frac{10^{-3} A}{1} \frac{10^{-2} Wb/cm²}{(-625 cm²/C)} = -62.5 μV \]

In this chapter we have discussed homogeneous semiconductors, without variations in doping and without junctions between dissimilar materials. In the following chapters we will be considering cases in which nonuniform doping occurs in a given semiconductor, or junctions occur between different semiconductors or a semiconductor and a metal. These cases are crucial to the various types of electronic and optoelectronic devices made in semiconductors. In anticipation of those discussions, an important concept should be established here regarding the demands of equilibrium. That concept can be summarized by noting that no discontunity or gradient can arise in the equilibrium Fermi level \( E_F \).

To demonstrate this assertion, let us consider two materials in intimate contact such that electrons can move between the two (Fig. 3-26). These may be, for example, dissimilar semiconductors, n- and p-type regions, a metal and a semiconductor, or simply two adjacent regions of a nonuniformly doped semiconductor. Each material is described by a Fermi–Dirac distribution function and some distribution of available energy states that electrons can occupy.

There is no current, and therefore no net charge transport, at thermal equilibrium. There is also no net transfer of energy. Therefore, for each energy \( E \) in Fig. 3-26 any transfer of electrons from material 1 to material 2 must be exactly balanced by the opposite transfer of electrons from 2 to 1. We will let the density of states at energy \( E \) in material 1 be \( N_1(E) \) and in material 2 we
will call $N_f(E)$. At energy $E$ the rate of transfer of electrons from 1 to 2 is proportional to the number of filled states at $E$ in material 1 times the number of empty states at $E$ in material 2.

\[
\text{rate from 1 to 2} = N_f(E)f_f(E)N_f(E)[1 - f_f(E)]
\]

where $f_f$ is the probability of a state being filled at $E$ in each material, i.e., the Fermi–Dirac distribution function given by Eq. (3-10). Similarly,

\[
\text{rate from 2 to 1} = N_f(E)f_f(E)N_f(E)[1 - f_f(E)]
\]

At equilibrium these must be equal.

\[
N_f(E)f_f(E)N_f(E)[1 - f_f(E)] = N_f(E)f_f(E)N_f(E)[1 - f_f(E)]
\]

Rearranging terms, we have, at energy $E$,

\[
N_f(E)f_f(E)N_f(E)[1 - f_f(E)] = N_f(E)f_f(E)N_f(E)[1 - f_f(E)]
\]

which results in

\[
f_f(E) = f_f(E).
\]

Therefore, we conclude that $E_f = E_f$. That is, there is no discontinuity in the equilibrium Fermi level. More generally, we can state that the Fermi level at equilibrium must be constant throughout materials in intimate contact. One way of stating this is that no gradient exists in the Fermi level at equilibrium

\[
\frac{dE_f}{dx} = 0
\]

We will make considerable use of this result in the chapters to follow.

**Problems**

1. The ionic radius of Na⁺ (atomic weight 23) and Cl⁻ (atomic weight 35.5) are 1.0 and 1.8 Å, respectively. Treating the ions as hard spheres, calculate the density of NaCl. Compare this with the measured density of 2.17 g/cm³.

2. Since the effective mass of electrons in a conduction band decreases with increasing curvature of the band according to Eq. (3-3), comment on the electron effective mass in the $\Gamma$-valley of GaAs compared to the indirect X or L valleys (see Fig. 3-6). How is this effective mass difference reflected in the electron mobility for GaAs and GaP shown in Appendix III? From Fig. 3-6(a), what would you expect to happen to the conductivity of GaAs if $\Gamma$-valley electrons drifting in an electric field were suddenly promoted to the L valley?

3. The electron bound to a donor is described in Example 3-3 as being in a hydrogen-like orbit. Calculate the radius of the ground-state orbit for this

4. Show that the probability a state $\Delta E$ above the Fermi level $E_f$ is filled is the probability a state $\Delta E$ below $E_f$ is empty.

5. From Eqs. (3-21) and (3-23), show that $E_f$ lies below the middle of the band gap by $kT\ln(m^2/m)^{1/2}$. Show that this displacement is small (e.g., compared to $kT$) for Si but significant for GaAs.

6. Derive Eq. (3-25) from Eqs. (3-15) and (3-19). If $n_i = 10^{10}$ cm⁻³ for Si and $T = 300$ K, where is $E_f$ with respect to $E_f$? Repeat for $p_i = 10^{10}$ cm⁻³.

7. Given $n_s$ and $n_d$ from Table 3-1, calculate the effective densities of states $N_s$ and $N_d$ for GaAs at 300 K (assume $m_s^* = m_d^* = m^*$ do not vary with temperature). Calculate the intrinsic carrier concentration and compare this with the value given in Fig. 3-17.

8. Calculate the band gap of GaAs from Eq. (3-23) and the plot of $n_i$ vs. $1000/T$ (Fig. 3-17). Hint: The slope cannot be measured directly from a semilogarithmic plot; read the values from two points on the plot and take the natural logarithm as needed for the solution.

9. A semiconductor device requires n-type material, it is to be operated at 400 K. Would Si doped with $10^{17}$ atoms/cm³ of arsenic be useful in this application? Could Ge doped with $10^{17}$ cm⁻³ antimony be used?

10. (a) A Si sample is doped with $10^{18}$ boron atoms/cm³. What is the electron concentration $n_e$ at 300 K?

(b) A Ge sample is doped with $5 \times 10^{17}$ Sb atoms/cm³. Using the requirements of space charge neutrality, calculate the electron concentration $n_e$ at 300 K.

11. Show that $t$ in Eq. (3-31) represents the mean time between scattering events.

12. Assume that a conduction electron in Si ($\mu_e = 1350$ cm²/V·s) has a thermal energy of $kT$, related to its mean thermal velocity by $E_v = (m_e kT)^{1/2}$. This electron is placed in an electric field of 10⁵ V/cm. Show that the drift velocity of the electron in this case is small compared to its thermal velocity. Repeat for a field of 10⁷ V/cm, using the same value of $\mu_e$. Comment on the actual mobility effects at this higher value of field.

13. A Si sample is doped with $5 \times 10^{18}$ cm⁻³ donors and $2 \times 10^{16}$ cm⁻³ acceptors. Find the position of the Fermi level with respect to $E_f$ at 300 K. What is the value and sign of the Hall coefficient?

14. How long does it take an average electron to drift 1 μm in pure Si at an electric field of 10⁵ V/cm? Repeat for 10⁷ V/cm.
3.15 (a) A Si sample is doped with $10^{16}$ cm$^{-3}$ boron atoms, and a certain number of shallow donors. The Fermi level is 0.36 eV above $E_F$ at 300 K. What is the donor concentration $N_D$? (b) A Si sample contains $10^{15}$ cm$^{-3}$ In acceptor atoms and a certain number of shallow donors. The In acceptor level is 0.16 eV above $E_F$, and $E_D$ is 0.26 eV above $E_F$ at 300 K. How many cm$^{-3}$ In atoms are un-ionized (i.e., neutral)?

3.16 A Si bar 0.1 cm long has a cross-sectional area of 0.01 cm$^2$. (a) What measurements would you use to find out the majority carrier type and concentration? (b) If the measurements in (a) show the sample is n-type with an electron concentration of $10^{18}$ cm$^{-3}$, and you measure the resistance to be 1Ω, what is the electron mobility? (c) How long does it take an average electron to drift the length of the bar in a field of 1 kV/cm?

3.17 A Ge sample is properly contacted and oriented in a 5 kG magnetic field as in Fig. 3-25. The current is 2 mA. The sample dimensions are $w = 0.25$ mm, $t = 50$ μm, and $L = 2.5$ mm. The following data are taken. $V_{PD} = -1.25$ mV and $V_{CE} = 85$ mV. Find the type and concentration of the majority carriers and its mobility. Note: 1 kG = 10$^3$ Wb/cm$^2$. Comment on the mobility value.

**Reading List**


Week 3

Thermal Oxidation
form electron concentration of $10^{19}$ cm$^{-2}/V \cdot s$.

was used with 10 V applied at the
red?

applied, what is the Hall voltage?

were formed by self-ion implant-
annealed at 550°C:
layer epitaxially?

were present, what time would be
required?

9

CHAPTER

Thermal Oxidation of Silicon and
Chemical Vapor Deposition of
Insulating Films

9.1 Introduction

Silicon has become the dominant semiconductor material in integrated-circuit technol-
ology because of our ability to grow a stable oxide with good control by relatively
easy methods. Silicon dioxide forms the basis of the planar technology. In industrial prac-
tice, silicon dioxide layers are frequently formed by thermal oxidation of Si in
the temperature range 900 to 1200°C.

The thermal oxidation process is one of the key steps in the fabrication of silicon
semiconductor devices and integrated circuits. Thermal oxides are used to mask
selectively against dopant diffusion, to passivate device junctions, and for device
isolation.

Figure 9.1a shows the cross section of a semirecessed oxide NMOS device (n-
channel metal--oxide--Si field-effect transistor). The sequence used to grow a layer
of recessed oxide is shown in Fig. 9.1b. A thin layer (10 to 20 nm) of SiO$_2$ is
grown on the wafer initially, followed by the deposition of silicon nitride (Si$_3$N$_4$)
and patterning of the wafer. Oxidation does not occur in areas covered by Si$_3$N$_4$,
due to the difficulty of oxygen and water vapor to diffuse through Si$_3$N$_4$. This
selective oxidation process results in the recessed oxide structure.

There are two commonly used chemical reactions in oxidation; one is due to
oxidation with dry oxygen gas (dry oxidation) and the other is due to oxidation
with water vapor (wet oxidation). Wet oxidation generally proceeds with a faster
rate than that of dry oxidation. The chemical reaction for dry oxidation is

$$\text{Si(solid)} + \text{O}_2(\text{gas}) \rightarrow \text{SiO}_2(\text{solid}). \quad (9.1)$$

The chemical reaction for wet oxidation is

$$\text{Si(solid)} + 2\text{H}_2\text{O}(\text{vapor}) \rightarrow \text{SiO}_2(\text{solid}) + 2\text{H}_2. \quad (9.2)$$

The heat of formation of SiO$_2$ at 25°C is generally accepted to be about 70
kcal/g atom (or 210 kcal/mol). Heats of reaction are reported in units of
kcal/g atom or kcal/mol in the literature, and 23.06 kcal/mol is equal to 1.0
eV/molecule.

Silicon dioxide has more than one crystalline form. The SiO$_2$ layers grown by
thermal oxidation, however, have no long-range crystalline order; they are amor-
oxidation both with oxygen and through the oxide layer to react with a deeper interface and of an inward oxide layer is to provide a clean air exposure prior to oxidation (in situ).

9.2 Reaction Kinetics

The oxidation model proposed by [Fig. 9.2] in [this model] and [surface, there are three steps in such a process:

1. The oxidizing species must reach the oxide-gas interface. The flux is given by

\[ F_1 \]

where \( h_{ci} \) is the mass-transfer coefficient for the oxidant in the bulk of the gas.

We now want to relate the concentrations of the oxidant in the solid to the partial pressure of the oxidant. Henry's law is that the equilibrium concentration in the solid is proportional to the partial pressure of the oxidant. Henry's law to the given situation is

\[ C_s = \frac{h_{ci} P}{RT} \]

\[ C_s \]

FIGURE 9.2 Model for the thermal oxidation of silicon and chemical vapor deposition of films.
oxidation both with oxygen and with water vapor, the oxidizing species move through the oxide layer to react with Si. The principal effect of a displaced and deeper interface and of an inward migration of the oxidizing species through the oxide layer is to provide a clean oxide–Si interface, free of contaminants due to air exposure prior to oxidation (interfacial impurities tend to end up on the oxide surface).

### 9.2 Reaction Kinetics

The oxidation model proposed by Deal and Grove (1965) is shown schematically in Fig. 9.2. According to this model, for the oxidizing species to reach the silicon surface, there are three steps in series:

1. The oxidizing species must be transported from the bulk of the gas to the oxide–gas interface. The flux is given by

   \[ F_1 = h_0(C_G - C_S) \]  

   (9.3)

   where \( h_0 \) is the mass-transfer coefficient and \( C_G \) and \( C_S \) are the concentration of the oxidant in the bulk of the gas and in the gas at the oxide surface, respectively. We now want to relate the concentrations of the oxidant in the gas to the concentrations of the oxidant in the solid (oxide) through Henry's law. The idea behind Henry's law is that the equilibrium concentration of a species in a solid is proportional to the partial pressure of that species in the surrounding gas. Applying Henry's law to the given situation, the concentration of the oxidant at the oxide surface (but within the oxide) is

   \[ C_o = K_H P_t \]  

   (9.4)

![Figure 9.2](image_url)  

**Figure 9.2** Model for the thermal oxidation of silicon. Direction of gas flow is normal to plane of paper. [After Grove (1967).]
where $C_o$ is the actual concentration of the oxidant at the oxide surface (within the oxide), $K_H$ the Henry’s law constant, and $P_s$ the partial pressure of the oxidant right next to the oxide surface. The concentration of the oxidant at the oxide surface (within the oxide), $C^*$, that would be in equilibrium with the partial pressure in the bulk of the gas $P_G$ is given by

$$C^* = K_H P_G.$$  \hspace{1cm} (9.5)

The driving force for $F_1$ is therefore proportional to the deviation from equilibrium (i.e., $C^* - C_o$). Using the ideal gas law,

$$PV = N_m kT.$$ \hspace{1cm} (1.3)

The gas-phase oxidant concentrations can be expressed as

$$C = \frac{N_m}{V} = \frac{P}{kT}.$$ \hspace{1cm} (9.5a)

where $P$ is the partial pressure of the oxidant in the gas, $N_m$ the number of molecules of the oxidant in the gas, and $C$ the gas-phase concentration expressed in units of molecules/volume. The concentration of oxidant in the bulk of the gas $C_G$ is

$$C_G = \frac{P_G}{kT}.$$ \hspace{1cm} (9.5b)

and the concentration of the oxidant at the oxide surface (but in the gas)

$$C_s = \frac{P_s}{kT}.$$ \hspace{1cm} (9.5c)

Substituting into eq. (9.3) yields

$$F_1 = h_G(C_G - C_s) = h_G \frac{(P_G - P_s)}{kT}.$$  

Using eqs. (9.4) and (9.5), $F_1$ can be expressed as

$$F_1 = h_G \frac{(P_G - P_s)}{kT} = h(C^* - C_o)$$ \hspace{1cm} (9.6)

where

$$h = \frac{h_G}{kT K_H}.$$  

2. The oxidizing species must diffuse across the oxide layer already present. The diffusional flux $F_2$ is given by

$$F_2 = D \frac{C_o - C_i}{x_o}.$$ \hspace{1cm} (9.7)

where $D$ is the diffusivity, $C_i$ the concentration of the oxidant at the oxide-Si interface, and $x_o$ is the thickness of the oxide at a given time, $t$.

3. The oxidizing species must diffuse through the oxide in order for oxidation to occur, the diffusivity of the oxidant $x_o(t)$, thus causing $C_i \to 0$. The flux is now diffusion controlled. In the absence of $k$, $x_o(t)$, $C_i \to C^* = C_o$. The oxide growth is therefore controlled. Combining eqs. (9.7) and (9.8), $F$ is given by

$$\frac{dx_o}{dt} = N k C_o,$$

thus the rate at which the oxide thickness $x_o$ grows is

$$\frac{dx_o}{dt} = N k C_o,$$

where $N$ is the number of oxidant molecules per unit volume. For dry oxidation, $N = 2 \times 10^{22}$ cm$^{-3}$ since two H$_2$O molecules...
3. The oxidizing species must then react with the Si at the oxide–Si interface. The flux to incorporate the oxidant into Si, \( F_3 \), is given by

\[
F_3 = k_i C_i
\]  

(9.8)

where \( k_i \) is the chemical reaction rate constant for oxidation at the oxide–Si interface.

It is important to recognize at this point that these three steps are in series; eventually, the slowest step among the three will control the entire process of oxidation. When a steady-state condition is established, all fluxes are equal (i.e., \( F = F_1 = F_2 = F_3 \)). Solving for \( C_i \) and \( C_o \), we have

\[
C_i = \frac{C^*}{1 + k_i/h + k_o x_o / D}
\]

(9.9)

and

\[
C_o = \frac{(1 + k_i x_o / D) C^*}{1 + k_i / h + k_o x_o / D}
\]

(9.10)

In thermal oxidation of Si, step 1 (\( F_1 \)) is rarely the rate-limiting step. We can assume that \( h \) is much larger than \( k_i \). With this assumption, eqs. (9.9) and (9.10) can be further simplified to

\[
C_i \approx \frac{C^*}{1 + k_i x_o(t) / D}
\]

(9.11)

and

\[
C_o \approx C^*.
\]

(9.12)

Let us consider the situation where step 2 (\( F_2 \)) is the rate-limiting step. For this to occur, the diffusivity of the oxidizing species, \( D \), is very small compared to \( k_i x_o(t) \), thus causing \( C_i \to 0 \). This is the case where the oxidation process is said to be diffusion controlled. In the opposite case where \( D \) is very large compared to \( k_i x_o(t) \), \( C_i \to C^* \approx C_o \). The oxidation process is said to be interfacial reaction controlled. Combining eqs. (9.7), (9.11), and (9.12) and eliminating \( C_i \), the flux \( F \) is given by

\[
F = \frac{DC_b k_i}{D + k_x x_o}
\]

(9.13)

Thus the rate at which the oxide layer grows is

\[
\frac{dx_o}{dt} = \frac{F}{N} = \frac{DC_b k_i}{N (D + k_x x_o)}
\]

(9.14)

where \( N \) is the number of oxidant molecules incorporated into a unit volume of oxide. For dry oxidation, \( N = 2.3 \times 10^{22} \text{ cm}^{-3} \), for wet oxidation \( N = 4.6 \times 10^{22} \text{ cm}^{-3} \) since two \( \text{H}_2\text{O} \) molecules are incorporated into one molecule of \( \text{SiO}_2 \).
Inspection of equation (9.14) reveals that oxidation is (1) controlled; when \( D \ll k_c x_0, F = \frac{d x}{dt} \) in two of the cases are illustrated schematically in Fig. 9.3, and \( x_o(t = 0) = x_i \), the general solution is

\[
x_o^2(t) = \frac{A}{2} t \quad \text{or} \quad x_o = \sqrt{\frac{A}{2} t}
\]

where \( A = 2Dk_i^{-1} \) (cm), \( B = 2Dk_i^{-1} t_i \).

Solving for the oxide thickness \( x_o \),

\[
x_o = \sqrt{\frac{A t}{2}}
\]

For short oxidation times, \( t + \tau \) \( x_o = (B/A)(t + \tau) \). For long oxidation times, \( x_o \) depends on the oxidation law, \( x_o^2 = B t \). It should be noted that when the oxide layer is thin, leading to a large thickness, \( x_o \), the diffusion process becomes the controlling step and the oxide thickness becomes large, resulting in a larger diffusion flux \( D/x_o(t) \).

The parabolic rate constant \( B \) is shown as a function of \( 1000/T \) in Figs. 9.4 and 9.5. The values of the rate constant \( B \) for the rate constant \( B \) is 1.24 eV for water vapor and 0.71 eV for oxygen. These are in good agreement with the values of 1.24 eV for oxygen and 0.71 eV for water vapor in fused silica (see Table 8.1). The linear rate constant \( (B/A) \) is about 1.24 eV for the energy of the surface, indicating that the rate constant \( k_o \) is determined by the surface recombination process.

The general relations for the parabolic diffusion forms are shown in Figs. 9.3 and 9.4.

In solving eq. (9.14), an initial condition is that for wet oxidation the value of the oxide thickness is not known. The initial condition is that the oxide thickness is zero.

**TABLE 8.1** Rate Constants for Wet Oxidation

<table>
<thead>
<tr>
<th>Oxidation Temperature (°C)</th>
<th>Rate Constant ( B ) (µm²/h)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1200</td>
<td>0.5</td>
</tr>
<tr>
<td>1100</td>
<td>0.75</td>
</tr>
<tr>
<td>1000</td>
<td>1</td>
</tr>
<tr>
<td>920</td>
<td>0.25</td>
</tr>
</tbody>
</table>

Source: After Deal and Grove (1965).
Inspection of equation (9.14) reveals that when \( D >> k_r x_0 \), \( F = C \cdot k_r \) (reaction controlled); when \( D << k_r x_0 \), \( F = D C_o x_o \) (diffusion controlled). The two limiting cases are illustrated schematically in Fig. 9.3. Using the boundary condition of \( x_o(t = 0) = x_i \), the general solution for eq. (9.14) is

\[
x_o^2 + A x_o = B(t + \tau)
\]

(9.15)

where \( A = 2Dk_r^{-1} \) (cm), \( B = 2DC_oN^{-1} \) (cm²/s), and \( \tau = x_i^2/B + x_i/(B/A) \).

Solving for the oxide thickness as a function of oxidation time, we obtain

\[
x_o = \frac{A}{2} \left( \left( 1 + \frac{t + \tau}{A^2/4B} \right)^{1/2} - 1 \right)
\]

(9.16)

For short oxidation times, \( t + \tau << A^2/4B \), we obtain the linear oxidation law, \( x_o = (B/A)(t + \tau) \). For long oxidation times, \( t >> A^2/4B \), we obtain the parabolic oxidation law, \( x_o^2 = Bt \). It should be noted that for short-time oxidation, the oxide layer is thin, leading to a large concentration gradient and large diffusional flux. Under such a circumstance, the interfacial oxidation reaction should be the rate-limiting step \([D/x_o(t) > k_r \text{ or } k_r \to 0]\) for oxidation. For long-time oxidation, the oxide thickness is large, resulting in a small concentration gradient and small diffusional flux \([D/x_o(t) < k_r \text{ or } D \to 0]\); the rate-limiting step is switched from a reaction-controlled mechanism to a diffusion-controlled mechanism.

The parabolic rate constant \( B \) and the linear rate constant \( (B/A) \) are plotted against 1000/T in Figs. 9.4 and 9.5, respectively. It can be seen that the activation for the rate constant \( B \) is 1.24 eV for dry oxidation and 0.71 eV for wet oxidation. These values are in good agreement with the activation energies for diffusion of O₂ and H₂O in fused silica (see Table 7.7), suggesting that the parabolic rate is associated with the diffusion of the oxidants. The temperature dependence of the linear rate constant \( (B/A) \) is about 2 eV (Fig. 9.5), close to the bond-breaking energy of Si, indicating that \( B/A \) is associated with the interfacial reaction rate constant \( k_r \). The general relationship for silicon oxidation [eq. (9.16)] and its two limiting forms are shown in Fig. 9.6. The rate constants are listed in Tables 9.1 and 9.2.

In solving eq. (9.14), an initial oxide thickness \( x_i \) is assumed. It should be noted that for wet oxidation the value of \( x_i = 0 \) (Nichollian and Brews, 1982). For dry oxidation, the initial condition \( x_i = 23 \) nm must be used in eq. (9.15). This is

<table>
<thead>
<tr>
<th>Oxidation Temperature (°C)</th>
<th>( A ) (µm)</th>
<th>( B ) (µm²/h)</th>
<th>( B/A ) (µm/h)</th>
<th>( \tau ) (h)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1200</td>
<td>0.05</td>
<td>0.720</td>
<td>14.40</td>
<td>0</td>
</tr>
<tr>
<td>1100</td>
<td>0.11</td>
<td>0.510</td>
<td>4.64</td>
<td>0</td>
</tr>
<tr>
<td>1000</td>
<td>0.226</td>
<td>0.287</td>
<td>1.27</td>
<td>0</td>
</tr>
<tr>
<td>920</td>
<td>0.50</td>
<td>0.203</td>
<td>0.406</td>
<td>0</td>
</tr>
</tbody>
</table>

Source: After Deal and Grove (1965).
FIGURE 9.5  Effect of temperature on the linear rate constant. [After Grove (1967).]

![Graph showing the effect of temperature on the linear rate constant.]

**TABLE 9.2** Rate Constants for Oxidation of Silicon in Dry Oxygen

\[ P_{O_2} = 760 \text{ torr}; \quad x_0 + Ax_o = B(t + \tau) \]

<table>
<thead>
<tr>
<th>Oxidation Temperature (°C)</th>
<th>A (μm)</th>
<th>B (μm²/h)</th>
<th>B/A (μm/h)</th>
<th>τ (h)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1200</td>
<td>0.040</td>
<td>0.045</td>
<td>1.12</td>
<td>0.027</td>
</tr>
<tr>
<td>1100</td>
<td>0.090</td>
<td>0.027</td>
<td>0.30</td>
<td>0.076</td>
</tr>
<tr>
<td>1000</td>
<td>0.165</td>
<td>0.0117</td>
<td>0.071</td>
<td>0.34</td>
</tr>
<tr>
<td>920</td>
<td>0.235</td>
<td>0.0049</td>
<td>0.0208</td>
<td>1.40</td>
</tr>
<tr>
<td>800</td>
<td>1.340</td>
<td>0.0011</td>
<td>0.0030</td>
<td>9.0</td>
</tr>
</tbody>
</table>

Source: After Deal and Grove (1965).

FIGURE 9.6 General relationship

[After Grove (1967).]

because of the fact that an extrapolation of the etch rate is, at t = 0 yields an intersect at 0.1 nm is believed to be a space-charge layer as a function of time at different oxidation pressures. These curves obey eq. (5) much faster rate in H₂O than in O₂, because of the properties of thermally grown oxide.

**TABLE 9.3** Properties of Thermally Grown Oxide

<table>
<thead>
<tr>
<th>Property</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>DC resistivity (Ω · cm), 25°C</td>
<td>10³</td>
</tr>
<tr>
<td>Density (g/cm³)</td>
<td>2.2</td>
</tr>
<tr>
<td>Dielectric constant</td>
<td>3.8</td>
</tr>
<tr>
<td>Dielectric strength (V/cm)</td>
<td>5−8</td>
</tr>
<tr>
<td>Energy gap (eV)</td>
<td>~8.2</td>
</tr>
<tr>
<td>Etch rate in buffered HF (nm/min)</td>
<td>~0.1</td>
</tr>
<tr>
<td>Infrared absorption peak (μm)</td>
<td>9.3</td>
</tr>
<tr>
<td>Linear expansion coefficient (°C⁻¹)</td>
<td>5.5</td>
</tr>
</tbody>
</table>

Source: After Wolf and Tauber (1986). *Buffered HF: 28 ml HF, 170 ml H₂O, 110°C.

because of the fact that an extrapolation of the oxide thickness $x_o$ versus oxidation time $t$ at $t = 0$ yields an intercept of 23 nm. The initial nonlinear growth of 23 nm is believed to be a space-charge effect. Figure 9.7 shows the oxide thickness as a function of time at different temperatures in wet (95°C H₂O) oxygen and dry oxygen. These curves obey eq. (9.16); and it is clear that oxidation proceeds at a much faster rate in H₂O than in O₂ in both the linear and parabolic regions. Some of the properties of thermally grown silicon dioxide are listed in Table 9.3.

### TABLE 9.3 Properties of Thermal Silicon Dioxide

<table>
<thead>
<tr>
<th>Property</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>DC resistivity (Ω · cm), 25°C</td>
<td>$10^{14} - 10^{16}$</td>
</tr>
<tr>
<td>Density (g/cm³)</td>
<td>2.27</td>
</tr>
<tr>
<td>Dielectric constant</td>
<td>3.8 – 3.9</td>
</tr>
<tr>
<td>Dielectric strength (V/cm)</td>
<td>$5 - 10 \times 10^6$</td>
</tr>
<tr>
<td>Energy gap (eV)</td>
<td>~ 8</td>
</tr>
<tr>
<td>Etch rate in buffered HF (nm/min)³</td>
<td>100</td>
</tr>
<tr>
<td>Infrared absorption peak (μm)</td>
<td>9.3</td>
</tr>
<tr>
<td>Linear expansion coefficient (°C⁻¹)</td>
<td>$5.0 \times 10^{-7}$</td>
</tr>
<tr>
<td>Melting point (°C)</td>
<td>~ 1700</td>
</tr>
<tr>
<td>Molecular weight</td>
<td>60.08</td>
</tr>
<tr>
<td>Molecules/cm³</td>
<td>$2.3 \times 10^{22}$</td>
</tr>
<tr>
<td>Refractive index</td>
<td>1.46</td>
</tr>
<tr>
<td>Specific heat (J/g · °C)</td>
<td>1.0</td>
</tr>
<tr>
<td>Stress in film on Si (N/m²)</td>
<td>$2 - 4 \times 10^8$</td>
</tr>
<tr>
<td>Compression</td>
<td></td>
</tr>
<tr>
<td>Infrared absorption peak (μm)</td>
<td>9.3</td>
</tr>
<tr>
<td>Linear expansion coefficient (°C⁻¹)</td>
<td>$5.0 \times 10^{-7}$</td>
</tr>
<tr>
<td>Thermal conductivity (W/cm · °C)</td>
<td>0.014</td>
</tr>
</tbody>
</table>

*Source: After Wolf and Tauber (1986).*

*Buffered HF: 28 ml HF, 170 ml H₂O, 113 g NH₄F.*
9.3 Factors Affecting Oxidation Kinetics

Besides the dry versus wet oxidation and the oxidation temperature, there are three factors commonly observed to affect the oxidation rate of Si: (1) orientation of the substrate, (2) pressure effects, and (3) impurity effects. The orientation dependence of the oxidation rate arises primarily from the total number of available Si atoms for oxidation as a function of the substrate orientation. Since the available number of Si atoms is related to the oxidation reaction at the interface between the oxide and the Si wafer, only the linear oxidation rate is expected to change as a function of orientation. In the diffusion-limited regime, the oxidation rate is independent of the substrate orientation. Figure 9.8 shows the oxidation kinetics on (100) and (111) Si at 1000°C. The oxidation rate for (111) Si is faster than that for (100) Si initially in the linear region. As the oxidation kinetics change from the linear rate to the parabolic rate, the difference between the two orientations diminishes.

In both the linear regime and the parabolic regime, the oxide thickness is related to the rate constant $B$, which is given by $2DC_pN^{-1}$ or more correctly, $2DC_p^*N^{-1}$. The equilibrium oxidant concentration $C^*$ is directly proportional to the partial pressure of the oxidant in the gas stream [eq. (9.5)]. The oxidation rate is therefore expected to increase with increasing pressure of the oxidant, as shown in Fig. 9.9.

Boron and phosphorus can cause an increase when the Si substrates are diffused p- and n-type layers. In these cases, the oxidation rate can increase when the Si substrates are diffused p- and n-type layers. The effect of boron and phosphorus can be increased with increasing boron concentration.

**FIGURE 9.7** Oxide thickness as a function of time with temperature as parameter. Solid lines represent wet oxygen; dashed lines represent dry oxygen. Silicon surface orientation is (111). [After Nicollian and Brews (1982).]

**FIGURE 9.8** Oxide thickness $x_t$ versus time $t$ in wet oxygen and (111)-oriented n-type silicon (NSI) at 1000°C. [After Nicollian and Brews]

**FIGURE 9.9** Wet oxide growth at 1000°C for Si substrate with a resistivity of 3 to 10 kΩ·cm. [After Su (1981). Reprinted, permission of Academic Press.]
Kinetics

At a constant temperature, there are three major factors affecting the oxidation rate of Si: (1) orientation of the Si substrate, (2) impurity concentration, and (3) oxidation temperature. The orientation dependence is determined by the number of available Si atoms at the Si-oxide interface, since the oxidation rate is expected to decrease as a function of the interface area. The oxidation rate is independent of oxidation temperature since the reaction is exothermic and occurs at near equilibrium conditions. Since the oxidation progress is linear, the oxide thickness is related to the oxidation time by the rate constant of the reaction.

Boron and phosphorus can cause the linear and parabolic oxidation rate to increase when the Si substrates are doped with these impurities to high concentrations. The effect of boron and phosphorus will be important when oxidizing over diffused p- and n-type layers. In this case the oxide will be thicker than elsewhere on the wafer (Deal and Sklar, 1965). The explanation of the increase in oxidation rate with increasing boron concentration is that boron is preferentially incorporated into the oxide.
into the silica during growth, thereby weakening the bond structure of the silica. Silicon and oxygen bonds are easily broken; oxygen and $H_2O$ molecules enter the silica more easily and diffuse through more easily. The cause of increased oxidation with increasing phosphorus concentration is apparently different from that for boron. There is very little incorporation of phosphorus into the silica. The influence of phosphorus is primarily on the reaction limited regime of Si oxidation, and is possibly due to a Si bond-breaking mechanism at the interface to facilitate the oxidation reaction.

Chlorine is another impurity that has large effects on the characteristics of the oxidation of Si. The presence of chlorine in the oxidation ambient increases the oxidation rate. Chlorine is added to a dry oxygen carrier gas in small amounts, up to 6% by volume, either as chlorine gas, anhydrous HCl, which oxidizes to form chlorine and water vapor, or as an organic molecule, such as trichloroethylene, which oxidizes to release chlorine. Anhydrous HCl is the most common way to introduce chlorine into the oxidation ambient at present. Figure 9.8 shows the effect of chlorine on the oxidation rates of (111) and (100) Si. There are several reasons for the enhanced oxidation rate with the addition of chlorine or HCl (Hirabayashi and Iwamura, 1973): (1) the enhanced diffusion of $O_2$ and $H_2O$ molecules in the oxide containing HCl (the diffusivities of $O_2$ and $H_2O$ are about two to three times higher than those in dry oxide); (2) enhanced reactions at the Si–SiO$_2$ interface due to the catalytic action of HCl; and (3) $H_2O$ molecules created by the reaction

$$2HCl + \frac{1}{2} O_2 \rightleftharpoons H_2O + Cl_2.$$

The oxidation rate is increased due to the enhanced diffusivity of $H_2O$ molecules through the growing oxide.

There are other impurities that may affect the oxidation kinetics of Si. For example, the implantation of about one monolayer of Ge (about $1 \times 10^{10} \text{ cm}^{-2}$) has been found to considerably increase the initial steam oxidation rate of Si (about 1.5 times faster). This is due primarily to the segregation of Ge at the SiO$_2$–substrate interface. As the thickness of oxide increases to the diffusion-limited regime, the oxidation rate is no longer enhanced (Holland et al., 1987).

### 9.4 Oxide Passivation by Chlorine

The addition of chlorine also appears to suppress the formation of stacking faults during oxidation. When wafers are oxidized in dry $O_2$, mixed with a proper amount of HCl, suppression of stacking fault generation and shrinkage of existing stacking faults take place, thus reducing the possibility of stacking faults intersecting $p$–$n$ junctions.

The presence of HCl in oxidizing atmospheres has the effect of passivating the oxide layer. One of the main causes of electrical instabilities observed in MOS structures is the migration of positively charged ions within the oxide. In particular, sodium and other alkali ions can move easily in SiO$_2$ even near room temperature under bias and temperature stressing. In bipolar devices, the presence of heavy metallic impurities in silicon is a primary cause of collector–emitter shorts. It has been observed that the addition of chlorine in the high-temperature phase of mobile ions and metallic impurities results in the removal of chlorine in the furnace atmosphere via the reaction

$$M_{(metal)} + Cl_{(g)} \rightarrow MCl_{(g)}.$$ 

The volatile compounds are then eliminated and recondensed to form the oxide layer. The oxide layer is therefore passivated by chlorine.

### 9.5 Redistribution of Impurities during Thermal Oxidation

It has been found that impurities are redistributed during growing thermal oxide. Some of the impurities may escape from the oxide layer. The concentration of impurities in the Si–SiO$_2$ interface may increase or decrease depending on the conditions. Because of the redistribution, the impurities may change accordingly. The redistribution coefficient $m$ is defined by

$$m = \frac{\text{equilibrium}}{\text{equilibrium}}.$$

When $m > 1$, the oxide rejects the impurity. In addition to the segregation of impurities, there is a diffusion of impurities through the oxide layer (no diffusion of impurities at the interface (Fig. 9.10a). Rapid diffusion and increase the amount of depletion layer. Diffusion through the oxide is rapid compared to the gaseous ambient that the oxide (Fig. 9.10c and d).

Redistribution may still occur if the oxide is not forming at the interface as thick as the amount of Si containing impurities.
Redistribution of Impurities During Thermal Oxidation

It has been found that impurities in the bulk of a Si wafer will redistribute near a growing thermal oxide. Some of the impurity atoms will be incorporated into the oxide layer. The concentration of impurity atoms within the Si located near the Si–SiO₂ interface may increase or decrease depending on the distribution coefficient. Because of the redistribution, the oxidation rate and the interface properties may change accordingly. The redistribution of dopant impurities is characterized by the segregation coefficient m, defined as

\[
m = \frac{\text{equilibrium concentration of impurity in Si}}{\text{equilibrium concentration of impurity in SiO₂}}.
\]  

(9.17)

When \( m > 1 \), the oxide rejects the impurity; when \( m < 1 \), the oxide takes up the impurity. In addition to the segregation coefficient, the redistribution behavior also depends on the diffusion rate of the impurity in the oxide. Figure 9.10 shows four different cases of impurity redistribution in Si due to thermal oxidation. These cases can be classified into two groups. In one group the oxide takes up the impurity \((m < 1)\), and in the other the oxide rejects the impurity \((m > 1)\). In each group the redistribution depends on how rapidly the impurity can diffuse through the oxide. For the case of \( m < 1 \), the silicon surface is depleted of the impurity near the interface (Fig. 9.10a). Rapid diffusion of the impurity through the SiO₂ will increase the amount of depletion (Fig. 9.10b). For the case \( m > 1 \), the diffusion rate of the impurity through the SiO₂ plays an even more significant role. When diffusion through the oxide is rapid, so many impurities may escape from the solid to the gaseous ambient that the overall effect will be a depletion of the impurity (Fig. 9.10c and d).

Redistribution may still occur even if \( m = 1 \). This is because the Si–SiO₂ interface is advancing into the Si substrate, converting the Si into SiO₂ about twice as thick as the amount of Si consumed. Thus more impurity atoms are required to
9.6 Properties of the Si–SiO₂ Interface

Associated with the thermally oxidized silicon system are several types of electrical charges which can adversely affect the device performance, reliability, and yield. These charges are summarized in Fig. 9.11. They include (1) the interface trapped charge \( Q_{it} \), (2) the oxide fixed charge \( Q_f \), (3) the oxide trapped charge \( Q_{ot} \), and (4) the mobile ionic charge \( Q_m \) (Deal, 1980).

The interface trapped charge, \( Q_{it} \), can trap positive or negative charges. High interface states. The origin of these interfaces with high density of surface states. The origin of these states is periodicity of lattice period at the Si–SiO₂ interface. In vacuum, \( Q_n \) is very high (about \( 10^{11} \) cm⁻²) due to oxide growth. For well-prepared surfaces, the oxide density of states can be neutralized by a low-temperature treatment. The density of \( Q_n \) can be as low as \( 10^{10} \) cm⁻².

The oxide fixed charge, \( Q_f \), is fixed in the oxide and cannot be discharged. These positive charges are thought to be fixed at the Si–SiO₂ interface. The value of \( Q_f \) is not greatly affected by the doping concentration in the silicon; however, it is strongly affected by the growth and the annealing processes. The density of Si–SiO₂ interface states is approximately \( 10^{13} \) cm⁻², regardless of the previous history of the wafer. Figure 9.12 shows the effect of wafer orientation and the annealing process on the oxide fixed charge, \( Q_f \). Figure 9.12 also shows that \( Q_f \) increases with 

\[
Q_f 
\]

oxidation at 1200°C followed by cooling in \( Q_f \). Figure 9.12 also shows that the addition of \( Q_f \) ranges between \( 10^{10} \) and 10\(^{12} \) cm⁻².

The oxide trapped charge, \( Q_{ot} \), and holes can be trapped in the oxide by various means, such as x-rays, avalanching in the oxide, and then thermal oxidation. The mobile ionic charge, \( Q_m \), is associated with the ionic charge in the oxide.

FIGURE 9.10 Four different cases of impurity redistribution in silicon due to thermal oxidation. [After Grove et al. (1964).]

FIGURE 9.11 Charges associated with the Si–SiO₂ interface.
The interface trapped charge, $Q_{it}$, is located at the Si–SiO$_2$ interface with either positive or negative charges. Historically, these states are called fast states or surface states. The origin of these states is due primarily to the disruption of the periodicity of the lattice at the Si–SiO$_2$ interface. On clean Si surfaces in ultrahigh vacuum, $Q_{it}$ is very high (about $10^{15}$ states/cm$^2$)—of the order of the density of surface atoms. For well-prepared SiO$_2$ on Si, most of the interface trapped charge can be neutralized by a low-temperature (450°C) hydrogen annealing. The value of $Q_{it}$ can be as low as $10^{10}$ cm$^{-2}$; typical values range between $10^{10}$ and $10^{13}$ cm$^{-2}$.

The oxide fixed charge, $Q_f$, is positive, fixed, and cannot be easily charged or discharged. These positive charges are located within 3 nm of the Si–SiO$_2$ interface. The value of $Q_f$ is not greatly affected by the oxide thickness or by the type or concentration in the silicon; however, it is a strong function of the substrate orientation and the annealing process. It has been suggested that incomplete oxidation of the Si near the Si–SiO$_2$ interface is the origin of the fixed oxide charge. The ratio of $Q_f$ under a given oxidation condition for silicon with (111), (110), and (100) orientation is approximately 3 : 2 : 1. It is also interesting to note that regardless of the previous history of a sample, the final heat treatment determines the value of $Q_f$. Figure 9.12 shows the oxide fixed charge density as a function of annealing, indicating that low $Q_f$ can be obtained by either a high-temperature dry oxidation (1200°C) followed by cooling in an inert ambient, or a low-temperature oxidation (<1000°C) followed by an anneal at the same temperature in either dry nitrogen or argon and then cooling down in the inert gas ambient. It has also been observed that the addition of HCl in O$_2$ reduces fixed charge density. Typical values of $Q_f$ ranges between $10^{10}$ and $10^{12}$ cm$^{-2}$.

The oxide trapped charge, $Q_{ot}$, is located within the oxide layer; both electrons and holes can be trapped in the oxide. These charges are caused by ionizing radiation (such as x-rays), avalanche injection, or by high currents induced through the oxide, thus generating electron–hole pairs in the oxide. In the absence of an
FIGURE 9.12 Oxide fixed charge density as a function of annealing temperature for annealing times of 1 hour or less, showing oxidation–annealing paths. Based on (111) silicon samples. [After Deal et al. (1967).]

electric field, the electrons and holes may recombine. In the presence of an electric field, charge buildup within the oxide may occur. Typical values of \(Q_{\text{ot}}\) range between \(10^{10}\) and \(10^{13}\) cm\(^{-2}\).

The mobile ionic charge, \(Q_{m}\), is due to impurities such as Na\(^+\), Li\(^+\), and K\(^+\). These ions can diffuse easily in the oxide (see Table 7.7). These mobile ions may be neutralized or "gettered" by the addition of chlorine in the oxidizing ambient. The value of \(Q_{m}\) may vary between \(10^{10}\) and \(10^{13}\) cm\(^{-2}\).

The presence of these charges causes the threshold voltage \(V_{T}\) of MOSFET devices to shift to unacceptable values (see Chapter 5). From a processing point of view, the control of these charges is absolutely necessary for reliable operation of MOS devices.

9.7 Deposition of Insulating Layers

In addition to thermal oxidation of silicon to form silicon dioxide, insulating layers such as SiO\(_2\) and Si\(_3\)N\(_4\) can be deposited by a number of chemical vapor deposition (CVD) processes. Chemical vapor deposition systems are either hot wall systems, where the reactors are resistance heated, or cold wall systems, where the reaction zones are heated by radio-frequency (RF) susceptors. These systems can be operated at atmospheric as well as at low pressures (commonly called low-pressure CVD). The reduction of pressure in the reaction chamber increases the mean free path of the reactant species; this leads to more uniform layers with fewer pinholes from slice to slice. The growth rate (about 10 to 15 nm/min) is slower than that for atmospheric pressure reactors (about 50 to 100 nm/min). In some systems, a plasma is used to enhance the chemical deposition temperatures (room-temperature) and uniformity of some of the CVD systems.

Deposited insulating layers are required to serve as insulators between two levels of conductors, to serve as implantation stoppers for diffusion and etch back in the fabrication of Si IC devices, and for the formation of dielectric layers between MOS field applied to the deposited surface to achieve high breakdown fields.

9.7.1 Silicon Dioxide Films

Silicon dioxide films can be formed by many methods. The most commonly used compound is 4-dichloroterephthalic acid, also known as Teepol, which reacts with water to produce a solution of sodium silicate:

\[
\text{Si(C}_{2}\text{H}_{4}\text{O})_{4} + 12\text{O}_2 \rightarrow \text{SiO}_2 + 6\text{CO}_2 + 6\text{H}_2\text{O}
\]

This reaction produces water as a byproduct, which is easily removed from the SiO\(_2\) layer, resulting in a perfect dielectric film.

Better-quality SiO\(_2\) films can be produced at lower temperatures (typically at 450°C) using the plasma-enhanced CVD technique:}

\[
\text{SiH}_4 + \text{O}_2 \rightarrow \text{SiO}_2 + 2\text{H}_2
\]
Summary of Thermal Oxidation
N. Cheung

Consumption of Si substrate
SiO$_2$ is an amorphous structure which contains $2.3 \times 10^{22}$ SiO$_2$ molecules/cm$^3$. Pure Si has an atomic density of $5 \times 10^{22}$ atoms/cm$^3$. Since the Si atoms all come from the Si substrate, for each µm of SiO$_2$ grown, 0.46 µm of Si substrate will be consumed. That means the SiO$_2$/Si interface is 0.46 µm below the original Si surface. In IC processing, oxides grown may be removed locally in different regions followed by re-oxidation at later processing steps. The resultant surface topography is no longer planar.

Grove Model for thermal oxidation
The Grove model gives accurate oxidation rate of thermal oxides if the oxide thickness is above 200 Å. However, thinner oxide growth will need more advanced models or based on empirical data. By balancing the oxygen flux coming the gas stream, the oxygen diffusion flux across the oxide already present and the oxygen flux consumed at the SiO$_2$/Si interface, we arrive at:

$$x_{ox}(t) = \frac{1}{2A} x_{ox}^2(t) + Ax_{ox}(t) = B (t + \tau) \quad \left[1 + \frac{4B(t+\tau)}{A^2}\right]^{1/2} - 1$$

where $\tau$ is an effective growth time corresponding to the initial oxide thickness: $\tau = \frac{x_i^2 + Ax_i}{B}$

The following are major conclusions from the Grove model:

0) The growth rate $\frac{dx_{ox}}{dt} = \frac{B}{A+2x_{ox}}$ slows down as $x_{ox}$ increases.

1) Oxidation rate can be characterized by the linear constant $B/A$ and the parabolic constant $B$ with:

$$B = P \times D_1 e^{E_1/kT} \quad B/A = P \times D_2 e^{E_2/kT}$$

where $P$ is the oxygen (or steam) pressure in atmospheres and $D_1, E_1, D_2, E_2$ are constants.

2) Oxidation rate slows down as the oxide grows to greater thickness

3) Presence of any initial oxide will yield less additional oxide than bare Si surface

4) For short oxidation times (i.e. thin oxide) with $(t+\tau) << A^2/4B$, $x_{ox}(t) = (B/A) (t+\tau)$ (linear with oxidation time). For long oxidation times (i.e. very thick oxide) with $(t+\tau) >> A^2/4B$, $x_{ox}(t) = \sqrt{Bt} \propto \sqrt{t}$

5) If initial oxide is deposited down by chemical vapor deposition (i.e., CVD oxide), as a first-order approximation, we can still them as thermal oxide to evaluate $\tau$ when further thermal oxidation is carried out.

Dependence of $B/A$ and $B$ on Processing Parameters

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Linear Constant $B/A$</th>
<th>Parabolic Constant $B$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Oxidation Pressure</td>
<td>linear with oxygen pressure (actually $\propto P^{0.8}$)</td>
<td>linear with oxygen pressure</td>
</tr>
<tr>
<td>Steam versus O$_2$</td>
<td>larger for steam oxidation</td>
<td>larger for steam oxidation</td>
</tr>
<tr>
<td>Si crystal orientation</td>
<td>$B/A(111):B/A(100) = 1.68:1$</td>
<td>independent of orientation</td>
</tr>
<tr>
<td>Dopant type and concentration in Si</td>
<td>increases with dopant concentration</td>
<td>insensitive</td>
</tr>
<tr>
<td>Addition of Cl-containing gas in oxidation ambient</td>
<td>insensitive</td>
<td>increases</td>
</tr>
</tbody>
</table>

Why steam oxidation is faster than dry oxidation

Although the diffusion constant $D$ for O$_2$ in SiO$_2$ is higher than that of H$_2$O in SiO$_2$, the higher values of the rate constants, $B$ and $B/A$, result from the fact that the value of $C^*$ is much higher for H$_2$O ($\sim 3 \times 10^{19}$/cm$^3$) than O$_2$ ($\sim 5 \times 10^{16}$/cm$^3$).
<table>
<thead>
<tr>
<th>Wet O₂ (X₁ = 0 nm)</th>
<th>Dry O₂ (X₁ = 25 nm)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>D₉</td>
</tr>
<tr>
<td>c₁ (100) silicon</td>
<td></td>
</tr>
<tr>
<td>Linear (B/A)</td>
<td>9.70 x 10⁶ µm/h</td>
</tr>
<tr>
<td>Parabolic (B)</td>
<td>386 µm³/h</td>
</tr>
<tr>
<td>c₂ (111) silicon</td>
<td></td>
</tr>
<tr>
<td>Linear (B/A)</td>
<td>1.63 x 10⁶ µm/h</td>
</tr>
<tr>
<td>Parabolic (B)</td>
<td>386 µm³/h</td>
</tr>
</tbody>
</table>

*Data from ref. [7].

**Wet and dry silicon dioxide growth**

---

**Temperature (°C)**

- **H₂O (640 torr)**
  - Eₐ = 0.78 eV

**Dry O₂**
- Eₐ = 1.23 eV

**1000/T (K⁻¹)**

---

**Wet and dry silicon dioxide growth for (100) silicon**

---

**Wet and dry silicon dioxide growth for (111) silicon**

---
Week 4

Ion Implantation
Typos for Chap 8 Ion Implantation

Page 222, bottom paragraph

The beam current = flux $F_i \bullet$ beam area $\bullet$ electronic charge/ion

Page 223, Equation 8.2

The integral expression should be multiplied by the beam area
8.1 Introduction

Integrated-circuit technology requires reproducible and controlled dopant concentrations in all active-device components, such as bipolar transistors and field-effect transistors. In a 5 × 5 mm Si chip containing memory arrays and logic elements, there may be a million transistors—all of which must be functional and properly doped with donors and acceptors. Dopants can be introduced by diffusion from gas ambients in quartz-tube diffusion furnaces as discussed in Chapter 7. For dopant control in diffusion, the surfaces must be clean and exposed to a uniform flow of dopants in the gas ambient; that is, the gas source of dopants must provide a uniform flux of dopants impinging on the semiconductor surface, and the transport of dopants across the gas—semiconductor surface should not be impeded by localized patches of contaminants or native oxide layers. From a production-line technology standpoint, the requisite control of surfaces and gas flow is difficult to achieve reproducibly to ensure that wafers after wafer receive the same amount of dopants.

To achieve a uniform and controlled introduction of dopants, ion implantation is used. In ion implantation, a beam of dopant ions of fixed energy—typically between 30 and 100 kiloelectron volts (keV)—is swept (or rastered) across the surface of the semiconductor (Fig. 8.1). The ions have a sufficiently high velocity, about 10^7 cm/s, so that they penetrate through the surface and come to rest within the semiconductor at depths of 10 to 100 nm below the surface. The penetration depths—the ion ranges—are determined by the energy of the incident ion and are not blocked by surface contamination or native oxide layers. Consequently, the total number \( Q_i \) of implanted dopants/cm^2 is given by the product of the flux \( F_i \) of incident ions/cm^2 · s and the implantation time \( t_i \) that the ion beam is incident on the sample:

\[
Q_i = F_i t_i \tag{8.1}
\]

A flux \( F_i \) of ions with positive charge \( q_i \) per ion represents a current that can be measured directly with a current meter connected between the sample and electrical ground (Fig. 8.1). The ion current \( q_i F_i \) may vary with time so that in ion-implantation systems, a charge integrator is used to give the total integrated charge

\[
Q_i = \frac{100 \times 10^{-6}}{200 \times 1.6 \times 10^{15}} \text{ cm}^2
\]

where \( q_i = 1.6 \times 10^{-19} \text{ C} \) for doubly ionized species. The number of ions swept across a 200-cm^2 area is

\[
Q_i = \frac{100 \times 10^{-6}}{200 \times 1.6 \times 10^{15}} \text{ cm}^2
\]

By control of the beam current and time, a concentration of \( 10^{15} \) and 5 × 10^{15}/cm^2 can be achieved. The uniformity of the dose is controlled between values of 100 and 1000 ppm over a range of six orders of magnitude in implantation time.
controlled dopant concent-
trates and field-effect trans-
munity arrays and logic ele-
ments, must be functional and properly intro-
duced by diffusion from gas
cussed in Chapter 7. For dopant
and exposed to a uniform flow of
dopants must provide a uniform
surface, and the transport of do-
not be impeded by localized
on a production-line technology
gas flow is difficult to achieve
the same amount of dopants.
Ion of dopants, ion implantation
ions of fixed energy—typically
swept (or rastered) across the
have a sufficiently high velocity,
the surface and come to rest within
low the surface. The penetration
energy of the incident ion and are
oxide layers. Consequently, the
by the product of the flux $F_i$
that the ion beam is incident

(8.1)

ion represents a current that can
between the sample and elec-
ny vary with time so that in ion-
give the total integrated charge

FIGURE 8.1 Schematic of an ion-implantation system target chamber showing
the ion beam incident on a Si wafer and the current integrator.

\[ Q_t = \int_0^t q_i F_i \, dt \]  

(8.2)

where $q_i = 1.6 \times 10^{-19}$ C for singly ionized ions and $q_i = 3.2 \times 10^{-19}$ C for
doubly ionized species. The number $Q_t$ of implanted dopants/cm$^2$ is given by

\[ Q_t = \frac{Q_i}{q_i A} \]  

(8.3)

where $A$ is the area of the implanted surface. A 100\,$\mu$A beam of singly ionized
ions swept across a 200-cm$^2$ area for 60 s gives

\[ Q_t = \frac{100 \times 10^{-6} \times 60}{200 \times 1.6 \times 10^{-19}} = 1.875 \times 10^{14} \text{ dopant ions/cm}^2. \]

By control of the beam current and implantation time, values of $Q_t$ between $5 \times 10^9$ and $5 \times 10^{13}$/cm$^2$ can be obtained for different applications. If the dopants
are distributed uniformly over 50 nm (500 Å), the dopant concentrations can be
controlled between values of $10^{15}$ and $10^{21}$ dopants/cm$^3$. This illustrates one of
the advantages of ion implantation—that dopant concentrations can be obtained
over a range of six orders of magnitude by adjusting the ion beam current and
implantation time.
FIGURE 8.2 Ion implantation into a Si wafer, where portions of the sample are masked by an SiO₂ layer whose thickness is greater than the ion range.

Another advantage of ion implantation is that selected areas can be implanted by use of masks which leave well-defined areas of the semiconductor exposed to the beam and other areas masked from the beam (Fig. 8.2). The thickness of the masks must be greater than the penetration depth or range of the ions. However, ion ranges are typically less than 100 nm, so that mask thicknesses of 200 nm serve to prevent the ions from reaching the semiconductor. The masks may be thermally grown oxide layers (SiO₂ on Si) or deposited layers of oxides, organic films, or metal layers. The mask layers are patterned by photolithographic techniques, as used in other stages of integrated circuit processes, and the mask material is removed in areas where the semiconductor wafer is to be exposed to the ion beam. After implantation, all the mask material is removed so that the wafer can receive further processing steps.

The penetration of energetic ions into the semiconductor damages the crystal structure. The ions collide with semiconductor atoms and displace them off lattice sites. The disorder can be sufficiently great so that an amorphous layer is formed. High-temperature (600 to 1000°C) processing is required to anneal the disorder and return the implanted semiconductor to a single-crystal state with a minimum number of defects.

8.2 Ion-Implantation Systems

The implantation system in concept is quite simple (Fig. 8.3) as it consists of an ion source, an acceleration tube, and a target chamber where the semiconductor wafers are held. Dopant atoms are introduced into the ion-source container from a gas or a vapor supplied by a liquid or solid heated in a small oven connected to the source. The dopant atoms in the source are ionized by energetic electrons emitted from a hot filament. The electrons collide with the cloud of atomic electrons around the positive nucleus of the dopant atom. In the collisions, electrons are knocked out of their atomic orbits and the atom is left in an ionized state with a positive charge of e (singly ionized) or 2e (doubly ionized). The positive dopant ions are then extracted through an aperture in the ion source and enter the acceleration tube.

The acceleration tube is an insulating column with a vacuum inside so that the ions from the source can be accelerated without suffering impact collisions with residual gas atoms in the column. The ion source is held at a positive voltage, the acceleration voltage potential $V_a$, and the length of column $L$, hence

$$v_f = \left(\frac{2 \times 10^5 \text{ eV}}{74.9 \times 1.6 \times 10^{-19} \text{ C}}\right)$$

where $M_i$ is the mass of the dopant (e.g., 40 for Si, 74.9 atomic mass units) acceleration potentials.

Ion-implantation systems are now available for production development. Every aspect of an ion-implanter has been refined to provide consistent and reliable operation. The basic schematic layout of a common system is shown in Fig. 8.3. The layout is immediately obvious: the extent from source to analyzing magnet. A large array of ion-implanters lines as many as 300 wafers at a time. Finally, the ion-implanter requires automatic placement of wafers into the source and out of the target chamber for removal.

An analyzing magnet is required to separate the dopant ion species but only one species is implanted at any given time. Ions with different masses and charge states are separated by their magnetic rigidity, $B$ = $\frac{mv}{q}$, where $m$ is the mass of the ion, $v$ is the velocity, and $q$ is the charge. Magnetic rigidity is inversely related to the magnetic field directed normal to their trajectories. This concept is discussed in Chapter 2, the forces...
where portions of the sample are larger than the ion range.

Selected areas can be implanted with the semiconductor exposed to (Fig. 8.2). The thickness of the mask of the ions. However, that mask thicknesses of 200 nm miconductor. The masks may be deposited layers of oxides, organic photoresists, or chemical techniques, and the mask material wafer is to be exposed to the ion is removed so that the wafer can miconductor damages the crystal ions and displace them off lattice at an amorphous layer is formed. required to anneal the disorder and the crystal state with a minimum number (Fig. 8.3) as it consists of a chamber where the semiconductor to the ion-source container from a heated in a small oven connected to be ionized by energetic electrons with the cloud of atomic electrons. In the collisions, electrons are (is left in an ionized state with a singly ionized). The positive dopant is the ion source and enter the acceleration voltage potential \( V_i \). The ions entering the column are accelerated along the length of column and exit into the target chamber with an energy \( q_i V_i \) and velocity \( v_i \):

\[
\frac{1}{2}m_i v_i^2 = q_i V_i
\]

where \( M_i \) is the mass of the dopant atom. For singly ionized As ions \( [M_i(As) = 74.9 \text{ atomic mass units}] \) accelerated through 100 kilovolts (kV), the velocity is

\[
v_i = \left( \frac{2 \times 10^5 \text{ eV} \times 1.6 \times 10^{-19}}{74.9 \times 1.66 \times 10^{-27} \text{ kg}} \right)^{1/2} = 5.07 \times 10^5 \text{ m/s}
\]

where one atomic mass unit = 1.66 \times 10^{-27} \text{ kg} and 1 eV = 1.6 \times 10^{-19} \text{ J}.

Ion-implantation systems are quite elegant and sophisticated after 20 years of development. Every aspect of the system from ion source to target chamber has been refined to provide controlled implantation conditions. Figure 8.4 shows a schematic layout of a commercial ion-implantation system. Two features are immediately obvious: the extent of the wafer transport area and the presence of an analyzing magnet. A large area is required for wafer transport because in production lines as many as 300 wafers per hour may be implanted. This wafer throughput requires automatic placement of wafers onto wafer holders as well as transport in and out of the target chamber.

An analyzing magnet is required because the ion source produces a variety of ion species but only one species (e.g., singly ionized As ions) is selected for implantation. Ions with charge \( q_i \) and velocity \( v_i \) will be deflected by a magnetic field directed normal to their path. As in the case of Hall effect measurements discussed in Chapter 2, the force \( \mathbf{F} \) on the ions is

\[
\mathbf{F} = q_i (v_i \times \mathbf{B})
\]
where \( B \) is the magnetic flux density. Charged particles in a homogeneous magnetic field move in a circular path of radius \( r \) as indicated in Fig. 8.5 for a magnetic field directed out of the page. The velocity \( v_i \) of the particle of mass \( M_i \) is unchanged by the deflection in the magnetic field and there is a centripetal force

\[
F = \frac{M_i v_i^2}{r}.
\] (8.6)

Combining eqs. (8.5) and (8.6), we have

\[
 r = \frac{M_i v_i}{q_i B}.
\] (8.7)

For an ion accelerated through a potential \( V_i \), the velocity \( v_i = (2q_i V_i/M_i)^{1/2} \), so that

\[
r = \frac{1}{B} \left( \frac{2M_i V_i}{q_i} \right)^{1/2}.
\] (8.8)

FIGURE 8.4 Schematic of a commercial ion-implantation system, the Nova-10-160, 10 mA at 160 keV.

For a given acceleration potential \( V_i \), the path is directly proportional to the radius of the circular path. Slightly ionized As ions accelerated through a 100-kV source have a radius of curvature of 5000 gauss, the radius is

\[
r = \frac{1}{0.5} \left( \frac{2 \times 74.9 \times 1}{1.6} \right).
\]

Consequently, magnetic analysis can be used for a reasonably compact magnet if the ion species are those that can be focused by a magnet with reasonable dimensions. With proper design, one ion species can be delivered into a wafer with high resolution.

8.3 Ion-Range Distribution

When an energetic ion penetrates a material, the distribution of energy loss with the atoms and electrons in the material is described by the energy loss function \( dE/dx \) of the ion. The energy lost by an ion as it moves through a material can be calculated by the energy loss mechanisms are discussed in the text. The penetration depth or range \( R \) of the ion in the material is the distance traveled before the rate of energy loss along the range

\[
R = \int_{0}^{R} \sqrt{2M_i \frac{dE}{dx}}
\]

where \( E_i \) is the incident energy of the ion. The sign of \( dE/dx \) is negative, as it is for all ions, although tabulated values are given for various materials.

FIGURE 8.5 Schematic of an analyzing magnet deflecting an ion beam.

FIGURE 8.6 An ion incident on a semiconductor surface gives a projected characteristic range.
For a given acceleration potential and magnetic flux density, the radius of the ion path is directly proportional to the square root of the mass-to-charge ratio. For singly ionized As ions accelerated through 10 kV, in a magnetic field of 0.5 Wb/m² (5000 gauss), the radius is

\[
r = \frac{1}{0.5} \left( \frac{2 \times 74.9 \times 1.66 \times 10^{-27} \text{ kg} \times 10^4}{1.6 \times 10^{-19}} \right)^{1/2} = 0.25 \text{ m}.
\]

Consequently, magnetic analysis of the ion beam can be carried out with a reasonably compact magnet if the ion energies or ion masses are not too great.

Aside from deflection in the magnetic analyzer for selection of the correct ion species, the ions are focused and steered along their paths by electrostatic lens and deflection plates. With proper design of ion optics, a monoenergetic ion beam of one ion species can be delivered into the target chamber.

### 8.3 Ion-Range Distributions

When an energetic ion penetrates a semiconductor it undergoes a series of collisions with the atoms and electrons in the target. In these collisions the incident particle loses energy at a rate \( dE/dx \) of a few to 100 eV per nanometer, depending on the energy and mass of the ion as well as on the substrate material. The energy-loss mechanisms are discussed in the next section, as we are concerned here with the penetration depth or range \( R \) of the ions (Fig. 8.6). The range \( R \) is determined by the rate of energy loss along the path of the ion,

\[
R = \int_{E_0}^{0} \frac{1}{dE/dx} \, dE \tag{8.9}
\]

where \( E_0 \) is the incident energy of the ion as it penetrates the semiconductor. The sign of \( dE/dx \) is negative, as it represents the energy loss per increment of path, although tabulated values are given as positive quantities.

**FIGURE 8.6** An ion incident on a semiconductor penetrates with a total path length \( R \) which gives a projected range \( R_p \) along the direction parallel to that of the incident ion.
The main parameters governing the range or energy loss rate are the energy $E_0$ and atomic number $Z_1$ of the ion and atomic number $Z_2$ of the substrate if we exclude the effect of the orientation of the crystal lattice. As the incident ion penetrates the semiconductor undergoing collisions with atoms and electrons, the distance traveled between collisions and the amount of energy lost per collision are random processes. Hence all ions of a given type and incident energy do not have the same range. Instead there is a broad distribution in the depths to which individual ions penetrate. The distribution in ranges is referred to as the range distribution or range straggling. Further, in ion implantation it is not the total distance $R$ traveled by the ion that is of interest but the projection of $R$ normal to the surface (i.e., the penetration depth or projected range $R_p$) (Fig. 8.6).

In the absence of crystal orientation effects, the range distribution is roughly Gaussian and as a first approximation we describe the projected range distribution $N(x)$ as a one-dimensional Gaussian characterized by a mean value, the projected range $R_p$, and a standard deviation $\Delta R_p$ from the mean:

$$N(x) = \exp \left[ -\frac{1}{2} \left( \frac{x - R_p}{\Delta R_p} \right)^2 \right]. \quad (8.10)$$

\[\text{FIGURE 8.7 Gaussian range distribution for implanted ions with } R_p = 2.35 \Delta R_p \text{ and a full width at half-maximum (FWHM) of } \Delta x_p.\]

For a Gaussian distribution (FWHM) is given by

$$\Delta x_p = 2(2\ln 2)^{1/2} \Delta R_p,$$

and the integral by

$$\int_0^\infty N(x) dx = 1.$$

In ion implantation, the number of ions $N(x)$ in atoms/cm$^3$ is given by

$$N(x) = \frac{Q_0}{(2\pi)^{1/2}} \exp \left[ -\frac{1}{2} \left( \frac{x - R_p}{\Delta R_p} \right)^2 \right].$$

A rough estimate of the average projected range obtained from

$$R_p = \frac{1}{N(x)} \int_0^\infty x N(x) dx = \frac{3}{4} \Delta R_p,$$

where $\Delta x_p = R_p$ for medium-mass ions.

Silicon dioxide, $SiO_2$, is often used to implant a dose of $10^{15}$ As ions/cm$^2$, $R_p$ is $10^{-5}$ cm (1000 Å), the value of $\Delta x_p = 10^{-3}$, so that the average projected range is $3 \times 10^{-3}$ cm. Thus one of the features of ion implantation is that the projected range is a reasonable rule of thumb. This is a key factor in determining the projected range of ions in $SiO_2$ as equal to $3 \times 10^{-3}$ cm.

In addition to the straggling in the surface, there is a fluctuation in the projected range. The fluctuation is defined by the range distribution at the surface. Implants into Si with a thick metal.
energy loss rate are the energy $E_0$ number $Z_2$ of the substrate if we
crystal lattice. As the incident ion ions with atoms and electrons, the sum of energy lost per collision
projectile and incident energy do not distribution in the depths to which rages is referred to as the range
implantation it is not the total but the projection of $R$ normal to the range $R_p$ (Fig. 8.6).
the range distribution is roughly the projected range distribution by a mean value, the projected mean:

$$
\frac{R_p}{R_p^2} \left( x - R_p \right)^2.
$$

(8.10)

For a Gaussian distribution (Fig. 8.7), the full width $\Delta x_p$ at half-maximum (FWHM) is given by

$$
\Delta x_p = 2(2 \ln 2)^{1/2} \Delta R_p = 2.35 \Delta R_p
$$

(8.11)

and the integral by

$$
\int_0^\infty N(x)dx = (2\pi)^{1/2} \Delta R_p.
$$

(8.12)

In ion implantation, the number $Q_i$ implanted ions/cm$^2$ is tightly controlled, and the number of relations in eqns. (8.10) and (8.12), the concentration depth distribution $N_f(x)$ in atoms/cm$^3$ is given by

$$
N_f(x) = \frac{Q_i}{(2\pi)^{1/2} \Delta R_p} \exp \left[ -\frac{1}{2} \left( \frac{x - R_p}{\Delta R_p} \right)^2 \right].
$$

(8.13)

A rough estimate of the average concentration $N_f(ave)$ of implanted ions can be obtained from

$$
N_f(ave) = \frac{Q_i}{\Delta x_p} = \frac{Q_i}{R_p}
$$

(8.14)

where $\Delta x_p = R_p$ for medium-mass ions such as As or P implanted into Si. If one implants a dose of $10^{15}$ As ions/cm$^2$ into Si at 200 keV so that the projected range $R_p$ is $10^{-5}$ cm (1000 Å), the value of the range straggling $\Delta R_p = 0.4 \times 10^{-5}$ cm and $\Delta x_p = 10^{-5}$, so that the average concentration $10^{15}/10^{-5} = 10^{20}$ As/cm$^3$. Thus one of the features of ion implantation is that high dopant concentrations can be achieved in the near-surface region—an ideal situation in the processing of integrated circuit components.

Projected ranges $R_p$ and range straggles $\Delta R_p$ of dopant ions in Si and GaAs are shown in Fig. 8.8. One sees immediately that the lighter ions such as boron, with their higher velocities, have a greater penetration in silicon than the heavier ions such as As. Further, the ranges of the ions increase nearly linearly with energy. Comparison of $R_p$ and $\Delta R_p$ values show that the expression $R_p = 2.35 \Delta R_p = \Delta x_p$ is a reasonable rule of thumb. The ranges of ions in GaAs are somewhat less than those at the same energy in Si.

Silicon dioxide, SiO$_2$, is often used as a masking material to prevent the penetration of ions into the Si below regions covered by SiO$_2$. The molecular density is $2.2 \times 10^{22}$ SiO$_2$ molecules/cm$^3$ and the mass density is 2.2 g/cm$^3$, a value close to that of Si, 2.33 g/cm$^3$. A very good approximation is to treat the projected ranges of ions in SiO$_2$ as equal to those in Si: $R_p(\text{SiO}_2) = R_p(\text{Si})$.

In addition to the straggles in the penetration depth $\Delta R_p$ along a direction normal to the surface, there is a fluctuation, or transverse straggles $\Delta R_t$, in the final position perpendicular to the incident ion direction. Transverse straggles is of importance in defining the penetration at the edge of a mask. The spread of ion distributions implanted into Si with a thick mask ($t > R_p$) with straight edges is shown in Fig.
8.9. The lateral distribution is influenced in the region within a few \( \Delta R_p \) of the edge of the mask. The values of \( \Delta R_p \) are somewhat greater than \( \Delta R_p \), but a first approximation is to set \( \Delta R_p = \Delta R_p \). Tapered or undercut mask edges will also lead to a lateral distribution.

The description of the ion distribution by use of \( R_p \) and \( \Delta R_p \) can be improved by the use of higher-order moments. A three-moment approach using two Gaussians with the same \( R_p \) but different \( \Delta R_p \) values has been used to fit experimental data. A more exact fit is found with a fourth moment approach: \( R_p \), \( \Delta R_p \), the third moment (skewness), and the fourth moment (kurtosis).

With single-crystal substrates of Si or GaAs, for example, the orientation of the ion beam with respect to the crystallographic axes of the substrate can have a pronounced effect on the range distribution. Figure 8.10 shows the range distribution for 100-keV As implanted with the beam aligned parallel (solid line) to the
the region within a few \( \Delta R \), of the few greater than \( \Delta R_{p} \), but a first undercut mask edges will also lead to the use of \( R_{p} \) and \( \Delta R_{p} \) can be improved approach using two Gaussians been used to fit experimental data. approach: \( R_{p} \), \( \Delta R_{p} \), the third (surtosis).

For example, the orientation of the axes of the substrate can have a Figure 8.10 shows the range distribution of nonaligned parallel (solid line) to the

---

**FIGURE 8.9** Schematic of the lateral distribution of ions implanted into Si with straight-wall oxide masks with a projected range \( R_{p} \) and lateral straggles \( \Delta R_{l} \).

**FIGURE 8.10** Range distributions for channeled ions implanted along the \( <100> \) axis of Si. The dashed line shows the Gaussian distribution for incident ions aligned away from any channeling direction.
<100> crystal axis and oriented away (dashed line) from any crystal axes or planes. As is evident in the figure, implantation along crystal axes can lead to a fraction of the total number of ions that penetrate several times \( R_p \).

The crystal orientation influence on ion penetration is called channeling or the channeling effect. When an ion trajectory is aligned along atomic rows, the positive atomic potentials of the line of atoms steer the positively charged ion within the open space or channels between the atomic rows. These channeled ions do not make close-impact collisions with the lattice atoms and have a much lower rate of energy loss, \( dE/dx \), and hence a greater range than those of nonchanneled ions. The depth distribution of channeled ions is difficult to characterize under routine implantation conditions. The channeling distribution depends on surface preparation, substrate temperature, beam alignment, and disorder introduced during the implantation process itself. The channeled ions that penetrate beyond \( R_p \) often have a distribution that falls off exponentially with distance as \( \exp \left( -x/\lambda_c \right) \), where \( \lambda_c >> R_p \). Since the concentration of implanted atoms near \( R_p \) is often large (\( \approx 10^{19} \) to \( 10^{20} \) atoms/cm\(^3\)), the junction depth is determined by the penetration of channeled ions.

The channeling effect requires that the incident ions be aligned within a critical angle \( \psi_c \) of the crystal axes or planes. The critical angle depends on the ion energy, ion species, and substrate, but is typically less than 5°. Consequently, the substrate holders are often tapered so that the wafers are mounted 7° off normal; this minimizes channeling effects. However, some ions originally incident at angles greater than the critical angle can be scattered into a channeling direction. It is difficult to avoid channeling effects completely unless the implanted region has been made amorphous by a previous implantation.

### 8.4 Energy-Loss Processes

The energy-loss rate \( dE/dx \) of an energetic ion moving through a solid is determined by collisions with the substrate atoms and electrons (Fig. 8.11). It is customary to distinguish two different mechanisms of energy loss: (1) nuclear collisions, in which energy is lost in displacements of the substrate atoms, and (2) electronic collisions, in which the energetic ion excites or ejects electrons from atomic orbitals. The energy loss rate \( dE/dx \) can be expressed as

\[
\frac{dE}{dx} = \frac{dE}{dx}_n + \frac{dE}{dx}_e
\]

where the subscripts \( n \) and \( e \) denote nuclear and electronic collisions, respectively.

Nuclear collisions can involve large discrete energy losses and significant angular deflection of the trajectory of the ion (Fig. 8.11). This process is responsible for the production of lattice disorder by the displacement of atoms from their lattice position. Electronic collisions involve much smaller energy loss per collision, negligible deflection of the ion trajectory, and negligible lattice disorder. As shown in Fig. 8.12, the relative importance of nuclear and electronic collisions at high energy \( E_i \) in which nuclear collisions dominate.

\[
\frac{dE}{dx}_n = \frac{dE}{dx}_e
\]

The values in Table 8.1 indicate that the relative importance of nuclear and electronic collisions is significant in determining the rate of energy loss over much of the trajectory.
moving through a solid is determined by the penetration of ions (Fig. 8.11). It is customary to consider loss: (1) nuclear collisions, in which atomic electrons, and (2) electronic collisions, respectively. The energy losses and significant another cause of energy loss per collision, negligible lattice disorder. As shown in

\[ \frac{dE}{dx} \]

Fig. 8.11, the relative importance of the two processes depends on the energy \( E \) (and atomic number \( Z \) of the ion): Nuclear collisions dominate at low energies and electronic collisions at high energies. Values are given in Table 8.1 for the energy \( E_1 \) in which nuclear collisions are maximum and for the energy \( E_2 \) at which \( \frac{dE}{dx} \) = \( \frac{dE}{dx} \) _n_.

The values in Table 8.1 indicate that both energy loss processes must be taken into account in determining the range of ions. For light ions in Si such as 100-keV boron, electronic collisions will be dominant over all the trajectory. For 100-keV As, nuclear energy loss can be a factor of 5 to 10 greater than electronic energy loss over much of the trajectory.

\[ \frac{dE}{dx} \text{ versus (Energy)}^{1/2} \]

FIGURE 8.11 An ion incident on a crystal lattice is deflected in nuclear collisions with the lattice atoms and also loses energy in collisions with electrons.

FIGURE 8.12 Rate of energy loss \( \frac{dE}{dx} \) versus (energy)\(^{1/2} \); showing nuclear and electronic loss contributions.
The relative magnitude of the two energy-loss mechanisms in Si is shown directly in Fig. 8.13. Both Figs. 8.12 and 8.13 show that the rate of electronic energy loss is proportional to the ion velocity

$$\frac{dE}{dx} = k_e(E)^{1/2}$$

(8.16)

where the coefficient $k_e$ does not depend on $E$. For Si, $k_e = 1$ in units of (eV)$^{1/2}$ per nanometer.

In the energy region near the range $E$, the energy loss $dE/dx$ may be treated as energy independent, with atomic density $N = 5 \times 10^{22}$ cm$^{-3}$.

$$\left.\frac{dE}{dx}\right|_n =$$

or

$$\left.\frac{dE}{dx}\right|_n =$$

with the screening radius $a = 1.4$ nm.

Subscripts 1 and 2 refer to ion and atom number and mass, respectively. The range of energetic particles undergoing stopping is calculated (see Feldman and Mayer, 1986). For Si ($Z_t = 14$, $M_s = 28$), eq. (8.14) gives $E = 550$ keV/μm, a value close to $R_p$.

In the case where nuclear stopping is the dominant mechanism, the range is estimated by

$$R_p = \int_{E_n}^{E} \left.\frac{dE}{dx}\right|_n$$

The estimates given in eq. (8.11) do give a quick approximation to $R_p$ [eq. (8.11)]. For more accurate values, eq. (8.14) should be used.

### 8.5 Implantation Damage

As an energetic ion slows down in a crystal, the number of collisions with the lattice may be transferred from the ion to the displaced atoms, which can then disperse forming a cascade of atomic collisions. This cascade consists of amorphous regions, and amorphous bands. As shown in Fig. 8.14, a heavy ion in electronic collisions and has an amorphous layer along its path. A heavy ion such as a He ion produces a dense cascade of atoms, and hence produces a dense cascade track. For both heavy and light ions, a dense cascade extends from the surface to the target material.
where the coefficient $k_e$ does not depend strongly on the ion species. The value of $k_e = 1$ in units of (eV)$^{1/2}$ per nanometer.

In the energy region near the maximum in nuclear energy loss, the value of $dE/dx$ may be treated as energy independent, with a value for stopping in a medium with atomic density $N = 5 \times 10^{22}$ atoms/cm$^3$ given by

$$
\frac{dE}{dx}\bigg|_n = N \frac{T_e}{2} e^2 a \frac{Z_1 Z_2 M_1}{M_1 + M_2}
$$

or

$$
\frac{dE}{dx}\bigg|_n = \frac{S Z_1 Z_2 M_1}{M_1 + M_2} \frac{eV}{nm}
$$

with the screening radius $a = 1.4 \times 10^{-2}$ nm, $e^2 = 1.44$ eV - nm, where the subscripts 1 and 2 refer to ion and substrate, and where $Z$ and $M$ are the atomic number and mass, respectively. This relation is derived for the rate of energy loss of energetic particles undergoing screened Coulomb collisions with a $1/r^2$ potential (see Feldman and Mayer, 1986). For phosphorus ions ($Z_1 = 15$, $M_1 = 31$) incident on Si ($Z_2 = 14$, $M_2 = 28$), eq. (8.17) gives a value for $dE/dx|_n = 550$ eV/nm = 550 keV/µm, a value close to that shown in Fig. 8.13.

In the case where nuclear stopping predominates, the projected range $R_p$ can be estimated by

$$
R_p = \int_{E_0}^{0} \frac{1}{dE/dx} \, dE = \frac{E_0}{dE/dx|_n}.
$$

The estimates given in eq. (8.16) to (8.18) are accurate only to 20 to 40% but do give a quick approximation to the range distribution if one uses $\Delta R_p = R_p/2.35$ [eq. (8.11)]. For more accurate values of $R_p$ and $\Delta R_p$, the values given in Fig. 8.8 should be used.

### 8.5 Implantation Damage

As an energetic ion slows down and comes to rest in a semiconductor, it makes a number of collisions with the lattice atoms. In these collisions, sufficient energy $T$ may be transferred from the ion to displace an atom from its lattice site. The displaced atoms can in turn displace other atoms, and so on—thus creating a cascade of atomic collisions. This leads to a distribution of vacancies, interstitial atoms, amorphous regions, and other types of defects as a result of nuclear collisions. As shown in Fig. 8.14, a light ion such as boron in Si loses energy primarily in electronic collisions and has only occasional large-energy transfer collisions along its path. A heavy ion such as Sb loses energy primarily in nuclear collisions and hence produces a dense cascade of collisions and hence disorder along the track. For both heavy and light ions, however, an amorphous layer can be formed, extending from the surface to the ion penetration depth, for sufficiently high num-
bers of implanted ions. In this amorphous layer there is no long-range crystallographic order, but the nearest neighbors can maintain their covalent, tetrahedral bonding with some distortion of the bond angles and lengths. In compound semiconductors such as GaAs, the collisions produce an intermixing of the Ga and As atoms so that localized, nonstoichiometric regions, Ga- and As-rich, are formed within the amorphous layer.

The amount of energy $T$ transferred can be determined by considering the kinematics of a collision between an energetic particle, identified by the subscript 1, and a stationary target atom, identified by subscript 2 (Fig. 8.15).

The energy transfers in collisions between two isolated particles can be solved by applying the principles of conservation of energy and momentum. For an incident energetic particle of mass $M_1$, the values of the velocity and energy are $v$ and $E_0$ ($E_0 = \frac{1}{2} M_1 v^2$), while the target atom of mass $M_2$ is at rest. After the collision, the values of the velocities $v_1$ and $v_2$ and energies $E_1$ and $E_2$ of the projectile and target atoms are determined by the scattering angle $\theta$ and recoil angle $\phi$. The notation and geometry for the laboratory system of coordinates are given in Fig. 8.15.

Conservation of energy and conservation of momentum parallel and perpendicular to the direction of incidence are expressed by the equations

$$\frac{1}{2} M_1 v^2 = \frac{1}{2} M_1 v_1^2 + \frac{1}{2} M_2 v_2^2$$  \hspace{1cm} \text{(8.19)}$$

$$M_1 v = M_1 v_1 \cos \theta + M_2 v_2 \cos \phi$$  \hspace{1cm} \text{(8.20)}$$

From these equations, the energy $E_2$ of the recoiling target atom is

$$E_2 = \frac{4M_1 M_2}{(M_1 + M_2)^2} \cos^2 \phi E_0.$$  \hspace{1cm} \text{(8.21)}$$

The energy $T$ transferred in the collision equals $E_2$ and the maximum energy transfer $T_{\text{max}}$ occurs for a head-on collision with $\phi = 0$,

$$T_{\text{max}} = \frac{4M_1 M_2}{(M_1 + M_2)^2} E_0.$$  \hspace{1cm} \text{(8.22)}$$

For 100-keV As ions ($M_1 = 75$) in GaAs, the energy transfer is 79 keV, indicating that an incident ion in individual collision transfers can be large, for $\phi = 0$, are relatively rare; however, they are the energy available to displace lattice atoms.

The minimum energy $E_d$, required to displace about 15 eV; that is, if a lattice atom is displaced. If the energy transfer is for another lattice atom ($T_{\text{max}} = E_0$ for the cascade we can assume that the energy to displace a neighboring atom. However, the number of atoms $N_D$ displaced

$$N_D = \frac{E_0}{E_d}$$

where $E_D$ is the amount of energy in ions and secondary recoils. If we consider ions in nuclear collisions produce $dE/dx|_n$. For the case of medium pendent (for energies around the $dE/dx|_n$ around $E_1$ in Table 8.1), so that

$$dE/dx|_n = \frac{\rho N_D}{ho}$$

Where $dE/dx|_n$ dominates over $dE/dx$. where $E_0$ is the energy of the incident...
duced along the individual paths of the amorphous region.

However, there is no long-range crystallinity, and the covalent, tetrahedral structures and lengths. In compound semiconductors, the intermixing of the Ga and As ions, Ga- and As-rich, are formed deterministically, identified by the subscript 1, 2 (Fig. 8.15). Isolated particles can be solved by considering the kinematic energy and momentum. For an ion, the velocity and energy are \( v \) of mass \( M_2 \) is at rest. After the collision, the energies \( E_1 \) and \( E_2 \) of the ion and recoiling atom system are given by the equations:

\[
\begin{align*}
E_2 & = \frac{1}{2} \left( \frac{M_1}{M_2} \right) v_1^2 \\
M_2 v_2 & = M_2 v_2 \cos \phi \\
M_2 v_2 & = -M_2 v_2 \sin \phi.
\end{align*}
\]

The initial and final energies of the incident ion is:

\[
E_1 = E_2 \phi E_0.
\]

For 100-keV As ions \( (M_1 = 75) \) incident on Si \( (M_2 = 28) \), the maximum energy transfer is 79 keV, indicating that substantial amounts of energy can be lost by the incident ion in individual collisions. Even for more glancing collisions, energy transfers can be large, for \( \phi = 45^\circ \), \( T = 0.395E_0 \). These large energy transfers are relatively rare; however, they do serve as an indication that there is sufficient energy available to displace lattice atoms.

The minimum energy \( E_d \) required to displace a Si atom is about 15 eV; that is, if a lattice atom receives less energy than \( E_d \), it will not be displaced. If the energy transfer is greater than \( 2E_d \), the displaced atom can displace another lattice atom \( (T_{max} = E_d \) for \( M_1 = M_2 \)). As an approximation in the collision cascade we can assume that the average energy required to displace an atom is \( 2E_d \). After an atom has been displaced, a much lower energy than \( E_d \) is required to displace a neighboring atom. However, in this book we use \( E_d \).

The number of atoms \( N_D \) displaced by an incident ion is then

\[
N_D = \frac{F_D}{2E_d}
\]

(8.23)

where \( F_D \) is the amount of energy deposited in nuclear collisions by the incident ions and secondary recoils. If we assume that all the energy lost by the incident ions in nuclear collisions produces displaced atoms, then \( F_D \) is proportional to \( dE/dx_{\text{in}} \). For the case of medium mass or heavier ions, \( dE/dx_{\text{in}} \) is energy-independent (for energies around the maximum in the nuclear energy loss curve; i.e., around \( E_1 \) in Table 8.1), so that

\[
F_D = \frac{dE}{dx} |_{R_p}
\]

(8.24)

Where \( dE/dx_{\text{in}} \) dominates over \( dE/dx_{\text{in}} \) and from eqs. (8.18) and (8.23),

\[
N_D = \frac{E_0}{2E_d}
\]

(8.25)

where \( E_0 \) is the energy of the incident ion.
During ion implantation, each ion produces a region of disorder around the ion track; as the implantation proceeds, the amount of disorder builds up until all the atoms have been displaced and an amorphous layer is produced over a depth \( R_p \), as shown in Fig. 8.14. The buildup and saturation of disorder are shown in Fig. 8.16 for 40-keV phosphorus ions incident on Si. In this example, about \( 4 \times 10^{14} \) phosphorus ions/cm\(^2\) are required to form an amorphous layer.

The number \( N_D \) of displaced atoms/cm\(^2\) is given by

\[
N_D = Q_I N_D
\]

where \( Q_I \) is the number of implanted ions/cm\(^2\). For the case where \( dE/dx \) is independent of energy and hence there is a uniform concentration \( N_d \) of displaced atoms per cm\(^3\), then from eqs. (8.23) to (8.26),

\[
N_d = \frac{N_D}{R_p} = \frac{Q_I}{dE/dx_n} \frac{1}{2E_d}.
\]

In regard to Fig. 8.16, for 40-keV phosphorus in Si, \( dE/dx_n = 500 \) eV/nm = \( 5 \times 10^9 \) eV/cm (Fig. 8.13) and \( E_d = 15 \) eV, so that for an ion dose \( Q_I = 3 \times 10^{14} \) ions/cm\(^2\),

\[
N_d = \frac{3 \times 10^{14} \text{ ions}}{cm^2} \cdot \frac{5 \times 10^9 \text{ eV}}{cm} \cdot \frac{1}{2 \times 15 \text{ eV}} = 5 \times 10^{22} \text{ atoms/cm}^2.
\]

The value of \( N_d = N_{SI} = 5 \times 10^{22} \text{ atoms/cm}^2 \), as expected from Fig. 8.16.

Except for low doses or light ions, we can anticipate that an amorphous layer is formed during the implantation process. This assumes that no recovery of lattice order occurs around the ion track. This condition is not met during high-dose-rate implantation, where the flux of energy to the substrate temperature. A milliamps of 100 W in the outer surface of the implanted wafer can cause stringent precautions are taken to ensure and holder. In such elevated-temperature occurs in the region around the ion interstitials into the surrounding matrix defects such as dislocation loops.

### 8.6 Annealing of Disorder

Irrespective of ion species or energy of the device atoms is always greater than the extent of disorder occurs during the implant rest in a disordered region. This disordered region contains extended defects, or it may be the nature of implanted defects in the disordered regions. Implantation electron and hole traps and recombination.

Of course, the objective of ion implantation is to create a layer with a controlled number of dopant atoms contributing one free dopant. To accomplish this the implanted dopant must reach the concentration of electrically determined concentration of dopants. Both reactants and reduction of disorder—can be achieved samples for temperature-time combinations so that the greater than the original range distribution in the next section; here, we discuss recovery of lattice order.

Implantation of donor ions into the silicon substrate via thermal annealing will form an implantation layer at the n-p junction serves to isolate the implant. Four-point probe measurements of the electrical activity of the sample is confined to the implant layer.

Lattice disorder can influence carrier mobility and consequently, Hall effect measurements are made to characterize
implantation, where the flux of energetic ions is sufficient to cause an increase in the substrate temperature. A milliampere (mA) current of 100-keV As ions dissipates 100 W in the outer surface of a Si wafer. With this heat dissipation, the surface of the implanted wafer can be heated to several hundred degrees unless stringent precautions are taken to ensure good thermal contact between the wafer and holder. In such elevated-temperature implantations, annealing of the disorder occurs in the region around the ion track. There can be a flow of vacancies and interstitials into the surrounding material, leading to the formation of extended defects such as dislocation loops.

8.6 Annealing of Disorder: Electrical Evaluation

Irrespective of ion species or energy, the number of initially displaced semiconductor atoms is always greater than the number of implanted ions. Even if annealing of disorder occurs during the implantation process, the implanted atom comes to rest in a disordered region. This disordered region may be amorphous, it may contain extended defects, or it may be a region with point defects. In any event, the electrical nature of implanted dopant ions is overwhelmed by the behavior of the disordered regions. Implantation produces a high-resistivity layer containing electron and hole traps and recombination centers.

Of course, the objective of ion implantation, with few exceptions, is to produce a layer with a controlled number of electrically active dopants, with each implanted dopant atom contributing one free carrier to the conduction or valence band. To accomplish this the implanted dopant must occupy a substitutional lattice site and the concentration of electrically degrading defects must be much less than the concentration of dopants. Both requirements—substitutional site location of dopants and reduction of disorder—can generally be met by heating the implanted samples for temperature–time combinations below that in which large-scale diffusion of dopants occurs. This thermal annealing step can achieve nearly complete electrical activity of implanted dopants in Si for anneal temperatures of 900 to 1000°C and times less than 30 minutes. In some cases it is desirable to use higher temperature–time combinations so that the implanted dopants are diffused to depths greater than the original range distribution. Diffusion of implanted species is discussed in the next section; here, we are concerned with electrical activity and recovery of lattice order.

Implantation of donor ions into a p-type semiconductor substrate followed by thermal annealing will form an implanted n-type layer (Fig. 8.17). The depletion layer at the n-p junction serves to isolate the n-type layer from the p-type substrate. Four-point probe measurements of the sheet resistance (Chapter 2) can be used to evaluate the electrical activity of the implanted layer. The measurement current I is confined to the implanted layer.

Lattice disorder can influence carrier mobilities, causing a deviation from the relation between carrier mobility and dopant concentration given in Fig. 3.11. Consequently, Hall effect measurements in conjunction with sheet resistance measurements are made to characterize the implanted layers. To make these
measurements, square-shaped test areas with contact regions at the corners are formed by lithographic techniques (generally by using oxide masks), as shown in the lower portion of Fig. 8.17. In such a symmetrical structure (often called a van der Pauw pattern) the current is passed through one pair of contacts and the voltage is measured across another pair of contacts. The sheet resistance $R_s$ in $\Omega/\square$ (see Section 2.5) is given by (Beadle et al., 1985)

$$R_s = \frac{\pi V_{sd}}{\ln 2 I_{12}} \tag{8.28}$$

where the subscripts indicate the contacts where voltages and currents are measured. The sheet Hall coefficient $R_{sh}$ [see eq. (2.34)] is determined for this configuration by measuring the voltage change $V_{13}$ normal to the current path $I_{24}$ when a magnetic field $B$ is applied perpendicular to the sample. The measured value of $R_{sh}$ is given by

$$R_{sh} = \frac{1}{B I_{24}} \tag{8.29}$$

where measured values are in units of volts, amperes, and webers/m$^2$.

Some insight into these measurements can be gained if we consider a layer uniformly doped to a depth $R_p$ with carriers with a concentration $n_i$ and mobility $\mu_i$ so that the sheet carrier concentration in a fully annealed sample, is given by

$$n_s = \frac{2}{B I_{24}}$$

and the sheet conductance $\sigma_s = \frac{1}{R_s}$

The sheet Hall coefficient (as usual) is then

$$R_{sh} = \frac{1}{B I_{24}}$$

so that

In the implanted layer, the carrier distribution form in depth so that

$$N_s = \frac{n_i}{B I_{24}}$$

and the sheet Hall coefficient is a function of $n_i$ and mobilities.

Thermal annealing of the implanted layers is necessary to remove the excess defects and conductance as the crystal order is restored.

FIGURE 8.18 Number of carriers in a Si for implantations that form layers with different concentrations of carriers. (dashed line) where an
contact regions at the corners are (using oxide masks), as shown in
metrical structure (often called a van
one pair of contacts and the voltage
the sheet resistance $R_s$ in $\Omega/\square$ (see

$$\text{(8.28)}$$

where voltages and currents are
eq (2.34)] is determined for this
$V_{13}$ normal to the current path $I_{24}$
lar to the sample. The measured

$$\text{(8.29)}$$

peres, and webers/m$^2$.
be gained if we consider a layer
with a concentration $n_f$ and mobility

$\mu_f$ so that the sheet carrier concentration $N_s = Q_f$, the number of implanted dopants
in a fully annealed sample, is given by

$$N_s = n_f R_p$$  \hspace{1cm} (8.30)

and the sheet conductance $\sigma_s = 1/R_s$ by

$$\sigma_s = \frac{1}{R_s} = e N_s \mu_f.$$  \hspace{1cm} (8.31)

The sheet Hall coefficient (assuming Hall and conductivity mobilities are equal)
is then

$$R_{sh} = \frac{1}{N_s e}$$  \hspace{1cm} (8.32)

so that

$$\mu_f = R_{sh} \sigma_s.$$  \hspace{1cm} (8.33)

In the implanted layer, the carrier concentration $n_f(x)$ and mobility are not uniform in depth so that

$$N_s = \int n_f \, dx$$  \hspace{1cm} (8.34)

$$\sigma_s = e \int n_f(x) \mu_f(x) \, dx$$  \hspace{1cm} (8.35)

and the sheet Hall coefficient is also a weighted average of carrier concentrations
and mobilities.

Thermal annealing of the implanted layer causes an increase in the sheet conductance as the crystal order is restored. Figure 8.18 shows the fraction $F_I = \frac{N_s}{\Omega_f}$

![Graph showing Number of carriers per incident ion as a function of anneal temperature in Si for implantations that form an amorphous layer (solid line) and implantations (dashed line) where an amorphous layer is not formed.]

**FIGURE 8.18** Number of carriers per incident ion as a function of anneal temperature in Si for implantations that form an amorphous layer (solid line) and implantations (dashed line) where an amorphous layer is not formed.
$N_c/Q_r$ of carriers versus anneal temperature for the case of donor-implanted Si, where the implanted layer is amorphous (solid line) or crystalline (dashed line) with residual defects. The latter case represents either a low-dose or elevated-temperature implant. For the amorphous case there is a pronounced increase in the electrical activity in the temperature range 500 to 600°C, where the amorphous layer reorders epitaxially on the substrate. Generally, there is a further increase in the fraction $F_r$ as the anneal temperature is increased to 900 or 1000°C. At these temperatures the residual defects are removed (see dashed curve) and the crystal lattice is restored to its original state before implantation.

Figure 8.18 represents an idealized case. The annealing behavior can be quite complex, reflecting the complex nature of the defects. In general, however, high temperatures are required to achieve complete annealing of the disordered layer. These temperatures are high enough so that one must also consider diffusion of the implanted dopants. In the next sections we discuss the reordering behavior (Section 8.7) of the amorphous layer—a relatively low temperature process—and then (Section 8.8) the diffusion of implanted dopants.

**8.7 Epitaxial Growth of Implanted Amorphous Silicon**

By use of self-ion implantation, Si ions implanted into $<100>$-oriented Si, one forms an amorphous layer several 100 nm thick on a single-crystal substrate. When the sample is annealed in a furnace at a fixed temperature, one-half the melting temperature (kelvin) of Si—about 550°C, the amorphous layer reorders on the underlying single-crystal substrate. As shown in Fig. 8.19, the thickness of the regrown layer increases linearly in time, thus indicating a constant growth velocity.

The velocity is about $10^{-8}$ cm/s at one-half the melting-point temperature and increases rapidly with temperature. The epitaxial reordering process is called “solid phase epitaxy” as it occurs at temperatures well below the melt temperature; in contrast to liquid phase epitaxy where growth occurs from the melt.

Measurements of the growth velocity $v_g$ of the crystal–amorphous interface are shown in Fig. 8.20. The measured velocities extend over nearly 10 orders of magnitude and can be characterized by a single-activation energy $E_A = 2.76$ eV, (Olson and Roth, 1988) so that

$$v_g = \text{Solid epitaxial growth}$$

**FIGURE 8.19** Solid-phase epitaxial regrowth versus anneal time for an amorphous implanted layer on 100 Si.

**FIGURE 8.20** Growth rate versus implanted amorphous Si on $<100>$.
8.7 / Epitaxial Growth of Implanted Amorphous Silicon 243

\[ u_g = u_0 \exp \left( \frac{-E_A}{kT} \right) \]  

(8.36)

where the preexponential factor \( u_0 = 3.68 \times 10^8 \) cm/s. The value of the activation energy is about half that observed in diffusion of dopants in Si. Consequently, regrowth of the implanted amorphous layer can be carried out at times short compared to those required for appreciable dopant diffusion.

During the epitaxial growth process, implanted dopants move onto substitutional lattice sites as the crystal–amorphous interface sweeps by their location. Substitutional concentrations of group III and V dopants can exceed the equilibrium solubility limits of the dopants, as the dopants are effectively frozen in the lattice at the low regrowth temperatures. A second anneal treatment at high temperatures of 900 or 1000°C (carried out to remove residual defects in the regrown layer or to diffuse the dopants) will allow the dopants to move off their substitutional sites. After this high-temperature process, there is a substitutional concentration of dopants.

Implanted Amorphous Silicon

Implanted Si is converted into \(<100>-\)oriented Si, one layer thick, on a single-crystal substrate. When implanted at a temperature, one-half the melting point of the amorphous layer reorders on the substrate. In Fig. 8.19, the thickness of the layer increases, indicating a constant growth velocity. The solid part of the melting-point temperature and its onset are 0°C. The reordering process is called "solid phase epitaxy," which occurs below the melt temperature; it is not a phase change from the melt.

The crystal–amorphous interface are labeled, and the growth rate extends over nearly 10 orders of magnitude, with an activation energy \( E_A = 2.76 \) eV, which is lower than the Si over Si growth.

---

**FIGURE 8.20** Growth rate versus \(1/kT\) for solid-phase epitaxial regrowth of implanted amorphous Si on \(<100>-\) Si. (From Olson and Roth, 1988.)
ants equal to the equilibrium solubility concentration, and the excess dopants above the solubility concentration have formed nonsubstitutional precipitates or clusters.

The presence of high concentrations of implanted dopants influences the epitaxial growth rate. As shown in Fig. 8.21, concentrations of phosphorus at levels greater than 0.1 atomic percent ($5 \times 10^{19}$ cm$^{-3}$) cause an increase in the growth rate. This increase is similar to the increase in the diffusion coefficient of dopants (see Section 7.7), which is attributed to an increase in the vacancy concentration in heavily doped silicon, where the Fermi level is near the conduction or valence-band edges. As one would anticipate, the growth rate is not increased by the implantation of both boron and phosphorus at equal concentrations. In this case, the implanted layer has equal concentrations of donors and acceptors and the Fermi level is near the center of the energy gap.

Implantation of oxygen ions tends to decrease the growth rate. If a sufficiently high concentration of oxygen is implanted, the growth rate will be slowed enough so that the remaining amorphous material recrystallizes in the form of a polycrystalline layer. The electrical characteristics of this polycrystalline layer are markedly inferior to those of an epitaxial regrown layer.

Polycrystalline layers are formed by implanting other species—notably the inert gas ions (such as Ne, Ar, and Kr), which coalesce to form internal gas bubbles. One can take advantage of these polycrystalline layers by implanting Ar into the back side of the wafers. During high-temperature annealing, impurities such as Cu that are already present in the wafer diffuse into the grain boundaries of the polycrystals. The removal of unwanted impurities greatly improves the quality of the material. The regrowth rate is also strongly dependent on the orientation of the sample. Figure 8.22 shows regrowth rates for different orientations. Samples with their (100) surfaces having the slowest growth rate. The orientation dependence has been explained by a model of grain boundary migration, where the growth velocity is maximum when the surface is perpendicular to the direction of the growth axis.

8.8 Diffusion of Implanted Ions

The implanted ion distribution can be described mathematically by the Fick's second law of diffusion, which is given by:

$$
\frac{\partial C(x,t)}{\partial t} = D \frac{\partial^2 C(x,t)}{\partial x^2}
$$

where $C(x,t)$ is the concentration of the implanted ions at position $x$ and time $t$, and $D$ is the diffusion coefficient. The solution to this equation for an impulse input is given by:

$$
C(x,t) = \frac{Q}{\sqrt{4 \pi D t}} e^{-x^2 / 4Dt}
$$

The implanted ion distribution can be expressed in eq. (8.13):

$$
N_i(x) = C(x) = \frac{Q}{\sqrt{4 \pi D t}} e^{-x^2 / 4Dt}
$$

where $Q_i$ is the implanted dose (in atoms per unit area), $t$ is the time of implantation, and $D$ is the diffusion coefficient.
Diffusion of Implanted Impurities

The implanted ion distribution can be described by a Gaussian function as given in eq. (8.13):

\[ N_i(x) = C(x) = \frac{Q_i}{\sqrt{2\pi} \Delta R_p} \exp \left[ -\frac{(x - R_p)^2}{2\Delta R_p^2} \right] \]

where \( Q_i \) is the implanted dose (ions/cm\(^2\)), \( R_p \) the projected ion range, and \( \Delta R_p \) the range straggling. To redistribute the implanted ions, a drive-in diffusion is
used, just as in the case of predeposition by diffusion. The concentration profile after the drive-in diffusion can be approximated by another Gaussian function, which includes the diffusion coefficient, \( D \):

\[
C(x, t) = \frac{Q_I}{\sqrt{2\pi (2\Delta R_p^2 + 4Dt)^{1/2}}} \exp \left[ \frac{-(x - R_p)^2}{2\Delta R_p^2 + 4Dt} \right].
\] (8.37)

This Gaussian function differs from the previous one [eq. (8.13)] only slightly [e.g., the \((2\Delta R_p^2)\) term is replaced by a \((2\Delta R_p^2 + 4Dt)\) term]. In this sense the range straggling and the redistribution of the implanted ions due to the drive-in diffusion can be viewed as the spreading of the impurity atoms around a mean location \( R_p \), the projected range. It is possible to make a quick estimate of the influence of diffusion during annealing by comparing \( \Delta R_p \) with \((Dt)^{1/2}\). Here, however, the diffusion coefficient may be enhanced in the high-dopant-concentration region and \( D \) may not be constant over the entire distribution.

When the implanted projected range is close to the surface, the implanted ions may escape from the sample during the drive-in diffusion. The escape of dopant atoms can be reduced or prevented by capping the sample surface with an oxide. With such a boundary condition imposed, the flux at the sample surface is zero, that is,

\[
\frac{\partial C}{\partial x} \bigg|_{x=0} = 0.
\]

The solution for this boundary condition can be constructed by adding two Gaussian equations, one identical to eq. (8.37) and the other with \( R_p \) placed at an image distance of \(-R_p\). This operation is symmetrical with respect to the plane of the sample, reflecting all out-diffusing atoms in the same manner:

\[
N(x, t) = C(x, t) = \frac{Q_I}{\sqrt{2\pi (2\Delta R_p^2)^{1/2}}} \exp \left( \frac{-(x - 2\Delta R_p)^2}{2\Delta R_p^2} \right).
\]

This Gaussian solution [eq. (8.38)] is plotted in Fig. 8.23 for \( 550 \) nm and \( \Delta R_p = 125 \) nm for various times.

The "reflection" of the out-diffusing ions near the reflecting surface to build up the peak concentration tends to move the peak concentration in the surface until 2\( D t \) is several times the range.

**EXAMPLE**

A Si p-channel MOSFET is implanted at 30 keV for threshold voltage. The implanted ions are annealed at 950°C for 30 minutes, which gives the peak concentration \( N_m \) and activation. Given \( R_p = 106 \) nm.

**Solution:**

\[
N_m = \frac{Q_I}{\sqrt{2\pi \Delta R_p^2}} = \frac{1}{\sqrt{2\pi}} \sqrt{\frac{Q_I}{\Delta R_p^2}}.
\]

The diffusion coefficient of boron is

\[
2Dt = 2 \times 8 \times 10^{-4} \text{cm}^2 \text{s}^{-1},
\]

\[
\Delta R_p^2 = (3.9 \times 10^{-6}) \text{cm}^2.
\]

Since 2\( D t \) is only about twice the range of the implanted species does not change the peak concentration is given by

\[
N_m = \frac{Q_I}{\sqrt{2\pi (2\Delta R_p^2 + 4Dt)^{1/2}}} = \frac{1}{\sqrt{2\pi (2 \times 1.5 \times 10^{-16})^2}}.
\]

The peak concentration has, therefore, reached \( N_m = 10^{16} \text{cm}^{-3} \) after the activation anneal.
diffusion. The concentration profile led by another Gaussian function,

\[
\exp \left[ -\frac{(x - R_p)^2}{2\Delta R_p^2 + 4Dt} \right]. \tag{8.37}
\]

vious one [eq. (8.13)] only slightly \( R_p^2 + 4Dt \) term. In this sense the implanted ions due to the drive-in the impurity atoms around a mean to make a quick estimate of the comparing \( \Delta R_p \) with \( (Dt)^{1/2} \). Here, inserted in the high-dopant-concentration entire distribution.

to the surface, the implanted ions in diffusion. The escape of dopant the sample surface with an oxide. flux at the sample surface is zero,

can be constructed by adding two and the other with \( R_p \) placed at an

image distance of \(-R_p\). This operation of adding two Gaussian functions symmetrical with respect to the plane of the sample surface \((x = 0)\) is equivalent to reflecting all out-diffusing atoms from the surface. This operation yields

\[
N_i(x, t) = C(x, t) = \frac{Q_i}{\sqrt{\pi} (2\Delta R_p^2 + 4Dt)^{1/2}} \left[ \exp \left\{ -\frac{(x + R_p)^2}{2\Delta R_p^2 + 4Dt} \right\} + \exp \left\{ -\frac{(x - R_p)^2}{2\Delta R_p^2 + 4Dt} \right\} \right]. \tag{8.38}
\]

This Gaussian solution [eq. (8.38)] is plotted in Fig. 8.23 for a case where \( R_p = 550 \text{ nm} \) and \( \Delta R_p = 125 \text{ nm} \) for various \( Dt \) values.

The “reflection” of the out-diffusing impurity atoms causes the concentration near the reflecting surface to build up, and as a result of this the location of the peak concentration tends to move toward the surface. It is clear from Fig. 8.23, however, that the peak of the concentration profile does not move toward the surface until \( 2Dt \) is several times the value of \( (\Delta R_p)^2 \).

**EXAMPLE**

A Si \( p \)-channel MOSFET is implanted with boron ions to a dose of \( 1 \times 10^{12} / \text{cm}^2 \) at 30 keV for threshold voltage adjustment. Calculate the peak concentration \( N_m / \text{cm}^3 \) of the as-implanted profile. After ion implantation, the sample is annealed at 950°C for 30 minutes to activate the implanted boron ions. Calculate the peak concentration \( N_m \) and the profile of the implanted species after the activation. Given \( R_p = 106 \text{ nm} \) and \( \Delta R_p = 39 \text{ nm} \).

**Solution:**

\[
N_m = \frac{Q_i}{\sqrt{2\pi} \Delta R_p^2} = \frac{1 \times 10^{12} / \text{cm}^2}{\sqrt{2\pi} (3.9 \times 10^{-6})^2} = 1 \times 10^{17} / \text{cm}^3.
\]

The diffusion coefficient of boron in Si at 950°C is about \( 8 \times 10^{-15} \text{ cm}^2 / \text{s} \);

\[
2Dt = 2 \times 8 \times 10^{-15} \times 30 \times 60 = 2.9 \times 10^{-11} \text{ cm}^2
\]

\[
\Delta R_p^2 = (3.9 \times 10^{-6})^2 = 1.5 \times 10^{-12} \text{ cm}^2.
\]

Since \( 2Dt \) is only about twice the value of \( \Delta R_p^2 \), the peak location of the implanted species does not change as a result of the activation annealing. The peak concentration is given by

\[
N_m = \frac{Q_i}{\sqrt{\pi} (2\Delta R_p^2 + 4Dt)^{1/2}}
\]

\[
= \frac{1 \times 10^{12}}{\sqrt{\pi} (2 \times 1.5 \times 10^{-11} + 2 \times 2.9 \times 10^{-11})^{1/2}} = 6 \times 10^{16} / \text{cm}^3.
\]

The peak concentration has, therefore, decreased from \( 1 \times 10^{17} / \text{cm}^3 \) to \( 6 \times 10^{16} / \text{cm}^3 \) after the activation annealing.
The straggling has also increased from $\Delta R_p$ (39 nm) to $(\Delta R_p^2 + 2D\tau)^{1/2}$ (66.3 nm). This increase in straggling means that the impurity profile has spread in both directions about $R_p$, accompanied by a decrease in peak concentration located at $R_p$ to $\sim 0.6$ of the initial maximum implanted concentration.

There are cases where the broadening of the implanted ion profile should be kept at a minimum, such as in the control of threshold voltage of MOS devices by implantation and in the formation of very shallow junctions by implantation. In this case the implanted dopant atoms are required to be activated without any significant amount of diffusion during the activation annealing. A process called rapid thermal processing (RTP) can be used to accomplish this requirement. In this process, heat is delivered to the specimen very rapidly, such that the temperature of the specimen can rise uniformly across the wafer to a high processing temperature (800 to 1000°C) in a short time ($\approx 1\ s$). The high temperature is then held for a short duration (seconds to minutes) and then the heat is turned off and the specimen is allowed to cool down also rather rapidly. The heat source of such a process is usually an array of lamps or slotted graphite sheets. The advantage of RTP is the rapid rise and fall time at the high processing temperature, so that the activation of the dopant can occur without significant diffusion.

GENERAL REFERENCES


PROBLEMS

8.1. For 100-keV implantation of 1 mA (10⁻³ A) over an area of 1 cm², what is the number of ions implanted?

(a) What is the number of ions implanted?
(b) What is the value of the implantation dose?
(c) What is the concentration of the implantation?
(d) At what depth would the implantation end? 

8.2. For 100-keV implantation of 1 mA (10⁻³ A) over an area of 1 cm², what is the value of the implantation dose?

(a) What is the number of ions implanted?
(b) What is the value of the implantation dose?
(c) What is the concentration of the implantation?
(d) At what depth would the implantation end?

8.3. For 100-keV As and Sb ions (1 nm):

(a) Calculate the nuclear energy range.
(b) Compare the calculated nuclear energy range with the projected range of the As and Sb ions.

8.4. For 100-keV boron in silicon, $dE/dx = ke^{1/2}$.

(a) From Fig. 8.13, find a value for $k$.
(b) Derive an expression for $dE/dx$.
(c) Compare your value with the given expression.

8.5. Assume that the diffusion coefficient of an implant 10¹⁵ As ions/cm² sec at $1000^\circ$C.

(a) What is the maximum As concentration drops by one order of magnitude?
(b) If the sample is annealed at $1000^\circ$C for 10 minutes, what is the As concentration?
(c) If the silicon contains 10²ⁱ ions/cm³ = $N_A$, what is the junction depth?
(39 nm) to \((\Delta R_p^2 + 2Dt)^{1/2}\) (66.3 nm) profile has spread in increase in peak concentration implanted concentration.

implanted ion profile should be stored voltage of MOS devices by wafer junctions by implantation. In red to be activated without any annealing. A process called complicate this requirement. In this rapidly, such that the temperature infer to a high-processing temperature as the temperature is then held in the heat is turned off and the cooled. The heat source of such a graphite sheets. The advantage of processing temperature, so that the cant diffusion.


**PROBLEMS**

8.1. For 100-keV implantation of As ions (singly ionized As) into Si at a current of 1 mA \((10^{-3} \, \text{A})\) over an area of 200 cm\(^2\) for 10 min:
(a) What is the number of implanted ions/cm\(^2\)?
(b) What is the value of the projected range \(R_p\) and range straggling \(\Delta R_p\)?
(c) What is the concentration of implanted As \((\text{As}/\text{cm}^3)\) at \(x = R_p\)?
(d) At what depth would the As concentration drop to \(10^{-2}\) of its value at \(x = R_p\)?

8.2. For 100-keV implantation of As\(^{3+}\) ions (singly ionized molecules):
(a) What is the velocity of the molecules?
(b) If the magnetic analyzing system has a radius of 0.2 m, what magnetic field is required?
(c) If the molecule breaks into two As ions when it hits the Si, what is the projected range of the As ions?

8.3. For 100-keV As and Sb ions incident on Si (assume that \(a = 1.4 \times 10^{-2}\) nm):
(a) Calculate the nuclear energy loss rate for a 1/r\(^2\) potential and estimate the range.
(b) Compare the calculated range values with those given in Fig. 8.8.
(c) If \(\Delta x_p = R_p\), find \(\Delta R_p\) and compare with values in Fig. 8.8.

8.4. For 100-keV boron in silicon, assume that electronic stopping dominates with \(dE/dx = kE^{1/2}\).
(a) From Fig. 8.13, find a value for \(k\).
(b) Derive an expression for the range.
(c) Compare your value with that given in Fig. 8.8.

8.5. Assume that the diffusion coefficient is \(10^{-14}\) cm\(^2\)/s for As at 1000°C. You implant \(10^{15}\) As ions/cm\(^2\) so that \(R_p = 10^{-5}\) cm and \(\Delta R_p = 4 \times 10^{-6}\) cm.
(a) What is the maximum As concentration and the depth at which the concentration drops by one-half?
(b) If the sample is annealed for 30 min at 1000°C, what is the maximum in the As concentration?
(c) If the silicon contains \(10^{18}\) acceptors/cm\(^3\) and the junction depth is at \(N_D = N_A\), what is the junction depth after implantation and after annealing?
In EECS143, we use a gaussian function to represent the ion implantation concentration depth profile:

\[ C(x) = \frac{\phi}{\sqrt{2\pi} \Delta R_p} \exp \left[ -\frac{(x - R_p)^2}{2\Delta R_p^2} \right] \]

where \( \phi \) is the implantation dose (in \#/cm\(^2\)), \( R_p \) is the projected range and \( \Delta R_p \) is the longitudinal straggle. This gaussian approximation to the single-peak implantation profile is good for first-order calculations only. It is reasonably good for sheet resistance calculations because the integral quantity \( R_s \) (= \( \frac{1}{q\mu\phi} \)) is less sensitive to details of the distribution. However, the gaussian function has too rapid a decay with distances from \( R_p \) and can lead to smaller junction depths \( x_j \) than the real profile. The rationale to choose the gaussian approximation is the simpler algebra and knowing that the gaussian function is a natural solution of the diffusion equation, which we have to deal with when further annealing steps are encountered in the processing. A better approximation for the implantation profile is the Pearson-IV distribution which requires the first four spatial moments of the distribution but such calculations will require numerical procedures.

Projected Range \( R_p \) and Longitudinal Straggle \( \Delta R_p \) for common dopants used in IC technology, B, P and As implanted into Si are shown in the following graphs (solid lines). The ranges (in Å) are also fitted to a polynomial (dashed lines) of the form:

\[ a_0 + a_1 E + a_2 E^2 + a_3 E^3 + a_4 E^4 \]

with \( E \) in keV

![Graph of Ion Energy vs. Projected Range & Straggle](image-url)
Transverse straggle $\delta R_t$

For common energies used in IC production (<200 keV), the transverse straggle ($\delta R_t$) is always larger than the longitudinal straggle ($\delta R_p$). The $\delta R_t$ values for B, P, and As are also attached for your reference. For a line-shape mask opening of width $2a$, the 2-D implantation profile is approximated by:
\[ C(x,y) = \frac{\phi}{\sqrt{2\pi \Delta R_p}} \exp \left( \frac{-(x-R_p)^2}{2\Delta R_p^2} \right) \left( \frac{1}{2} \left[ \text{erfc} \left( \frac{y-a}{\sqrt{2\Delta R_t}} \right) - \text{erfc} \left( \frac{y+a}{\sqrt{2\Delta R_t}} \right) \right] \right) \]

Note that for an infinite opening (i.e., \( a \to \infty \)), \( C(x,y) \) reduces to the one-dimensional case \( C(x) = \frac{\phi}{\sqrt{2\pi \Delta R_p}} \exp \left( \frac{-(x-R_p)^2}{2\Delta R_p^2} \right) \), as expected.

**Ion Channeling**

To minimize ion channeling effect, the Si substrate is usually tilted by 7° with respect to the ion beam but the \( \cos (7°) \) correction to projected depths is close to unity and is usually neglected in calculations.

**Implantation into other Substrates**

Poly-Si is pure Si, so it has identical ranges as single-crystal Si. Ranges in SiO\(_2\) is about several percent smaller than that of Si. For IC processing designs, we are primarily concerned with the dopant profile in Si. When we deal with problems involving implantation through SiO\(_2\) into Si, we usually treat the SiO\(_2\) having the same energy stopping power as Si to simplify the calculations. Detailed profile data of ions in many substrate can be found in tables published by Gibbons et al or from Monte Carlo Simulators such as TRIM. [J.F. Zeigler and J.P. Biersack, “The Stopping and Range of Ions in Solids,” Pergamon Press, 1985].

The following six plots show calculated values by TRIM for \( R_p \) and \( \delta R_p \) of B, P and As into Photoresist, Si\(_3\)N\(_4\) and SiO\(_2\).
Implantation profiles through multilayer structures

The profile calculations generally require advanced techniques such as Boltzmann transport equation or Monte Carlo simulation. For implant masking calculations, most times we only care about the fraction of implant dose not passing through the mask thickness, not the detailed depth profile into the underlying substrate. The transmission factor $T$ is equal to:
\[ T = \frac{\phi_{\text{transmission}}}{\phi} = \frac{1}{2} \text{erfc} \left[ \frac{d - R_p}{\sqrt{2} \Delta R_p} \right] \]

where \( d \) is the mask thickness, \( R_p \) and \( \delta R_p \) correspond to values of the ion through the mask material, and \( \phi_{\text{transmission}} \) is the dose of ions that penetrate pass through the mask. The complementary error function \( \text{erfc}(x) \) is plotted in Figure 4.4 of Jaeger.

**Example:**

SiO\(_2\) is used as the implantation mask and we assume the SiO\(_2\) stopping power is identical to that of Si.

For 200 keV Boron, find the required oxide thickness \( d \) such that the transmission factor \( T = \frac{1}{2} \text{erfc} \left[ \frac{d - R_p}{\sqrt{2} \Delta R_p} \right] \) is: (i) \( 10^{-5} \), (ii) \( 10^{-4} \), and (iii) \( 10^{-3} \).

Transmitted fraction \( \frac{\phi_f}{\phi} = \frac{1}{2} \text{erfc}(z) \) where \( z = \frac{d - R_p}{\sqrt{2} \Delta R_p} \)

(i) \( \frac{\phi_f}{\phi} = 10^{-5} \Rightarrow z = \text{erfc}^{-1}(2 \times 10^{-5}) = 3.02 \)

(ii) \( \frac{\phi_f}{\phi} = 10^{-4} \Rightarrow z = \text{erfc}^{-1}(2 \times 10^{-4}) = 2.64 \)

(iii) \( \frac{\phi_f}{\phi} = 10^{-3} \Rightarrow z = \text{erfc}^{-1}(2 \times 10^{-3}) = 2.20 \)

\[ d \text{ values are:} \]

(i) \( = \sqrt{2}\Delta R_p \cdot z + R_p = \sqrt{2} \times 0.093 \times 3.02 + 0.53 = 0.93 \mu m \)

(ii) \( = \sqrt{2} \times 0.093 \times 2.64 + 0.53 = 0.877 \mu m \)

(iii) \( = \sqrt{2} \times 0.093 \times 2.20 + 0.53 = 0.82 \mu m \)

Si\(_3\)N\(_4\) is more effective than SiO\(_2\) in blocking implantation by about 15%. Implantation into photoresist is often encountered in IC processing. Since we use the photoresist as an implantation mask, they are usually made sufficiently thick to completely block the implanted dopants. For this reason, the ranges of dopants into photoresist are not given here but they can be looked up in tables. Roughly, a photoresist layer should be 1.8 times the thickness of a SiO\(_2\) blocking layer. The rule-of-thumb to achieve sufficient blocking is to have the masking layer thickness larger than \( R_p + 5\delta R_p \) of that ion into the masking material.

**What happen to charges carried by the ions?**

One common question asked is the charge state of the ions once they enter the solid. The answer is positive ions will be neutralized by electrons in the solid instantaneously (e.g. Si) or after annealing steps (e.g. thick SiO\(_2\)). We need the positive charge on the ions so that we can manipulate the accelerating energy. In principle, any particle with positive charge, negative charge, or in the neutral state will give the same implantation profile provided they have identical kinetic energy.
Week 5

Dopant Diffusion
Typo
Wolf, Chap 9

Eqn (9.16) $C(x,t)$ should have $\exp[-x^2/4Dt]$ dependence, not $\sqrt{Dt}$. 
Chapter 9
DIFFUSION in SILICON

Diffusion is the phenomenon by which one chemical constituent moves within another as a result of the presence of a chemical gradient (more accurately, a chemical potential gradient). The diffusion of controlled impurities or dopants in silicon is the foundation of pn junction formation and device fabrication. In the early days of transistor and IC processing, dopants were supplied to the silicon by chemical sources. These dopants were then thermally diffused to the desired depth by subjecting the wafers to an elevated temperature treatment (900–1200°C) in a furnace. However, with the onset of submicron device fabrication, ion implantation became the standard method for introducing dopants into Si, and chemical-source diffusion was essentially abandoned. However, interest in this method was revived to some degree, primarily for forming ultra-shallow junctions (<100 nm). The chemical techniques being investigated for this application include rapid vapor doping (RVD) and P-GILD. They are described later in this chapter.

Nevertheless, after the dopants are introduced into silicon wafers by ion implantation they still need to be electrically activated. Such electrical activation requires a high-temperature operation called an anneal. Dopant diffusion also occurs during these activation anneals, as well as during any other high-temperature processes that may take place after the anneal (e.g., thermal oxidation, BPSG reflow). This implies that a high-temperature treatment will affect the distribution of all dopants already in the silicon. Therefore, it must be possible to trace the doping profile of all of the previous implants up to the last high-temperature thermal process. The control of the doping profile becomes ever more difficult as the devices shrink in size, and for that reason a fundamental understanding of diffusion remains important for ULSI fabrication.

Unfortunately, diffusion in silicon is not accounted for simply by solving the mathematical equations of diffusion. In this regard, the diffusion of dopants in silicon differs from that of diffusion in metals. Since silicon is a semiconductor, and the diffusing dopants are electrically charged, one must account for the interaction of the dopants with silicon point defects. These interactions are controlled by equilibrium reactions relating the balance of charge from all sources in the silicon crystal (i.e., intrinsic electron-hole pairs, n-type and p-type dopants, and the various charged defects). Silicon defects are discussed in Chap. 2.

In the early days of silicon technology, the junction depths in devices were typically as deep as 2000–4000 nm and their fabrication was relatively uncomplicated. With current generations of technology, the required junction depth has been scaled to 50–100 nm and future requirements will dictate junctions as shallow as 20 nm. This reduction in junction depth has made the fabrication of diffused junctions a major challenge. Notable advances in ion implantation (see Chap. 10) and thermal processing technology (RTP) have gone a long way to enabling the formation of such shallow junctions. Accompanying the fabrication of ultra-shallow junctions is the challenge of developing reliable analytical techniques for determining junction depths and concentration profiles. These techniques are also discussed in this chapter. Table 9-1 shows the National Technology Roadmap for Semiconductors requirements for the range of minimum junction depth necessary for each successive technology generation from 1997 to 2012.

In this chapter the following topics are covered: a) mathematics governing the mass transport phenomena of diffusion (Fick’s equations and their solutions); b) the temperature dependence
of diffusion coefficients; c) the diffusion coefficients of the substitutional impurities; d) the simulation of one-dimensional doping profiles with SUPREM; e) atomistic models of diffusion in Si; f) diffusion in polycrystalline-Si; g) diffusion in SiO₂; h) anomalous diffusion effects; i) transient enhanced diffusion (TED); and j) measurement techniques used in diffusion studies.

9.1 THE MATHEMATICS OF DIFFUSION

In this section the basic differential equations of simple diffusion and their solutions for some special cases of diffusion in Si are described. The meaning of the diffusion coefficient $D$ is defined and a method to experimentally determine its value is discussed. (The terms diffusion coefficient, diffusivity, and diffusion constant are used interchangeably.) The atomic nature of the Si matrix and the interaction of the dopants with Si defects is reviewed in a subsequent section.

9.1.1 Fick’s First Law

The basic mathematical tools for treating diffusion were elucidated in 1855 when Fick proposed that equations analogous to Fourier’s heat-flow equations, should also apply to the diffusive flow of matter. The first law posited by Fick states that if a concentration gradient $\partial C/\partial x$ of an impurity exists in a finite volume of a matrix substance, then there exists a tendency for the impurity material to move in such a direction as to decrease the gradient. When diffusion occurs, the impurity concentration at any location becomes a function of both distance and time. One can represent the concentration profile by $C(x,t)$. If the flow persists for a sufficiently long time, the material will become homogeneous and the net flow of matter will cease. Fick went on to state that the flow of the impurity, represented by the flux $J$ is proportional to the concentration gradient, $\partial C/\partial x$ and that the constant of proportionality is defined as the diffusivity $D$ of the impurity in that particular matrix. Fick’s statements can be represented mathematically as:

$$J = -D \frac{\partial C(x,t)}{\partial x} \quad (9.1)$$

Equation 9-1 implies that as the concentration gradient decreases, the flux or diffusion decreases. The flux, $J$, is defined as the mass of material moving per unit area, per unit time. The units used in Eq. 9-1 are shown in Eq. 9-1a:

$$J \, (\text{gm/cm}^2\text{sec}) = -D \, (\text{cm}^2/\text{sec}) \frac{\partial C(\text{gm/cm}^3)}{\partial x(\text{cm})} \quad (9.1a)$$

Equation 9-1a shows the units for $D$ as cm²/sec, but $D$ can also be given in other units (e.g., $\mu$m²/hr, or m²/sec). The constant $D$ depends on the diffusion temperature, the diffusing species (also called the diffusant), and the concentration of the diffusant. Other factors such as the ambient conditions during diffusion and the presence of other dopants also affect the measured value of $D$, and they will be discussed in ensuing sections of this chapter.

9.1.2 Fick’s Second Law

Fick’s Second Law in one-dimension is now derived from the first law. In order to obtain useful results one needs to determine what happens to the concentration gradient with the progression
of time. To do this refer to Fig. 9-1, where the amount of dopant entering and leaving a finite volume of matrix is illustrated. Consider a finite volume of matrix material \( \Delta V \), having a thickness \( \Delta x \) and a unit cross-sectional area (i.e., \( \Delta V = 1 \cdot \Delta x \)). Material enters or leaves this volume only in one dimension (i.e., only in the \( z \)-direction). We define \( J_1 \) as the flux of material entering the volume \( \Delta x \), and \( J_2 \) as the flux leaving the same volume element. If the volume element is very thin (i.e., \( \Delta x \to 0 \)), then the difference in the fluxes is:

\[
J_2 - J_1 = -\Delta x \left( \frac{\partial J}{\partial x} \right)
\]  

(9.2)

Since the mass of material that enters the element in unit time \( (J_1) \) is different than the amount which leaves \( (J_2) \), the concentration within the element must also be changing with time. The change of the concentration with respect to time is given by:

\[
\frac{\partial C}{\partial t} = \frac{J_2 - J_1}{\Delta x}
\]  

(9.3)

By rearranging Eq. 9-2 to get \( (J_2 - J_1)/\Delta x \) on one side, and substituting the other side into Eq. 9-3 one gets:

\[
\frac{\partial C(x, t)}{\partial t} = -\frac{\partial J}{\partial x}
\]  

(9.4)

Now substitute \( J \) from Eq. 9-1 into Eq. 9-4, to obtain Fick's Second Law, which can be written:

\[
\frac{\partial C(x, t)}{\partial t} = \frac{\partial}{\partial x} \left( D \frac{\partial C}{\partial x} \right)
\]  

(9.5)

Equation 9-5 relates the change in concentration with respect to time, as a function of the changing concentration gradient. If the diffusion coefficient \( D \) is independent of position (concentration) then \( D \) can be brought outside the partial differential and Eq. 9-5 can be rewritten as:

\[
\frac{\partial C(x, t)}{\partial t} = D \left[ \frac{\partial^2 C(x, t)}{\partial x^2} \right]
\]  

(9.6)

The fact that \( D \) in Eq. 9-6 is assumed to be independent of position implies it is also independent of concentration (which varies with position as a result of the concentration gradient). As it turns out, for low concentrations of diffusant the assumption that \( D \) is independent of position is valid, and Eq. 9-6 can be used. But for high diffusant concentrations one must use Eq. 9-5.

**9.1.3 Solutions to Fick's Second Law**

Two solutions to Fick's Second Law when \( D \) is independent of concentration (i.e., Eq. 9-6 prevails), will be examined. The first solution relates to the case where dopants are introduced into the silicon from a vapor source. The source is considered to bring a constant supply of dopant atoms to the silicon surface where they diffuse into the silicon at the process temperature.

![Fig. 9-1 Schematic showing an element of volume with the flux \( J_1 \) entering and \( J_2 \) leaving.](image)
It is assumed that the source maintains a constant value of surface concentration during this entire diffusion process. In diffusion technology, this type of dopant incorporation is termed **predeposition**, and is used to introduce a known quantity of impurities into the silicon. As noted earlier, this step in the past was carried out by diffusion from a chemical source, but now has been largely replaced by ion implantation. The pre-deposited dopant (whether chemically or ion implanted) is then redistributed in the silicon by a further thermal treatment called a **drive-in or redistribution diffusion**. Despite the reliance on ion implantation for the pre-deposition step, it is still illustrative to solve Fick’s differential equation for the pre-deposition boundary conditions. Moreover, chemical source pre-deposition is still being used for some deep junction applications and is now under consideration again for some ultra-shallow junction applications.

**9.1.3.1 Chemical Pre-Deposition:** In order to solve the second-order differential equation (Eq. 9.6) it is necessary to establish initial and boundary conditions. For the chemical predeposition step the initial condition is that the concentration of dopant in the silicon at \( t = 0 \) for all values of \( x \) is equal to 0. This condition is written as:

\[
C(x, 0) = 0 \quad (9.7)
\]

The two boundary conditions for this process step are established next. The first boundary condition (Eq. 9.8) states that the concentration of dopant at \( x = 0 \) (i.e., the concentration at the surface of the Si), for any time during the diffusion, is fixed at a value \( C_s \) for the entire diffusion cycle. The units of \( C_s \) are atoms/cm\(^2\) and the condition is described mathematically as:

\[
C(0, t) = C_s \quad (9.8)
\]

The second boundary condition states that the concentration of dopant at \( x = \infty \) is equal to 0 for any time, or:

\[
C(\infty, t) = 0 \quad (9.9)
\]

The solution to the partial differential equation (Eq. 9.6) with the application of the initial and boundary conditions is given by Eq. 9.10:

\[
C(x, t) = C_s \text{erfc} \left( \frac{x}{2\sqrt{Dt}} \right) \quad (9.10)
\]

where \( D \) is the diffusion coefficient, \( x \) is the distance coordinate, \( t \) is the time of the diffusion, and \( \text{erfc} \) is the complementary error function. The complementary error function is tabulated\(^3\) for values of the dummy variable \( z = x/2(\sqrt{Dt}) \). The denominator of Eq. 9.10, \( 2(\sqrt{Dt}) \), is termed the **diffusion length**, which is representative of how far atoms move during a given thermal step. The term is also referred to as "root \(Dt\". The units of "root \(Dt\" are length (cm) and its value is used when comparing the thermal exposure or thermal budget of a process (see Chap. 8).

Figure 9-2 shows a plot of the concentration profile for a predeposition for several different times at a constant temperature. \( D \) is assumed to be constant and the profile only depends on the time of the diffusion. As the diffusion time is increased the surface concentration remains constant at \( C_s \), while the dopants move deeper into the silicon. The total amount of dopant introduced into the silicon also increases with time. The **junction depth**, \( x_p \), can be defined as the point at which the concentration of the diffusant equals the background doping in the Si. The total amount of dopant that accumulates in the silicon \( (C_0) \) during the predeposition can now be calculated by integrating Eq. 9.10 with respect to \( x \) over all of space (in one-dimension) and obtain Eq. 9.11.

\[
Q_0 = \int_{-\infty}^{+\infty} C(x, t) \, dx = \int_{-\infty}^{+\infty} C_s \text{erfc} \left( \frac{x}{2\sqrt{Dt}} \right) \, dx = \frac{2}{\sqrt{\pi}} C_s \sqrt{Dt} \quad (9.11)
\]
Fig. 9-2 Schematic representation of the diffusion profile for constant source diffusion (pre-deposition). Note the surface concentration remains constant with increasing diffusion time.\textsuperscript{72} From S.K. Ghandi, \textit{VLSI Fabrication Principles}. Reprinted with permission of John Wiley & Sons.

The junction depth, $x_j$, can be calculated by determining the depth (value of $x$) at which the diffusing dopant concentration just equals the background concentration in the silicon, $C_{\text{sub}}$. This is accomplished by solving Eq. 9-10 for $x$ when $C = C_{\text{sub}}$:

$$x_j = (2\sqrt{Dt}) \text{erfc}^{-1} \left( \frac{C_{\text{sub}}}{C_s} \right) \quad (9.12)$$

where $\text{erfc}^{-1}$ is the inverse of the complementary error function. Both Eqs. 9-11 and 9-12 contain the "root Dt" term.

\subsection{9.1.3.2 Drive-in Diffusion}

Next consider what happens to the dopant atoms that have been introduced during the predeposition if the silicon is subjected to additional thermal treatment. We would expect that the total dose, $Q_0$, in the silicon should remain constant and the diffusant would move deeper into the silicon. This process is termed the \textit{redistribution} or \textit{drive-in} diffusion. The drive-in is used to move the impurities to the desired junction depth following the predeposition. In order to obtain a closed-form solution of Fick's Second Law, it is necessary to assume that the initial dopant is confined to a thin layer at the surface. As a matter of fact, the solution assumes that the dopant $Q_0$ exists as a $\delta$-function in a thin rectangular region at the surface. The total quantity of impurity present is fixed at $Q_0$ (atoms/cm$^3$). For the drive-in case, the initial condition is given by Eq. 9-13:

$$C(x,0) = 0 \quad \text{for } x > \delta \quad (9.13)$$

and the two boundary conditions are given by Eqs. 9-14 and 9-15:

$$\int_0^{+\infty} C(x,t)dx = Q_0 \quad (9.14)$$

and

$$C(\infty,t) = 0 \quad (9.15)$$

The boundary condition in Eq. 9-14, states that the total amount of dopant is fixed at $Q_0$. The solution to Fick's Second Law, under these conditions is:

$$C(x,t) = \frac{Q_0}{\sqrt{\pi Dt}} \exp \left( \frac{-x^2}{4\sqrt{Dt}} \right) \quad (9.16)$$
This solution has the form of a Gaussian distribution. The surface concentration, \( C_s \), which decreases as a function of time in this case, is determined from Eq. 9-17, when \( x = 0 \):

\[
C_s = C(0, t) = \frac{Q_0}{\sqrt{\pi}Dt}
\]

(9.17)

From Eq. 9-17 one can see that as diffusion time increases the surface concentration decreases.

The junction depth for the drive-in case is calculated in a manner similar to that used to obtain Eq. 9-12 and is given by:

\[
x_j = 2\sqrt{D_t}\left(\ln\frac{Q_0}{C_{sub}}\sqrt{\piDt}\right)^{1/2}
\]

(9.18)

Figure 9-3 shows a plot of the concentration profile \( C(x) \) for the Gaussian case for several drive-in diffusion times at a constant temperature. The key points to note are that the value of \( C_s \) decreases and the junction depth moves deeper into the silicon as the diffusion time increases.

An example of how to use these two diffusion models is now provided. In the first example the diffusion profile, the junction depth, and total dopant concentration after a short predeposition are each calculated. In the second example the information from the previous example is used to establish the diffusion profile and junction depth after a brief drive-in.

**EXAMPLE 9-1:** Calculate: a) diffusion profile; b) junction depth; & c) total amount of dopant introduced after boron predeposition performed at 950°C for 30 min, in a neutral ambient. Assume the substrate is n-type with a background doping level of \( 1.5 \times 10^{16} \) atoms/cm\(^3\) and the B surface concentration reaches solid solubility \( (C_s = 1.8 \times 10^{20} \) atoms/cm\(^3\)).

**SOLUTION:** a) Under the conditions of the predeposition, the diffusion profile is controlled by the erfc of Eq. 9-10. The profile can be found by using the erfc curve of Fig. 9-4. In order to use this curve, values of \( x/(2\sqrt{D_t}) \) must be calculated for values of \( x \). Corresponding values of \( C(x)/C(0) \) [where \( C(0) \) is the B surface concentration \( C_s \), are found from Fig. 9-4, and finally values of \( C(x) \) are calculated. First, the value of the B diffusion coefficient at 950°C is found from:

\[
D_B(950°C) = 0.76 \exp\left(-\frac{3.46 \text{ eV}}{k T} 1223 \text{ K}\right) [\text{cm}^2/\text{sec}] = 3.0 \times 10^{-15} \text{ cm}^2/\text{sec}.
\]

Next calculate the diffusion length, \( z = 2\sqrt{D_t} \) for the process:

\[
z = 2 \left(3 \times 10^{15} \text{ cm}^2/\text{sec} \cdot 1.8 \times 10^{3} \text{ sec}\right)^{1/2} = 4.65 \times 10^{-6} \text{ cm}.
\]

Now set up the following table to assist in the calculation of the diffusion profile:

<table>
<thead>
<tr>
<th>( x ) (cm ( \times 10^{-4} ))</th>
<th>( x/(2\sqrt{D_t}) )</th>
<th>( C(x)/C(0) )</th>
<th>( C(x) ) (atoms/cm(^3))</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0</td>
<td>1</td>
<td>( 1.8 \times 10^{20} )</td>
</tr>
<tr>
<td>0.05</td>
<td>1.07</td>
<td>0.13</td>
<td>( 2.3 \times 10^{19} )</td>
</tr>
<tr>
<td>0.075</td>
<td>1.67</td>
<td>0.018</td>
<td>( 3.2 \times 10^{18} )</td>
</tr>
<tr>
<td>0.10</td>
<td>2.15</td>
<td>0.0022</td>
<td>( 3.9 \times 10^{17} )</td>
</tr>
<tr>
<td>0.15</td>
<td>3.23</td>
<td>0.000005</td>
<td>( 9.0 \times 10^{14} )</td>
</tr>
</tbody>
</table>

b) The junction depth is defined as the value of \( x \) where the concentration of the diffusion profile equals the substrate doping \( (1.5 \times 10^{16} \) atoms/cm\(^3\)). When \( C(x) = 1.5 \times 10^{16} \) atoms/cm\(^3\), then \( C(x)/C(0) = 8.3 \times 10^{-5} \). Next calculate \( x_j/(2\sqrt{D_t}) \) from Fig. 9-4, to get:

\[
x_j/(2\sqrt{D_t}) = 2.8; \text{ or } x_j = 2.8 \cdot 4.65 \times 10^{-6} \text{ cm} = 1.3 \times 10^{-5} \text{ cm} = 0.13 \mu\text{m}
\]

c) The total amount of dopant is calculated from Eq. 9-11:

\[
Q_o = C_s(2\sqrt{D_t})\sqrt{\pi} = 1.8 \times 10^{20} \cdot 4.62 \times 10^{-6}/1.77 = 4.7 \times 10^{14} \text{ atoms/cm}^2.
\]
**EXAMPLE 9-2:** Calculate the diffusion profile and junction depth after the predeposition of Example 9-1 is subjected to a neutral ambient drive-in at 1050°C for 60 min.

**SOLUTION:** For the drive-in conditions of this example the Gaussian distribution of Eq. 9-16 (the curve labeled *Gaussian* in Fig. 9-4) controls the profile. First calculate the diffusion length \( \sqrt{2D_t} \) for the new diffusion conditions:

\[
D_B (1050°C) = 0.76 \ e^{(-3.46 \ eV/k \ 1323 \ K)} = 3.3 \times 10^{-14} \ \text{cm}^2/\text{sec}.
\]

Next calculate the value of the surface concentration \( C(0) \) from Eq. 9-17:

\[
C(0) = Q_0/\sqrt{\pi D t} = 4.7 \times 10^{14} \ \text{atoms/cm}^2/(\pi \times 3.3 \times 10^{-14} \ \text{cm}^2/\text{sec} \cdot 3600 \ \text{sec})^{1/2}
\]

\[
= 2.44 \times 10^{19} \ \text{atoms/cm}^3
\]

Set up the following table to assist in the calculation:

<table>
<thead>
<tr>
<th>( x ) (cm x 10^{-4})</th>
<th>( x/(2\sqrt{D t}) )</th>
<th>( C(x)/C(0) )</th>
<th>( C(x) ) (atoms/cm^3)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0</td>
<td>1</td>
<td>2.44 x 10^{19}</td>
</tr>
<tr>
<td>0.1</td>
<td>0.45</td>
<td>0.75</td>
<td>1.83 x 10^{19}</td>
</tr>
<tr>
<td>0.2</td>
<td>0.91</td>
<td>0.60</td>
<td>1.46 x 10^{19}</td>
</tr>
<tr>
<td>0.3</td>
<td>1.36</td>
<td>0.17</td>
<td>4.10 x 10^{18}</td>
</tr>
<tr>
<td>0.4</td>
<td>1.82</td>
<td>0.038</td>
<td>9.20 x 10^{17}</td>
</tr>
<tr>
<td>0.5</td>
<td>2.27</td>
<td>0.006</td>
<td>1.46 x 10^{17}</td>
</tr>
<tr>
<td>0.6</td>
<td>2.73</td>
<td>0.0004</td>
<td>9.80 x 10^{15}</td>
</tr>
<tr>
<td>0.7</td>
<td>3.18</td>
<td>0.00004</td>
<td>9.80 x 10^{14}</td>
</tr>
</tbody>
</table>

Next calculate \( x_j \) in a manner similar to the one used in Example 9-1.

\[
C(x_j)/C(0) = 1.5 \times 10^{16}/2.44 \times 10^{19} = 6 \times 10^{-4}.
\]

Next find the corresponding value of \( x/(2\sqrt{D t}) \) from Fig. 9-4, which is \( x_j/(2\sqrt{D t}) = 2.7 \) or:

\[
x_j = 2.7 \cdot 2.2 \times 10^{-5} \ \text{cm} = 5.9 \times 10^{-5} \ \text{cm} = 0.59 \ \mu\text{m}.
\]
Fig. 9-4 Normalized concentrations \([C/C(0)]\) as a function of normalized distance \((dC/dt)\) for the erfc and Gaussian curves.

9.1.3.3 Drive-In From An Ion Implantation Predeposition: If the predeposition is from an ion implanted source (which produces a doping concentration with a near-Gaussian distribution close to the surface), the Gaussian profile will move when subjected to a high temperature anneal (see Chap. 10 on Ion Implantation). The solution to Eq. 9-6 (concentration-independent diffusion) with an initially implanted Gaussian profile has been reported for the cases of drive-ins performed in a neutral ambient, and an oxidizing ambient. For drive-ins in an oxidizing ambient, the solution to Eq. 9-6 is difficult to obtain in closed form, since it involves a moving boundary problem. Consequently, numerical methods must be employed to obtain the solutions (see Chap. 2, Vol. 3 of this series).

9.1.4 Concentration Dependence of the Diffusion Coefficient

The diffusion profiles that were calculated in the previous section are valid for the case when the diffusion constant is not a function of the doping concentration. That is why solutions of Eq. 9-6 were used, rather than those for Eq. 9-5. Constant-value D prevails when the doping concentration is lower than the intrinsic-carrier concentration \(n_i\) at the diffusion temperature, and conversely, concentration dependent diffusion occurs when the doping level is greater than the intrinsic carrier concentration. The first case is termed intrinsic diffusion while the latter is called extrinsic diffusion. To determine where extrinsic diffusion starts one needs to know the intrinsic carrier concentration at the diffusion temperature. Equation 9-19 is the expression used to calculate the intrinsic carrier concentration in silicon:

\[
n_i = p_i = \sqrt{N_v N_c} \exp\left(-\frac{E_g(T)}{2kT}\right) \text{ (cm}^2\text{)}
\]

(9.19)

Here, \(N_v\) and \(N_c\) are the density of states at the valence and the conduction band edges, respect-
ively, \( E_g \) is the energy gap of Si, \( k \) is Boltzmann’s constant, and \( T \) is the temperature in kelvins (K). The value of \( E_g \) depends somewhat upon temperature and the doping concentration and these factors must be taken into account when determining the intrinsic carrier concentration. The bandgap dependence on temperature in the range of 800–1100°C, and is given by:

\[
E_g(T) = E_g(0) - bT
\]  

(9.20)

where \( E_g(0) = 1.46 \text{ eV} \) and \( b = 2.97 \times 10^{-4} \text{ eV/K} \).

There is a “bandgap narrowing” effect, \( \delta E_g \), that occurs for heavily doped silicon. A “lumped” model has been developed to provide a value for the narrowing effect.\(^5\) This model is not based upon first principles, since no simple physical model has been found to explain the effect. One proposal is that the lattice strain caused by the presence of heavy doping induces the band gap to change. When bandgap narrowing, \( \delta E_g/2kT \) is included, the effective intrinsic carrier concentration \( n_{\text{eff}} \) is given by:

\[
N_{\text{eff}} = N_i \left( \frac{\delta E_g}{2kT} \right)
\]  

(9.21)

Under extrinsic conditions, \( N \) depends on the doping concentration, and since the concentration changes with distance (i.e., a concentration gradient exists), \( N \) also changes with distance. To obtain a solution to Fick’s Law one can expand Eq. 9-5 to read:

\[
\frac{\partial C(x,t)}{\partial t} = \frac{\partial}{\partial x} \left( D \frac{\partial C}{\partial x} \right) = \frac{\partial D}{\partial x} \frac{\partial C}{\partial x} + D \frac{\partial^2 C}{\partial x^2}
\]  

(9.22)

The \( \partial D/\partial x \) term makes Eq. 9-22 an inhomogeneous differential equation making it difficult to solve. Fortunately, a graphical procedure termed the Boltzmann-Matano analysis has been developed, which renders a solution for the diffusion coefficient, \( D \), as a function of concentration.\(^7\) It does not, however, explicitly provide the diffusion profile \( C(x,t) \), as was obtained for the intrinsic diffusion case. The Boltzmann-Matano method allows \( D(C) \) to be calculated from the experimentally determined concentration profile (C vs. x plot). There are several methods that are commonly used to obtain the concentration profile and they are discussed later in this chapter.

### 9.2 Defects and Dopant Diffusion

It has been realized for some time that the diffusion of dopant atoms in Si is strongly coupled to the presence of point defects. This early realization came about when it was found that the diffusion coefficient of the dopant atoms did not follow the behavior predicted by Fick’s Laws, but rather exhibited certain anomalous effects. These effects will be described in detail in later sections but are briefly mentioned here. The first anomaly was that the diffusivity of boron was greatly enhanced when phosphorous was subsequently diffused into the Si (the so-called emitter push effect). The second was that some dopants exhibited a higher diffusivity when the drive-in diffusion was performed in an oxidizing ambient (oxidation-enhanced diffusion or OED). Since these effects were initially observed, other anomalous effects have been found, and as a group their behavior can be explained by the presence of point defects in the silicon. Consequently, the types and interaction of point defects and dopants in silicon are now discussed.

#### 9.2.1 Point Defects in Silicon

Chapter 2 described some of the point defects that can exist in silicon, particularly at the temperatures normally associated with diffusion. Some of these defects are now briefly reviewed and how they relate to the diffusion of dopant atoms in silicon is discussed. The nomenclature used in the excellent treatise on diffusion by Fahey, Griffin, and Plummer is employed.\(^8\)
The first type of point defect to consider is the defect which arises from the introduction of impurity atoms into silicon. An impurity atom, such as a dopant atom that sits on a lattice site, is termed a substitutional defect. Even though the atom resides on a lattice site it is either larger or smaller than the silicon atom, and thereby causes a local perturbation in the periodicity of the lattice. Any perturbation to the lattice periodicity is termed a defect. It turns out that most of the common dopant atoms (P, As, Sb, B, In, Ga) dissolve into silicon as substitutional atoms. That is, they reside on silicon lattice sites in place of silicon atoms. These dopant atoms are also remarkable for their high solubility in silicon when compared to other elements that also dissolve in silicon. The solubilities of these dopants are discussed in Chap. 1.

A dopant atom on a substitutional site is given the designation A and can be either a donor or an acceptor. When a dopant atom occupies a substitutional site it contributes either an electron to the conduction band, or a hole to the valence band. In either case, the dopant atom then becomes an ion. It is the ion (a charged atom) that must diffuse in the silicon. It is not surprising that since the dopant atom (or ion) is charged it would interact with other charged species (e.g., crystalline point defects and free carriers).

As discussed in Chap. 2, there are other point defects that may be present at the diffusion temperature, namely: a) vacancies; b) interstitials; and c) interstitialcies. Such lattice defects are designated as X, where X can be either a vacancy or interstitial-type defect. The vacancy defect is defined as a missing Si atom or an empty lattice site and is designated as V. The presence of a vacancy requires the covalent bonds around that site to reconfigure to accommodate the defect. The V defect itself can become charged by capturing one or two electrons (or a hole), or can remain neutral. As discussed in Chap. 2, four vacancy defects have been identified:

1. The neutral vacancy, V^0
2. The singly-charged negative vacancy (in which one electron is captured), V^-1
3. The doubly-charged negative vacancy (in which two electrons are captured), V^-2
4. The singly-charged positive vacancy (in which one hole is captured), V^+

Figures 9-5a1, 9-5a2, and 9-5a3 show how the unsatisfied bonds reconfigure to accommodate the various defects in the silicon lattice.

The next type of defect discussed is the self-interstitial, which can be described as a Si atom on an interstitial site. Self-interstitials are given the designation I. The term self-interstitial is used to distinguish these defects from those involving extrinsic or foreign atoms (which can also exist in the lattice as interstitials). Figures 9-5b1 and 9-5b2 show the bonding around a tetrahedral and a hexagonal interstitial site, respectively. The last type of defect (not often recognized) is called the interstitialcy. The interstitialcy is also designated by I. The interstitialcy defect differs from the interstitial defect in that it has two "associated" atoms in non-substitutional sites, around a substitutional site. The two atoms constituting the interstitialcy can consist of: a) both Si atoms (an I-defect); or b) a Si atom and a dopant atom (an AI-defect). The interstitialcy defects can remain neutral or become charged, and have the following configurations:

1. A neutral interstitialcy with two silicon atoms, I^0.
2. A singly-charged positive interstitialcy comprised of two silicon atoms, I^+1.
3. An associated dopant and silicon interstitial that is neutral, (AI)^0. This type of defect behaves as an acceptor site.
4. An associated dopant and silicon interstitial that is singly-negatively charged, (AI)^-1. This type of defect also behaves as an acceptor site.
5. An associated dopant and silicon interstitials that is singly-positively charged, (AI)^+1. This type of defect behaves as a donor site.
**Fig. 9-5** Schematic of how the bonds may reconfigure to accommodate various defects in the silicon lattice. a) the vacancy defects, b) interstitial defects, and c) interstitialcy defects. If both dark spheres in (c) are silicon the defect is the (I) defect but if one is silicon and one the dopant atom the defect is (AI). Reprinted with permission of the American Physical Society.

Figures 9-5c1 and 9-5c2 show the atomic arrangement of interstitialcy defects. The distinction between the interstitial and the interstitialcy is often not emphasized since it is difficult to discern between the two defect types by experimentation. As a result, both of these defects are referred to as *interstitial-type* defects.

Now that the three defects (i.e., the vacancy V, the interstitial-type I, and the dopant impurity A) have been described, it is appropriate to investigate the interaction among these defects, as this will help in understanding how such interactions may affect diffusion.

The following associated dopant-atom/defect configurations may exist in the lattice: a) a dopant atom is on a lattice site (A); b) a doping atom is adjacent to a vacancy (AV); and c) a dopant atom is one of the atoms of the interstitialcy (AI). If the dopant atom actually occupies the interstitial site, it is designated A_i. It is the dopant-atom/defect pair that migrates during diffusion in one or more of the following configurations: AV, AI, or A_i. The coordinated dopant-atom/defects can be formed by the following reversible equilibrium reactions. The “Law of Mass Action” applies to these reactions.

\[ I + A \leftrightarrow AI \]  \hspace{1cm} (9.23a)

or:

\[ I + A \leftrightarrow A_i \]  \hspace{1cm} (9.23b)

\[ V + A \leftrightarrow AV \]  \hspace{1cm} (9.24)

\[ I + AV \leftrightarrow A_i \]  \hspace{1cm} (9.25)

\[ V + AI \leftrightarrow A \]  \hspace{1cm} (9.26)

or \[ V + A_i \leftrightarrow A \]
There is a further equilibrium reaction that should also be mentioned, and that is the creation and annihilation of vacancies and interstitials via the Frenkel defect:

\[ 0 \leftrightarrow V + I \]  \hspace{1cm} (9.27)

The \( \leftrightarrow \) symbolizes that the reaction is reversible, and since the concentrations of I and V can be affected by external forces (i.e., point defects arising from other sources), these reactions may be favored in one or the other direction. For example, if a large concentration of interstitials is created by oxidation, then the forward direction of Eq. 9-23 is favored, and the concentration of AI increases. The increase in AI concentration results in increased dopant diffusion.

A brief description of Eqs. 9-23 to 9-27 is given next. The reaction in Eq. 9-23 represents an exchange between a substituational atom A and the interstitial(cy) I. It is termed the kick-out reaction, since the Si-I kicks out a dopant atom from a substituational site and forms an AI pair. The reaction in Eq. 9-24 describes the dopant-vacancy reaction and governs vacancy-controlled diffusion. The reactions given in Eqs. 9-25 to 9-27, describe recombination reactions between I and V defects. The reaction in Eq. 9-26 describes a dissociative reaction termed the Frank-Turnbull mechanism. This mechanism is not considered to be important for Si diffusion since it is believed to have a low probability of occurrence. Finally, the forward reaction of Eq. 9-27, describes the formation of a Frenkel defect (vacancy and interstitial) as is described in Chap. 2.

In the past it was considered that dopant diffusion in silicon was controlled solely by the vacancy mechanism, but now it is well established through numerous investigations that both I- and V-type mechanisms contribute to dopant diffusion. As it turns out, most of the common dopants have a dominant I-diffusion-controlled mechanism. The evidence for this conclusion is discussed in detail in a later section.

The reactions dealt with above do not give the origin of the defects. There are two possibilities of defect formation that were described in Chap. 2, namely: a) an atom leaving a lattice site and creating a V and I, and b) a silicon atom diffusing from the surface of the silicon into the bulk to create an interstitial or a silicon atom that was on a lattice site moving to the surface (and vanishing by becoming part of the disorder at the surface) and creating a vacancy in the bulk. Another important source of defect production is lattice damage resulting from ion implantation. As a result, the diffusion of dopants from an ion implanted source can substantially differ from that of a chemical source.

### 9.2.2 Temperature Dependence of the Diffusion Coefficient Under Intrinsic Conditions

The temperature dependence of the diffusion coefficient, D, for intrinsic diffusion for the common dopant atoms will now be examined. However, before we proceed, it is instructive to first inspect the temperature dependence of the self-diffusion of silicon. Most self-diffusion studies use radio-active isotopes of silicon to determine the diffusion profile. Even to this day, however, the basic mechanism of Si self-diffusion is not well understood.

Some results from a study that examined the effect of hydrostatic pressure on Si self-diffusion show the dominant mechanism is interstitial-type-dominated diffusion.\(^9\) The most interesting aspect about silicon self-diffusion is that the activation energy for the process is about 1 eV greater than that of dopant diffusion. For some reason the movement of silicon-defect pairs takes more energy than that of dopant-defect pairs. The diffusion coefficient for Si self-diffusion \(D_{self}\) is given by:\(^10\)

\[ D_{self} = 1400 \exp (-5.01 \text{ eV}/kT) \text{ cm}^2/\text{sec} \]  \hspace{1cm} (9.28)
### Table 9-2 Compiled Values of Apparent Activation Energies for Diffusion of Common Dopants

<table>
<thead>
<tr>
<th>Donors</th>
<th>$Q_A$(eV)</th>
<th>Acceptors</th>
<th>$Q_A$(eV)</th>
</tr>
</thead>
<tbody>
<tr>
<td>P</td>
<td>3.51–3.67</td>
<td>B</td>
<td>3.25–3.87</td>
</tr>
<tr>
<td>As</td>
<td>4.05–4.34</td>
<td>Al</td>
<td>3.36</td>
</tr>
<tr>
<td>Sb</td>
<td>3.8–4.05</td>
<td>Ga</td>
<td>3.75</td>
</tr>
<tr>
<td></td>
<td></td>
<td>In</td>
<td>3.60</td>
</tr>
</tbody>
</table>

#### 9.2.3 Intrinsic Diffusion Coefficients

Numerous experimental measurements have been made on the effect of temperature on the diffusion coefficients ($D_A$) of the common dopants during intrinsic diffusion. In general the resulting data can be represented by an Arrhenius equation.

$$D_A = D_{Ao} \exp \left(-\frac{Q_A}{kT}\right)$$  \hspace{1cm} (9.29)

Here $D_A$ is the diffusion coefficient of the dopant atom, $D_{Ao}$ is the pre-exponential term or *frequency factor*, which is related to the frequency of the lattice vibrations (i.e., the frequency at which atoms strike the potential barrier that they must overcome to move in the lattice), and $Q_A$ is the *apparent activation energy* of the diffusion process (which is related to the height of the energy barrier that the impurity must overcome in order to move within the lattice).

The value of $Q_A$ for the common dopants has been compiled by Wohlbier.\(^{11}\) This data is reproduced in Table 9-2, which was extracted from the summary of Wohlbier's work, given by Fahey, Griffin, and Plummer.\(^{8}\) Readers interested in additional details are referred to that article, which lists the original references, or to the Wohlbier monograph. Although the values of apparent activation energy vary somewhat, depending on the particular study and dopant, all of the values range between 3.3 to 4.3 eV.

#### 9.2.4 Fast Diffusers in Silicon

There is another class of impurities that diffuse in silicon much more rapidly than the electrically active dopant atoms. These impurities are called the *fast diffusers* and their diffusion coefficients at 900°C are listed in Table 9-3. The fast diffusers, as evidenced by their name, diffuse much faster in silicon than do the dopants atoms. The fast diffusers (e.g., Fe, Cu, Pt) move great distances in silicon during high temperature processing (thereby increasing their probability of being trapped at an extended defect and causing increased leakage current in a device). The effect of the fast diffusers on device properties is discussed in Chap. 2. The apparent activation energies (0.2–2 eV) for the fast diffusers are much lower than for the substitutional impurities. At one time it was believed that the low values of activation energy implied interstitial diffusion, while the high activation energy values of the dopant atoms were due to vacancy diffusion. This analysis is not correct, and it is now well established that both vacancies and interstitials play a role in diffusion in silicon.

#### 9.3 Atomistic Models of Diffusion

Some of the possible ways a dopant atom may move in the crystalline lattice are next explored from an atomistic perspective. The most favorable configuration for a dopant atom to diffuse in silicon is when it is associated with a nearest-neighbor defect. Only a fraction of the dopant atoms are associated with defects in this manner, but it is this fraction that contributes to
Table 9-3 DIFFUSION COEFFICIENTS of FAST DIFFUSANTS in SILICON @ 900°C

<table>
<thead>
<tr>
<th>Element</th>
<th>D (cm²/sec)</th>
<th>Element</th>
<th>D (cm²/sec)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Na</td>
<td>1.6 x 10⁻³</td>
<td>Ag</td>
<td>2.0 x 10⁻³</td>
</tr>
<tr>
<td>K</td>
<td>1.1 x 10⁻³</td>
<td>Pt</td>
<td>1.6 x 10⁻²</td>
</tr>
<tr>
<td>Cu</td>
<td>4.7 x 10⁻²</td>
<td>Fe</td>
<td>6.2 x 10⁻³</td>
</tr>
<tr>
<td>Au</td>
<td>1.1 x 10⁻³</td>
<td>Ni</td>
<td>1.0 x 10⁻¹</td>
</tr>
<tr>
<td>Aui</td>
<td>2.4 x 10⁻⁴</td>
<td>C₂</td>
<td>7.0 x 10⁻²</td>
</tr>
<tr>
<td>Aui₅</td>
<td>2.8 x 10⁻³</td>
<td>H₂</td>
<td>9.4 x 10⁻³</td>
</tr>
</tbody>
</table>

contributes to the diffusion. This fraction has been calculated as:¹²

\[ C_{AX} = \theta_{AX} \frac{C_A C_X}{C_S} \exp\left(\frac{E_{AX}^b}{kT}\right) \]  \hspace{1cm} (9.30)

where \( C_A \) is the dopant concentration; \( C_X \) is the unassociated defect concentration; \( C_S \) is the concentration of lattice sites (5 x 10²²/cm³), \( E_{AX}^b \) is the binding energy of the AX defect. The \( \theta_{AX} \) factor accounts for the number of equivalent ways of forming the defect AX at a particular site (e.g., \( \theta_{AV} = 4 \)). There have been numerous investigations to determine the value of \( E_{AX}^b \) and the reader is referred to the Fahey treatise for further detail.⁸

There are several potential ways dopant atoms may diffuse in the lattice, namely: a) direct dopant exchange (no point-defect interaction); b) vacancy-dopant interaction; c) interstitial-dopant interaction; and d) interstitial-interstitial mechanism. These mechanisms are very similar and it is difficult to discern which is actually operative during diffusion.

First consider the possibility that mechanism (a) is responsible for controlling diffusion in Si. That is, what is the likelihood that diffusion proceeds by substitutional dopant atoms exchanging positions with adjacent silicon atoms (the direct exchange mechanism, shown schematically in Fig. 9-6a)? For diffusion to continue after the first interchange, the dopant atom must exchange positions again with the next Si atom. If the dopant atom instead exchanges positions with the original Si atom it will not make progress diffusing within the lattice. The continuation of the direct interchange process is unlikely because the activation energy associated with the breaking of Si bonds is high, and at least six Si bonds must be broken for each exchange to occur.

On the other hand, it is much more energetically favorable if the substitutional impurity is adjacent to a silicon vacancy and the exchange happens between the vacancy and the dopant atom. When this exchange occurs, it is termed the vacancy mechanism \( (A + V \leftrightarrow AV) \), as schematically shown in Fig. 9-6b. The uncomplicated motion involved in the vacancy mechanism led early investigators to believe that the vacancy interchange should dominate dopant diffusion in silicon. This belief was supported by the fact that vacancies dominate diffusion in metals. Consequently, detailed models were developed to explain diffusion in silicon according to the vacancy mechanism.¹² ¹³

Diffusion via the vacancy mechanism, however, does not occur by a single interchange of a vacancy with a dopant atom. If that was indeed the case, there would be no net diffusion. That is, in order for the dopant to diffuse over a long range, the vacancy must diffuse away from the dopant atom at least to the third nearest neighbor. Only in this way can it return to the dopant atom via a different path, thus allowing diffusion to continue further along. It can be shown that
Fig. 9-6 Dopant movement at the atomic level: a) direct exchange; b) vacancy mechanism; c) substitutional-interstitial mechanism via "kick-out"; d) substitutional-interstitial mechanism via "kick-out". Reprinted with permission of the American Physical Society.
the activation energy for the diffusion of the AV associated pair is less than that of A diffusing by itself, and as a result the vacancy mechanism is more energetically favorable than the dopant atom diffusing by direct interchange.

The interstitialcy mechanism \((I + A \leftrightarrow AI)\) can be represented on an atomic level as a dopant atom in a substitutional site that is approached by a silicon interstitialcy (Fig. 9-6c). Recall that the interstitialcy defect has two associated atoms in non-substitutional sites around a substitutional site. In the case of the silicon interstitialcy both atoms are silicon. One of the silicon interstitialcy atoms knocks a dopant atom out of a substitutional site and forms a dopant interstitialcy. The dopant atom then moves toward a silicon atom on a lattice site and kicks out the silicon atom and forms a reconfigured dopant-interstitialcy. Unlike the vacancy mechanism, which requires that the AV pair partially dissociate, the interstitialcy mechanism does not have the same restrictions. The energy for this type of motion has been determined to be less than for self-diffusion by about \(\Delta Q_{AI}\) (the attractive potential between A and I), where \(\Delta Q_{AI} \approx -1\, \text{eV}\).

The final atomistic mechanism used to explain dopant diffusion is the "pure" interstitial mechanism \((I + A_x \leftrightarrow A)\). In this mechanism (see Fig. 9-6d), a silicon interstitial replaces a dopant atom on a substitutional site to form a dopant interstitial. The dopant interstitial moves through the interstices of the silicon lattice until it finds a substitutional site to occupy. When the dopant atom occupies the substitutional site it forms another silicon interstitial. This mechanism is very similar to the interstitialcy mechanism discussed above and as a result it is difficult to separate it from the interstitialcy mechanism.

The general consensus among investigators in the field regarding the atomistic mechanism controlling the movement of the common dopants in silicon is as follows: Phosphorus and boron have the largest amount of interstitial-type diffusion, while arsenic exhibits both vacancy and interstitial diffusion, and antimony seems to be dominated by vacancy-controlled diffusion. How these conclusions were reached is discussed in a subsequent section.

### 9.4 DIFFUSION MODELLING

It is important to be able to predict the movement of dopant atoms in Si during thermal processing with mathematical simulations. Simulation of the diffusion and redistribution of the dopant allows process development and integration engineers to substantially reduce the number of process experiments required during technology development. There are several simulation programs that are commonly used, with the most common being SUPREM (Stanford University Process Engineering Model). The modeling of diffusion in one-dimension is covered by SUPREM III, while the extension to two-dimensions is covered by SUPREM IV. SUPREM IV, of course, is also capable of treating the one-dimensional case. Some of the models that are used in SUPREM III and their extension to SUPREM IV will now be discussed.

#### 9.4.1 SUPREM III

The models in SUPREM III are empirically determined and are derived from the observed dopant-diffusion interactions in silicon. The basis of the diffusion model in SUPREM III is that diffusion is governed by the one-dimensional continuity equation, which accounts for the atom flux from the concentration gradient plus the enhanced flux that results from the built-in electric field formed from ionized impurities in a concentration gradient, and is given by Eq. 9-31:

\[
\frac{\partial C}{\partial t} = \frac{\partial}{\partial x} \left( D \frac{\partial C}{\partial x} \right) + \frac{q}{kT} \left[ \frac{\partial}{\partial x} \left( D C \frac{\partial \phi}{\partial x} \right) \right]
\]

\[\text{(9.31)}\]
Where \( D \) is the dopant diffusion constant, and \( C \) and \( C_i \) are the total and electrically charged dopant concentrations, respectively. The potential \( \phi \) that results from the concentration gradient is given by:

\[
\phi = kT/q \left[ \ln(n/n_i) \right]
\]  

(9.32)

where \( n \) is the carrier concentration resulting from dopant ionization and \( n_i \) is the intrinsic carrier concentration at the diffusion temperature. In many real problems the local charge neutrality is maintained and the second term on the right hand side of Eq. 9-31 can be dropped, resulting in Fick's Second Law. More simply, if \( D \) does not depend on the concentration of the dopant then the simplified Fick's Second Law holds as expressed in Eq. 9-33, below, which is the same as Eq. 9-6.

\[
\frac{\partial C(x,t)}{\partial t} = D \left[ \frac{\partial^2 C(x,t)}{\partial x^2} \right]
\]  

(9.33)

If the values of \( D \) are known, the solutions to Eqs. 9-31 and 9-33 will provide accurate simulations of the dopant concentrations for many applications.

There are two approaches used to obtain the values of \( D \). In the first approach the value of \( D \) is determined from measured diffusion profiles and then plotted for various temperatures and doping concentrations. Solutions to Eq. 9-31 and 9-33 are then fitted to the experimental plots by finding values of \( D \) that give solutions that match the \( D \) values extracted from the observed data. This type of model is known as a **phenomenological diffusion model**, because it is based on observation.

The second type of model assumes that the correct values of \( D \) can be obtained directly from the physics of the dopant-defect interaction. This type of model is referred to as a **point-defect-based diffusion model**. One such advanced model uses a kinetic Monte Carlo simulation to obtain the formation and diffusion energies associated with the defects, dopants, and impurities. The simulations, however, are number-crunching intensive and require a high-end workstation or a supercomputer to be able to get results in a reasonable time. The attraction of these models, however, is that they do not require diffusion profiles and empirical curve-fitting to predict the outcome. The description of such models is beyond the scope of this text and the interested reader is referred to the literature.\(^{15,16}\)

Phenomenological models have been implemented into SUPREM III and other 1-dimensional process simulators, such as PREDICT and RECIPRO. These models furnish an adequate fit to 1-dimensional diffusions, but are not successful at predicting the two-dimensional diffusion so critical for advanced devices with small feature sizes. The two-dimensional case requires the use of SUPREM IV. Also the phenomenological models are not adequate at predicting diffusion behavior of such defect-dominated phenomena as oxidation-enhanced diffusion (OED).

### 9.4.2 SUPREM III Models for Boron,Arsenic,Phosphorus,and Antimony Diffusion

The models used in SUPREM III are based on the **vacancy model under non-oxidizing conditions**, proposed by Fair and Tsai.\(^7\) These models, however, do not accurately reflect what is occurring on an atomic scale (since diffusion is related to both dopant/interstitialcy and dopant/vacancy interactions). However, the Fair-Tsai models are nonetheless very useful, since they provide an accurate representation of the diffusion profile for the common dopants.

The Fair-Tsai Model assumes that the diffusivity of an ionized dopant atom is based on the sum of the diffusivities of neutral vacancies and ionized vacancies, weighted by the probability
of their existence. According to the model there are four possible states of a vacancy: 1) the neutral vacancy \( V^0 \); 2) the single-negatively-charged vacancy \( V^- \); 3) the double-negatively-charged vacancy \( V^{-2} \); and 4) the single-positively-charged vacancy \( V^{+1} \). During extrinsic diffusion each contribution must be modified by the ratio of the doping level to the intrinsic carrier concentration raised to the power (including the correct sign) of the charge state. For example the contribution of the double-negatively-charged vacancy is modified by \((n/n_i)^2\), while that of the single-positively-charged vacancy by \((n/n_i)\). Thus, the effective diffusion coefficient under non-oxidizing conditions can be calculated from the sum of all the individual vacancy components. The effective \( D \) is given by:

\[
D = D^0 + D^- \left(\frac{n}{n_i}\right) + D^{-2} \left(\frac{n}{n_i}\right)^2 + D^+ \left(\frac{n_i}{n}\right)
\]

(9.34)

where the individual diffusivities on the right hand side of the equation correspond to the interaction between dopant atoms and neutral or charged vacancies.

**9.4.3 Modeling Intrinsic Diffusion**

Intrinsic diffusion is dominant when \( n \leq n_i \) at the diffusion temperature. In such cases the effective \( D \) is independent of dopant concentration and depends only on the diffusivities of the individual defects. Thus, when intrinsic diffusion is being modeled, Eq. 9-34 is re-written as:

\[
D^i = D^0 + D^- + D^{-2} + D^+
\]

(9.35)

and \( D^i \) is referred to as the *intrinsic effective diffusion coefficient*. The models used to determine \( D^i \) for boron, arsenic, phosphorus, and antimony are examined next.

**9.4.3.1 Boron:** Since boron diffuses as a negatively charged “atom” the model assumes that diffusion occurs primarily by the interaction of the boron with the neutral and positively charged vacancies. In that case the intrinsic diffusion coefficient of boron \( D_{b}^i \) can be represented by:

\[
D_{b}^i = D_{bV}^0 + D_{bV}^{+1}
\]

(9.36)

where:

\[
D_{bV}^0 = 0.037 \exp(-3.46 \text{ eV}/kT) \quad \text{cm}^2/\text{sec}
\]

(9.37a)

and:

\[
D_{bV}^{+1} = 0.72 \exp(-3.46 \text{ eV}/kT) \quad \text{cm}^2/\text{sec}
\]

(9.37b)

resulting in Eq. 9-38 for the total intrinsic diffusivity for boron \( D_b^i \):

\[
D_b^i = 0.76 \exp(-3.46 \text{ eV}/kT) \quad \text{cm}^2/\text{sec}
\]

(9.38)

**9.4.3.2 Arsenic:** The intrinsic diffusion of arsenic is assumed to be controlled by neutral and negative vacancies and the intrinsic diffusion coefficient for As \( (D_{as}^i) \) is given by:

\[
D_{as}^i = D_{as}^0 + D_{as}^{-1} = 0.066 \exp(-3.44 \text{ eV}/kT) + 12.0 \exp(-4.05 \text{ eV}/kT) \quad \text{cm}^2/\text{sec}
\]

(9.39)

**9.4.3.3 Phosphorus:** The intrinsic diffusion of phosphorus is assumed to be controlled by the interaction of the impurity ions with neutral vacancies only. Thus, \( D_p^i \) is given as:

\[
D_p^i = D_p^0 = 3.85 \exp(-3.66 \text{ eV}/kT) \quad \text{cm}^2/\text{sec}
\]

(9.40)

**9.4.3.4 Antimony:** The intrinsic diffusion of antimony is assumed to be dominated by interaction between neutral and single negatively charged vacancies. Thus, \( D_{st}^i \) is given by:
Fig. 9-7 Intrinsic diffusion coefficients vs. temperature for: a) boron; b) phosphorus. Reprinted with permission of North Holland Publishing Co.

\[ D_{\text{Si}} = D_{\text{Si}}^0 + D_{\text{Si}}^{-1} = 0.214 \exp \left( -3.65 \text{ eV/kT} \right) + 13 \exp \left( -4.08 \text{ eV/kT} \right) \text{ cm}^2/\text{sec} \]  

(9.41)

Figures 9-7 (a–d) show the Arrhenius plots of the intrinsic diffusion coefficients of B, P, As, and Sb. As can be seen from the plots, the agreement between the measured data and the calculated curves are quite good.

9.4.4 Extrinsic Diffusion Coefficients for the Common Dopants (As, P, and B)

An extensive amount of work has been done relating to the modeling of extrinsic diffusion of the common dopants As, P, and B. The earliest models developed by Fair were phenomenological expressions based on the result of careful determination of diffusion profiles. However, these models did not account for the effects of both I and V on dopant diffusion. Uematsu developed an integrated diffusion model based on measurable parameters that can predict the diffusion profiles of these dopants. The model considers contributions from the vacancy mechanism (V), the kick-out mechanism (I), and the Frank-Turnbull mechanism (F-T). The details of the Uematsu model are beyond the scope of this text and the reader is referred to the original paper. Here only the contributing defects for each of the common dopants are noted.

9.4.4.1 Arsenic: Figure 9-8 shows a typical diffusion profile for arsenic. There are several features that are common to As diffusion profiles. First, arsenic is a relatively slow diffuser and shows a relatively abrupt profile. Such features are desirable when fabricating shallow junctions for advanced CMOS and bipolar devices. Consequently arsenic is the most widely used n-type dopant in ULSI processing and is the mainstay for the n-channel shallow source/drain diffusion.
Fig. 9-7 Intrinsic diffusion coefficients vs. temperature for: c) arsenic; d) antimony. Reprinted with permission of North Holland Publishing Co.

In advanced CMOS devices and the emitter diffusion in bipolar and BiCMOS devices. The second observation is that the total As concentration ($C_{As}$) is frequently greater than the electrically active ($C_T$) concentration, particularly at high doping levels. It has been found that the amount of inactive As increases with the fourth power of the active concentration (i.e., $C_{As} / C_T \propto C_T^4$). Various models have been proposed to explain this phenomenon, including As-vacancy complex, small As clusters, and As-precipitate formation and dissolution. The arsenic-

Fig. 8-8 Arsenic diffusion profile. Note the steep diffusion front and the difference between the total As concentration $C_T$ and the active As concentration, $C_{As}$, for high As concentrations. Excellent agreement exists between simulated and measured profiles. Courtesy of American Physical Society.
vacancy cluster model appears to have the most support. Regardless of the mechanism, the fact that the amount of electrically-active As may change during processing must be taken into account. To obtain meaningful As profiles both SIMS (for $C_{T}$) and SRP (for $C_{AA}$) should be used.

As discussed previously As diffusion has nearly equal contributions from both I and V. Unlike boron diffusion, which is controlled by the presence of silicon self-interstitials, the I-component in As is controlled by neutral As-interstitials ($A_{S}^{0}$). The vacancy contribution is dominated by the presence of $V^{-2}$ and $V^{-1}$. The complex formation (As complexes) is dominated by the neutral vacancy complex, $(A_{S}^{0}V)^{0}$. Excellent agreement is observed between this model and experimental results as shown in Fig. 9-8.

**9.4.4.2 Boron Diffusion:** Figure 9-9 shows a typical profile for B diffusion. B is a relatively fast diffusant, and consequently great care must be exercised when attempting to fabricate shallow $p$-channel source/drain regions. The diffusion of boron from an implanted source is further increased by transient enhanced diffusion (discussed below). The formation of shallow boron diffusions is a fertile area of research and several techniques are candidates to achieve such shallow junctions. The most prominent approach is the use of *ultra-low-energy* (*ULE*) ion implantation (<1 keV), followed by high-ramp-rate rapid thermal annealing (see Chap. 10). Other approaches such as P-GILD, solid-source diffusion, plasma doping, and RTP gas doping are also under consideration.

Boron diffusion is known to be dominated by the “kick-out” mechanism. The diffusion profile can be modeled when three parameters are used, namely, neutral B interstitial (B$^{0}$), a neutral self-interstitial (I$^{0}$), and a positive self-interstitial (I$^{+}$). As shown in Fig. 9-9, excellent agreement between the model and experimental results have been obtained.

**9.4.4.3 Phosphorus Diffusion:** Although phosphorus is less commonly used in advanced devices, it still finds some application for the source/drain contacts in CMOS and for emitter diffusion in bipolar devices. Phosphorus diffuses quite normally in the intrinsic regime and is a relatively fast substitutional diffuser. When the dopant concentration exceeds the intrinsic carrier concentration, however, the in-diffusion behaves anomalously. By examining the P-diffusion profile in Fig. 9-10 one observes that at high concentrations the diffusion profile becomes quite complex and has a shape with three distinct regions: a) the *high concentration* or plateau region;
b) the transition or kink region; and c) the low concentration, or tail region. Figure 9-10 provides a schematic representation of these regions. No simple equation has been found to fit the profile and numerous models have been expounded to explain this strange behavior.

The most well known model was proposed by Fair and Tsai. A key element of that model is that a singly-charged phosphorus ion-vacancy pair, (PV)\(^1\) controls phosphorus diffusion. This defect is believed to have a negative charge, and is present in the high concentration region. The defect pair becomes important when it dissociates as it diffuses into the kink and tail regions. The dissociation is believed to create excess vacancies, which in turn enhances the phosphorus diffusion in those regions of the profile. The problem with this model is that I-defects are known to dominate P diffusion. Hu et al., described a model for the kink and tail regions that results from a two-stream (interstitial and vacancy) diffusion process.\(^{71}\)

More recently, Uematsu applied the unified model to explain anomalous P diffusion.\(^68\) In his model the diffusion is described in terms of a interstitial and a vacancy mechanism. For high surface concentration the vacancy mechanism (by way of a V\(^-2\)) governs the diffusion in the plateau region, while the kick-out mechanism through I and P\(_i^0\) dominates in the kink and tail regions. The changeover from the vacancy to the kick-out mechanism is what is believed to be responsible for the kink and tail profiles observed in P diffusion. For low phosphorus surface concentrations only the kick-out mechanism is operative and no kink and tail are observed.

**9.4.5 Modeling Diffusion with SUPREM IV**

The most advanced diffusion model in SUPREM IV is a physics-based model that depends on the interaction of point defects (silicon self-interstitials and vacancies) and the diffusant dopants. This model is capable of predicting diffusion in an oxidizing ambient. SUPREM IV is currently supported by the commercial suppliers Avanti/TMA (under the name TSUPREM4) and Silvaco, Inc.\(^{18}\) The advanced model in SUPREM IV was first developed to solve the coupled oxidation and diffusion problem while using the same grid. Readers interested in the equations used in this model are referred to the suppliers of these programs or the literature.

![Graph showing phosphorus diffusion profile](image_url)

**Fig. 9-10** Phosphorus diffusion profile for several surface concentrations. Note the presence of the “kink” and “tail” regions only for the high concentration diffusions. Excellent agreement exists between simulated and measured (SIMS) profiles.\(^{68}\) Courtesy of American Physical Society.
Two-dimensional modeling of diffusion is simulated by calculating the local diffusion constants based on the point defect and impurity concentrations at the location of interest. The more accurately the local defect concentration is known, the more accurately the local value of diffusion can be determined. Since the concentration of the point defects may depend on the geometry at a local region, it may be necessary to simulate the movement of the oxide interface and redistribution of the impurities if diffusion takes place in an oxygen-bearing environment.

To accurately model diffusion in two dimensions the following information is usually required:

1. The nature of the point defects,
2. How far the point defects move into the bulk,
3. How the point defects interact and recombine at the surface of the silicon and in the bulk regions,
4. The predominant mechanism that drives the diffusion (i.e., is it interstitialcy or vacancy).

The diffusion equations used in SUPREM IV are non-linear, since the value of D and built-in electric field depend upon the point defects and the carrier concentration (ionized dopants). Hence, a full solution requires the simultaneous solution of a set of non-linear coupled differential equations for the dopants and the defects. SUPREM IV solves these equations using numerical methods. The solutions produce several models that are useful for obtaining diffusion profiles in one- and two-dimensions. Models from SUPREM IV are briefly discussed below.

The simplest fastest model to simulate is the FERMI Model. It assumes that the point-defect concentrations depend only on the position of the Fermi level in the silicon. The actual point-defect concentrations are not calculated in this model. The FERMI model is most useful when the speed of the simulation is more important than its accuracy. The model does not work for oxygen-enhanced diffusion (OED), high-concentration diffusion, or implant-damage effects.

The TRANS Model accounts for the two-dimensional simulation of point defects. This model includes such events as the generation of point defects at interfaces, the diffusion of point defects into the bulk, and the recombination of defects at interfaces and in the bulk. The model is capable of simulating OED, but not high-concentration effects (such as the phosphorus kink and tail). It is also capable of simulating implant damage but not as well as the so-called FULL model. The TRANS model is typically used for routine simulations.

The FULL Model is the most accurate and requires significant computer CPU time to perform simulations. The FULL model does everything the TRANS model does, plus it simulates the effects of dopant diffusion on point-defect concentrations. It includes such refinements as pair saturation and dopant-assisted recombination effects. This model simulates high-concentration effects, OED, and implant-damage effects. Since it requires significant computer time, use of the FULL model is only recommended when maximum accuracy is necessary, or when high-concentration effects or implant-damage effects are pertinent.

9.5 DIFFUSION IN POLYCRYSTALLINE SILICON

Polycrystalline silicon is used as the gate electrode (conductor) in CMOS devices, as the emitter in bipolar devices, and as high-value load resistor in SRAMs (see Chap. 6 for more information on polysilicon). CMOS devices require that the polysilicon be heavily doped (degenerate) for maximum conductivity and to minimize the so-called “poly depletion” effect. The doping level
of the polysilicon at the polysilicon/oxide interface has a strong effect on the device threshold voltage. In deep-submicron CMOS devices it is necessary to dope some of the polysilicon with boron and some with arsenic on the same circuit. In advanced bipolar applications the polysilicon serves as the emitter contact diffusion source. Most often the emitter is doped with arsenic, although phosphorus may be used in less advanced applications. High-value resistors require controlled and repeatable resistivity of the polycrystalline-Si films.

Impurity diffusion in polycrystalline thin films, in general (and in polysilicon in particular), is quantitatively different from diffusion in single crystal material. The difference arises as a result of the grain boundaries present in polycrystalline films. That is, bulk diffusion is the controlling mechanism in single crystal material, while grain-boundary diffusion predominates in polycrystalline films. The diffusivity along grain boundaries can be up to 100 times faster than the diffusivity in bulk material.

Polycrystalline-silicon films are made up of grains with sizes ranging from 50 to 300 nm. The diffusion of the dopant within the grains is comparable to that of single-crystal silicon. But, the dopant atoms also diffuse much more rapidly along the grain boundaries, and then diffuse into the grains. Since the grains are small, only a short time is required for the dopant (which is entering from all sides of the grain) to fully diffuse within the grain. As a result, the overall diffusion is controlled by the grain boundaries, the grain structure, and the preferred orientation of the film. These quantities, in turn, depend upon such deposition conditions of the film as temperature, deposition rate, thickness, and post-deposition annealing treatments.

Each measurement of the diffusion constant D in polysilicon depends very much on the film’s history, and published results cannot be taken universally. Data has been obtained for the diffusivity of phosphorus, boron, and arsenic in polysilicon, and these diffusivity values are listed in Table 9-4. It has been found that phosphorus and arsenic can precipitate at grain boundaries, resulting in reversible changes of resistivity upon annealing. That is, the precipitates can dissolve at elevated processing temperatures and go into solution in the grains, thereby lowering the resistivity of the film. The final resistivity of the film thus depends not only on the doping level in the grains, but upon the grain size, any precipitates at the grain boundaries, and the presence of other defects that can reduce the mobility of the carriers.

### 9.6 DIFFUSION IN SILICON DIOXIDE

Silicon dioxide serves as a diffusion mask for both chemical and ion-implanted predeposition. Its sole purpose, in these cases, is to keep the diffusant atoms away from regions where they are not wanted. Silicon nitride (Si₃N₄) can be used for similar purposes. The diffusivities of the

<table>
<thead>
<tr>
<th>Dopant</th>
<th>D₀(cm²/sec)</th>
<th>E (eV)</th>
<th>D (cm²/sec)</th>
<th>T (°C)</th>
</tr>
</thead>
<tbody>
<tr>
<td>As</td>
<td>8.6 x 10⁻⁴</td>
<td>3.9</td>
<td>2.4 x 10⁻¹⁴</td>
<td>800</td>
</tr>
<tr>
<td>As</td>
<td>0.63</td>
<td>3.2</td>
<td>3.2 x 10⁻¹⁴</td>
<td>950</td>
</tr>
<tr>
<td>B</td>
<td>1.5–6.0 x 10⁻³</td>
<td>2.4–2.5</td>
<td>9.0 x 10⁻¹⁴</td>
<td>900</td>
</tr>
<tr>
<td>B</td>
<td></td>
<td></td>
<td>4.0 x 10⁻¹⁴</td>
<td>925</td>
</tr>
<tr>
<td>P</td>
<td>6.9 x 10⁻¹³</td>
<td></td>
<td></td>
<td>1000</td>
</tr>
<tr>
<td>P</td>
<td>7.0 x 10⁻¹³</td>
<td></td>
<td></td>
<td>1000</td>
</tr>
</tbody>
</table>
common dopants in SiO$_2$ can be calculated from measured profiles assuming solutions to Fick's second law, with the appropriate boundary conditions.

The Group III and V elements are known to form glassy networks with SiO$_2$ and accordingly their diffusivity strongly depends on their concentration. The diffusivities of these elements are very low for concentrations less than 1%, and generally do not need to be considered in detail. A recent paper reports on the diffusion of phosphorus from a phosphorus vapor source into thermal oxides.$^{28}$ The diffusivity of phosphorus is given by:

$$D_p = 3.79 \times 10^{-9} \exp\left(-\frac{2.3}{kT}\right) \text{ (cm}^2/\text{sec)}$$  (9.42)

and it was found to be independent of the phosphorus concentration. The solubility of phosphorus in the glass, in the temperature range of 1000$^\circ$C, was found to vary between $3 \times 10^{20}$ and $2 \times 10^{21}$ cm$^{-3}$. The diffusion mechanism is described as phosphorus dissolving in the interstitial sites as P$_2$, where it is incorporated into the network of SiO$_2$. The P$_2$ exchanges sites with silicon atoms and continues diffusing through the silicon sites.

### 9.6.1 Boron Penetration of Thin Gate Oxides

In the case of thin gate oxides, the diffusion of boron through the gate oxide must be considered. This effect is termed boron penetration. During boron penetration, dopant atoms from the heavily-boron-doped polysilicon can diffuse through the thin gate oxide layer and into the device channel, where its presence can then change the device threshold. Also, boron incorporated in the gate oxide during the diffusion can degrade the oxide breakdown characteristics and charge trapping rate.$^{29}$ It is further noted that the diffusion of boron through oxide is enhanced by the presence of hydrogen and fluorine in the oxide. Fluorine can be incorporated into the gate oxide if the boron is introduced into the polysilicon using BF$_2$ as the implantation source.

There are two main ways to reduce boron penetration or diffusion through a thin oxide. The most common approach is to incorporate nitrogen into the gate oxide. This technique is known to reduce boron diffusion by the nitrogen bonding in the glassy network, which in turn impedes the boron flux through the glass. A model has been developed to explain why the presence of nitrogen in the oxide hinders the boron diffusion.$^{30}$ The model explains that the value of $D_0$ decreases and that of $Q_A$ increases as the nitrogen concentration increases.

The diffusivities (at 900$^\circ$C) of the common dopants in SiO$_2$ are listed in Chap. 8, Table 8-5. There is a class of materials that are fast diffusants in SiO$_2$, including H$_2$, Hc, OH$^-$, Na, O$_2$, and Ga. Values of $D$ greater than $10^{-13}$ cm$^2$/sec have been determined for these elements.

### 9.7 ANOMALOUS DIFFUSION EFFECTS

Any deviation from in the diffusion behavior predicted by Fick's Law may be considered as anomalous behavior. The most common anomalous diffusion effects include: a) the effect of high dopant concentration on diffusion behavior (extrinsic diffusion), as discussed earlier; b) the enhancement of diffusion as a result of the built-in electric field; c) the effect of sequential diffusions including the “emitter push effect”; d) lateral diffusion under a window; e) oxidation-enhanced (and retarded) diffusion, and f) transient enhanced diffusion. To a large extent these phenomena are dependent upon the interaction of the dopant atoms with silicon point defects. Their study helps better understand the interactions between dopants and defects.
9.7.1 Electric Field Enhancement

During diffusion at elevated temperatures, the diffusing species are usually ionized. As a result, a gradient of charge (or more formally, a gradient in the Fermi level) will develop in the crystal. The gradient of the Fermi level will result in a built-in electric field. The electric field always operates in the direction to enhance the diffusion of the ionized impurities.

For the case of diffusion of a dopant-defect (AX) pair in the presence of an electric field, the total diffusion flux, \( J \), will have both a classical diffusion component and an electric-field drift component. The flux of a dopant-defect pair (AX) in the presence of an electric field is then given by:

\[
J_{AX} = -D_{AX} \frac{\partial C}{\partial x} + Z_{AX} \mu_{AX} C_{AX} E \tag{9.43}
\]

where \( Z_{AX} \) defines the charge state of AX dopant/defect pair and \( \mu_{AX} \) is the mobility of AX. Assuming Boltzmann statistics pertain and that the Einstein relation between the mobility and the diffusivity also applies, the calculation of the enhancement of the diffusivity due to an electric field is rather straightforward, but lengthy. The reader is referred to the original reference for the complete solution.\(^{31}\)

The enhanced diffusivity in an electric field is given by:

\[
D_A = h \left[ D_{A^+X^o} + D_{A^+X^-} \cdot \left( \frac{n}{n_i} \right) + D_{A^+X^-} \cdot 2 \left( \frac{n}{n_i} \right)^2 \right] \tag{9.44}
\]

where \( h \) is defined as:

\[
h = 1 + \frac{C_{A}}{2n_i} \left[ \left( \frac{C_A}{2n_i} \right)^2 + 1 \right]^{-1/2} \tag{9.45}
\]

The value of \( h \) varies from 1 for the case of \( n \ll n_i \), and reaches 2 for the case of \( n \gg n_i \). The field enhancement may not only result from the presence of the diffusing dopant but also from the presence of another dopant in the crystal from a previous diffusion step.

9.7.2 Emitter Push Effect

Since the performance of an \( npn \) bipolar transistor depends precisely upon the width of the base region, the diffusion of the emitter must be accurately controlled. In the early days of bipolar device fabrication it was found that the measured base width, after the emitter drive, was larger than predicted from the diffusion coefficients of both the phosphorous and the boron. During the emitter drive, the boron (under the phosphorus) diffused deeper into the silicon than expected. This effect became known as the emitter push (or emitter dip effect). The emitter push effect is depicted schematically in Fig. 9-11.

![Fig. 9-11](Image)

**Fig. 9-11** Enhanced base diffusion under the emitter resulting from a heavy phosphorus diffusion. (Emitter push [or dip] effect).
Fair proposed that the enhanced boron diffusion under the emitter results from a combination of effects, related to the creation and dissociation of PV pairs. The dissociation of the pairs near the boron increases the vacancy concentration and thereby the boron diffusion. It is now known that interstitials play a key role in both boron and phosphorus diffusion. Other experiments have shown that there is indeed a net interstitial supersaturation below the phosphorus diffused layer and these excess interstitials enhance the boron diffusion.

Fortunately the importance of the phosphorus emitter push effect has diminished, since arsenic has replaced phosphorus as the emitter dopant. This change was made because arsenic allows shallower and better-controlled junctions to be formed. The use of arsenic, however, is not without its problems, since arsenic tends to form clusters and also tends to interact with the boron. In some cases arsenic was found to retard the diffusion of the boron (emitter pull).

### 9.7.3 Lateral Diffusion Under Oxide Windows

Since the majority of diffusion processes are carried out in masked areas one must be concerned not only with vertical diffusion, but also with lateral diffusion under the mask. This is not truly an anomalous effect since a concentration gradient exists in the lateral direction, as well as in the vertical direction. Thus, diffusion should proceed in both directions (two-dimensions).

One of the key processing issues related to lateral diffusion occurs when boron is diffused into the source-drain regions of a CMOS device, and the boron diffusion proceeds laterally under the edge of the gate electrode. There are two harmful effects that arise from lateral diffusion under the gate edge. First, the gate/drain overlap capacitance increases, and this can cause a decrease in the circuit performance. Second, the length of the active channel region (the distance between the source and drain) will decrease. If this decrease is excessive, it is possible to reach what is termed a “punchthrough” condition. When punchthrough occurs either the leakage current of the device substantially increases or the device will cease to function. Figure 9-12 shows the effect of boron lateral diffusion in the source/drain region of an MOS device.

For intrinsic diffusion under both constant-source (i.e., pre-deposition) conditions and limited-source (i.e., drive-in) conditions, the lateral penetration is found to be about 75–85% of the vertical diffusion depth (Figs. 9-13a and 9-13b). At high-concentration (i.e., extrinsic) diffusion, as well as for shallow diffusions, lateral penetration is found to extend only 65–70% of the vertical diffusion distance. There is an additional anomalous effect caused by the stress generated at the edge of an oxide or nitride window. Both enhanced and retarded diffusions

---

**Fig. 9-12** The effect of lateral diffusion in an MOS device. The channel length is reduced to less than the polysilicon line size by the amount of lateral diffusion.
Fig. 9-13 Diffusion contours at the edge of an oxide window: a) constant source diffusion; b) limited source diffusion.

...sions have been observed as a result of elastic strain near the window edge. It is known that strain can induce defects which affect the diffusivity of the dopants in silicon.

It is extremely difficult to directly measure the lateral diffusion in shallow junctions and the electrical performance of a MOS device may be used as an indication of lateral diffusion. TEM has been employed for this task, but it is difficult to discern the precise junction location. Some new techniques are beginning to be used for such measurements and they will be discussed in a later section.

9.7.4 Oxidation-Enhanced Diffusion (OED)

It has been known for some time that when diffusion takes place under oxidizing conditions the motion of certain dopants is greatly enhanced, as compared to the same diffusion in a neutral ambient. To be more precise, the diffusion of both boron and phosphorus are enhanced during oxidation. While arsenic diffusion is only slightly enhanced under oxidizing conditions, the diffusion of antimony (Sb) is retarded under the same oxidation conditions. The general name for this phenomenon is oxidation-enhanced diffusion (OED). There also exists a subsidiary effect called nitridation retarded diffusion (NRD), which will be discussed subsequently.

It is important to examine OED from two different perspectives. First of all, the amount of OED that occurs during a drive-in step must be known to accurately model diffusion. From a practical point of view one needs to take OED into account when defining a diffusion schedule. Failure to do so could lead to a much deeper junction than anticipated. It is also important to consider the possibility that different regions on the device may exhibit differing amounts of OED, since the amount of OED depends on the oxidation rate and different parts of the device may have different oxidation rates.

The second perspective is that the study of OED provides a window on how silicon vacancies and interstitials interact with the common dopants during diffusion. There have been a myriad of investigations of OED in silicon, and the treatise by Fahey, Griffin, and Plummer provides an excellent critical review of the OED literature through 1989. Some of their conclusions are briefly described here, but readers interested in the details of the analysis should refer to the Fahey, et al., paper.

As far back as 1974, S.M. Hu recognized that the growth of oxidation-induced stacking faults, OISF’s (see Chap. 2) and OED had a common origin. Direct observation of stacking faults in the TEM revealed that they consisted of silicon interstitial type defects, bounded by Frank dislocations. It was also known that the stacking faults grew during oxidation and the
growth was controlled by silicon atoms attaching themselves to the ends of the dislocation. The silicon atoms were injected into the bulk of the silicon from the surface during the oxidation process. Armed with the fact that interstitials enter the silicon during oxidation it was postulated that these same interstitials also affect OED.

Based on various OED experiments over the past 20 years a convincing and repeatable correlation between defects and dopants emerged. The conclusions are:

1. Since the diffusion of boron and phosphorus are enhanced by oxidation, the dominant (>50%) mechanism controlling their diffusion is the presence of self-interstitials,
2. Since the diffusion of antimony is retarded during oxidation the dominant mechanism controlling its diffusion is the presence of vacancies,
3. Since the diffusion of arsenic is not greatly enhanced or retarded during oxidation, there is no dominant mechanism, and both vacancies and interstitials play an important role.

The literature is replete with studies on how OED is affected by temperature and time. To perform these studies it is necessary to measure and quantify OED. The measurement technique used is shown in Fig. 9-14. Here the measurement is performed by comparing the diffusion of a dopant during oxidation with the same dopant when the surface is not being oxidized (at the same time and in the same sample). This is accomplished by masking part of the wafer with silicon nitride. It is well known that oxide will not grow under the nitride. The OED is then determined from the difference in the junction depth between the masked and unmasked regions.

The results of these studies are briefly summarized here. The major conclusion is that OED depends on the oxidation rate and the dependence follows the power law:

\[ \text{OED} \propto (dx_{ox}/dt)^n \]  

where \( (dx_{ox}/dt) \) is the oxidation rate and \( n \) is an experimentally determined fitting parameter, with a value less than 1, and typically in the range of 0.2–0.3.

When the oxidation rate is increased (for instance, during wet oxidation or during high temperature growth), it is expected that the relative amount of OED should decrease. The

![Fig. 9-14](image_url) A schematic showing how to measure oxygen enhanced diffusion (OED).
relative reduction of the OED with higher temperatures has been confirmed. There were also
studies performed on the effect of silicon crystal orientation dependence on OED. In these
studies it was found that the relative amount of OED decreased in the order (111), (110), and
(100) during dry oxidation, in line with the expected oxidation rates. It was also found that for
oxidation temperatures in excess of 1150°C, the diffusion of boron was actually retarded, rather
than enhanced, implying the injection of vacancies at this high temperature. At the same
time antimony exhibited enhanced diffusion, confirming the presence of excess vacancies.

There have also been a series of experiments that studied OISF shrinkage and diffusion
during the direct nitridation of silicon by ammonia (NH₃). It was found that during nitridation,
OISF actually shrink in size. This shrinkage could result from a vacancy reaching the OISF and
changing places with a silicon atom, thus making the stacking fault smaller. It is believed the
vacancies are injected from the surface of the silicon during the nitridation. The mechanism for
the creation of the vacancies is not yet well understood. One idea is that the stress-relief
mechanism prevalent during the high-temperature nitridation is the formation of Frenkel pairs.
The silicon interstitials from the Frenkel pairs are consumed by the nitride, while the vacancies
enter the silicon where they react with the interstitials from the OISF.

If one considers the results of diffusion studies on the common dopants during nitridation
one discovers that the diffusion of boron and phosphorus are retarded, while that of antimony is
enhanced. This effect is parallel, but opposite, to that observed during oxidation. Since the
diffusion of boron and phosphorus is dominated by interstitial defects, their undersaturation in
the bulk will result in a reduced diffusivity. On the other hand, since the diffusion of antimony is
dominated by vacancies it is not surprising to find enhanced diffusivity along with super-
saturation of vacancies. The study of both OED and NRD provide a set of self consistent results
which show a clear linkage between the diffusion of the dopants and the presence of defects in
the silicon.

9.7.5 Transient Enhanced Diffusion (TED)

After an ion implantation step it is necessary to thermally treat (anneal) the silicon to activate
the implanted ions (i.e., to move the dopant atoms to substitutional sites), and to remove residual
implantation damage. During this anneal a new type of anomalous diffusion is observed, called
transient enhanced diffusion or TED. The study of TED has become an area of fertile research,
and by understanding TED the knowledge of the interaction between dopant atoms and defects
has also been increased.

Before the discussion of the models that have been posited to explain TED is undertaken,
some of the features that are common to this phenomenon will be described. First, TED is
associated with the ion implantation damage and does not occur if the dopant is introduced by
chemical means. Second, the diffusion enhancement is transient. That is, it exists for a finite
amount of time after which "normal" diffusion returns. Finally, TED takes place in the
temperature range where little or no atomic movement is expected (670–900°C). As a matter of
fact, as much as 100 nm of boron dopant movement has been measured in this low temperature
range.

These observations will now be discussed in terms of the proposed mechanism for TED.
Briefly, TED is explained by assuming that silicon interstitials are injected from the implant
damaged region into the bulk of the material. Furthermore, this interstitial injection occurs at
relatively low temperatures for some finite time. The interstitials, once in the bulk, enhance the
diffusion of boron and phosphorus. The sources of these interstitials are {311} defects, which
Fig. 9-15 A comparison of P-diffusion SIMS profiles "as-implanted" (initial) vs. after a furnace anneal at 800, 1000, and 1100°C. There is more diffusion occurring at 800°C than at 1100°C as a result of TED.\textsuperscript{73} coalesce after ion implantation damage. During the anneal cycle, these \{311\} defects begin to dissolve, and during their dissolution they emit interstitials into the silicon.

Figure 9-15 shows a SIMS plot of the phosphorus concentration as a function of distance in both the "as-damage-implanted" case and after anneal at several temperatures. The astonishing result is that diffusion proceeds farther at 800°C than at 1100°C. The fact that so much atomic movement can occur at such low temperatures implies that TED can take place during the loading and unloading of the wafers in a conventional furnace. Before a furnace annealing process even starts, the atoms may have moved too far to make useful junctions. This is one of the reasons why RTA has rapidly replaced furnaces for annealing the implanted atoms for shallow junctions. Fig. 9-16 displays a plot of the enhanced diffusion of phosphorus as a function of time at temperature for two different levels of implant damage. The conclusions that can be drawn from this curve are: a) the enhancement decreases rapidly with time, and b) for short times the enhancement is independent of the amount of damage. A phenomenological equation can be written for the time dependence of the enhanced diffusivity as:\textsuperscript{42,44}

\[ D_E = D_0 + D_{TED} \exp(-t/\tau) \]  \hspace{1cm} (9.47)

Fig. 9-16 Amount of enhanced diffusion as a function of time at temperature for two different levels of implantation damage.\textsuperscript{74} Reprinted with permission of the American Physical Society.
where $D_B$ is the enhanced diffusivity; $D_i$ is the intrinsic diffusivity; $D_{TED}$ is the enhancement factor at the onset of TED; $t$ is the time into the TED event; and $\tau$ is the time constant for the enhanced diffusion. For short times and low temperatures, the total diffusion is dominated by the $D_{TED}$ term.

It is difficult to measure TED directly from the SIMS profile in the implanted silicon since the measurement is confounded by too many variables, including: a) the TED itself; b) the damage profile; c) the stationary component of the profile; and d) the variation of the interstitial profile in the implant region. A novel technique has been developed to directly measure TED. The technique uses CVD or MBE epitaxially-doped “superlattice” regions on float zone (FZ) silicon. These samples consist of regularly spaced spikes of dopant which are used as “marker” layers to monitor the atom movement during the TED event. A damage implant (either silicon or germanium) is then made into the sample silicon to create the required defects. The distance of the damage layer has been well characterized.

Figure 9-17 shows a SIMS profile of the as-deposited and diffused profiles of these spikes (boron doped in this case), as a function of depth after the introduction of ion implantation damage and anneal. The damage was introduced by a silicon implantation and the furnace anneal was performed at 815°C for 10 and 30 min. Also shown on that curve is the depth (about 100 nm) of the implant damage. The boron profile before diffusion shows uniform sharp peaks as a result of the doping spikes that were grown-in during the low-temperature MBE. The results after the damage implant and anneal show that some of the boron spikes significantly broaden as a result of the thermal treatment. The amount of broadening is a direct measure of the diffusivity of the boron. From Fig. 9-17 one sees that the boron markers located within and adjacent to the implant damage remain immobile during annealing, while those farther away show broadening (indicating diffusion). The amount of broadening increases, and then decreases, as it moves farther away from the implant damage region. The broadening that is observed at the 350-nm marker corresponds to a 1000 times increase over the normal diffusivity of boron. The fact that no additional broadening takes place after 10 min testifies to the fact that the TED is a transient phenomenon. The duration of the transient, on the other hand, does strongly depend on the annealing temperature and is found to decrease from over 5 hours at 670°C to less than 10 minutes at 815°C. These times are related to the constant $\tau$ in Eq. 9-47.
The observation that the concentration peaks of the boron in or near the implanted region remain stationary, suggests that near the implant damage the boron atoms form immobile clusters (Blₙ where n = 1, 2, or 3) with the excess silicon interstitials. The results of the marker experiments (no broadening) along with the fact that the electrical concentration of B is generally much less than the chemical concentration in that region, provides additional evidence of B-I cluster formation.

The source of the interstitials and their time dependence is now discussed. Eaglesham et al., performed high resolution TEM studies of silicon in the implanted region after very short time anneals in order to observe the defect structure. The appearance of an extended defect is quite obvious, and the defect has been identified as the rodlike or [311] defect. These defects are known to be present during oxidation and ion implantation of silicon. It is generally agreed that the [311] defect consists of a condensation of interstitials which form five- and seven-member rings. Another way of looking at these defects is that they consist of an extra monolayer of hexagonal silicon in the lattice that has no dangling bonds. As a result, these defects are relatively stable, except at high temperatures (> 950°C).

Using TEM, Eaglesham et al., were able to measure the time dependence of the density and size of the [311] defects during short-time anneals and then to calculate the density of silicon interstitials. Figure 9-18 shows the time dependence of the silicon interstitial density (the so-called evaporation rate) for several anneal temperatures. The evaporation rate was found to have an exponential dependence on time, with time constants that correspond very well to the duration of the TED as described in Eq. 9-47. Recall that the duration of TED decreased as the temperature increased. This correspondence strongly suggests that TED results from the growth and dissolution of [311] defects that are formed as a result of ion implantation. During the dissolution there are a large number of silicon interstitials formed. They greatly enhance the diffusion rate of boron, (and phosphorus) through the "kick-out" or interstitialcy mechanism.

It has been found that the presence of substitutional carbon in silicon in the range of 5 × 10¹⁸/cm³ effectively suppresses TED without affecting the electrical activity of the boron. It is believed that the carbon atoms absorb the injected interstitials and therefore the interstitials are not available to contribute to TED.

Although TED has been studied most extensively with boron diffusion, both phosphorus and antimony also demonstrate the effect. In order to get TED to occur in antimony it is

![Fig. 9-18](image)

**Fig. 9-18** Evaporation of interstitials from [311] defects as a function of time at several temperatures.
necessary to achieve vacancy supersaturation in the region of the diffusion. This can be
accomplished using MeV silicon implantations.

9.8 DIFFUSION SYSTEMS AND DIFFUSION SOURCES

Regardless of the predeposition process (chemical or ion implantation) a drive-in (or
redistribution) step is usually required to move the dopants in the silicon (i.e., to attain electrical
activation, and reduce the implantation damage). This drive-in step is performed either in a
diffusion furnace or a rapid thermal annealing (RTA) system. (These technologies are both
discussed in detail in Chap. 8.) However, RTA allows dopant activation to occur using a smaller
thermal exposure than can be obtained with a furnace. As a result, the use of RTA for annealing
ion-implanted shallow source/drain junctions in CMOS device processing has become more
popular. There are some advanced device flows that use only RTA for drive-in diffusions.

Not all fabrication processes, however, use ion implantation for dopant introduction. Some
still use chemical procedures. For example, backside phosphorus gettering (Chap. 2) uses a
POCl₃ source and a diffusion furnace, while bipolar base diffusions may be predeposited in
diffusion furnaces sourced with boron nitride (BN) discs. Nevertheless, ion implantation is the
standard technology for introducing dopants into Si wafers, and this topic is covered in Chap.
10. New technologies are still being developed for creating shallow junctions in ULSI devices,
and a discussion of these follows. Older technologies for chemical diffusion of dopants into Si
are covered in the 1st edition of this text (© 1986); interested readers are referred there for this
information.

9.8.1 Advanced Diffusion Technologies

Advanced technologies have been studied for introducing dopants into a silicon device. These
new techniques include: a) plasma immersion ion implantation doping (PII); b) rapid vapor
phase doping (RVD); and c) gas immersion laser doping (GILD). In this section, RVD and
GILD will be briefly discussed. The subject of PII is covered in Chap. 10.

9.8.1.1 Rapid Vapor Doping (RVD): Rapid vapor doping (RVD) is emerging as an alternative to
ion implantation doping for some advanced applications, such as capacitor doping for DRAMs,
side-wall doping in deep trenches, and even for shallow source/drain junctions in CMOS
applications. Unlike vapor doping in a furnace, RVD does not depend on the formation of a
doped glass layer on the surface of the wafer. For RVD, the doping proceeds directly from the
gas phase into the solid. For that to occur the dopant must first be adsorbed on the silicon
surface, followed by solid-state diffusion. The potential advantages of the RVD technique over
ion implantation (particularly for shallow-junction applications) are that RVD does not suffer
from the effects bedeviling implantation, namely: channeling, lattice damage, or wafer charging.

Rapid vapor doping uses rapid thermal processing (RTP) to quickly and uniformly heat the
wafer to the desired temperature while the wafer is exposed to a gaseous dopant source. The
precise control of the gas flow and dilution are necessary to ensure a uniform dopant
concentration across the wafer surface. The final dopant surface concentration (Cₛ) and junction
depth (xₐ) depend on the gas phase concentration of dopant and the duration and temperature of
the RTP treatment. The gas phase concentration is controlled by diluting the source gas (e.g.,
phosphine) with a hydrogen carrier gas. There are some disadvantages of RVD, such as: a)
a relatively high thermal budget is required for the surface reaction and solid-state diffusion to get
the dopants into the silicon; and b) "hardmasks" are required for selective doping.
There have been reports describing the RVD fabrication of shallow boron junctions from vapor sources. Kiyota et al., reported forming ultra-shallow boron junctions by flowing B₂H₆ in H₂ at 900°C for 30 sec.⁴⁹ This process resulted in a junction depth of 30 nm and a surface concentration of 5.8 x 10¹⁹ cm⁻³. This level of doping is adequate for fabricating base regions in bipolar devices or the source/drain extensions in CMOS devices. Short-channel CMOS devices produced using RVD showed improved performance over devices produced by conventional ion implantation.⁵⁰ Similar studies using phosphine (PH₃) diluted in hydrogen at 900°C reported junction depths <100 nm and surface concentrations between 3.0 x 10¹⁸ and 1.3 x 10¹⁹/cm³.⁵¹

9.8.1.2 Gas Immersion Laser Doping (GILD): The origins of GILD technology extend back more than 15 years, when lasers were first used to locally melt silicon immersed in a doping gas (such as PF₅ or BF₃).⁵²,⁵³ The technology has been refined since that time, and it is now possible to fabricate devices using GILD technology. The basis of GILD technology is to shine high energy density (0.5–2.0 J/cm²) eximer laser light (308 nm) for a short pulse time (20–100 ns) onto a silicon wafer immersed in a dopant gas phase. The energy is absorbed into the first 10 nm of the silicon surface, where it dissipates by melting the surface layer. The thin melted layer absorbs and dissolves the dopant atoms directly from the gas phase. Since the diffusivity of the dopants in the liquid phase is about 8 orders of magnitude faster than in the solid phase, the dopants rapidly and uniformly diffuse throughout the molten layer. When the laser power is turned off, the thin molten layer epitaxially recrystallizes (freezes). During recrystallization the dopant is incorporated into the silicon lattice on substitutional sites. Consequently, the majority of the dopant is activated immediately after cool down and no further anneal is required. The liquid-to-solid transformation occurs at an extremely high rate (> 3m/s) and it is possible for the amount of active dopant incorporated in the silicon to exceed the equilibrium concentration (supersaturation). As long as the wafer is not subjected to subsequent elevated temperature processing, the excess dopant will stay in solution.

Since the time that the wafer is exposed to the high-energy laser beam is short, the bulk of the wafer remains cool and little or no solid-state diffusion occurs. As a result the dopant is confined only to the molten region and any dopants that were previously introduced into the silicon remain unperturbed. The depth of the junction is controlled by the depth of the molten silicon, which depends on the energy density and pulse time of the laser treatment.

Since the concentration of the dopant is uniform throughout the “diffused” region the concentration profile tends to be more box-like rather than erfc or Gaussian. Consequently, the profile gives an abrupt junction. Figure 9-19 shows the diffusion profile after a GILD of phosphorus in a p-type substrate. The box-type profile is apparent. A 30 ns pulse results in a junction depth of 30 nm with a surface concentration 4x10²⁰/cm³. As the pulse time increases the junction depth gets deeper.

A major innovation towards making GILD a production-worthy technology was the introduction of projection-GILD or P-GILD. This technology utilizes a step-and-repeat camera, similar to those described for projection lithography in Chap. 13. Eximer laser light is shined through a dielectric mask. The light from the mask is demagnified (e.g., a 4x reduction), focused, and projected onto the wafer. The wafer is stepped after each field of the mask is exposed (or melted, in this case). The amount of doping incorporated into the layer increases as the number of laser exposures increase, while the depth of the junction increases with the increase in the energy of each pulse. The dopant is incorporated only in the regions that are melted (i.e., only beneath the regions of the mask that transmit the laser light).
Fig. 9-19 Spreading resistance profile (SRP) for a phosphorus doped P-GILD formed in a p-type doped substrate. Note that $x_j$ increases with surface melt time and the diffusion profile forms a "box" type profile with no anomalous tail. $^\text{53}$

### 9.9 MEASUREMENT TECHNIQUES FOR DIFFUSED (AND ION-IMPLANTED) LAYERS

The diffusion coefficient $D$ is experimentally determined from the concentration profile in the silicon subsequent to chemical predesposition (or ion implantation) and drive-in. In order to determine accurate values of $D$, reliable measurements of the junction depth, sheet resistance, and diffusion profile must be feasible. Sheet-resistance measurements have become the standard method for "in-line" monitoring of diffusion processes in IC production facilities. These measurements are performed immediately after a diffusion step. In this section some of the important measurement techniques for determining sheet resistance and doping profiles are described.

#### 9.9.1 Sheet Resistance Measurements

The *sheet resistance*, $R_s$ of a diffused layer is the resistance exhibited in a square (i.e., a region of equal length and width) of that layer, which has a thickness $x_j$ (junction depth). The value of $R_s$ is expressed in units of ohms/sq ($\Omega/$sq), and is related to the *average resistivity* $\rho$ of a diffused layer by:

$$R_s (\Omega/sq) = \rho \text{ (ohm-cm)/(cm)}$$  \hspace{1cm} (9.48)

The layer (or film) resistivity is a material property that depends on the number of free carriers and the mobility of these carriers. The average resistivity $\rho$ is found from:

$$\rho = [q \int_0^{x_j} \mu (C_A(x)) C_A(x) dx]^{-1}$$  \hspace{1cm} (9.49)

where $q$ is the charge on a carrier $\mu (C_A)$ is the carrier mobility (which is a function of the carrier concentration), and $C_A(x)$ is the diffusion profile.
The value of $R_s$ is readily obtained by measuring the resistance of the diffused layer using the 4-point probe technique. Figure 9-20 shows the co-linear probe arrangement that is the basis for the ASTM standard, which is the standard for measuring $R_s$. To make this measurement a current, $I$, is forced between the outer two probes, and the voltage drop, $V$, is measured between the two inner probes. When all of the probes are equally spaced, the resistivity is given by:

$$R_s = 2\pi s F (V/I) \tag{9.50}$$

where $s$ is the probe spacing, and $F$ is a correction factor (that arises from the fact that the sample is not semi-infinite, but has a finite thickness and diameter). When the sample thickness $t$ is small, relative to the probe spacing (as in the case of a diffusion, where $t = x_d$), then Eq. 9-50 can be rewritten as:

$$\rho = 4.532 t \left( \frac{V}{I} \right) \quad \text{or} \quad R_s = \frac{D}{t} = 4.532 \left( \frac{V}{I} \right) \tag{9.51}$$

There is an additional set of correction factors related to the finite diameter of the wafer and the effect of the probes being close to the edge of the wafer. The correction factor is 1 if the probes are at least 3x their spacing from the wafer edge. When the probes are closer to the edge, the correction factor decreases to a value of 0.5-0.7, depending on the orientation of the probes to the wafer edge. The factors are listed in Ref. 64. For a standard probe spacing of 0.0625 inches, the probes need at least 4.7 mm from the wafer edge in order not to require a correction factor. To prevent erroneous readings the probe spacings must be accurately known, and if they are not equal, that must also be taken into account in the correction factor.

The sheet resistance measurement is first performed with the current in the forward direction, and then in the reverse direction (in order to minimize errors due to thermoelectric heating and cooling effects). The two $(V/I)$ readings are averaged. The measurements should also be made at several current levels, until the proper level is found. That is, if the current is too low, the values of the forward and reverse readings will differ, while if it is too high, PR heating will cause the measured reading to drift with time. The ASTM standard also recommends best current levels, which depend on the resistivity range of the diffused layer. The amount of pressure applied to the probes during measurement is also important, since excessive pressure can drive the probes past the junction, while too little pressure results in poor electrical contact to the sample.

Automated 4-point probe equipment is currently available from several vendors. The equipment is capable of measuring the sheet resistance in numerous places on the wafer. The

---

**Fig. 9-20** Co-linear four-point probe arrangement for the measurement of resistivity.
system has a built-in algorithm to calculate the sheet resistance at these points and to use the appropriate correction factor for each position. The data can be presented in a graphical format using a *contour map* of resistivity. In this way the process engineer can look for process variations that are wafer dependent. The program can also calculate and depict statistical information about the sheet resistance variation. Such tools are used for in-line process control in many wafer fabs.

It is possible to estimate the doping profile of a diffusion using the *differential sheet resistivity* technique. In this technique, the sheet resistivity is measured initially and then again after a small amount of the silicon surface is removed. Anodic oxidation, followed by an etch of the anodic film, is used to control the amount of silicon removed in each step. The sheet resistivity will increase as dopant is removed, since the amount of doping is decreasing and the thickness of the diffused layer is decreasing with each step. The rate of the increase in $R_s$ depends on the resistivity profile shape. The resistivity depends on the dopant concentration and the carrier mobility. If the mobility is known for each value of resistivity, the carrier concentration profile can be obtained with this technique. Note that the differential-sheet-resistivity method is tedious, and is only used in special cases where other profiling techniques are not available.

### 9.9.2 Capacitance-Voltage (C-V) Measurements

The carrier concentration profile can be measured directly within a diffused region by using the C-V technique. The use of C-V measurements to determine the diffusion profile is predicated on the measurement of the reverse bias capacitance of an $n^+/p$ or $p^+/n$ junction as a function of voltage. The capacitance of a junction is given by:

$$
C(V) = \left[ \frac{q\varepsilon_{Si}C_A}{2} \right]^{1/2} \left[ V_{bi} \pm V_R - \frac{2kT}{q} \right]^{1/2} \quad (9.52)
$$

where $\varepsilon_{Si}$ is the permittivity of silicon, $C_A$ is the substrate doping concentration, $V_{bi}$ is the built-in potential of the junction, and $V_R$ is the applied reverse bias voltage.

The technique requires the formation of a very shallow $n^+$ or $p^+$ diffusion over the surface of the diffused layer of interest. The doping in the heavily doped region must be more than 100 times that in the lightly doped region (where the profile is measured) for the technique to work. The capacitance is measured while applying a variable reverse dc voltage to the junction. The instantaneous value of $C_A$ is then calculated from Eq. 9.52. The depth at which $C_A$ is measured is obtained from the value of the applied reverse voltage, assuming the distance equals the depletion width of a one-sided junction. If those conditions hold, the depth of the measurement is given by:

$$
x = \sqrt{\frac{2\varepsilon_{Si}(V_{bi} + V_R)}{qC_A}} \quad (9.53)
$$

The major limitation of this technique is that the initial values of $C_A$ that are measured are at least as deep as the value of $x_j$ for the $n^+$ or $p^+$ diffusion plus the zero-bias depletion width (which extends into the substrate). The depth of the profiling is limited to the reverse breakdown voltage of the junction. Therefore, the C-V technique has little value for determining the doping profile in shallow junctions. The technique also fails short when abrupt profiles are measured, since fictitious tails are often observed near the steep edge.$^{56}$
9.9.3 Spreading Resistance Profiling (SRP)

The use of spreading resistance SR (i.e., the resistance associated with the divergence of the current lines emanating from a small-tipped electrical probe that is placed onto the silicon), was first proposed by Mazur and Dickey for measuring the thickness of diffused or epitaxial layers, and establishing the impurity profile for these structures.\(^{57}\) The technique is referred to as spreading resistance profiling (SRP). In this technique, the SR of a reproducibly formed point-contact is measured. This can be accomplished by using one, two, or three probe configurations. The two-probe method has met with the most success on commercially available equipment. SR measures only the electrically active dopant concentration, and other techniques, such as SIMS, are needed if the total (active + inactive) concentration profile is required.

To make spreading resistance measurements, a known current is applied between the probes, and the voltage drop is measured across these probes to obtain a spreading resistance \(R_{SR} = V/I\). Most of the voltage drop occurs at a distance that is only a few times the probe radius, and therefore the value of \(R_{SR}\) is measured in a very small volume of silicon immediately under the probe. In the two-probe method, the Si resistivity \(\rho\) is related to the \(R_{SR}\) by the relationship:

\[
\rho = 2 R_{SR} a
\]  \hspace{1cm} (9.54)

where \(a\) is an empirical quantity which is related to the effective electrical contact radius (it is not however the probe tip radius). The value of \(a\) is determined by measuring \(R_{SR}\) on a well-characterized material of known resistivity. An ASTM Standard has been developed for conducting SR measurements, and the reader is referred to this reference for details.\(^{58}\)

In order to use spreading resistance the probe tips must be well conditioned. This is achieved by stepping the probes at least 500 times on a silicon substrate that has been ground with fine grit. The probe tips must also be calibrated to obtain the empirical value of \(a\). This is best done on a wafer that has a well-documented resistivity (determined using a more direct technique). It is important that the surface of the calibration wafer and the surface of the unknown sample be identically prepared.

![Graph](image)

**Fig. 9-21** Spreading resistance profile measurement showing the capability of measuring very shallow junctions. Courtesy of Applied Materials
The main use of SR is to determine doping profiles of diffused or epitaxial layers. This is accomplished by angle lapping the wafer and then making SR measurements along the length of the lapped surface. Knowing the angle of the taper gives the depth as a function of distance from the surface. The same precautions used to lap a wafer for junction depth measurements (see next section) must be obeyed when beveling a specimen for SR measurements, including the use of the laser measurement to determine the angle. Elaborate correction techniques and correction factors (CF) have been developed to convert the values of $R_{SR}$ into carrier concentration values, particularly in multilayer samples. The values of the CFs depend on the nature of the layer and the proximity of the measurement to the junctions.

By beveling the sample to extremely low angles (<0.05 degrees), it is now possible to perform SRP on very shallow junctions. The skill required to accurately make these measurements is restricted to a few commercial laboratories.\textsuperscript{59} Fig. 9-21 shows a SR profile for a shallow junction boron diffusion, with an $x_J$ of only 30 nm. The most difficult part of the measurement is determining where the top surface ends and the bevel starts. Depositing a layer of polysilicon on the top surface can aid in finding where the bevel begins.

9.10 JUNCTION DEPTH MEASUREMENTS: PHYSICAL TECHNIQUES

9.10.1 Angle Lap and Stain

The angle-lap and stain method is effective for determining the depth of some $pn$ junctions, and the test method is available in ASTM Standard Test Method F-110-84.\textsuperscript{60} The method works well on relatively deep, highly-doped junctions, but is not accurate for shallow junctions. Wu et al. have reported that the accuracy of this technique can be significantly improved (to ±20 nm of the true metallurgical junction), if the proper lapping and staining precautions are taken.\textsuperscript{61}

The angle-lapping technique requires the grinding of a small angular bevel on the silicon wafer. This is accomplished by mounting the wafer on an angle block, and moving the assembly over a flat glass onto which a lapping-compound slurry has been poured. To obtain small angles (≤1°) of good quality, the following practices should be followed:

1. The angle-lap assembly should be massive, to eliminate rocking during the lapping process. Small pieces of silicon affixed to the bottom of the holder, where they act as runners, are found to be helpful in minimizing the rocking, as well.

2. A slurry consisting of the lapping compound (alumina of ≤0.3 μm particle size, mixed with water) is used. The slurry should be changed often to keep it free from dirt and Si chips.

3. Any metallization on the wafer surface should be removed prior to lapping, since it smears when lapped, and will interfere with the subsequent staining process.

4. If the region of interest does not have a reasonably thick (>300 nm) protective oxide or nitride layer, edge rounding can occur. This can make it difficult to locate the position of the surface with certainty. Thus, it is good practice to chemically deposit ~500 nm of low temperature silicon dioxide onto such surfaces prior to lapping.

After lapping has been performed (which brings the junction to the surface), a staining solution is used to delineate either the $n$-type area or the $p$-type area. One such stain consists of a copper sulfate ($CuSO_4$) solution for staining the $n$-region. The staining occurs as a result of the following events: a) a drop of $CuSO_4$ is dispensed onto the junction; b) at the same time the junction is illuminated by an intense light source which causes it to become forward-biased.
Fig. 9-22 Schematic diagram of the setup for small angle measurements with a laser.

c) this produces a voltage drop across the junction; and d) Cu atoms plate onto the n-side of the junction (Cu⁺⁺ → Cu⁺ + 2 holes). A different staining technique is used to stain p-regions. It employs a solution of 100 cc of HF and a few drops of nitric acid (HNO₃).

A large source of error in junction depth measurements arises from the uncertainty of the lapping angle on the sample. Since the sample is mounted to the block by wax, the angle on the wafer may not coincide exactly with that on the block. Therefore, a technique to determine the angle on the wafer after lapping (and staining) is required. Such a technique is provided by the laser reflection system shown schematically in Fig. 9-22. In this technique the laser beam is collimated, and then made to pass through a hole in a screen. The screen is typically ~140 inches from the sample. The sample is mounted such that reflections from the beveled edge and the surface are occurring at the same time. The sample orientation is then adjusted to achieve a condition such that reflections from these two surfaces are roughly equidistant from the primary beam. Under these conditions the angular magnification \( m \) from the lapped surface is given by:

\[
m = \left(\frac{l^2 + (s/4)^2}{s/2}\right)
\]

where \( l \) and \( s \) are defined in Fig. 9-22. If the beveled angle is small, so that \( l \gg s \)

\[
m \approx 2x \frac{l}{s}
\]  

(9.56)

Typically, \( l = 140 \) inches and \( s = 5 \pm 0.1 \) inches, so that \( m \) can be measured to 2% accuracy for angles about \( 1^\circ \). After obtaining the magnification factor \( m \), the junction depth is found from:

\[
x_j = x/m
\]

(9.57)

where \( x_j \) is the depth of the stained junction measured along the beveled surface. The value of \( x_j \) is typically determined from photographs taken at ~1000x magnification. The accuracy to which \( x_j \) can be measured on the photograph varies between \( 0.5-1 \mu \text{m} \). For a \( 1^\circ \) angle lap, this corresponds to an experimental uncertainty in \( x_j \) between 10–20 nm. The angle-lap and stain accuracy is also limited by the resolution of the staining procedure, which in some cases may not accurately delineate the metallurgical junction. Strong illumination helps to assure that staining provides an accurate indication of the junction edge.

Another factor influencing stained junctions is the wedge effect. For small angle laps (<\( 1^\circ \)), there is some question as whether enough dopant ions exist in the \( 1^\circ \) wedge to truly stain the junction, especially if the top layer is lightly doped. Strong illumination helps to uniformly plate the Cu on the n-side of the depletion region to within ~20 nm of the metallurgical junction.

### 9.10.2 Groove and Stain

A technique related to the angle-lap and stain method for measuring \( x_j \) is that of groove and stain. This technique has the advantage over angle lapping in that it rapidly exposes the junction. By rotating a wheel or ball impregnated with diamond grit (and ranging in diameter
from 0.75–1.5 in) against the wafer at a predetermined location, a groove will be cut in the silicon. If the groove is sufficiently deep, the junction will be exposed. After the groove is formed, the wafer is cleaned and the junction stained using the same methods discussed above.

The *arcuate trigonometric* method uses a groove of precise radius, $R$. The groove width at the surface of the wafer, $W_1$, is determined using a measuring microscope, as is the width at the junction, $W_2$ (see Fig. 9-23). The junction depth, $x_j$, is then calculated from the following trigonometric relationship:

$$x_j = A_2 - A_1$$  \hspace{1cm} (9.58)

or:

$$x_j = [R^2 - (W_1/2)^2]^{1/2} + [R^2 - (W_2/2)^2]^{1/2}$$  \hspace{1cm} (9.59)

The advantage of this technique is that the multiplying ratio of the groove's width-to-depth expands the layer thickness into the easy measuring range of an optical microscope. Commercial grooving instruments manufactured by Philtec Instruments Company are available.\textsuperscript{52} The Philtec measurement tool is advertised as being precise to ±1% for junction depths as shallow as 50 nm. The major sources of error in this method arise from the following assumptions: a) the groove is exactly cylindrical, and b) the edge of the groove can be accurately determined when viewed in the microscope.

### 9.10.3 Secondary Ion Mass Spectroscopy

*Secondary ion mass spectroscopy* (SIMS) provides a direct method for measuring the impurity profile of a diffused (or epitaxial) layer. The method is described in detail in the first edition of this volume. Briefly, this technique directly measures the total amount of impurity in the silicon matrix as the sample surface is slowly sputtered away. The technique has a depth resolution of 5–10 nm and it measures both electrically active and inactive impurity levels of all impurities present.

The SIMS profile is obtained by measuring the secondary ion yield of the sputtered atoms as a function of time. If it is assumed that the sputtering rate is linear, then the sputtering time can be directly related to the amount of material removed, and therefore, to the depth into the sample. To obtain quantitative data about impurities present either external standards or a self-standard are required. These are needed because the ion yield for a particular species depends on the matrix, as well as on the primary ion beam source and energy. The ion yields of the external standard are measured at the same time as the ion yields from the test sample, and a linear dependence between the ion yield and concentration is assumed to determine the unknown
concentration. The self-standard uses the ion yield of the silicon matrix to provide a standard against which the impurity is measured. This works quite well for high impurity concentrations (~10^{20} atoms/cm^3). The SIMS technique can also be used to profile impurities remaining in oxide masks, or to ascertain the impurity incorporated into an oxide during oxidation. By measuring the impurity concentration in the oxide (and in the silicon near the Si/SiO_2 interface), the segregation coefficient can be established. However, it should be repeated that the SIMS technique cannot differentiate between active and inactive dopants, but rather measures both. To obtain the active dopant profile a technique such as SRP is required.

In order to measure the profile of ultra-shallow junctions, it is necessary to accurately determine the dopant concentration near the silicon surface in the range of 2–100 nm. There is an issue associated with the SIMS method, called atomic mixing, that makes measurements close to the surface difficult. In atomic mixing the primary ion strikes a sample atom, knocking that atom off its lattice site. This atom can leave its site with enough energy to knock out other atoms which, in turn, knock out others. This is called the cascade effect. As a result of the cascade effect, dopant atoms that are knocked forward will no longer be in their correct position. The cascade problem can be ameliorated by using primary ion beams with low energy and at low angles of incidence to the substrate. The most common primary beam used for shallow junction measurements is Cs^+ at 2 keV energy, impacting the sample at a low angle of incidence (about 30°). Some advanced SIMS systems use the more sensitive quadrupole mass-analyzer to analyze the secondary ions, which improves the sensitivity of the measurement.

9.10.4 Two-Dimensional Depth Profiling

The use of SIMS to directly measure lateral doping under the gate is not possible since there are too few dopant atoms present to be detected by this technique. In addition, the spatial resolution of the SIMS technique is too coarse for small geometry devices. There are two other techniques available to make such measurements. The first uses cross-sectional high resolution transmission electron microscopy (HRTEM). After preparing a cross-sectional TEM sample, for the HRTEM technique, the sample is etched in a selective chemical etch (nitric acid and hydrofluoric acid) that has differing etch rates depending on the doping levels. A concentration gradient, then, would etch to different depths depending on the doping concentration. The different depths of etching reveal themselves with different contrast in a TEM micrograph. The

![Fig. 9-24 Cross-section HRTEM of an etched As implanted and diffused junction showing 2-D diffusion under the polysilicon gate.](image-url)
DIFFUSION IN SILICON  367

contrast level is calibrated to an unpatterned section of the wafer, where the doping level has been measured by SIMS. Figure 9-24 shows a TEM micrograph of the source/drain extension in a CMOS device near the gate structure. The lateral diffusion under the gate is evident. The main problem with this technique is that the contrast is lost at doping levels below $10^{19}/\text{cm}^3$.

The second technique uses scanned probe microscopy such as atomic force microscopy (AFM) or scanning capacitance microscopy (SCM). The fundamentals of these techniques are described in the literature. Such techniques also use selective etching of the junctions and calibration of the measured scanning signal against a standard. The standard is again obtained from SIMS analysis on an unpatterned section. The AFM technique is based on measuring differences in height in the etched region. The AFM technique has better sensitivity to lower doping concentrations but poorer spatial resolution than does the HRTEM technique. The scanning capacitance measurement uses an ultraline probe to measure the variation in capacitance in the lateral direction. The measured capacitance depends on the doping level which is also calibrated to a SIMS measurement on an unpatterned wafer. These techniques are in their early stage of development and work continues on their improvement.

REFERENCES

18. Avant/TMA and Silvaco, Inc.
19. D.J. Eaglesham et al., “(311) Defects in Ion Implanted Silicon: The Cause of Transient Diffusion
59. For example, Solid State Measurements, Inc and Solecon Laboratory.


**PROBLEMS**

1. Determine the solid solubility and the diffusion coefficient of: a) boron at 950°C; b) phosphorus at 1050°C; and c) arsenic at 1100°C.

2. During predeposition, what parameter controls the concentration of dopant at the surface of the wafer?

3. Predeposition is carried out on a silicon wafer at 975°C for 30 minutes in the presence of an excess of phosphorus. The substrate is boron doped to a concentration of $10^{17}$ atoms/cm$^3$: a) Find $C_S$ for this procedure; b) Calculate $\sqrt{Dt}$ for this procedure; c) Calculate the dopant concentration in the silicon at the following depths: $x = 0$, $x = 0.1 \mu$m, $x = 0.2 \mu$m, $x = 0.3 \mu$m, $x = 0.4 \mu$m, and $x = 0.5 \mu$m. Plot this concentration on log-linear graph paper; d) Determine the depth at which the metallurgical pn junction occurs; e) Determine the junction depth if a 2 Ω-cm p-type substrate is utilized; f) Find the total amount of dopant incorporated into the wafer per unit area.

4. Assume the wafer of Problem 3 is used in this problem. The objective will be to determine the doping profile as a function of depth after a drive-in step of 50 minutes at 1100°C has been carried out: a) Calculate $\sqrt{Dt}$; b) Find the final surface concentration, $C_S$; c) Find the values of the dopant concentration at the following depths: $x = 0$, $x = 0.5 \mu$m, $x = 1.0 \mu$m, $x = 1.5 \mu$m, and $x = 2.0 \mu$m. Plot this data on log-linear graph paper; d) Determine the depth at which the metallurgical pn junction occurs.

5. A diffusion furnace is operated at 950°C ± 1°C. Assuming a diffusion profile described by a Gaussian distribution, what is the corresponding spread in the diffusion lengths due to the uncertainty in the temperature?

6. Show that the logarithm of the sheet resistance after a predeposition is approximately a linear function of the reciprocal. This indicates an Arrhenius behavior (see Appendix 3).

7. Verify that equations 9-10 and 9-16 satisfy Fick's Second Law (Eq. 9-6) and the appropriate initial and boundary conditions.

8. The base region of a npn transistor is fabricated according to the following process sequence:
   
   Starting material: $p$-type, $10^{15}$ boron atoms/cm$^3$;  
   Oxidation: 80 minutes at 1200°C, wet oxygen;  
   Open base region window;  
   Predeposition of phosphorus, 30 minutes at 800°C;  
   Drive-in diffusion: 50 minutes at 1200°C, dry oxygen.

The measured V/I after the drive-in step is 1.4 Ω, and the junction depth is measured to be 4 μm deep. Calculate: 1) The oxide thickness over the base window and in the field regions; 2) The phosphorus dopant profile after the drive-in diffusion; 3) The total number of impurities per unit area of base.

9. When determining if the intrinsic diffusivity of an impurity is applicable at a given temperature, it is necessary to know the intrinsic carrier concentration, $n_i$. Using Eq. 9-19 (and the fact that $\sqrt{N_e N_h} = 3.87 \times 10^{16}$ T$^{3/2}$, and $E_g(T)$ is found using Eq. 9-20), plot the value of $n_i$ versus $T$ from $T = 600^\circ$C to $1200^\circ$C, making the calculation every 100°C.

10. To prevent wafer warpage that would occur due to a rapid decrease in temperature, the temperature in a diffusion furnace is linearly reduced from 1100°C to 600°C over a period of 30 minutes. Calculate the effective diffusion time at the initial diffusion temperature for an arsenic diffusion in silicon.
Week 6

Lithography
11

Lithography and Etching

Lithography is the process of transferring patterns of geometric shapes on a mask to a thin layer of radiation-sensitive material (called resist) covering the surface of a semiconductor wafer. These patterns define the various regions in an integrated circuit such as the implantation regions, the contact windows, and the bonding-pad areas. The resist patterns defined by the lithographic process are not permanent elements of the final device but only replicas of circuit features. To produce circuit features, these resist patterns must be transferred once more into the underlying layers comprising the device. The pattern transfer is accomplished by an etching process which selectively removes unmasked portions of a layer. In this chapter we consider the lithographic process and the various lithographic methods.\(^1\) We shall also consider both the wet chemical etching and dry etching techniques.\(^2\)\(^3\)

11.1 OPTICAL LITHOGRAPHY

The vast majority of lithographic equipment for integrated circuit (IC) fabrication is optical equipment using ultraviolet light (\(\lambda \approx 0.2\) to 0.4 \(\mu\)m). In this section we consider the exposure tools, the masks, and the resists used for optical lithography. We also consider the pattern transfer process which serves as a basis for other lithographic systems. We shall first briefly consider the clean room, because all lithographic processes must be performed in an ultraclean environment.

11.1.1 The Clean Room

An IC fabrication facility requires a clean processing room, especially in the area used for lithography. The need for such a clean room arises because dust particles in the air can settle on semiconductor wafers and lithographic masks and can cause defects in the devices that result in circuit failure. For example, a dust particle on a semiconductor surface can disrupt the single-crystal growth of an epitaxial film, causing the formation of dislocations. A dust particle incorporated into the gate oxide can result in enhanced conductivity and cause device failure due to low breakdown voltage. The situation is even more critical in the lithographic area. When dust particles adhere to the surface of a photomask, they behave as opaque patterns on the mask, and these patterns

will be transferred to the underlying photoresist on the mask. Figure 1 shows three different situations that can occur. Particle 1, located near a pattern edge and not over a metal runner, will result in the formation of a pinhole defect. Particle 2, located near a pattern edge and over a metal runner, will result in a short circuit. Particle 3 can lead to missing regions and render the circuit useless.

In a clean room, the total number of particles is tightly controlled along with the distribution of the particle-size distribution curve. An ISO 100 clean room is one which has a particle count of 10 particles of 0.5 \(\mu\)m or larger per cubic meter of air. In the ISO system, the particle size decreases, a more stringent cleanroom is required when the minimum feature size decreases. For most IC fabrication applications, the minimum feature size is 0.3-0.4 \(\mu\)m. Since the minimum feature size decreases, a more stringent cleanroom is required when the minimum feature size decreases. For most IC fabrication applications, the minimum feature size is 0.3-0.4 \(\mu\)m. Since the minimum feature size decreases, a more stringent cleanroom is required when the minimum feature size decreases.

Problems

1. If we expose a 125-mm wafer for 10 seconds under a certain condition at 30 m/min, how many particles are removed by the class 10 clean room?
11.1 Optical Lithography

will be transferred to the underlying layers along with the circuit patterns on the mask. Figure 1 shows three dust particles on a photomask. Particle 1 may result in the formation of a pinhole in the underlying layer. Particle 2 is located near a pattern edge and may cause a constriction of current flow in a metal runner. Particle 3 can lead to a short circuit between the two conducting regions and render the circuit useless.

In a clean room, the total number of dust particles per unit volume must be tightly controlled along with the temperature and humidity. Figure 2 shows the particle-size distribution curves for various classes of clean rooms. A class 100 clean room is one which has a dust count of 100 particles (with particles diameters of 0.5 μm or larger) per cubic foot. This corresponds to 3500 particles (with particle sizes 0.5 μm or larger) per cubic meter, as indicated in parentheses in Fig. 2. Since the number of dust particles increases as particle size decreases, a more stringent control of the clean room environment is required when the minimum feature lengths of ICs are reduced to the 1-μm range. For most IC fabrication areas, a class 100 clean room is required, that is, the dust count must be about four orders of magnitude lower than that of ordinary room air. However, for the lithographic area, a class 10 clean room or one with a lower dust count is required.

Problem

If we expose a 125-mm wafer for 1 min to an air stream under a laminar-flow condition at 30 m/min, how many dust particles will land on the wafer in a class 10 clean room?

![Fig. 1 Various ways in which dust particles can interfere with photomask patterns.](image-url)
11.1 Optical Lithography

resist film on a semiconductor wafer, patterned on successive masks using previously defined patterns on the wafers that can be exposed per hour.

There are basically two optical projection printing.沙滩 print contact with one another as in proximity printing. Figure 3a shows a resist-coated wafer is brought into contact with the mask for a fixed time. The interfacial adhesion caused by dust on the wafer can be imaged by the mask. The imaged mask and results in defects in the wafer.

To minimize damage, Figure 3b shows the basic setup. except that there is a small gap, 10 μm, during exposure. The small gap feature edges on the photomask (opaque mask feature, fringes are in the shadow region) and the resolution.

In shadow printing, the minimum feature size $L$ is given by

$$L = \frac{\lambda}{2n \sin \theta}$$

where $\lambda$ is the wavelength of the exposure light, $n$ is the refractive index of the mask and the wafer and $\theta$ is the angle at which the light hits the wafer.

11.1.2 Exposure Methods

The pattern transfer process is accomplished by using a lithographic exposure tool. The performance of an exposure tool is determined by three parameters: resolution, registration, and throughput. The resolution is the minimum feature dimension that can be transferred with high fidelity to a

Fig. 2 Particle-size distribution curve.

Solution

For a class 10 clean room, there are 350 particles (0.5 μm or larger) per cubic meter. The air volume that goes over the wafer in 1 min is

$$\left(30 \text{ m/min}\right) \times \frac{0.125 \text{ m}^2}{2} \times 1 \text{ min} = 0.368 \text{ m}^3.$$  

The number of dust particles (0.5 μm or larger) contained in the air volume is

$$350 \times 0.368 = 128 \text{ particles}.$$  

Therefore, if there are 200 IC chips on the wafer, the particle count amounts to one particle on each of 64% of the chips. Fortunately, only a fraction of the particles that land adhere to the wafer surface, and of those only a fraction are at a critical location critical enough to cause a failure. However, the calculation indicates the importance of the clean room.

Fig. 3 Schematics of optical shadow (a) and proximity printing (b).
11.1 Optical Lithography

resist film on a semiconductor wafer; registration is a measure of how accurately patterns on successive masks can be aligned (or overlaid) with respect to previously defined patterns on the wafer; and throughput is the number of wafers that can be exposed per hour for a given mask level.

There are basically two optical exposure methods: shadow printing and projection printing. Shadow printing may have the mask and wafer in direct contact with one another as in contact printing, or in close proximity as in proximity printing. Figure 3a shows a basic setup for contact printing where a resist-coated wafer is brought into physical contact with a mask, and the resist is exposed by a nearly collimated beam of ultraviolet light through the back of the mask for a fixed time. The intimate contact between resists and mask provides very high resolution ($\sim 1 \mu m$). However, contact printing suffers a major drawback caused by dust particles. A dust particle or a speck of silicon dust on the wafer can be imbedded into the mask when the mask makes contact with the wafer. The imbedded particle causes permanent damage to the mask and results in defects in the wafer with each succeeding exposure.

To minimize mask damage, the proximity exposure method is used. Figure 3b shows the basic setup. It is similar to the contact printing method except that there is a small gap, 10 to 50 $\mu m$, between the wafer and the mask during exposure. The small gap however results in optical diffraction at feature edges on the photomask (i.e., when light passes by the edges of an opaque mask feature, fringes are formed, and some light penetrates into the shadow region) and the resolution is degraded to the 2- to 5-$\mu m$ range.

In shadow printing, the minimum linewidth that can be printed is roughly

$$l_m \approx \sqrt{\lambda g}$$

when $\lambda$ is the wavelength of the exposure radiation and $g$ is the gap between the mask and the wafer and includes the thickness of the resist. For
\[ \lambda = 0.4 \, \mu m \text{ and } g = 50 \, \mu m, \text{ the minimum linewidth is } 4.5 \, \mu m. \] 
If we reduce \( \lambda \) to 0.25 \( \mu m \) (wavelength of 0.2 to 0.3 \( \mu m \) are the deep-UV spectral region) and \( g \) to 15 \( \mu m \), \( l_m \) becomes 2 \( \mu m \). Thus, there is an advantage in reducing both \( \lambda \) and \( g \). However, for a given distance \( g \), any dust particle with a diameter larger than \( g \) potentially can cause mask damage.

To avoid the mask damage problem associated with shadow printing, projection printing exposure tools have been developed to project an image of the mask patterns onto a resist-coated wafer many centimeters away from the mask. To increase resolution, only a small portion of the mask is exposed at a time. The small image area is scanned or stepped over the wafer to cover the entire wafer surface. Figure 4a shows a 1:1 wafer scan projection system.\(^6,7\) A narrow, arc-shaped image field \( \sim 1 \, mm \) in width serially transfers the slit image of the mask onto the wafer. The image size on the wafer is the same as that on the mask. This scanning concept can be extended to two dimensions whereby a small symmetrically shaped image field is scanned in an overlapping raster fashion, as shown in Fig. 4b. These techniques are called scanning projection. The small image field can also be stepped over the surface of the wafer by two-dimensional translation of the wafer, which remains stationary. After the exposure at the next chip site and the processing, the wafer is translated in both directions. The magnification of the wafer image is determined by the demagnification ratio \( M:1 \) (e.g., \( \times 5, \times 1 \)) in the projection system. The 1:1 optical systems are easier to design and construct than reduction systems, but it is much easier to achieve a 1:1 than it is at a 10:1 or a 5:1 demagnification.

The resolution of a projection technique not only depends on the wavelength of light \( \lambda \), which is given by

\[ N = \frac{\lambda}{2n} \]

where \( n \) is the index of refraction of the resist (\( n = 1 \)), and \( \theta \) is the half-angle of the light at the wafer as shown \(^5\) in Fig. 4c. The depth of focus, \( \Delta z \), which can be expressed as

\[ \Delta z = \frac{\pm l_m/2}{\tan \theta} \]

Equation 2 indicates that resolution improves as either reducing the wavelength of light \( \lambda \) or as \( \tan \theta \) increases, which indicates that the depth of focus is very dependent on \( NA \) than by decreasing \( \lambda \). This effect may be used in optical lithography. Typically, a system with a 1.5-\( \mu m \) resolution, while step-and-repeat systems can achieve 0.1-\( \mu m \) or better resolution.

\[ \Delta z = \frac{\pm l_m/2}{\tan \theta} \]

Fig. 5 Shaded drawing of the relationship of the mask and wafer.
Lithography and Etching

A linewidth is 4.5 μm. If we reduce the deep-UV spectral region) there is an advantage in reducing, e.g., any dust particle with a diametrical loose film is a problem. However, with shadow printing, projected to project an image of the many centimeters away from the portion of the mask is exposed at a stepped over the wafer to cover the wafer scan projection system.\(^6\) A in width serially transfers the slit edge size on the wafer is the same as can be extended to two dimensions. A planer field is scanned in an overlap-lapping techniques are called scanning be stepped over the surface of the wafer by two-dimensional translations of the wafer only, while the mark remains stationary. After the exposure of one chip site, the wafer is moved to the next chip site and the process is repeated. Figures 4c and 4d show the partitioning of the wafer image by step-and-repeat projection with the demagnification ratio M:1 (e.g., 10:1 for a 10 times reduction on the wafer) or at 1:1, respectively. The demagnification ratio is an important factor in our ability to produce both the lens and the mask from which we wish to print. The 1:1 optical systems are easier to design and fabricate than a 10:1 or a 5:1 reduction systems, but it is much more difficult to produce defect-free masks at 1:1 than it is at a 10:1 or a 5:1 demagnification ratio.

The resolution of a projection system is given by

\[ l_m \approx \frac{\lambda}{NA} \]  

where \( \lambda \) is again the exposure wavelength and \( NA \) is the numerical aperture, which is given by

\[ NA = \bar{n} \sin \theta \]  

with \( \bar{n} \) the index of refraction in the image medium (usually air, where \( \bar{n} = 1 \)), and \( \theta \) is the half-angle of the cone of light converging to a point image at the wafer as shown\(^5\) in Fig. 5. Also shown in the figure is the depth of focus, \( \Delta z \), which can be expressed as

\[ \Delta z = \frac{\pm l_m/2}{\tan \theta} \approx \frac{\pm l_m/2}{\sin \theta} = \pm \frac{\bar{n} \lambda}{2(NA)^2}. \]  

Equation 2 indicates that resolution can be improved (i.e., smaller \( l_m \)) by either reducing the wavelength or increasing \( NA \) or both. However, Eq. 4 indicates that the depth of focus degrades much more rapidly by increasing \( NA \) than by decreasing \( \lambda \). This explains the trend toward shorter wavelengths in optical lithography. Typically, scanning projection systems can achieve 1.5-μm resolution, while step-and-repeat projection systems are capable of 1-μm or better resolution.

11.1 Optical Lithography

![Fig. 5 Simple image system.\(^5\)](image-url)
11.1.3 Masks

For discrete devices or small-scale (up to 100 components/chip, SSI) to medium-scale (up to 1000 components/chip, MSI) integrated circuits, the first step in the generation of patterns is to draw a large composite layout of the mask set. This is typically $100 \times$ (i.e., 100 times) to $2000 \times$ the final size. The composite layout is broken into mask levels that correspond to the IC process sequence such as the isolation region on one level, the gate electrode on another, and so on. Artwork is drawn for each masking level. The artwork is reduced to a $10 \times$ glass reticle by using a reduction camera. The final mask is made from the $10 \times$ reticle using a system similar to the 10:1 step-and-repeat projection printing system described previously. All chip sites on the mask are exposed with identical patterns in the form of a matrix, and each of the sites contains a complete device or circuit pattern for that mask level. Figure 6 shows a mask on which patterns of geometric shapes have been formed. A few secondary-chip sites are also included in the mask; these sites are used for process evaluation.

For large-scale ($10^3$ to $10^5$ components/chip, LSI) or very-large-scale ($>10^5$ components/chip, VLSI) integrated circuits, this approach is not practical because of the large amount of time needed and the unavoidable human error involved in generating such artwork. Instead, we use computer-aided design (CAD) systems in which designers can completely describe the circuit layout electrically. The digital data produced by the CAD system then drives a pattern generator (e.g., an electron beam machine to be considered in Section 11.2) that transfers the patterns directly to photosensitized masks.

**Fig. 6** A glass IC photomask.¹

11.1 Optical Lithography

For feature sizes of the order of micrometers, plates covered with a soft-surface photoresist are used. For feature sizes, masks are made from thin metal films, such as chromium or iron oxide. The mask defect density. Mask defects can be created during manufacturing or during subsequent lithographic processing. The defect density has a profound effect on the yield of good chips per wafer. The yield $Y$ can be calculated as a first-order approximation, the yield $Y$ is expressed as

$$Y = \frac{D}{N}$$

where $D$ is the average number of defects per unit area of an IC chip. If $D$ remains constant, then the final yield $Y$ is

$$Y = \frac{1}{N}$$

Figure 7 shows the mask-limited yield as a function of chip size for various values of $N = 10$ levels.
11.1 Optical Lithography

For feature sizes of the order of 5 \( \mu \text{m} \) or larger, masks are made from glass plates covered with a soft-surface material such as an emulsion. For smaller feature sizes, masks are made from glass covered with hard-surface materials such as chromium or iron oxide. One of the major concerns on masks is the defect density. Mask defects can be introduced during the manufacture of the mask or during subsequent lithographic processes. Even a small mask defect density has a profound effect on the final IC yield. The yield is defined as the ratio of good chips per wafer to the total number of chips per wafer. As a first-order approximation, the yield \( Y \) for a given masking level can be expressed as

\[ Y \approx e^{-DA} \] \hspace{2cm} (5)

where \( D \) is the average number of "fatal" defects per unit area and \( A \) is the area of an IC chip. If \( D \) remains the same for all mask levels (e.g., \( N = 10 \) levels), then the final yield becomes

\[ Y \approx e^{-NDA} \] \hspace{2cm} (6)

Figure 7 shows the mask-limited yield for a 10-level lithographic process as a function of chip size for various values of defect densities. For example, for

![Fig. 7](#) Yield for a 10-mask lithographic process with various defect densities per level.
\( D = 0.5 \text{ defect/cm}^2 \), the yield is 22% for a chip size of 30 mm\(^2\), and it drops to about 1% for a larger chip of 90 mm\(^2\). Therefore, inspection and cleaning of masks are important to achieve high yields on large chips. Of course, an ultraclean processing area is mandatory for lithographic processing.

### 11.1.4 Photoresist

The photore sist is a radiation-sensitive compound. Photore sist s can be classified as positive and negative, depending on how they respond to radiation. For positive resists, the exposed regions become more soluble and thus more easily removed in the development process. The net result is that the patterns formed (also called images) in the positive resist are the same as those on the mask. For negative resists, the exposed regions become less soluble, and the patterns formed in the negative resist are the reverse of the mask patterns.

Positive photore sist s consist of three components: a photosensitive compound, a base resin, and an organic solvent. Prior to exposure, the photosensitive compound is insoluble in the developer solution. After exposure, the photosensitive compound absorbs radiation in the exposed pattern areas, changes its chemical structure, and becomes soluble in the developer solution. Upon development, the exposed areas are removed.

Negative photore sist s are polymers combined with a photosensitive compound. After exposure, the photosensitive compound absorbs the optical energy and converts it into chemical energy to initiate a chain reaction. This reaction causes crosslinking of the polymer molecules. The crosslinked polymer has a higher molecular weight and becomes insoluble in the developer solution. Upon development, the unexposed areas are removed. One major drawback of a negative photore sist is that in the development process the whole resist mass swells by absorbing developer solvent. This swelling action limits the resolution of negative photore sist s.

Figure 8a shows a typical exposure response curve and image cross section for a positive resist. The response curve describes the percent resist remaining after exposure and development versus the exposure energy. Note that the resist has a finite solubility in its developer, even without exposure to radiation. As the exposure energy increases, the solubility gradually increases until, at a threshold energy \( E_T \), the resist becomes completely soluble. The sensitivity of a positive resist is defined as the energy required to produce complete solubility in the exposed region. Thus, \( E_T \) corresponds to the solubility sensitivity. In addition to \( E_T \), a parameter \( \gamma \), the contrast ratio, is defined to characterize the resist:

\[
\gamma = \left[ \ln \left( \frac{E_T}{E_1} \right) \right]^{-1}
\]

where \( E_1 \) is the energy obtained by drawing the tangent at \( E_T \) to reach 100% resist thickness as shown in Fig. 8a. A larger \( \gamma \) implies a more rapid solubility

of the resist with an incremental increase to sharper images.

The image cross section in Fig. 8a shows the edges of a photomask image and the image after development. The edges of the optically projected positions of the mask resist image corresponds to the processing energy equals the threshold energy.

Figure 8b shows the exposure response for a negative resist. The negative resist uses a developer solution for exposure energies. Above \( E_T \), more of the resist film is soluble twice the threshold energy and is completely soluble in the developer. The sensitivity of the exposed region is required to retain 50% of the positive resist.

### Problem

Find the parameter \( \gamma \) for the photore sist.
Lithography and Etching

11.1 Optical Lithography

![Graph of Resist Response]

Fig. 8 Exposure response curve and cross section of the resist image after development. (a) Positive photoresist. (b) Negative photoresist.

of the resist with an incremental increase of exposure energy and results in sharper images.

The image cross section in Fig. 8a illustrates the relationship between the edges of a photomask image and the corresponding edges of the resist images after development. The edges of the resist image are generally not at the vertically projected positions of the mask edges due to diffraction. The edge of the resist image corresponds to the position where the total absorbed optical energy equals the threshold energy $E_T$.

Figure 8b shows the exposure response curve and image cross section for a negative resist. The negative resist remains completely soluble in the developer solution for exposure energies lower than the threshold energy $E_T$. Above $E_T$, more of the resist film remains after development. At exposure energies twice the threshold energy, the resist film becomes essentially insoluble in the developer. The sensitivity of a negative resist is defined as the energy required to retain 50% of the original resist film thickness in the exposed region. The parameter $\gamma$ is defined similarly as in Eq. 7 except that $E_1$ and $E_T$ are interchanged. The image cross section for the negative resist (Fig. 8b) is also influenced by the diffraction effect.

**Problem**

Find the parameter $\gamma$ for the photoresists shown in Fig. 8.
For the negative resist (Kodak 747), we have $E_T = 7 \text{ mJ/cm}^2$ and $E_I = 12 \text{ mJ/cm}^2$:

$$\gamma = \ln \left( \frac{E_I}{E_T} \right)^{-1} = \ln \left( \frac{12}{7} \right)^{-1} = 1.9.$$  

For the positive resist (AZ 1350J), $E_T = 90 \text{ mJ/cm}^2$ and $E_I = 45 \text{ mJ/cm}^2$:

$$\gamma = \ln \left( \frac{E_T}{E_I} \right)^{-1} = \ln \left( \frac{90}{45} \right)^{-1} = 1.4.$$  

Table 1 lists a few commercially available negative and positive photoresists. (The resists for electron beam and X-ray lithographies are also listed.) Note that although they have comparable $\gamma$ values, the negative resist has much lower threshold energy. Therefore, for an optical lithographic system with a given exposure power (in mW/cm$^2$), much shorter times are needed to expose a wafer using a negative resist, permitting a larger wafer exposure throughput per hour. However, as mentioned before, negative resists usually swell after development, resulting in poor resolution. Positive resists, on the other hand, do not swell so that better resolution can be obtained; but they require much larger exposure energy and consequently longer exposure time, resulting in lower throughput.

To improve resolution of a resist image, thinner resist films are generally preferred. However, because of the topography of the underlying layers in ICs, thicker resist films are required to cover the steps, thus degrading the resolution. The multilayer resist systems were developed to separate the functions of image formation and step coverage. Figure 9 shows a trilayer resist system. 

A very thick resist layer 2 to 3 μm thick is applied to planarize the topography. A thin layer of silicon dioxide is then deposited to a thickness of 1000 Å.

<table>
<thead>
<tr>
<th>Lithography</th>
<th>Name</th>
<th>Type</th>
<th>Sensitivity $E_I$</th>
<th>$\gamma$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Optical</td>
<td>Kodak 747</td>
<td>Negative</td>
<td>9 mJ/cm$^2$</td>
<td>1.9</td>
</tr>
<tr>
<td></td>
<td>AZ-1350J</td>
<td>Positive</td>
<td>90 mJ/cm$^2$</td>
<td>1.4</td>
</tr>
<tr>
<td></td>
<td>PR102</td>
<td>Positive</td>
<td>140 mJ/cm$^2$</td>
<td>1.9</td>
</tr>
<tr>
<td>e-Beam</td>
<td>COP</td>
<td>Negative</td>
<td>0.3 μC/cm$^2$</td>
<td>0.45</td>
</tr>
<tr>
<td></td>
<td>GeSe</td>
<td>Negative</td>
<td>80 μC/cm$^2$</td>
<td>3.5</td>
</tr>
<tr>
<td></td>
<td>PBS</td>
<td>Positive</td>
<td>1 μC/cm$^2$</td>
<td>0.35</td>
</tr>
<tr>
<td></td>
<td>PMMA</td>
<td>Positive</td>
<td>50 μC/cm$^2$</td>
<td>1.0</td>
</tr>
<tr>
<td>X-Ray</td>
<td>COP</td>
<td>Negative</td>
<td>175 mJ/cm$^2$</td>
<td>0.45</td>
</tr>
<tr>
<td></td>
<td>DCOPA</td>
<td>Negative</td>
<td>10 mJ/cm$^2$</td>
<td>0.65</td>
</tr>
<tr>
<td></td>
<td>PBS</td>
<td>Positive</td>
<td>95 mJ/cm$^2$</td>
<td>0.5</td>
</tr>
<tr>
<td></td>
<td>PMMA</td>
<td>Positive</td>
<td>1000 mJ/cm$^2$</td>
<td>1.0</td>
</tr>
</tbody>
</table>

Finally, a thin resist layer (~ 400 Å) planarization and the opacity (not shown) of a positive resist layer is optimized for high contrast. When Ge$_2$Se$_3$ light, lateral diffusion of silver of this thickness to give a very large contrast. In practice, about 0.5 μm have been printed in Ge$_2$Se$_3$ exposure equipment.

### 11.1.5 Pattern Transfer

Figure 10 illustrates the steps of transferring a pattern etched in a semiconductor wafer that has been thermally grown SiO$_2$ on Si with a mask placed in a clean room which transfers the image to the wafer. The wafers are then held on a vacuum spindle, and ~ 11 of wafer. The wafer is then rapidly...
Lithography and Etching

We have $E_T = 7 \text{ mJ/cm}^2$ and

$$\frac{18}{7} = 2.6 \quad \text{and} \quad \frac{45}{20} = 2.3.$$

Table negative and positive photo and X-ray lithographies are also stable $\gamma$ values, the negative resist (e.g., for an optical lithographic system $\gamma$/cm$^2$), much shorter times are resist, permitting a large wafer as mentioned before, negative etching in poor resolution. Positive so that better resolution can be more energy and consequently throughput.

Thinner resist films are generally aply of the underlying layers in the steps, thus degrading the resolution. Developed to separate the functions Fig. 4 shows a trilayer resist system. Applied to planarize the topography, resist to a thickness of 1000 Å.

<table>
<thead>
<tr>
<th>Sensitivity</th>
<th>$\gamma$</th>
</tr>
</thead>
<tbody>
<tr>
<td>9 mJ/cm$^2$</td>
<td>1.9</td>
</tr>
<tr>
<td>90 mJ/cm$^2$</td>
<td>1.4</td>
</tr>
<tr>
<td>140 mJ/cm$^2$</td>
<td>1.9</td>
</tr>
<tr>
<td>0.3 μC/cm$^2$</td>
<td>0.45</td>
</tr>
<tr>
<td>80 μC/cm$^2$</td>
<td>3.5</td>
</tr>
<tr>
<td>1 μC/cm$^2$</td>
<td>0.35</td>
</tr>
<tr>
<td>50 μC/cm$^2$</td>
<td>1.0</td>
</tr>
<tr>
<td>175 mJ/cm$^2$</td>
<td>0.45</td>
</tr>
<tr>
<td>10 mJ/cm$^2$</td>
<td>0.65</td>
</tr>
<tr>
<td>95 mJ/cm$^2$</td>
<td>0.5</td>
</tr>
<tr>
<td>1000 mJ/cm$^2$</td>
<td>1.0</td>
</tr>
</tbody>
</table>

11.1 Optical Lithography

Finally, a thin resist layer ($\sim 4000 \text{ Å}$) is applied (Fig. 9a). Because of the planarization and the opacity (nonreflecting) of the lower layer, the top photosist layer is optimized for high-resolution imaging. After the top photoresist is developed, it serves as a mask to etch the thin silicon dioxide layer (Fig. 9b), which in turn serves as a mask for etching of the thick bottom resist layer (Fig. 9c).

Inorganic materials such as germanium selenide (Ge$_x$Se$_{1-x}$) can be used as a negative resist. When Ge$_x$Se$_{1-x}$ film is coated with silver and exposed to light, lateral diffusion of silver occurs which results in an edge-sharpening effect to give a very large contrast ratio ($\gamma = 3.5$). Lines and spaces as small as 0.5 μm have been printed in Ge$_x$Se$_{1-x}$ resist using step-and-repeat optical-exposure equipment.

11.1.5 Pattern Transfer

Figure 10 illustrates the steps of transfer of IC patterns from a mask to a semiconductor wafer that has an insulating layer formed on its surface (e.g., thermally grown SiO$_2$ on Si with a thickness of 0.1 to 1 μm). The wafer is placed in a clean room which typically is illuminated with yellow light, since photoresists are not sensitive to wavelengths greater than 0.5 μm. The wafer is held on a vacuum spindle, and ~ 1 cm$^3$ of liquid resist is applied to the center of wafer. The wafer is then rapidly accelerated up to a constant rotational
11.1 Optical Lithography

dissolved in the developer, as shown. Photoresist development is usually done in a developer solution. The wafer is then rinsed in deionized water (at \( \sim 100 \) to \( 180^\circ \)C) before being returned to the substrate. The wafer is then baked in an oven (e.g., using solvent, or equivalent), and peeling or lifting (or pattern) is done with a tool (e.g., Fig. 10e).

For the negative photoresist, the same process is followed, except that the unexposed areas are not protected and the other side of Fig. 10e is the reverse of the positive resist.

The insulator image can be used to form the pattern, for example, an ion implantation can be used in the oxide region, but not the area covered by the insulator. A duplicate of the design pattern on the mask (for a layout) or its complementary pattern (for an etch) is fabricated by aligning the next mask on the wafer and repeating the lithographic transfer process. This process requires 5 to 10 separate masks and steps.

A related pattern transfer process is lift-off (see Fig. 11). A positive resist is used to form the pattern (Fig. 11a and 11b). The film (e.g., aluminum) is then deposited on the substrate (Fig. 11c); the film thickness is controlled by the resist thickness. Those portions of the film on the wafer surface that are not protected by the resist are etched away.
11.1 Optical Lithography
dissolved in the developer, as shown in the left side of Fig. 10c. The photoresist development is usually done by spraying the wafer with the developer solution. The wafer is then rinsed and dried. After development, a postbaking (at \( \sim 100 \) to \( 180^\circ\mathrm{C} \)) may be required to increase the adhesion of the resist to the substrate. The wafer is then put in an ambient that etches the exposed insulating layer but does not attack the resist (e.g., buffered hydrofluoric acid is a typical SiO\(_2\) etchant), as shown in Fig. 10d. Finally, the resist is stripped (e.g., using solvent, or plasma oxidation), leaving behind an insulator image (or pattern) that is the same as the opaque image on the mask (left side of Fig. 10e).

For the negative photoresist, the procedures described are also applicable, except that the unexposed areas are removed. The final insulator image (right side of Fig. 10e) is the reverse of the opaque image on the mask.

The insulator image can be used as a mask for subsequent processing. For example, an ion implantation can be done to dope the exposed semiconductor region, but not the area covered by the insulator. The dopant pattern is a duplicate of the design pattern on the photomask (for a negative photoresist) or its complementary pattern (for a positive photoresist). The complete circuit is fabricated by aligning the next mask in the sequence to the previous pattern and repeating the lithographic transfer process. Typically, to fabricate an IC requires 5 to 10 separate masks and lithographic transfer steps.

A related pattern transfer process is the lift-off technique, shown in Fig. 11. A positive resist is used to form the resist pattern on the substrate (Fig. 11a and 11b). The film (e.g., aluminum) is deposited over the resist and the substrate (Fig. 11c); the film thickness must be smaller than that of the resist. Those portions of the film on the resist are removed by selectively dissolving

\[
(8) \quad \text{10,000 rpm to give a uniform film } \theta_a \text{. This procedure is also applicable to considerably more processing steps.}
\]

In a pre-exposure baking (typically the photoresist film and to improve alignment with respect to the mask in is exposed to UV light as shown in photoresist. The exposed resist is

![Fig. 11 Lift-off process for pattern transfer.](Image)
the resist layer in an appropriate liquid etchant so that the overlying film is lifted off and removed (Fig. 11d). The lift-off technique is capable of high resolution and is used extensively for discrete devices (e.g., high-power MESFET). However, it is not as widely applicable for very-large-scale integration, in which dry etching is the preferred technique.

11.2 ELECTRON BEAM, X-RAY, AND ION BEAM LITHOGRAPHIES

Various types of advanced lithographies for IC fabrication are shown\textsuperscript{11} in Fig. 12. As we have discussed in Section 11.1, the majority of IC exposure tools are optical systems using ultraviolet light (Fig. 12a). Optical exposure tools are capable of approximately 1-µm resolution, 0.5-µm registration, and a throughput of 50 to 100 wafers per hour. Electron beam lithography, shown in Fig. 12b, is primarily used to produce photomasks; relatively few are dedicated to direct wafer exposure (i.e., exposure of the resist directly by focused electron beam without a mask). Because of electron backscattering, electron beam exposure systems are limited to a practical minimum feature length of about 0.5 µm with 0.2 µm registration. X-Ray lithography, shown in Fig. 12c, has 0.5-µm or better resolution and 0.5-µm registration. However, it requires a complicated mask and is not yet used to produce ICs in volume. Ion beam lithography, shown in Fig. 12d, offers patterned-doping capability and very high resolution (\(\sim 100\) Å); this method is still in its initial development stage.

\begin{figure}[h]
\centering
\includegraphics[width=0.7\textwidth]{fig12.png}
\caption{Types of advanced lithographic methods. (a) Optical lithography. (b) Electron beam lithography. (c) X-ray lithography. (d) Ion beam lithography.\textsuperscript{11}}
\end{figure}

11.2.1 Electron Beam Lithography

Figure 13 shows a schematic of an electron beam lithography system. Electron gun is a device that can generate electron beam with a high current density. A tungsten thermionic cathode (LaB\textsubscript{6}) can be used to produce an electron beam that is focused by magnetic lenses or magnetic coils. The electron beam is then directed to the substrate using a deflection coils and a computer-controlled scanner. The target is then scanned over a predetermined area, and the exposed area then develops the resist layer. Because the beam size is much smaller than the substrate diameter, a precise pattern can be printed onto the substrate.

The advantages of electron beam lithography are: micromachining and submicron resist geometric control, greater depth of focus, and direct patterning of the substrate. The disadvantage is that electron beam lithography is slow. The throughput is approximately five wafers per hour. This throughput is adequate for the small laboratory runs that require small numbers of custom ICs, however, for maskless direct writing, the throughput could be increased significantly in future designs.

\begin{figure}[h]
\centering
\includegraphics[width=0.7\textwidth]{fig13.png}
\caption{Schematic of an electron beam lithography system.}
\end{figure}
11.2 Electron Beam, X-Ray, and Ion Beam Lithographies

11.2.1 Electron Beam Lithography

Figure 13 shows a schematic of an electron beam lithography system. The electron gun is a device that can generate a beam of electrons with a suitable current density; a tungsten thermionic-emission cathode or single-crystal lanthanum boride (LaB₆) can be used for the electron gun. Condenser lenses are used to focus the electron beam to a spot size of 0.01 to 0.1 μm in diameter. Beam-blanking plates (to turn the electron beam on and off) and beam deflection coils are computer-controlled and operated at high rates (in MHz or higher rates) to direct the focused electron beam to any location in the scan field on the substrate. Because the scan field (typically 1 cm) is much smaller than the substrate diameter, a precision mechanical stage is used to position the substrate to be patterned.

The advantages of electron beam lithography include the generation of micron and submicron resist geometries, highly automated and precisely controlled operation, greater depth of focus than that available from optical lithography, and direct patterning on a semiconductor wafer without using a mask. The disadvantage is that electron beam lithographic machines have low throughput—approximately five wafers per hour at less than 1-μm resolution. This throughput is adequate for the production of photomasks, in situations that require small numbers of custom circuits, or for design verification. However, for maskless direct writing, the machine must have the highest possible...
throughput and therefore the largest beam diameter possible consistent with the minimum device dimensions.

There are basically two ways to scan the focused electron beam: raster scan and vector scan. In a raster scan system, the patterns are written by a beam that moves through a regular pattern (vertically oriented, as shown in Fig. 14a). The beam scans sequentially over every possible location on the mask and is blanked (turned off) where no exposure is required. All patterns on the area to be written must be subdivided into individual address, and a given pattern must have as a minimum increment an interval that is evenly divisible by the beam address size.

In the vector scan system as shown in Fig. 14b, the beam is directed only to the requested pattern features and jumps from feature to feature, rather than scanning the whole chip, as in raster scan. For many chips, the average exposed region is only 20% of the chip area, so we can save time by using a vector scan system. To further increase the writing speed, a variable-beam-size system was developed in which the patterns beam has a rectangular cross section of variable size and aspect ratio.

**Fig. 14** Comparison of raster scan and vector scan writing schemes.}

**Fig. 15** Schematic of positive and negative resists used in electron beam lithography.

### 11.2 Electron Beam, X-Ray, and Ion Beam Resist

**Electron Resist** Electron resist is similar to that of a photoresist: a change is induced in the resist by radiation. For positive resists, electrons cause photochemical damage to the resist material, which causes chemical bonds to be broken and the resist to fragment, as shown in Fig. 15a. A positive resist is a polymer that becomes more soluble in the irradiated area. The irradiated area is then developed with a developer solution that attacks the low-molecular-weight fragments. The nonirradiated polymer is then washed out to leave the pattern.

For a negative resist, the entire polymer is made soluble by irradiation, as shown in Fig. 15a. The nonirradiated polymer is protected by a positive resist, which is then washed away to leave the pattern. For both positive and negative resists, the develop is a Bayer process: expose a positive resist, develop the resist, and then remove the develop. Positive resists are typically composed of negative resists that are made soluble by irradiation. Negative resists also have a long development time, and many resists have poor sensitivity and resolution.

**Polymer Linking** A polymer linking is shown in Fig. 15a. A three-dimensional structure with a third dimension is an example of a polymer linking. The nonirradiated polymer is divided into two parts, each of which is made soluble by irradiation. For example, a polymer linking is used to make a polymer linking that is soluble in the irradiated area.

The characteristic response curve of the resist to radiation is shown in Fig. 15a. The sensitivity is defined as the area to give complete development at a given energy. The energy required at which 50% of the original polymer is soluble is called the energy required to develop the resist. The resist sensitivities and resolutions are much lower for negative resists than for positive resists (e.g., PMMA). A positive resist has a higher energy threshold, and a negative resist has a much shorter energy range over which resists have a higher sensitivity.

**The Proximity Effect** In optical lithography, the diffraction of light. In electron beam lithography, by diffraction (because the waves are damped and higher energies are less than those that penetrate the resist film and expose the resist). These collisions lead to energy losses, and the incident electrons spread out as they traverse the resist. Some of their energy is lost or they leave the resist. Figure 16a shows computed energy loss of 20 keV incident at the resist. The electron trajectories have been projected onto two dimensions, and the electrons are distributed in a diameter of the same order of magnitude (≈ 3.5 μm). Also, there are many electron trajectories and travel backward from the resist, and leave the material.
11.2 Electron Beam, X-Ray, and Ion Beam Lithographies

**Electron Resist**

Electron resists are polymers. The behavior of an electron resist is similar to that of a photoresist, that is, a chemical or physical change is induced in the resist by radiation. This change allows the resist to be patterned. For a positive electron resist, the polymer-electron interaction causes chemical bonds to be broken (chain scission) to form shorter molecular fragments, as shown in Fig. 15a. As a result, the molecular weight is reduced in the irradiated area. The irradiated area can be dissolved in a developer solution that attacks the low-molecular-weight material. Common positive electron resists are poly(methyl methacrylate), called PMMA, and poly(butene-1 sulfone), called PBS. Positive electron resists have resolutions of 0.1 μm or better.

For a negative electron resist, the irradiation causes radiation-induced polymer linking as shown in Fig. 15b. The crosslinking creates a complex three-dimensional structure with a higher molecular weight than that of the nonirradiated polymer. The nonirradiated resist can be dissolved in a developer solution that does not attack the high-molecular-weight material. Poly(glycidyl methacrylate-co-ethyl acrylate), called COP, is a common negative electron resist. COP, like a negative photoresist, also swells during development so the resolutions are limited to about 1 μm.

The characteristic response curves of electron resists are similar to that of photoresists. The sensitivity is defined as the electron dose required per unit area to give complete development of a positive resist or the electron dose required at which 50% of the original resist has been retained for a negative resist. The resist sensitivities and γ values are listed in Table 1. Note that the negative resist COP is much more sensitive than the positive resist PMMA, and that a much shorter exposure time is required for COP. However, slow resists (e.g., PMMA) have a higher resolution than fast resists (e.g., COP); so there is a trade-off between sensitivity and resolution.

**The Proximity Effect**

In optical lithography, the resolution is limited by diffraction of light. In electron beam lithography, the resolution is not limited by diffraction (because the wavelengths associated with electrons of a few keV and higher energies are less than 1 Å) but by electron scattering. When electrons penetrate the resist film and underlying substrate, they undergo collisions. These collisions lead to energy losses and path changes. Thus, the incident electrons spread out as they travel through the material until either all of their energy is lost or they leave the material due to backscattering.

Figure 16a shows computed electron trajectories of 100 electrons with initial energy of 20 keV incident at the origin of a 0.4-μm PMMA film on a thick silicon substrate. The electron beam is incident along the z-axis and all trajectories have been projected onto the xz plane. This figure shows qualitatively that the electrons are distributed in an oblong pear-shaped volume with a diameter of the same order of magnitude as the electron penetration depth (~ 3.5 μm). Also, there are many electrons that undergo backscattering collisions and travel backward from the silicon substrate into the PMMA resist film and leave the material.
Figure 16 (a) Simulated trajectories of 100 electrons in PMMA for a 20-keV electron beam. (b) Dose distribution for forward scattering and backscattering at the resist-substrate interface.

Figure 16b shows the normalized distributions of the forward scattering and backscattering electrons at the resist-substrate interface. Because of the backscattering, electrons effectively can irradiate several micrometers away from the center of the exposure beam. Since the dose of a resist is given by the sum of the irradiations from all surrounding areas, the electron beam irradiation at one location will affect the irradiation in neighboring locations. This phenomenon is called the proximity effect. The proximity effect places a limit on the minimum spacings between pattern features. To correct for the proximity effect, patterns are divided into smaller segments. The incident electron dose in each segment is adjusted so that the integrated dose from all its neighboring segments is the correct exposure dose. This approach further decreases the throughput of the electron beam system, because of the additional computer time required to expose the subdivided resist patterns.

11.2.2 X-Ray Lithography

X-Ray lithography uses a shadow printing method similar to optical proximity printing. The X-ray wavelength (4 to 50 Å) is much shorter than the wavelength of the ultraviolet light (2000 to 4000 Å) used for optical lithography. Therefore, diffraction effects are reduced and higher resolutions can be obtained. For an X-ray wavelength of 5 Å and a gap of 40 μm, the minimum line width obtained from Eq. 1 is less than 0.2 μm. Compared to the electron beam system, X-ray lithography can allow parallel exposure, as opposed to the exposure of菲.

Figure 17 shows an X-ray system generated by an electron gun incident on X-rays with a wavelength of 4.4 Å. The X-rays are filtered into a helium-filled chamber to the X-ray mask and semiconductor. They are then moved to the right side of the chamber.

Figure 18 shows the geometric effect of finite size of the X-ray source (with a gap g, a penumbral effect results.
Lithography and Etching

11.2 Electron Beam, X-Ray, and Ion Beam Lithographies

beam system, X-ray lithography can have a higher throughput because it uses parallel exposure, as opposed to the serial exposure approach of electron beam lithography.

Figure 17 shows an X-ray system. A 25-keV, 5-kW electron beam generated by an electron gun is incident upon a palladium target that emits X-rays with a wavelength of 4.4 Å. The X-rays pass through a beryllium window into a helium-filled chamber to the mask and wafer. Helium is used in the chamber because air is a strong absorber of X-rays. As shown on the left side of Fig. 17, the X-ray mask and semiconductor wafer are first aligned with each other. They are then moved to the exposure position as shown on the right side of the figure.

Figure 18 shows the geometric effects on X-ray lithography. Because of the finite size of the X-ray source (with diameter $a$) and the finite mask-to-wafer gap $g$, a penumbral effect results. The penumbral blur $\delta$ on the edge of the
11.2 Electron Beam, X-Ray, and Ion Beam Lithography

A silicon wafer is used as the substrate. A polyimide film (also ~6 μm) is deposited on the substrate. A thin layer of gold (≈ 4.4 Å) is used as an X-ray absorber. Masking is performed using electron beam lithography. The photoresist and the silicon substrate is removed by etching it from the back to form the device.

We can use electron beam resist for fine resolution, which is absorbed by an atom, the emission of an electron. The excited atom emits an X-ray having a different wavelength than the one absorbed by another atom, and this results in the emission of electrons equivalent to one being irradiated from any of the processes. Once the excitation or chain scission will occur, depending on the X-ray resist, which would be attractive X-ray resist is DCOP (2-dimethylaminoethyl methacrylate-co-ethyl acrylate), (~10 mJ/cm²).

11.2.3 Ion Beam Lithography

Ion beam lithography can achieve high-resolution features, and therefore scatter less than electron beams. It is also used in a focused-beam deposition process. The depth of the resist is similar to photolithography.

Figure 19 shows the computed depth of the resist after ion implantation at 60 keV into PMMA.

\[
\delta = a \frac{g}{L} \quad (9)
\]

where \( L \) is the distance from the source to the X-ray mask. If \( a = 3 \text{ mm}, \ g = 40 \mu \text{m}, \) and \( L = 50 \text{ cm} \), the penumbral blur is of the order of 0.2 μm. The blurring at the edge of a feature causes a loss of resolution in the subsequent processing. Another geometric effect is the lateral magnification error, due to the finite gap \( g \) and the nonvertical incidence of the X-ray flux. The projected images of the mask are shifted laterally by an amount \( d \), called runout:

\[
d = r \frac{g}{L} \quad (10)
\]

where \( r \) is the radial distance from the center of the wafer. The runout is zero at the center of the wafer but increases linearly toward the wafer edge. For a 125-mm wafer, the runout error can be as large as 5 μm (assuming \( g = 40 \mu \text{m} \) and \( L = 50 \text{ cm} \)). This runout error must be compensated during the mask making process.

The insert of Fig. 18 shows the cross section of an X-ray mask. The construction of an X-ray mask is much more complicated than that of a photomask.
Lithography and Etching

11.2 Electron Beam, X-Ray, and Ion Beam Lithographies

A silicon wafer is used as the substrate. A boron nitride film ($\sim 6 \, \mu m$) is deposited on the silicon substrate, followed by a spun-on polyimide film (also $\sim 6 \, \mu m$). Because gold has a relatively high absorption coefficient at 4.4 Å, a gold film ($\sim 0.6 \, \mu m$) is deposited on the polyimide to serve as an X-ray absorber. Mask patterns are defined in the gold film using electron beam lithography. The patterned wafer is bonded to a Pyrex ring, and the silicon substrate is removed (except underneath the Pyrex ring) by etching it from the back to form the membrane structure shown in Fig. 18.

We can use electron beam resists as X-ray resists. This is because when an X-ray is absorbed by an atom, the atom goes to an excited state with the emission of an electron. The excited atom returns to its ground state by emitting an X-ray having a different wavelength than the incident X-ray. This X-ray is absorbed by another atom, and the process repeats. Since all the processes result in the emission of electrons, a resist film under X-ray irradiation is equivalent to one being irradiated by a large number of secondary electrons from any of the processes. Once the resist film is irradiated, chain crosslinking or chain scission will occur, depending on the type of resist. Table 1 lists some X-ray resists along with their sensitivities and $\gamma$ values. One of the most attractive X-ray resists is DCOPA (dichloropropyl acrylate and glycidyl methacrylate-co-ethyl acrylate), because it has a very low threshold ($\sim 10 \, mJ/cm^2$).

11.2.3 Ion Beam Lithography

Ion beam lithography can achieve higher resolution than optical, X-ray, or electron beam lithographic techniques because ions have a higher mass and therefore scatter less than electrons. Ion beam lithography can be operated in a shallow printing mode using PMMA as an ion beam resist. Ion beams can also be used in a focused-beam direct-writing mode. We have considered a few resistless-pattern generation approaches using an ion beam in Section 10.6.

Figure 19 shows the computer-simulated trajectories of 50 H$^+$ ions implanted at 60 keV into PMMA and various substrates. Note that the

![Fig. 19 Trajectories of 60-keV H$^+$ ions traversing through PMMA into Au, Si, and PMMA.](image)

\[ \text{Distance (} \mu \text{m)} \]

\[ \text{Depth (} \mu \text{m)} \]

The runout of the wafer. The runout is zero normally toward the wafer edge. For a}$^g \sim 40 \, \mu m$ be compensated during the mask generation of an X-ray mask. The con-

\[ (9) \]

\[ (10) \]
spread of the ion beam at a depth of 0.4 \mu m is only 0.1 \mu m in all cases (compare with Fig. 16a for electrons). The backscattering is completely absent for silicon substrate, and there is only a small amount of backscattering for the gold substrate.

Ion beam lithography is still in its initial development stage. Its major advantage is high spatial resolution due to absence of proximity effect. Ion beam lithography will become important when the minimum feature dimension is reduced to about 0.2 \mu m and below.

### 11.3 Wet Chemical Etching

The throughput of electron beam lithography is limited by the time required to transfer the designed pattern to the wafer. This can be increased by using a scanning electron beam technique or by increasing the exposure time. For example, a 5:1 step-and-repeat or electron beam method can be used for the most critical mask levels, while 1:1 scanning projection can be used for the other levels.

Figure 20 shows the estimated resolution for various lithographic systems as a function of 125-mm wafer throughput. For feature sizes of 1 \mu m or larger, the 1:1 scanning projection system has the highest throughput. For 1- to 0.5-\mu m feature sizes, the 5:1 step-and-repeat system has the highest throughput. For feature sizes below 0.5 \mu m, the electron beam and optical lithography systems are more efficient.

### 11.3.1 Wet Chemical Etching

Wet chemical etching is used for the remaining levels of the process. The wafer is chemically cleaned and scrubbed to remove any remaining material. The wet etching process is then carried out. The etchant solution reacts with the silicon substrate, dissolving it away. The etchant solution is then rinsed out, and the remaining material is dried. The process is then repeated for the next level of the fabrication process.
Week 7

Thin Film Deposition
Chapter 13

Chemical Vapor Deposition

The last two chapters discussed physically based methods for depositing thin films: evaporation and sputtering. They are called physical processes because the techniques did not involve chemical reactions. Rather, they produce a vapor of the material to be deposited by heating (evaporation) or by energetic ion bombardment (sputtering). While most metal films for silicon ICs are deposited using these methods, they have major problems associated with step coverage. This is a particular concern with submicron technologies where very small contacts require the coverage of high aspect ratio features. Furthermore, these techniques are not well suited to the deposition of insulating or semiconducting films. This chapter will discuss methods of thin film deposition based on chemical reactions. The chemicals are normally supplied through the use of gaseous compounds. By definition, the chemical bonding state of the final product is different than the precursor gasses.

Chemical vapor deposition (CVD) has become extremely popular and is the preferred deposition method for a wide range of materials. Thermal CVD also forms the basis for most epitaxial growth in IC manufacturing. Modifications of simple thermal CVD processes provide alternate energy sources such as plasmas or optical excitation to drive the chemical reactions, allowing the deposition to occur at low temperature. For a comprehensive review of CVD for IC fabrication, the reader is referred to Sherman [1]. Unfortunately, CVD does not lend itself to simple analytic explanation. Both the gas flow in the reactor and the chemical reactions require a detailed numerical analysis that is both reactor and process dependent. This chapter will begin by introducing a simple CVD system that will be used in the following two sections to discuss the equations that would have to be solved in order to understand the chemical reactions and flows in the chamber. The remainder of the chapter will discuss various types of deposition systems and specific gas chemistries used to deposit materials of interest.

13.1 A Simple CVD System for the Deposition of Silicon

To begin to understand CVD processes, consider the simple reactor shown in Figure 13.1. The reactor consists of a tube with a rectangular cross section. The walls of the tube are maintained at a temperature $T_w$. A single wafer rests on a heated susceptor in the center of the tube. This susceptor is maintained at $T_s$, where in many cases $T_s \gg T_w$. In order to discuss a simple but representative process we will use the decomposition of silane gas ($\text{SiH}_4$) to form polycrystalline silicon. Assume that the gas flows through the tube from left to right. Since the silane will begin to decompose when it approaches the
hot susceptor, the concentration of silane, and therefore the deposition rate, may decrease along the length of the tube. To improve the uniformity of the deposition, the silane can be mixed in an inert carrier gas. A common diluent for silane is molecular hydrogen (H₂). Assume the chamber is fed a mixture of 1% SiH₄ in H₂. Not only are the use of diluents common practice in real systems, it also avoids further complications in the chemistry of the reaction, since at typical deposition conditions very little of the hydrogen can decompose. Finally assume that the temperature of the gas as it enters the tube is the same as the wall temperature. The reaction products and any unreacted silane flows out of the tube at the right. The flows in the chamber will be slow enough (see Section 9.3) that the pressure in the chamber can be considered uniform.

The overall reaction that must occur is

\[
\text{SiH}_4(g) \rightarrow \text{Si}(s) + 2\text{H}_2(g)
\]  \hspace{1cm} (13.1)

where the quantity in parenthesis is \((g)\) for gas phase and \((s)\) for solid. The detailed process by which this overall reaction occurs is much more complex. One of the distinctions drawn regarding CVD is the location of the reaction that liberates a solid atom or cluster of atoms from a gaseous source. If that reaction occurs spontaneously in the gas above the wafer, it is called a homogeneous process. Such processes are generally undesirable. Using the deposition of silane as an example, excessive homogeneous reactions will result in large silicon particles in the gas phase that gradually accumulate on the wafer. The result is a deposit with poor surface morphology and inconsistent properties. In real systems, such deposits have a poorly controlled composition and may have significant contamination from residual gases in the chamber. This chapter will therefore emphasize processes that are heterogeneous. That is, processes that operate in such a manner as to greatly favor the formation of solids only at surfaces. Even for processes that are run this way, homogeneous reactions are still important. For example, in the deposition of silane, the homogeneous production of silylene (SiH₂) is a crucial process because it is generally believed it is the silylene rather than silane itself that adsorbs on the surface of the wafer and produces the solid silicon. The distinction here is that the homogeneous reaction produces a gaseous, not a solid, product. The chapter will focus initially on the simplest type of heterogeneous reaction, which is run in a cold wall chamber (like the one shown in Figure 13.1), where all deposition reactions occur at the surface of the wafer.

Generally, the steps that occur during a chemical vapor deposition process include: (1) the transport of the precursors from the chamber inlet to the proximity of the wafer, (2) reaction of these gasses to form a range of daughter molecules, (3) transport of these reactants to the surface of the wafer, (4) surface reactions to release the silicon, (5) desorption of the gaseous byproducts, (6) transport of the byproducts away from the surface of the wafer, and (7) transport of the byproducts from the reactor. Even if the discussion is limited to thermal CVD in this very simple deposition system, understanding each of these steps is a formidable task. To simplify matters, the problem is often divided in half. The next section will focus on the chemical reactions that occur in the reactor, both in the gas phase and at the surface of the wafer. The following section will discuss the flow of gases in the reactor. Choosing to study a system that contains only a small concentration of the reactant gas (1% SiH₄ in H₂) allows this separation to be fairly realistic. The thermal and mechanical properties of the gas are relatively unaffected by any chemical reaction in the small, active component.
13.2 Chemical Equilibrium and the Law of Mass Action

Focusing for the moment on CVD processes that involve long times and many collisions between molecules, the chemical composition at each point in the reactor approaches equilibrium. To understand chemical equilibrium consider a unit volume of the gas somewhere in the chamber (Figure 13.2). Assume that the volume is small enough that the temperature and chemical composition in this volume is uniform. One of the reactions that one might think of is

\[ \text{SiH}_4(g) \rightleftharpoons \text{SiH}_2(g) + 2\text{H}(g) \quad (13.2) \]

The double arrow indicates that the reaction proceeds in both directions. Chemical equilibrium is reached when the concentration of each species is constant, even if the gasses take an arbitrarily long time traversing this unit volume. (Actually this reaction is unlikely, but it is instructive.)

Assume for the moment that this is the only reaction that occurs. The law of mass action says that

\[ K_p(T) = \frac{p_{\text{SiH}_2}p_{\text{H}}^2}{p_{\text{SiH}_4}} \quad (13.3) \]

where \( p \) refers to the partial pressure of the subscripted species, and \( K_p(T) \) is a reaction equilibrium constant that depends only on the temperature. The atomic hydrogen term is squared because of the 2 in front of the atomic hydrogen term of the species balance equation (Equation 13.2). The equilibrium constant generally follows an Arrhenius function

\[ K_p(T) = k_p e^{\Delta G/RT} \quad (13.4) \]

where \( \Delta G \) is the change in the Gibbs free energy in the reaction.

Assume that \( K_p(T) \) is known for this process. There are three unknowns (the three partial pressures) and only one equation. Solving for them requires two more equations. The total pressure of the reactor \( P \) is a constant whose value is normally known. It is the sum of the partial pressures

\[ P = p_{\text{SiH}_4} + p_{\text{SiH}_2} + p_{\text{H}} + p_{\text{H}_2} \quad (13.5) \]

Assuming, for example that the chamber is run at atmospheric pressure, the partial pressure of \( \text{H}_2 \) is the same as its inlet partial pressure (0.99\%P) since it is assumed to be inert. The final equation comes from the inlet flow. One can use the Si/H ratio as

\[ \frac{Si}{H} = \frac{f_{\text{SiH}_4} + 2f_{\text{H}_2}}{4f_{\text{SiH}_4} + 2f_{\text{H}_2}} = \frac{p_{\text{SiH}_4}}{4p_{\text{SiH}_4} + 2p_{\text{H}_2} + p_{\text{H}}} \quad (13.6) \]

where the \( f \) terms are the inlet flows, also assumed to be known.

This expression has not taken any other reactions into account. For example, in actual CVD, silicon is consumed
from the gas phase. In that case, the partial pressure of silicon cannot be completely determined by the inlet flows. Instead, one must consider the inlet flow to be a source of silicon containing molecules and the deposition surface to be a sink. Then, the silicon containing molecule flux, which depends on the flow field and diffusion, must be calculated.

To begin to develop a more realistic picture, some of the reactions that would have to be included are [2]:

\[
\text{SiH}_4(g) \rightleftharpoons \text{SiH}_2(g) + \text{H}_2(g) \quad (13.7)
\]

\[
\text{SiH}_4(g) + \text{SiH}_2(g) \rightleftharpoons \text{Si}_2\text{H}_6(g) \quad (13.8)
\]

\[
\text{Si}_2\text{H}_6(g) \rightleftharpoons \text{H}_2\text{SiH}_3(g) + \text{H}_2(g) \quad (13.9)
\]

Other reactions are, of course, possible, and \emph{a priori} one cannot decide which reactions to include. Instead, one must find the equilibrium constant for each possible reaction and ignore only those reactions for which the \(K_p(T)\) values are negligibly small. Finding the equilibrium partial pressures, therefore, requires an equilibrium constant for each of the three reactions listed above and the solution of a set of coupled algebraic equations.

**Example 13.1**

Assume that the gas AB is introduced into a reactor and that the only chemical reaction that occurs in the chamber is:

\[
\text{AB} \rightleftharpoons \text{A} + \text{B} \quad (13.10)
\]

If the process is run at 1 atm (760 torr) and a temperature of 1000 K and the process reaches chemical equilibrium, calculate the partial pressure of each species. The equilibrium constant for this reaction is given by

\[
K(T) = 1.8 \times 10^9 \text{ torr} \cdot e^{-\frac{2.0eV}{kT}} \quad (13.11)
\]

At 1000 K the equilibrium constant is calculated to be 0.15 torr. Then

\[
0.15 = \frac{p_A p_B}{p_{AB}} \quad (13.12)
\]

and the total pressure \(P\) is the sum of the partial pressures:

\[
P = p_A + p_B + p_{AB} \quad (13.13)
\]

We have three unknowns but only two equations. Since A and B are both created by the dissociation of the inlet gas, there must be an equal number of each. It would be reasonable to assume that the partial pressures are equal. Then

\[
p_A^2 + 0.3p_A - 0.15 \cdot P = 0 \quad (13.14)
\]

Setting \(P = 760 \text{ torr}\) and solving gives \(p_A = p_B = 10.5 \text{ torr}\) and \(p_{AB} = 739 \text{ torr}\).
The discussion thus far makes a crucial approximation: that all species are in chemical equilibrium. To understand the limitation of this approximation, consider what happens as the pressure of the chamber is reduced. At a low enough pressure, the mean free path of the molecules approaches the width of the chamber (see Section 10.1). If these gas phase collisions do not occur, the species in the gas often do not reach thermal equilibrium and therefore cannot achieve chemical equilibrium. Furthermore, since the gas molecules have a distribution of energies, a large number of collisions must occur in each unit volume for the gasses to potentially reach chemical equilibrium. Therefore, a characteristic chamber length such as the distance between the gas injector and the susceptor must be at least a few orders of magnitude larger than the mean free path. Depending on the particular reaction equations involved, some processes may reach equilibrium, others will not. Processes that do not reach chemical equilibrium are called \textit{kinetically controlled processes}. Typically low pressure CVD is kinetically controlled, while atmospheric CVD may be in equilibrium.

To begin to understand this class of deposition, reconsider Reaction 13.7. This is a prototypical reaction in silicon CVD. When the reaction is kinetically controlled, it is written as

$$\text{SiH}_4(g) \xrightleftharpoons{\text{K}_r}{\text{K}_f} \text{SiH}_2(g) + \text{H}_2(g) \quad (13.15)$$

where \(\text{K}_f\) and \(\text{K}_r\) are the forward and reverse reaction rate coefficients.

By writing a similar expression for each of the chemical reactions, it is possible to construct a differential equation for the time rate of change of the concentration (or partial pressure) of all of the chemical species. For example, if only Reactions 13.7 through 13.9 are considered, the time rate of change of silane is given by

$$\frac{d}{dt} C_{\text{SiH}_4} = -k_{f1} C_{\text{SiH}_4} + k_{r1} C_{\text{SiH}_2} C_{\text{H}_2} - k_{f2} C_{\text{SiH}_4} C_{\text{SiH}_2} + k_{r2} C_{\text{SiH}_6} \quad (13.16)$$

where the subscript 1 on the first two terms refers to the reaction given in Equation 13.7, and the subscript 2 refers to the reaction listed in Equation 13.8. Equation 13.9 does not involve silane and so does not enter into Equation 13.16. One can construct similar equations for the other species. If the \(k(T)\)'s are known, we are left with a set of coupled first order differential equations that can be solved for the rate of change of each chemical species in the unit volume. If one also takes into account diffusion of chemical species due to concentration gradients, the residence time of the gasses in the unit volume are known (through the flow velocities), and the value of the temperature for each of the unit volumes is known, one could begin to solve for a map of the chemical species in the chamber. While it is possible to solve for the species balance in very simple systems, real CVD systems stretch the capabilities of even supercomputers [3]. Furthermore, many of the important rate coefficients, even in the most commonly used chemical systems, also are not presently known accurately enough to produce meaningful information. For that reason, we will set aside attempts at meaningful quantitative discussion of the CVD reactions and use this introduction as a basis for a more qualitative discussion.

While the chemical reactions in the gas phase are at least qualitatively understood, the situation at the surface is much less clear. Part of the problem are the tools that can be used to investigate the gas sample over some finite volume. When these same techniques are applied to the surface, there is insufficient signal. Methods that have been used to develop a detailed understanding of the surface...
during high vacuum growth cannot be used in chambers under typical CVD conditions. Although new techniques have emerged in the last few years, the picture that we have is still tentative and qualitative.

The last two sections suggest that for the deposition of silicon from silane, several types of silicon containing species may strike the surface of the wafer. In the boundary layer model, one can readily calculate the flux of the molecules if a boundary condition is established for the concentration at the surface. It is common practice to define a phenomenological parameter called the *sticking coefficient*, which varies from zero for molecules that are completely reflected from the surface to unity for molecules that irreversibly adsorb. In the case of silylene, it is commonly assumed that the sticking coefficient is one. On the other hand, it has been shown that the probability of an incident silane molecule sticking and reacting is given by [4]

\[
y_{\text{SiH}_4} = 0.054e^{-0.81eV/RT}
\]  

(13.17)

At typical deposition conditions, the sticking coefficient is only about \(10^{-6}\); however, the flux of silane to the surface is much larger than the flux of silylene, so that both molecules may contribute to the deposition rate.

Once the molecule is on the surface, a chemical reaction must occur to remove the silicon atom and free the hydrogens. Taking silylene as an example, the molecule first adsorbs:

\[
\text{SiH}_2(g) \rightleftharpoons \text{SiH}_2(a)
\]  

(13.18)

The overall surface reaction must take the form

\[
\text{SiH}_2(a) \rightleftharpoons \text{Si}(s) + \text{H}_2(g)
\]  

(13.19)

where \((a)\) refers to adsorbed species and \((s)\) refers to atoms that have been incorporated into the solid. Because of the high concentration of \(\text{H}_2\) in the gas, it is believed that the surface is covered with either physically adsorbed \(\text{H}_2\) (at low temperature) or chemically adsorbed \(\text{H}\) (at high temperature). These surface species must be desorbed to allow Reaction 13.19 to proceed. The desorption process also follows an Arrhenius behavior, so that the surface will have a concentration of vacancies, whose density increases with increasing temperature. The adsorbed silylene can diffuse across the passivated surface (Figure 13.3) until it finds such a vacancy at which point it will bond and eventually lose its hydrogen atoms. This diffusion across the surface plays an important role in CVD processes. When the surface diffusion is large (of order millimeters), the deposition is very uniform. When the surface diffusion length is short, a less uniform deposit will result. As with physical deposition processes, surface diffusion increases exponentially with temperature, and so film uniformity can generally be improved by heating the wafer.

![Figure 13.3](image-url) A simple model of the surface of the wafer during silane CVD includes adsorbed SiH₄ and SiH₂.
13.3 Gas Flow and Boundary Layers

The second area that needs to be understood for CVD is gas flow dynamics. The gas flow in the reactor is important because it determines the transport of the various chemical species in the chamber, and it plays a significant role in the temperature distribution in the gas in many reactors. Furthermore, the temperature distribution will also affect the flow. If the mean free path of the gas is much smaller than the chamber geometries, the gas can be treated as a viscous fluid. Furthermore, if the flow velocities are much less than the speed of sound (low Mach numbers) the gas can be considered incompressible. Nearly all CVD systems operate in pressure and flow regimes that make these approximations valid. Finally, as a starting point assume that the gas velocity is low enough that the gas flows along the contours of the chamber. The flow is said to be laminar and it can be well described by the mechanical properties of the gas.

If the reactor is a circular tube and all surfaces are at the same temperature, the problem can be simplified considerably. Assume that the gas is introduced with a uniform velocity \( U_\infty \) at the left end of the tube (Figure 13.4). One important feature of gas flows is that the gas velocity must be zero at all surfaces. Because of the finite gas viscosity, the flow velocity must vary smoothly from zero at the walls to some maximum value at the center. This change from a uniform or plug flow, to a fully developed tube flow occurs over a distance \( z_c \):

\[
z_c = \frac{a}{25} N_{Re}
\]

(18.20)

where \( a \) is the radius of the tube, and \( N_{Re} \) is a dimensionless quantity known as the Reynolds number. The Reynolds number \( (N_{Re}) \) is given by

\[
N_{Re} = \frac{U_\infty L}{\mu} = \frac{U_\infty L}{\eta}
\]

(18.21)

where \( L \) is a characteristic length of the chamber (such as the radius \( a \)), \( \mu \) is the kinematic viscosity, \( \rho \) is the mass density of the gas, and \( \eta \) is the dynamic viscosity of the gas [5]. When \( Re \) is low, the flow in the tube is dominated by the finite viscosity effects and so is parabolic across the chamber. The velocity then is given by

\[
v(r) = \frac{1}{4 \eta} \frac{dp}{dz} (a^2 - r^2)
\]

(18.22)

where \( dp/dz \) is the pressure gradient across the tube, which is assumed to be small. At very large \( Re \), the gas cannot support the large velocity gradients required for fully developed laminar flow, and so the flow becomes turbulent. The transition between laminar and turbulent flows depends on the gas. For example, when \( Re > 2300 \) in \( H_2 \), the flow is turbulent [6].
Example 13.2

An LPCVD tube operating at 10 torr has an inlet gas flow of 1000 sccm (standard cubic centimeters per minute) of nitrogen. At the reactor temperature of 1000 K, the dynamic viscosity of nitrogen is 0.04 gm/cm-sec. The reactor is 200 mm in diameter. Estimate the length required for fully developed flow and calculate \( v(r) \) after the flow is fully developed.

According to 13.20 and 13.21,

\[
z_v = \frac{d^2}{25} \frac{U_w}{\eta} \rho
\]

The mass density can be calculated using the ideal gas law

\[
\rho = m_{N_2} \frac{P}{kT} = 5.8 \times 10^{-6} \text{ gm/cm}^3
\]

If \( U_w \) is taken as the plug flow velocity,

\[
U_w = 1000 \text{ cm}^3/\text{min} \frac{1 \text{ min}}{60 \text{ sec}} \frac{1000 \text{ K}}{760 \text{ K}} \frac{760 \text{ torr}}{10 \text{ torr}} \frac{1}{\pi(10 \text{ cm})^2} = 15 \text{ cm/sec}
\]

and

\[
z_v = \frac{(10 \text{ cm})^2}{25} \frac{15 \text{ cm/sec}}{5.8 \times 10^{-6} \text{ gm/cm}^3} \frac{5.8 \times 10^{-6} \text{ gm/cm}^3}{0.04 \text{ gm/cm-sec}} = 0.0085 \text{ cm}
\]

Since the total flow must be constant,

\[
\int 2\pi v(r) dr = 15 \text{ cm/sec} \pi 100 \text{ cm}^2
\]

one can readily show that

\[
v(r) = 30 \text{ cm/sec} \left[ 1 - \frac{r^2}{a^2} \right]
\]

To begin to understand the flow fields of a more complicated chamber, return to Figure 13.1, but for the moment keep the system at a uniform temperature. The wafers will rest on the bottom surface of the chamber. The previous discussion suggests that the gas flow velocity must go to zero at the surface of the wafer. The standard textbook picture has the height of the chamber large enough to have a large \( N_{Re} \) and so a broad region of uniform gas velocity approximately equal to \( U_w \). To further simplify the behavior of flow, it is customary to approximate the parabolic fall off of the gas
velocity in the vicinity of the wafer as a boundary layer of width $\delta(z)$ where, for a flat surface whose normal is perpendicular to the flow direction

$$\delta \approx 5 \frac{\sqrt{\frac{\mu z}{U_w}}}{\sqrt{U_w}} \quad (13.23)$$

In this model, the gas flow in the boundary layer is zero, while outside the boundary layer the flow velocity is $U_w$ (Figure 13.5).

If deposition is occurring at the surface of the wafer, the deposition gasses must diffuse through the stagnant boundary layer. The boundary layer thickness, therefore, may play a critical role in determining the deposition rate. Notice that, according to Equation 13.23, for a flat surface the boundary layer thickness increases as $z^{1/2}$. To maintain a uniform boundary layer thickness, CVD systems in which gas transport plays a significant role in the deposition rate often tilt the deposition surface with respect to the flow direction. The wafers then rest on a wedge shaped susceptor. The tilt angle of the wedge must be optimized to obtain the best uniformity for a particular CVD process.

Gas phase diffusivities are much less temperature sensitive than bulk diffusivities. One common form is given by Hammond [7]

$$D_e \approx T^{3/2} \frac{P_e}{P} \quad (13.24)$$

where $P_e$ and $P$ are the partial pressure of the diffusing species and the total pressure, respectively, at the edge of the stagnant layer. One of the ways to distinguish CVD processes that are limited by diffusion through the gas is to measure the temperature dependence of the deposition rate. If $P_e$ can be considered independent of temperature, the deposition rate of such a mass transport limited reaction would increase only weakly with increasing temperature.

The gas flow in CVD reactors is often nonlaminar and may involve recirculations and roll cells. One of the most common sources of these flows is natural convection. As the gas flows past hot surfaces it expands. The expansion is described by an equation of state such as the ideal gas law:

$$\rho = \frac{nm}{V} = \frac{Pm}{kT} \quad (13.25)$$

where $m$ is the molecular mass. In so doing, the mass density decreases. (Recall that the pressure is fixed.) The hot gas tends to float or rise in the reactor with respect to cooler gasses. This effect, called natural con-
convection, must also be considered when calculating the flow of gasses in a real chamber under deposition conditions [8]. The effects are most pronounced when heavy molecules are used at near atmospheric pressures. Conversely, natural convection has almost no effect for low pressure H$_2$ ambients.

Figure 13.6a shows the calculated flow fields for a horizontal reactor with a square cross section. The calculations were done by discretizing the reactor in 3-D and solving for momentum conservation in the form of the Navier-Stokes equation over this grid. Gasses are injected at the back (upper left) end of the box and flow toward the front (lower right) end of the box. The lower surface is a hot susceptor. The hot susceptor causes the gas at the center of the reactor to rise, leading to transverse

Figure 13.6b  Roll cells in hemispherical tubes at 760 torr (left) and 160 torr (right) (after Takahashi et al. reprinted by permission of the publisher, The Electrochemical Society).
roll cells that develop along the length of the reactor. Again, a
quantitative discussion of the exact flow in any reactor is well beyond
the scope of this text, but these principles can be used to develop a
qualitative understanding of a variety of chamber geometries and CVD
processes. Figure 13.6b shows TiO$_2$ particles in a carrier
gas as they flow through semicircular tubes. The left-hand
figure shows the flow at 760 torr and the right-hand picture shows the flow at 160 torr [8]. Obviously
the roll cells are greatly diminished at low pressure.

In other reactors, natural convection shows up as
recirculation cells in which part of the gas flow has a
significant velocity component in the opposite direction
of the intended flow (Figure 13.7). Recirculation cells
may also arise when the cross-sectional area of the
chamber changes abruptly. These cells can trap gas
components that would otherwise be swept from
the growth system. The effect can lead to an incomplete
flushing of the chamber. If the composition of the de-
posit changes, this change will be much less abrupt than
if the reactor and process are designed to avoid recirc-
ulation cells.

### 13.4 Evaluation of the Simple CVD System

Now examine what happens if the deposition rate in our example CVD reactor is measured as a
function of temperature. The prediction of the growth rate and growth rate uniformity requires exten-
sive fluid dynamics and chemical concentration calculations. Instead we will work backward by
developing a qualitative understanding of experimental results. The deposition rate of the silicon
containing precursor (silane) will be used as a growth parameter. Figure 13.8 shows the results. At
low wafer temperature the deposition rate increases exponentially with decreasing inverse temperature.
In this regime the limiting step of the process is some reaction rate. This may be in the gas phase or
at the surface. The information provided is insufficient to determine whether the gas is in chemical
equilibrium or is kinetically controlled. Processes operated in this region are called reaction rate
limited. CVD systems that operate in this region must have excellent temperature control and tem-
perature uniformity. These systems are primarily large batch systems where many wafers are processed
at relatively low rates. Flow dynamics in this chamber are less of a concern, except to the extent that
they contribute to temperature nonuniformity across the wafer. For that reason, this type of reactor
often heats not only the wafer but the walls as well. It is therefore called a hotwall batch CVD reactor.

At high wafer temperature, the growth is limited by the arrival rate of the growth species. It is
often assumed that the limitation is diffusion across the boundary layer. As already pointed out, the
temperature dependence of the gas phase diffusivity is much lower than the activation energy of the
chemical reaction. Processes operating in this regime are referred to as mass transport limited. Then
the concentration of the deposition gas or gasses controls the deposition rate. Remember that the rate
of production of these gasses from the precursors may also have a temperature dependence. CVD
systems that operate in mass transport limited regime must have excellent control of the gas flows,
and the chamber geometry must be designed so as to ensure a uniform transport to all parts of all
wafers. As a result, these systems often are single wafer or small batch systems.
In addition to deposition rate and uniformity, CVD films must be examined with regard to stress, step coverage, and composition. Just as is the case for sputtered films, CVD layers with large compressive or tensile stress may crack, particularly when they cover steps. The step coverage of CVD films is usually very good even for high aspect ratio features. Plasma enhanced CVD films however, may be somewhat re-entrant. This will be discussed later in the chapter. One of the major concerns with CVD layers is chemical composition. In the silane decomposition process that has been discussed so far, the deposited silicon may have a high concentration of hydrogen. This leads to low density films that may etch more quickly than pure silicon films. Residual gasses in the chamber, such as oxygen or water vapor, may also react with the silicon to form SiO₂ layers with high resistivities. In many processes, compounds such as SiO₂ or Si₃N₄ are the desired result. In that case, not only are contaminants a possibility, but also the stoichiometry of the deposited films may differ from the ideal. For example, it is common to anneal a CVD oxide in oxygen at high temperature to densify the film. This step adds oxygen and moves the film closer to stoichiometric SiO₂. Plasma enhanced CVD (PECVD) films are particularly prone to stoichiometry concerns. One common and easy method to check the density of deposited oxides is to measure the wet chemical etch in dilute HF solutions. As deposited, PECVD films frequently have etch rates ten times those of thermal oxides.

13.5 Atmospheric CVD of Dielectrics

Some of the earliest CVD processes were done at atmosphere pressure (APCVD) because of the large reaction rates and the simplicity of the CVD system, particularly for the deposition of dielectrics. While the deposition of silicon from silane, as discussed earlier in the chapter, has been run at atmospheric pressure, the uniformity is poor. It is easy to obtain good uniformity at low pressure, and so APCVD is generally reserved for thick dielectrics, where deposition rates in excess of 1000 Å/min make the process very attractive.

Figure 13.9 shows a simple continuous feed atmospheric CVD reactor. Wafers travel from cassette to cassette on a heated chain track. Wafer temperatures may be anywhere from 240 to 450°C [9]. The gasses are injected from a showerhead above the wafers. When the oxygen to silane gas flow ratio is at least 3:1, stoichiometric SiO₂ will result. SiH₄ and O₂ is a very volatile mixture, which reacts readily. Without a sufficient flow of a diluent such as N₂, this reaction proceeds in the gas phase, resulting in poor morphology.
As will be discussed in Chapter 15, it is often desirable to deposit silicon dioxide films with 4 to 12% phosphorus. These phosphosilicate glasses (PSG) soften and reflow at moderate temperatures, smoothing wafer topology and gettering many impurities. PSG can be formed in an atmospheric process by adding phosphine (PH₃). Figure 13.10 shows a typical plot of the deposition rate of PSG versus temperature and oxygen to hydrogen flow rates. For high oxygen concentration ambients (30:1), the deposition rate increases sharply with temperature and is probably reaction rate limited. For low oxygen containing ambients (2.5:1), the growth rate actually decreases slightly with increasing temperature. The phosphorus content of the film can be controlled by changing the phosphine to silane ratio. Due to the toxicity of phosphine and silane, APCVD systems designed for PSG deposition are generally housed in a vented cabinet. To improve uniformity and step coverage, many PSG and borophosphosilicate glass (BPSG) processes now use organometallic sources such as TEOS (tetraethylxosilane or Si(OC₂H₅)₄). TEOS and ozone can also be used to deposit SiO₂ at about 400°C [10]. TEOS is supplied as a stable, inert, high vapor pressure liquid that is used in a bubbler (see Section 3.8). One of the advantages of using TEOS is the elimination of the need for some of the hazardous chemical handling. The lines from the bubbler must be heated to prevent deposition on the walls of the tubing. Various alternate organometallics have also been investigated. Hexamethyldisiloxane ((CH₃)₃-Si-O-Si-(CH₃)₃), a linear disiloxane, has also shown excellent characteristics comparable to TEOS [11], but the deposition rate was found to depend on the substrate material. Similar films have also been deposited from hydridosiloxanes [12] with wet oxygen at about 500°C.

The major drawback of APCVD is particle formation. While particle formation in the gas phase can be controlled by adding a sufficient amount of N₂ or another inert gas, heterogeneous deposition can also oc-

**Figure 13-10** Deposition rate of PSG in an APCVD system (after Kern and Rosler, ©1977, AIP).

**Figure 13-11** Showerhead design used to minimize deposition at the nozzle by maintaining an inert curtain between the reactants.
cur at the gas injectors. Even if the growth rate of these particles is low, after a number of wafers the particles will become large enough to flake off and fall on the wafer surface. To avoid this problem, the showerhead may be segmented so as to keep the reactant gases separated until they are injected into the chamber. Figure 13.11 shows a showerhead arrangement designed to reduce this problem.

13.6 Low Pressure CVD of Dielectrics and Semiconductors in Hot Wall Systems

A variety of system geometries have been used for low pressure CVD (LPCVD). Figure 13.12 shows a sample of some of the most common. We can divide the reactors into hot and cold wall systems. Hot wall systems have the advantages of uniform temperature distributions and reduced convection effects. Cold wall systems are able to reduce deposition on the walls. These deposits can lead to depletion of the deposition species and the formation of particles, which may flake off the walls and fall on the wafers. Deposits on the walls also lead to memory effects: the deposition on the wafer of material previously deposited on the walls. For that reason hot wall reactors must be dedicated to the growth of a particular film.

Virtually all polycrystalline silicon and a considerable amount of dielectric deposition is done in horizontal flow hot wall systems. Instead of using an inclined susceptor, the wafers are close packed vertically, like wafers in thermal oxidation systems. In order to achieve reasonable deposition uniformity in such a system, the process must be designed so as to keep the reaction strictly controlled by the deposition kinetics [13]. Instead of a diluent gas, the use of low pressures (0.1 to 1.0 torr) reduces gas phase nucleation. This process is commonly called LPCVD.

Figure 13.13 shows a photograph of an LPCVD system. Like furnaces, they generally are built in banks of four tubes. In this bank, the upper tube has been removed to show the heater coils. The gasses are controlled at the back of the tube using mass flow controllers and routed to the front of the furnace. Depending on the process to be run, the gasses are either injected through a ring at the front of the tube, or they are
plumbed to a tube that runs the length of the furnace and injects gasses uniformly across the load. Due to the amount of deposition on the walls, most production systems have soft landing cantilever loaders to minimize particulate formation and flaking during the load/unload process. After loading the furnace is closed with an o-ring sealed door. The tube is flushed with an inert gas such as N₂ and pumped to a medium vacuum. The furnace is ramped to the deposition temperature if it is not already idling at that temperature, and the deposition gasses are switched on. The deposition is allowed to proceed for a predetermined time, then the furnace is again flushed with N₂, the pressure raised to an atmosphere, and the wafers unloaded.

A recent innovation in the LPCVD area is the introduction of vertical chambers. Similar to vertical oxidation/diffusion tubes, these systems have several advantages over standard tubes. Since the wafers are all held by gravity, the wafer to wafer spacing in the reactor is more uniform. Convective effects are more uniformly distributed across the wafer. As a result of these advantages, vertical LPCVD systems can routinely achieve uniformities of better than 2% in the deposition of undoped poly and silicon nitride [14]. Vertical CVD systems are more easily integrated into automated factories since the wafers do not have to be tipped to the vertical allowing easier robotic handling. Perhaps the most important advantage of vertical LPCVD systems is the reduced particle counts. The cost of these systems is considerably higher than conventional LPCVD systems however.

Another newer type of hot wall LPCVD system that has widespread application to silicon IC manufacturing is the hot wall cross-flow reactor. In this reactor, the wafers are placed vertically in closely spaced cassettes, arranged so that fresh gas flows past each wafer. This reduces particle formation and improves uniformity. The system requires extensive quartzware maintenance, however.

Most LPCVD poly is done with silane in furnaces at temperatures ranging from 575 to 650°C. The activation energy for poly deposition is about 1.7 eV. It is believed that the deposition rate is limited by H₂ desorption from the silicon surface. Typical polysilicon deposition rates are 100 to 1000 Å/min, so common deposition times are tens of minutes. When the gas is injected at the front of the tube, the furnace may be programmed with a small temperature gradient (25°C) from front to back. This allows the higher reactivity at the back of the tube to compensate for silane depletion. Temperature ramping has the undesirable effect of producing wafers with larger poly grains at the back of the tube than at the front; however, after high temperature annealing the crystal structure of the films is indistinguishable [15]. It is common to obtain thickness uniformities of 5% across a batch of 100 large diameter wafers [16].

Figure 13.14 shows the deposition rate of poly as function of pressure for a range of temperatures [17]. The morphology of the deposit is a sensitive function of the deposition conditions. When the deposit is performed at temperatures below 600°C, it is usually amorphous [15], but may be polycrystalline if the deposition rate is low enough. These layers can be crystallized into poly layers when annealed at low temperatures [18].

LPCVD polysilicon can be N⁺ doped by implantation, solid or POCl₃ diffusion, or in situ doping.
through the addition of arsine (AsH₃) or PH₃. Impurity concentrations approaching 10²¹ cm⁻³ can be achieved in poly layers doped in this manner [19]. It is common to achieve resistivities of less than 1 mΩ-cm in both in situ doped and diffused polysilicon. The major difficulty with in situ doping is that the addition of PH₃ degrades the deposition uniformity, particularly near the edge of the wafer. Recent modeling efforts have suggested that this may be the result of an increased silylene concentration [20]. It is also critical to include the effects of temperature variations across the load, particularly when considering the first and last few wafers [21].

Aside from particle control, one of the major limitations of traditional LPCVD furnaces is the difficulty of integrating them into automated, cluster tool environments. The most common example is polysilicon. It would be desirable to grow a gate oxide, deposit the poly, and dope the poly in a single cluster tool. Such systems can be made by using a number of simple single wafer reactors that can be fed by a central robot arm. In Applied Material’s Precision 5000®, for example, wafers can be shuttled among up to five chambers (Figure 13.15). The wafer enters the load locked area, where it is transferred to a rapid thermal chamber for oxidation. After oxidation, it is transferred to one of two poly deposition chambers. After deposition, it travels to another rapid thermal chamber for phosphorus diffusion using a POCl₃ source.

Various chemistries have been used for hot wall deposition of oxide in LPCVD reactor including silane and oxygen, dichlorosilane (SiCl₂H₂ or DCS) and nitrous oxide (N₂O), and the decomposition of TEOS. The silane and oxygen process can be run at substrate temperatures of less than 500°C. This means that it can often be used to deposit a layer of oxide over an aluminum layer. Films deposited at these temperatures are found to contain significant quantities of silanol (SiOH), hydride

---

**Figure 13-15** A CVD cluster tool showing the central robot and one of the single wafer processing stations (photo courtesy Applied Materials).
(SiH), and water [22]. As with the corresponding APCVD process, the major limitations are particulate generation and low deposition rates [12]. Furthermore, due to the low substrate temperature, the step coverage of these films is often unacceptable.

The DCS and nitrous oxide process [23] must be run at temperatures of about 900°C. Uniformity and step coverage are excellent, and etch rates are close to those of thermal SiO₂. Although these films are virtually free of hydrogen, they do contain measurable amounts of chlorine, which can lead to etching of underlying polysilicon layers. The DCS and nitrous oxide deposition process has a strong nonlinear pressure dependence. Deposition nonuniformities across the tube are common. Special injectors and temperature gradients can be used to attempt to provide a more uniform deposition.

A popular LPCVD oxide process is the thermal decomposition of TEOS. Figure 13.16 shows the deposition rate of oxide as a function of TEOS partial pressure [24]. Typical deposition temperatures of 650 to 750°C are low enough that redistribution of dopants in the substrate is not a concern, but the temperature is still too high to use over aluminum layers. Lower temperature depositions have been reported with alternate organometallics such as diacetoxydimethylsilane, which decomposes at temperatures as low as 450°C [25] and provides excellent step coverage, but is not yet used extensively.

The deposited stress in undoped LPCVD oxides are about 1–3 × 10⁹ dyne/cm². Low temperature depositions are tensile, while high temperature depositions are compressive. The refractive index of thermal SiO₂ is 1.46 [26], while deposited oxides are generally higher. It is often found that the higher refractive index of deposited oxides correlates with a low mass density and high etch rates in buffered HF. As with APCVD, dilute mixtures of phosphine or phosphine and diborane [27] in hydrogen can be added to the TEOS LPCVD process [28] or the SiH₄ and O₂ process [29] to produce doped oxide layers. The addition of phosphine generally increases the deposition rate, but degrades the deposition uniformity, while adding diborane to form BPSG sometimes increases the deposition rate and has little effect on the uniformity.

While silicon nitride can be deposited from silane and a nitrogen containing precursor, LPCVD silicon nitride is most commonly deposited from mixtures of DCS and ammonia (NH₃). Typical deposition temperatures are 700 to 900°C [12]. The activation energy for this process is 1.8 eV. The deposition rate increases with the DCS flow rate. Due to DCS depletion effects, a ramped profile is usually required for this process.

Two common techniques to check the composition of the films are the measurement of the refractive index with ellipsometry and the measurement of the etch rate in buffered HF. The refractive index is usually between 1.8 and 2.2, with 2.0 being the ideal value. High refractive indices indicate
a silicon rich film. Low refractive indices indicate the presence of oxygen, often due to a vacuum leak, a contaminated gas, or an incomplete pumpdown. The presence of oxygen is also indicated if the etch rate of the film exceeds 1 nm/min in 49% HF. Other common impurities in Si$_3$N$_4$ include hydrogen and approximately 0.4% Cl [30]. The authors of this work also found that these films have a 10 to 20 Å layer of SiO$_2$. LPCVD silicon nitride films often have high tensile stress ($10^{10}$ dyne/cm$^2$). Increasing the concentration of DCS in the growth ambient can decrease the stress, but results in films with a considerable amount of excess silicon. Due to the high concentration of silicon containing species (primarily SiCl$_3$), gas phase nucleation is likely during the LPCVD of Si$_3$N$_4$, leading to a high concentration of particles formed in the gas.

### 13.7 Plasma Enhanced CVD of Dielectrics

In many applications, it is necessary to deposit films at very low substrate temperatures. The deposition of SiO$_2$ over aluminum and the deposition of Si$_3$N$_4$ capping layers over GaAs are two common examples. To accommodate these lower substrate temperatures, an alternate energy source must be applied to the gaseous and/or adsorbed molecules. While photo-enhanced deposition has been experimentally demonstrated, and even seen some limited use in production, the primary nonthermal energy source used to drive CVD reactions is the RF plasma. Plasma enhanced chemical vapor deposition (PECVD) systems have the added advantage of using ion bombardment of the surface to provide energy to the adspecies to allow them to diffuse further along the surface, without a high substrate temperature. As a result, the process is very good at filling small features. Chapter 10 introduced the concept of plasmas. The application of glow discharges to sputtering and etching have already been discussed. This section will discuss the use of plasmas to enhance the deposition rate in CVD processes. Since the deposition of insulating layers is of primary interest, only RF discharges need to be considered.

There are three basic types of PECVD systems as shown in Figure 13.17. In each system, the RF frequency chosen is normally less than 1 MHz, although PECVD oxide can be deposited at 13.56 MHz. The first PECVD systems were cold wall parallel plate reactors. Gasses are either injected at the edge or through an upper electrode showerhead, and exhausted through a port.
at the center, or the gas is injected at the center and exhausted around the edges. As the wafer diameter has increased, the low throughput and marginal uniformity of these systems has obviated their use for silicon IC production. Due to the small wafer diameters and the limited number of wafers per batch however, this style of reactor is often preferred for GaAs technologies.

For silicon IC manufacturing with large diameter wafers, the currently preferred technique is a parallel plate hot wall system. Similar in appearance to an LPCVD tube, the wafers are mounted vertically on conductive graphite electrodes of alternating polarity. The substrate temperature is controlled as in any furnace, although it is much cooler than it would be for a comparable LPCVD process. Although the throughput of this type of reactor is better than the parallel plate plasma reactor, it is much less than that of a standard LPCVD system.

Hot wall batch PECVD systems suffer from the same types of gas depletion/uniformity and particle problems as their thermal counterparts. For that reason, there has been a resurgence of interest in cold wall PECVD systems. To increase the throughput, a number of deposition stages may be put into a single vacuum system, or several single wafer chambers may be run in parallel with a robot arm to feed the chambers. One manufacturer does the deposition sequentially in five stages. Not only does this improve throughput, uniformity of close to 1% is seen as well [31].

To deposit high quality layers at low substrate temperatures, remote plasmas have recently been introduced. These reactors use a conventional or electron cyclotron resonance (ECR) plasma [32] to dissociate or crack one or more of the precursors. These sources were covered in Chapter 10. One application of ECR is to dissociate N₂ to form atomic nitrogen that readily reacts with silane to form Si₅N₅ with virtually no ion bombardment of the substrate. The silane can be introduced outside the plasma [33]. Because of the high reactivity of the atomic species, a large substrate temperature is not necessary to drive the reaction and obtain dense films [34]. Good silicon dioxide films have also been demonstrated at temperatures as low as 120°C [35].

The low pressure of an ECR plasma (about 0.1 torr) results in long mean free paths and, therefore, poor step coverage. If the system is designed to allow a significant amount of ion bombardment of the surface however, the deposited species will be continuously sputtered, allowing the fill of high aspect ratio features. One of the primary limitations of ECR deposition systems is the high concentration of particles produced in the plasma. This has been addressed recently by creating particle traps and/or particle absorbent chamber surfaces. To improve the low deposition rates of ECR, production systems can be fabricated with a large number of the remote plasma injectors in parallel in the same vacuum chamber.

The PECVD of silicon nitride has been used in GaAs implantation for many years. In silicon technologies, this process has also been used as a final passivation or scratch protect layer as one of the final steps of the process [36]. The process is run at 300 to 400°C using mixtures of a diluent gas such as Ar or He, SiH₄, and either NH₃ or N₂. Both hot wall and cold wall reactors have been used (Figure 13.17).

Figure 13.18 shows plots of the deposition rate, mass density, and atomic composition as a function of the fraction of ammonia in the gas flow [37]. The deposition rate is relatively insensitive to gas composition. The maximum in the density occurs at a Si/N ratio of 0.75, the correct value for stoichiometric silicon nitride. Increasing ammonia flow however, increases the hydrogen concentration, which at about 20%, is typical for PECVD Si₅N₅. Increasing the substrate temperature decreases the hydrogen content in the film. The use of N₂ rather than NH₃ also decreases the hydrogen content [38].

One of the major applications for deposited dielectrics is to form the insulator between metal interconnect levels in an IC. Particularly when aluminum is used for the metal, PECVD provides the
necessary low temperature deposition. A problem with PECVD nitrides in this application is the relatively large dielectric constant of $\text{Si}_3\text{N}_4$. When used as an insulator between two metal layers, this will result in a large node capacitance and therefore, a slower circuit speed. To improve circuit performance, the nitride can be replaced with a PECVD oxide. PECVD silicon dioxide processes can be run using silane and an oxidizer. $\text{O}_2$ can be used, but the reaction between silane and $\text{O}_2$ does not need a plasma to drive it. As a result, considerable homogeneous nucleation occurs in the inlet nozzles and in the gas above the wafer resulting in high particle counts and poor morphology. $\text{CO}_2$ can be used but $\text{N}_2\text{O}$ is the preferred oxidant to avoid carbon incorporation. It has also been reported that the addition of He as a diluent improves deposition uniformity and reproducibility [39].

Oxides deposited with this technique have high concentrations (1 to 10%) of hydrogen [40]. It is also common to find substantial amounts of water and nitrogen [41]. The exact composition depends critically on the chamber power and the gas flows. As shown in Figure 13.19, increasing the plasma power increases the deposition rate but also reduces the density [42]. As a result of the ease of silicon oxidation reaction, low plasma power densities provide large deposition rates. It has also been shown that if the power density of the plasma is large enough to ensure a sufficient concentration of atomic oxygen in the plasma to completely oxidize the silane, good films with a high dielectric strength are observed [43]. Post deposition high temperature bakes may be used to reduce the hydrogen content and densify the film. These bakes can also be used to control film stress [44] but often PECVD processes are chosen precisely because such high temperature steps cannot be tolerated.

One of the interesting features of PECVD films is that it is possible to change the composition of the films continuously from oxide to nitride by changing the gas flow. By adding increasing amounts of $\text{N}_2\text{O}$ to a mixture of $\text{SiH}_4$, $\text{NH}_3$, and He in a 13.56 MHz cold wall PECVD system, films were deposited whose refractive index varies smoothly from that of nitride to that of oxide [45]. This allows the technologically interesting possibility of stacks and graded composition films.

PECVD of silicon dioxide may also be done using a TEOS source [46],[47]. One of the motivations for this approach is the danger associated with silane use. At sufficiently large $\text{O}_2$ to TEOS inlet flow ratios, the residual carbon contamination in the deposited films can be made extremely small [48], and reasonable values of the index of refraction and dielectric constant [49] can be obtained. Substrate temperatures are well below 400°C. As with thermally deposited TEOS films, the stress of the layer can be controlled over a broad range [50] by a post deposition bake. Doped layers such as PSG and BPSG are also increasingly looking toward the use of organometallics such as tetrabutylphosphine and trimethylborane to reduce the use of hydrides.
13.8 Metal CVD

The last sections described the attributes of CVD: excellent step coverage and the potential for low substrate temperature depositions. One of the most serious step coverage concerns is that of metal as it goes into contacts. Particularly for deep submicron devices, the step coverage of sputter deposited films into increasingly high aspect ratio features has become unacceptable unless very high temperatures are used. Furthermore, in order to ensure coverage of the metal over the contact, sidewalls must be carefully tapered during the etch process. To accommodate this widening of the contact, typical metal lines must have caps (Figure 13.20). These caps can significantly reduce wiring density. Finally, the topology created by tapering the contact accumulates so that coincident contacts or “plugs” are not allowed. If, on the other hand, metal CVD is available, vertical contact structures can be used, filling the contact and decreasing the surface topology. No cap is required, and step coverage is much less of a concern. It is therefore highly desirable to develop processes for metal CVD. Of the various metals that have been attempted, the greatest success has been achieved with tungsten [51].

Much of the early work in tungsten CVD was done in standard horizontal LPCVD tubes [52],[53].
Tungsten was found not to adhere to the tube walls and so particulation was a serious problem. Even more of a concern however, was the fact that thin layers of tungsten could effectively block the IR radiation from the coils that heat the wafers. This effect resulted in poor uniformity and reproducibility. As a result of these problems, most tungsten CVD is now carried out in cold wall reactors [54]. Keeping the chamber walls below about 150°C is critical to the success of the process due to the high reactivity of the precursors. Tungsten sources include WCl₆ [55],[56]; W(CO)₆ [57],[58]; and WF₆ [59]. Of these, only WF₆ is a liquid at room temperature. The others are high vapor pressure solids. Most tungsten CVD is therefore run using WF₆ with a H₂ carrier gas. Deposition temperatures are usually less than 300°C.

The simplest type of tungsten deposition process is the blanket CVD of tungsten. Since blanket CVD W films do not adhere well to the oxide, a thin adhesion layer must first be deposited. One of the most commonly used layers is sputtered TiW, which is often used as a barrier metal under aluminum interconnect. TiN has also received considerable attention, particularly since recent demonstrations of the CVD of these layers [60],[61]. It has been shown however, that the deposition of W over TiN has a substantial initiation time, during which no film is formed unless SiH₄ is used during the growth initiation phase [62].

Although typical deposition rates are rather slow, the best step coverage for W CVD is obtained using WF₆ and H₂ [63]. This is a major concern since one of the primary applications for CVD tungsten films is the filling of high aspect ratio contacts. If the films have poor step coverage, a void will be left in the center of the contact as the top closes. Even a small taper of the contact will ensure complete filling (Figure 13.21). The overall reaction for this process is

\[ WF₆ + 3H₂ \rightarrow W + 6HF \]

The tungsten layer deposited in this manner may be left in place and used for the interconnect level, or it may be etched back leaving only the filled contacts or plugs. In the latter case, a sacrificial nitride layer may be used on top of the oxide to avoid roughening the surface of the oxide during the etchback (Figure 13.22).

To avoid the etchback altogether, it is also possible to deposit tungsten selectively in the exposed contact windows. Selective tungsten is expected to fill any aspect ratio structure, since the deposition occurs from the bottom of the structure up, rather than from the sides. This has the additional advantage of reducing the cost of the deposition, which is considerable. The problem with these processes is that maintaining selectivity is extremely difficult [64]. Deposits tend to form on the dielectric and may also creep up the walls of the contact.

Aside from tungsten CVD, there are two primary areas of metal CVD research: silicides for gate electrodes, local interconnect, and to serve as W adhesion layers, and alternate interconnect metals such as aluminum and copper to replace the use of sputtered AISi and AlCuSi altogether. In the former
category, the primary material is TiS "$2. TiS "$2 is deposited from a low pressure mixture of Si:H$_4$ and TiCl$_4$ in a H$_2$ carrier. Typical deposition temperatures are 600 to 750°C and deposition rates are several thousand Å/min, making the deposition process suitable for use with single wafer systems [65]. Selective deposition is also possible. This is attractive for a salicide replacement (see Chapter 16), since the very shallow junctions employed by deep submicron CMOS do not tolerate much silicon consumption. This process can even be run in standard LPCVD tubes [66], but as with tungsten the use of such systems for metal deposition leads to eventual problems with thermal uniformity. Furthermore, the process window for selective deposition is very narrow [67]. If the process is carried out in a PECVD reactor, the films deposited over silicon tend to be rough unless subsequently annealed at high temperature [68].

The two films that have received the most attention as potential replacements for sputtered aluminum are CVD copper and CVD aluminum. Because of its current use in IC interconnect, aluminum was one of the first films studied [69]. Due to the grain structure, electromigration [70] and junction spiking [71] are much less severe in CVD aluminum than in sputtered films. The most commonly used precursor is tri-isobutyl-aluminum (TIBA), because of the low carbon content in LPCVD deposited layers. Typical deposition temperatures are 200 to 300°C, and deposition rates are hundreds of Å/min. It is also possible to deposit these films selectively as well [72]. Of course, the deposition of the desired alloys AlSi and AlCuSi are much more difficult.

The most popular precursor for copper CVD are the Cu β-diketones. Deposition temperatures are typically 100 to 200°C [73]. Although the films have very low resistivity, there is currently no acceptable way to anisotropically etch copper. As a result, much of the copper CVD work has been directed toward selective deposition. Here the concept would be to deposit and pattern a thin seed layer, then selectively deposit the copper on top of that. To achieve good selectivity, the deposition temperature must be kept low and the surface must be very clean [74].

13.9 Summary

This chapter reviewed the basic chemistry and fluid mechanics of chemical vapor deposition. Although an accurate quantitative description of any reaction is highly process and reactor dependent, this introductory material can be used to develop a qualitative understanding of the deposition process. Atmospheric pressure CVD (APCVD) is only widely used for SiO$_2$ deposition. The major problem with these processes is particle formation either at the injector nozzles or through homogeneous nucleation. Low pressure CVD (LPCVD) of polysilicon and silicon nitride is widely used in silicon technologies. Standard furnace tube arrangements provide good uniformity and high throughput. To
improve the uniformity further, cross-flow reactors have been developed. The deposition of oxide can be done with silane and oxygen or nitrous oxide, but the currently preferred method is the thermal decomposition of TEOS. Plasma enhanced CVD is also popular when low deposition temperatures are required. These films tend to have poor stoichiometry and suffer from high etch rates unless subsequent anneals are done. To obtain high density PECVD films, after glow reactors have been developed. The CVD of metals is a new and important area that allows the fabrication of high density interconnect. Currently, tungsten is the most popular CVD metal system.

Problems

1. Repeat the calculations done in Example 13.1 if the reaction of interest is

\[ AB_2 \rightleftharpoons A + 2B \]

Assume that the equilibrium constant is unchanged.

2. Repeat the calculations done in Example 13.2 assuming that the gas is hydrogen \( (\eta = 30 \text{ gm/cm} \cdot \text{sec}) \).

3. Briefly describe the advantages and disadvantages of APCVD.

4. Assume that you wanted to deposit NaCl (table salt) on the wafer. What precursors might you use? What problems could you foresee?

5. A particular process is reaction rate limited at 700°C and the activation energy is 2 eV. At this temperature the deposition rate is 1000 Å/min. What would you guess that it would be at 800°C? If the measured deposition rate at 800°C is much less than this prediction, what might your conclusion be? How would you prove it?

6. A process is run in a standard horizontal LPCVD tube, where the wafers are loaded on edge in a standard slotted boat. What factors might explain a reduction in the deposition rate from the front of the tube to the back? From the edge of each wafer to the center? What would you do to try to improve the uniformity in each case?

7. A post-deposition anneal called a densification step is often used to reduce the etch rate of CVD SiO₂. The densification is typically run at 900–1000°C. The step is not normally done for PECVD films although they would benefit from the anneal. Explain why the process is not done for these films.

References

27. C. L. Ramiller and L. Yau, "Borophosphosilicate Glass for Low Temperature Reflow," *Semi-


The most critical use of a vacuum is to remove undesirable reactive gases from a process.

The term “vacuum” describes an environment in which the gas pressure is less than the ambient atmospheric pressure. Since the pressure results from the force exerted by gas molecules hitting a surface, the gas pressure is dependent on the species of the gas molecules and their temperature (energy). Typically, a gas consists of a mixture of species. For example, the Earth’s atmosphere consists of mostly nitrogen (78%) and oxygen (21%), with small amounts of other gases and varying amounts of water vapor and particulates. The amounts of these gases and vapors are given by their “partial pressures.” Under standard atmospheric conditions, the total gas pressure is 760 Torr, where 1 Torr is the pressure required to support a column of mercury 1 mm in height in a manometer tube.

In PVD processing, one of the principal uses of the vacuum environment is to control the amount of reactive (contaminant) gas in the deposition chamber. The presence of reactive gases is generally undesirable except when deliberately depositing a film of compound material. The amount of reactive gas that can be tolerated depends on the desired purity of the deposited film material. The Figure shows the flux of gas that impinges on a surface as a function of partial pressure of the gas. Also shown is the time to form a monolayer of gas on the surface if all the molecules striking the surface stick to the surface. Note that at $10^{-6}$ Torr (about $10^{-9}$ atmospheres of pressure) about one monolayer of gas will impinge on a surface per second. Taking 10 Å per second as a typical vacuum deposition rate for titanium and a partial pressure of oxygen of $10^{-6}$ Torr, this would mean that about 1/3 of the deposited material would be TiO if all the impinging oxygen reacted with the depositing titanium. This may or may not be acceptable for a given application.

The gas pressure is also important to collisions in the gas phase. If the mean free path for collision is greater than the vaporization source-to-substrate distance, the deposition will be “line of sight.” If there is scattering in the gas phase, the incident direction of the “adatom” flux will be randomized. The direction of the incoming flux is important to film growth and surface coverage. The Figure shows the mean free path for collision in the gas as a function of pressure. Note that a relatively poor vacuum of $10^{-3}$ Torr results in a mean free path of 1 meter.

Gases and vapors in the deposition chamber originate from a variety of sources, including:

- Residual (unpumped) atmospheric gases.
- Desorption from chamber surfaces, primarily water vapor.
- Real leaks—paths to the external ambient.
• Virtual leaks—paths to internal trapped volumes.
• “Outgassing” of materials, e.g., water vapor from polymers.
• Vaporization of high-vapor pressure materials in the system.
• Permeation of gases through materials, e.g., water vapor through polymer O-rings.
• “Brought-in” with source materials, substrates and fixtures—the amount depends on the surface preparation techniques used to clean and handle the surfaces and materials.

These sources of gases and vapors are minimized by proper vacuum engineering, proper operation and maintenance of the deposition system, and proper cleaning and handling of surfaces and materials that go into the deposition chamber.

The principal contaminant in a good vacuum system is usually water vapor that is adsorbed on surfaces. This water vapor is often a processing variable that must be controlled in order to have a reproducible deposition process. PVD deposition chambers should be “let-up” to atmosphere using dry gas and should be open to the ambient for as short a time as possible to reduce water adsorption on surfaces. Water vapor can be kept from condensing on interior surfaces of the system by keeping the surfaces warm while the system is opened to the ambient atmosphere. Surfaces to be placed in the deposition system should be dried using a drying agent, such as anhydrous alcohol, and stored in a desiccated chamber. Materials, such as polymers, which absorb water should be vacuum baked and outgassed before being placed in the deposition chamber.

Water vapor does not easily desorb from surfaces in the vacuum system, but, in production processing, several techniques can be used to aid in its release from surfaces. These include:

• Flushing the deposition chamber with hot, ultra-dry gas during the pumpdown process; i.e., a pump-flush-pump cycle.
• Establishing a plasma discharge in the chamber during the pumpdown process; this produces “ion scrubbing” of the surfaces in the chamber.

Water vapor is strongly adsorbed on surfaces, and it is not easily pumped from the deposition chamber, particularly if the water molecules must undergo numerous collisions with surfaces before the vapor enters the vacuum pump. The “pumping” of water vapor in the deposition chamber can be aided by using cold surfaces in the chamber that “freeze-out” the water. These large-area surfaces are cooled to about -150°C using refrigerants. At -150°C the vapor pressure of water is about 10^-14 Torr and the cold ice does not provide a significant source of water vapor.

REFERENCES

Reproduced with permission of the Society of Vacuum Coaters from the SVC Educational Guides to Vacuum Deposition Technology.

The 1999 edition of these Guides contains 90 individual, stand-alone, two-page guides on different aspects of the equipment and technology associated with vacuum coating. Two Guides will be published each month in this magazine. The complete set of Guides can be ordered from the Society of Vacuum Coaters for $50 (plus $10 shipping and handling fee for airmail shipment outside the USA) by contacting SVC at <svcinfo@svc.org> or by calling 505-856-7188. The complete list of Guide titles can be found in SVC’s ad elsewhere in this magazine.

---

Rigaku
Announces An Expansion of the Superseal Product Line
Made in the USA

More Seal for your Cents Starting at just $395

- 314" wide, 450mm
- 72" slit, 1800mm dia.
- 60" wide, 400mm
- 15" wide, 300mm dia.
- High Vacuum
- Medium Vacuum
- Reduced Vacuum
- Roughing Vacuum

Rigaku Vacuum Products • 199 Rosewood Drive • Danvers, MA 01923
(Phone) 978-777-2446 • (Fax) 978-777-3594 • www.rigaku.com

For Free Literature, Circle 39

Vacuum Technology & Coating • Jan/Feb 2000
13
Deposition of atoms freed from a bombarded target works best in a vacuum.

When an atom of a surface \( M_{\text{target}} \) is struck by an energetic particle \( M_{\text{incident}} \) of roughly the same size, the energy and the momentum that are transferred depend on the relative masses \( M_i \) to \( M_t \) of the particles and the angle of collision with respect to a line connecting their center-of-masses, as is shown in Figure 1. The maximum energy is transferred when the particles have the same masses and the collision path is along a line joining their centers. Glancing angles and mass differences give lower energy transfer. The "recoil angle" of the struck particle from the collision depends on the angle of collision. Whether the bombarding particle reflects from the surface or continues into the surface also depends on the angle of collision.

![Diagram](image)

*Figure 1. Interaction of energetic particles of atomic dimensions with a solid surface.*
Elastic Collision (Hard Spheres)

\[ E_f = \frac{4M_i M_f}{(M_i + M_f)^2} \cos^2 \theta \]

\[ \left( \frac{E_f}{E_i} \right)_{\text{max}} \text{ when } M_i = M_f \]

Figure 2: Energy transfer during collision of particles.

If the surface atom that has been struck attains enough energy, it will strike other atoms in the near-surface region and a "collision cascade" will develop. Multiple collisions can result in some momentum being directed back toward the surface; and if the energy attained by a surface atom that is struck from below is sufficient, it can be physically ejected from the surface, i.e., sputtered. Generally, sputtered particles are neutral, but sputtering from compound or alloy materials can yield ions depending on the relative electronegativities of the materials in the surface. Most of the energy that is transferred by the bombarding particle appears as heat in the near-surface region.

The collisions in the collision cascade displace atoms from their lattice positions and create lattice defects. These lattice defects can trap gaseous atoms injected into the surface by the bombardment. This "implantation" or "gas trapping" can easily approach several atomic percent in the near-surface region when energetic bombarding species are used. The displaced atoms can be "stuffed" into interstitial lattice positions and can create compressive stresses in the near-surface region. Figure 2 summarizes the effects that take place in the near-surface region of a sputtering target.

Energetic particles that are reflected from the surface typically are neutral, but they can be charged positive or negative depending on the relative electronegativities of the surface and the reflected particle. Since the energy lost depends on the relative masses of the incident and target particles, the amount of energy loss at a particular reflection angle can be used to determine the mass of the surface atom. This effect is used in the surface elemental analytical technique of "Ion Scattering Spectrometry" (ISS).

At all but the lowest bombarding energies, the flux of atoms that are
sputtered from the surface leave the surface with a cosine distribution. They typically have average kinetic energies of about 10 times that of thermally vaporized atoms and have a high energy “tail” in the energy distribution that can be several tens of eV. For example, copper (M = 64) evaporated at 1500°C will have an average atom energy of 0.2 eV, while 600 eV mercury-sputtered copper has an average energy of 4 eV, which is the equivalent of an evaporation temperature of about 40,000°C.

The “sputtering yield” is the number of surface atoms that are sputtered for each incident energetic bombarding particle. The sputtering yield depends on the relative masses, the angle-of-incidence of the bombarding species, and the chemical bond strength of the surface atoms. The most common inert gas used for sputtering is argon (M = 40). Figure 3 shows the relative sputtering yields from several materials sputtered with argon ions at various energies. As the angle-of-incidence of the bombarding particles becomes off-normal, the sputtering yield can increase two to three times, up to a point where the bombarding particles transfer little momentum because of the high collision angle, and the sputtering yield drops off rapidly. Under these conditions, most of the bombarding species are reflected from the surface.

The apparent sputtering yield can be affected by the surface topography since, in sputtering a rough surface, some of the sputtered particles are “forward sputtered” and redeposited on the surface. This means that the sputtering yield can decrease with time if the surface roughens with use. At high gas pressures, some of the sputtered material will be scattered back (“backscattered”) to the surface.

Since the sputtering process removes each solid-surface atomic layer consecutively, if there is no diffusion, the composition of the vapor flux leaving the surface is the same as the composition of the bulk of the material being sputtered. This allows the sputter-vaporization of alloy materials that cannot be thermally evaporated because of greatly differing vapor pressures of the constituents.

Often, surfaces to be sputtered have a surface layer composed of a reacted material, such as an oxide or a nitride. Since the chemical bonding of the compound materials is stronger than that of the elemental material, the sputtering yield is initially low until the surface layer is removed. Also, if reactive gases are present, they can continuously “poison” the target surface, giving low sputtering yields.

Particles that are sputtered or reflected from the surface at low gas pressures will travel in a line-of-sight path, with no collisions, to condense or bombard a surface, such as the substrate. If the gas pressure is higher, collisions can take place, thus reducing the energy of the particles and scattering them from a line-of-sight path. If there are enough collisions, the energetic particles are “thermalized” to the energy of the ambient gas. Energetic gaseous particles can bombard the surface of the growing film with enough energy to affect the film formation process and the properties of the deposited film material. Thus it makes a difference in film properties whether the sputter deposition is done at a low gas pressure (<3 mTorr) or a higher gas pressure (>5 mTorr).

Reproduced with permission of the Society of Vacuum Coaters from the SVC Educational Guides to Vacuum Deposition Technology.

The 1999 edition of these Guides contains 90 individual, stand-alone, two-page guides on different aspects of the equipment and technology associated with vacuum coating. Two Guides will be published each month in this magazine. The complete set of Guides can be ordered from the Society of Vacuum Coaters for $50 (plus $10 shipping and handling fee for airmail shipment outside the USA) by contacting SVC at svcinfo@svc.org or by calling 508-856-7188. The complete list of Guide titles can be found in SVC’s ad elsewhere in this magazine.
Appendix D

Gas Systems

In this appendix we consider some basic ideas of gas systems that are common to the processes of Chapters 5, 6, 7, 9, 10, and 12. These systems distribute gases under controlled pressure and flow rates. Our interests will center on systems working near atmospheric pressure; space limitations prevent a discussion of systems in the vacuum range, where pressures are at least a few decades below atmospheric value.

D.1 Basic Concepts

Consider a volume $V$ in a tube through which gas is flowing at a constant rate and temperature, as in Figure D–1(a). The pressure difference across the volume is $(p_2 - p_1)$, and the average pressure is $\langle p \rangle$. The general gas law teaches that

$$\langle p \rangle V = nkT$$  \hspace{1cm} (D.1-1)

where

$n = \text{number of gas molecules in volume } V$

$k = \text{Boltzmann's constant} = 1.38 \times 10^{-23} \text{ J/K} = 8.62 \times 10^{-5} \text{ eV/K}$

$T = \text{absolute temperature in Kelvins (K)}$

Unfortunately, pressure units are not standardized. For pressures below atmospheric value the deprecated unit torr is still in common use. It is equal to 1 mm of Hg within two parts in $10^7$. The millibar (mbar) and Pascal (Pa = N/m$^2$) also are common. Pounds/in.$^2$ (PSI) and atmosphere (atm.) are
Thus the throughput may be defined as the molecular flow rate at standard temperature of 25°C (300 K). Typical units are torr-l/s.

The small tube in Figure D-1(a) impedes the free flow of gas, and its effect on the throughput is characterized by a parameter

$$C = \text{conductance} = \frac{Q}{p_1 - p_2} \quad (D.1-8)$$

Note that this equation is analogous to that for the electrical conductance of Figure D.1(b), viz

$$G = \frac{I}{V_1 - V_2} \quad (D.1-9)$$

The electrical quantities in Eq. (D.1-9) are given here in roman type to avoid confusion with their gas analogs. Note the analogs: C and G, Q and I, and p and V.

### D.2 Conductance Calculations

Gas at near atmospheric pressure typically behaves like a fluid in flowing through a tube and is said to be in the **viscous range**. For practical purposes this is defined by [6]

$$L_m/d < 0.01, \quad \text{Viscous Range} \quad (D.2-1)$$

where

$$L_m = \text{mean free path of the gas}$$

= average distance traveled by a molecule between collisions with other molecules

$$= \frac{5 \times 10^{-3}}{p^*}, \quad \text{cm} \quad (D.2-2)$$

where $p^*$ is the pressure in torr.

In this viscous domain, molecules in a pipe are more likely to collide with each other than with the pipe walls; hence, they will drift along the pipe under the influence of a pressure difference between the ends. Consider what pressure a 2-in.-diameter pipe drops out of the viscous range. From the last two equations we may reduce the viscous domain criterion to

$$p^* > \frac{5 \times 10^{-3}}{10^{-2}d} = \frac{5 \times 10^{-3}}{10^{-2}(2.54)} = 0.0984\text{torr}$$
The conductance of a cylindrical pipe or tube when it is in the viscous domain can be derived from Poiseuille's equation for viscous flow, and in the commonly used bastard units is [6]

$$ C = \frac{180 d^3 \langle p^* \rangle}{L} \text{ l/s (Viscous Range)} \quad (D.2.3) $$

where

$$ C = \text{conductance in l/s of a cylindrical pipe } L \text{ cm long and of diameter } d \text{ in cm, for air at room temperature} $$

and \( \langle p^* \rangle = \text{average pressure in the pipe in torr} \). By virtue of the analog between Eqs. (D.1-8) and (D.1-9) it is apparent that the total conductance of two tubes of different diameters in series is the product over the sum of the individual conductances, with the smaller controlling.

A circular hole or orifice between two parts of a gas system, or a change in effective tube diameter, has a nonzero value of conductance, and hence can support a pressure difference. Use was made of this concept in Section 8.4.3. It also is the basis for valve operation: the lowered conductance can maintain a pressure difference while reducing the flow rate.

In the transition range, where

$$ 0.01 \leq \frac{L_m}{d} \leq 1 \quad (\text{Transition Range}) \quad (D.2.4) $$

and for a pressure difference of at least one decade between the two sides of the orifice, the conductance is given by [6]

$$ C = 15 A, \quad \text{l/s} \quad (D.2.5) $$

where \( A = \text{area of the circular hole in cm}^2 \). For example, a 1-in.-diameter orifice has a conductance of 76 l/s.

At typical high vacuum pressures of several decades below atmospheric value, gas behavior changes markedly and is in the molecular range, which is defined by [6]

$$ \frac{L_m}{d} > 1.0, \quad (\text{Molecular Range}) \quad (D.2.6) $$

where \( d \) in the case of a cylindrical tube is its diameter and \( L_m \) is the mean free path of the gas molecules, both expressed in the same units. Because \( L_m \) exceeds the vessel size in this range, a molecule has a greater probability of colliding with the vessel walls than with another molecule. Under this condition, molecules will not drift along a pipe under the influence of a pressure difference between the pipe ends, because this mechanism depends on intermolecular collision. Molecular motion here is determined solely by thermal agitation, so the magnitude and direction of the molecular velocities are probabilistic. A pressure difference cannot force molecules along a pipe; they must move through the tube by their own thermal motion.

In this range the cylindrical pipe conductance becomes [6]:

$$ C = \frac{12.2 d^3}{L} \quad (\text{Cylindrical Pipe}) \quad (D.2.7) $$

For the circular aperture

$$ C = 11.7 A \quad (\text{Circular Aperture}) \quad (D.2.8) $$

where, as before, \( C \) is in l/s, \( L \) and \( d \) are in cm, and the aperture area \( A \) is in cm².

### D.3 Gas Supply Systems

Consider some supply systems for processing gases and how gases from different sources are mixed. Emphasis will be on those systems where sources are gases or liquids at room temperature. Among the former are oxygen, nitrogen, argon, silane (SiH₄), arsenic (AsH₃), and phosphine (PH₃). Typical liquid sources at room temperature are phosphorus oxychloride (POCl₃ or pockle), silicon tetrachloride (SiCl₄), boron tribromide (BB₃), and trichloroethane (TCA). These liquids have nonzero vapor pressures near room temperature and hence will provide a vapor or gas phase when in a confining vessel. This vapor then may be used for processing.

#### D.3.1 Supply Tanks

Gases of desired composition and purity in high-pressure tanks or cylinders are available from vendors. The tanks come equipped with a tank valve that prevents gas escape. The user connects a pressure-regulator/pressure-gage/valve combination to this for metering and flow control (Figure D–2). Tanks and regulators are available with different fittings and threads, some being left-handed, to minimize the chance of interconnecting dangerous and explosive gas combinations. For example, different sizes and threads prevent connection of an oxygen regulator, with some carryover oxygen, to a tank of hydrogen, a gas combination that might explode. Therefore, the user must match threaded regulators and gages to the tanks properly.
Table D–1. Fitting Numbers for Certain Gases

<table>
<thead>
<tr>
<th>CGA #</th>
<th>Gas</th>
</tr>
</thead>
<tbody>
<tr>
<td>320</td>
<td>CO₂</td>
</tr>
<tr>
<td>350</td>
<td>AsH₃, B₂H₆, H₂, PH₃, SiH₄ (high pressure)</td>
</tr>
<tr>
<td>510</td>
<td>SiH₄ (low pressure)</td>
</tr>
<tr>
<td>540</td>
<td>O₂</td>
</tr>
<tr>
<td>580</td>
<td>Ar, He, N₂</td>
</tr>
<tr>
<td>590</td>
<td>Industrial air</td>
</tr>
<tr>
<td>660</td>
<td>Freon 12</td>
</tr>
</tbody>
</table>

Courtesy of Matheson Gas Products, Inc., Secaucus, NJ [7].

Fortunately, standards exist for the threaded fittings, which are identified by a CGA (Compressed Gas Association) number. Numbers for a few typical gases are given in Table D–1. More complete lists are available in vendors' literature [7, 8]. Note that gases with the same fitting number are compatible and no danger arises if they are mixed.

Processing tanks come in several sizes, and the pressure gage often has a scale calibrated in cubic feet of gas [7, 8]. This calibration gives the gas volume at standard conditions of 70°F and 1 atmosphere, and will far exceed the physical volume of the tank. The relationship between the two volumes may be calculated with help from the gas law.

Say a full tank of oxygen has an indicated gage pressure of 2000 PSIG, corresponding to a gas volume of 220 ft³ under standard conditions of 70°F and 14.7 PSI. What is the volume of the tank?

From the gas law we have that

\[ P_T V_T = P_S V_S \]

where \( T \) refers to the tank and \( S \) to standard conditions. Note that \( P_T \) is the gage pressure plus atmospheric pressure. Then

\[ V_T = \frac{P_S V_S}{P_T} = \frac{14.7(220)}{(2000 + 14.7)} = 1.6 \text{ ft}^3 \]

This tank closely matches a Matheson 1-A size of 1.55 ft³ rated internal volume, but note that tank sizes vary among vendors.

Improper use can damage a tank regulator, so manufacturers' instructions should be followed exactly. Special care also is required in handling high-pressure gas cylinders, since they can act like bombs. They must not fall or collide against each other; they never should be left free-standing without support. Safety requires that they be strapped or chained to a solid support such as a table, wall, or gas cabinet. Also, great care must be exercised in moving cylinders to avoid bumping them. They should be moved on a cart to which they are firmly fastened.

D.3.2 Single Tank

Consider a gas cylinder containing a gas \( G_1 \) flowing at pressure \( p_1 \) and temperature \( T_1 \) as shown in Figure D–3(a). (Symbols are defined in Figure D–4.) The volumetric flow rate can be set manually and read on a Rotameter

![Figure D–3. Representative gas supply systems. (a) Single gas supplied from a tank. (b) Two gases supplied from separate tanks and mixed. (c) Two mixed gases supplied from a single tank. (d) A tank-supplied carrier gas mixed in a bubbler with the vapor from a liquid source.](image-url)
unchanged, and the total volumetric flow rate to the diffusion tube is their sum \((F_{n1} + F_{n2})\).

Second, consider the pressures. Dalton’s law requires that \(p_i\) in the tube be the sum of the partial pressures of the two gases less any pressure drops in the tubing. If the sum \(p_i\) is greater than \(p_a\), the pressure difference must be absorbed by the tube of small diameter in the end cap. The condition that \(p_i > p_a\) is desirable because it prevents backflow from atmosphere into the tube.

### D.3.4 Single Tank, Mixed Gases

A number of semiconductor processing gases are toxic and/or pyrophoric (can ignite spontaneously on contact with air). Silane, for example, is pyrophoric if it mixes with air in a volume ratio of between 4 and 96\% of silane. Hence, for safety reasons it is common practice to purchase tanks of, say, 2\% silane in an inert gas such as nitrogen or argon. Then, even if this mixture is inadvertently released into air, the percentage of silane will remain below the 4\% danger margin.

Such a mixed-gas system is shown in Figure D–3(c). Dalton teaches that each component gas behaves as if it alone occupies the total volume \(V\). Using \(p_1\) and \(p_2\) as the partial pressures of the two gases, we have

\[
p_1V_1 = n_1kT \quad \text{and} \quad p_2V_2 = n_2kT
\]

Let

\[
r = \frac{n_1}{n_2} \quad \text{ratio of gas } G_1 \text{ to gas } G_2
\]

Adding, we get

\[
(p_1 + p_2)V = (n_1 + n_2)kT
\]

But the sum of the partial pressures is the total pressure \(p\), which is indicated by gages, so

\[
pV = (n_1 + n_2)kT \left(1 + \frac{1}{r}\right)n_1kT
\]

If volume \(V\) is flowing in time \(t\)

\[
pF_v = \left(1 + \frac{1}{r}\right)F_{n1}kT
\]
or the molecular flow rate of the lesser component is

\[ F_{n1} = \frac{p F_v}{(1 + 1/r) kT} \]  \hspace{1cm} (D.3-4)

Note again that \( F_{n1} \) is set by adjusting the total \( F_v \), which can be determined from a flowmeter.

### D.3.5 Vapor from a Liquid Source; Bubbler

Figure D–3(d) shows a situation in which the tank-supplied gas serves as a carrier component \( G_2 \) for a vapor \( G_1 \) coming from a liquid source in a bubbler (Figure 5–15). Gas temperature in the vessel is \( T \); the liquid source temperature is \( T_L \). Then within the volume \( V \) of the vessel we can write for the carrier gas \( G_2 \ :

\[ p_2 V = n_2 kT \]

and for the liquid source's vapor \( G_1 \)

\[ p_1 V = n_1 kT \]

Note that \( T \) is the same for both gases (and so is \( V \) by virtue of Dalton's law), but \( p_1 \) is a function of the liquid-phase source temperature, \( T_L \), and is obtained from vapor pressure tables or curves for the source in question—at e.g., Appendix F.6.8. The tank regulator determines \( p_2 \).

Dividing the equations, and dividing again by time \( t \), we have for the molecular ratio of the source gas to the carrier

\[ r = \frac{n_1}{n_2} = \frac{F_{n1}}{F_{n2}} = \frac{p_1}{p_2} \]  \hspace{1cm} (D.3-5)

These equations can be combined with those of the earlier sections to solve flow problems in diffusion, chemical vapor deposition, and other semiconductor processes.

Further information on gas supplies are given in appropriate sections throughout the text and references [9].

### D.4 Gas Distribution Systems

Some of the gases used in silicon processing are hazardous and impose special conditions on the distribution system that transports them from supply to process chamber. A system typical of atmospheric-pressure epitaxy is illustrated in Figure D–5. The additional components needed for a liquid source such as SiCl₄ are shown in Figure D–3(d).

Consider a few of the features in the figure. Each gas is supplied from a tank with a shutoff valve. Both p- and n-type dopants are shown for versatility. Silane and the dopant supplies are diluted with an appropriate gas, either inert or compatible with the system, to keep concentrations down to a safe level. The tanks are enclosed in a gas cabinet that is connected to the vent line to prevent buildup of gas pressure due to leaks in the tanks and valves. As an added safety measure, facilities may be provided to purge the gas cabinet with CO₂ or another fire extinguishing gas (e.g., Halon) in the event of fire. Appropriate temperature and/or smoke sensors would be provided to turn on the extinguisher supply if needed.

Pressure regulators and flow meters are provided for checking and setting the flow rates of the several gases individually. The check valves are one-way devices to prevent backflow of gases from the vent or reactor lines toward the flow meters and supply tanks. This eliminates cross-contamination among the supply lines.
A FAIRY TALE: VAPOR PRESSURE DATA OF THE ELEMENTS
by
R.E. Honig (retired)
Haverford, PA 19041-1095

Once upon a time, some four decades ago, Industrial Laboratories LOVED doing RESEARCH, among them RCA LABORATORIES - now defunct, alas! Still, practical results were expected, not tomorrow, but eventually - perhaps ten years down the line. At that time, High-Purity Semiconductor Materials were of great interest, and in this context, I had become interested in their Vaporization Characteristics: Vapor Pressures, Transition Temperatures and Energies, and Clustering. It soon became apparent that most data in this field were quite ancient, dating back to pre-WW II collections, while clustering of the elements of interest had never been observed until we did our mass spectromatic studies in the early fifties.

Thus, I decided, in 1957, to take a little time, on the side, to collect and evaluate all recent data available and then present them graphically in the most suitable format. This turned out not to be the classical \( \log p \ vs. \ T \) presentation which compresses the high-temperature end so much that it is not feasible to show more than a few elements on the same sheet. Instead, I adopted the \( \log p \ vs. \ T \) format which avoids that difficulty. A thorough library search yielded a large volume of recent data that had to be selected, treated with the help of a 1957 vintage computer, and plotted point-by-point on large graph sheets. The special log-log grid used first had to be carefully drawn in by hand, line by line, in our Drafting Room. Vapor pressure data and transition points selected for a total of 57 solid and liquid elements could be accommodated on two large charts in the 1957 edition (1) which became an instantaneous success. Three years later, a similar collection, evaluation, and presentation was made and published by the author and H.O. Hook for the gaseous elements and some of the more common molecular gases (2). Since by 1962 a considerable volume of new information had become available for many additional elements, a second edition covering a total of 79 solid and liquid elements was published (3). Finally, a third edition was undertaken in 1969 by Honig and Kramer to include all the latest, up-to-date information for a total of 81 elements (4). This publication presented the Vapor Pressure Data in tabular and graphical form, as well as Tables of Melting and Vaporization Energies. For about three decades, public interest in this information continued to be very strong, causing RCA Laboratories to distribute many thousands of reprints.

VAPOR PRESSURE CURVES OF THE ELEMENTS

Temperature (°C) vs. Vapor Pressure (mm Hg)

BOILING POINT
Week 8

Etching
Chapter 10
Etching

Chapter 1 showed the need for selective etching, whereby an etchant attacks one material, such as SiO$_2$, while leaving relatively unaffected adjacent materials, such as Si and Al. Etching processes are discussed in greater detail in this chapter.

10.1 Types; Processes
Etches may be classified by their physical state: liquid, dry plasma, or vapor. The E-gas used in chemical vapor deposition (Section 9.4) is an example of a vapor phase etchant.

Another classification is based on the manner in which the etch attacks monocrystal materials: in isotropic etching the action is independent of the crystal orientation, and hence of direction; in anisotropic etching different crystal planes etch at different rates. We shall see some practical examples of anisotropic etching (e.g., in fabricating V-shaped grooves for VMOS devices).

All etching processes, irrespective of the physical state in which they take place, involve three basic events: (1) movement of the etching species to the surface to be etched, (2) chemical reaction to form a compound that is soluble in the surrounding medium, and (3) movement of the by-products away from the etched region, allowing fresh etchant to reach the surface. Both (1) and (3) usually are referred to as diffusion, although convection may be present, too. The slowest of these processes primarily determines the etch rate, which thus may be diffusion- or chemical-reaction-limited. Most liquid etching processes used commercially in silicon technology are reaction-limited.

Different chemistries are required for selectively etching the materials used in silicon technology. Many sources are available in the literature for specific applications [1, 2, 3].
10.2 Wet Etching

In wet etching the masked substrate contacts the actual etching agents in liquid solution. Two of the most common components in electronic grade etch come diluted with water: hydrofluoric acid (HF) is 49% and nitric acid (HNO₃) is 70% in H₂O. These give etch rates too high for good control in many applications, so further dilution, to moderate the etching action, is necessary. Water may serve as the diluent or diluting agent, but sometimes acetic acid (CH₃COOH) will be used with HNO₃ because it buffers against HNO₃ depletion.

In order to remove depleted solution and by-products from the substrate surface and replenish the etchant, the system is agitated gently, often in a circular swirling action, or N₂ may be bubbled through the mixture. Machines that do this automatically are available commercially. Alternatively, the etchant may be sprayed onto the wafer.

Most Si technology etches are acid-based, so the etching tank effluent is acidic, and requires special disposal precautions. For the usual acids—such as nitric, acetic, and sulfuric—neutralization in a catch basin before disposal is adequate. On the other hand, HF requires special treatment and must not be discharged into a sewer system.

10.3 Oxide Etching

To make windows in SiO₂ without affecting the underlying Si, a selective etch is needed. Hydrofluoric acid (HF) attacks SiO₂ but leaves Si unaffected at room temperature, so HF is the basis of SiO₂ etches. The overall etching reaction is [4]

\[
\text{SiO}_2 + 6 \text{HF} + 6 \text{H}_2\text{O} \rightarrow \text{SiO}_2 + 6 \text{H}_3\text{O}^+ + 6 \text{F}^- \quad (10.3-1)
\]

\[
\rightarrow \text{H}_2\text{SiF}_6 + 8 \text{H}_2\text{O} \quad (10.3-2)
\]

The by-product H₂SiF₆ (fluorsilicic acid) is water-soluble and can move away from the region of chemical reaction. Equation (10.3-1) shows production of the hydronium ion (H₃O⁺), which really is the basis for the acid action [4, 5]. The etch rate may be moderated by adding more water because it lowers the hydronium ion concentration.

During etching, HF tends to deplete rapidly so ammonium fluoride (NH₄F) is added to form a buffered solution, typical proportions being 10NH₄F:1HF. Since NH₄F is a solid, it is used in a water solution; a typical mix would be 6(40%NH₄F):1(49%HF). The etch rate for this solution at room temperature is around 1600 Å/min, too fast to control depth precisely by adjusting the etch time. But oxide thickness versus oxidation time is well controlled; hence the thin gate oxide of an MOS transistor is produced by etching away a thick oxide completely and thermally regrowing a new oxide of the desired thickness, rather than by partially etching away the thick oxide.

A surfactant or wetting agent may be added to the solution to obtain a more uniform etch over the entire area. Typical are Dow Chemical Company Triton-XR or 3M Company FG-95. The usual concentration is very small: 1 part of surfactant to roughly 500 parts of etchant solution.

Fast-acting etches tend to remove photoresist (PR) from the oxide. This effect can be reduced by spinning a layer of HMDS (hexamethyldisilazane) onto the oxide and drying it before the PR is applied (Section 11.8).

Even though HF stops etching when it reaches Si, there is no visible indication that the SiO₂ etch is completed. One can make a good estimate by knowing the etch rate, oxide thickness, and etch time, but a direct visual check can be made after the wafer is removed from the solution, and rinsed in water. Silicon is hydrophobic and is wetted by neither water nor HF, while both wet SiO₂ and PR. Thus the etch is completed if the windows are water-free. Usually the wafer is returned to the etch for another half-minute or so (to insure clean window edges), rinsed, and dried with dry N₂.

Caution: Remember—HF attacks the human body and, in contrast to other common acids, does not cause pain on contact. Proper gloves should be worn, HF fumes should be avoided, and proper eye protection should be in place. All operations involving HF must be performed in a fume hood. Remember that HF attacks glass; use plastic or Teflon beakers.

The etch rate is affected by strength of the etchant, temperature, and SiO₂ doping. Boron-doped oxide etches slowly, phosphorus-doped, very rapidly. The latter fact must be kept in mind when contact windows are etched in the oxide after a phosphorus diffusion. A 5H₂O:1 buffered HF etch is typical for B-doped oxides, while more water is added for P-doped—up to 20H₂O:1 buffered HF for heavy doping, such as in emitters.

Steam-grown oxides are less dense than dry-grown oxides and etch faster. The etch rate for B-doped oxide may be increased by a pre-etch immersion in HBF₄ (fluoroboric acid). Apparently this treatment modifies the glass structure [6].

10.4 Isotopic Si Etch

The isotropic etching of silicon is based on a two-step process. First, the Si is oxidized to SiO₂ by a strong oxidizer, usually HNO₃:

\[
3 \text{Si} + 4 \text{HNO}_3 \rightarrow 3 \text{SiO}_2 + (\text{by-products}) \quad (10.4-1)
\]
Second, $\text{SiO}_2$ is dissolved by HF into a water-soluble compound, as described in the last section.

$$3 \text{SiO}_2 + 18 \text{HF} + \text{(by-products)} \rightarrow 3 \text{H}_2\text{SiF}_6 + 8 \text{H}_2\text{O} + 4 \text{NO}$$  

While we think of the reaction as a two-step sequence, actually the two reactions take place simultaneously, giving the overall reaction

$$3 \text{Si} + 4 \text{HNO}_3 + 18 \text{HF} \rightarrow 3 \text{H}_2\text{SiF}_6 + 8 \text{H}_2\text{O} + 4 \text{NO}$$

The etch rates of the $\text{HNO}_3$/HF combination can be high — the maximum at room temperature being roughly 800 $\mu$m/min, for the ratio $4.5\text{HNO}_3:5.5\text{HF}$ [6]. This value is too high for good control. Changes in the proportions in either direction slow the rate. For example, $7\text{HNO}_3:3\text{HF}$ drops the rate to roughly 120 $\mu$m/min because the reaction tends toward HF limitation [2].

The etch rate is moderated more often by adding water or $\text{CH}_3\text{COOH}$ (acetic acid), which act primarily as diluents (although the acid also provides buffering against loss of $\text{HNO}_3$, and in sufficient quantity may be used to stop the etching action completely). Typical component ratios run in the range $5\text{HNO}_3:3\text{CH}_3\text{COOH}:3\text{HF}$ to $7\text{HNO}_3:7\text{CH}_3\text{COOH}:1\text{HF}$. Curves of isetch-rate versus content are available in many references [4, 7].

PolySi often is used for interconnects and gates in MOS devices. Since polySi presents all possible crystal orientations to the etchant, an isotropic etch should be used. A typical formulation is $10\text{HNO}_3:1\text{HF}:10\text{H}_2\text{O}$.

10.5 Anisotropic Si Etches

The three principal silicon lattice planes have different combinations of atomic density and number of bonds normal to the plane (Table 2–1). Some etches sense these differences and attack the planes at different rates, giving anisotropic etching. Theoretical reasons for this selective action are meager, but a plane that combines high atomic density and a small number of bonds etches more slowly. On this basis we expect etch rates to decrease from $\langle 100 \rangle$ to $\langle 110 \rangle$ to $\langle 111 \rangle$ planes.

Anisotropic Si etches usually are not based on $\text{HNO}_3$/HF mixtures but rather on $\text{NaOH}$ (sodium hydroxide), $\text{KOH}$ (potassium hydroxide), $\text{N}_2\text{H}_4$ (hydrazine), or exotic organics, all complexed with water, alcohol, or other organics such as catechol [$\text{C}_6\text{H}_4\text{(OH)}_2$ = pyrocatechol]. Many formulations are described in the literature [8, 9].

A quasi-anisotropic silicon etch (Canadian patent No. 903650) is based on $\text{As}_2\text{O}_3$ (arsenic trioxide) in reflexed orthophosphoric acid ($\text{H}_3\text{PO}_4$), and yields vertical side walls irrespective of the wafer orientation [10]. The bottom profile may be controlled, however. Figure 10–1 shows a deep well etched in $\langle 111 \rangle$ Si as used frequently for dielectric isolation in IC fabrication. Increasing the fraction of $\text{As}_2\text{O}_3$ can change the bottom profile from A to B, so a flat bottom may be obtained. The dashed lines at the sides show the effect of undercutting beneath the mask when conventional etch is used.

10.6 Anisotropic Si Etch: Applications

Anisotropic etching is used in several silicon processing applications. We shall consider only three of these, briefly, by way of illustration. These techniques must be used judiciously, however, to avoid unwanted crystal-orientation effects. For example in etching mesa structures, $\langle 331 \rangle$ and $\langle 211 \rangle$ planes may be exposed, with a resultant cutting away of the mesa corners.

10.6.1 V-Groove Etching

A whole technology, VMOS, has been developed that is based on the ability to etch a V-shaped groove in silicon. A $\langle 100 \rangle$ wafer is used as the substrate, as shown in Figure 10–2(a). A window of width $w$ is cut in the masking medium, nominally $\text{SiO}_2$, with one pair of window sides parallel to the $\langle 110 \rangle$ directions, as shown in Figure 10–2(c); this will be parallel to the normal wafer flat. Then the wafer is immersed in an etch that attacks the $\langle 100 \rangle$ planes faster than the $\langle 111 \rangle$ planes. The dashed line $\alpha$ shows the groove after a short etch interval. The etch proceeds at a high rate vertically by acting on $\langle 100 \rangle$ planes. The $\langle 111 \rangle$ faces, however, are exposed on the sides, faces that etch very slowly; hence, as shown in (b), the etch effectively "stops" at the two exposed $\langle 111 \rangle$ planes that are inclined at 54.7° wrt the $\langle 100 \rangle$ planes.

The dashed line $\beta$ shows the groove at a later time, and finally the end result of the etch is shown by the solid line where no $\langle 100 \rangle$ plane is left exposed. The etch action effectively stops when this limit is reached, because only the slow-etching $\langle 111 \rangle$ faces are exposed.

An interesting result comes about because of the self-limitation at the bottom of the groove. If the etch is allowed to proceed until the V is formed, the vertical depth $d$ of the groove is controlled by the window width $w$. Reading from Figure 10–2(b) we note that
10.6.2 Narrow-Groove Etching

An anisotropic etch of (44 wt% KOH): H₂O at 85°C can etch ⟨110⟩ planes up to 400 times faster than ⟨111⟩ planes. Hence, if rectangular mask windows are aligned parallel to ⟨111⟩ planes on a ⟨110⟩ wafer, it is possible to get very narrow grooves with minimum undercut. Kendall has shown the importance of proper window alignment and has etched grooves only 0.4 μm wide [13].

High-C capacitors and vertical, multijunction solar cells are applications that use several deep, parallel, narrow grooves. The capacitors are fabricated by oxidizing the sides of the narrow grooves and then coating the oxide with deposited metal to give several capacitor segments in parallel.

10.6.3 Plane and Defect Revealing Etches

Certain anisotropic etches may be used to reveal wafer crystal orientation and crystalline defects. The former application has been illustrated in Figure 2–11(a) whereas a ⟨111⟩ triangle has been revealed by application of Sirle etch (Section 2.6). Other etches, such as Secco and Wright, will pit a wafer to reveal defects in the crystal structure.

10.7 Other Insulator Etches

After SiO₂ (or silicate glasses), the two most commonly used insulators in silicon technology are Si₃N₄ (silicon nitride), and polyimide (PI).

Si₃N₄: A typical nitride etch comprises an 85% solution of H₃PO₄ (orthophosphoric acid) in water. This formulation etches both Si₃N₄ and SiO₂ at rates of 100 and 25 Å/min, respectively, at 180°C. Unfortunately, it also attacks PR, so pattern etching can be a problem. This may be circumvented by depositing an SiO₂ layer by CVD on top of the nitride, and covering this with PR. (This idea has been considered in Section 5.9.2 in connection with nitride masking.) Dry plasma etching, which is discussed later in this chapter, is preferred, because it provides greater selectivity in etching the two layers.

Polyimide: Polyimide (PI) is a generic name for a family of organic polymers that contains the imide group shown in Figure 10–3(a). It is mixed with solvents in liquid form and is spun on the wafer as photoresists are (Section 11.8). Typical solvents for the polyimides are NMP (methyl pyrrolidinone) and DMF (dimethyl formamide). Sometimes these are used in mixture with acetone or methanol.

The thickness for a given formulation is determined by the spin rate and time. The layer is dried and cured at 300°C; subsequently, it is stable up to 500°C. The material is a good between-layer insulator, and is used for wafer passivation and especially for planarization [14, 15]. With proper additives it also may serve as a photoresist.

In multilayer structures, the top wafer surface becomes less flat as more windows are etched and additional layers are applied. Eventually, nonplanar-

Figure 10–2. A V groove is formed with an anisotropic etch. (a) Pertinent crystal planes. (b) Angles of the V groove. (c) Top view of Figure 2–4 showing the proper groove orientation.

\[
\tan \frac{70.6}{2} = \frac{w}{2d}
\]

whence

\[d = 0.706w\] (10.6.1)

If a square mask window is aligned as shown in (c), the two ends also will be ⟨111⟩ planes, so a V is formed in the second direction also. On the other hand, if the mask is rectangular, the V will appear normal to the shorter dimension w as in Figure 10–2(c). The groove will be flat-bottomed normal to the longer dimension, as shown by one of the dashed lines in Figure 10–2(a).

If the mask is not aligned exactly, there will be some undercutting beneath the window edges. The usual practice is to use a 5-min dip in 10% HF, after the anisotropic etch is completed, to remove any SiO₂ overhanging the edges [11].

Wafers of ⟨110⟩ orientation also may be used for V-groove etching, but ⟨100⟩ wafers are readily available so they are preferred. Many references discuss how the V-grooves are used in MOS technology [9, 12].
for masking, and special photoresists for metal etching, such as Kodak KTFR and KMER, are available.

10.9 Lift-Off Technique

In most photolithographic processes, such as the etching of windows in SiO₂, the PR is applied on top of the existing layer that is to be etched. The PR is exposed and developed before the wafer is immersed in the etchant. For thin metal films an alternative process is available: lift-off. A similar process was described in Section 9.10, but a different masking material is used here.

Lift-off is illustrated in Figure 10–4, where the PR is applied, exposed, and developed before the metal deposition. A postbake is not used (Section 11.8). The metal layer is deposited by CVD or PVD over the entire surface, windows included, as in Figure 10–4(b). The wafer then is immersed in PR stripper that can act laterally under the metal located atop the PR. The stripping action removes the PR and overlying metal with it. Metal deposited through the PR windows remains in place as shown in (c). Positive PR is preferred, because it can support finer details and no acids are involved in its use. Note that in lift-off the metal itself is not actually etched.

10.8 Metal Etches

Metal etches abound, but care must be used to ensure that a proper selective type is chosen so that other materials on the wafer are not attacked. Typical formulations are given in handbooks. Conventional PR techniques are used

Figure 10–4. The lift-off technique. (a) Patterned photoresist. (b) Metal is deposited over the entire surface. (c) The PR is dissolved and removed leaving the desired metal pattern.

10.10 Wafer Cleaning

Technically, wafer cleaning is not etching, but the process is included here because it may involve etching, and some steps are related chemically to etching.
The processing steps for a given batch of wafers may be spread over some two months or even longer. Most of that time the wafers are in storage between processing steps. Clean wafers can become oxidized and pick up dust and other contaminants within hours or so of exposure to air. Hence, on return to the production line they should be cleaned just prior to each processing step. In the next sections we consider some of the common wafer cleaning methods.

Four general types of contaminants are found on wafers:

1. Dirt and dust. This type is the result of wafer sawing, polishing, and dicing, all of which produce particles of various sizes. Dirt (actually detritus) may fall from plastic storage boxes, while hair and minute skin particles come from personnel. Chemicals also can contribute particulate contamination, but are available with low particle count.

2. Unwanted oxides. Silicon on exposure to air at room temperature can grow a very thin (< 100 Å) native oxide in just a few minutes. Also, some oxidizing agents that are used in processing can produce a thin oxide on exposed Si. As discussed in Section 10.3 they may be removed by HF-based etches.

3. Organics. These result from certain processing chemicals, and particularly from fatty acids deposited if skin contacts wafers. Lange gives a summary of contamination sources and discusses contamination by human skin [16].

4. Metallic residues and ions. These come mainly from processing chemicals. For example, analytical reagent grade HNO₃ assays out the heavy metals as Pb at 0.05, and Fe at 0.06 ppm (parts per million). The same grade of acetone is rated at < 0.1 ppm for Cu, Ni, and Fe, with the heavy metals at < 0.5 ppm. Electronic grade chemicals have smaller, but nonzero, amounts of trace metals. If these remain on the wafer surface, they may be introduced into the silicon during high-temperature operations such as oxidation, diffusion, and chemical vapor deposition. Metal ions can lower breakdown voltages and change MOS threshold voltages, hence they should be removed before high-temperature operations are started.

Generally the dirt and dust contaminants can be removed by water solutions and some sort of mechanical scrubbing action; chemistry is unnecessary. The three remaining types of contaminants require chemical reactions with oxidizers, acids, organics, or alkalis.

10.10.1 Scrubbing

Mechanical wafer scrubbing is performed by automatic machines that involve rotating brushes, high-pressure water jets, or sonic agitation [17]. Not only does it remove dust and dirt, scrubbing also improves adhesion of photoresist to the wafer.

In operation the brush bristles are forced down toward the wafer, so they are bent to make a line contact with it. The pressure along that contact line can be very high. If dry operation were permitted, the bristles would damage the wafer if they were harder than it. If the wafer were harder, the bristles would break and contribute more particles to be cleaned away. Moreover, such particles tend to lodge in etched windows or other low spots, making removal more difficult; thus it is essential that water be present so bristle and wafer never make contact. Preferably, a surfactant (detergent) should be added to lower surface tension and to improve wetting of both wafer surface and bristles. The detergent also acts as a degreasing agent. Unfortunately, bacteria thrive on detergent, so frequently NH₄OH is added to kill them. To further aid the wetting process, the bristles are made from hydrophilic (water-like) material such as nylon or propylene. The high pressure between the wafer surface and the water layer produces a rapid cleaning action, a few seconds being enough to bring about thorough cleaning [18]. After being scrubbed and rinsed, the wafers are spun dry.

The problems associated with brushes may be bypassed by replacing them with high-pressure spray jets. The water solution is forced from the jets at 300 to 3000 psi, with the jets located not over half an inch from the wafer surface. As the jets move relative to the wafer, a vigorous scrubbing action is supplied by the liquid itself. An added advantage is that the jet spray can remove very small particles that brush bristles would miss.

Ultraclean water has very high resistivity. As the water flows past the metal jets at high velocity, friction can build up a static electric charge (an example of triboelectricity). Additives to the bath help dissipate this charge; sometimes CO₂ (carbon dioxide) is bubbled through the liquid for this purpose.

Vigorous liquid scrubbing action also can be produced by cavitation—the formation of bubbles when shock waves, launched into the liquid from a sonic generator, hit the wafer surface. These bubbles do the actual scrubbing. Generators for this purpose operate around 0.8 MHz. Lower frequencies, in the 20 to 80 kHz range, which are more typical of conventional ultrasonic cleaning, cannot remove the submicron particles encountered in semiconductor processing. A mix of NH₄OH : H₂O : H₂O often is added to aid the bubbles in wetting the wafer surface [17, Part 3].

10.10.2 Degreasing

Contaminants of the three other types require chemical reactions to dislodge them from the wafer surface. As mentioned in the last section, greases, such as fingerprints, lightly bound to the surface sometimes can be removed by scrubbing with detergent solutions. Organic solvents can be used, too. Simple dip-and-dry procedures do not work very well, as the solvent rapidly becomes
Figure 10–5. A small vapor degreaser. Solvent fumes are condensed in the condenser region.

polluted as successive contaminated wafers are dipped into it. An alternative is vapor degreasing.

A simple form of vapor degreaser is shown in Figure 10–5. An organic solvent such as TCE (trichloroethylene)—or preferably TCA (trichloroethane) for safety reasons—is placed in the beaker, whose top is surrounded by a finned, aluminum cooling sleeve. The solvent is boiled, the vapors are cooled enough in the sleeve region to condense, and the liquid returns for boiling again. This is a reflux process because the solvent flows back into the sump.

A wafer to be cleaned is held in the solvent vapor, which dissolves the grease and condenses. The liquid falls and carries the dissolved materials back into the sump. The wafer is not dipped into the liquid, so it does not pick up dissolved, carry-over contaminants. Rather, it contacts only clean solvent vapor while the sump collects more and more of the contaminants. Another advantage is that the reflux action is efficient in preventing vapor loss by condensing the vapor before it can rise above the top of the beaker and cooling sleeve. This extends the useful life of a given batch of solvent, with a corresponding cost reduction.

After vapor degreasing with TCE or TCA, the wafer surface must be made water-compatible again, so it is rinsed in acetone, in methyl or isopropyl alcohol, and finally in water before drying. The reasons for this sequence are given in Section 3.14.1.

10.10.3 Organic/Inorganic Removal

General cleaning to remove organics and inorganics usually involves a strong oxidizer such as hydrogen peroxide ($\text{H}_2\text{O}_2$). This reagent must be handled carefully because the 30% concentration used for semiconductor work is much stronger than the dilute 3% form available at the local drugstore. The peroxide usually is combined with either an acid or a base. One example is given here. (Standard references, such as those given earlier for etching, should be consulted for more details, as the actual, multistep procedures used for cleaning can be rather long and involved.)

Piranhka etch, $7\text{(conc } \text{H}_2\text{SO}_4) : 3(30\% \text{H}_2\text{O}_2)$, attacks organics vigorously. The wafer is immersed for 10 min at 125°C, and then rinsed in $\text{H}_2\text{O}_2$. Organic traces are removed by a subsequent dip in $\text{NH}_4\text{OH} : \text{H}_2\text{O}_2$ or $\text{HCl} : \text{H}_2\text{O}_2$. A stop is available commercially for protecting equipment from piranha.

10.11 Wet versus Dry Etching

Dry etching provides an alternative to the methods already discussed; it involves gases, rather than liquids. Consider the pros and cons of the two etching types. The equipment required for wet etching is less expensive, its cost running only some 10% of that for dry plasma etching [19]. Equipment for both types is easy to automate, and both are able to etch several different materials with only a change in the etchant being used. The cost factor is important, so wet etching is used where it is adequate; even in 1986 over 85% of chips manufactured in the United States were processed by wet etching.

Wet etching also provides better selectivity. In plasma etching, a single etch medium that attacks, say, $\text{SiO}_2$, also attacks $\text{Si}$, but with different etch rates. Dry etching can give smaller line width (i.e., the size of the windows and interconnect lines). Wet etching undercutters more, so it is more isotropic, and minimum line width is limited; control is more difficult, and fewer devices can fit on a chip of a given size. Undercutting also raises interconnect resistance. The finer line structure available with anisotropic plasma etch thus allows much greater packing density on wafers and becomes predominant for widths of 2 $\mu$m or less.

Both input and output of the dry etching system are in gas form; this provides some distinct advantages over wet etch systems: Many of the gases used are safe, nontoxic, and easy to handle in vendor-supplied cylinders, and the effluents raise fewer pollution problems. We have seen earlier that dry etching provides better resolution, or smaller line widths, and a clean/etch/strip sequence can be handled more easily in a single reactor system than in wet etching. Temperature effects, which are inevitable since chemistry is involved, are much smaller with dry etch than with wet—where only a 1 to 2% change in etch rate per degree Celsius is the norm. Finally, wet etching presents more critical problems in effluent disposal since it usually involves liquid acids as against gases in the other type (Section 10.2).

We tend to think in terms of high-tech applications of silicon technology, VLSI and the like, which involve sub-micron geometries. Yet the geometries
of the vast majority of chips are not this small. Wet etching will be around for a long time. Murray gives an interesting table comparing the two etching types [19].

10.12 Dry Etching

In contrast to wet etching, dry etching is based on the use of one or more gases as the vehicle for the etching species. Gases are neutral; hence ionization, usually in a plasma (Appendix C), is required to release the etching species. A vacuum system is required for initial pumpdown before the gas is introduced, as is the plasma generating equipment. By-products are gaseous and can be removed by pumping.

There are two categories of dry etching, based on the properties of the chosen gas. (1) In reactive etching the ionized gas has radicals that react chemically with the material being etched. Gases can be chosen for different chemical reactions, so reactive etching tends to be selective but not necessarily anisotropic. (2) In nonreactive etching an inert gas such as argon (Ar) is used. The ionized argon atoms, Ar+, are accelerated by a large electric field toward the material to be etched and dislodge the atoms there by momentum transfer. This is a physical/mechanical process called sputtering. Ionized inert gases support no chemistry, so nonreactive etching tends to be nonselective, but may be anisotropic. There are other intermediate types of dry etching that combine physical and chemical processes, RIE (reactive ion etching) being an example. This form combines the properties of chemical etching and sputtering.

All types of dry etching require the neutral gas to be broken down into radicals or ions. In semiconductor work the conversion usually is accomplished by establishing a plasma that gives a copious supply of the desired species. Since a plasma is used, these types are called plasma etching.

Another type of reactive etching, not based on a plasma, utilizes a photochemical process to produce the reactive species by exciting the appropriate gas with a laser beam [20]. Because choice of chemistry may be made, this type tends to be selective, but it is not anisotropic.

10.13 Plasma Etching Reactors

In this section we consider some of the basic forms of plasma etching reactors and the reasons why each is suited to a particular type of etching. Other configurations are described in the literature. Appendix C discusses the formation of plasmas and related information.

10.13.1 Basic Reactor Types

Three typical reactor configurations used for plasma etching are shown in Figure 10–6. All of the configurations are r-f driven and require a vacuum chamber or vessel, fed by a suitable gas supply and exhausted with a vacuum pump to remove effluent gases and maintain the proper pressure within the chamber. The gas system details, which include means for regulating the gas flow at very low values, are not shown. The barrel reactor shown in (a) was the first type to find commercial use in the semiconductor industry. It is unique in that its electrodes are outside the vacuum vessel and so are not in contact with the discharge gas. The barrel reactor also may be driven by an inductance coil surrounding the vessel.

In some applications a perforated aluminum cylinder, called a tunnel, is placed concentrically between the vessel walls and the wafers. It is shown
dashed in Figure 10–6(a), and acts to confine the plasma outside the tunnel so that ions do not reach the wafers.

The reasons for having different reactor configurations may be explained in terms of the d-c voltage or self-bias that appears across the electrodes. A description of how this voltage is developed is given in Appendix C.3.

10.13.2 Batch Processing

In any commercial process, such as plasma etching or ion implantation, that requires processing in an evacuated chamber, one must reckon with the time required for pumpdown. It is directly related to chamber volume. Shall a large chamber that can process a large batch of wafers for each pumpdown cycle be used, or will a small chamber which processes one wafer at a time but has a short pumpdown time be better? This problem also was encountered in Chapter 8. In the early days of plasma etching the large-batch/large-chamber method was preferred. The trend now is toward a one-wafer-at-a-time approach.

An important factor here has to do with the uniformity of etch. If many wafers are exposed simultaneously to a single plasma between plane parallel electrodes, it is difficult to have every wafer in exactly the same environment, so etch uniformity suffers. The hex reactor of shape similar to that of Figure 9–2(d) or the cylindrical configuration of Section 12.8.2 overcomes this problem very well. Another method uses several wafers on a rotatable carousel, so that only one wafer at a time is exposed to the plasma, even though several are in the chamber.

The present trend toward wafers of larger diameter tips the scales in favor of processing one wafer at a time in the chamber. With this design approach, the electrodes and chamber may be shaped to provide excellent uniformity of etch over the entire wafer surface. The wafers may be shuffled into and out of the small chamber without breaking vacuum by using a lock-and-shuffle mechanism in principle not unlike that shown in Figure 8–7. These systems often are referred to as the load-lock type. Another advantage they have is that the chamber and gas are not exposed to air during the wafer shuffle cycle, thereby eliminating the problems associated with chloride gases contacting air (Section 10.15.2).

The principal exceptions to this trend of processing wafers one at a time are the hex and barrel reactors; however, the barrel now is used only for PR ashing (the plasma removal of photoresist).

10.14 Endpoint Detection

A single plasma may etch both mask and underlying material, but not necessarily at the same rate. If the mask etch rate is lower, endpoint detection is essential so the etch can be stopped when the mask windows are completed. Endpoint detection methods monitor some system parameter that changes significantly when the mask etch stops or the underlying etch begins. We briefly consider five typical methods [21].

10.14.1 Optical Spectroscopy

The plasma-generating gas glow discharges are characterized by a color that depends upon the gas being used. For example N₂ gives a pinkish glow, while O₂ has a bluish color. It turns out that atoms and free radicals such as F and CF₃⁺, respectively, also can become excited and subsequently emit radiation of characteristic wavelength. Some typical values are: F⁺, 7040 Å; CO⁺, 2977, 4835, and 5198 Å; and N⁺, 6740 Å. The superscripted asterisk here represents an excited species [22]. One such emission may be monitored with a typical setup shown in Figure 10–7(a), where the monochromator serves as an adjustable filter. A simpler system uses a fixed optical filter to pass the radiation from a single species. The intensity in both cases is then monitored by an appropriate photodiode or photocell.

If an active species such as F⁺ is monitored, its concentration drops during etch, causing a decrease in the monitored radiation. On the other hand, if a by-product such as CO (carbon monoxide) is monitored, its emission intensity will increase during etch. Thus, the endpoint can be detected as a change in radiation intensity at the particular wavelength being monitored.

If the complete spectrum is plotted by a spectrophotometer (recorder in the figure), the method also may be used as a leak detector. For example if a peak occurs at the hydrogen value of 6563 Å, the odds are large that water is getting into the system, unless H₂ is a supplied or by-product gas. The presence of N₂, indicated by a peak at 6740 Å, usually signals an air leak in the system, unless N₂ is being supplied.

10.14.2 Mass Spectroscopy

Another similar means of endpoint detection uses a mass spectograph, as described in Section 8.4.2, to separate out one species by its mass. The monitored quantity essentially gives a count of the selected species. The spectrometer generally is located between the reactor chamber and the exhaust pump (i.e., at some distance from the discharge so it cannot sense short-lifetime species). Usually an etching by-product is measured [23].

10.14.3 Laser Interferometry/Reflection

The reflecting properties of the wafer surface can change as a result of etching. This change may be sensed for endpoint detection by focusing a laser beam
on a spot on the wafer and monitoring the intensity of the reflected beam. If the etched surface is a thin film, interferometry is preferred. A typical setup is shown in Figure 10–7(b). Due to its high reflectance, aluminum works well with this method.

The reflecting method is inadequate for batch processing because only a small spot on one wafer is monitored. Also, if very fine lines are being etched, the change in reflection from the test spot will be very small. This can be circumvented by using masks that provide for a fairly large test spot. Careful orientation of the wafer in the reactor is needed so that the wafer is reached by the laser beam.

10.14.4 Self-Bias Detection

This is a very simple system, which requires only a high impedance d-c voltmeter connected across the electrodes. The magnitude of the self-bias voltage (Section C.3) is influenced by changes in the discharge, so a change in the concentration of any species will cause a variation in the voltage reading. The variation may be small, however, and give less sensitivity than methods that monitor a single species in the discharge.

10.14.5 Pressure Sensing

The total pressure in a discharge is affected by a change in the partial pressure of any of its components; hence, the pressure in the discharge or effluent may be monitored with a vacuum gage to sense the endpoint. This is a simple method to realize, but precautions are necessary. Small variations in the gas flow rate into the reactor will cause small pressure variations, resulting in noise in the gage’s electrical output. This may be filtered out with a simple RC filter having a time constant of about 5 s. Chlorine-based etches can contaminate the sensing element in certain gage types, causing false readings. This can be remedied by heating the gage periodically with an air gun.

10.15 Reactive Plasma Etching

In reactive plasma etching the source gas is broken down by the plasma into excited elements, here represented by a superscripted asterisk, and into charged and/or neutral free radicals (groups of atoms). These, for example F* or CF$_3^+$, are then available to do the actual chemical etching. Reaction by-products are gases at the operating temperature, and their partial pressures are kept high enough to be removed by the system’s vacuum pump. Gases are replenished by leaking them (introducing them at low flow rates) into the system to balance the pump action. Two type of reactors are used for reactive plasma etching: the barrel and the parallel plate electrode (PPE) forms, shown in Figure 10–6(b) and (c), respectively.

The choice of gases for use in reactive dry etching has been largely empirical. A partial list of these gases is given in Table 10–1, which shows that CF$_4$ (Frecon-14) etches a wide range of materials. It is one of the fluorocarbons that were developed originally as nontoxic, noncorrosive, and nonflammable refrigerants, and was one of the first gases used for plasma etching. Their
Table 10-1. Partial List of Plasma Etching Gases

<table>
<thead>
<tr>
<th></th>
<th>Si</th>
<th>polySi</th>
<th>SiO₂</th>
<th>Si₃N₄</th>
<th>Al₂O₃</th>
<th>GaAs</th>
<th>PR</th>
</tr>
</thead>
<tbody>
<tr>
<td>CF₄</td>
<td>√</td>
<td>√</td>
<td>√</td>
<td>√</td>
<td>√</td>
<td>√</td>
<td></td>
</tr>
<tr>
<td>CF₄/O₂</td>
<td></td>
<td>√</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CF₄/Ar</td>
<td></td>
<td>√</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CF₄/H₂</td>
<td></td>
<td>√</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CF₄/N₂</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CCl₂F₂</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CCl₄</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CCl₄/Ar</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>C₁₂F₆</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>C₁₄F₈</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>C₄F₈</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>HF</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>O₂</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>BCl₃</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CClF₃</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CF₃Br</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CHF₃</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

*Note: (√) indicates low etch rate. Compiled from several sources.*

The properties enhance the fluorocarbons' value as etchant source gases, but, as we shall see, they sometimes lack selectivity. Unfortunately, recent experience indicates that the related chlorofluorocarbons (CFCs) affect the earth's ozone layer.

**10.15.1 Fluorocarbons**

Consider some basic reactions resulting from a CF₄ plasma. A copious supply of free electrons (e⁻) is present in a plasma, and some of them enter into the fundamental reaction that gives free radicals, namely

\[
CF₄ + e^- → F^* + CF₃^+ + 2 e^-
\]

It has been determined that the F* species has a long lifetime (τ) on the order of 0.1 to 1 s, depending upon the partial pressure of CF₄. This means that atomic F* can exist outside the plasma proper. In contrast, the CF₃⁺ radical, which is a strong reducing agent, has a relatively short τ of roughly 10 μs, and so cannot exist outside the plasma.

These facts have strong implications for the barrel etcher. The charged species CF₃⁺ is prevented from reaching the wafers by lifetime and tunnel-shielding effects; hence etching by CF₄ in the barrel reactor must be by the excited atomic form F*.

Consider some reactions of the two radicals F* and CF₃⁺ with various materials to be etched. It is atomic fluorine that provides the etching of silicon:

\[
Si + 4 F^* → SiF₄
\]

(10.15-2)

The resulting compound SiF₄ is a gas at etch temperatures and pressures, and so may be removed by the system pump. From τ considerations the silicon need not be located directly in the plasma for etching to take place, so a barrel etcher can be used.

Silicon dioxide also is etched by a CF₄ plasma. There is disagreement on the etching action, but the consensus holds that the active radical is CF₃⁺. The overall reaction may be written in the somewhat ambiguous form

\[
SiO₂ + CF₃⁺ → SiF₄ + CO + CO₂
\]

(10.15-3)

where the CF₃⁺ comes from the basic plasma reaction of Eq. (10.15-1).

In typical processing an SiO₂ mask on Si is frequently encountered. Since CF₄ etches both materials, the problem of relative etch rates arises. If mask windows are to be etched in the oxide, the etch will not stop automatically on reaching the silicon, as it did with liquid HF; hence it is desirable to have the SiO₂ : Si etch rate ratio, R_{SiO₂:S}, as large as possible. On the other hand, if silicon is to be etched through oxide mask windows, we desire R_{SiO₂} to be as small as possible (or R_{SiO₂} = 1/R_{SiO₂} as large as possible). What can be done to satisfy these contradictory conditions?

If the barrel etcher is used, the wafers are shielded from the charged CF₃⁺ species, but they can be reached by the long-lifetime, excited F* atoms; hence the silicon etches faster, with a typical R_{SiO₂} value of 1/10.

If a PPE (parallel-plane-electrode) reactor is used, the wafers are in the plasma, so both species can etch; experience shows an R_{SiO₂} value of roughly one, so for etching SiO₂ the etch ratio must be enhanced in some manner. One approach adds a scavenger gas that consumes the F* species before it can
Table 10–2. Etching Gas Nomenclature

<table>
<thead>
<tr>
<th>Gas</th>
<th>Name</th>
</tr>
</thead>
<tbody>
<tr>
<td>BCl₃</td>
<td>boron trichloride</td>
</tr>
<tr>
<td>CBrF₃</td>
<td>bromotrifluoromethane</td>
</tr>
<tr>
<td>CClF₃</td>
<td>chlorotrifluoromethane</td>
</tr>
<tr>
<td>CCl₂F₂</td>
<td>dichlorodifluoromethane</td>
</tr>
<tr>
<td>CCl₂O</td>
<td>carbonyl chloride, phosgene⁴</td>
</tr>
<tr>
<td>CF₄</td>
<td>tetrafluoromethane</td>
</tr>
<tr>
<td>CHF₃</td>
<td>fluoroform</td>
</tr>
<tr>
<td>CHF₂F</td>
<td>fluoromethane, methyl fluoride</td>
</tr>
<tr>
<td>C₂Cl₂F₄</td>
<td>dichlorotetrafluoroethane</td>
</tr>
<tr>
<td>C₂ClF₃</td>
<td>trichlorotrifluoroethane</td>
</tr>
<tr>
<td>C₂F₆</td>
<td>perfluoroethane</td>
</tr>
<tr>
<td>C₃F₈</td>
<td>perfluoropropane</td>
</tr>
<tr>
<td>C₄F₈</td>
<td>perfluorocyclobutane</td>
</tr>
<tr>
<td>HF</td>
<td>hydrogen fluoride</td>
</tr>
<tr>
<td>NF₃</td>
<td>nitrogen trifluoride</td>
</tr>
<tr>
<td>SF₆</td>
<td>sulfur hexafluoride</td>
</tr>
<tr>
<td>SiCl₄</td>
<td>silicon tetrachloride</td>
</tr>
<tr>
<td>SiF₄</td>
<td>silicon tetrafluoride</td>
</tr>
</tbody>
</table>

⁴extremely dangerous

attack the Si, leaving the CF₃⁺ free to etch the oxide. Hydrogen is one such gas since it reacts with F* to form HF in gas phase. Principal etching of the oxide will be by CF₃⁺, however.¹

Another approach is to change the source gas. The C/F ratio in the gas affects the etch rate ratio. Singer cites the following data: for CF₄, R₉₅ = 1; C₂F₆, 3; C₃F₈, 5; CHF₃, 10. This is an example of how an empirical choice of gas can solve an etching problem. Singer also discusses the etching chemistry of these gases and the variation of etch rates with process parameters [25].

If the scavenging gases such as H₂ or CHF₃ deplete the F* too much, a condition known as polymerization can occur. This means that molecules of an unsaturated reaction compound at the wafer surface can join to form more complex molecules. These may form solid residues that can retard or even stop the etching action. Singer gives the curves of Figure 10–8, which show the parameter regions where this effect comes into play [25].

On the other hand, if Si is to be etched through an oxide mask in the PPE setup, the etch rate for Si should be significantly higher than that for SiO₂. A mixture of CF₄ with about 8% oxygen added is used for this, since it gives an etch rate ratio of (111)Si to SiO₂ of about 17, and of polysilicon to the oxide of about 25 [25, 26]. One of the mechanisms postulated for this effect is that free electrons in the plasma collide with molecular O₂ to form atomic O*. This is not a chemical reaction, but then intermediate chemical steps form fluorinated radicals such as COF*. These react at the surface to give an overall result of

\[ \text{CF}_4^+ + \text{O}_2^+ \rightarrow 4 \text{F}^* + \text{CO}_2^+ \]  (10.15-4)

or possibly

\[ 2 \text{CF}_4^+ + 2 \text{O}^* \rightarrow 8 \text{F}^* + 2 \text{CO}^+ \]  (10.15-5)

In either event, the production of the silicon etching species F* is enhanced, resulting in the higher Si etch rate. No such increase is caused in the production of CF₃⁺ that etches the oxide. In all of the foregoing reactions it is easy to identify an active or by-product species suitable for endpoint detection as discussed in Section 10.14.

Silicon nitride (Si₃N₄) also may be etched in a CF₄ based plasma with the F* species acting as the actual etchant. Following the reaction of Eq. (10.15-1) the result is:

---

¹ Anhydrous HF has been used to etch SiO₂ selectively in the DryOx process. Even though this does not involve a plasma, it uses HF at temperatures and pressures common to plasma etching. Bersin and Reichelderfer state that HF in the 150 to 190°C temperature range and at pressures from 0.1 to 30 torr does not etch Si [24]. Etching under a mask also is possible. DryOx is a trademark of the International Plasma Corporation.
\[ \text{Si}_3\text{N}_4 \downarrow + 12 \text{F}^* \uparrow \rightarrow 3 \text{SiF}_4 \uparrow + 2 \text{N}_2 \uparrow \]  

(10.15-6)

There is a problem in that F*, the active agent for \text{Si}_3\text{N}_4, also etches Si, and at a faster rate. Typically, the etch rate in a barrel reactor will be less than that for Si but about four times that of SiO_2 [27]. Hence, an oxide may be used as an etching mask for the nitride. Since the same species etches both Si and its nitride, it follows that steps taken to raise (or lower) the etch rate of Si relative to SiO_2 will have the same effect on the nitride. The substitution of SiF_4 improves the nitride/silicon selectivity, however. With 4% oxygen added, the etch rates of Si and SiO_2 are reduced with respect to \text{Si}_3\text{N}_4, that of Si by a factor of about 6. This combination also will etch polyimide [28].

Consider metal etching with CF_4. Both tungsten (W) and molybdenum (Mo) react with atomic F* to give hexafluorides that are volatile at etching temperatures and pressures.

\[ \text{W} \downarrow + 6 \text{F}^* \uparrow \rightarrow \text{WF}_6 \uparrow \]  

(10.15-7)

\[ \text{Mo} \downarrow + 6 \text{F}^* \uparrow \rightarrow \text{MoF}_6 \uparrow \]  

(10.15-8)

Reference to Table 10-1 shows that Al is not etched by fluorocarbon types of chemistry, a fortunate fact because it provides a form of selectivity. The table also shows that chloride chemistry does work with Al.

10.15.2 Chlorides

Aluminum is a commonly used metal in silicon technology and requires a suitable selective etch. Chlorides such as \text{CCl}_4, \text{BCl}_3, and \text{SiCl}_4 are used rather than fluorocarbons because of selectivity. These chlorides serve a double function.

When aluminum is exposed to a thin layer, of about 30 Å, of Al_2O_3 (aluminum oxide) grows very quickly, even at room temperature [29]. This native oxide is resistant to a large range of chemicals and must be removed before the aluminum itself can be etched. Even though CF_4 will not attack Al, it is used sometimes for a 2-min oxide removal before the Al etch is started. But chlorine (Cl_2) alone cannot remove the oxide, e.g., this reaction is not possible:

\[ 2 \text{Al}_2\text{O}_3 \downarrow + 6 \text{Cl}_2 \uparrow \rightarrow 4 \text{AlCl}_3 \uparrow + 3 \text{O}_2 \uparrow \]  

(10.15-9)

because the free energy of reaction is \( \Delta G = +105.5 \text{ kcal/mole} \).

Fortunately most chlorides can etch both the native oxide and Al. For example, for etching Al_2O_3 with \text{CCl}_4:

\[ 2 \text{Al}_2\text{O}_3 \downarrow + 3 \text{CCl}_4 \uparrow \rightarrow 4 \text{AlCl}_3 \uparrow + 3 \text{CO}_2 \uparrow \]  

(10.15-10)

where \( \Delta G = -16.8 \text{ kcal/mole} \) and the reaction can proceed [30]. The behavior of \text{BCl}_3 is similar to that of \text{CCl}_4.

The gases \text{CCl}_4 and \text{BCl}_3 are used most often for etching aluminum. Their etching chemistries are quite similar, so the choice between them is based largely on their other properties. Both are considered to be analogous to \text{CF}_4; when they collide with free electrons in the plasma these radicals will result:

\[ \text{CCl}_4 \uparrow + e^- \rightarrow \text{Cl}^* \uparrow + \text{CCl}_3^* \uparrow + 2 e^- \]  

(10.15-11)

and

\[ \text{BCl}_3 \uparrow + e^- \rightarrow \text{Cl}^* \uparrow + \text{BCl}_2^* \uparrow + 2 e^- \]  

(10.15-12)

In either case it is the atomic Cl* that provides the etching action. This is a comparatively short-lived species and so cannot exist outside the plasma proper. For this reason, chloride-based etching of aluminum is not practicable in a barrel reactor with tunnel; a parallel plate (PPE) configuration, where the wafer is in contact with the discharge, is necessary. The 2550 Å spectral line of Cl can be used for endpoint detection.

The actual etching reaction is

\[ \text{Al} \downarrow + 3 \text{Cl}^* \uparrow \rightarrow \text{AlCl}_3 \uparrow \]  

(10.15-13)

The AlCl_3 (aluminum chloride) by-product is a solid at room temperature and sublimes at roughly 180°C; so substrate heating is used to ensure that the chloride volatilizes.

The reactions of both Eqs. (10.15-11) and (10.15-12) can run in the reverse direction, causing Cl* to recombine with \text{CCl}_3* or \text{BCl}_2*, as the case may be, to form the original source chloride. These are three-body reactions and require a surface—of the wafer, of the fixtures, or of the walls and base plate of the reactor. Since the reverse reactions will reduce the quantity of Cl* available for etching, they lower the etch rate. Therefore, good scavenging of Cl_3* or BCl_2* by the system pump is important.

10.15.3 PR Ashing

Dry process removal of photoresists, PR ashing, is often carried out in the barrel etcher. Oxygen is used to ash conventional organic resists and plasma excitation results in atomic oxygen (O*), which ideally oxidizes the PR into gases such as CO, CO_2, and H_2O that are removed by the system's vacuum pump. Complete removal is essential so that no solid residue remains on the wafers. An argon purge is often used after the stripping process. Typically, an ashing cycle for a 1-μm-thick resist layer may last 15 to 40 min, with the actual
time depending upon such factors as the load (number of wafers in the batch) and the applied power [31].

An oxidizing agent weaker than O₂ is desired for stripping resists on sensitive films (e.g., the chromium on some photomasks) to avoid damaging them. Wet air (air bubbled through water) may be used for this.

The ashing time may be reduced by using other gases or by preheating the wafers to 200 to 300°C before insertion into the etcher barrel. This heat treatment can cut the ashing time in half. Also, the CF₄/O₂ mixture can speed the process, although care must be used to ensure that the F⁺ species will not damage exposed Si or Si₃N₄. The CF₄/O₂ mixture also will strip polyimide-based resists.

Special care is required if the photos resist has been used to mask aluminum during etching. Conventional negative resists take up free chlorine-bearing radicals. If the resist is exposed to moist air after the etch, HCl will form and possibly etch the aluminum further. Also the resist becomes more resistant to stripping. To avoid this problem, the wafers should be immersed in distilled water for a minimum of 2 min after etching. The strip process follows after drying. Conventional positive resists, due to different chemistry, do not exhibit this effect, but they should be stripped immediately after the etch cycle.

Irving cites these advantages of dry over wet PR stripping [31]:

1. It is a cooler process.
2. Stripping time is independent of resist history, e.g. pre- and post-bake times.
3. It requires fewer steps.
4. Less airborne contamination is present since the process is isolated in a partial vacuum.

10.16 Anisotropic Etching

Plasma etching with reactive gases at relatively high pressures, say > 10⁻¹ torr, usually is isotropic: the etch proceeds horizontally as well as vertically relative to the wafer surface. For small geometries, with line widths of 2 μm or less, anisotropic etching normal to the wafer surface is desired. This may be achieved with two forms of plasma etching: nonreactive ion or sputter etching with gas in the 10⁻⁴ torr pressure range, and reactive ion etching (RIE) with reactive gas in the 10⁻² to 10⁻¹ torr range.

10.17 Nonreactive Ion Etching: Sputtering

In ion (or sputter) etching a nonreactive gas such as argon (Ar) is ionized by an r-f plasma in a PPE configuration as shown in Figure 10–6(b). A d-c self-bias voltage is present (Section C.3). The Ar⁺ ions are accelerated toward the wafers, which are on, and in contact with, the grounded negative electrode. If the d-c self-bias is large enough, the argon ions will hit the wafer with sufficient momentum to dislodge material, giving sputter etching.

The gas pressure can determine if the etching will be anisotropic or not. The mean free path \( L_m \) (the average distance the ions travel between collisions) is given by

\[
L_m = \frac{5 \times 10^{-3}}{p^*}, \text{ cm}
\]

where \( p^* \) is the gas pressure in torr (= 1 mm Hg).

Typically, \( d \), the interelectrode spacing, will run a few, say 5 cm, and \( d \), the spacing from the cathode to the near edge of the plasma, will be a small fraction of this, say \( d \approx 1 \text{ cm} \).

Say \( p^* = 1 \text{ mtorr} (= 1 \text{ micron}) \), so that \( L_m = 5 \text{ cm} \). This is large enough wrt \( d \) that the positive ions (\( i^+ \)) probably will not collide with other ions in moving from the plasma to the negative electrode and so will travel in essentially straight lines normal to the cathode. If the electric field is large enough to cause sputtering, the etch will tend to be anisotropic.

On the other hand, say \( L_m < d \). For example, if \( p^* = 20 \text{ mtorr}, L_m = 0.25 \text{ cm} \). Then an ion will suffer several direction-changing collisions in moving to the cathode, so anisotropic etching cannot be expected. The use of r-f, as opposed to d-c, excitation favors operation at lower pressures, so anisotropic etching can be achieved.

Since no chemical action is involved, all materials exposed to the ions are etched: oxide, nitride, aluminum, and silicon, but not at the same rate. Generally, the selectivity among the materials is poor because it depends on their relative sputtering rates alone, and not on chemical reaction. For this reason sputter etching is seldom used in IC processing.

When ions hit the wafer under usual sputtering conditions, they give up some kinetic energy to heat. To prevent an excessive rise in temperature, the wafer-holding electrode is usually water cooled. This also minimizes the formation of an arc discharge (Appendix C.1).

10.18 Reactive Ion Etching (RIE) [32]

In RIE reactive gas in the pressure range of 10⁻² to 10⁻¹ torr is used with the reactor configurations shown in Figure 10–6(c).² In brief, the advantages of PPE and sputter etching are combined to give a combination of reactive chemistry and mild sputtering by positive ions. For a midrange pressure of

²The effect of comparative electrode size on voltage magnitudes is discussed in Appendix C.3.
5 \times 10^{-2} \text{ torr} the mean free path is \( L_m = 10^{-1} \text{ cm} \). At these pressures the distance between the plasma sheath and the wafers is even smaller, so ions arrive primarily in the direction parallel to the \( \mathcal{E} \) field. Also, the vertical features on the wafer surface are much smaller than the interelectrode spacing, being at the most a few micrometers, so they do not disturb the direction of the \( \mathcal{E} \) field, which remains normal to the wafer surface. As a result, the sputtering action is anisotropic.

The reactive chemistry and ion sputtering are not necessarily by the same species; a case in point is Si. Recall from Section 10.15.1 that F\(^+\) is the species that etches silicon in a CF\(_4\) plasma. Since this is a charge-neutral species, it cannot gain momentum due to the \( \mathcal{E} \) field and cannot sputter etch; some other mechanism must be involved.

While we tend to think of ion enhancement of etching by direct sputtering, some alternate mechanisms are proposed, all of which are based on the formation of volatile reaction products. These mechanisms are (1) polymer removal, (2) creation of dangling bonds at the surface that accelerate the formation of free-radical reactions, and (3) localized surface heating by ion impact that can accelerate chemical surface reactions and the volatilization of compounds.

Consider some facts related to these mechanisms. The creation of dangling bonds at the wafer surface facilitates reaction with free radicals there to form oxides, fluorides, or chlorides—depending upon the source gas composition. The reaction rates of neutral species at a surface (e.g., F\(^+\) or CF\(_2\)) have an exponential temperature dependence of the form \( \exp(-b/T) \); therefore, wafer temperature can affect rates at the wafer surface.

Production rates of species in the plasma are relatively independent of the wafer surface temperature; rather, they depend on electron impacts with gas molecules, and so on the free electron concentration in the plasma. If the power delivered to the plasma is increased, free electron concentration increases, so positive ion concentration increases, too. This tends to increase any ion-enhanced process.

The formation of polymers on wafer surfaces depends on the wafer material, its surface condition (whether damaged or having dangling bonds) and the source gas composition. For example, most polymers are carbon based. Carbon is available in many gases, such as CF\(_4\), CHF\(_3\), or CCL\(_4\), but RIE gases, such as SF\(_6\) and NF\(_3\), have no carbon. Yet any organic present, such as PR, can furnish carbon, and C-polymer films may form. Subsequently these may be converted to volatile compounds such as CO, CO\(_2\), or H\(_2\)O. Si-based polymers usually do not convert to volatiles, however; therefore, they may be used as an etch stop.

Polymers form by three commonly used gases obey the hierarchy CF\(_4\) < CHF\(_3\) < CH\(_3\)F. Where polymerization is not wanted, say for sidewall protection of Si, a gas should be chosen, if possible, that forms heavier ions, such as CCL\(_x^+\) rather than the lighter BCL\(_y^+\).

Polymer formation on the RIE reactor electrodes also can affect wafer etch rates and their constancy. Electrode polymers can act as sources of contaminating particles when sputtered. Tailoring of the gas may decrease polymer formation; Egitto et al. show that the addition of CO\(_2\) or O\(_2\) to CHF\(_3\) has this effect [33]. Apparently these added gases scavenge excess CF\(_3\) and CF\(_3^-\) radicals that are active in polymer formation.

We shall consider some typical etching applications that are based on the foregoing mechanisms. Remember that the choice of reactive and other source gases is largely empirical. In these examples relative, rather than absolute, etch rates are given because the latter depend on too many parameters—such as electrode geometry and size, and the wafer batch size or load.

### 10.18.1 Silicon Dioxide

The etching of silicon dioxide provides a good example of the combined effects of chemical and sputter actions by RIE in CF\(_4\) plasmas. The oxide often is on or near silicon, so relative etch rates are of concern. We saw earlier that \( R_{ox} \), the oxide-to-silicon etch rate ratio, is relatively small and that the etch is isotropic in PPE, being entirely chemical in action. Under RIE conditions, however, \( R_{ox} \) increases because the CF\(_3^+\) radical selectively attacks the SiO\(_2\) by both chemical and sputtering actions.

If F\(^+\), the species that chemically etches the silicon, were scavenged, \( R_{ox} \) would increase because of the lowered silicon etch rate. This effect has been demonstrated by experiments with the gas combination CF\(_4\)/H\(_2\) [34]. As the percentage of H\(_2\) is increased, the etch rates of both Si and SiO\(_2\) decrease but that of Si does so much faster. The \( R_{ox} \) ratio increases up to a maximum of 35:1 at 40% H\(_2\). At higher percentages a C-based polymer film, which does not sputter well, forms on the wafers and causes the chemical etch by CF\(_3^+\) to decrease. Note that the Si rate decreases because of the scavenging effect of H\(_2\) on F\(^+\).

Ephrat also reports that AZ1350B, a Shipley optical, positive photoresist, and PMMA (polymethylmethacrylate), a positive resist used with electron and ion beams, also etch more slowly as the H\(_2\)% increases and at a lower rate that SiO\(_2\), so the selectivity of silicon dioxide etching wtr these two resists rises with H\(_2\) percentage. Silicon nitride's etch behavior in the CF\(_4\)/H\(_2\) gas mixtures parallels that of SiO\(_2\) [34].

Results similar to those with CF\(_4\) are obtained if CHF\(_3\) is used in place of CF\(_4\)/H\(_2\) since hydrogen is available to scavenge the F\(^+\) species. Polymer formation may occur also, in fact CHF\(_3\) is well known as a polymer precursor. More information on the behavior of CHF\(_3\) when mixed with other gases is given later.
Consider an RIE mechanism other than the combined chemical and direct sputter etching of SiO₂. This mechanism assumes that while CF₄ does not adsorb to oxide and silicon surfaces readily, the plasma-produced species F* and CF₂⁺ do [35]. One of these latter, CF₂⁺ (difluorocarbene), forms a polymer on Si that does not sputter well, so the chemical etching by F* on silicon is limited. When CF₂ is adsorbed to SiO₂, however, it releases F, which reacts chemically with silicon in the SiO₂ and carbon, which reacts chemically with oxygen in the SiO₂ to form volatile compounds such as CO and CO₂. The mechanism proposed here differs from the earlier one, but the overall result is the same: \( R_{OS} > 1 \) in CF₄. The sputter action aids in the formation of volatile compounds on the oxide, but is not effective on the polymer formed on the silicon.

This model is consistent with the result that carbon in fluorocarbon gases has less effect on the etching of SiO₂ than on Si. The effects on SiO₂ depend on the radical CF₂ that forms volatile products on the oxide, but C forms a hard-to-sputter polymer film on the Si; hence a higher C : F ratio in the source gas raises \( R_{OS} \) as in CF₃F₆ and C₃F₈ over that for CF₄ [36].

The foregoing example illustrates that halocarbon plasma species can react with wafer materials to form polymers that inhibit chemical reactions. If some of these polymers are sputtered off by the accelerated ions, etching will proceed on the sputtered surfaces, but not on surfaces where the film remains [37].

Different polymer films may exhibit different effective etch rates. For example, if Si is exposed to oxygen during RIE, its etch rate \( r_S \) drops. On the other hand, with chlorine substituted for O₂, \( r_S \) rises. This is attributed to differences in the polymer films on the Si [35].

The effect of different etch rates may be clarified by the example shown in Figure 10–9. A cross section of silicon masked with oxide is shown at (a). The CF₃⁺ ion direction is vertical as shown. The F⁺ species attacks the Si at “xxx” and forms a polymer that is removed by CF₃⁺ sputtering; thus F⁺ continues to etch the Si with the polymer forming on both the sidewalls and bottom of the well as shown at (b). Again, the ions sputter away the polymer at the bottom. Since they arrive vertically, however, they cannot reach and sputter the sidewall polymer film. This action continues with the final result shown at (c). There is no undercut and so no lateral etching of Si under the mask; therefore, the Si etching is anisotropic.

Results with CHF₃ can give some insight into the models. Alone, CHF₃ permits RIE of silicon dioxide but at a lower rate than the fluorocarbons. Even though F⁺ is scavenged by the hydrogen, plasma products, most likely CF₂⁺, cause more polymerization on the oxide; sputter rate there decreases, so \( R_{OS} \) decreases. Chang has reported results for CHF₃ mixed with O₂ and CO₂ [37]. For 9 CHF₃/1 O₂, O₂ apparently lowers the CF₂⁺ concentration and raises the F⁺ concentration. Direct chemical etch of Si increases, with less chance of silicon polymer growth, so \( R_{OS} \) drops compared to the value with CHF₃ alone.

Figure 10–9. Ion direction and polymer formation produce etching anisotropy.

With 9 CHF₃/1 CO₂ the CO₂ aids formation of volatile compounds at the SiO₂ and less polymer film develops there. Chang reports good anisotropy with \( R_{OS} > 20 : 1 \).

10.18.2 Silicon

The etching of silicon is of great importance, especially of polysilicon, which in modern MOS ICs is used for gates and some levels of interconnects. In both applications the polySi may be highly doped to lower its sheet resistance. Doping levels affect etch rates, a phenomenon related to the change in Fermi level as doping concentration changes [38].

We already have some data in the preceding section about fluorocarbon etching of Si in the RIE setup. Letting \( r_S \) denote the silicon etch rate, we can summarize as follows:

1. In CF₄, \( r_S \) is low.
2. Species that scavenge F⁺ lower \( r_S \). Such species are obtained by adding H₂ to the CF₄ or by substituting CHF₃.
3. The addition of O₂ to either CF₄ or CHF₃ enhances the production of F⁺ causing \( r_S \) to rise.
4. Polymer formation on Si is enhanced by raising the H : F ratio. This is consistent with (2) and lowers \( r_s \).

We now consider some alternative chemistries for Si etching. Recall that the silicon-to-oxide etch rate ratio is \( R_{SO} = 1/R_{SO} \). Sulfur hexafluoride (SF\(_6\)) may replace CF\(_4\). Under a photoresist mask \( r_s \) is high but undercutting is present so the etch is isotropic. The addition of up to 30\% of CFCl\(_3\) (fluorotrichloromethane) to the SF\(_6\) makes the etch anisotropic, apparently the effect of the chlorine addition.

If SF\(_6\) is used to etch Si under an SiO\(_2\) mask, \( r_s \) remains high but etching is isotropic, a condition that can be remedied by adding one of the chlorides. No carbon is present to combine with the oxygen in the SiO\(_2\) to form volatiles. Also Cl\(_2\) does not react appreciably with SiO\(_2\) so the oxide etch rate is low, and \( R_{SO} \) is high.

10.18.3 Aluminum

Aluminum and its copper and silicon alloys also may be etched anisotropically under RIE conditions [39, 40]. As we have seen, chlorides form the basis for the etch process. Since highly corrosive by-products are formed, they frequently are frozen out from the system exhaust by being passed through a cold trap chilled with liquid nitrogen, or they are converted chemically.

A number of chlorine-bearing gases are used—e.g., Cl\(_2\), BCl\(_3\), CCl\(_4\), and other fluorocarbons—but any one used alone gives poor results, so mixtures, including other additives, are used commonly. Consider the BCl\(_3\)/Cl\(_2\) combination. The first gas is needed to remove the native Al\(_2\)O\(_3\) oxide. It also acts as a getter for the O\(_2\) released in reduction of the oxide, and so helps to prevent subsequent oxidation, but its aluminum etch rate is low. Chlorine in the 25 to 50\% range is added as a remedy, but too much will make the etch isotropic.

While Cl\(*)\ is the principal active etchant, anisotropy of etch is aided by BCl\(_3\) radicals that are formed in the plasma and react at the wafer surface, and by surface bombardment by high-energy positive ions. Since carbon is not present, polymer formation is difficult. Anisotropy also is aided by the addition of SiF\(_4\) and reduction of the self-bias voltage. The fluoride also cleans up undesirable residues in this manner. Aluminum chlorides and oxychlorides (remember O\(_2\) is released by Al\(_2\)O\(_3\) etching) are usual by-products. These are hygroscopic and form acidic compounds with water vapor on exposure to air, compounds that can severely corrode the vacuum system and fixtures. Usual practice follows the etch cycle with a gas change to a fluorine-bearing type, causing the chlorides to become fluorides which are not hygroscopic. Wafer-shufflers that do not require exposure of the vacuum system to air after each etch cycle further reduce the water vapor problem.

Copper and silicon in the aluminum alloys can lower the aluminum etch rate \( r_s\). Silicon tetrafluoride (SiF\(_4\)) or small amounts of O\(_2\) or He can reduce this problem. When CVD SiO\(_2\) is used to mask the Al, raising plasma power lowers the Al/Si to SiO\(_2\) etch rate ratio because increased ion bombardment of the oxide raises \( r_s\).

A chlorocarbon, CCl\(_4\), in place of Cl\(_2\) with the BCl\(_3\), gives another chemistry. The carbon permits the formation of CCl\(_3\) radicals that attack the Al\(_2\)O\(_3\), and aids in forming sidewall-protecting polymers. The organic residues on the wafer after etch can be oxidized by adding a small amount of oxygen with the chlorides, or by a separate O\(_2\) ashing step.

10.19 Radiation Damage

In both sputter and reactive ion etching high-energy ions strike the wafers, so we expect forms of radiation damage that are common to ion implantation. Ion bombardment in RIE is combined with chemical reaction effects. As a general rule, these effects may be reduced by lowering the self-bias voltage to lower incident ion energy, but this lowers the ion directionality and so affects the degree of anisotropy. Other reactor configurations that involve the use of magnetron action can also help [41]. (Section 12.11). Consider some of these effects beginning with ones that are simpler to correct.

Incident ions also can sputter materials from the walls and fixtures of the system, especially from the cathode that supports the wafers. This sputtered material can fall back onto the wafers, causing contamination. The effect is minimized by (1) reducing self-bias to lower ion momentum (as noted above this may be counterproductive for etching), (2) using metals for the fixtures that have low sputter yields, such as aluminum or stainless steel (but Al must not be used with chlorides since it will be etched), and (3) protecting the fixtures with a low-yield, passivating coating.

Combined chemical and sputter effects leave polymeric residues on the wafers. Post-etch ashing in an O\(_2\) plasma can eliminate these by oxidizing them into volatile products. This can be a no-\( \delta \)-field, isotropic process [42].

Bombardment by ions and photons from the plasma can affect the surface density of states, \( N_{SS} \), on silicon. This becomes important in silicon gate MOS
(SIGMOS) technology when the polysilicon gate is etched, because device transconductance and threshold voltage are changed by $N_{SS}$ shifts. Changes in $N_{SS}$ also can affect the barrier height in Schottky diodes, and the behavior of ohmic contacts between Si and a metal such as Al or Au. Thus, all these properties are affected by the choice of plasma gases and the self-bias voltage. In RIE, CF$_4$ increases $N_{SS}$ more than do SiCl$_4$/Ar or CHF$_3$.

Ion and photon bombardment of oxides can create traps in the oxide, too. These may be neutral or may become positive by trapping a hole. In either case, drifting MOS threshold voltage can result. For large geometry SIGMOS devices these traps may be removed by a furnace anneal in N$_2$, at about 1000°C, but small-scale devices need a shorter anneal such as afforded by lasers.

A catastrophic example of radiation damage is the destructive dielectric breakdown of a SIGMOS gate oxide when the anisotropic etching of the polySi gate is terminated [43]. A typical cross section of the materials involved is shown schematically in Figure 10–10(a). The electric charges developed by positive ions and self-bias during RIE are shown on the pertinent layers; the dielectric PR and gate oxide regions behave as capacitances and are shown at (b) as two equivalent capacitors in series with $C_E$. During RIE the largest voltage will appear across the external capacitor, $C_E$. Typically, gate oxides are thin, say 1000 Å or less, making them very susceptible to dielectric breakdown. The trend toward ever smaller devices means ever thinner gate oxides, which exacerbates the problem. When the rf is turned off to terminate the etch, a transient readjustment of charge takes place—positive charge in the gate layer and negative charge from $C_E$ flowing toward the gate oxide.

As a result the equivalent oxide charge $Q_{ox}$ on the capacitance $C_{ox}$ increases. Since $V_{ox} = Q_{ox}/C_{ox}$, $V_{ox}$ increases with $Q_{ox}$. Due to the small size of $C_{ox}$, this voltage can be large, so large, in fact, that the gate oxide can rupture, a condition that becomes more likely as device size decreases. For example when the size of an MOS device is scaled down linearly by a factor of $1/S$, where $S$ is greater than one, all linear dimensions are multiplied by $1/S$. Oxide thickness $d$ scales by $1/S$ and area by $1/S^2$; hence the capacitance, $C_{ox} = eA/d$, also scales by $1/S$. Its decrease makes $V_{ox}$ even larger for a given $Q_{ox}$.

Watanabe and Yoshida suggest these remedies. (1) Use an insulated cathode so the wafer does not contact that electrode. Remember that with r-f excitation dc does not flow in the series circuit; electrical contact between the back side of the wafer and the electrode is not required. This effectively introduces a fourth capacitor in the string, located between $C_{ox}$ and $C_E$. (2) At the end of the etch cycle, raise the gas pressure before turning off the r-f power. This lowers the self-bias, less voltage is on $C_E$ before the transient, and the transient voltage across $C_{ox}$ is lowered.

### 10.20 Compound Semiconductors

Gallium and indium fluorides tend to be nonvolatile at normal etching temperatures, even though arsenic and phosphorus are highly volatile. As a result, gallium arsenide (GaAs), indium phosphide (InP), and similar compound semiconductors cannot be etched in barrel and PPE configurations under isotropic conditions in fluoride plasmas such as CF$_4$ and SF$_6$/O$_2$. Rather, chlorine-bearing gases are used, such as BCl$_3$/Cl$_2$, CF$_2$Cl$_2$, CF$_2$Cl$_2$/O$_2$, and CCl$_4$ for GaAs and Cl$_2$/O$_2$, Cl$_2$/O$_2$/Ar, and CCl$_4$ for InP, because Ga- and In-chlorides are salts with high vapor pressures at etch temperatures, and As- and P-chlorides are liquids or gases.

Both gallium arsenide and its native oxide, "GaAs-oxide," will etch in COCl$_2$, PCI$_3$, and HCl. The gas combination 8 Cl$_2$/2 O$_2$ will etch both InP and GaInAs.

Under RIE conditions, where additional ion bombardment is present, GaAs will etch in CF$_4$ and in CHF$_3$. The availability of C here allows polymer formation, and bombardment generates surface defects. A study of GaAs Schottky diodes shows that RIE produces changes in the barrier heights and breakdown voltages, probably the results of radiation damage. Further changes in device characteristics can come about due to loss of stoichiometry in the compound semiconductor layers under bombardment [41].

An interesting case arises in the fabrication of gallium arsenide phosphide (GaAsP) LEDs (light emitting diodes). Since As and P have very high vapor pressures, encapsulation of the compound semiconductor, usually in Si$_3$N$_4$, is necessary to prevent dissociation at diffusion temperatures. The nitride also can serve as a diffusion mask and it is patterned with windows by RIE in

![Figure 10–10. Dielectric breakdown in SIGMOS devices. (a) The stack of materials. (b) Equivalent series circuit of three capacitors. After Watanabe and Yoshida [43]. (Courtesy of Solid State Technology, PennWell Publishing Company, Copyright 1984.)](image-url)
Problems

10–1 A thin Al interconnect layer is deposited on an oxide-coated Si wafer.

(a) What wet etch can be used to selectively pattern the Al layer? Need it be anisotropic? Explain.
(b) Repeat for the SiO₂ layer.

10–2 You wish to etch a row of holes, 5 μm in diameter, completely through a 20-mil-thick Si wafer.

(a) What wafer orientation should be used? Why?
(b) Recommend three wet etches that may be used.
(c) Explain how this technique may be used to fabricate an inductance “coil.”
   Use sketches.
(d) Do you expect a large or small L value? Why?

10–3 Part of a wafer is to have Al interconnects, the other part Al. Explain the advantages of the lift-off technique for patterning the metals in this situation. Positive resist is used.

10–4 How does a load-lock system help in spot monitoring during plasma etch?

10–5 Explain why N₂ and O₂ have different glow discharge colors.

10–6 A W layer is to be selectively dry etched without attacking some Al interconnects. Specify the gas(es) and reactor type to be used. Explain.

10–7 Explain why the positive ions travel a distance significantly less than d in sputter etching.

10–8 Verify Eq. (C.3.1).

References

Chapter 11

Lithography

Etching is used to remove unwanted material and to delineate the regions where material is to be added. The pattern of these regions is specified by the device or circuit designer. The fundamental function of lithography is to transfer the pattern to the wafer itself.

One version of photolithography was outlined in Chapter 1, where the pattern, stored in the emulsion of a glass based photomask, was transferred to the wafer photoresist (PR) by ultraviolet (UV) radiation in a contact aligner. Alternative versions are discussed in this chapter.

Lithography is named for the specially prepared stone (Greek: lithos) that has been used to transfer a drawing to the printed page for over 200 years. Much of the technology and terminology in the semiconductor industry is derived from the graphic arts or other industries that specify dimensions in British units. Thus we find mixed units in semiconductor industry usage: for example, mils and μin as well as Å, mm, μm, and nm.

11.1 Pattern Transfer Processes

In Chapter 1 we considered use of (PR) as a stop to delineate regions to be etched. Our concern now is to transfer the proper pattern to the resist. Figure 11–1 shows several alternative ways to do this, tracing the major steps from the designer’s sketch to the patterned resist. The actual transfer during exposure (print or write) is by radiation. The figure shows two wavelength ranges: Path I uses the ultraviolet (UV), where the whole process is referred to as photolithography. Path II uses a nonoptical range, furnished by x-rays or e-beams, where wavelengths are shorter by orders of magnitude.

The semiconductor industry strives to reproduce ever smaller features on chips, i.e., to increase the resolution, or fineness of pattern detail. In this context, we shall use ρ to designate the smallest feature dimension, line width, or space, that can be reproduced in the process. The smallness of ρ is limited...
Week 9

Metallization
Chapter 15

Device Isolation, Contacts, and Metallization

This chapter will cover two of the most basic functions of an IC. The first half will discuss device isolation, that is, the ability of the technology to allow each device to operate independently of the state of the other devices. Unless the technology is limited to building discrete devices, this is an essential function. The second half of the chapter will discuss device interconnection, including the metal-to-semiconductor contact. Again, only a brief review of the device physics will be provided to point out the implications for the technology.

15.1 Junction and Oxide Isolation

To fabricate ICs, some sort of isolation module must be developed [1]. The metrics of such a module are density, process complexity, yield, planarity, and parasitic effects. Trade-offs exist among these parameters. The result of this simple consideration is that no single isolation module is the most appropriate for all applications. A high-density, high-complexity module is the most appropriate choice for DRAMs, but a simpler, lower density isolation module can be used for low-cost TTL. This section will present several options for device isolation in silicon and GaAs technologies and evaluate each according to the metrics listed above.

The first ICs were bipolar. To understand the necessity for isolation, consider Figure 15.1. From the discussion on unit processes, it is easy to understand how a modest bipolar technology can be constructed. To do this, simply diffuse a deep p layer and a shallower N+ layer into an n-type substrate. The substrate acts as the common collector. Assume that an insulating layer is then deposited on the substrate, contacts are patterned and etched, and the interconnect is applied and patterned. The first question is, how close can the transistors be placed; that is, what is the packing density?

The emitters of two adjacent transistors are automatically isolated from each other by the fact that each emitter is totally enclosed in the base diffusion. For the base layers to be isolated from each other, there must be a large energy barrier between the holes in the two base regions. As the height of this barrier decreases, the leakage between the bases increases exponentially. A simple and convenient measure of isolation is that the depletion layers associated with the two base collector junctions do not
touch. To get an estimate, assume a simple one-sided step junction. Then the depletion layer thickness is

$$W_D = \sqrt{\frac{2k_e e_s}{qN_D} (V_{bi} + V_{CB})} \quad (15.1)$$

where $k_e$ is the relative permittivity of silicon, $V_{bi}$ is the built in voltage of the junction, which for nondegenerate doping is given by

$$V_{bi} = \frac{kT}{q} \ln \frac{N_A N_D}{n_i^2} \quad (15.2)$$

and $n_i$ is the intrinsic carrier concentration (about $10^{10}$ cm$^{-3}$ at room temperature). For a typical collector concentration of $10^{16}$ cm$^{-3}$, this reduces to

$$W_D = 0.36 \, \mu\text{m} \sqrt{V_{bi} + V_{CB}} \quad (15.3)$$

The isolation distance then is just $2W_D$, or about 1.8 $\mu$m for a maximum $V_{BC}$ of 5 V, or about 2.4 $\mu$m, to ensure isolation for a 10-V $V_{BC}$. Of course, one must also consider lateral diffusion. The base regions must be at least this far apart at the end of the process, not just on the photomasks.

To extend the approach to completely isolate the transistors, simply add a third diffusion to form the collector and use p-type substrates. To minimize the capacitance associated with the collector and to ensure that the collector doping is independent of variations in the substrate doping, a lightly doped substrate must be employed. Assume that the substrate is now $10^{15}$ cm$^{-3}$ p-type. Then

$$W_D = 1.14 \, \mu\text{m} \sqrt{V_{bi} + V_{CS}} \quad (15.4)$$

For a 10 V bias (2× a supply voltage of 5 V), this would require about a 4-$\mu$m spacing around each diffusion or a total of 8 $\mu$m between devices after processing is complete. Taking into account lateral diffusion, a 12-$\mu$m collector-to-collector spacing might be required. In these processes, it is essential to ensure that the junctions remain reverse biased. This idea of using reverse-biased junctions to isolate the devices, first patented in 1959 [2], was the earliest practical method of device isolation.

To get an estimate for the isolation length required for a technology, consider the densest portion of the circuit. Then referring to Figure 15.2, read off the maximum permissible isolation distance. For example, if the densest portion of the circuit has $10^3$ transistors/cm$^2$ and 50% of the area is active in that region, the device separation must be 10 $\mu$m or less. In many designs, the circuit density is limited not by the isolation distance, but by the metal density. The density goal of the isolation
module, then, is that the separation necessary to make
the interconnect, and not the device isolation, the lim-
iting factor for the circuit density. Generally, for a ran-
don logic circuit, the isolation distance should be less
than twice the first metal pitch (pitch is minimum
linewidth + minimum line space). For memories and
other highly structured logic, the isolation distance
should be no more than the first metal pitch.

The other important point to make about Figure
15.2 is that for isolation-limited circuits, the circuit
density is a sensitive function of the isolation distance.
Reducing the isolation distance by a factor of 3 allows
an order of magnitude increase in the circuit density. Consequently, companies that deal in isolation
driven circuits such as memories, have developed elaborate isolation techniques.

Density, however, is not the only consideration in designing the isolation module. Go back to
Figure 15.1, but add an insulating layer and a metal line that happens to pass over, but does not con-
tact both transistors (Figure 15.3). A parasitic MOS transistor now exists. The collectors of the two
transistors act as the source and drain of the MOSFET, and the metal line acts as the gate. If a large
enough positive bias exists on the line, the surface of the area under the line may invert, turning on
the parasitic MOSFET. This effect will short out the two collectors, even if they are sufficiently far
apart so that the depletion regions do not touch.

Neglecting oxide charge, the MOSFET threshold for NMOS transistors on p-type substrates is
given by (see Section 16.1)

\[ V_T = \phi_{ms} + 2\phi_f + \frac{k_f}{k_{ox}} \sqrt{\frac{4qN_A}{k_T \epsilon_0}} \phi_f \]  \hspace{1cm} (15.5)

where

\[ \phi_f = \frac{kT}{q} \ln \left[ \frac{N_A}{n_i} \right] \]  \hspace{1cm} (15.6)

and \( \phi_{ms} \) is the metal semiconductor work function that depends on the type of metal used in the inter-
connect and the doping concentration in the substrate. For aluminum on p-type substrates, \( \phi_{ms} \) varies
from about \(-0.8 \) to \(-1.0 \) V. Since the intrinsic carrier concentration increases with temperature, the
parasitic threshold voltage may shift by several volts when operating at high temperatures. Further-
more, the source-to-drain current of an MOS device increases exponentially with gate voltage in the
subthreshold regime. Normal practice is to make the threshold of these parasitic transistors at least
2 \( \times \) (and preferably 3 \( \times \)) the supply voltage. This ensures that they will not turn on, even in the pres-
ence of an excess supply voltage and/or voltage spikes on the supply line, nor will the leakage cur-
rents be excessive.

Figure 15.4 shows a plot of threshold voltage versus the substrate concentration with oxide
thickness as a parameter. The oxide thickness was varied from 0.2 to 1.0 \( \mu \)m in 0.2-\( \mu \)m increments.
Also included is the threshold voltage if a total fixed charge of \( 10^{11} \) cm\(^{-2} \) is present. To achieve a suita-
ably large parasitic threshold, one must select a thick field oxide and/or a large substrate concentra-
tion. The large substrate concentration degrades the performance due to junction capacitance. The
thicker oxide improves both performance and parasitic turn on, but obviously due to the effect of


Figure 15.4  $V_T$ as a function of the $N_n$ assuming no $\phi_{ms} = 0$. Solid lines are a perfect interface, dotted lines are for $N_n = 10^{11}$ cm$^{-2}$.

Assuming a 2-μm resolution lithography such as might be used in such a simple technology and recognizing that lateral diffusion that would occur, the final width of the guard ring is over 5 μm. This requirement ultimately limits improvements in packing density.

We can now analyze this result in terms of the metrics described above. Obviously, junction isolation is simple and produces a planar isolation. Because of its simplicity, it would have a high yield. Density, however, is not large and must be traded off against the parasitic capacitance of the collector substrate junction. Increasing the substrate concentration increases the density but also increases the capacitance. Guard rings can improve the situation, particularly concerning the prevention of turning on parasitic MOS devices. For these reasons, in spite of its low density, junction isolation is still used in some low-cost, low-density applications.

15.2 LOCOS Methods

The most straightforward way to produce a thick field oxide is by growing one before device fabrication, then etching holes in the oxide and fabricating the devices in these holes. This approach has two serious shortcomings. The first is the topology that is created. The step coverage for subsequent depositions will be poor and the photolithography will suffer. This is extremely serious if small features are to be printed. The second drawback is less obvious. On lightly doped substrates, a guard ring must be implanted to increase the parasitic threshold voltage. Unless very high energies are used, the implant must be done before the oxidation. Diffusion during oxidation may also be enhanced by point defects released during the oxidation process. Combined with alignment tolerance requirements, this will significantly reduce the density of the IC.

The isolation approach that has become the standard of silicon IC fabrication is local oxidation of silicon or LOCOS [3]. Local oxidation is essentially an outgrowth of junction isolation and
addresses both the isolation and the parasitic device formation concerns. A thin oxide is first grown and a layer of Si$_3$N$_4$ deposited on the wafer, usually by LPCVD. After the nitride is patterned, a field implant may be done to increase the threshold voltage of the parasitic MOSFET. Then the photoresist is stripped and the wafer is oxidized (Figure 15.6). The nitride acts as a barrier to the diffusion of the oxidant, preventing oxidation in selected regions of the silicon. A thin oxide will also be grown on top of the nitride. This is important because it limits the minimum nitride thickness to about 1000 Å and because the oxide must be removed before the nitride can be stripped after the field oxidation.

Since oxidation consumes 44% as much silicon as it grows, the resultant oxide is partially recessed and has a gradual step onto the field that is easy for subsequent layers to cover. If the silicon is etched before the field implant, the field oxide can be made fully recessed, resulting in a nearly planar surface. Figure 15.7 shows the growth of a local oxide along with a cross-sectional view of a completed LOCOS structure. The process leaves a characteristic bump on the surface, followed by a gradually narrowing oxide tail into the active area. The structure is called a bird’s beak for obvious reasons. The bump, or bird’s head, is particularly pronounced in recessed structures.

The purpose of the thin pad oxide layer under the nitride is to reduce the stress that occurs in the silicon substrate during oxidation. This stress is due to the mismatch of the thermal expansion coefficients of the substrate and the nitride and due to the volumetric increase of the growing oxide. At high temperature, viscous flow of the oxide greatly reduces the stress. A great deal of work has gone into optimizing the thicknesses of the oxide and nitride layers. If the stress exceeds the yield strength of silicon, it will generate dislocations in the substrate. A thicker pad oxide will lower the stress in the substrate. The minimum pad oxide thickness that can be tolerated without dislocation...
formation is about one-third the thickness of the nitride [4]. This defect protection must be traded off against increased lateral encroachment of the oxide, which occurs due to lateral diffusion of the oxidizing species through the pad oxide. A nitride to thermal pad oxide thickness ratio of 2.5:1 produces a lateral encroachment or bird’s beak, approximately equal to the thickness of the field oxide.

One concern of the LOCOS process is the white ribbon or Kooi nitride effect [5]. In this situation, a thermal oxynitride forms at the surface of the silicon under the edges of the nitride pad. White ribbon is caused by the reaction of Si$_3$N$_4$ with the high-temperature wet ambient to form NH$_3$, which diffuses to the silicon/ silicon dioxide interface where it dissociates. The surface texture caused by these nitrides can be seen as a white ribbon around the edges of the active area. This defect leads to a reduced breakdown voltage in subsequent thermal oxides (such as gate oxides) in the active region.

The existence of the bird’s beak has two important consequences from a device standpoint. Often the active region defines the edge of the device in at least one direction. Then encroachment reduces the active width of the device, reducing the amount of current that a transistor will drive. A more subtle effect is due to the field doping. The field oxidation causes the field implant to diffuse into the edge of the active region. Figure 15.8 shows a schematic of an MOS transistor along its width, immediately under the gate. If the transistor is narrow enough, the additional dopant from the field diffusion will increase the threshold voltage of the device, reducing its drive current. Known as the narrow channel effect, it is important in extremely dense technologies such as memories.

Various methods have been proposed to modify the LOCOS process to reduce the bird’s beak length. The simplest is the use of materials other than a thermal oxide for the pad layer. A sandwich
of thermal oxide and polysilicon is extremely effective in reducing lateral encroachment [6] and has become quite popular. The polysilicon absorbs the excessive stress resulting from the use of thin pad oxides and very thick nitrides. Typically about 500 Å of polysilicon is deposited on top of 150 Å of thermal oxide, followed by 1500 Å of nitride. This structure produces a bird’s beak that is less than half the thickness of the field oxide [7]. Experiments describing the optimization of the poly-buffered process can be found in Ghezzo et al. [8] and Guldí et al. [9]. Although the length of the bird’s beak is reduced, the poly-buffered process does not solve the problem of lateral diffusion of the field dopant. To gain the full benefit of such a process, the field oxidation should be done at high pressure in a wet ambient to minimize the lateral diffusion. White ribbon effects in poly-buffered LOCOS is a concern, however. Care must be taken when etching back the nitride/poly/oxide layers [10] to avoid it.

First proposed in 1982, sidewall masked isolation (SWAMI) has generated a great deal of interest due to its ability to fabricate thick, highly planar field oxides with little encroachment. The SWAMI process is shown in Figure 15.9. A thermal pad oxide is grown, and a nitride layer is deposited. The layers are patterned and the silicon is etched to a depth of about half the desired oxide thickness. This can be done wet chemically with KOH, which produces a 60° slope on <100> substrates, or it can be done in a properly designed plasma etch. A second layer of nitride is then deposited and anisotropically etched. This leaves a sidewall of nitride on the slopes of the silicon, which effectively seals the active region. A thermal oxidation is done for the field oxide. Finally, the nitride and pad oxide are removed. SWAMI can produce thick field oxides with minimal encroachment, but retains the dopant diffusion problems of all LOCOS-based modules. Its complexity and the development of alternative isolation techniques have precluded SWAMI’s widespread application.

15.3 Trench Isolation

It became obvious during the 1980s that neither LOCOS nor any of its variations would be acceptable for ICs with transistor densities much greater than $10^7 \text{ cm}^{-2}$. This can be seen easily if one considers transistor size. Typical internal logic devices have a width-to-length ratio of about 4:1. For a 1-μm gate
length, this corresponds to a width of 4 μm. In that case, a lateral encroachment of 0.3 μm per side might be undesirable since it reduces the width of the transistor by about 15%, but it is probably acceptable. On the other hand, when the gate length is 0.18 μm, this same encroachment is almost 85%, a clearly unacceptable amount. It is now believed that the absolute minimum isolation distance for advanced LOCOS processes is about 0.8 μm from the edge of one N⁺/p junction to another [11]. The ultimate limitation is not surface inversion or simple punchthrough, but a reachthrough effect known as drain-induced barrier lowering.

Many new isolation approaches have been developed around the idea of etching away part of the substrate and refilling it with an insulator. These can be divided into two classes. Shallow etches with small aspect ratios were considered first [12]. These techniques can be thought of as similar to recessed LOCOS, but a deposited rather than a thermal oxide is used to fill the field regions. After deposition, planarization must be done to remove the unwanted oxide from over the active regions (Figure 15.10). Initially, to be effective, an additional photomask step was required for oxide removal. Due to problems of misalignment this was difficult to achieve reliably. Furthermore it is difficult to prevent inversion along the etched sidewalls [13], since the field implant is normally done at close to normal incidence. In a CMOS process this inversion may cause excess leakage between source and drain. Due to their geometry, bipolar transistors are not as prone to this problem as MOSFET’s.

Deep trench isolation modules (Figure 15.11) use trenches of fixed width. Typical dimensions are 0.18 to 1.0 μm in width and 2 to 5 μm in depth although trenches with widths of 0.25 μm and depths of 10 μm have been demonstrated [14] (see Figure 15.12). The smaller trench widths are particularly attractive for memory applications. The process is fabricated by starting from a standard LOCOS structure. After nitride patterning the trenches are etched. Trench isolation puts extreme demands on the etch process. It must have smooth walls at no more than 85° with respect to the plane of the wafer. More taper is highly desirable [15]. The trench etch is typically done by simultaneously depositing SiO₂ while etching silicon anisotropically. This creates a small cusp of SiO₂ at the top of the trench. The thickness of this cusp increases with time, producing the desired taper. The walls cannot undercut the mask, and should end in a rounded bottom. Sharp corners at the bottom of the trench will result in excess stress during oxidation, and ultimately defects in the oxide [16]. Next a field implant is done. Preventing sidewall inversion becomes even more difficult as the aspect ratio of the structure increases. An important feature of the implant, therefore, is that the beam be perpendicular to the surface of the wafer. As discussed in Chapter 5, new implanters have been developed to accomplish this [17]. Another, albeit much less popular, approach to solving this problem is the use of rapid thermal processing combined with planar dopant sources [18] or a doped CVD glass [19].

The implant is followed by a thin local oxidation (typically less than 1000 Å). In applications in which the trench fill is to be used as a charge storage capacitor, it is common practice to use much
thinner oxides to increase the capacitance. When the oxide is less than about 100 Å, however, the capacitor leakage current increases [20]. Finally, a layer of polysilicon is deposited and etched back. If the polysilicon is thick enough, it will fill the groove. Etching back to the substrate will leave this fill in place. A second thermal oxidation can be used to complete the process by oxidizing the upper part of the polysilicon in the groove. Often this is done as part of a LOCOS process. This minimizes leakage, it allows the fabrication of arbitrary isolation lengths (see below), and it offsets the junctions from the wall of the trench. To

**Figure 15.11** Deep trench isolation process schematic.

**Figure 15.12** Cross-sectional scanning electron micrograph of an extreme aspect ratio deep isolation process (after Rajeevakraum et al., (copy) 1991 IEEE).
minimize lateral diffusion, the channel stop implant associated with the LOCOS may be done after oxidation [21].

It is important to emphasize that the trench must be of a fixed width for a deep trench process to work. This can be made to work with the minimum N⁺/P⁺ separation, but if the design also calls for larger separations in some areas, an additional isolation (such as a standard LOCOS) technique must also be used. The result is a complex process that is difficult to control. If the deposition process is not done properly, a void may form in the center of the trench, trapping some material. This may also occur if too steep an angle is etched into the substrate during the silicon RIE. The void may pose a reliability problem. Deep trench isolation technology has demonstrated N⁺-to-P⁺ spacings less than 2 µm and N⁺-to-N⁺ spacings less than 0.5 µm. Referring to Figure 15.2, such an isolation technique will be adequate for device densities well in excess of 10⁹ cm⁻².

Deep trench isolation is extremely difficult to manufacture and it is difficult to integrate with random logic when arbitrary device spacings must be accommodated. The development of chemical mechanical polishing (CMP) has made previously rejected shallow trench isolation (STI) a viable process since it can remove the excess deposited oxide without a lithography step. As shown in Figure 15.13, the process begins with a pad oxide of 100 to 150 Å, followed by a layer 1500 to 2000 Å of LPCVD Si₃N₄. Next a field is patterned and the nitride, oxide, and silicon are etched. Typical etch depths are about 0.5 µm thick. Trench sidewalls are etched at 75 to 80°. If desired, a field implant can then be done now to prevent inversion under the trench, or it can be delayed until after the polishing step is sufficiently high energy is used. Next a thin (150 to 200 Å) layer of SiO₂ is grown thermally to reduce the etch damage on the sidewalls and round off some of the corners. A 0.9- to 1.1-µm layer of SiO₂ is then deposited, usually by high-density PECVD, and CMP is used to remove the excess oxide. The nitride serves as a polish stop for this step. Finally the nitride is removed and the pad oxide stripped in HF [22].

The integration of STI presents a number of challenging problems. Some of these problems are associated with the upper corner. Typically in an MOS transistor, the gate polysilicon stripe extends onto the field oxide to ensure a separation between the source and drain. If the corner of the STI is too sharp, the trench sidewall will invert (due to field concentration) leading to excess subthreshold

![Figure 15.13](image_url) Schematic of a shallow trench isolation module (after Chaterjee et al., used with permission, APS, 1997).
leakage. This is especially a problem if the CMP is overdone, that it, if the top of the planarized oxide is below the top of the silicon [23]. To avoid this problem the trench walls must be properly tapered and the top corner must be rounded. To achieve the desired rounding, the pad oxide is selectively removed by undercutting the nitride layer, then oxidized using high oxidation temperatures (~1100°C) and/or an ambient containing HCl. CMP dishing leads to a thinning of the field oxide [24] and may lead to a design rule for the maximum isolation distance and/or the use of dummy active areas [25].

15.4 Silicon on Insulator Isolation Techniques

The ideal method of device isolation would be to completely encase each device in an insulating material. Several techniques exist for doing this in silicon. Generically they are called silicon on insulator (SOI). All of these methods have suffered from problems related to defect density. For that reason SOI technologies in silicon have been relegated to small markets such as radiation hardened devices where extreme isolation is required. In some measure, this has changed in the last few years. Several techniques have improved dramatically and ULSI applications have been demonstrated. Furthermore, the use of a buried insulator can reduce device parasitic capacitance, increasing circuit speed. The most promising method, SIMOX, has been presented in Chapter 5 and so will not be repeated here. Instead we will concentrate on several other processes.

One of the first SOI methods developed is called dielectric isolation (DI). The DI process sequence, shown in Figure 15.14 [26], was developed to build high-voltage telecommunication ICs that required electrically isolated bidirectional switches. It has since gained popularity for other high-voltage and radiation hard digital applications. Deep grooves are first etched in the surface of the wafer. The wafer is oxidized and a very thick (>200-μm) layer of polysilicon is deposited. If substrate contact is needed in some islands, windows can be opened in the oxide before deposition. The deposition can be done by conventional CVD or by using a molten silicon spray deposition (MSSD) process [27]. The wafer is turned over and mechanically ground until the grooves penetrate through the wafer. Finally, the wafer is chemically polished and devices are fabricated in the isolated islands.

Dielectric isolation has several severe drawbacks. The wafers are not as planar as normal starting material. This not only impacts further processes such as lithography, but also results in varying silicon island thicknesses across the wafer. Wafers made using the DI process are expensive. Typical costs are well over $100 per 100-mm wafer, before any device fabrication. Finally, if KOH is used to etch the v-grooves, the isolation density is not very large.

Wafer bonding is an alternate approach to producing SOI wafers (Figure 15.15). In this process, two wafers are pressed together at high temperature until they fuse [28]. Alternatively, the wafers can be fused at low temperature by anodic bonding [29, 30]. If the wafers are oxidized before bonding, a layer of oxide remains at the center of the fused wafer. The wafer can be ground back down to thicknesses of 2 to 3 μm using standard grinding and polishing techniques. If thinner layers are required, additional processing can be done to

![Diagram of dielectric isolation (DI) process for forming silicon on insulator.](image)
produce submicrometer semiconducting films on top of the oxide. Devices can be isolated with a simple etch process that produces single-crystal islands on top of the insulating oxide. Also, as with DI isolation, wafer bonding suffers from high cost, but can produce high-density, well-isolated structures in CZ grade silicon.

15.5 Semi-insulating Substrates

There are far fewer papers considering device isolation in GaAs technologies than there are for silicon isolation. Due to the availability of semi-insulating substrates, one can easily achieve high-density, planar isolation. The only remaining question is how to separate the conducting islands from each other. The fact that a semi-insulating substrate is used in most GaAs ICs means that the devices are inherently radiation hard. A significant market for GaAs ICs involves defense and space-based applications that demand such hardening.

Most of the early semi-insulating GaAs material was made by chromium doping. Both chromium and oxygen have an energy level very close to the center of the GaAs band. If chromium is added in high concentrations (about $10^{17}$ cm$^{-3}$ in older material, now more commonly in the mid $10^{16}$ range) to the boule as it is grown, semi-insulating wafers will result. The chromium adds discrete states near the center of the band (Figure 15.16). Consider what happens if a single donor atom is added to a semi-insulating wafer with no other dopants. Normally, the dopant atom will ionize and the electron will go into the conduction band, even though the conduction band of the crystal is a higher energy state. The probability of ionization is nearly 1 at room temperature for most dopants because of the large number of empty states near the conduction band edge as opposed to the single state at the donor atom. For chromium-doped GaAs, much lower energy states exist at the chromium sites. The sites will therefore ionize, accepting the donor atom’s electron. Since the states are discrete and widely spaced, the electron cannot move. The crystal is therefore semi-insulating and the carrier concentration is approximately the intrinsic carrier concentration. Resistivities of $10^6$ Ω-cm have been reported. Not until the dopant concentration approaches that of the chromium will there be significant conduction in the crystal. Conduction will also occur when the injected carrier density exceeds the trap concentration. This may occur under illumination, or in GaAs metal semiconductor field effect transistors (MESFETs) in a process known as sidegating. The effect is seen as a reduction in the current of a device if an adjacent device is turned on. Sidegating is a problem for high-density GaAs ICs.
Many GaAs technologies, therefore, begin by producing a thin conducting layer on top of the semi-insulating substrate. Device isolation can be achieved by simply etching through the conducting layer, leaving islands of semiconducting GaAs. This simple method is known as mesa isolation. It has the disadvantage that the surface is no longer planar. To avoid this problem, many GaAs technologies use ion implantation. This method implants the field regions of the wafer with hydrogen or another ion. A $10^{13}$ cm$^{-2}$ dose and 100 keV energy are typical implant parameters. This implant is sufficient to destroy the lattice, producing resistivities as high as $10^7$ Ω-cm. The surface is also quite planar. A drawback of proton implantation is that the crystal will be regrown if any further processing is done at temperatures above 350°C. Proton implantation must therefore be done near the end of the process.

Semi-insulating substrates are not produced in silicon. Due to the smaller bandgap of silicon, it has a maximum resistivity of less than $10^8$ Ω-cm at room temperature. Many of the noble and transition metals have states near the center of the silicon gap, but their diffusivities are very large ($10^4$ μm$^2$/hr at 800°C and $10^6$ μm$^2$/hr at 1100°C for gold). As a result, it is very difficult to keep the impurity in the substrate from diffusing into the active region during processing. If the impurity reaches device junctions, enormous increases in leakage will result since the states that render the wafer semi-insulating are also very efficient recombination centers. In a recent paper, however, semi-insulating silicon has been grown epitaxially using oxygen diluted in argon along with silane [31].

In many GaAs applications, chromium doping has been replaced by nonstoichiometric liquid encapsulated Czochralski (see Section 2.7) wafers. Figure 15.17 shows the resistivity of the GaAs produced this way as a function of the arsenic mole fraction in the melt [32]. For compositions close to stoichiometric, the wafers are semi-insulating with bulk resistivities of $10^7$ to $10^8$ Ω-cm. The semi-insulating property is believed to be due to the presence of mid-$10^{16}$ cm$^{-3}$ deep donor levels called EL2 sites [33]. The atomic structure of these sites is believed to be related to arsenic atoms on gallium sites (As$_{Ga}$). Although these defects introduce a large etch pit density in undoped semi-insulating wafers, they do not seem to affect IC yield. The B$_2$O$_3$ plug used in LEC growth also tends to getter silicon impurities that would otherwise dope the GaAs substrate n-type. The EL2 center has the interesting property that when illuminated by visible or near-IR light while at low temperature, the defect state is transformed into a higher energy metastable state [34]. This second state has very different properties than the original defect. EL2 centers near the surface of the wafer may also be removed by high-temperature hydrogen anneals.

One interesting comparison between GaAs and silicon isolation methods is that GaAs technologies do not generally use low dielectric constant insulators, whereas they are emphasized in most silicon technologies to minimize parasitic capacitance. Many GaAs technologies are designed for microwave applications. Then, it is not the absolute capacitance that is critical but the characteristic impedance of the interconnect. This must be impedance matched to the

---

**Figure 15.17** The resistivity of GaAs substrates as a function of the melt stoichiometry (after Ferry).
devices to ensure the maximum power transfer. To achieve this characteristic impedance, the GaAs wafers are often thinned after processing, and a ground plane is deposited on the back of the wafer. Through wafer vias are used to contact the ground plane.

15.6 Schottky Contacts

Later sections of the book will deal with the fabrication of specific transistor structures. To efficiently connect semiconductor devices to the outside world, contacts must be made between the semiconductor and metal lines. Two types of contacts are often used: ohmic contacts and rectifying or Schottky contacts. In ideal ohmic contacts, the current varies linearly with the applied voltage. It is also implicitly assumed that ohmic contacts have low resistance. To transfer as much current as possible from the device to charge the various capacitances of the circuit, the contact resistance must also be a small fraction of the device resistance. Schottky contacts, however, should act as perfect diodes. They should have a very low resistance in the forward direction and an infinite resistance in the reverse direction. They should have a well-defined and perfectly reproducible “on voltage.” The reader should be aware that although semiconductor technologists often use these descriptions for metal to semiconductor contacts, real contacts are neither perfectly ohmic nor perfectly rectifying. If due care is taken in the fabrication, however, good approximations of the ideal can be achieved. This section will deal with the fabrication of Schottky contacts. The next section will discuss process modules for making ohmic contacts.

Schottky barrier contacts have a variety of applications related to voltage clamping and controlled diode drops. It has the advantage of turning on (and off) more quickly than a p–n diode, and the turn on voltage can be selected during the fabrication sequence. Figure 15.18 shows a typical application of a Schottky diode in a bipolar silicon technology. The base and collector of an NPN bipolar transistor are clamped through the Schottky diode. The turn on voltage of the diode is set to a few tenths of a volt less than the turn on of the base collector diode. If the base is suddenly driven higher than the collector, the base voltage will be limited to the Schottky diode on voltage. This prevents the transistor from being driven into a strong saturation condition from which it would take a long time to recover.

Figure 15.19 shows a band diagram for a Schottky diode on a p-type substrate. \( \phi_m \) is the metal work function, the voltage necessary to remove an electron from the surface of the metal. \( \chi \) is the electron affinity, the voltage necessary to remove an electron from the conduction band minimum of the semiconductor. \( \phi_s \) is the voltage difference between the Fermi energy of the semiconductor and the vacuum level. When the metal and the semiconductor are brought into contact, charge will flow from one to the other until an electric field builds up that is sufficient to prevent further current flow. This occurs when the barrier at the interface reaches \( \phi_{m} - \phi_{s} \). If a Schottky contact is needed, the metal must be chosen to create such a barrier. If, for example, this metal were chosen for a Schottky contact to a lightly doped n-type semiconductor, electrons in the metal, which has a higher Fermi energy than the semiconductor, flow readily into the semiconductor. This leads to an electric field that will assist the movement of carriers across the interface. Then a Schottky diode will not be formed. Ideally, if \( \phi_{m} > \phi_{s} \) and the semiconductor is n-type, or if \( \phi_{m} < \phi_{s} \) and the semiconductor is p-type, a Schottky diode will result.

The current across a Schottky diode is determined by thermionic emission. That is, it is determined by the fraction of carriers that has sufficient energy
Band diagram for an ideal Schottky contact: before contact (right) and after contact (left).

to surmount the barrier. It is given by the equation

\[ I = I_o \left( \exp \frac{qV}{nkT} - 1 \right) \quad (15.7) \]

where

\[ I_o = RT^2A \exp \left[-\frac{\phi_m}{kT}\right] \quad (15.8) \]

In this equation \( A \) is the area and \( n \) is the ideality factor of the diode. \( R \) is Richardson's constant, which has values of 110 and 32 A cm\(^{-2}\) K\(^{-2}\) for n- and p-type silicon, respectively, and 8 and 74 A cm\(^{-2}\) K\(^{-2}\) for n- and p-type gallium arsenide. It is clear from this equation that the current density depends strongly on the barrier height.

Figure 15.20 shows the measured barrier heights for both silicon and gallium arsenide as a function of the metal work function. The experimental points should fall on a line with unity slope. The offset would be due to the different electron affinities and (potentially) Fermi levels for the two semiconductors. Instead of a unity slope, a slope of about 0.25 is seen for silicon and about 0.1 is seen for GaAs. From an empirical point of view we need to modify Equation 15.8 to read

\[ I_o = RT^2A \exp \left[-\frac{\phi_b}{kT}\right] \quad (15.9) \]

where \( \phi_b \) is the effective barrier height.

To understand this discrepancy, one needs to go back to the basic assumptions in the model. In presenting these equations a perfect interface was implicitly assumed. In most processes that make
Schottky diodes, the metal is not deposited under ultrahigh vacuum conditions. Rather, an active region of the semiconductor is first exposed, for example, by etching a hole in an insulating layer. Then the wafer is placed in a vacuum system to deposit the metal layer. During this time the surface of the wafer will oxidize and receive a coating of carbon and other contaminants from the residual pump oil vapor in the chamber and from the air in the room. To improve the process, many sputtering systems can presputter the surface of the wafer before deposition. This, however, does not ensure that a perfect surface is achieved. The sputtered species are not volatile; many will redeposit on the surface of the wafer. Furthermore, material will be sputtered wherever the plasma touches a surface. Thus, it would not be surprising to find Fe, Ni, and Co on the surface of the wafer. The plasma also damages the surface of the wafer. Both contamination and damage have an impact on the Schottky diode. Even without these defects, the semiconductor surface itself represents an important “ imperfection” from the standpoint of the electronic structure. Unless the metal can be epitaxially grown on the surface of the semiconductor, the barrier height and the metal semiconductor work function may be considerably different.

PtSi is commonly used as a Schottky contact for lightly doped n-type silicon. The cross section of a Schottky clamped bipolar transistor is shown in Figure 15.21. The diode is formed by first etching holes down to bare silicon. A dip in dilute HF is normally done just before deposition. A platinum layer is then sputtered onto the surface of the wafer. Typical thicknesses are 300 to 600 Å. It is critically important in obtaining a good surface morphology that the silicon surface is clean before deposition [35], and that little oxygen and water vapor are present during the deposition step. Recall that from kinetic theory the flux of residual gasses is given by

$$ J = \frac{p}{\sqrt{3kTm}} $$

(15.10)

For nitrogen at 10^{-6} torr, the bombardment rate is 0.17 atoms Å^{-2} sec^{-1}. Thus, if one assumes that all of the atoms that strike the surface of the wafer stick, a monolayer of atoms will deposit within 1 sec.
It is therefore impossible to ensure even a reasonably clean surface unless ultrahigh vacuum methods are used. Platinum will reduce a thin oxide however, breaking it up in the anneal. As a result, a standard presputter is sufficient to ensure that a uniform reaction will proceed. It is also critically important to ensure that no oxygen leak exists into the sputtering system. The deposition rate is normally kept low to achieve good process control and uniformity. A small air leak or a significant water partial pressure in the chamber can produce highly oxygen contaminated films. This leads to high resistivity and poor reproducibility in the diode characteristics.

After deposition, the wafer is annealed at about 550°C in a furnace to allow the film to react with the silicon. Toward the end of the process, the furnace is switched to an oxidizing ambient and a thin layer of SiO₂ is grown over the platinum silicide. The platinum itself does not oxidize. After reaction, the wafer is immersed in a dilute solution of aqua regia at 85°C to selectively remove the unreacted platinum. The oxide over the silicide protects it during this etch. Finally, the wafer is dipped in HF to remove the oxide. After silicide formation, the wafer temperature must not be allowed to exceed 800°C, otherwise the morphology of the PtSi layer quickly degrades. Since the silicon surface is consumed during the reaction process, a clean interface results between the semiconductor and the wafer. Typical PtSi to lightly doped n-type Si barrier heights are 0.85 ± 0.05 V.

Schottky diodes are more widely used in GaAs technologies than in silicon. The fundamental reason is that GaAs does not have the exceptional oxide that silicon has. Thus, the basic GaAs field effect device is not the MOSFET but the MESFET. For the MESFET to work well, the gate electrode must make a Schottky contact with the channel. The pinchoff voltage of the FET depends directly on the barrier voltage of the Schottky diode. Furthermore, all of the metals used in GaAs must be fully deposited; they are not reacted with the substrate. Figure 15.22 shows a typical cross section of a GaAs MESFET. WSi and WSi/W are commonly used gate electrodes, although aluminum, chromium, titanium, and molybdenum have also been used. CoAl alloys can also be produced that have good thermal stability and Schottky barrier heights as large as 0.9 V [36]. WSi, can be deposited using CVD [37], by sputtering from a composite target, or more commonly by cosputtering W and Si [38]. The optimal composition appears to be WSi₀.₆ [39]. Gallium does not diffuse rapidly in WSi unlike many metals, preventing degradation of the interface. WSi is also attractive for this application for the same reason polysilicon is attractive for MOSFETs in silicon. It is a refractory metal that will tolerate subsequent high-temperature processing [40]. This means that the gate can be used as an implant mask for the source and drain, self-aligning these diffusions to the gate (see Chapter 17). An 0.8% atomic alloy of Al in W is also an attractive gate material for GaAs MESFETs. It is thermally stable, has a larger barrier than W [41], and has lower resistance than WSi [42].

As shown in Figure 15.20, the barrier voltage for a Schottky diode in GaAs is about 0.8 V and is nearly independent of the metal. Many compound semiconductors have similar behavior. This represents a serious limitation for MESFET circuits. InP based materials, for example, have several electronic advantages over GaAs, but there is no suitable material that will provide a Schottky barrier larger than 0.4 V [43], virtually precluding their use in MESFETs. Even in GaAs, the gate cannot be forward biased much beyond 1 V or significant gate leakage will result. This leakage increases power consumption and reduces performance. The Schottky barrier also depends strongly on the interface quality. A poor interface results in a larger ideality factor and a reduced barrier height. It has been shown that a sputter cleaning of the interface can greatly reduce the nonuniformity of the pinchoff voltage of GaAs MESFETs [44].

![Figure 15.22](image_url) Typical GaAs MESFET structure.
Several methods have been attempted to increase the Schottky barrier height for GaAs. One technique is to use a narrow band of doping called a charge sheet or delta doping at the surface of the semiconductor [45]. The carriers will be depleted, but the remaining ionic charges are sufficient to shift the barrier height by as much as a few tenths of an eV [46]. Another method is to use a monolayer of alternate materials [47] or heterostructures with a wide bandgap material (AlGaAs) between the GaAs and the WSi. This latter technique produces a barrier height of 1.0 to 1.2 V [48]. Either method requires the use of MBE or MOCVD growth.

15.7 Implanted Ohmic Contacts

Often one is interested in making low-resistance ohmic contacts to the semiconductor. The specific contact resistance can be defined as

$$R_c = \left[ \frac{\partial J}{\partial V} \right]_{V=0}^{-1} \quad (15.11)$$

From Equation 15.9, for thermionic emission this becomes

$$R_c = \frac{k}{qRT} \exp \left[ \frac{\phi_b}{kT} \right] \quad (15.12)$$

Since the barrier height depends logarithmically on the dopant concentration in the substrate, $R_c$ should decrease linearly as the doping concentration increases. This is true up to doping concentrations of $10^{18}$ to $10^{19}$ cm$^{-3}$. For more heavily doped substrates $R_c$ depends strongly on the doping concentration, falling by orders of magnitude. The reason for this behavior is that thermionic emission over the energy barrier is no longer the dominant transport mechanism.

The width of the depletion region at zero applied bias can be calculated using

$$W_D = \frac{2k_\varepsilon \varepsilon_0 \phi_b}{qN_A} \quad (15.13)$$

For heavily doped substrates this width becomes sufficiently small that carriers can tunnel through the barrier (Figure 15.23) rather than be limited by thermionic emission over the barrier. The specific contact resistance in this region can be approximated by

$$R_c \approx A_o \exp \left[ \frac{C_2 \phi_b}{\sqrt{N_D}} \right] \quad (15.14)$$

where

$$C_2 = \frac{4\pi}{h} \sqrt{m^*_e \varepsilon_0} \quad (15.15)$$

$h$ is Planck’s constant and $m^*_e$ is the effective mass of the electron in the semiconductor. In contacts to p-type substrates $m^*_e$ is replaced by $m^*_h$, the effective mass of the hole.
From Equations 15.14 and 15.15, it is apparent that one should make the substrate doping as large as possible to reduce the contact resistance. Alternately, it is possible to make the barrier height as small as possible. This approach has as its limitation the fact that both polarities are contacted in many technologies. One or the other will have a large barrier height. To form ohmic contacts then, very heavily doped junctions must be made in one or both types of semiconductor. Generally, the higher the concentration, the lower the specific contact resistance. Values as low as $10^{-7}$ Ω-cm² are often obtained in semiconductor fabrication.

Most of the work in silicon ohmic contacts has been done with aluminum. Aluminum has the property that it will react readily with SiO₂ to form a thin layer of Al₂O₃, which promotes adhesion between the silicon dioxide and the aluminum. This reaction also can assist in the formation of an ohmic contact.

To form a low-resistance contact with silicon, most aluminum metallization processes include as the last step a low-temperature anneal or sinter. Typically done at 450°C, this sinter has several effects. As the aluminum reduces the native oxide, it breaks it up, diffusing the oxygen atoms back into the bulk of the aluminum and allowing fresh Al to diffuse to the metal semiconductor interface. Figure 15.24 shows the diffusion rate of aluminum through aluminum oxide. For a 30-min sinter at 450°C, the aluminum will penetrate about 10 Å into the aluminum oxide. Just as with Pt deposition,
it is important to ensure that the oxide at the contact interface is as thin as possible. Contacts are often dipped in a very dilute HF solution immediately before loading the wafers into the deposition chamber. Once in the chamber, a sputter clean is often carried out immediately before deposition.

While the sinter step is essential to obtaining low resistance contacts, it has undesirable side effects. Figure 15.25 shows a phase diagram for the Al/Si system. As the inset shows, if pure aluminum is heated to 450°C and a source of silicon is present, silicon will begin to dissolve in the aluminum until it reaches a concentration of about 0.5%. If the sample is heated to 525°C, silicon will dissolve to about 1%. The source for this silicon of course, is the wafer. Although this does not seem to be a large amount of silicon, the metal line is an enormous sink. Once dissolved in the aluminum, silicon will diffuse rapidly along the grain boundaries, moving silicon away from the contact. The aluminum in turn moves into the holes to fill the voids left by the silicon. Spikes of aluminum can penetrate into the wafer as deep as 1 μm. If the spike penetrates an electrical junction, the result is a short circuit. Figure 15.26 shows a sequence of diagrams that depicts junction spiking.

There are several methods that have been used to reducing spiking liability in semiconductors. The simplest is to ensure that deep junctions are employed. This is not always desirable from a device standpoint. The next solution is the use of dilute alloys of aluminum/silicon instead of pure aluminum. If the concentration of silicon in the aluminum exceeds the solid solubility at the anneal temperature, little spiking will occur. For 500°C anneals typical Si alloy concentrations range from 1 to 2%. The use of Al/Si alloys also presents problems. The first is silicon condensation. The fact that the dissolved silicon concentration exceeds the solid solubility at low temperature means that there will be a driving force to condense silicon nodules, which are typically 0.5–1.5 μm in diameter. These nodules form between the aluminum grain boundaries and at the metal/semiconductor interface. Since aluminum is an acceptor in silicon, these nodules are heavily doped p-type. If the contact is between n-type silicon and metal, this p layer can significantly increase the contact resistance, particularly if the contact is small. Furthermore, the existence of silicon nodules between the grain boundaries of the metal line represents a serious reliability concern. For narrow lines, the nodules approach the full cross-sectional area of the wire. When large currents are passed through the line, significant local heating can occur, eventually leading to a failure of the line.

For very shallow junctions (less than about 0.2 μm), silicon-doped aluminum is no longer effective. This can occur due to silicon condensation that lowers the silicon concentration in the metal. It can also be related to electromigration induced contact
failures. In this phenomenon, electrons moving under the influence of an imposed electric field accelerate to high energies and then collide with an aluminum atom. In some cases, the collision transfers enough momentum to the aluminum atom to drive it into the substrate. The effect is known to increase as the junction area decreases. To make reliable ohmic contacts to very shallow junctions, a barrier metallization must be employed. A thin layer of the barrier metal is deposited beneath the aluminum. This metal must have the property that the diffusivity of silicon and the interconnect metal are both low at typical sintering temperatures. It must also be highly conductive and have good adhesion between the semiconductor and the metal. One of the first widely used barrier metal is an alloy of titanium and tungsten. Typically 1000 to 2000 Å of 10 wt% Ti/90 wt% W is first sputter deposited. The aluminum is then deposited without breaking vacuum. Several common improvements to this basic process include the use of PtSi or TiSi2 under the TiW to improve the contact resistance. PtSi is formed as discussed previously. TiSi2 can be formed in a furnace [49] or in a multi-step process in a rapid thermal processor [50]. One disadvantage of TiW processes is that the material flakes off the walls of the sputtering system [51]. These flakes can reduce the IC yield. Refractory metal nitrides are now more widely used as barrier metals. TiN is the most widely used, but TaN and WN are also used as barriers. These films can be deposited through reactive sputtering or through rapid thermal annealing of a deposited TiN layer in a nitriding ambient. In this process, the Ti or TiW is sputtered in the presence of nitrogen. The nitrogen is incorporated in the grain boundaries and further reduces the diffusion rate. It is believed that this reduction in diffusivity is due to a passivation effect that dramatically reduces the grain boundary diffusion coefficient. Contacts made this way are stable up to temperatures of 500°C. Alternatively one can use amorphous metal nitrides to avoid grain boundary effects.

All sputtered barrier metals suffer from a common shortcoming for deep submicron devices. In these devices, the contact holes have large aspect ratios. In such a topology, sputtering will not produce a uniform deposit at the bottom of the contact. The barrier metal will be very thin near the edges. Furthermore, the barrier metal does little to improve the planarity of the contact. As a result, the deposited metal will also have poor step coverage going into and out of the contact. One contact method that has received considerable use is chemical vapor deposited (CVD) tungsten (Figure 15.27), as discussed in Section 13.8.

### 15.8 Alloyed Contacts

Unlike silicon, which uses implanted contacts almost exclusively, most ohmic contacts in GaAs and other compound semiconductors are alloyed [52]. To form an ohmic contact to n-type GaAs, the wafers must first be cleaned. A typical process is to rinse the wafer in organic solvents, followed by a DI rinse. Various metallizations have been used for ohmic contacts to n-type GaAs including GeMoW [53] and GeWSi2Au [54]. The most common contact, however, is NiAuGe. In this process, the wafer is placed in an evaporator where 1000 to 1500 Å of a eutectic of 88 wt% Au/12 wt% Ge is deposited, followed by 100 to 500 Å of Ni. Layers of a refractory barrier metal and gold may be added to the top of the Ni to reduce the resistance of the metal lines. The ohmic contact is formed by annealing the wafer for 30 min at 450°C in an $\text{H}_2/\text{N}_2$ mixture. Contact resistivities of $10^{-6} \ \Omega\cdot\text{cm}^2$ have been reported using this technique.
During alloyed contact formation, Au reacts with substrate Ga to form various alloys, leaving behind a large concentration of Ga vacancies. Ge diffuses into the GaAs, occupying the Ga sites and doping the GaAs heavily n-type [55]. The specific contact resistance of ohmic contacts formed in this way depends inversely on the doping concentration in the lightly doped substrate [56] instead of depending on the doping concentration at the contact, which can be as large as $5 \times 10^{19}$ cm$^{-3}$. It has been found that the germanium does not penetrate the surface of the contact uniformly. Instead, contact is made in small pockets of hemispherical radius $r$ (Figure 15.28) that have been observed electrically [57] and by using TEM [58]. Then, the specific contact resistance will have two components, one due to the tunneling contact resistance that depends on the Ge doping level in the GaAs ($N_D$) and one due to the spreading resistance of the Ge inclusions. The total specific contact resistance then is

$$R_c = A_s \exp \left[ \frac{C_s \Phi_b}{\sqrt{N_D}} \right] + D^2 \frac{\rho}{\pi r}$$  \hspace{1cm} (15.16)$$

where $D$ is the mean distance between the pockets and $\rho$ is the resistivity of the substrate. This suggests that it is desirable to have a small separation between these inclusions. In practice, however, that is difficult to achieve.

Alternatively, the specific contact resistance can be lowered by increasing the dopant concentration near the contact (as opposed to at the contact). The most common method is ion implantation. Implanted alloyed contacts are now commonly used in GaAs technologies. One problem that arises in such technologies is simultaneously maintaining both Schottky and ohmic contacts. Figure 15.29 shows the cross section of a MESFET with ion-implanted contacts. It is important that the implantation is offset from the gate. If it extends under the gate, gate leakage will increase substantially. Although a deposited dielectric can be used, followed by an implantation directly into the contact, this approach leaves a high series resistance between the FET and the contact. As will be described in a later section, this seriously reduces the device performance. To achieve this self-aligned implantation for the contact formation, several schemes have been developed. These will be covered in Chapter 17.

Alloyed ohmic contacts to p-type GaAs has been studied in less detail, in part due to the low hole mobility of GaAs. The same general procedure is used, but the usual metallization employed is Au/Zn [59] with 5 to 15 wt% zinc. The metal structure may be evaporated or sputtered [60]. Since Zn is a Group II material, it will dope the GaAs p-type when it resides on a Ga site. An ohmic contact may be formed in a furnace or in a rapid thermal processor [61]. As with n-type GaAs, specific resistances less than $10^{-6}$ $\Omega$-cm$^2$ have been achieved. The thermal stability and the reliability of these contacts are somewhat suspect, however, due to the large vapor pressure of Zn.

An alternate approach to forming low-resistance ohmic contacts to GaAs is to use indium in the
deposited layers, usually close to the GaAs surface. The indium does not dope the GaAs. Rather, it forms Ga$_{1-x}$In$_x$As compounds where $x$ varies smoothly from 1 to 0 through the interface. These compounds have lower barriers and so improved ohmic behavior [62]. This type of contact is also more thermally stable than traditionally formed contacts [63].

### 15.9 Multilevel Metallization

This section will deal with the problem of making the connection between the contact and the bonding pad. Silicon and GaAs technologies have taken very different approaches in this area. In large part, this is because the tasks for which these technologies have been designed are so different. Most silicon technologies have been designed to achieve high levels of integration. Many GaAs technologies have been optimized for high-speed analog operation, with only a secondary emphasis on density. These technologies often simply place a single layer of gold on top of the Ni to reduce the interconnect resistance. This layer is deposited directly on top of the wafer after the transistors have been fabricated. The most important criterion is that the interconnect have a controlled characteristic impedance that is matched to the device input and output impedance. As the number of transistors on digital GaAs circuits has increased over the last few years, however, digital GaAs-based technologies have had to use multiple layers of interconnect. In doing so, they have sometimes grafted the same metallization approaches that have been used in silicon technologies for many years onto the basic MESFET technology. For that reason, we will approach the metallization process primarily as technology independent.

A few comments about the metrics for the interconnect process module can be made. The most critical for digital circuits is capacitance. Digital switching speed is proportional to the capacitance on each node as

$$\tau \propto \frac{V_{\text{swing}} C_{\text{node}}}{I_{\text{drive}}}$$  \hspace{1cm} (15.17)

Alternatively, for very long runs, the speed may be controlled by the $RC$ time constant of the wire. In either case, controlling the node capacitance is essential.

As technologies have improved an increasingly large part of this node capacitance is due to the wire. This can be capacitance between the wire and the substrate or the capacitance between the wires. As the wiring density increases, the wires get narrower. In principle, the wire-to-substrate capacitance should be proportional to the wire width. In practice, edge effects make the capacitance larger than what is predicted from a simple parallel-plate model. This occurs when the linewidth approaches the oxide thickness. Due to fringing capacitance, the reduction of metal linewidth does not reduce the line to substrate capacitance as much as one would expect. Furthermore, the wire-to-wire capacitance increases as the metal spacing decreases. This line to line capacitance is inversely proportional to the line to line spacing. The result of these considerations is that there is a minimum in capacitance for a particular density. There are two ways to accommodate the amount of interconnect required: decreasing the pitch or increasing the number of interconnect layers. Due to crosstalk noise, metal pitch cannot be decreased too severely [64]. As a result, the number of levels of interconnect on modern IC processing is constantly increasing. The current state of the art for silicon is six to eight layers of metal [65]. There is every indication that this will continue to increase to as many as 10 levels.

As lithography and etching have continued to improve, it is now possible to form lines whose spacings are much less than their thickness. In this situation, one could substantially lower the
wire-to-wire capacitance by using thinner metal. This has driven leading edge technologies to abandon Al in favor of Cu. A 0.40-μm-thick layer of pure Cu has the same sheet resistance as 0.65 μm of Al/Cu. Naturally, one must take severe precautions to ensure that Cu does not get into the silicon. Fabs will dedicate not only process equipment and wafer-handling supplies, but often entire bays to copper work to minimize any chance of cross-contamination.

Example 15.1

A 0.25-μm metal line is 500 μm long. It is on top of 0.5 μm of SiO₂, and there are two more identical lines, one on each side. The line-to-line spacing is 0.25 μm. This space is also filled with SiO₂. Neglecting fringing effects, calculate wire-to-wire and wire-to-substrate capacitances for 0.40-μm-thick Cu and 0.65-μm-thick Al/Cu.

In either case the wire-to-substrate capacitance is

\[ C_{w-s} = \frac{(5.0 \times 10^{-2} \text{ cm}) \times (2.5 \times 10^{-5} \text{ cm}) \times 3.9 \times 8.84 \times 10^{-14} \text{ F/cm}}{5 \times 10^{-5} \text{ cm}} \]

\[ = 8.6 \text{ fF} \]

For the aluminum wire,

\[ C_{w-w} = 2 \times \frac{(5.0 \times 10^{-2} \text{ cm}) \times (6.5 \times 10^{-5} \text{ cm}) \times 3.9 \times 8.84 \times 10^{-14}}{2.5 \times 10^{-15} \text{ cm}} \]

\[ = 90 \text{ fF} \]

For the copper wire,

\[ C_{w-w} = 55 \text{ fF} \]

In microwave circuits, the important parameter is the wire impedance. To achieve efficient power transfer, the characteristic impedance of the wire must match the input impedance of the device. Two primary methods have been employed to achieve a controlled impedance and a stable ground connection. One common approach is called a microstrip line. To do this, the wafer is first thinned to achieve the desired characteristic impedance for the design wire width. Next, a layer of metal is deposited on the back side of the wafer. Holes are etched through the wafer to make contact to this solid ground plane. These through holes are patterned and etched from the back side of the wafer using an infrared aligner. A second method for fabricating controlled impedance lines uses a coplanar waveguide. Here two ground wires are run on either side of the signal wire. This does not require any back side processing, but it decreases the density of the IC and does not provide as stable a ground plane.

Most millimeter microwave ICs (MMICs) often have only a small number of transistors. Many discrete microwave components are also manufactured. In these technologies, wire density is not typically an issue. One layer of metallization is used to interconnect the devices. The Schottky gate
metal is also often used as a first layer for parallel plate capacitors and, where necessary, the Ni/AuGe serves as the second layer of metallization.

In modern silicon technologies a distinction is drawn between global or “true” interconnect and local interconnect. Although the resistivity of aluminum is low enough that long runs do not degrade performance in most circumstances, the same cannot be said of polysilicon. Its resistivity is typically $10^{-4}$ $\Omega$-cm. Silicides, which can be run directly on top of the polysilicon to shunt the poly resistance, still have resistivities that are much larger than aluminum (Table 15.1). A rough estimate of the delay can be made using a lumped capacitance model. In this model, the line is thought of as one side of a parallel plate capacitor. Then

$$C = \frac{LW_{ox} \varepsilon_\alpha}{t_{ox}}$$  \hspace{1cm} (15.18)

and

$$R = \rho_{net} \frac{L}{W_{net}}$$  \hspace{1cm} (15.19)

Therefore,

$$RC = \rho_{net} \frac{\varepsilon_{\alpha} L^2}{t_{net} t_{ox}}$$  \hspace{1cm} (15.20)

Figure 15.30 shows the lumped delay as a function of length for several metals running over 1 $\mu$m of oxide. The important feature to understand here is that the delay depends on the square of the length of the interconnect. Designers can therefore use moderate resistance materials to join adjacent transistors or even adjacent cells if they are close together. These materials are often called local interconnect. These local interconnects have several advantages over aluminum or copper that is used on the upper layers of interconnect. They can tolerate high-temperature processing. Local polysilicon interconnect is often available for no additional processing cost since it may already be used for the gates of MOSFETs or the emitter of a bipolar structure. Similarly modern devices often use silicides, as will be discussed in later chapters. These materials can also be used for local interconnect with little additional process complexity.

Aluminum-based alloys have been the metallization of choice for silicon IC technologies. One of its limitations, junction spiking, has already been described. A second limitation for Al and AlSi alloys is electromigration. Electromigration is the movement of conducting atoms as a result of
momentum transfer from current-carrying electrons (Figure 15.31). This movement gives rise to a net flux of metal atoms. If a nonzero divergence in the atomic flux exists anywhere along the line, metal atoms will be depleted or accumulated. Upon depletion, open circuits form. Hillocks are formed by accumulation. If these hillocks get large enough, adjacent lines or even overlying lines can be shorted together. For aluminum metallization systems in particular, this represents a serious reliability problem. A circuit that works well initially will wear out and eventually fail in the field.

A phenomenological description of the electromigration process is given by Black’s equation [66]:

$$MTF = A J^{-n} \exp \left[ \frac{-E_A}{kT} \right]$$

(15.21)

where $MTF$ is the median time to failure, $J$ is the current density, $n$ is a fitting parameter, typically about 2 [67], $E_A$ is the activation energy, and $A$ is a constant. $E_A$ depends on the diffusivity of the metal atom. At high temperatures (above 350°C), the activation energy for aluminum closely matches the self diffusivity of aluminum. At lower temperatures where ICs operate, however, the activation energy for electromigration is smaller. At these temperatures grain boundary diffusion along the facets of the aluminum crystals dominates.

A common metal atom flux divergence mechanism is shown in Figure 15.32. At some point in the line, three crystals of aluminum come together. One crystal face leads into this intersection and two lead out. As a result, atoms experiencing grain boundary diffusion will tend to diffuse more easily out of this point in the line than into it. Over time a void will form. A second source of flux divergence is the presence of a temperature gradient. Because of the temperature dependence of the diffusivity, there will be a net movement of material away from the hotter portion of the wire. This will ultimately produce a void. Electromigration failures of this sort are often seen near bonding pads or large contacts where a large thermal sink exists. Electromigration in contacts is further aggravated by metal thinning as it goes over the topology.

Accelerated testing for electromigration can be done by holding the wafer at high temperatures and/or high current densities, measuring the MTF, and extrapolating back to use conditions. The result is a rule for the maximum dc current that can be allowed in a line to ensure a certain lifetime. A rather conservative number often used for Al/Si is $10^5$ A/cm² for a 20-yr life. For pulsed dc stress, Black’s equation can be modified by multiplying the current density by the duty factor [68]. Various models have been used to explain ac electromigration results. In the commonly used average current
model, the same formula that is used for pulsed dc testing [69] is found appropriate for both unidirectional and bidirectional stressing.

An interesting question in electromigration is what happens when the linewidth approaches the grain size. Typical grains in aluminum metallization after the final sinter are of order 1 μm, so many advanced technologies have linewidths less than the grain size. Early reports suggested a sharp rise in lifetime due to this so-called bamboo effect [70]. On closer examination, however, it was found that while the median time to failure increased, there was still a sizable fraction of wires that failed in much shorter times. These failures were attributed to the statistical distribution of grain sizes in the wire and ultimately to the existence of an intersection between three crystals.

To increase its current-carrying capability, a different metallization system must be used. One commonly employed method is the addition of 1 to 4% copper to the aluminum. The copper atoms reduce grain boundary diffusion effects in aluminum and so greatly increase the activation energy for electromigration. Current densities as high as 10⁶ A/cm² are considered reliable with many Al/Cu metallurgies. A difficulty with the addition of Cu is plasma etching. Normally, aluminum interconnect is etched in a chlorine plasma. Silicon also etches well in these systems due to the high vapor pressure of the various chlorosilanes. Copper chlorides, on the other hand, do not tend to form gasses readily at temperatures below 175°C. Since most photoresists will not tolerate such high temperatures, the wafers cannot be heated sufficiently during the etch to remove all of the copper. As a result, most of the copper in the metallization is left behind after a plasma etch. Chlorinated plasma etch processes tend to corrode the remaining metal due to the presence of AlCl₃ etch residue and water vapor in the air. When an Al/Cu alloy is used, AlCu₂ is formed in the metal. This compound has a galvanic response with Al that drives the corrosion process. As a result, care must be taken to ensure that all of the chlorine has been removed from AlCu wafers that have been plasma etched. This is sometimes done with an inert plasma after the chlorine plasma is completed.

Another reliability concern for fine pitch metallization systems is stress-induced voiding [71]. This effect can open narrow slits [72] or wider features called wedges [73]. The effect arises due to the difference in the thermal expansion coefficients of the metal and the encapsulating dielectric layers and is aggravated at very small metal linewidths. As with electromigration, stress-induced voiding is significantly reduced by the introduction of small amounts of Cu in Al [74] or by using refractory metals instead of aluminum.

Several other metal systems have been proposed for use in silicon ICs. Layers of Al and a refractory metal such as Ti or TiN [75] have been suggested to reduce electromigration. The processing is more difficult, however. The most promising material at this point is tungsten. As mentioned in a previous chapter, the increasing contact aspect ratio is making the attainment of an adequate sputtering process extremely difficult. One possible solution is the use of contact plugs made from CVD tungsten [76, 77]. As discussed in Section 13.8, CVD tungsten is typically deposited in a vacuum system from WF₆ or WF₆/SiH₄ mixtures. It is possible to deposit the tungsten either selectively [78] or nonselectively and etch back the film as with trench isolation (Figure 15.27). It is found that tungsten
deposition consumes some silicon; however, this can be reduced by the addition of SiH₄ during deposition. The process can be incorporated into a single multichamber system [79]. In this application, an underlayer of TiW or TiN is first sputtered to seal the source/drain surfaces from attack during the tungsten deposition [80] and to serve as an adhesion promoter between the metallization and the oxide [81]. A second possibility is to keep the tungsten layer and to use it as a layer of interconnect. This is very attractive, particularly for local interconnect. Since tungsten has a very low self-diffusivity at operating temperatures, electromigration is not a concern. It appears likely that the use of tungsten as the first layer of metallization will become increasingly more common. Selective nickel silicidation of polysilicon can also be used to form these contact plug structures [82].

15.10 Planarization and Advanced Interconnect

Planarized interconnect process modules have become very popular as digital technologies have driven into the submicron, due to the convergence of several factors. The most obvious one is the photolithography. As the imaging lenses have tended to higher numerical apertures, the depth of focus has decreased. Furthermore, nonplanarized interconnect typically has not only the highest steps in the process, but these steps are often nearly vertical. One way of coping with the problem was to impose large design rules for the minimum pitch (linewidth plus line space) that resulted in increasing the number of required interconnect layers. If the lowest level of interconnect had a pitch of x, the second level might have a pitch of 1.5x, the third level might have a pitch of 3x, and so on. If finer pitch can be achieved with planarized interconnect modules, however, their additional process complexity may be more than offset by their ability to reduce the number of interconnect layers that are required.

At the same time, the step height often cannot be reduced. The result is higher aspect ratio structures that are difficult to cover. Not only must the metal have sufficient step coverage into and out of contacts, it must also have good step coverage as it goes over metal lines on lower levels. There can be particularly severe topology as the second level metal exits a via down to the first layer of metal. In an extreme example called a plug (Figure 15.33), the via may be coincident with a contact from the first layer metal down to the substrate.

An early, popular planarization approach was the sacrificial oxide etchback method [83]. As shown in Figure 15.34, a thick oxide is first deposited over a patterned metal layer. Next, a layer of photoresist is spun onto the wafer. The wafer is then put into a plasma etch system. The etch ambient is a mixture of O₂ and CF₄ or another fluorinated species. The mixture is set to provide nearly identical etch rates of resist and oxide. Since the resist is thinner over the first metal lines, the oxide will be exposed here first. The etch is done until most of the resist has been removed. A second layer of oxide may then be deposited. Another etchback cycle can also be run if required.

Since these processes follow the resist profile, they are only locally planarizing. That is, they smooth out steps but do not produce a truly planar surface. For feature sizes greater than about 50 µm, the planarizing effect is lost [84], although temperature baking of the resist tends to improve this somewhat. The thickness of the oxide will also vary across the device. This means that a subsequent contact etch process must have a high selectivity and must not significantly

![Diagram](image.png)

**Figure 15.33** Nonplanarized plug structure consisting of a coincident contact and via.
etch laterally. This second criterion can be difficult to achieve if the etch process must also slope the sidewalls of the oxide.

More recently these local planarization techniques have been replaced by chemical mechanical polishing, as discussed in Chapter 11.

From the previous discussion it is apparent that for many integrated circuits, the interconnect capacitance is the dominant term in the delay equation. Decreasing the interconnect capacitance decreases the node capacitance and therefore increases the circuit speed. The most straightforward way to do this is to lower the dielectric constant (permitivity) of the insulator used between the metal layers, often called the interlayer dielectric (ILD). When an external electric field is applied to a simple dielectric, the valence electron cloud is displaced from the nuclei and core electrons. This creates a dipole that changes the electric field. The permitivity is a measure of the strength of the field caused by this dipole. The more tightly bound the valence electrons, the lower the permitivity. Of course, the lowest relative permitivity possible, one, is that of empty space. There are two approaches then to decreasing the dielectric constant of a film: (1) use a material with tightly bound electrons, and/or (2) use an open film structure with a large amount of void space. Both have been tried for advanced interconnect insulators.

The simplest approach is to modify CVD SiO₂, lowering the permitivity by increasing the localization of the electrons participating in the silicon bond. Replacing O with F does this. The higher the fluorine concentration, the lower the permitivity. Fluorine can be added readily in a CVD process by using NF₃ or some other fluorine-containing species in the reactor. As shown in Table 15.2 "pure" CVD SiO₂ has a permitivity of 4.1 to 4.2. This can be reduced to almost 3.2 by adding enough fluorine, but the etch rate becomes very large. The addition of F weakens the glass, noticeably reducing the hardness and changing the elastic modulus for F concentrations above 6%. At high fluorine concentrations films also absorb water, which can lead to metal corrosion [85]. This ultimately makes it difficult to stop the metal etch and to CMP the insulator controllably. Very high F-concentration films may be used with upper and/or lower "hard" layers consisting of SiO₂ without any fluorine; otherwise the lowest usable permitivity of fluorinates oxide is about 3.4.

The next step in reducing the permitivity of the ILD is to abandon SiO₂-based materials altogether. The next lowest permitivity material is diamond-like carbon (DLC). DLC can be deposited from CH₄ or other carbon-containing species in a PECVD reactor (Applied Materials, trade name for DLC is Black Diamond™). DLC has a permitivity ranging from 2.7 to 3.3 depending on deposition.
Table 15.2 Effect of adding fluorine using C₂F₆ in a TEOS/O₂ PECVD process

<table>
<thead>
<tr>
<th>Si–F atom%</th>
<th>C₂F₆ Flow Rate (slm)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0.8</td>
</tr>
<tr>
<td>ε</td>
<td>4.2</td>
</tr>
<tr>
<td>Stress (MPa)</td>
<td>-176</td>
</tr>
<tr>
<td>Etch rate*</td>
<td>110</td>
</tr>
</tbody>
</table>

The permittivity decreases linearly with F concentration.

*nm/min in 10:1 BHF.

conditions. Diamond-like carbon does not adhere well to most layers, however, so a stack of adhesion layers must be used [86]. Typically this includes thin (~100-Å) layers of SiOₓ and SiOₓ where x < 2. To further reduce the permittivity of DLC, one can again add fluorine to the film. Fluorinated carbon films are amorphous and can reduce the dielectric constant to a value as low as 2.0. Insulators with permittivities in this range can also be formed using spin-on organic insulators such as some types of polyimide [87]. Fluorinated polyimide can be used for extremely low permittivity films [88]. Generically these very low permittivity ILDs have problems with low mechanical strength, moisture absorption, poor dimensional stability, low breakdown strength, increased leakage current, poor thermal stability, increased thermal expansion coefficients, low thermal conductivity, and outgassing leading to an effect known as via poisoning [89]. Polyimides also have problems with CVD tungsten since the fluorine that is formed as a decomposition by-product can easily attack the insulator leading to void formation and/or low breakdown strength.

The alternate approach to low permittivity insulators is the use of low-density materials. In that case the dielectric constant is well described by a simple effective media approximation:

\[
\varepsilon = \varepsilon_x \times \varepsilon_{oxide} + (1 - \varepsilon_x) \times 1
\]

where \( \varepsilon_x \) is the filling factor. Thus an SiO₂ film with a filling factor of 0.3 has an effective permittivity of about 1.9. A simple way to form this type of film is to apply a slurry of SiO₂ particles in a liquid carrier. After spin-on application the film is heated to form a gelatinous material (Figure 15.35) [90]. Next higher temperature steps are used to drive off most of the solvent and the film is annealed to set up the desired electrical properties. The result is sometimes called a silica aerogel. (More properly that is a xenogel. Aerogels are dried supercritically at high pressure and show less shrinkage.) This type of film can be produced with up to 95% porosity (\( \varepsilon_x = 0.05 \)) resulting in a permittivity of 1.1 [91]. The structural stability of these films, however, is extremely questionable.

Figure 15.35 Process sequence for a typical aerogel film (after Ramos et al.).
Figure 15.36  Two options for performing a dual damascene process: (A) single damascene; (B) dual damascene (after Price et al., Used with permission. Thin Solid Films, 1997).
As suggested earlier, the interconnect capacitance may be lowered by using thinner layers of low-resistivity interconnect. Silver has the lowest room temperature resistance, but it does not adhere well to SiO₂, but diffuses rapidly through it. Furthermore it has poor electromigration resistance. The low resistivity and improved electromigration resistance of copper [92] (about 10X the maximum current density of AlCu) have driven its use as an interconnect material. One of the major problems with copper is the lack of an adequate copper etch process. This has been solved through the use of the damascene process. The interlayer dielectric is first etched anisotropically. Next a thin barrier metal such as Ti, Ta, TaN, or TiN is deposited [93]. Copper is then deposited to a thickness greater than the oxide recess, filling the trenches. Finally the excess copper is removed by CMP [94].

The simple damascene process has been extended to a dual damascene process by using the inlaid copper to form both interconnect lines and via fills. There are two ways to run a dual damascene process. As shown in Figure 15.36, the ILD is first etched down to the previous metal layer, the via is filled, and the excess removed by CMP. Next the interconnect recess is patterned and etched, and a second copper fill and CMP are done. Alternatively, both the via and line recess can be done before copper deposition. Often a thin Si₃N₄ etch stop is inserted in the ILD to mark the top of the via. This process has the disadvantage that the smallest feature, the via, must be patterned in the bottom of the interconnect recess. For small vias this can be extremely challenging.

Copper can be deposited in a variety of ways, but plating and ionized metal plasma PVD are preferred for the damascene process. For plating processes a thin copper seed layer may have to be deposited on top of the barrier metal to initiate the plating. Electroplating copper can have problems with deposition uniformity, however, due to the dependence of the deposition rate on the local current density.

The technique that is most widely used for planarization is chemical mechanical polishing (Chapter 11). In this technique a thick layer of SiO₂ is first deposited on top of the metal interconnect. The oxide is then polished to a flat finish before the contact mask. This technique is now applied on virtually all metal layers, allowing the use of a fine pitch on every layer.

15.11 Summary

In this chapter, the process modules of device isolation, contact formation, and interconnection were presented. The simplest isolation techniques involve junction isolation. Various LOCOS-based methods have been widely used, but suffer from lateral encroachment and incomplete isolation at small junction separations. Trench-based methods have become popular for submicron technologies. For GaAs technologies, nearly all device isolation is accomplished with semi-insulating substrates. Conducting islands can be created via proton implantation or mesa etching.

Contacts are divided into rectifying and ohmic. For rectifying contacts, the barrier height is a sensitive function of the nature of the metal/semiconductor interface. For ohmic contacts, achieving a low contact resistivity requires a large doping concentration at the metal/semiconductor interface. Most silicon technologies achieve heavy doping by implantation. Self-aligned silicides (salicides) have been developed to reduce the series resistance of shallow junctions in silicon. Many GaAs technologies use alloyed contacts, sometimes with an implantation, to achieve acceptably low specific contact resistivities.

High-performance interconnect requires the use of low-resistivity metal on top of low-capacitance dielectrics. Aluminum alloys are the most widely used for silicon-based technologies, although copper is now beginning to replace it. Gold is generally used for GaAs technologies. CVD SiO₂ is the most commonly used dielectric, although lower permittivity films such as polyimide are promising for future applications.
Problems

1. In some microwave applications, the collector of a bipolar transistor is heavily doped to allow the device to be run at large dc-bias currents. Assume that for the simple IC shown in Figure 15.1, the n-type collector concentration is $2 \times 10^{17}$ cm$^{-3}$. If the maximum reverse bias ever applied to the base collector junction is 5 V, what is the minimum base to base spacing required to isolate the transistors?

2. If the field oxide is 5000 Å thick, what will the threshold voltage of the parasitic device be? Assume no interface states or metal semiconductor work function and a substrate concentration of $2 \times 10^{19}$ cm$^{-3}$.

3. A LOCOS structure is used to isolate two N$^+$ regions. A metal line runs over the 0.5-μm-thick LOCOS oxide forming a parasitic MOSFET. Assume that both N$^+$ junctions and the line are all at the same voltage. The substrate concentration ($N_A$) is $2 \times 10^{16}$ cm$^{-3}$ (p-type), the oxide is ideal, and the junctions are 3.0 μm apart. Calculate and show the parasitic threshold voltage and the voltage necessary to punch through between the two N$^+$ regions. Which one will occur at the lower voltage? What substrate concentration would you need to ensure that neither occurred below 10 V.

4. Consider the deep trenches shown in Figure 15.13. If the substrate is uniformly doped at $1 \times 10^{19}$ cm$^{-3}$ and the trenches are 5 μm deep, how large a reverse bias is required to deplete the trench all the way to the bottom?

5. Calculate the specific contact resistance at room temperature for a contact to $10^{17}$ cm$^{-3}$ n-type GaAs if the metallization has a barrier height of 0.8 eV. (Hint: Differentiate Equation 15.9 and evaluate at $V = 0$.)

6. An implanted ohmic contact is formed n-type silicon. When the doping concentration at the metal/semiconductor interface is $1 \times 10^{20}$ cm$^{-3}$, the specific contact resistance $5 \times 10^{-6}$ Ω-cm$^2$. A new source/drain implant is being implemented to increase the doping concentration at the metal/semiconductor interface to $2 \times 10^{20}$ cm$^{-3}$. Assume that $A_0$ is fixed, the barrier height is 0.6 eV, and $m^* = 1.18 \times m_p$. Calculate the specific contact resistance that you would expect for the new technology. Calculate the resistance of a vertical-flow $0.5 \times 0.5$-μm contact for each of the two technologies.

7. A triple-diffused bipolar technology is fabricated using $10^{14}$ p-type substrates. The maximum bias that either collector would ever see is 10 V. Assume simple-step junctions and calculate the required separation between devices. (Calculate the distance between the collectors after fabrication. The design of the collectors must be further apart than this to allow lateral diffusion during the collector drive.) To improve matters, a p$^+$ guard ring will be added to the process. The guard ring must be at least 3 μm deep to be effective. Assume that the guard ring pattern is 2 μm wide, that the final guard ring diffusion must never be allowed to contact the collector, and the alignment tolerance between the collector and guard ring is 1 μm. What is the new minimum separation between the collector diffusions?

8. Memory makers found that replacing long word lines made from polysilicon with a salicide could decrease the access time of the device. Assume that the line is 1 cm long, the oxide is 1 μm thick, the poly and silicide are each 0.5 μm thick, and the resistivities of the two films are $10^{-3}$ and $10^{-4}$ Ω-cm, respectively. Use the simple lumped RC model to determine the delay associated with each of the two lines. (The dielectric constant of SiO$_2$ is $3.9 \times 8.84E-14$ F/cm.)

9. An ohmic contact is to be made to n-type GaAs using an alloyed Ni/Au/Ge contact on an ion-implanted channel layer. The resistivity of the conducting GaAs layer is 0.01 Ω-cm. In the region of the alloyed contact, the surface concentration is $1 \times 10^{19}$ cm$^{-3}$. The mass of
an electron in GaAs is \(0.067 \times m_e\). The barrier height of the metal is 0.75 V, and the constant \(A\) for GaAs in the exponential expression for the ohmic contact resistance is \(10^{-8} \text{ \Omega-cm}^2\). The contact is found to form pits at the metal/semiconductor interface. These pits have radii of 300 Å with a pit density of \(10^9 \text{ cm}^{-2}\).

(a) What is the specific contact resistance in ohm-cm²?
(b) What would the specific contact resistance be if the diffused impurity concentration under the contact is raised to \(3 \times 10^{19} \text{ cm}^{-3}\)?

10. An advanced metallization process is proposed for high-density silicon-based ICs. This process will use several new materials. Identify one advantage and one disadvantage for each new material: (a) CVD tungsten, (b) evaporated copper, and (c) spin on polyimide.

References


Week 10

Layout Design Rules
Chapter 3

Layout of CMOS Integrated Circuits

To design a digital integrated circuit, one usually starts with a circuit schematic. This provides the topology of the network that implements the logic. Layout design is the next step in the process. It is directed toward the problem of translating the schematic into a set of patterned layers that form the integrated structure in a silicon substrate. All of the electrical properties of the circuit are established in this phase of the design sequence.

This chapter examines the concepts and details used to implement CMOS circuits in silicon.

3.1 Introduction to Physical Design

Several equivalent viewpoints may be used to describe an integrated circuit. To a circuit designer, a chip is the physical realization of an electronic network. A logic designer, on the other hand, may choose to view the chip as a device that performs functions specified by logic diagrams, function tables, or an HDL\(^1\) file. Figure 3.1 illustrates how different people might view the same thing. Regardless of the abstraction used, in the final analysis, an integrated circuit is really an intricate physical object that has been carefully designed and fabricated.

Physical design in VLSI deals with the procedure needed to realize a circuit on the surface of a semiconductor wafer. Starting with the electrical network schematic, computer tools are used to create the necessary patterns on each layer in the 3-dimensional structure. Once the drawings are completed, the information can be used to fabricate the masks needed in the processing line.

---

\(^1\) HDL is an acronym that stands for **hardware description language**.
effect is important when determining the minimum spacing $S$ between adjacent
doped lines.

Depletion effects also influence the value of $S$. As shown in Figure 3.4, a depletion region exists at every pn junction. Let us assume for simplicity that the junction has a step-doping profile where the impurity concentration changes abruptly from $N_d$ on the n-side to $N_a$ on the p-side. With a reverse-bias voltage of $V_R$, the total depletion width $x_d$ can be computed from

$$
x_d = x_0 \sqrt{1 + \frac{V_R}{V_{bi}}}, \quad (3.1)
$$

where

$$
x_0 = \frac{2e_s}{q} \frac{V_{bi}}{N_a \ln\left(\frac{N_a}{N_d}\right)} \quad (3.2)
$$

is the zero-bias value of the total depletion width, and

$$
V_{bi} = \left(\frac{kT}{q}\right) \ln\left(\frac{N_a N_d}{n_i^2}\right) \quad (3.3)
$$

is the built-in voltage. Table 3.1 provides a list of useful numerical values for basic calculations. Note that the intrinsic concentration $n_i$ applies only to silicon at room temperature ($T = 300^\circ$ K). To calculate the p-side depletion width $x_p$ shown in the drawing, we use

$$
x_p = \left(\frac{N_d}{N_a + N_d}\right) x_0 \sqrt{1 + \frac{V_R}{V_{bi}}} \quad (3.4)
$$

Since this increases with the reverse bias voltage, the minimum spacing distance $S$ often accounts for the worst-case situation, i.e., when $V_R = V_{DD}$. From this discussion, it is not surprising that the minimum width and spacing for $n^+$ and $p^+$ regions are usually larger than those for a polysilicon line.

\footnote{"Lateral" means in a direction parallel to the surface.}
3.3.2 Contacts and Vias

Contacts and vias are used to provide electrical connections between different material layers. In general, contacts are necessary connections to access the various regions of silicon, while vias are used between two interconnect layers to simplify the layout. When formulating design rules for these types of objects, two important considerations arise: the physical size of the oxide cuts, and the spacing needed around the connection on the layers.

### TABLE 3.1 Useful constants

<table>
<thead>
<tr>
<th>Symbol/Parameter name</th>
<th>Value</th>
<th>Units</th>
</tr>
</thead>
<tbody>
<tr>
<td>q</td>
<td>1.6 × 10^{-19}</td>
<td>Coulombs</td>
</tr>
<tr>
<td>ε_s=ε_0</td>
<td>(11.8)(8.854 × 10^{-14})</td>
<td>Farads/cm</td>
</tr>
<tr>
<td>ε_ox=ε_r=ε_0</td>
<td>(3.9)(8.854 × 10^{-14})</td>
<td>Farads/cm</td>
</tr>
<tr>
<td>n_i</td>
<td>1.45 × 10^{10}</td>
<td>cm^{-3}</td>
</tr>
<tr>
<td>(kT/q)</td>
<td>0.0258</td>
<td>Volts</td>
</tr>
</tbody>
</table>

Let us first examine the dimensions of a contact region. The geometry is shown in Figure 3.5. It is apparent that the minimum size is limited by the lithographic process. However, this does not imply that one uses the largest contacts possible, as other considerations come into play. If the contact cut is too large, then it may be difficult to attain complete coverage when depositing the upper layer. Large contact cuts may result in cracks or voids, that may in turn lead to a circuit failure. To overcome this problem, it is common to restrict the dimensions of contact cuts to pre-specified values that can be reliably made in the fabrication process.

Now, consider the problem of spacing x around an oxide cut as shown in Figure 3.6. We must specify the minimum distance between the edge of an oxide cut and the edge of a patterned region to allow for misalignment tolerances in the masking steps. These are generically classified as **registration errors**, and arise because it is not possible to align the mask with arbitrary precision.
small misalignment where the polysilicon does not traverse the entire active area. Since the ion implant will dope all of the exposed substrate, the resulting structure shown in (c) has the drain and source n+ regions merged into one. Electrically, the drain and source are shorted, so the device cannot control the current flow, i.e., the switching action has been lost.

The same consideration applies to a MOSFET where the n+ region changes shape as shown in Figure 3.10. The channel width W is a critical design parameter, so that the spacing s between the poly and n+-edges must be large enough to ensure that the MOSFET still has the proper value if small registration errors occur.

### 3.3.4 Bloats and Shrinks

The drawings produced by a layout editor provide the basic view of an integrated circuit that are used to extract the equivalent device parameters. It is therefore important to understand the correlation between what is shown on the computer monitor when compared with the actual die after fabrication.

In general, the final size of a physical layer will be different from the dimensions specified by the mask that created the pattern. Two obvious examples are

- **ACTIVE**: Encroachment reduces the size of the usable active area.
- **Doped n+ or p+**: Lateral doping effects increase the size of these regions.

In addition, the physical process of etching a material layer is anisotropic, with both vertical and lateral etching present. Although the lateral etch rate can be reduced.
using various techniques, it cannot in general be reduced to zero. The effect of anisotropic etching on a polysilicon layer is illustrated in Figure 3.11.

The question that naturally arises at this point is "What does the layout drawing represent relative to the finished chip?" In other words, will the chip patterns be identical to those shown by the layout editor, or are size adjustments necessary? In the early days of chip design, one had to increase or decrease the size of the layout drawing to view the actual chip dimensions. However, it is now common for the chip fabricator to subject the masks to bloats (increases in object sizes) and shrinks (decreases in the object size) as needed to compensate for the difference between the mask dimensions and the resulting size on the chip. When this is done, then the layout editor displays a reasonably accurate view of the finished circuit.

These considerations are particularly important to designing a MOSFET. Although the two critical dimensions $L$ (the channel length) and $W$ (the channel width) are related to the drawn mask values, the values are different as shown in Figure 3.12. The channel length $L$ that is required in the current flow equations is reduced from the drawn value $L'$ by

$$L = L' - 2L_o$$ \hspace{1cm} (3.5)

where $L_o$ is the overlap distance from lateral diffusion effects. In a similar manner, the channel width $W$ is smaller than the drawn ACTIVE width $W'$ because of active area encroachment. This is where the usable size of the active area is reduced because of oxide growth underneath the edges of the nitride pattern. If the encroachment per side is $(\Delta W)$, then

$$W = W' - 2(\Delta W)$$ \hspace{1cm} (3.6)

gives the proper electrical value. This can become confusing when entering the device dimension into a circuit simulation program. For example, SPICE parameters can be entered using either the drawn or physical values so long as the remaining data values are consistent. This will be discussed in more detail in the next chapter.

---

1 The gate overlap $L_o$ is also known as the lateral diffusion length $L_D$. 

---

Design Rule Basics

3-9
3.4 Types of Design Rules

Geometrical design rules are a set of minimum widths, spacings, and layout guidelines needed to create the masks. There are two ways to specify these dimensions:

- **Specific Values**: All dimensions are stated in standard unit of length, such as the micron;
- **Scalable**: Distances are specified as multiples of a metric $\lambda$ that has dimensions of length. The actual value of $\lambda$ is adjusted to correspond to the limitations of the process line.

Both approaches are common in CMOS VLSI. Scalable rules have the advantage that they can be adjusted to several different processing lines by changing the value of $\lambda$. However, this does not come without cost. Since every distance is specified as a multiple of $\lambda$, the numerical value is dictated by the worst-case situation. This generally decreases the compaction density of the circuit compared to what is attainable if the parameters are specified in an absolute metric such as microns.

In general, there are three main classes of design rule specifications. These are

- Minimum Width,
- Minimum Spacing, and,
- Surround.

Surround rules apply to objects placed within larger objects (such as contacts). Every layer has a minimum width and minimum spacing value, while surrounds are specified as required.

3.5 CMOS Design Rules

In this section, we will examine a basic set of CMOS design rules to understand the
presentation and meaning of each type of rule. This set has been provided in the setup technology file as \texttt{ledit.tdb}, and is also available with the name \texttt{morn20.tdb}. The DR set describes the MOSIS Orbit 2-micron double-poly, double-metal, n-well CMOS process; the technology name for this process is SCNA (for Scalable CMOS N-Well Analog). For the purposes of this discussion, we will not list all layers. In particular, the POLY2 layer is not shown explicitly here to simplify the discussion.

A complete design rule set contains all of the geometric limits for mask layout. This includes the minimum feature sizes and minimum spacings on each mask, and also provides layer-to-layer spacings when necessary. In order to list the rules in an easy-to-find format, they are presented according to the order of the masks used in the processing. The primary design rule layers for the morn20 technology are listed in Table 3.2.

<table>
<thead>
<tr>
<th>Mask Number</th>
<th>Mask Layer</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.</td>
<td>NWELL</td>
</tr>
<tr>
<td>2.</td>
<td>ACTIVE</td>
</tr>
<tr>
<td>3.</td>
<td>POLY</td>
</tr>
<tr>
<td>4.</td>
<td>SELECT</td>
</tr>
<tr>
<td>5.</td>
<td>POLY CONTACT</td>
</tr>
<tr>
<td>6.</td>
<td>ACTIVE CONTACT</td>
</tr>
<tr>
<td>7.</td>
<td>METAL1</td>
</tr>
<tr>
<td>8.</td>
<td>VIA</td>
</tr>
<tr>
<td>9.</td>
<td>METAL2</td>
</tr>
<tr>
<td>10.</td>
<td>PAD</td>
</tr>
<tr>
<td>11.</td>
<td>POLY2</td>
</tr>
</tbody>
</table>

When you are using L-Edit, the design rules corresponding to the technology are always loaded into your file, and are Saved when you save your work. A text listing of the design rules can be obtained using the keyboard command \texttt{Alt-W}; there is no Menu Bar equivalent. This action writes a text file named \texttt{filename.rul} to the working disk drive that provides a listing of all layers and rules in ASCII format. The list also provides information on derived layers.

Design rule sets are most easily understood by providing a text list in conjunction with simple drawings to illustrate each value. These are broken into groups corresponding to each basic layer. The layer number N. is used to identify the group, and each dimensional specification is labelled accordingly, e.g., N.1, N.2, and so on. In order to clarify some of the fine points involved with design rules, they will be presented in two different forms. First, we will examine a simplified set of rules that provide basic information on minimum widths and spacings, and then look at MOSFET layout rules. This gives a general idea of what the rules mean.
This is followed by a more complete set of layout statements that correspond to those used by L-Edit in performing the DRC algorithm.

### 3.5.1 Basic Rules

The most fundamental layout guidelines limit the smallness of each material layer, and provide the basis for device design. The values in this design rule set are scalable according to the metric \( \lambda \). Numerically, \( \lambda = 1 \ \mu \text{m} \) for the 2 \( \mu \text{m} \) technology. However, these rules also apply to the mhp_n12.tdb (named SCN) 1.2-micron, single-poly, double-metal process with a metric of value \( \lambda = 0.6 \ \mu \text{m} \).

#### Minimum Widths and Spacings

This group of rules are those that specify the minimum linewdths and minimum spacings permitted on the primary layers summarized in Table 3.3 and illustrated in Figure 3.13.

**TABLE 3.3 Minimum Width and Spacing Rules**

<table>
<thead>
<tr>
<th>Layer</th>
<th>Type of Rule</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>POLY</td>
<td>Minimum width</td>
<td>( 2\lambda )</td>
</tr>
<tr>
<td></td>
<td>Minimum spacing</td>
<td>( 2\lambda )</td>
</tr>
<tr>
<td>ACTIVE</td>
<td>Minimum width</td>
<td>( 3\lambda )</td>
</tr>
<tr>
<td></td>
<td>Minimum spacing</td>
<td>( 3\lambda )</td>
</tr>
<tr>
<td>NSELECT</td>
<td>Minimum width</td>
<td>( 3\lambda )</td>
</tr>
<tr>
<td></td>
<td>Minimum spacing</td>
<td>( 3\lambda )</td>
</tr>
<tr>
<td>METAL1</td>
<td>Minimum width</td>
<td>( 3\lambda )</td>
</tr>
<tr>
<td></td>
<td>Minimum spacing</td>
<td>( 3\lambda )</td>
</tr>
<tr>
<td>METAL2</td>
<td>Minimum width</td>
<td>( 3\lambda )</td>
</tr>
<tr>
<td></td>
<td>Minimum spacing</td>
<td>( 4\lambda )</td>
</tr>
</tbody>
</table>

![Figure 3.13. Minimum width and minimum spacing.](image)

#### MOSFET Layout

The basic layout rules for MOSFETs are illustrated in Figure 3.14, and the values are summarized in Table 3.4. Dimensions that deal with POLY and N+/P+ apply equally to both nFETs and pFETs. Contacts (cuts in the oxide that allow electrical...
connections between two layers) to N+/P+ are called ACTIVE CONTACTS, while POLY CONTACTS provide access to the POLY layer (which must be in field regions).

In this technology, p-channel MOSFETs must be inside n-well regions, and sufficient spacing must be provided between opposite-polarity regions (i.e., between n+ and p+ sections) as well as between P+ regions and the NWELL edge. Spacing requirements also apply between different n-well regions; in general, a larger space is needed if the n-wells are biased at different voltages, due to the voltage dependence of depletion regions.

**TABLE 3.4 MOSFET Layout Rules**

<table>
<thead>
<tr>
<th>RULE</th>
<th>Meaning</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>POLY Overlap</td>
<td>Minimum extension over ACTIVE</td>
<td>2(\lambda)</td>
</tr>
<tr>
<td>POLY-ACTIVE</td>
<td>Minimum Spacing</td>
<td>1(\lambda)</td>
</tr>
<tr>
<td>MOSFET Width</td>
<td>Minimum N+/P+ MOSFET W</td>
<td>3(\lambda)</td>
</tr>
<tr>
<td>ACTIVE CONTACT</td>
<td>Exact Size</td>
<td>2(\lambda \times 2\lambda)</td>
</tr>
<tr>
<td>ACTIVE CONTACT</td>
<td>Minimum Space to ACTIVE Edge</td>
<td>2(\lambda)</td>
</tr>
<tr>
<td>POLY CONTACT</td>
<td>Exact Size</td>
<td>2(\lambda \times 2\lambda)</td>
</tr>
<tr>
<td>POLY CONTACT</td>
<td>Minimum Space to POLY Edge</td>
<td>2(\lambda)</td>
</tr>
</tbody>
</table>

### 3.5.2 Mask-Based Design Rule Set

At the layout level, design rules apply to the masks involved in the chip patterning process. L-Edit employs a complete set of rules that apply to both masking layers and derived layers. Using derived layers allows us to make the connection between the individual patterns and the material layers that are important to the operation of transistors and other devices.

Figures 3.15, 3.16, and 3.17 provide most of the SCNA design rule set using

![Diagram of MOSFET layout rules](image)

**Figure 3.14.**
MOSFET layout rules.
1.0 NWELL
  1.1 Minimum Width . . . . . . . . 10
  1.3 Minimum Spacing . . . . . . . . 6

2.0 ACTIVE (N+, P+)
  2.1 Minimum Width . . . . . . . . 3
  2.2 Minimum Spacing . . . . . . . . 3
  2.3 Drain/Source ACTIVE to NWELL
  2.3a P+ ACTIVE to NWELL . . . . . 5
  2.3b N+ACTIVE to NWELL . . . . . 5
  2.4 CONTACT to NWELL EDGE
  2.4a P+ in SUB to NWELL . . . . . 3
  2.4b N+ in WELL to NWELL . . . . 3

3.0 POLY
  3.1 Minimum Width . . . . . . . . 2
  3.2 Minimum Spacing . . . . . . . . 2
  3.3 Gate Extension out of ACTIVE . . 2
  3.4 Extension (MOSFET)
    3.4a nMOSFET Drain/Source . . . . 3
    3.4b pMOSFET Drain/Source . . . . 3
  3.5 POLY to ACTIVE Spacing . . . . 1

N+ = (NSELECT) AND (ACTIVE)
P+ = (PSELECT) AND (ACTIVE)

Figure 3.15. SCNA design rules.
4.0 NSELECT and PSELECT

4.2 ACTIVE - SELECT Spacing
  4.2a ACTIVE in SELECT . . . . 2
  4.2b ACTIVE in SELECT to
    ACTIVE in next SELECT . . . 2

4.4 Minimum Dimensions
  4.4a NSELECT Minimum Width . . 2
  4.4b PSELECT Minimum Width . . 2
  4.4c NSELECT Minimum Space . . 2
  4.4d PSELECT Minimum Space . . 2
  4.5 PSELECT overlap of NSELECT . 0

5.0 POLY CONTACT
  5.1 Exact Size . . . . . . . . . 2 x 2
  5.2 Field Poly Overlap of
    POLY CONTACT . . . . . . . . . 2
  5.3 Spacing . . . . . . . . . . . 2

6.0 ACTIVE CONTACT
  6.1 Minimum Width . . . . . . . 2
  6.2 Field Poly Overlap of
    POLY CONTACT . . . . . . . . . 2
  6.3 Spacing . . . . . . . . . . . 2

7.0 METALL
  7.1 Minimum Width . . . . . . . 3
  7.2 Minimum Spacing . . . . . . . 3
  7.3 Overlap of POLY CONTACT . 1
  7.4 Overlap of ACTIVE CONTACT . 1

Figure 3.16. SCNA design rules (continued).
Week 12A

CMOS Process Integration

**EXERCISES**

1. Prove that CMOS inverter delay time is minimal if $L_p = L_n$ and $W_p = \sqrt{\mu_n / \mu_p W_n}$.

2. A CMOS inverter has the following dimensions: $t_{ox} = 200 \text{ Å}; t_{poly} = 3000 \text{ Å}; L_p = L_n = 1 \mu\text{m}; W_p = 2W_n = 10 \mu\text{m}; \mu_n = 2 \mu_p = 600 \text{ cm}^2/\text{V-sce}; I_{ov} = 0.1 \mu\text{A}$, and $V_n = -V_p = 1 \text{V}$. There are four $0.5 \mu\text{m}$-wide metal lines at the output node at a length of $250 \mu\text{m}$ each. They are $1 \mu\text{m}$ apart on a field oxide with $1 \mu\text{m}$ thickness. Determine the Miller capacitance ($C_m$) and stray capacitance ($C_s$), then calculate total delay time for charging and discharging the output node.

3. If the $p$-channel device in the CMOS inverter described in Exercise 2 has a leakage current of $1 \mu\text{A}$ when the gate is off, calculate the output voltage when its input is at $5 \text{V}$.

4. For a CMOS buffer made of an inverter chain with load capacitance of $C_L$ and internal inverter input capacitance of $C_i$, prove that the total delay is minimal if each inverter is larger than the preceding one by a factor of $e$, where $e$ is the base of natural logarithms, and the number of inverters is $\ln(C_L/C_i)$.

5. An NMOS pass gate is connected by a $5 \text{V}$ clock voltage, a $5 \text{V}$ input voltage and the output load is $5 \text{pF}$. The transistor characteristics are: $V_t = 0.5 \text{V}; W/L = 10/1; t_{ox} = 200 \text{ Å}; \mu_n = 500 \text{ cm}^2/\text{V-sce};$ and, the body effect coefficient ($\gamma$) is $0.5 \text{ V}^{1/2}$. What voltage level can the output reach in a reasonable charging time? What is the estimated charging time?

6. For the pass gate in Exercise 4, if the subthreshold swing of this transistor is $100 \text{ mV/dec}$ and threshold voltage is defined at $1 \mu\text{A}$, and $\gamma = 0$, estimate the additional time to charge the output to $5 \text{V}$.

7. What are the advantages and disadvantages of a domino CMOS logic in comparison with a static CMOS logic? Is $n$- or $p$-well a better technology to implement the domino logic?

---

**CMOS PROCESS TECHNOLOGY**

This chapter discusses CMOS process technology, emphasizing process architectures, meaning the masking sequences and overall process integrations, rather than individual process modules. First introduced is the relationship between a circuit layout generated by designers and a device cross-section at various key processing steps. Then various CMOS process architectures are described: $p$-well, $n$-well, twin-tub and retrograde well processes. Comparisons will be made between these different processes and their impacts on circuit performance will be discussed. Reasons for selecting a particular process architecture or process module will be presented in this chapter, and further explained in later chapters in conjunction with device design. CMOS devices built on sapphire and other insulators are also described. Finally, bipolar and CMOS integration, called BiCMOS, is discussed extensively. Bipolar device characteristics, BiCMOS operation, BiCMOS process technology and BiCMOS applications are all included in the discussion.

### 5.1 PROCESS ARCHITECTURE

A standard IC process involves transferring an IC layout to Si wafers. This process is normally done by printing IC layers from a set of masks through photolithography followed by subsequent processing such as etching, im-
plantation, diffusion, etc. The sequence of applying the photolithographic masks is important for the fabrication of devices and ICs. In general, a bottom-up approach (meaning the layers at the bottom are formed first) is used because the layers are stacked during IC processing. However, options exist in permutating the masking sequence and even the total number of masks. Trade-off commonly occurs between the number of masks and process complexity.

The art of defining mask count and sequence and overall process integration will thereby be referred to as process architecture. A process architect is responsible for fabricating ICs by integrating the necessary process modules in the appropriate order. Mask counts, process simplicity and device performance are important considerations in establishing a good process architecture.

To establish a process architecture, the correspondence between the IC layout and device cross-section at any process step must be fully understood. A CAD program, SIMPL-2 (SIMulated Profiles from the Layout—Version 2), has been developed for this purpose. Figure 5.1 shows an example for a CMOS process. The layout of a CMOS inverter and its corresponding cross-sections are generated by the SIMPL-2 program. The bold horizontal line on the layout is the user-specified "cut-line" along which a cross-sectional view is generated.

A CMOS process architecture has to provide both $n$- and $p$-channel transistors on a Si wafer. As a result, various choices exist in defining the process architecture. Forming a $p$-well in an $n$-type substrate is one option, or an $n$-well can be created in a $p$-substrate. Other alternatives are twin-tub, retrograde-well, and quad-well technology. Which one should be chosen for VLSI has been a controversial subject.

One reason for the CMOS controversy is that two device types used in the technology perform differently, and at least one of them must be located in a well. As described in Chapter 2, an $n$-channel transistor delivers more drain current because electrons move faster than holes; however, unwanted substrate current in an $n$-channel device is also higher due to higher impact ionization. Figure 5.2 shows these two fundamental differences between $n$- and $p$-channel devices. The devices in the well suffer from higher junction capacitance and stronger body effect. Substrate current in the well is also harder to collect. A device technologist may think that if one type of device has to suffer from CMOS integration, it ought to be the better device ($n$-channel in this case) to balance the performance between the two devices. A circuit designer, on the other hand, may want the better device to be maintained or further optimized and choose to avoid using the other type. One well-known example is the domino logic (described in Chapter 4), which uses only one $p$-channel device for several $n$-channel transistors. Whether CMOS circuits should be made truly complementary depends on the performance difference between the two device types. The ratio of the $p$-channel saturation current to the $n$-channel saturation current increases from 1/4 to 1/2 when the devices are scaled to 1 µm [Fig. 5.2] because electron and hole velocities start saturating to a common asymptote. Below that, the $p$-channel current approaches the $n$-channel current. Thus the twin-tub approach, which discriminates neither $n$- nor $p$-channel devices, would be attractive. In the following sections, we will discuss all of these options and compare their pros and cons.
5.2 P-WELL PROCESS

Conventional CMOS was realized by putting n-channel transistors in a p-well formed by diffusing boron atoms into an n-type substrate. P-channel devices were made outside the well, in the n-substrate. Figure 5.3 shows the layout masks and the corresponding device cross-sections of a typical p-well CMOS process. The first mask defines p-well regions by opening windows in an oxide masking layer. Boron atoms are introduced into these windows by a shallow ion implantation followed by a high-temperature diffusion. A p-type well with a depth of a few microns is then formed (Fig. 5.3a). A layer of nitride is then deposited and patterned to cover the active areas in which transistors are to be built. LOCOS (LOCal Oxidation of Si) is used to grow a thick (0.5–1.0 μm) field oxide among the active areas (Fig. 5.3b). After etching away the nitride layer, a thin gate oxide (200–500 Å) is grown on the active areas as the gate oxide. A layer of polysilicon material is deposited and doped to n-type by phosphorus. The poly-layer is then patterned and reactive-ion-etched to form MOS gates for both n- and p-channel FETs (Fig. 5.3c). To form source/drain (S/D) regions, a layer of photoresist is patterned to have large windows opened on one type of FETs, e.g., p-channel FETs.

Boron ions are implanted into the p-channel active areas forming p+ source and drains (Fig. 5.3d). Poly-gates in the active areas block boron ions going to the channel region, resulting in p+ self-aligned to the poly-gates.

A similar process is applied to form n+ source and drains for n-channel FETs by implanting arsenic ions through an n+ mask (Fig. 5.3e), which is basically the complement of the p+ mask. Following the stripping of the photoresist layer, a layer of oxide is Chemically Vapor Deposited (CVD) and contact holes are patterned and etched on gates, p+ region, and n+ regions to make electric contacts (Fig. 5.3f). A layer of metal (normally aluminum)
of more processing steps. Details of this process will be discussed in Chapter 6.

### 5.3 N-WELL PROCESS

Recent development has resulted in an n-well technology\(^7,10\) with a structure identical to Fig. 5.3g if n- and p-labels are interchanged in the figure. The idea is to make n-channel FETs in a low-resistivity p-type substrate rather than a more heavily doped p-well. N-channel devices formed by this CMOS process are equivalent to FETs produced by NMOS technology. Device design and process architecture for NMOS technology is therefore transferable to this CMOS process.

Process architecture is similar to p-well architecture with the exception that phosphorus atoms are used to form n-wells in a lightly doped p-substrate. In addition, the field region outside the n-wells should be doped to enhance isolation for n-channel FETs. This additional doping can be done with boron implantation before LOCOS.\(^11\) The boron dose should be sufficiently high to offer high field threshold for n-channel transistors, but still low enough so that the boron concentration is fully compensated by the n-well dopants under the field oxide. Consequently, a reasonable field threshold for p-channel devices can be maintained. The finished CMOS device structure is shown in Fig. 5.4.

Another alternative is to implant boron atoms only in the n-channel field region to avoid boron/phosphorus compensation. However, this process would require an extra mask and slightly larger layout area for maintaining the separation between the n-well and the p-type channel stop. The process sequence for this architecture is shown in Fig. 5.5. Notice that the n-channel field implant mask overlaps the actual n-well edge (Fig. 5.5b), which extends beyond the n-well mask edge due to lateral diffusion. Shallow implants for adjusting n- and p-channel threshold voltages are also included.

![Figure 5.4](image) Cross-section of a finished n-well CMOS structure with self-aligned p-type channel stop.
Sec 5.4 p-Well versus n-Well

The extra dose needed for the n-channel can be fixed at the $V_{Tn}$-adjust ion implant. The remaining process steps follow conventional MOS process.

5.4 P-WELL VERSUS N-WELL

Even though no process clearly wins today in selecting a technology, it is useful to know the pros and cons of the two distinct well configurations and their impacts on circuit applications.

Present CMOS technology offers 1–2 μm design rules. At these dimensions, n-channel devices, when compared with p-channel devices, provide about twice the driving current but almost four orders of magnitude more substrate current. These fundamental differences arise from carrier mobility and the impact ionization rate (described in Chapter 2), and they affect technology selection in many ways. Another fundamental limit is that the doping concentration in the well has to be higher than in the starting substrate, thus resulting in higher junction capacitance and more body effect for devices made in the well.

Material-related issues and process considerations serve as practical constraints. For example, epitaxy material types that have been used to reduce latch-up susceptibility have different implications. It has been shown that p-type epi grown on p+ substrate provides a longer (msec) minority lifetime than the case of n-epi on n+ substrate, due to better intrinsic gettering. But, out-diffusion of p-on-p+ is much more severe than that of n-on-n+. Because boron diffuses much faster than arsenic or antimony, this point of view, DRAMs or other dynamic circuits should be built on p-on-p+ epi (meaning n-well technology), whereas static circuits ought to be made in n-on-n+ epi (p-well) for sharper epi interface, hence better latch-up protection.

It is also known that, during oxidation, boron segregates into oxide but phosphorus piles up at the silicon surface. This phenomenon and the fact that the oxide fixed charge is normally positive make the field region among n-channel transistors more sensitive to inversion problems. While a p-well process can use the well itself as an n-channel channel stop, an n-well process has the burden to produce separate p-type channel stops for n-channel devices. Moreover, if ion implantation is used rather than diffusion to form a well (a so-called retrograde-well), it is easier to form a p-well than an n-well because boron ions penetrate deeper than arsenic or phosphorus ions for a given implant energy.

The choice of well type depends highly on circuit applications. For n-MOS rich circuits, such as domino logic or cascade voltage switch logic, n-well technology, which allows n-channel transistors to be built in the substrate, should be chosen. On the other hand, the p-well approach may be a better choice for pure static logic to balance the performance of the two device
types. $P$-well also favors devices that require an isolated $p$-region, e.g., $n$-channel FETs for analog input.

In the case of RAMs, if alpha-particle induced SER (Soft Error Rate) is a major limitation in RAM scaling, the RAM cells ought to be made inside a well. For DRAMs, $p$-channel arrays in an $n$-well might be suitable because $p$-channel devices have low substrate current whereas the high substrate current in $n$-channel devices can be easily sunk when they are in substrate. The $p$-channel speed disadvantage does not degrade DRAM performance greatly because DRAM sensing is limited primarily by the amount of charge stored. SRAM sensing is, however, different and it depends on current provided by the SRAM cell. Hence, high gain $n$-channel transistors are more desirable for pass gates and drivers in a cell. A cell should be put in a $p$-well for low soft error rate as described in Chapter 1. For high voltage applications, such as EPROM and EEPROM, an $n$-well is appropriate because sinking $n$-channel substrate current is critical.

### 5.5 TWIN-TUB PROCESS

The twin-tub approach forms two separate wells for $n$- and $p$-channel transistors in a lightly doped substrate. The complete device structure is shown in Fig. 5.6 in which the substrate can be either $n$- or $p$-type. The original claim for this structure was that doping profiles in each well could be set independently; hence both device types would be optimized. In a single-well scheme, the substrate doping concentration must be an order of magnitude lower than the well concentration but meanwhile, must be high enough to prevent puncture-through for the devices made in the substrate. This claim is not always true, because state-of-the-art MOS technology uses a shallow implant to prevent punchthrough without raising the entire substrate doping concentration. In terms of lower junction capacitance and less body effect, this method actually produces better devices than those attainable with the twin-tub approach.

The major advantage of the twin-tub approach is the flexibility of selecting substrate type ($n$ or $p$) with no effects on transistor performance; the latchup behavior however, will not be identical. This flexibility may be important in implementing designs with different applications. In addition, self-aligned channel stops can be easily implemented with this approach. Consequently, spacing between an $n$- and a $p$-channel device can be reduced for high density circuits.

![Figure 5.6 Twin-tub CMOS structure](Ref. 2).

![Figure 5.7 Twin-tub CMOS process.](a) $P^+$ implant for $n$-tub; (b) $B^+$ implant for $p$-tub; and (c) well diffusion and strip oxide (Ref. 2, © 1980, IEEE).
Figure 5.7 shows the process sequence at the front-end of the twin-tub process. A layer of patterned nitride is used to mask a phosphorus implant for n-tubs. Subsequent to the growth of a masking oxide, the nitride layer is removed and the oxide is then used as a mask during boron implantation for p-tubs. With this method, n- and p-tubs are formed using only one mask. Moreover, the two tubs are mutually self-aligned. The remaining process follows active area definition using LOCOS and similar process steps described for p-well process architecture.

As CMOS technology advances to submicron dimensions, the twin-tub approach may become more attractive for the following reasons. Because the two device types perform similarly in the half-micron regime, it makes sense to provide symmetric n- and p-channel devices. Because the doping concentration will be scaled up at these dimensions anyway, whether devices are made in the well or substrate makes only a marginal difference. Submicron technologies such as trench isolation and epi substrate work well with the twin-tub approach. For example, trench sidewall inversion is less likely when both sidewalls are butted against highly doped wells. This problem will be discussed in Chapter 7. Moreover, when epi is used, this approach offers greater flexibility in choosing n-on-n⁺ or p-on-p⁺, and even n-on-p⁺ or p-on-n⁺ if BiCMOS (bipolar/CMOS) chips are implemented.

### 5.6 RETROGRADE-WELL PROCESS

Conventional wells are formed by diffusion, which is an isotropic process, meaning impurity atoms diffuse laterally as well as vertically. Lateral diffusion takes up Si area resulting in poorer packing density. High-energy ion implantation, when used for well formation, provides minimal lateral spread because of the anisotropic nature of the implantation process. In a retrograde-well process architecture, wells are formed after active area definition and LOCOS. As a result, the lateral spread of the well is further minimized because the high-temperature LOCOS process is done prior to well formation. Figure 5.8 shows the cross-section of an implanted p-well in comparison with a conventional diffused well. In this example, the implanted p-well is formed by boron implantation at 400–600 KeV followed by a brief 30 min. anneal at 1000°C instead of a 20 hr. drive-in at 1100°C used for conventional well formation. Notice that the lateral spread associated with the p-well is greatly reduced. This decrease in spread has led a reduction of p⁺-to-n⁺ spacing from 12 μm to 9 μm. As a result, a 4K-bit CMOS SRAM was shrunk by 25% in all linear dimensions using this process.

Unlike a diffused profile in which peak concentration is always at the Si surface, the peak of the implanted profile is buried at a certain depth (depending on the implant energy) inside the Si substrate and the impurity concentration decreases as it approaches the Si surface. This type of profile

![Figure 5.8](image)
is called retrograde profile, and the implanted well is often referred to as the retrograde well. Comparison of conventional and implanted p-well profiles are shown in Fig. 5.9. An n-type retrograde-well process has also been proposed using an even higher implant energy: 700 KeV phosphorus. As shown in Fig. 5.10, the process involves lifting-off a layer of evaporated Si film and using a Si/Si$_3$N$_4$/SiO$_2$ multi-layer film as a mask. As a result, the p-type channel stop is self-aligned to the n-well, but at the expense of process complexity. A simpler retrograde n-well process was recently demonstrated.

Figure 5.11 shows the front end of the process. Similar to a standard nMOS process, nitride is used as a mask for blanket field boron implant and LOCOS. Next, a resist mask is used to implant phosphorus ions for n-well formation. Because the phosphorus concentration for the n-well portion underneath the field oxide is an order of magnitude higher than the boron concentration there, the resultant net impurity is n-type with $10^{17}$ cm$^{-3}$ concentration. Quadruple-well structure uses two very shallow wells in the field region as channel stops in addition to the two relatively deep wells for active transistors. Both deep wells are retrograde. Figure 5.12 shows the process architecture in which liftoff is used. However, liftoff can be eliminated by using an extra mask.

Figure 5.9 Comparison of conventional and retrograde p-well profiles (Ref. 4, © 1981, IEEE).

Figure 5.10 A retrograde n-well CMOS process. (a) High energy $p^{++}$ implant to form n-well with a resist mask; (b) the same resist is used to lift off Si; (c) the remaining Si and another resist mask are used to implant $B^+$ for a p-type channel stop; (d) transfer of the resist mask to the underlying nitride; and (e) LOCOS and strip resist, then $BF_2$ for channel implants (Ref. 16, © 1985, IEEE).
Figure 5.11 A simplified retrograde n-well process. (a) Nitride mask for active area and blanket $B^+$ implant for p-type channel stop; (b) photoresist mask for n-well; and (c) high energy $P^+$ implant and any additional shallow implants in the n-well (Ref. 11, © 1986 IEEE).

The major benefit of retrograde-well processes is high packing density through the reduction of $p^+\text{-}to\text{-}n^+$ spacing. This approach is desirable for radiation-hard applications because of the need for high field threshold voltages at a small isolation spacing and it is also scalable for half-micron technology. Other advantages are:

1. Providing a retarding electrical field that reduces the current gain of the vertical bipolar transistor.
2. Increasing conductivity at the bottom of the well, which decreases well resistance, hence enhancing latchup resistance. It also increases breakdown voltage of vertical punchthrough between the drain in the well and the substrate.

Figure 5.12 Fabrication sequence for quadruple well CMOS. (a) Define active area by oxide windows; (b) a thick, preferable 2-level photoresist mask for n-well implant; (c) evaporate and lift off a layer of Al; (d) implant $B^+$ for p-wells; and (e) strip Al and follow standard processes to complete the CMOS structure (Ref. 8, © 1984 IEEE).
(3) Reducing junction capacitance and body effect if the implant energy is sufficiently high (approaching 1 MeV) to move the highly doped region away from the channel.

(4) Enhancing latchup immunity, because its low thermal cycle is more compatible with a shallow epitaxial layer, which is crucial for latchup resistance.

These benefits will be discussed in Chapter 8. Moreover, for BiCMOS applications, the highly conductive layer near the bottom of the retrograde well can also be used as a buried layer if bipolar devices are made in the well.

Concerns for retrograde-well processes are the need for thick photoresist as a masking material for high energy implant, annealability of Si damage caused by high energy ions, and other practical considerations for high-volume production. Although the retrograde-well technology has been shown in laboratories, it should be mentioned that manufacturability has not been established in production due to the need of high-throughput implanters operating at ion energies above 400 KeV.

## 5.7 CHOICE OF PROCESS ARCHITECTURES

In addition to considering process simplicity and device performance, the choice of a particular CMOS process architecture strongly depends on circuit applications. As described above, a $p$-well process may be desirable for SRAM design if alpha particle induced soft error is the main concern. On the other hand, for most logic circuits designed with Domino logic, an $n$-well approach is more suitable. The relationship between process choice and circuit design is summarized in Table 5.1. For analog circuits, high gain $n$MOS devices are preferred over $p$MOS devices. In addition, the substrate is often connected to the source to avoid body effect; therefore, isolated $p$-wells are required.

Process choice also relies on the historical development of a Si house; for example, a well-established NMOS facility may have chosen the $n$-well approach because all of the $n$-channel processes are portable, whereas an established CMOS manufacturer may maintain its original $p$-well technology so that existing designs can be used. Scaling of CMOS to half-micron dimensions may make twin-tub or retrograde-well technologies more attractive for the reasons previously mentioned.

## 5.8 SOS TECHNOLOGY

All of the process architectures described so far are based on bulk CMOS technology from which most CMOS ICs are fabricated. However, another CMOS technology involves making circuits on a sapphire instead of a silicon substrate. Sapphire is an insulating material with its lattice sufficiently matched to silicon so that silicon can be epitaxially grown on the sapphire substrate. This structure, referred to as silicon on sapphire (SOS), offers some unique advantages for CMOS technology, especially for military and aerospace applications.

<table>
<thead>
<tr>
<th>Technology Consideration</th>
<th>Impact on Design</th>
<th>$p$-Well</th>
<th>$n$-Well</th>
<th>Twin-Tub</th>
</tr>
</thead>
<tbody>
<tr>
<td>Optimizing $n$MOS performance</td>
<td>$n$MOS rich, dynamic circuits</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Balancing $n$MOS and $p$MOS performance</td>
<td>Static</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sinking substrate current caused by impact ionization</td>
<td>DRAM and E'PROM</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Collecting α-particle-induced carriers</td>
<td>RAM cells in a well</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Using low $I_{on}$ FETs for memory arrays</td>
<td>DRAM cells</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Using high-gain drivers and pass gates</td>
<td>SRAM cells</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Eliminating body effect</td>
<td>Analog input</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Switching substrate type</td>
<td>Design flexibility</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Scaling down to submicron</td>
<td>Future VLSI</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Using epi: $n/n^+$</td>
<td>Latch-up-free</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>$p/p^+$</td>
<td>DRAM</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Using trench isolation</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Using retrograde well</td>
<td>High density</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Integrating bipolar on CMOS chip</td>
<td>BiCMOS</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Using well-established CMOS technology</td>
<td>Existing designs</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Modifying existing $n$MOS process</td>
<td>$n$MOS compatible</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
In a CMOS/SOS process, sapphire wafers with a thin (0.3–0.5 \( \mu \)m) Si epi layer are commonly used as a starting material. Most IC facilities purchase them from SOS wafer suppliers. SOS wafers in general are smaller in size (2–3 inches in diameter) and cost much more than bulk Si wafers. Crystal defects also exist at or near the Si/sapphire interface and techniques such as solid-phase epitaxial re-growth\(^\text{17}\) have been attempted to improve SOS material quality before IC processing.

As shown in Fig. 5.13, a CMOS/SOS process normally starts with making Si islands by etching grooves down to the sapphire substrate. These islands are made on the insulating substrate and separated by air. Using ion implantation with photoresist masks, \( p^-\) and \( n^-\) type islands are formed for \( n^-\)

![CMOS/SOS process diagram](image)

**Figure 5.13** CMOS/SOS island process. (a) Island definition by etching Si; (b) \( B^+ \) / I for nMOS; (c) \( P^- \) / I for pMOS; (d) Poly gate definition; (e) As \(^+\) / I for nMOS S/D; (f) \( B^- \) / I for pMOS S/D; and (g) contacts and metal to complete the process.

and \( p^-\) channel MOSFETs, respectively. Subsequent process steps are similar to the bulk CMOS process discussed in Sec. 5.2. An obvious advantage of this process is that MOSFETs are nicely isolated with a minimum of isolation spacing. However, the corresponding topography has made step coverage of poly or metal lines over the island edges difficult. An alternative is to use a nitride mask for partial Si etching (Fig. 5.14(a) and recessed LOCOS. This process results in a planar structure at the expense of the LOCOS-induced bird’s beak. The remaining process steps are identical to a standard CMOS process. Both process architectures require a minimum of eight masks consisting of: active area, \( p^-\) island, \( n^-\) island, poly-gate, \( n^+\), \( p^+\), contacts and metal.

In addition to the advantage of device isolation provided by the insulating substrate, SOS offers several other benefits. First, latchup does not exist because there is no \( pnnpn\) path in SOS. Second, because its S/D regions are located directly on the sapphire substrate, junction capacitance and soft error are minimized. Third, with the thick (300 \( \mu \)m) insulating substrate, metal to substrate capacitance is low. Finally, the corresponding process can be simpler because neither well nor channel stop is needed.

Several disadvantages are associated with this technology in addition to the wafer size and cost previously mentioned. Crystal defects at the Si/sapphire interface always produce lower electron mobility in SOS, even though its hole mobility is comparable to that of bulk Si. Moreover, higher leakage current is normally observed in SOS due to the parasitic back-channel FET.

![LOCOS oxidation diagram](image)

**Figure 5.14** CMOS/SOS isoplanar process. (a) Nitride mask for active area, then etching half of the Si layer away; and (b) LOCOS oxidation to form recessed (i.e., planar) field oxide.
Another problem is related to the parasitic capacitance of metal interconnects. In spite of the lower metal-to-substrate capacitance \( (C_s) \) when compared with bulk Si, SOS has higher metal-to-metal capacitances \( (C_1 \) and \( C_2 \) in Fig. 5.15) because the high dielectric constant of sapphire \( (\varepsilon_{\text{sapphire}} = 11, \varepsilon_{\text{oxide}} = 3.9) \) provides a stronger capacitive coupling between metal lines. This inter-metal capacitance becomes a dominant capacitance as the metal lines and spacings are scaled down for VLSI interconnects. Consequently, as shown in Fig. 5.15, the total interconnect capacitance in SOS is comparable or higher than that of bulk for one-micron CMOS circuits. Additionally, because the interconnect capacitance is dominant in total parasitic capacitance for VLSI, one-micron SOS circuits do not have a speed advantage over their bulk counterparts. Table 5.2 shows a checklist of various considerations for VLSI circuits in favor of CMOS/SOS, bulk CMOS or basically no difference.

**TABLE 5.2** Comparison of SOS and Bulk Si Technologies in Various Considerations

<table>
<thead>
<tr>
<th>Considerations</th>
<th>Favoring SOS</th>
<th>Favoring Bulk</th>
<th>No Difference</th>
</tr>
</thead>
<tbody>
<tr>
<td>Density</td>
<td>x</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Isolation and latch-up</td>
<td>x</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Radiation hardness</td>
<td>x</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Capacitance to Sub.</td>
<td>x</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Soft error</td>
<td>x</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Leakage</td>
<td>x</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Noise margins</td>
<td>x</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mobility</td>
<td>x</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Substrate defects &amp; floating sub.</td>
<td>x</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Materials cost/availability</td>
<td>x</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Inter-wire capacitance</td>
<td>x</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total parasitic capacitance</td>
<td>x</td>
<td></td>
<td></td>
</tr>
<tr>
<td>VLSI speed</td>
<td>x</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
5.9 SOI TECHNOLOGY

Insulating substrates other than sapphire have also been used for fabricating CMOS ICs, commonly referred to as silicon-on-insulator (SOI) technology. In this technology, a crystalline Si film is grown on a layer of oxide or nitride formed on Si wafers. The idea is based on inexpensive and readily available substrate. Furthermore, the use of stacked, thin films opens the potential of three-dimensional ICs. Many SOI techniques such as dielectric isolation with substrate removal,\textsuperscript{20} Full Isolation by Porous Oxidized Silicon (FIPOS),\textsuperscript{21} Silicon Implanted with OXygen (SIMOX),\textsuperscript{22} Epitaxial Lateral Overgrowth (ELO),\textsuperscript{23–25} and polysilicon recrystallization\textsuperscript{26} have been proposed. Those that are important for CMOS ICs will be discussed in this chapter.

5.9.1 Epitaxy on Insulator

In the SIMOX process shown in Fig. 5.16, high-dose, low energy oxygen or nitrogen atoms are implanted into a Si substrate. At a subsequent high-temperature anneal, these atoms react with Si forming an SiO\textsubscript{2} or Si\textsubscript{3}N\textsubscript{4} film buried in, but near, the surface of the Si substrate. The thin single-crystal Si layer above the insulator is then used as a seed for growing a thicker Si epi-layer. Good electrical characteristics have been obtained for CMOS devices and circuits built on implanted-buried nitride\textsuperscript{27} and oxide\textsuperscript{28} wafers (Fig. 5.17). Leakage current is not excessive but still higher than bulk CMOS, especially at high drain biases. The throughput of implanting oxygen or nitrogen to achieve atomic concentrations at 10\textsuperscript{18} or 10\textsuperscript{17} cm\textsuperscript{-2} has been a practical concern. The application of this technology for VLSI was recently made possible through the demonstration of a 4 K bit CMOS SRAM.\textsuperscript{29} In this work, 18 wafers were implanted in 6.5 hours through the use of a high current implanter.

Figure 5.18 shows the process steps for lateral epi growth over oxide.\textsuperscript{24} A Si epitaxial layer is grown around the SiO\textsubscript{2} island and laterally expanded over the island. A conventional epi reactor can be used for this process, thus no new equipment is needed. \textit{p}MOSFETs made with this technique

![Graph showing subthreshold characteristics for MOSFETs made by SIMOX processes.](image)}
have shown characteristics similar to those in bulk devices.\textsuperscript{24,25} However, minority lifetimes for the ELO materials were found to be an order of magnitude lower than those of bulk control wafers. The process is still in an early stage.

\textbf{Figure 5.17} Continued.

---

**5.9.2 Recrystallization of Polysilicon**

Another SOI technique that has received considerable attention is the recrystallization of polysilicon film. This technique involves recrystallizing, or at least enlarging the grains, of a polysilicon film by scanning a heat source over the film (Fig. 5.19).\textsuperscript{30} Heat sources used for this purpose include lasers, electron beams, graphite strip heaters, and incoherent light sources.

CW laser scanning has been effective in forming larger poly grains. The laser scans a small energy spot at a very short (msec) radiation time. The advantages are rapid heat dissipation and self-limiting energy absorption. Thus, disturbance of devices beneath the SOI layer is minimal. The disadvantages are low throughput and difficulty in producing large-area, single-crystal films. It normally needs a capping layer to control the amount of energy absorption and minimize surface-tension effects.

Electron beam scanning is less sensitive to the capping material and its thickness. It also offers higher power density and can emulate a line source.
by scanning the beam rapidly (MHz rates) in the x direction and moving slowly (10 cm/sec) in the y direction. Grains as large as 50 × 120 μm² have been achieved with this technique. Higher throughput is expected, but commercial equipment is not readily available. Charging of the insulating substrate and e-beam radiation damage are other concerns. As shown in Table 5.3, CMOSFETs fabricated in laser or e-beam recrystallized films are superior compared to those made in as-deposited poly films and their mobility values approach those of bulk Si.

Black body radiation from a strip heater has produced large-area single crystals in which CMOS devices comparable to bulk devices have been demonstrated. A sample is first heated to a high background temperature, then a moving line source produced by the strip heater melts poly-Si films and provides lateral epitaxial growth (Fig. 5.20) at a high throughput. The corresponding high substrate temperature and long radiation time (seconds) however tend to destroy the devices in the substrate. Contamination from graphite is another consideration. Incoherent light sources such as arc lamps are probably cleaner than graphite heaters, but only offer limited power.

**TABLE 5.3** Characteristics of MOSFETs Fabricated in Single-Crystal Si and in Poly Si (Ref. 30, © 1982, IEEE).

<table>
<thead>
<tr>
<th></th>
<th>n-Channel</th>
<th>p-Channel</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Mobility</td>
<td>$V_t$</td>
</tr>
<tr>
<td></td>
<td>(cm²/V-sec)</td>
<td>(V)</td>
</tr>
<tr>
<td>Single-crystal silicon</td>
<td>670</td>
<td>1</td>
</tr>
<tr>
<td>Recrystallized polysilicon</td>
<td>&gt;300</td>
<td>1−2</td>
</tr>
<tr>
<td>Fine-grain polysilicon</td>
<td>10−20</td>
<td>10−20</td>
</tr>
</tbody>
</table>

5.9.3 CMOS/SOI for 3-D Integration

SOI technology opens up the possibility of fabricating devices in more than one layer, leading to three-dimensional (3-D) integration. In principle, this integration will overcome the shortcoming of the poor packing density associated with CMOS and take the advantage of its low power characteristic, which is absolutely needed in a 3-D IC. Combining CMOS with SOI has the potential of offering a high-density and low-power technology for future VLSI. However, many technological barriers exist in forming an IC with multiple layers.

As previously discussed, laser or e-beam recrystallization seems more likely in making 3-D ICs because this technique does not greatly disturb the underlying devices in terms of dopant diffusion and oxide damage. A CMOS inverter with one type of FET in the SOI layer and another type in the substrate has been demonstrated using laser recrystallization. The CMOS inverter has two different configurations. The first employs the idea of flipping the p-channel FET on top of the n-channel with a two poly gate joint (Fig. 5.21a). This Joint-gate MOSFET (JMOS) has been successfully demonstrated. However, the difficulties in making the top MOSFET with self-aligned S/D and forming a good SOI layer on a thin gate oxide have made this structure hard to scale for VLSI. The second configuration is to simply stack one FET on top of the other with an insulator in between (Fig. 5.21b). Although one more poly layer is used, the difficulties of JMOS formation do not exist. The corresponding process is also more controllable. The stacked structure has been realized with laser and e-beam recrystallization.
To date, a 1.1K gate array\textsuperscript{36} and a 256 bit SRAM\textsuperscript{37} have been demonstrated for 3-D CMOS/SOI. The 8K bit parallel array multiplier in the gate array exhibits complete operation. The 256-bit SRAM is configured with the NMOS memory cells in the bottom layer and the CMOS peripheral circuits in the top layer as shown in Fig. 5.22. Complete memory operation including the intralayer and interlayer data transfer has been demonstrated. Both these circuits were fabricated with a laser recrystallized SOI layer. An even larger circuit, a 64K SRAM, was demonstrated\textsuperscript{38} using p-channel loads made in a non-crystalline polysilicon SOI layer. The as-deposited polysilicon was hydrogen passivated rather than beam recrystallized for manufacturability. Although the p-channel polysilicon FETs had low mobility, the resultant CMOS configuration provided lower static power than that of a conventional NMOS RAM with poly load resistors. On the other hand, SOI advantages such as latchup free, high density, high alpha-particle immunity also existed in this structure.

Despite the fact that sizable CMOS circuits have been demonstrated in two-layer SOI, 3-D VLSI using CMOS/SOI technology has a long way to go for production. A reliable and high-throughput recrystallization process for SOI layers is needed. Intra- and inter-layer interconnection with planarization must be developed. High yield, fast turn-around and reliability circuits are other concerns for VLSI fabrication.

### 5.10 Bipolar/CMOS Integration—BiCMOS

As described in Chapter 1, CMOS offers low power which is attractive for VLSI, especially in digital applications. However, the speed of CMOS, although comparable with NMOS, is slower than what bipolar can provide. It is particularly true when heavy-loading or long interconnects need to be driven. A bipolar transistor not only can deliver a large current, it also has well-controlled turn-on voltage, good noise margin and small logic swing for...
SALICIDE PROCESS
SILICIDED SOURCE-DRAIN TECHNOLOGY

(a) Diffused Well

- Form Standard Device Up To Diffusions
- Protect Sidewalls Of Poly-Si With Oxide
- Deposit Blanket Metal
- React To Form Metal Silicide Over Silicon Regions
- Remove All Free Metal

(b) Retrograde Well

- Diffused well structure. The large depletion width inside the well necessitates the use of wider field oxide spacing to prevent punch-through.

- Retrograde well formed by high-energy implantation. The implanted dopant reduces the well depletion region which leads to smaller field oxide spacing. Shallower well depths can also be used with this retrograde feature.

Figure 1

Figure 5.20 A retrograde n-well CMOS process. (a) High energy boron implant to form n-well with a resist mask; (b) the same resist is used to lift off Si; (c) the remaining Si and another resist mask are used to implant B⁺⁺ for a p-type channel stop; (d) transfer of the resist mask to the underlying oxide; and (e) LOCOS and strip resist, then BF⁺⁺ for channel implants (Ref. 16, © 1985, IEEE).
Week 12

MOS Devices
8.3.1 Basic Operation

The basic MOS transistor is illustrated in Fig. 8-8 for the case of an n-type channel formed on a p-type Si substrate. The n⁺ source and drain regions are diffused or implanted into a relatively lightly doped p-type substrate, and a thin oxide layer separates the Al metal gate from the Si surface. No current flows from drain to source without a conducting n channel between them, since the drain–substrate–source combination includes oppositely directed p-n junctions in series.

When a positive voltage is applied to the gate relative to the substrate (which is connected to the source in this case), positive charges are in effect deposited on the gate metal. In response, negative charges are induced in the underlying Si, by the formation of a depletion region and a thin surface region containing mobile electrons. These induced electrons form the channel of the FET, and allow current to flow from drain to source. As Fig. 8-8c suggests, the effect of the gate voltage is to vary the conductance of this induced channel for low drain-to-source voltage, analogous to the JFET case. For a given value of Vₒ there will be some drain voltage V₅ for which the current becomes saturated, after which it remains essentially constant.

An important parameter in MOS transistors is the threshold voltage V₅, which is the minimum gate voltage required to induce the channel. In general, the positive gate voltage of an n-channel device (such as that shown in Fig. 8-8) must be larger than some value V₅ before a conducting channel is induced. Similarly, a p-channel device (made on an n-type substrate with p-type source and drain implants or diffusions) requires a gate voltage more negative than some threshold value to induce the required positive charge (mobile holes) in the channel. There are exceptions to this general rule, however, as we shall see. For example, some n-channel devices have a channel already with zero gate voltage, and in fact a negative gate voltage is required to turn the device off. Such a “normally on” device is called a depletion-mode transistor, since gate voltage is used to deplete a channel which exists at equilibrium. The more common MOS transistor is “normally off” with zero gate voltage, and operates in the enhancement mode by applying a gate voltage large enough to induce a conducting channel.

The MOS transistor is particularly useful in digital circuits, in which it is switched from the “off” state (no conducting channel) to the “on” state. The control of drain current is obtained at a gate electrode which is insulated from the source and drain by the oxide. Thus the d-c input impedance of an MOS circuit can be very large.

Both n-channel and p-channel MOS transistors are in common usage. The n-channel type illustrated in Fig. 8-8 is generally preferred because it takes advantage of the fact that the electron mobility in Si is larger than the mobility of holes. In much of the discussion to follow we will use the n-channel (p-type substrate) example, although the p-channel case will be kept in mind also.

8.3.2 The Ideal MOS Capacitor

The surface effects that arise in an apparently simple MOS structure are actually quite complicated. Although many of these effects are beyond the scope of this discussion, we will be able to identify those which control typical MOS transistor operation. We begin by considering an uncomplicated idealized case, and then include effects encountered in real surfaces in the next section.

Some important definitions are made in the energy band diagram of Fig. 8-9. The work function characteristic of the metal (see Section 2.2.1) can be defined in terms of the energy required to move an electron from the Fermi level to outside the metal. In MOS work it is more convenient to use a modified work function ϕₜ₉ for the metal–oxide interface. The energy ϕₜ₉ is measured from
the metal Fermi level to the conduction band of the oxide. Similarly, \( q \Phi \) is the modified work function at the semiconductor-oxide interface. In this idealized case we assume that \( \Phi_m = \Phi_i \), so there is no difference in the two work functions. Another quantity that will be useful in later discussions is \( \phi_f \), which measures the position of the Fermi level below the intrinsic level \( E_f \) for the semiconductor. This quantity indicates how strongly p-type the semiconductor is (see Eq. (3-25)).

The MOS structure of Fig. 8-9 is essentially a capacitor in which one plate is a semiconductor. If we apply a negative voltage between the metal and the semiconductor (Fig. 8-10a), we effectively deposit a negative charge on the metal. In response, we expect an equal net positive charge to accumulate at the surface of the semiconductor. In the case of a p-type substrate this occurs by \textit{hole accumulation} at the semiconductor-oxide interface.

Since the applied negative voltage \textit{depresses} the electrostatic potential of the metal relative to the semiconductor, the electron energies are \textit{raised} in the metal relative to the semiconductor. As a result, the Fermi level for the metal \( E_{Fe} \) lies above its equilibrium position by \( qV \), where \( V \) is the applied voltage.

Since \( \Phi_m \) and \( \Phi_i \) do not change with applied voltage, moving \( E_{Fe} \) up in energy relative to \( E_f \) causes a tilt in the oxide conduction band. We expect such a tilt since an electric field causes a gradient in \( E_f \) (and similarly in \( E_c \) and \( E_v \)) as described in Section 4.4.2.

\[
\phi_f(x) = \frac{1}{q} \frac{dE_f}{dx} \quad \text{(see 4.26)}
\]

\(1\) On the MOS band diagrams of this section we show a break in the electron energy scale leading to the insulator conduction band, since the band gap of SiO\(_2\) (or other typical insulators) is much greater than that of the Si.

\(2\) Recall that an electrostatic potential diagram is drawn for positive test charges, in contrast with an electron energy diagram which is drawn for negative charges.

\[ p = n \mu \text{ carriers} \quad \text{(see 3-25)} \]

The energy bands of the semiconductor bend near the interface to accommodate the accumulation of holes. Since
it is clear that an increase in hole concentration implies an increase in \( E_u - E_F \) at the semiconductor surface.

Since no current passes through the MOS structure, there can be no variation in the Fermi level within the semiconductor. Therefore, if \( E_u - E_F \) is to increase, it must occur by \( E_c \) moving up in energy near the surface. The result is a bending of the semiconductor bands near the interface. We notice in Fig 8-10a that the Fermi level near the interface lies closer to the valence band, indicating a larger hole concentration than that arising from the doping of the p-type semiconductor.

In Fig. 8-10b we apply a positive voltage from the metal to the semiconductor. This raises the potential of the metal, lowering the metal Fermi level by \( qV \) relative to its equilibrium position. As a result, the oxide conduction band is again tilted. We notice that the slope of this band, obtained by simply moving the metal side down relative to the semiconductor side, is in the proper direction for the applied field, according to Eq. (4-26).

The positive voltage deposits positive charge on the metal and calls for a corresponding net negative charge at the surface of the semiconductor. Such a negative charge in p-type material arises from depletion of holes from the region near the surface, leaving behind uncompensated ionized acceptors. This is analogous to the depletion region at a p-n junction discussed in Section 5.2.3. In the depleted region the hole concentration decreases, moving \( E_c \) closer to \( E_F \), and bending the bands down near the semiconductor surface.

If we continue to increase the positive voltage, the bands at the semiconductor surface bend down more strongly. In fact, a sufficiently large voltage can bend \( E_c \) below \( E_F \) (Fig. 8-10c). This is a particularly interesting case, since \( E_v \to E_F \) implies a large electron concentration in the conduction band.

The region near the semiconductor surface in this case has conductivity properties typical of n-type material, with an electron concentration given by Eq. (3.25a). This n-type surface layer is formed not by doping, but instead by inversion of the originally p-type semiconductor due to the applied voltage. This inverted layer, separated from the underlying p-type material by a depletion region, is the key to MOS transistor operation.

We should take a closer look at the inversion region, since it becomes the conducting channel in the FET. In Fig. 8-11 we define a potential \( \phi \) at any point \( x \), measured relative to the equilibrium position of \( E_F \). The energy \( q\phi \) tells us the extent of band bending at \( x \), and \( \phi \) represents the band bending at the surface. We notice that \( \phi = 0 \) is the flat band condition for this ideal MOS case (i.e., the bands look like Fig. 8-9). When \( \phi > 0 \), the bands bend up at the surface, and we have hole accumulation (Fig. 8-10a). Similarly, when \( \phi < 0 \), we have depletion (Fig. 8-10b). Finally, when \( \phi \) is positive and larger than \( \phi_s \), the bands at the surface are bent down such that \( E_c(x = 0) \) lies below \( E_F \), and inversion is obtained.

While it is true that the surface is inverted whenever \( \phi \) is larger than \( \phi_s \), a practical criterion is needed to tell us whether a true n-type conducting channel exists at the surface. The best criterion for strong inversion is that the surface should be as strongly n-type as the substrate is p-type. That is, \( E \) should be as far below \( E_v \) at the surface as it is above \( E_F \) far from the surface. This condition occurs when

\[
\phi(x = 0) = 2\phi_s = 2kT \ln \frac{N_d}{N_i} \tag{8-14}
\]

A surface potential of \( \phi_s \) is required to bend the bands down to the intrinsic condition at the surface \( (E_c = E_V) \), and \( E_F \) must then be depressed another \( \phi_s \) at the surface to obtain the condition we call strong inversion.

The electron and hole concentrations are related to the potential \( \phi(x) \) defined in Fig. 8-11. Since the equilibrium electron concentration is

\[
N_e = n_e \exp\left(-\frac{E_F - E_V}{kT}\right) = n_e \exp(-\phi(x)/kT) \tag{8-15}
\]

we can easily relate the electron concentration at any \( x \) to this value

\[
n = n_e \exp\left(-\phi(x) - \phi(x_0)\right) = n_e \exp\left(-\phi(x) - \phi_s\right) \tag{8-16}
\]

and similarly for holes

\[
p = p_h \exp\left(-\phi(x) - \phi(x_0)\right) = p_h \exp\left(-\phi(x) - \phi_s\right) \tag{8-17}
\]

at any \( x \). We could combine these equations with Poisson's equation (8-18) and the usual charge density expression (8-19) to solve for \( \phi(x) \)

\[
\frac{\partial^2 \phi}{\partial x^2} = \frac{\rho(x)}{\varepsilon} \tag{8-18}
\]

\[
\rho(x) = q\left(N_d^+ - N_i^- + p - n\right) \tag{8-19}
\]

The charge distribution, electric field, and electrostatic potential for the inverted surface are sketched in Fig. 8-12. For simplicity we use the depletion
approximation of Chapter 5 in this figure, assuming complete depletion for
0 < x < W, and neutral material for x > W. In this approximation the charge
per unit area due to uncompensated acceptors in the depletion region is
\( qN_a W \).

The positive charge \( Q_n \) on the metal is balanced by the negative charge \( Q_i \) in
the semiconductor, which is the depletion layer charge plus the charge due to
the inversion region \( Q_o \).

\[
Q_n = -Q_i = qN_a W - Q_o \tag{8-20}
\]

The width of the inversion region is exaggerated in Fig. 8-12 for illustrative
purposes. Actually, the width of this region is generally less than 100 Å.

Thus we have neglected it in sketching the electric field and potential distribu-
tion. In the potential distribution diagram we see that an applied voltage \( V \)
appears partially across the insulator \( (V_i) \) and partially across the depletion region
of the semiconductor \( (\phi_i) \).

\[
V = V_i + \phi_i \tag{8-21}
\]

The voltage across the insulator is obviously related to the charge on either
side, divided by the capacitance

\[
V_i = \frac{-Q_i}{\varepsilon_i} = \frac{Q_i}{C_i} \tag{8-22}
\]

where \( \varepsilon_i \) is permittivity of the insulator and \( C_i \) is the insulator capacitance per
unit area. The charge \( Q_i \) will be negative for n-channel, giving a positive \( V_i \).

Using the depletion approximation, we can solve for \( W \) as a function of \( \phi_i \).
(Prob. 8.8) The result is the same as would be obtained for an n-p junction
in Chapter 5, for which the depletion region extends almost entirely into the
p region:

\[
W = \frac{2\varepsilon_i \phi_i}{qN_a} \frac{1}{\varepsilon_i} \tag{8-23}
\]

This depletion region grows with increased voltage across the capacitor
until strong inversion is reached. After that, further increase in voltage result
in stronger inversion rather than in more depletion. Thus the maximum value
of the depletion width is

\[
W_m = \left( \frac{2\varepsilon_i \phi_i}{qN_a} \right)^{1/2} = 2 \left( \frac{e_k T}{qN_a} \ln(N_D/N_A) \right)^{1/2} \tag{8-24}
\]

using Eq. (8-14). We know the quantities in this expression, so \( W_m \) can be
calculated.

In this chapter we will use charge per unit area \( Q \) and capacitance per unit area \( C \) to
avoid carrying \( A \) throughout the discussion.
EXAMPLE 8.1

Find the maximum width of the depletion region for an ideal MOS capacitor on p-type Si with \( N_a = 10^{16} \text{ cm}^{-3} \).

**SOLUTION**

The relative dielectric constant of Si is 11.8 from Appendix III. We get \( \phi_r \) from Eq. (8-14).

\[
\phi_r = \frac{kT}{q} \ln \frac{N_a}{n_i} = 0.0259 \ln \frac{10^{16}}{1.5 \times 10^{10}} = 0.347 \text{ V}
\]

Thus

\[
W_m = 2 \left( \frac{\varepsilon \varepsilon_r}{q N_a} \right) \left( \frac{(11.8)(8.85 \times 10^{-14})(0.347)}{(1.6 \times 10^{-9})(10^{10})} \right)^{1/2} = 3.01 \times 10^{-3} \text{ cm} = 0.301 \mu\text{m}
\]

The charge per unit area in the depletion region \( Q_d \) at strong inversion is

\[
Q_d = -qN_a W_m = -2(\varepsilon \varepsilon_r \phi_r)^{1/2}
\]

(8-25)

The applied voltage must be large enough to create this depletion charge plus the surface potential \( \phi_s \). The threshold voltage required for strong inversion, using Eqs. (8-14), (8-21), and (8-22), is

\[
V_r = \frac{Q_d}{C} + 2\phi_r \quad \text{(ideal case)}
\]

(8-26)

This assumes the negative charge at the semiconductor surface \( Q_s \), at inversion is mostly due to the depletion charge \( Q_d \). The threshold voltage represents the minimum voltage required to achieve strong inversion, and is an extremely important quantity for MOS transistors. We will see in the next section that other terms must be added to this expression for real MOS structures.

The capacitance–voltage characteristics of this ideal MOS structure (Fig. 8-13) vary depending on whether the semiconductor surface is in accumulation, depletion, or inversion. For negative voltages, holes are accumulated at the surface (Fig. 8-10a). As a result, the MOS structure appears almost like a parallel-plate capacitor, dominated by the insulator properties \( C_i = \varepsilon_i / d \). As the voltage becomes positive, the semiconductor surface is depleted. Thus a depletion-layer capacitance \( C_d \) is added in series with \( C_i \).

\[
C_d = \frac{\varepsilon_i}{W}
\]

(8-27)

where \( \varepsilon_i \) is the semiconductor permittivity and \( W \) is the width of the depletion layer from Eq. (8-23). The total capacitance is

\[
C = \frac{C_i C_d}{C_i + C_d}
\]

(8-28)

The capacitance decreases with positive voltage as \( W \) grows, until finally

---

In the p-channel (n-type substrate) case, for which \( \phi_r \) is negative, we use \( Q_s = +qN_a W_m = \left( \varepsilon \varepsilon_r |\phi_r| \right)^{1/2} \).

---

EXAMPLE 8.2

Using the conditions of Example 8.1 and a 100-Å-thick SiO₂ layer, we can calculate major points on the \( C-V \) curve of Fig. 8-13. The relative dielectric constant of SiO₂ is 3.9.

\[
C_i = \frac{\varepsilon_i}{d} = \frac{3.9(8.85 \times 10^{-14})}{10^{-8}} = 3.45 \times 10^{-7} \text{ F/cm}^2
\]

\[
Q_d = -qN_a W_m = -(1.6 \times 10^{-9})(10^{10})(0.301 \times 10^{-4}) = -4.82 \times 10^{-4} \text{ C/cm}^2
\]

\[
V_r = \frac{Q_d}{C_i} + 2\phi_r = \frac{4.82 \times 10^{-4}}{3.45 \times 10^{-4}} + 2(0.347) = 0.834 \text{ V}
\]

At \( V_r \):

\[
C_d = \frac{\varepsilon_i}{W} = \frac{11.8(8.85 \times 10^{-14})}{0.301 \times 10^{-4}} = 3.47 \times 10^{-4} \text{ F/cm}^2
\]

\[
C_{\text{total}} = \frac{C_i C_d}{C_i + C_d} = \frac{3.45 \times 3.47}{3.45 + 3.47} = 3.15 \times 10^{-5} \text{ F/cm}^2
\]
### 8.3.3 Effects of Real Surfaces

When MOS devices are made using typical materials (e.g., Al–SiO₂–Si), departures from the ideal case described in the previous section can strongly affect \( V_T \) and other properties. First, the work function of Al is not the same as that of Si. Second, there are inevitably charges at the Si–SiO₂ interface and within the oxide which must be taken into account.

**Work Function Difference.** We expect \( \Phi_m \) to vary depending on the doping of the semiconductor. Figure 8-14 illustrates the work function potential difference \( \Phi_m = \Phi_r - \Phi_k \) for Al on Si as the doping is varied. We note that \( \Phi_m \) is always negative for this case, and is most negative for heavily doped p-type Si (i.e., close to the valence band).

If we try to construct an equilibrium diagram with \( \Phi_m \) negative (Fig. 8-15a), we find that in aligning \( E_F \) we must include a tilt in the oxide conduction band (implying an electric field). Thus the metal is positively charged and the semiconductor surface is negatively charged at equilibrium, to accommodate the work function difference. As a result, the bands bend down near the semiconductor surface. In fact, if \( \Phi_m \) is sufficiently negative, an inversion region can exist with no external voltage applied. To obtain the flat band condition pictured in Fig. 8-15b, we must apply a negative voltage to the metal (\( V_{bb} = \Phi_m \)).

**Interface Charge.** In addition to the work function difference, the equilibrium MOS structure is affected by charges in the insulator and at the semiconductor–oxide interface (Fig. 8-16). For example, alkali metal ions (particularly Na⁺) can be incorporated inadvertently in the oxide during growth or subsequent processing steps. Since sodium is a common contaminant, it is necessary to use extremely clean chemicals, water, gases, and processing environment to minimize its effect on dielectric layers. Sodium ions introduce positive charges \( Q_m \) in the oxide, which in turn induce negative charges in the semiconductor.

![Figure 8-14](image-url) **Figure 8-14** Variation of the metal-semiconductor work function potential difference \( \Phi_m \) with substrate doping concentration, for Al–Si

![Figure 8-15](image-url) **Figure 8-15** Effect of a negative work function difference \( \Phi_m < 0 \): (a) band bending and formation of negative charge at the semiconductor surface, (b) achievement of the flat band condition by application of a negative voltage.

![Figure 8-16](image-url) **Figure 8-16** Effects of charges in the oxide and at the interface:
(a) definitions of charge densities \( Q_r / C_m \) due to various sources;
(b) representing these charges as an equivalent sheet of positive charge \( Q_s \) at the oxide–semiconductor interface. This positive charge induces an equivalent negative charge in the semiconductor, which requires a negative gate voltage to achieve the flat band condition.

The effect of such positive ionic charges in the oxide depends upon the number of ions involved and their distance from the semiconductor surface (Prob. 8.13). The negative charge induced in the semiconductor is greater if the Na⁺ ions are near the interface than if they are farther away. The effect of this ionic charge on threshold voltage is complicated by the fact that Na⁺ ions are relatively mobile in SiO₂, particularly at elevated temperatures, and can thus drift in an applied electric field. Obviously, a device with \( V_T \) dependent on its past history of...
voltage bias is unacceptable. Fortunately, Na contamination of the oxide can be reduced to tolerable levels by proper care in processing. The oxide also contains trapped charges \( Q_t \) due to imperfections in the SiO\(_2\).

In addition to oxide charges, a set of positive charges arises from interface states at the Si–SiO\(_2\) interface. These charges, which we will call \( Q_{\text{tr}} \), result from the sudden termination of the semiconductor crystal lattice at the oxide interface. Near the interface is a transition layer \( (\text{SiO}_2) \) containing fixed charges \( Q_f \). As oxidation takes place in forming the SiO\(_2\) layer, Si is removed from the surface and reacts with the oxygen. When the oxidation is stopped, some ionic Si is left near the interface. These ions, along with uncompleted Si bonds at the surface, result in a sheet of positive charge \( Q_p \) near the interface. This charge depends on oxidation rate and subsequent heat treatment, and also on crystal orientation. For carefully treated Si–SiO\(_2\) interfaces, typical charge densities due to \( Q_{\text{tr}} \) and \( Q_f \) are about \( 10^{10} \) charges/cm\(^2\) for samples with \( \{100\} \) surfaces. The interface charge density is a factor of ten higher on \( \{111\} \) surfaces.

For simplicity, we will include the various oxide and interface charges in an effective positive charge at the interface \( Q_i \) (C/cm\(^2\)). The effect of this charge is to induce an equivalent negative charge in the semiconductor. Thus an additional component must be added to the flat band voltage:

\[
V_{FB} = \Phi_{FB} - \frac{Q_i}{C_i} \quad \text{(8-29)}
\]

Since the difference in work function and the positive interface charge both tend to bend the bands down at the semiconductor surface, a negative voltage must be applied to the metal relative to the semiconductor to achieve the flat band condition of Fig 8-16b.

### 8.3.4 Threshold Voltage

The voltage required to achieve flat band should be added to the threshold voltage equation (8-26) obtained for the ideal MOS structure (for which we assumed a zero flat band voltage)

\[
V_T - \Phi_{FB} = \frac{Q_f}{C_i} + \frac{Q_i}{C_i} + 2\Phi_p \quad \text{(8-30)}
\]

Thus, the voltage required to create strong inversion must be large enough to first achieve the flat band condition \( (\Phi_{FB} \text{ and } Q_i/C_i \text{ terms}) \), then accommodate the charge in the depletion region \( (Q_f/C_i) \), and finally induce the inverted region \( (2\Phi_p) \). This equation accounts for the dominant threshold voltage effects in typical MOS devices. It can be used for both n-type and p-type substrates if appropriate signs are included for each term (Fig 8-17). Typically \( \Phi_{FB} \) is negative, although its value varies as in Fig. 8-14. The interface charge is positive, so the contribution of the \(-Q_i/C_i\) term is negative for either substrate type. On the other hand, the charge in the depletion region is negative for ionized acceptors (p-type substrate, n-channel device) and is positive for ionized donors (n-type substrate, p-channel). Also, the term \( \phi_p \), which is defined as \((E_T - E_p)/q\) in the neutral substrate, can be positive or negative, depending on the conductivity type of the substrate. Considering the signs in Fig 8-17, we see that all four terms give negative contributions in the p-channel case. Thus we expect negative threshold voltages for typical p-channel devices. On the other hand, n-channel devices may have either positive or negative threshold voltages, depending on the relative values of terms in Eq (8-30).

All terms in Eq. (8-30) except \( Q_i/C_i \) depend on the doping in the substrate. The terms \( \Phi_{FB} \) and \( \phi_i \) have relatively small variations as \( E_T \) is moved up or down by the doping. Larger changes can occur in \( Q_f \), which varies with the square root of the doping impurity concentration as in Eq. (8-25). We illustrate the variation of threshold voltage with substrate doping in Fig 8-17. As expected from Eq (8-30), \( V_T \) is always negative for the p-channel case. In the n-channel case, the negative flat band voltage terms can dominate for lightly doped p-type substrates, resulting in a negative threshold voltage. However,
for more heavily doped substrates, the increasing contribution of \( N_a \) to the \( Q_d \) term dominates, and \( V_f \) becomes positive.

We should pause here and consider what positive or negative \( V_f \) means for the two cases. In a p-channel device we expect to apply a negative voltage from metal to semiconductor in order to induce the positive charges in the channel. In this case a negative threshold voltage means that the negative voltage we apply must be larger than \( V_f \) in order to achieve strong inversion. In the n-channel case we expect to apply a positive voltage to the metal to induce the channel. Thus a positive value for \( V_f \) means the applied voltage must be larger than this threshold value to obtain strong inversion and a conducting n-channel.

On the other hand, a negative \( V_f \) in this case means that a channel exists at \( V = 0 \) due to the \( \Phi_m \) and \( Q \) effects (Figs. 8-15 and 8-16), and we must apply a negative voltage \( V_f \) to turn the device off. Since lightly doped substrates are desirable to maintain a high breakdown voltage for the drain junction, Fig. 8-17 suggests that \( V_f \) will be negative for n-channel devices made by standard processing. This tendency for the formation of depletion mode (normally on) n-channel transistors is a problem which must be dealt with by special fabrication methods to be described in Section 8.3.6.

**EXAMPLE 8.3**

We can calculate \( V_f \) for the MOS structure described in Examples 8-1 and 8-2, including the effects of flat band voltage. If Al is used for the gate, Fig. 8-14 indicates \( \Phi_m = -0.95 \) V for \( N_a = 10^{20} \text{ cm}^{-3} \). Assuming an interface charge of \( 5 \times 10^{11} \text{ cm}^{-2} \), we obtain

\[
V_f = \Phi_m + 2\phi_f - \frac{1}{C_i} (Q + Q_d) \]

\[
= -0.95 + 0.694 - \frac{(5 \times 10^{11} \times 1.6 \times 10^{-19}) - 4.82 \times 10^{-4}}{345 \times 10^{-8}} \]

\[
= -0.14 \text{ V} \]

This value corresponds to the \( N_a = 10^{20} \text{ cm}^{-3} \) point in Fig. 8-17 for the n-channel case.

### 8.3.5 The MOS Field-Effect Transistor

The MOS transistor is also called a surface field-effect transistor, since it depends on control of current through a thin channel at the surface of the semiconductor (Fig 8-8). When an inversion region is formed under the gate, current can flow from drain to source (for an n-channel device). In this section we analyze the conductance of this channel and find the \( I_d - V_d \) characteristics as a function of gate voltage \( V_g \). As in the JFET case, we will find these characteristics below saturation and then assume \( I_d \) remains essentially constant above saturation.
In a thin region near the surface the depletion charge is not the same as in the bulk material. At point x we have

\[ I_d(x) = \mu_n ZQ_d(x) \Delta V \]  

(8-35)

Integrating from drain to source,

\[ \int_{V_D}^{V_S} I_D(x) dx = -\mu_n ZC_i \int_{V_D}^{V_S} (V_D - V_T - \frac{1}{2}V_D) dV \]

\[ I_D = \frac{\mu_n ZC_i}{L} \left[ (V_D - V_T) V_D - \frac{1}{2}V_D^2 \right] \]  

(8-36)

In this analysis the depletion charge \( Q_d \) in the threshold voltage \( V_T \) is simply the value with no drain current. This is an approximation, since \( Q_d(x) \) varies considerably when \( V_D \) is applied, to reflect the variation in \( V_T \) (see Fig. 8-8b). However, Eq. (8-36) is a fairly accurate description of drain current for low values of \( V_D \), and is often used in approximate design calculations because of its simplicity. A more accurate and general expression is obtained by including the variation of \( Q_d(x) \). Performing the integration of Eq. (8-35) using Eq. (8-33) for \( Q_d(x) \), one obtains (Prob. 8.14)

\[ I_D = \frac{\mu_n ZC_i}{L} \left[ (V_D - V_T - 2\phi_F) \right] V_D - \frac{2\sqrt{2e_d N}}{C_i} [(V_D + 2\phi_F)^2 - (2\phi_F)^2] \]  

(8-37)

The drain characteristics that result from these equations are shown in Fig 8-8c. If the gate voltage is above threshold \( V_G > V_T \), the drain current is described by Eq. (8-37) or approximately by Eq. (8-36) for low \( V_D \). Initially the channel appears as an essentially linear resistor, dependent on \( V_D \). The conductance of the channel in this linear region can be obtained from Eq. (8-36) with \( V_D \ll (V_G - V_T) \):

\[ g = \frac{\partial I_D}{\partial V_D} = \frac{Z}{L} \mu_n C_i (V_G - V_T) \]  

(8-38)

where \( V_G > V_T \) for a channel to exist.

As the drain voltage is increased, the voltage across the oxide decreases near the drain, and \( Q_d \) becomes smaller there. As a result the channel becomes pinched off at the drain end, and the current saturates. The saturation condition is approximately given by

\[ V_D = V_G - V_T \]  

(8-39)

The drain current at saturation remains essentially constant for larger values of drain voltage. Substituting Eq. (8-39) into Eq. (8-36), we obtain

\[ I_D = \frac{Z}{2L} \mu_n C_i V_D^2 \]  

(8-40)

for the approximate value of drain current at saturation.

The transconductance in the saturation range can be obtained approximately by differentiating Eq. (8-40) with respect to the gate voltage:

\[ g_m = \frac{\partial I_D}{\partial V_G} = \frac{Z}{L} \mu_n C_i (V_G - V_T) \]  

(8-41)

The similarity of this expression with Eq. (8-38) is due to the approximations used. A more exact analysis reveals a difference in the subtractive term \( V_T \).

The derivations presented here are based on the n-channel device. For the p-channel enhancement transistor the voltages \( V_D, V_G, \) and \( V_T \) are negative, and current flows from source to drain (Fig. 8-19).

8.3.6 Control of Threshold Voltage

Since the threshold voltage determines the requirements for turning the MOS transistor on or off, it is very important to be able to adjust \( V_T \) in designing the device. For example, if the transistor is to be used in a circuit driven by a 3-V battery, it is clear that a 4-V threshold voltage is unacceptable. Some applications require not only a low value of \( V_T \), but also a precisely controlled value to match others in the circuit.

All of the terms in Eq. (8-30) can be controlled to some extent. The work function potential difference \( Q_m \) is determined by choice of the gate conductor material, \( \phi_F \), depends on the substrate doping; \( Q_r \) can be reduced by proper oxidation methods and by using Si grown in the [100] orientation, \( Q_d \) can be ad-

\[ + \quad \text{Drain} \quad \text{current–voltage characteristics for enhancement transistors. (a) for n-channel} \quad V_G, \quad V_D, \quad \text{and} \quad I_D \text{are positive; (b) for p-channel all these quantities are negative.} \]
justed by doping of the substrate, and $C_r$ depends on the thickness and dielectric constant of the insulator. We shall discuss here several methods of controlling these quantities in device fabrication.

**Silicon Gate Technology.** A straightforward method for reducing $C_r$ is to deposit Si on the gate electrode instead of Al. Silicon can be deposited in a silane reactor (Section 13.4), and heavily doped deposited Si can approximate the desired properties of a metal electrode. Since the deposition occurs onto the insulator layer, the resulting Si layer is polycrystalline.

The use of polycrystalline Si as a gate conductor provides a close match between $D_n$ and $D_f$ (depending on the Fermi level position in the two materials). There are additional advantages of this method that make it attractive for MOS integrated-circuit applications. Unlike Al, a Si gate layer can be raised to high temperature. This allows considerable flexibility in device processing, as we shall see in Section 8.3.8.

**Control of $C_r$.** Since a low value of $V_f$ is usually desired, a thin oxide layer is used in the gate region to increase $C_r = ε_r/ε$ in Eq. (8.30). From Fig. 8-17 we see that increasing $C_r$ makes $V_f$ less negative for p-channel devices and less positive for n-channel with $-Q_r > Q_r$. For practical considerations, the gate oxide thickness is generally 40–100 Å (4–10 nm) in modern devices having submicron gate length. An example of such a device is shown on the page opposite the title page of this book. The gate oxide, easily observable in this micrograph, is 40 Å thick. The interfacial layer between the crystalline silicon and the amorphous $SiO_2$ is also observable.

Although a low threshold voltage is desirable in the gate region of a transistor, a large value of $V_f$ is needed between devices. For example, if a number of transistors are interconnected on a single Si chip, we do not want inversion layers to be formed inadvertently between devices (generally called the field). One way to avoid such parasitic channels is to increase $V_f$ in the field by using a very thick oxide. Figure 8-20 illustrates a transistor with a gate oxide 10 nm thick and a field oxide of 0.5 $μm$. Such a thick oxide layer can be deposited by chemical vapor deposition (e.g., the oxidation of silane).

**EXAMPLE 8.4** Consider an Al–$SiO_2$–Si p-channel device with $N_d = 10^{18}$/cm$^3$ and $Q_r = 5 \times 10^{10}$ C/cm$^2$. Calculate $V_f$ for a gate oxide thickness of 0.01 $μm$ and repeat for a field oxide thickness of 0.5 $μm$.

**Figure 8-20** Thin oxide in the gate region and thick oxide in the field between transistors for $V_f$ control.

Values of $φ_r$, $Q_r$, and $Q_f$ can be obtained from Examples 8.2 and 8.3 if we use appropriate signs as in Fig. 8-17a. The value of $C_r$ for the thin oxide case is the same as in Example 8.2. From Eq. 8-14, $Φ_m = -0.25 \text{ V}$.

$$V_f = -0.25 - 0.694 - \frac{8 \times 10^{-3} + 4.82 \times 10^{-4}}{34.5 \times 10^{-4}} = -1.1 \text{ V}$$

This value corresponds to that expected from Fig. 8-17b. In the field region where $d = 0.5 \mu m$,

$$V_f = -0.944 - \frac{5.62 \times 10^{-4}}{6.9 \times 10^{-9}} = -9.1 \text{ V}$$

The value of $C_r$ can also be controlled by varying $ε_r$. For example, Si$_3$N$_4$ has a relative dielectric constant of about 7, compared with 3.9 for $SiO_2$. Sandwich structures of $SiO_2$ covered with Si$_3$N$_4$ provide good Si–$SiO_2$ interface properties along with higher $ε_r$. Problems arise, however, with charges at the oxide-nitride interface. We shall discuss control of these charges in Section 9.5 as a means of programming $V_f$, values in integrated circuits.

**Threshold Adjustment by Ion Implantation.** The most valuable tool for controlling threshold voltage is ion implantation (Section 5.1.4). Since very precise quantities of impurity can be introduced by this method, it is possible to maintain close control of $V_f$. For example, Fig. 8-21 illustrates a boron implantation through the gate oxide of a p-channel device such that the implanted peak occurs just below the Si surface. The negatively charged boron acceptors serve to reduce the effects of the positive depletion charge $Q_r$. As a result, $V_f$...
becomes less negative. Similarly, a shallow boron implant into the p-type substrate of an n-channel transistor can make $V_T$ positive, as required for an enhancement device.

**EXAMPLE 8-5**

For the p-channel transistor of Example 8-4, calculate the boron ion dose $F_B$ ($B$ ions/cm$^2$) required to reduce $V_T$ from $-1.1$ V to $-0.5$ V. Assume that the implanted acceptors form a sheet of negative charge just below the Si surface.

**SOLUTION**

$$-0.5 = -1.1 + \frac{9F_B}{C_0}$$

$$F_B = \frac{3.45 \times 10^{-7}}{1.6 \times 10^{-19}} (0.6) = 1.3 \times 10^{12} \text{ cm}^2$$

For a beam current of 10 μA scanned over a 650-cm$^2$ target area,

$$10^{-9} (\text{C/s}) \times 10^{12} \text{ (ions/cm}^2\text{)} \times 1.6 \times 10^{-19} (\text{C/ion})$$

The implant time is $t = 13.5$ s.

If the implantation is performed at higher energy, or into the bare Si instead of through an oxide layer, the impurity distribution lies deeper below the surface. In such cases, the essentially gaussian impurity concentration profile cannot be approximated by a spike at the Si surface. Therefore, effects of distributed charge on the $Q_t$ term of Eq. (8-30) must be considered. Calculations of the effects on $V_T$ in this case are more complicated, and the shift of threshold voltage with implantation dose is often obtained empirically instead.

The implantation energy required for shallow $V_T$ adjustment implants is low (50–100 keV), and relatively low doses are needed. A typical $V_T$ adjustment requires only about 10 s of implantation for each wafer, and therefore this procedure is compatible with large-scale production requirements.

If the implantation is continued to higher doses, $V_T$ can be moved past zero to the depletion-mode condition (Fig. 8-22). This capability provides considerable flexibility to the integrated-circuit designer, by allowing enhancement- and depletion-mode devices to be incorporated on the same chip. For example, a depletion-mode transistor can be used instead of a resistor as a load element for the enhancement device. Thus an array of MOS transistors can be fabricated in an IC layout, with some adjusted by implantation to have the desired enhancement mode $V_T$ and others implanted to become depletion loads.

**8.3.7 Substrate Bias Effects**

In the derivation of Eq. (8-36) for current along the channel, we assumed that the source $S$ was connected to the substrate $B$ (Fig. 8-19). In fact, it is possible to apply a voltage between $S$ and $B$ (Fig. 8-23). With a reverse bias between the substrate and the source ($V_S$ negative for an n-channel device), the depletion region is widened and the threshold gate voltage required to achieve inversion must be increased to accommodate the larger $Q_t$. A simplified view of the result is that $W$ is widened uniformly along the channel, so that Eq. (8-25) should be changed to

$$Q_t' = -\left[2e\phi_N(2\phi_T - V_S)\right]^{1/2}$$

The change in threshold voltage due to the substrate bias is

$$\Delta V_T = \frac{\sqrt{2e\phi_N}}{C_0} \left[(2\phi_T - V_S)^{1/2} - (2\phi_T)^{1/2}\right]$$

If the substrate bias $V_S$ is much larger than $2\phi_T$ (typically $-0.6$ V), the threshold voltage is dominated by $V_S$ and

$$\Delta V_T = \frac{\sqrt{2e\phi_N}}{C_0} \left(-V_S\right)^{1/2}$$

(n channel)
where \( V_s \) will be negative for the n-channel case. As the substrate bias is increased, the threshold voltage becomes more positive. The effect of this bias becomes more dramatic as the substrate doping is increased, since \( \Delta V_t \) is also proportional to \( \sqrt{N_s} \). For a p-channel device the bulk-to-source voltage \( V_s \) is positive to achieve a reverse bias, and the approximate change \( \Delta V_t \) for \( V_s \approx 2 \phi_F \) is

\[
\Delta V_t \approx -\frac{\sqrt{2e\rho D}}{C_v} V_s^{1/2} \quad \text{(p channel)} \quad (8-45)
\]

Thus the p-channel threshold voltage becomes more negative with substrate bias.

The substrate bias effect (also called the body effect) increases \( V_t \) for either type of device. This effect can be used to raise the threshold voltage of a marginally enhancement device \( (V_t \approx 0) \) to a somewhat larger and more manageable value. This can be an asset for n-channel devices particularly (see Fig. 8-17). The effect can present problems, however, in MOS integrated circuits for which it is impractical to connect each source region to the substrate. In these cases, possible \( V_t \) shifts due to the body effect must be taken into account in the circuit design.

### 8.3.8 Capacitance Effects and Self-Aligned Transistors

The high-frequency operation of MOS transistors depends upon the capacitances associated with various parts of the device. The intrinsic MOS capacitance due to the gate region itself is essentially that of the MOS capacitor of Fig. 8-13, taking into account the effects of threshold voltage. Since the inversion region is contacted by the source and drain in the transistor, the supply of carriers at high frequency is not limited by thermal generation as in the MOS capacitor. Thus the capacitance after inversion is \( C_v \) (Fig. 8-24). For a 100-Å gate oxide thickness, this capacitance is \( 34.5 \times 10^{-14} \) F/cm². Assuming a gate region with \( L = 1.0 \) μm and \( Z = 5 \) μm, the gate capacitance is 0.017 pF.

![Figure 8-24](image)

**Gate capacitance-voltage characteristic of a p-channel MOS transistor with \( V_s = 0 \).**

In addition to the intrinsic MOS capacitance, there are several parasitic capacitances which limit the high-frequency operation of MOS transistors. The junction capacitance associated with the junctions between source and drain and the substrate is described by Eq. (5-61). This capacitance is often comparable to the MOS gate capacitance, and varies with junction voltage. Therefore, the junction capacitances are significant in high-frequency or fast-switching operations. Another parasitic capacitance arises from metal interconnections over the field oxide. Since the field oxide is generally thick, this stray capacitance is most important in circuits with considerable metallization.

A particularly troublesome stray capacitance results from the overlap of the gate with the source and drain regions (Fig. 8-25a). If the gate does not extend to the source and drain diffusions, an incomplete channel will be formed, and the device will not operate. To avoid this possibility, standard Al gate processing calls for some overlap of the gate past the source and drain edges. This overlap must be large enough to allow for mask registration tolerances and variation in lateral diffusion in the source and drain diffusion step. Because of this overlap, a parasitic capacitance develops between the gate and source and between gate and drain. Since the gate is usually the input and the drain is the output in typical circuit configurations, this overlap capacitance introduces an undesirable feedback effect between input and output at high frequencies (sometimes called the Miller effect). To avoid this problem, several methods of achieving self-aligned gate regions have been developed.

One method for avoiding the gate overlap capacitance is to use ion implantation to shape the source and drain with the edges of the gate (Fig. 8-25b). In this method the gate metal is made narrower than the distance between the gate and source diffusions. An ion implantation is then performed, using the metal gate as a mask, to extend the source and drain regions to the edges of the gate. This procedure takes advantage of the fact that the ion beam arrives essentially perpendicular to the surface. The primary disadvantage of this self-aligned gate procedure is the requirement of a relatively high dose to achieve sufficient conduction in the implanted source and drain extension, and the need for subsequent annealing of the radiation damage. Since the Al metal gate cannot be raised to high temperature, annealing must be done in the range of 500°C Al-

![Figure 8-25](image)

Reduction of parasitic capacitance due to gate overlap: (a) standard processing with overlap allowed for mask registration tolerance, (b) extension of source and drain to the edges of the Al gate by boron implantation.
Though reasonable annealing can result, the residual damage does reduce the conductivity of the implanted regions.

The silicon gate method described in Section 8.3.6 has the additional advantage of fairly good gate alignment. In this procedure the polycrystalline Si gate material is deposited over the gate oxide, and windows for the source and drain diffusion are opened in the Si film and the underlying oxide (Fig. 8-26). Since the polycrystalline Si can be raised to high temperature, the source and drain diffusion can be done with the gate in place. Two important advantages result from this method—the gate oxide is covered immediately after growth by the deposited Si, thereby protecting it during subsequent processing. In addition the edges of the gate serve to define the windows for the source and drain diffusion. As a result, the overlap is only that which results from lateral diffusion. The mask registration tolerance is eliminated.

An additional advantage of the Si gate procedure is that the deposited Si can also be used for interconnections to adjacent devices (Fig. 8-26c). After covering the polycrystalline Si with oxide, another layer of metallization (e.g., Al) can be added. Such multilayer interconnection is of great value in the layout of complex integrated circuits. The reduction of gate-substrate work function difference, the self-aligned gate feature, protection of the gate oxide during most processing steps, and the ease of multilayer interconnection all provide attractive advantages for polycrystalline Si gate technology.

**Figure 8-26**

Silicon gate technology: (a) polycrystalline Si deposited over SiO₂; (b) diffusion of source and drain using the Si gate as a mask to define the channel length; (c) after further processing, the poly-Si gate and interconnection regions are covered with SiO₂. The source and drain regions can now be contacted by Al metal, and Al interconnections can pass over the burred poly-Si interconnections.

### 8.3.9 Short-Channel Effects

When MOS transistors are made very small, as required in many high-speed devices and integrated circuits, the size of the depletion region near the drain in saturation can be comparable to the channel length (Fig. 8-27). When this occurs, the saturation current does not remain constant with increased $V_D$. In deriving Eq. (8.40) we assumed that the channel length $L$ was not changed by the formation of the depletion region at the drain end beyond pinch-off. For short channels the spreading of this depletion region with increased $V_D$ beyond pinch-off causes the effective channel length $L'$ to decrease. Since $I_D$ (sat.) is inversely proportional to channel length in Eq. (8.40), the drain current increases as the drain depletion region spreads into the channel. As a result, $I_D$ (sat.) increases somewhat with $V_D$, rather than remaining constant. Velocity saturation effects (Section 8.2.3) are also likely to occur in short-channel devices.

A variety of other effects occur as the MOS device dimensions are reduced. For example, as Fig. 8-28 illustrates, the source and drain depletion regions can intrude into the channel even without bias, as these junctions are brought closer together in short-channeled devices. This effect is called charge sharing, and drain-induced barrier lowering can result in loss of gate control.

**Figure 8-27**

Short-channel effects.
(a) Reduction of the effective channel length $L'$ in saturation; (b) increase of $I_D$ beyond pinch-off due to the reduction of $L'$.

**Figure 8-28**

Intrusion of the source and drain depletion regions into the channel as the device dimensions are reduced from the long-channel case (dashed lines). The effect of these depletion regions in reducing the usable channel length is called charge sharing, and drain-induced barrier lowering can result in loss of gate control.
sharing, but it should probably be called charge hugging, because the source and drain in effect take part of the channel charge, which would otherwise be controlled by the gate. As the drain depletion region continues to increase with bias, it can actually interact with the source-to-channel junction and lower the potential barrier. This problem is known as drain-induced barrier lowering (DIBL). When the source junction barrier is reduced, electrons are easily injected into the channel and the gate voltage no longer has control of the drain current. Under extreme conditions of encroaching source and drain depletion regions, the two can meet. This punch-through effect results in a continuous depleted region from drain to source. Carriers injected at the source are driven to the drain by the high field between the two electrodes. The current is called space-charge-limited current and is proportional to \( V_G^2/L^2 \).

An effect that is exacerbated by short-channel devices is the subthreshold current, which arises from the fact that some electrons are induced in the channel even before strong inversion is established. For the low electron concentrations typical of the subthreshold regime, we expect diffusion currents (proportional to carrier gradients) to dominate over drift currents (proportional to carrier concentrations). For very short channel lengths, such carrier diffusion from source to drain can make it impossible to turn off the device below threshold. The subthreshold current is made worse by the DIBL effect just mentioned, which increases the injection of electrons from the source.

Electric fields tend to be increased at small geometries, since device voltages are difficult to scale to arbitrarily small values. As a result, various hot carrier effects appear in short-channel devices. The field in the reverse-biased drain junction can lead to impact ionization and carrier multiplication. The resulting holes contribute to the substrate current and some may move to the source, where they lower the source barrier and result in electrons injected from the source into the p-region. In fact, n-p-n transistor action can result within the source-channel-drain configuration and prevent gate control of the current.

Another hot electron effect is the transport of energetic electrons over (or tunneling through) the barrier into the oxide. Such electrons can become trapped in the oxide, where they change the threshold voltage and the I-V characteristics of the device. Hot electron effects can be reduced by reducing the doping in the source and drain regions, so that the junction fields are smaller. However, lightly doped source and drain regions are incompatible with small geometry devices because of contact resistance and other problems. A common design called the lightly doped drain (LDD) uses two doping levels, with heavy doping over most of the source and drain areas but with light doping in a region adjacent to the channel (Fig. 8.29). The LDD structure decreases the field between the drain and channel regions, thereby reducing injection into the oxide, impact ionization, and other hot electron effects.

**Problems**

8.1 Modify Eqs. (8-1) through (8-4) to include effects of the contact potential \( V_C \). Define a true pinch-off voltage \( V_T \) to distinguish this case from \( V_T \) defined in Eq. (8-3).

8.2 Modify Eqs. (8-6) through (8-9) to include \( V_C \). Let \( V_T \) be defined as in Eq. (8-3), and call the true pinch-off voltage \( V_T \).

8.3 Assume the JFET shown in Fig. 8.4 is a Si and has p' regions doped with \( 10^{10} \) acceptors/cm² and a channel with \( 10^{16} \) donors/cm³. If the channel half-width \( w \) is 1 \( \mu \)m, compare \( V_T \) with \( V_C \). What voltage \( V_C \) is required to cause pinch-off when \( V_T \) is included? With \( V_C = -3 \) V, at what value of \( V_T \) does the current saturate?

8.4 If the ratio \( Z/L = 10 \) for the JFET of Prob. 8.3, and \( \mu_n = 1000 \) cm²/V-s, calculate \( I_D(\text{sat}) \) for \( V_T = 0, -2, -4 \), and -6 V. Plot \( I_D(\text{sat}) \) vs. \( V_C(\text{sat}) \).

8.5 For the JFET of Prob. 8.4, plot \( I_D \) vs. \( V_D \) for the same three values of \( V_C \). Terminate each plot at the point of saturation.

8.6 Show that Eq. (8-8) results from integration of Eq. (8-7).

8.7 Redraw Figs. 8.10, 8.11, and 8.12 for the p-channel (n-type substrate) case.

8.8 Show that the width of the depletion region in Fig. 8.12 is given by Eq. (8-23). Assume the carriers are completely swept out within \( W \), as was done in Section 5.2.3.

8.9 An Al-gate n-channel MOS transistor is made on a p-type Si substrate with \( N_a = 10^7 \) cm⁻³. The SiO₂ thickness is 100 Å in the gate region, and the effective interface charge \( Q \) is \( 5 \times 10^{10} \) C/cm².

(a) Find \( W_D \), \( V_B \), and \( V_T \).

(b) Sketch the \( G-V \) curve for this device and give important numbers for the scale.

---

**Figure 8.29** The lightly doped drain (LDD) structure, in which the source and drain regions are doped more lightly near the channel to reduce the electric fields and therefore reduce hot electron effects.
8.10 An Al-gate p-channel MOS transistor is made on an n-type Si substrate with \( N_s = 5 \times 10^{17} \) cm\(^{-3}\). The SiO\(_2\) thickness is 100 Å in the gate region, and the effective interface charge \( Q_i \) is \( 5 \times 10^{10} \) q C/cm\(^2\). Find \( W_s \), \( V_{gs} \), and \( V_{sd} \). Sketch the \( C-V \) curve for this device and give important numbers for the scale.

8.11 Find the threshold voltage for a Si n-channel MOS transistor with \( N_s = 10^{17} \) cm\(^{-3}\), \( \Phi_a = -0.95 \) V, \( Q_i = 10^{10} \) q C/cm\(^2\), and an SiO\(_2\) thickness \( d = 200 \) Å. Repeat for a p-channel device \( (N_d = 10^{17} \) cm\(^{-3}\) \) with the same parameters (except for \( \Phi_a \), which can be calculated from the change in \( E_g \)).

8.12 Calculate sufficient points to construct a figure such as Fig. 8-17b for the transistor parameters shown.

8.13. (a) Find the voltage \( V_{gs} \) required to reduce to zero the negative charge induced at the semiconductor surface by a sheet of positive charge \( Q_s \), located \( x \) below the metal. (b) In the case of an arbitrary distribution of charge \( p(x') \) in the oxide, show that

\[
V_{gs} = -\frac{1}{C_0} \int_0^d x' p(x') dx'
\]

8.14 Obtain Eq. (8-37) by integration of Eq. (8-35), using Eq. (8-33) for \( Q_s \).

8.15 Find the dose of boron (ions/cm\(^2\)) required to form (a) an n-channel enhancement transistor with \( V_t = +1 \) V, and (b) a p-channel depletion transistor with \( V_t = +1 \) V, for substrate doping of \( 10^5 \) cm\(^{-3}\) in each case and the parameters of Fig. 8-17b. Assume that the implanted boron resides just below the Si surface and all impurities are ionized.

8.16 Calculate the substrate bias required to achieve enhancement-mode operation with \( V_t = +0.5 \) V for the n-channel device of Example 8-3. Comment on the practicality of this method of threshold control for thin-oxide transistors.

8.17 When an MOS transistor is biased with \( V_{gs} > V_{ds} \) (sat), the effective channel length is reduced by \( \Delta L \) and the current \( I_{ds} \) is larger than \( I_{ds} \) (sat), as shown in Fig. 8-27. Assuming that the depleted region \( \Delta L \) is described by an expression similar to Eq. (8-23) with \( V_g = V_{ds} \) (sat), show that the conductance beyond saturation is

\[
g_d = \frac{\partial I_{ds}}{\partial V_{ds}} = I_{ds} \text{(sat)} \frac{L}{L - \Delta L}
\]

and find the expression for \( g_d \) in terms of \( V_{ds} \).

8.18 Show that the conductance of the differential element in Fig. 8-18 is \( \frac{dQ}{L} \frac{1}{2/L} \).

8.19 Sketch the \( C-V \) curve for an MOS capacitor with parameters of the device discussed in Example 8-3. What is the relation of this curve to Fig. 8-13?

8.20 The flat band voltage is shifted to \(-2\) V for an Al-SiO\(_2\)-Si capacitor with parameters discussed in Example 8-2. Redraw Fig. 8-15 for this case and find the value of interface charge \( Q_i \) required to cause this shift in \( V_{fb} \), with \( \Phi_a \) given by Fig. 8-14.

8.21 Plot \( I_d vs V_d \) for several values of \( V_{gs} \) for the thin-oxide p-channel transistor described in Example 8-4. Use the p-channel version of Eq. (8-36), and assume that \( I_d \) (sat) remains constant beyond pinch-off. Assume that \( m = 200 \) cm\(^2\)/V.s, and \( Z = 10 \) L.

8.22 A typical figure of merit for high-frequency operation of MOS transistors is the cutoff frequency \( f_c = \frac{g_m}{2\pi C_L Z} \), where the gate capacitance \( C_L \) is essentially \( C \) over most of the voltage range. Express \( f_c \) above pinch-off in terms of materials parameters and device dimensions, and calculate \( f_c \) for the transistor of Prob. 8,21, with \( L = 1 \) \( \mu \)m.

8.23 From Fig. 8-28 it is clear that the depletion regions of the source and drain junctions can meet for short channels, a condition called punch-through. Assume the source and drain regions of an n-channel Si MOSFET are doped with \( 10^2 \) donors/cm\(^2\) and the 1-\( \mu \)m-long channel is doped with \( 10^3 \) acceptors/cm\(^2\). If the source and substrate are grounded, what drain voltage will cause punch-through?

**READ LIST**


Dambkes, H., "Gallium Arsenide HEMTs for Low-Noise GHz Communications Engineering" Microwave Journal 20 (September-October 1989) 1-6


Frein, W. R. "Gallium Arsenide Transistors" Scientific American 257 (August 1987) 80-87

Inoue, K., "Recent Advances in InP-Based HEMT/HBT Device Technology" Fourth International Conference on Indium Phosphide and Related Materials (April 1992) 10-13

Kubat, M. Power Semiconductors. Berlin: Springer-Verlag, 1994

Week 14

MEMS Process Integration
Chapter 3: 

MECHANICAL TRANSDUCERS

1. INTRODUCTION

Numerous micromachined mechanical transducers have been developed or demonstrated, and many of the truly “micromachined” sensor devices being shipped today are mechanical (as opposed to sensors based on conventional integrated circuits that, through inherent mechanisms, sense light, temperature, etc.). This chapter focuses on the phenomena that can be sensed or acted upon with micromachined mechanical devices, but the reader should be aware that important discussion of mechanical transduction is presented in other chapters, since it appears as an intermediate mechanism in many devices. Also, some categories of mechanical devices, such as valves, flow restrictors, pumps, etc., are covered in the separate Microfluidic Devices chapter.

An important addition to “conventional” uses of common semiconductors (mainly silicon) began in the 1970’s as some engineers realized that silicon possessed remarkable mechanical properties. The first mechanical uses of silicon were in the area of strain gauges, where their extremely high sensitivities compared to competing methods quickly won them market share. This was followed by the gradual development of a large variety of applications, and today the dominant micromachined mechanical transducers are pressure sensors and accelerometers. Petersen (1982) provides an excellent overview of the wide variety of silicon applications in mechanical devices that had already been studied by the early 1980’s. (This classic review article is probably the most cited paper in the field of micromachining.)

This chapter begins with a concise review of basic mechanics, followed by a discussion of basic mechanisms. The remainder of the chapter progresses through sensing mechanisms, sensor designs, actuation mechanisms, and specific actuators, including resonators and electrical relays.
2. BASIC MECHANICS

This section is provided to familiarize the reader with some of the basic concepts and terminology of mechanical design and analysis. It is not intended to replace a thorough review of mechanics, and one is encouraged to refer to C and Timoshenko (1990) or Popov (1968). However, as a convenient reference, as an aid to understanding the mechanical scaling laws applicable to microstructure, some of the most relevant concepts have been included.

2.1 AXIAL STRESS AND STRAIN

When a force is applied to a surface, that surface is said to be stressed. The average value of this stress equals the loading force, \( F \), divided by the area, \( A \), on which it is applied,

\[
\sigma = \frac{F}{A} \text{ (typically in N/m}^2\text{, or Pa)}
\]

Forces that act perpendicular to the surface are called axial or normal forces and produce axial or normal stresses. By convention, tensile forces, which pull on the surface, are given a positive sign. Compressive stresses, which push on a surface, are negative. It should be noted that \( \sigma \) is used to represent normal strain and \( \tau \) represents shear stress, as discussed below.

![Illustration of the relationship of an applied force to a beam and the resulting dimensional change. Note that the assumption here is that there is no result dimensional change in \( D \), which is generally not the case, as explained below. A strained beam is longer by \( \Delta L \) from its original length, \( L_0 \). Courtesy K. Honer.](image-url)
When subjected to a stress, materials literally get pushed (or pulled) out of shape. Strain, $\varepsilon$, is a measure of this deformation (within the elastic limit for a material), and equals the change in length, $\Delta L$, divided by the original length, $L_o$, of an object,

$$\varepsilon = \frac{\Delta L}{L_o}$$

(It should be noted that the term microstrain is often used, and simply refers to $\varepsilon \times 10^6$. This is convenient since in many practical situations, strains are in the 1 to 100 microstrain range.)

Most materials of interest obey Hooke’s law; that is, they deform linearly with load. Since load is proportional to stress and deformation is proportional to strain, stress and strain are linearly related. The proportionality constant that relates them is known as the elastic modulus or Young’s modulus of a material, and is usually given the symbol, $E$,

$$E = \frac{\text{stress}}{\text{strain}} = \frac{\sigma}{\varepsilon} \quad \text{(typically in N/m}^2)$$

This parameter is relatively constant below the plastic flow point of a given material. The higher the elastic modulus of a material, the less it deforms for a given stress, and thus the stiffer it is. For example, an incompressible material (i.e., one that would not deform under any stress) would have an infinite Young’s modulus,

$$E = \lim_{\frac{\Delta L}{L} \to 0} \frac{A}{\frac{\Delta L}{L}} = \infty$$

while a “soft” material would deform considerably for a given amount of stress, so its modulus of elasticity would be quite small.

For example, $E$ for Si is 190 GPa (1 Pa = 1 n/m$^2$), 73 GPa for SiO$_2$ (quartz), and 1,035 GPa for diamond. The Young’s moduli of several other materials used in micromachining are given in the Micromachining Techniques chapter. It should be pointed out that the elastic moduli for crystalline materials is dependent on their orientation. However, the calculation of orientation-dependent moduli is beyond the scope of this book and one can usually use average values for initial calculations to good effect. In addition, for most materials, the moduli for tension and compression are equal, but their maximum values of stress in these states may not be equal.
2.2 SHEAR STRESS AND STRAIN

Shear stress is stress due to force applied in parallel to surfaces of an object, as opposed to normal force applied to surfaces for axial stress. Shear stress is denoted by the symbol, $\tau$, to distinguish it from axial stress,

$$\tau = \frac{F}{A} \quad \text{(typically in N/m$^2$)}$$

Illustration of shear strain, $\gamma$, of a rectangular element resulting from a load parallel to the top surface. In order to balance the load on the element, the support anchor is also applying a load, $F$, in the opposite direction, as well as vertical forces directed up on the right side and down on the left (not shown). Courtesy K. Honer.

Shear strain, $\gamma$, is slightly different from axial strain. Whereas axial strain is a measure of linear deformation, shear strain can be thought of as related to an angle that a deformed element's sides make with respect to its original shape (as illustrated above). As in the axial case, shear strain is linearly proportional to shear stress. However, the proportionality constant is different. This new constant, $G$, is called the shear modulus of elasticity,
of an object, shear stress is

\[ G = \frac{\text{shear stress}}{\text{shear displacement angle (rad)}} = \frac{\tau}{\gamma} = \frac{F}{A \Delta X / L} \] (typically in N/m²)

For isotropic materials (those having identical properties in every direction, which is not the case for most single-crystal materials), the shear modulus, G, is related to the elastic modulus, E, by,

\[ E = 2G(1 + \mu) = 3K(1 - 2\mu) \]

where \( \mu \) is Poisson's ratio and K, the bulk modulus, is defined as the ratio of hydrostatic stress to volume compression,

\[ K = \frac{\text{hydrostatic stress}}{\text{volume compression}} = \frac{F}{A \Delta V / V} \] in N/m²

The bulk modulus of a material represents its volume change under uniform pressure. In general, solids are less compressible (larger K) than liquids due to their rigid atomic lattices (e.g., for water, \( K = 2.0 \times 10^9 \) N/m², for aluminum, \( K = 7 \times 10^{10} \) N/m², and for steel, \( K = 14 \times 10^{10} \) N/m² (from Giancoli (1989)).

### 2.3 POISSON'S RATIO

When a material is subjected to an axial load, it deforms in the direction of the load. However, it may also deform in directions perpendicular to the load as shown in the figure below. When subjected to a tensile load, the length of an object will typically increase and its girth will decrease. When compressed, its length will decrease and its girth will increase.

In this situation, there are two strains, one axial (\( \varepsilon_a \)) and one transverse (\( \varepsilon_t \)),

\[ \varepsilon_a = \frac{\Delta L}{L_o} \quad \text{and} \quad \varepsilon_t = \frac{\Delta D}{D_o} \]

and as illustrated below, the longitudinal strain is tensile and the transverse strain is compressive (\( \varepsilon_t \) and \( \varepsilon_a \) will usually be of the opposite sign). Note that the sign convention is \( \varepsilon_a > 0 \) for tension and \( \varepsilon_a < 0 \) for compression. Poisson's ratio is the ratio of the transverse strain to the axial strain,
\[ v = \frac{\text{transverse strain}}{\text{longitudinal strain}} = -\frac{\varepsilon_t}{\varepsilon_l} = -\frac{\Delta D}{D_o} \frac{\Delta L}{L_o} \]

and \( v \) (or \( \mu \)) is always defined as a positive value (as noted above, \( \varepsilon_l \) and \( \varepsilon_t \) usually be of the opposite sign). Typical values of Poisson’s ratio are 0.2 to 0.5 most materials. For most metals, Poisson’s ratio is \( \approx 0.3 \). Rubbers have a Poisson ratio closer to 0.5, which corresponds to a conservation of volume. Cork has Poisson’s ratio near zero, which is why it was chosen as a plug for wine bottles (i.e., it will not expand when strained during insertion).

Illustration of Poisson’s ratio. In addition to straining in the direction of the load, a material also deforms in directions perpendicular to the load. Courtesy Honer.
2.4 COMMONLY USED DEFLECTION EQUATIONS FOR MICROSTRUCTURES

Most micromechanical structures are based on deflection of rectangular cross-sectional beams, torsion of rectangular cross-sectional beams, or deflection of membranes. This is to be expected due to the fact that the thin films from which the structures are fabricated tend to take on such cross sections when patterned. Naturally, the degree to which they are actually "rectangular" depends on the anisotropy of the etch used to form them, but to a first approximation, the rectangular assumption is usually satisfactory. The following subsections present the most common equations for linear deflection of these three basic structures.

2.4.1 STATIC BEAM EQUATIONS

Many common microstructures are based on deflection of beams. They are used in the suspension of rigid plates or by themselves as cantilever devices and are a natural choice for bearing-less motion. For the standard beam equations to be valid, the beams must have lengths at least an order of magnitude greater than their other dimensions and the deflections must be small relative to their lengths as well.

Most micro-beams are constrained at one or two points and subjected to either a point load or an evenly distributed load. The equations that describe the load versus deflection of a beam for these boundary conditions are shown below, for a point load, $F$, in N, or a distributed load, $p$, in N/m.
### Point Load

\[
y(x) = \frac{F}{6EI} \left(3x^2L - x^4\right)
\]

\[
\sigma_{\text{max}} = \frac{FLt}{2I}
\]

### Distributed Load

\[
y(x) = \frac{\rho x^2}{24EI} \left(6L^2 - 4Lx + x^2\right)
\]

\[
\sigma_{\text{max}} = \frac{\rho L^2t}{4I}
\]

\[
y(x) = \frac{Fx}{48EI} \left(3Lx - 4x^3\right)
\]

\[
\sigma_{\text{max}} = \frac{FLt}{8I}
\]

\[
y(x) = \frac{\rho x^2}{24EI} (L - x)^2
\]

\[
\sigma_{\text{max}} = \frac{\rho L^2t}{12I}
\]

\[
\text{where,}
\]

\[
L = \text{length of the beam, in m}
\]

\[
t = \text{thickness of the beam, in m}
\]

\[
I = \text{bending moment of inertia, which for a beam of a rectangular cross section given by,}
\]

\[
I = \frac{1}{12} wt^3 \quad \text{(in m}^4\text{)}
\]

where \(w\) is the width of the beam, in m. It is important to note that no matter \(w\) the boundary conditions or loading, the deflection is proportional to the load. It is a consequence of a linear material and small deflections. The same is true torsional structures and membranes.
2.4.2 STATIC TORSION EQUATIONS

In addition to bending, beams may be twisted about their axes. This is known as torsion and is illustrated in the figure below.

*Illustration of torsion of a rectangular cross-sectional beam about its primary axis. Courtesy K. Honer.*

The constitutive equation for torsional structures deflecting through an angle $\theta$ (in radians) is,

$$\theta = \frac{TL}{KG}$$

where $T$ is the applied torque, and $K$ is a constant depending on geometry.

For a circular beam, $K$ is given by,

$$K = \frac{1}{2} \pi r^4$$

where $r$ is the radius of the beam.

For a rectangular beam of dimensions $x_0$ and $y_0$, $K$ is given by,

$$K = \frac{x_0 y_0^3}{16} \left[ \frac{1}{3} - 0.21 \frac{y_0}{x_0} \left( 1 - \frac{y_0^4}{12x_0^4} \right) \right] \text{ for } x_0 > y_0$$

There are equations in the literature for other shapes. Useful references are Hopkins (1987), Roark and Young (1989), Avallone and Avallone (1996), and Gere and Timoshenko (1990).
2.4.3 STATIC PLATE EQUATIONS

Plates are similar to beams except that plates have widths comparable to their lengths and are typically much thinner than their lengths or widths. In addition, in plates there may be stress gradients in the z-direction (thickness). In these structures, Poisson's ratio becomes important as the lateral strain serves to stiffen the plates. Thus a plate will have less curvature than a beam under equivalent load — approximately \((1 - v^2)\) as much for a plate bending in only one direction. One way of looking at this is that for a simple beam being deflected downward, material closer to the bottom surface of the beam will expand, while the upper portion will contract. A plate can be considered to be an array of parallel beams facing each other, and their interaction opposes these dimensional changes, resulting in greater resistance to bending.

Membranes are generally thinner than plates and typically have fixed (constrained) boundaries (e.g., a drumhead). The stress in the z-direction is typically uniform (not always the case in microstructures). Membranes usually deflect in two directions. Under uniform pressure, \(P\), in N/m\(^2\), a circular membrane with radius \(R\) deflect as follows:

\[
\delta_{\text{max}} = \frac{3Pr^4(1-v^2)}{16Et^3}
\]

The above equation assumes that the membrane is under zero initial stress. This is usually not the case. Residual tensile stresses will tend to reduce the maximum deflection, while residual compressive stresses may result in buckling even in the absence of a load. Micromachined membranes are usually intentionally under moderate tensile stresses.

2.5 DYNAMICS

In the above sections, the loads on the structures were assumed to be constant over time, and no such assumption is made here. The basics of dynamic response for microstructures is described to illustrate that a static analysis may be insufficient to ensure a safe design.

Any lumped mechanical structure can be modeled as a simple mass on a spring. For this case, the governing equation is,
nparable to widths. In  
knness). For  
in serves to  
er equivalent  
e direction.  
downward,  
upper part  
ems fused  
resulting in  
fixed (con-  
is typically  
y deflect in  
mbrane will  
press. This is  
maximum  
even in the  
nally under  
be constant  
ic response  
insufficient  

\[
m \frac{d^2 x}{dt^2} + b \frac{dx}{dt} + kx = F_{\text{external}}
\]

where,
\[
m = \text{mass, in kg}
\]
\[
b = \text{damping coefficient, in (N·s)/m}
\]
\[
k = \text{spring constant, in N/m}
\]
\[
F_{\text{external}} = \text{applied force, in N = (kg·m)/s}^2
\]

Damping arises from any losses that the system experiences when it moves. It can be due to aerodynamic drag of the mass or from small plastic deformations taking place in the spring, among other things. The external load can be any force, such as gravity, electrostatic attraction, or thermal expansion. If, however, \( F_{\text{external}} \) has the form of a sine wave, the movement of the mass will also be sinusoidal at the same frequency (since this is a linear system). The magnitude and relative phase of this sine wave is a strong function of frequency for such a second-order system (a typical log-log plot of this dependence, called a Bode plot, is shown below).

Analyzing this second-order system, one obtains a second-order (low-pass filter) force-to-displacement response of the form (where \( s \) is the Laplace variable),

\[
H(s) = \frac{1}{m} \frac{m}{s^2 + \frac{b}{m} s + \frac{k}{m}}
\]

The electrical analog of this system is an RLC circuit, as illustrated below. This circuit has an electrical transfer function (volts-to-volts) given by,

\[
H(S) = \frac{V_o}{V_i} = \frac{1}{\frac{1}{L} C} = \frac{\frac{\omega_0^2}{Q}}{s^2 + \frac{1}{RC} s + \frac{1}{LC}} = \frac{\frac{\omega_0^2}{Q}}{s^2 + \frac{\omega_0^2}{Q} s + \omega_0^2}
\]

with a natural frequency and quality factor, \( Q \), (a measure of how "lossless" the system is in terms of energy — high \( Q \) systems tend to resonate for long periods of time once energized) given by,
\[ Q = \frac{\text{energy stored per cycle}}{\text{energy lost per cycle}} = \omega_o RC \quad \text{where} \quad \omega_o = \frac{1}{\sqrt{LC}} \]

Schematic of the electrical RLC analog of a second-order mechanical system.

Typical gain and phase response plots (a Bode plot) for a second-order system.
Three important parameters for a second-order system are DC gain, natural or resonant frequency, and quality factor. These can be calculated for mechanical and electrical systems as shown in the table below.

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Mechanical System</th>
<th>Electrical System</th>
</tr>
</thead>
<tbody>
<tr>
<td>DC Gain</td>
<td>$\frac{1}{k}$</td>
<td>1</td>
</tr>
<tr>
<td>Natural Frequency, $\omega_0$</td>
<td>$\sqrt{\frac{k}{m}}$</td>
<td>$\frac{1}{\sqrt{LC}}$</td>
</tr>
<tr>
<td>Quality Factor, Q</td>
<td>$\omega_0 \frac{m}{b} = \frac{\sqrt{km}}{\sqrt{b^2}}$</td>
<td>$\frac{R^2C}{\sqrt{L}}$</td>
</tr>
</tbody>
</table>

*Table showing key parameters for mechanical and electrical second-order systems.*

While not discussed in detail here, it should be noted that for torsional systems, there is a similar governing equation,

$$I \frac{d^2\theta}{dt^2} + b \frac{d\theta}{dt} + k\theta = T_{\text{external}}$$

where $I$ is the second moment of inertia.

For frequencies below the natural frequency, the system will respond with close to the DC gain. At the natural frequency, the response is $Q$ times the DC response and is $90^\circ$ out of phase with the input force (or voltage). At frequencies above the natural frequency, the response falls off by 40 dB per decade and phase approaches a phase of $-180^\circ$.

Tuning forks have a very high $Q$, which is why they resonate at predominately one frequency. This is good for deliberate resonators, as discussed in Section 7, but it brings up a potential material failure problem for other mechanical structures. A comprehensive static analysis, assuming a maximum force, may indicate a safe amount of deflection and hence stress. However, at resonance the deflection, and hence stress, are $Q$ times as large as at steady state. Increasing the damping of the system helps to ensure that out of control oscillations do not cause the structure to exceed its maximum stress levels.
2.6 THERMAL NOISE

As discussed above, damping is desirable unless a structure should deliberate resonate. Unfortunately, anywhere damping is present there is also noise. This directly analogous to the electronic case, wherein pure LC circuits (infinite resistance can have infinite Q in theory). In practice, resistive damping limits Q quite severely. These electrical resistances all generate Johnson noise (thermal noise), which is in spectral density ("white"), but is generally shaped by the transfer function of circuit it is in.

In the case of mechanical noise, the molecules in any material at a temper above absolute zero are constantly vibrating. This vibration produces small random motions in microstructures. When the position of a microstructure in a sense mechanism, this random vibration shows up as noise analogous to Johnson noise electronic circuits. The equipartition theorem states that the value of this noise is

\[
\frac{1}{2} k \langle x^2 \rangle = \frac{1}{2} m \langle v^2 \rangle = \frac{1}{2} k_b T
\]

where,
\[
\langle x^2 \rangle = \text{mean squared average displacement}
\]
\[
\langle v^2 \rangle = \text{mean squared average velocity}
\]

\[k_b = \text{Boltzmann's constant} = 1.38066 \times 10^{-23} \text{ J/K} \quad (\text{note that here } k_b \text{ is used, rather than } k, \text{ to avoid confusion with the spring constant, } k)
\]

\[T = \text{temperature in K}
\]

(Note that this is an idealized model, derived based on the assumption of a thermally driven, statistical ensemble of non-interacting particles, e.g., a gas. In this case should be treated as an approximation.)

In terms of the spectral density of the noise force, one can use Nyquist relation to obtain,

\[F_{\text{noise}} = \sqrt{4k_b b T} \quad (\text{in } \frac{N}{\sqrt{\text{Hz}}})
\]

where simply replacing the damping, b, with the electrical resistance, R, allows this formula to be used for electrical noise. In fact, the same techniques used to model noise in mechanical structures by taking advantage the parallelism between mechanical structures and circuits. The output noise of a mechanical system is white noise filtered by the force-to-displacement response of the system and the electrical transfer function in the electronic case. Noise issues are discussed in more detail in Section 7 below, and the reader is also referred to Gabrielson (1993, 1995) and Motchenbacher and Connelly (1993).
3. MECHANICAL PROPERTIES OF MATERIALS

3.1 MATERIAL FAILURE

As indicated above, when a material is stressed, it undergoes a strain proportional to the stress. This is true up to a point. If the stress is above the yield stress, the material will deform significantly more and will be permanently deformed when the stress is removed. If, however, the stress is above the ultimate stress, the material will fail completely and break into separate sections if not constrained.

For some materials, such as many metals, the ultimate stress is significantly greater than the yield stress. These materials will bend before they break and are referred to as ductile. Materials such as silicon, however, have ultimate stresses that are virtually the same as their yield stresses. These materials will break suddenly and without warning. These materials are brittle. The table below shows the yield strengths for some common materials, with more examples provided in the Micromachining Techniques chapter.

<table>
<thead>
<tr>
<th>Material</th>
<th>Yield Strength (MPa)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Al</td>
<td>170</td>
</tr>
<tr>
<td>Steel</td>
<td>2,100</td>
</tr>
<tr>
<td>W</td>
<td>4,000</td>
</tr>
<tr>
<td>Si</td>
<td>7,000</td>
</tr>
<tr>
<td>Quartz (SiO₂)</td>
<td>8,400</td>
</tr>
<tr>
<td>Diamond</td>
<td>53,000</td>
</tr>
</tbody>
</table>

*Brief table of example yield strengths of materials. From Petersen (1982).*

Determining when a structure will fail is not as simple as summing up the contributions from all loads and checking to see if they exceed the yield stress. For example, some materials fail at lower stresses in shear, others in tension, making the sign of the applied load particularly important. However, checking to make sure stresses are below the yield stress is a good way to eliminate grossly erroneous designs quickly.

A material can also fail even if it is never exposed to a load greater than its yield strength. Over time, cyclic loads can cause fatigue at stresses significantly
below yield values. When a material fatigues, tiny cracks appear as a result of local stress and grow with each cycle until the material breaks. For some materials used in micromachined structures, fatigue failures have been documented and for others such as the < 100 nm metal films used in various mechanical light modulators, large strains can be applied without failure over trillions of cycles (this is discussed in the Optical Transducers chapter). General rules for the fatigue properties of microscale materials are not yet clear, but several groups are actively investigating these issues. Connelly and Brown (1991, 1992) and Brown, et al. (1993, 1994) used deliberate stress concentration regions on resonators to show that the presence of moisture during mechanical cycling had a major influence on crack growth and failure. The effect of stress level on fatigue life was also demonstrated, consistent with intuition from macroscopic devices.

### 3.2 GENERAL MATERIALS CONSIDERATIONS

It is critical to note that the commonly available tabulated values of mechanical properties of materials are generally derived from bulk specimens, and thus may not be very relevant to the materials and scales used in micromachined devices. For example, one can find wide variations in published values of parameters such as Young's moduli for polysilicon, and these are likely to be different due more to the process-specific nature of polysilicon properties than to experimental error. Also, as mentioned above, crystalline materials have anisotropic mechanical properties. Even if the properties were consistent, since the Young's moduli for many of the materials of interest are very high, there are practical issues that interfere with measurements of these materials, such as compliance of measuring tools.

Micromachined structures are often laminates. In macroscopic composite it is well known that the overall properties are not simply a linear interpolative between the properties of the individual constituent materials (e.g., fiberglass behave neither like glass fibers nor like epoxy, nor like anything in between). This is a key concept to keep in mind when analyzing the mechanical (and other) properties of microstructures.

There will be considerable ongoing benefit to more basic science work in the area of mechanical and other properties of thin-film and small-scale materials as the micromachining industry continues to expand. Such efforts have, to some extent, been hampered by the fact that process parameters greatly influence these properties and may be difficult to reproduce on one set of equipment, let alone on the equipment of others. In a great many cases, improved knowledge of material properties would greatly reduce the amount of "cut-and-try" engineering that is common in micromachining endeavors despite the emergence of modeling tools.
3.3 MECHANICAL CHARACTERIZATION OF THIN-FILMS

For many micromachined transducers and structures, the mechanical properties of applied thin-films can be critical to their success, as discussed in Ristic, et al. (1994). For example, stresses or stress gradients can cause thin-film structures (such as surface micromachined polysilicon or metal devices) to warp to the point that they are useless. It is often also necessary to determine other mechanical properties such as Young’s moduli, Poisson’s ratios, etc. This section is concerned with direct and indirect means of measuring or estimating these parameters.

3.3.1 STRESS MEASUREMENT

It should be noted that stresses can be due to the deposition process(es) themselves or to the post-deposition process history of a thin-film. With appropriately designed experiments, it is usually possible to determine the stage at which such stresses arise.

The typical macroscopic method for measuring stresses in thin-films is to quantify optically the differential “bow” or curvature of a wafer before and after the deposition of the film. Commercial instruments that operate on this principle can provide stress resolutions in the MPa range without difficulty. A useful discussion of this technique is provided in Flinn (1988). Microscale methods for measuring such stresses can make use of a variety of methods, including deflection of thin-film structures freed by removal of a sacrificial layer, external mechanical probing of free regions of a film, pressure-based membrane-deflection methods, or resonant frequency measurement.

UNIFORM STRESSES

For films wherein the stress is uniform in the direction perpendicular to the substrate surface, a free region of the film will relax to a non-stressed state, with an accompanying dimensional change.

![Illustration of (left) dimensional change (expansion) of a simple cantilever with a uniform, compressive residual stress when released, and (right) the buckling of a doubly supported beam in the same circumstances.](image)
For a simple cantilever, the length, $L$, will change an amount $\Delta L$, and this can potentially be measured directly relative to a fixed landmark or vernier on the substrate, as demonstrated by Fan, et al. (1988). If the Young’s modulus of the film, $E$, is known, the stress, $\varepsilon$, can be estimated (since $\Delta L/L$ is the strain, $\sigma$) using

$$\varepsilon = \frac{\sigma}{E}$$

As discussed in Ristic, et al. (1994), this type of measurement is difficult with films having large Young’s moduli since the strains are very small. However, one advantage of this approach is that both compressive and tensile stresses can be measured. Test structures that geometrically amplify the strains somewhat can be used in some cases, but the total deflections still generally remain small for reasonable device sizes (see Allen, et al. (1987) and Mehregany, et al. (1987)). If both high-resolution lithography and electron microscopy are available, such measurement approaches may be useful for stress estimates.

For estimating compressive stress, a simple technique is to fabricate doubly supported beams of various lengths and determining which have buckled at a given stress level. As demonstrated by Guckel, et al. (1985), the stress can be estimated in this case using the equation for the critical buckling strain,

$$\varepsilon_{CR} = \frac{\pi^2 t^2}{KL^2}$$

where $t$ is the beam thickness, in m, and $K$ is a constant with value $3 < K < 1$, where the low and high bounds represent a doubly supported beam and a cantilever beam, respectively. In this case, the constant $K$ needs to be determined, as well as the Young’s modulus, in order to obtain a stress value from the estimated strain.

For estimating tensile stress, the “ring-and-beam” test structure (illustrated below) introduced by Guckel, et al. (1988a, 1988b) and Guckel and Burns (1990) can be used. When released from the substrate so that the structure is supported at the ends of two supporting members, the film will relax to relieve residual tensile stress by contracting. As the ring contracts, the central beam will be compressed and, if the compressive forces are sufficient, will buckle. The critical value of strain required for this buckling is given by,

$$\varepsilon_{CR} = \frac{1}{3G \left( \frac{\pi t}{2R_{CR}} \right)^2}$$

where $R_{CR}$ is the ring radius, in m, and $G$ is the geometry-dependent ratio of strain in film to strain in crossbar. With a scaled array of cantilever and/or ring-in-beam
test structures, the approximate value of the residual film stress can be determined by observing (typically by optical microscopic methods, such as interference microscopy) which length beams have buckled.

![Illustration of top and side views of “ring-and-beam” test structures for measuring tensile residual stresses. The geometry used converts tension on the ring into compressive forces on the central beam, causing it to buckle. After Guckel, et al. (1988a, 1988b) and Guckel and Burns (1990).](image)

**NONUNIFORM STRESSES (STRESS GRADIENTS)**

While it is generally possible to design microstructures that are tolerant of uniform residual stresses, the more serious problem is stress gradients. During processing (deposition, thermal cycling, etc.), such gradients can be formed in thin films. Such variations in stress lead to internal bending moments that cause the films to deform (“curl”) when released (converting the stress gradient into a strain gradient). As described in Fan, et al. (1990), the net bending moment is given by,

\[ M = \int_0^l \sigma(y) w \left(y - \frac{t}{2}\right) dy \]

where,
- \( y \) = height above the bottom of the film, in m
- \( \sigma(y) \) = stress at height \( y \), in N/m²
- \( w \) = width of structural member in which the moment is present, in m
- \( t \) = thickness of film, in m

In practice, when such gradients are present, they are generally seen as upward or downward curvature of released structures, and if severe enough, they may preclude their use.

For a simple cantilever, the vertical deflection at a distance \( x \) from the support, \( \delta(x)/x \), is given by (Roark and Young (1989)).
\[
\frac{\delta(x)}{x} = K + \frac{(1 - v^2)}{2EI} Mx
\]

where,
- \( K \) = constant determined by boundary conditions at the support end
- \( E/(1 - v^2) \) = biaxial modulus of the film (compensating Young’s modulus for stiffening of the beam during stretching)
- \( I \) = moment of inertia of the cantilever about the z-axis, in \( m^4 \)
- \( M \) = internal bending moment defined above, in \( N\cdot m \)

Such deflections can be measured using optical and other microscopic means, can theoretically be used to estimate the value of \( M \).

Fan, et al. (1990) proposed the use of Archimedean spirals as test struct for measuring the average values of such gradients.

Illustration of a spiral test structure for stress gradients, showing (at right) various of the relaxed states of the spiral for positive and negative strain gradients (view as slices through the spiral along the plane indicated at left).

If the stress gradient is positive (increasingly tensile with increasing height above the substrate), the spiral will assume an open bowl shape. If the stress gradient is negative (increasingly compressive with increasing height above the substrate), the spiral will assume a dome shape. Note that to cover stress gradient of both polarities, test spirals with anchors to the substrate at the center and outer turn must be included so that their relaxed states are above the substrate. In the one can also extract information as to the magnitude of the strain gradient measuring the height of the free end of the spiral, its rotation, and contraction the diameter of the spiral (unfortunately, these parameters may be quite tedious measure). In practice, however, spiral structures are most convenient as indica
of the presence and sign of significant residual stress gradients. For further information on stress gradient effects on a variety of thin-film microstructures, the reader is referred to Howe (1994).

3.3.2 MEASUREMENT OF OTHER MECHANICAL PROPERTIES

DIRECT MEASUREMENT

As is commonly done for macroscopic specimens, key mechanical properties of materials (Young’s modulus, Poisson’s ratio, and tensile strength) can be measured directly by applying carefully controlled, uniform stress to a specimen and measuring the resulting dimensional changes. By measuring the resulting dimensional changes in the directions of, and perpendicular to, the applied stress, the Young’s modulus and Poisson’s ratio can be computed.

Sharpe, et al. (1996, 1997) reported on a relatively simple technique for measuring these parameters for thin-film specimens. Their approach is based on the fabrication of a released, doubly supported cantilever beam of the thin-film to be tested. Markers of different reflectivity from the thin-film under test are fabricated on the beam, and laser interferometry is used to measure their displacements relative to each other. As described in Sharpe, et al. (1996, 1997), test specimens were fabricated by depositing the test layer on a silicon wafer and back-side etching to form the released, doubly supported membrane. In this example, the authors used two layers of polysilicon, one 2.0 µm thick and the other 1.5 µm thick, both deposited by LPCVD and annealed at 1,050°C for 1 hour to reduce the grain size and stress. Prior to deposition of the polysilicon, the substrates were coated with silicon nitride and PSG (phosphosilicate glass) on both sides. To release the beams, the wafers were cut into individual dice, the back-side polysilicon and PSG were removed, the back-side silicon nitride was patterned, the top side of the wafer was protected by waxing it to a carrier, and the silicon bulk was etched using an unspecified wet etchant. With the bulk etching complete, the PSG beneath the polysilicon test film was removed using HF and the die was removed by dissolving the wax in an organic solvent. Key to practical handling of the final dice was the inclusion of support strips of bulk silicon that prevent forces from being applied to the test beam before it is mounted in the stress test assembly. Once a test die was mounted (with adhesive), the support strips were cut using a diamond saw.
Illustration (not to scale) of a bulk micromachined, doubly supported cantilever with removable support strips (at top and bottom of top view) for handling prior to stress testing. Adapted from Sharpe, et al. (1997).

The measurements were made using an air-bearing supported mechanism with a piezoelectric translator and a load cell. Movement of optical markers on the test specimen’s surface was detected by measuring the changes in fringe pattern from laser light diffracted from each edge of a marker on the polysilicon’s surface. It should be noted that although Sharpe, et al. (1997) reported only on the use of this direct method on polysilicon, it could, in principle, be applied to any thin-film material from which a doubly supported cantilever could be fabricated.

Sharpe, et al. (1997) also present an excellent summary of the mechanical properties of polysilicon, with average values that are in agreement with many previously reported results. The reader should be warned that for any such mea
surements, process-, doping-, and thermal-history-dependent variations must be considered when making comparisons. The results Sharpe, et al. (1997) presented were derived from polysilicon fabricated using the Microelectronics Center of North Carolina’s “MUMPS” polysilicon foundry process. Those values were Young’s modulus $= 169 \pm 6.15$ GPa, Poisson’s ratio $= 0.22 \pm 0.011$, and tensile strength $= 1.20 \pm 0.150$ GPa. It should be noted that their paper also summarizes similar results from 13 other references and brief comments on the methods used. An excellent discussion of the structural and mechanical properties of polysilicon deposited using various methods can be found in Krulevitch (1994).

INDIRECT MEASUREMENT

As alternatives to direct measurement of mechanical properties of thin films, several indirect methods are available. Several of these are reviewed in Sharpe, et al. (1997) and Ristic, et al. (1994). For example, if a cantilever can be electrostatically (easiest for conductive films) or photothermally driven (e.g., with a laser beam) with a sinusoidal waveform and its vibration amplitude is measured (typically using a deflected laser beam), a plot of amplitude versus frequency will show a characteristic resonance at a frequency $f_R$. From this information, the Young’s modulus of the film can be estimated using,

$$E = \frac{2\pi f_R^2 AL^4 \rho}{3.52I}$$

where,
- $A =$ cross-sectional area, in $m^2$
- $L =$ cantilever length, in $m$
- $\rho =$ film density, in $kg/m^3$
- $I =$ moment of inertia of the cantilever, in $m^4$

A doubly supported beam can also be used in a similar fashion, with the resonant frequency determined by the tension in the beam.

4. BASIC MECHANISMS AND STRUCTURES

As a building block for fabricating sensors or machines on a microscale, mechanisms are often necessary to couple energy between actuators and sensors and/or the outputs of a mechanical system. A useful definition of “mechanism” (from Mehregany, et al. (1988) is the “means for transmitting, controlling, or constraining relative movement.” It is possible to make mechanisms such as joints, linkages, gears, sliders, hinges, etc., using micromachining techniques.
of PDMS (or by covering them with tape prior to oxidation). Using externally powered swirling of the water bath, a variety of different shapes were self-assembled into ordered arrays. Yeh and Smith (1994a, 1994b) demonstrated the fluid-powered self-assembly of GaAs light-emitting diodes (LEDs) into etched holes in a silicon substrate.

As an alternative to such self-assembly techniques, the batch transfer of microcomponents from one substrate or assembly to another could be used to construct more complex micromachines. Such a technique was demonstrated by Cohn, et al. (1996) using the transfer of separately fabricated polysilicon structures to a single-crystal silicon substrate. In principle, this approach could be extended to realize multi-level transferred structures, but it does not appear to have the ability to replace the folding of hinged structures, for example, in terms of achievable object complexity. However, such batch-transfer approaches do not require any further interaction with the resulting microstructures.

5. MECHANICAL SENSORS

As mentioned above, there is a tremendous variety of direct mechanical sensors that have and could be micromachined. In order to study this vast area, it makes sense to begin with a discussion of the basic mechanical sensing mechanisms (e.g., for force, displacement, strain, etc.) and then see how these mechanisms can be applied to realize a wide range of micromachined sensors.

5.1 SENSING MECHANISMS

The following table summarizes the commonly used mechanical sensing mechanisms in micromachined devices. Important considerations when choosing such a mechanism include the need for local (or even monolithically integrated) circuitry, whether or not the transduction mechanism is DC-responding (piezoelectric sensors are the only common variety that are generally not), temperature coefficients, long-term drift, overall system complexity, and others.
externally powered in a silicon...

...transfer of be used to structures extended to have the achievable equire any...

mechanical fast area, it mechanisms inisms can

\[ \text{Mechanism} \quad \text{Parameter Sensed} \quad \text{Needs Local Circuits?} \quad \text{DC Response?} \quad \text{Complex System?} \quad \text{Linearity} \quad \text{Issues} \\
\begin{array}{|c|c|c|c|c|c|c|}
\hline
\text{Metal Strain Sensor} & \text{strain} & \text{NO} & \text{YES} & + & +++ & \text{• low sensitivity} \\
& & & & & & \text{• very simple} \\
\hline
\text{Piezoresistive Strain Sensor} & \text{strain} & \text{NO} & \text{YES} & + & +++ & \text{• temperature effects can be significant} \\
& & & & & & \text{• easy to integrate} \\
\hline
\text{Piezoelectric} & \text{force} & \text{NO} & \text{NO} & ++ & ++ & \text{• high sensitivity} \\
& & & & & & \text{• fabrication can be complex} \\
\hline
\text{Capacitive} & \text{displacement} & \text{YES} & \text{YES} & ++ & & \text{• very simple} \\
& & & & & & \text{• extremely low temperature coefficients} \\
\hline
\text{Tunneling} & \text{displacement} & \text{YES} & \text{YES} & +++ & poor & \text{• sensitive to surface states} \\
& & & & & & \text{• drift performance not yet proven} \\
\hline
\text{Optical} & \text{displacement} & \text{NO} & \text{YES} & +++ & +++ & \text{• rarely employed in mechanical microsensors} \\
\hline
\end{array} \\

Table comparing some of the major properties of the mechanical sensing mechanisms commonly used in micromachined devices.

5.1.1 RESISTIVE AND PIEZORESISTIVE STRAIN SENSORS

Strain sensors are an integral part of many micromachined devices, serving to measure strain or, indirectly, displacement of structures. A strain gauge is a conductor or semiconductor that is fabricated on or bonded directly to the surface to be measured. Changes in gauge dimensions result in proportional changes in resistance in the sensor. This is partly due to stretching (changes in dimension) and partly due to the piezoresistive effect, discovered by Lord Kelvin in 1856. As might be expected, the sensitivity of gauges can be quite different, depending on their design. Strain gauges of all types can be very linear over considerable ranges of strain, making them attractive in a variety of applications.

In general, the sensitivity is expressed by the gauge factor (dimensionless),

\[ \text{GF} = \frac{\Delta R}{\Delta L} \frac{R}{L} = \frac{\Delta R}{\varepsilon R} \]

(here longitudinal strain, \( \varepsilon_L \), is used). One can use partial derivatives to derive a general expression for the gauge factor in terms of the physical parameters of the
strain gauge. This begins with the derivation of a relation between resistance changes in its underlying parameters, as seen in,

\[ R = \frac{\rho L}{A} \text{ in } \Omega \]

where,
\( \rho \) = resistivity, in \( \Omega \cdot \text{cm} \)
\( L \) = length, in cm
\( A \) = cross-sectional area, in cm

Differentiating the resistance equation, one obtains,

\[ \frac{dR}{R} = \frac{dL}{L} + \frac{d\rho}{\rho} - \frac{\rho L}{A^2} \frac{dA}{A} \]

which can be divided by the above equation for resistance to obtain,

\[ \frac{dR}{R} = \frac{dL}{L} + \frac{d\rho}{\rho} - \frac{dA}{A} \]

It must be noted that from this point on in the derivation it becomes geometric specific, and a cylindrical wire is assumed for simplicity. It is useful to introduce Poisson’s ratio to express the relative dimensional change in diameter, \( D \), versus length, \( L \),

\[ \nu = -\frac{\varepsilon_L}{\varepsilon_D} = -\frac{\Delta D}{D} \approx -\frac{\Delta L}{L} \frac{dL}{L} \]

where, for a cylinder, the area and diameter are related through,

\[ A = \frac{\pi D^2}{4} \quad \text{and} \quad \frac{dA}{A} = \frac{2dD}{D} \]

which in turn allows Poisson’s ratio to be written and rearranged to obtain,

\[ \frac{dA}{A} = -2\nu \frac{dL}{L} \]

finally allowing the differential form of the resistance to be written,
\[ \frac{dR}{R} = (1 + 2\nu) \frac{dL}{L} + \frac{dp}{\rho} \]

where the first term represents the dimensional effect and the second term represents the piezoresistive effect (change in resistivity of the material of the strain gauge). From this, the gauge factor can be expressed in terms of these parameters as,

\[ GF = \frac{dR}{R} \frac{R}{dL} = \frac{R}{\varepsilon_i} (1 + 2\nu) + \frac{\rho}{\varepsilon_i} \]

Naturally, such a derivation can be carried out for strain gauges with non-cylindrical shapes.

As shown in the table below, the gauge factors of different types of strain gauges can be vastly different, due mainly to whether or not they have a significant piezoresistive effect (as do the semiconductor types).

<table>
<thead>
<tr>
<th>Type of Strain Gauge</th>
<th>Gauge Factor</th>
</tr>
</thead>
<tbody>
<tr>
<td>Metal Foil</td>
<td>1 to 5</td>
</tr>
<tr>
<td>Thin-Film Metal</td>
<td>(\approx 2)</td>
</tr>
<tr>
<td>Bar Semiconductor</td>
<td>80 to 150</td>
</tr>
<tr>
<td>Diffused Semiconductor</td>
<td>80 to 200</td>
</tr>
</tbody>
</table>

*Table comparing the gauge factors of different types of strain gauges.*

**METALLIC STRAIN GAUGES**

For metals, \(\rho\) does not vary significantly with strain (as long as the cross-sectional dimensions are much larger than the grain size), and \(\nu\) is typically in the range of 0.3 to 0.5, for gauge factors on the order of two. However, in practice, macroscopic metal strain gauges often have gauge factors higher than this, so it appears that some piezoresistive effect and/or change in total wire volume comes into play. Whether or not this would be helpful in micromachined strain gauges is most likely a moot point since the much larger gauge factors of piezoresistive strain gauges make them nearly ubiquitous in this domain.
Metal strain gauges may be made from thin wires or metal films (thin-film strain gauges) that may be directly fabricated on top of microstructures. Thin-film metal strain gauges are easier to fabricate (photolithographically) and allow for more complex shapes. They are generally built on flexible plastic substrates (sometimes self-adhesive) and can be glued onto a surface.

Illustration of typical metal strain gauge designs. Three strain gauges arranged as shown in the lower part of the illustration allow \( \sigma_x, \sigma_y, \) and \( \tau_{xy} \) to be resolved at a given location (this is a standard configuration). After Norton (1989).

**SEMICONDUCTOR STRAIN GAUGES**

In semiconductor strain gauges, the piezoresistive effect is very large, leading to much higher gauge factors. P-type silicon has a gauge factor up to 200 and n-type has a negative gauge factor, down to \( \approx -140 \). Strain gauges can be locally fabricated in bulk silicon through ion implantation or diffusion, or the entire substrate can be used as the sensor. Unfortunately, these semiconductor strain gauges also have much higher-temperature coefficients of resistivity, making temperature compensation more important. (For example, one can use a Wheatstone bridge with a reference strain gauge that is not deformed to compensate, as long as all bridge elements are isothermal.)
s (thin-film. Thin-film
d allow for
tates (some-

The detailed theory of piezoresistivity is not covered herein, but good expla-
nations can be found in Kanda (1982), Yamada et al. (1982), and Middelhoek and
Audet (1989). In short, the effective mobilities of majority carriers in a piezoresistive
material are affected by stress (this effect is highly orientation dependent). For
p-type materials, the effective mobility of holes decreases so the resistivity increases.
For n-type materials, the effective mobility of electrons increases so the resistivity
decreases. The observed change in mobility results from the strain-induced distortions
of the energy band structure, and this can be calculated quite accurately if necessary.

While the strong temperature dependence of the gauge factor for single-crystal
semiconductor strain gauges makes their use sometimes more difficult, polycrystall-
line and amorphous silicon are useful alternatives (which are not anisotropic). The
total resistance of polycrystalline silicon is determined by the resistance of the
silicon grains and that of the grain boundaries, the latter being the most important
aspect. Within the grains, the resistivity behaves essentially like that of the single-
crystal material, and so as the temperature increases, the mobility decreases and the
resistivity increases. At grain boundaries, depletion regions develop due to charge
trapping, and here as the temperature increases, more carriers can overcome these
boundaries, decreasing the resistivity. By balancing these effects (i.e., changing
the dose of ion implant doping), the net temperature coefficient can be adjusted to
nearly zero. It is worth pointing out that silicon (and polysilicon and amorphous
silicon) are centrosymmetric and not piezoelectric (unless stressed). Piezoresistive
behavior is completely different from piezoelectric properties (discussed below).

5.1.2 PIEZOJUNCTION EFFECT

A related effect to piezoresistance is the piezojunction effect, which is a
marked shift in the I-V characteristic of a p-n junction when mechanical stresses
are applied to it. One can also fabricate pressure sensitive tunnel diodes, MOSFETs,
MESFETs, etc. While this effect has been discussed in the micromachining literature,
little use has been made of it in micromachined transducers. Friedrich, et al.
(1997) presented a piezojunction-based strain sensor where reverse I-V characteris-
tics, dominated by band-to-band tunneling, were modulated by applied strain.

5.1.3 PIEZOELECTRIC EFFECT

Piezoelectricity is a phenomenon in which a mechanical stress on a material
produces an electrical polarization and, reciprocally, an applied electric field produces
a mechanical strain. The Curies discovered the effect in 1880, and the first useful
applications were made by Cady in 1921 with his work on quartz resonators (see
Cady (1964)). The effect can certainly be used to sense mechanical stress (or,
indirectly, displacement, etc.) and as an actuation mechanism. However, a key
potential limitation of this transduction mechanism is that the piezoelectric effect
produces a DC charge (polarization), but not a DC current. Thus such transducers
are inherently incapable of providing a DC response. The limited low-frequency response of piezoelectric devices is primarily due to parasitic charge leakage paths and can be significantly improved through micromachining and directly coupling piezoelectric outputs to MOSFET gates (see Chen, et al. (1982, 1984), discussed below in the accelerometer section).

Illustration showing the generation of incremental charge (and hence voltage) on metallized electrodes on opposite sides of a slab of piezoelectric material, in response to the application of force.

Centrosymmetric crystals such as silicon and germanium are not piezoelectric. If such materials are strained, the effective centers of the positive and negative charges do not move with respect to each other, preventing the formation of dipole as required for piezoelectricity. Thus, materials whose crystal structures lack center of symmetry are required for the piezoelectric effect to be possible. Thus, to fabricate silicon-based piezoelectric transducers, a suitable material must be deposited on the devices. (It is possible to make centrosymmetric crystals such as silicon exhibit a weak piezoelectric effect by applying a uniaxial stress, but the effect is too small for any practical applications.) The III-V and II-VI compounds (e.g., GaAs, CdS, ZnO, etc.) are not centrosymmetric and have bonds that are partly covalent and partly ionic in nature, and thus are piezoelectric. Materials such as CdS and ZnO can be deposited by co-evaporation or sputtering, with the sputtered ZnO being more common approach.

Piezoelectricity, pyroelectricity, and ferroelectricity all derive from one single physical cause: the existence of an electric polarization vector, \( \mathbf{P} \). As a rule of thumb, if a crystal is piezoelectric, almost always it will be pyroelectric and ferroelectric too (there are very few exceptions in some exotic materials). This points out another factor limiting the use of piezoelectric sensing for low frequencies and DC, since most suitable materials exhibit considerable temperature errors due to
their pyroelectric behavior. This effect can, however, be mitigated through the use of compensation capacitors made from the same piezoelectric material but left unstrained (Chen, et al. (1982, 1984)).

In terms of sensitivity to stress, piezoelectric materials are commonly characterized by the charge sensitivity coefficients, $d_{ij}$, (in units of C/N), which relates the amount of charge generated at the surfaces of the material (of area A) on the i axis to the applied force, F, on the j axis,

$$\Delta Q_i = d_{ij} \Delta F_j = d_{ij} \Delta \sigma A$$

From this, the voltage change across the conductive plates (at a spacing x) can be written as,

$$V = \frac{Q}{C} = \frac{Qx}{\varepsilon_0 \varepsilon_r A} \quad \Rightarrow \quad \Delta V_i = \frac{d_{ij} \Delta F_j x}{\varepsilon_0 \varepsilon_r A}$$

The piezoelectric effect is reversible, such that the application of a voltage $\Delta V$ gives rise to a corresponding force, $\Delta F$, and resulting dimensional change $\Delta L$. This is commonly used in piezoelectric actuators. Typical values for $\Delta L$ vary between $10^{-10}$ and $10^{-7}$ cm/V. Thus, to obtain displacements on the order of micrometers ($\mu$m), voltages exceeding 1,000 V are often necessary, unless stacked actuators or mechanical motion amplification methods are used.

<table>
<thead>
<tr>
<th>Material</th>
<th>Type</th>
<th>Piezoelectric Constant pC/N</th>
<th>Relative Permittivity ($\varepsilon_r$)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Quartz</td>
<td>single crystal</td>
<td>$d_{33} = 2.33$ [2, 3]</td>
<td>4.5 [2], 4.0 [3]</td>
</tr>
<tr>
<td>Polyvinylidene fluoride (PVDF)</td>
<td>polymer</td>
<td>$d_{31} = 20$, $d_{32} = 2$, $d_{33} = -30$ [2]</td>
<td>12 [1, 2]</td>
</tr>
<tr>
<td>Barium titanate (BaTiO$_3$)</td>
<td>ceramic (Perovskite crystal)</td>
<td>$d_{31} = 78$ [1, 2], $d_{33} = 190$ [3]</td>
<td>1,700 [1, 2], 4,100 [3]</td>
</tr>
<tr>
<td>Lead zirconate titanate (PZT)</td>
<td>ceramic</td>
<td>$d_{31} = 110$ [1, 2], $d_{33} = 370$ [3]</td>
<td>1,200 [1], 300 to 3,000 [3]</td>
</tr>
<tr>
<td>Zinc oxide (ZnO)</td>
<td>metal oxide</td>
<td>$d_{33} = 246$ [4]</td>
<td>1,400 [4]</td>
</tr>
</tbody>
</table>

Common piezoelectric materials with applications in micromachining include quartz (basis of quartz crystal oscillators and resonators); polyvinylidene fluoride (PVDF); lead zirconate titanate (PZT); perovskite crystals such as barium titanate lithium niobate, etc.; III-V compounds such as GaAs, GaP; and II-VI compounds such as ZnO, ZnS, ZnSe, etc. Relevant properties of several piezoelectric materials are given in the table above.

PZT is a ceramic with a high value for the piezoelectric strain constant. However, it is somewhat difficult to deposit as a thin film. Polyvinylidene fluoride (PVDF) and ZnO are most often used in the microfabrication of piezoelectric transducers.

PVDF, which is a carbon-based polymer, is usually deposited as a spin cast film from a dilute solution in which PVDF powder has been dissolved. As for most piezoelectric materials, processing after deposition greatly affects the behavior of the PVDF film. For example, heating and stretching can increase or decrease the piezoelectric effect. PVDF and most other piezoelectric films require a polarization after deposition (called poling). This is done by the application of a large electric field for a few hours using electrodes deposited on both sides of the film. The polarization (which is closely related to the piezoelectric strain constant) at the end of this process depends on the time integral of the applied field up to a saturation level. It is important to note the electro-spray PVDF deposition method of Asahi et al. (1993) discussed in the Optical Transducers chapter. Alternatively, Swartz and Plummer (1979) described the attachment of a prefabricated PVDF sheet to an array of MOSFETs using epoxy (in their application, ultrasound receivers, acoustic energy impinging upon the PVDF generated charges that were sensed by the MOSFETs). These approaches eliminate the need for a poling step, and may make PVDF a more attractive thin-film piezoelectric material than the others.

ZnO is the most common piezoelectric material used in microfabrication. Unlike PVDF, it can be sputter-deposited as a polycrystalline thin film with its c-axis (along which piezoelectricity is strongest) perpendicular to the surface of the substrate. Pure Zn is usually sputtered in an O₂/Ar plasma to form ZnO. ZnO has also found broad applications as a pyroelectric material.

The required high voltages make piezoelectric materials poor actuators for displacements in the micron regime, but they are precise actuators on the sub-nanometer scale (the reverse argument shows that small displacements will cause large detectable voltages making piezoelectric materials very sensitive sensors). Piezoelectric actuation is ideal, however, for scanning tunneling and scanning force microscopes, where small, precise displacements are required. Piezoelectric materials are also very useful in micromachined transducers such as surface acoustic wave (SAW) devices, accelerometers, microphones, etc.
5.1.4 CAPACITIVE SENSING

Perhaps one of the most important, and oldest, precision sensing mechanisms is capacitive (or electrostatic). The physical structures of capacitive displacement sensors are extremely simple (one or more fixed plates, with one or more moving plates). The inherent nonlinearity of most capacitive sensors is often overshadowed by their simplicity and very small temperature coefficients. With the monolithic integration of signal conditioning circuitry, the additional problem of measuring often miniscule capacitance changes in the face of large parasitics is mitigated. Several potential capacitive sensing modes are illustrated below.

Illustration of four different possible capacitive sensing modes. After Cobbold (1974). In case A, the distance between the plates is varied, which leads to an extremely nonlinear transfer function. In case B, an intermediate plate is moved relative to two fixed plates, providing a fixed total capacitance and lending itself to differential measurements. Cases C and D represent two modes (single-ended and differential, respectively) where the overlap area of the plates is varied with position, providing far greater linearity (impeded to some extent, nonetheless by fringing field effects). Another mode of operation, varying the dielectric constant between the plates (e.g., by moving a dielectric slab between them) is illustrated as case E.

The basic parallel-plate capacitor equation (assuming no fringing fields) is,

\[ C = \frac{\varepsilon_0 \varepsilon_r A}{d} \] in F
where,
\( \varepsilon_0 \) = dielectric constant of free space = \( 8.854188 \times 10^{-14} \) F/cm
\( \varepsilon_r \) = relative dielectric constant of the material between the plates
A = overlapping plate area, in cm
\( d \) = plate separation, in cm

Similarly, as is often the case in micromachined structures, for \( n \) dielectric layers of a relative dielectric constant, \( \varepsilon_{ir} \), the overall capacitance is,

\[
C = \frac{\varepsilon_0 A}{\frac{d_1}{\varepsilon_{r1}} + \frac{d_2}{\varepsilon_{r2}} + \ldots + \frac{d_n}{\varepsilon_{r\infty}}} \quad \text{in F}
\]

Capacitive sensor structures are relatively simple to fabricate. As illustrated above, one can vary \( d, \varepsilon, \) or \( A \), providing very nonlinear (in the former two cases or quite linear position-to-capacitance transfer functions (in the latter case). While macroscopic capacitive transducers of nearly any imaginable shape can be (largely have been) implemented, this is not the case for micromachined devices. Membrane-type capacitive devices (e.g., microphones or pressure sensors) straightforward to fabricate, but they are extremely nonlinear, since \( d \) varies. Capacitors are commonly used in surface micromachined devices, and theoretically based on varying the overlapping area for greater linearity. However, at such scales (particularly for surface micromachined devices), fringing fields become very significant or even dominant. Thus, the parallel-plate capacitor equation is only useful for first-order estimates at best. Varying the dielectric constant between the plates (e.g., by moving a slab of a different dielectric constant for the ambient between the plates) appears not to have been employed in micromachined devices often, and there seems to be little incentive to do so given the relative ease of using the other two modes. One noteworthy exception is the class of humidity and chemical sensors in which the dielectric constant of a sensitive layer is varied in relation to the concentration of analyte.

Despite these potentially difficult issues, a key redeeming feature of capacitive transducers is their near lack of a temperature coefficient (as long as the material of the gap has a low-temperature coefficient of its dielectric coefficient, e.g., as with vacuum do). According to Baxter (1997), the temperature coefficient of dielectric constant of air at 1 atm and \( 20^\circ \text{C} \) is \( \approx 2 \) ppm/°C for dry air and \( 7 \) ppm/°C for moist air. However, the change in dielectric coefficient of air versus pressure is more sizable at 100 ppm/atm. If the dielectric between the plates is a gas, a stable pressure (or vacuum), the dominant (and often minor) thermal effect on capacitance is due to differential thermal expansion of the structures themselves. An additional advantage of capacitive sensing is the fact that it is non-contact. A classic paper on capacitive transduction is Foldvari and Lion (1964), show
many basic configurations of mechanical to capacitive devices (although the circuits discussed, e.g., the diode twin-T network, are dated in specific implementation, they are still very relevant). Furthermore, a book entirely dedicated to capacitive sensing is Baxter (1997).

While capacitive transduction is inherently less noisy than resistive (with its attendant thermal, or Johnson noise), the noise performance of capacitive versus piezoresistive approaches is not always better, particularly since surface micromachining approaches often result in extremely small (femto or attofarad) capacitances. In such cases, the necessary interface electronics (and Brownian motion of the extremely low-mass capacitive structures) often supplies enough noise to eliminate any potential signal-to-noise ratio (SNR) advantage of capacitive sensing (one needs only examine current commercial piezoresistive versus capacitive accelerometers to verify this).

Changing capacitance can be measured using a number of well-known circuit techniques, such as 1) charge-sensitive amplifiers, 2) charge-redistribution techniques, 3) impedance measurements (measuring the impedance in a bridge or other configuration), 4) RC oscillators (making the unknown capacitance the time-constant determining capacitance in an oscillator and measuring the frequency), and 5) direct charge coupling (e.g., use the moving plate as the gate of a field-effect transistor (FET)). In general, these circuits can (and often must) be integrated with the capacitive sensors themselves, or at least positioned nearby to minimize the effects of parasitic capacitances. Examples of capacitive sensor interface circuits for micromachined sensors can be found in Park and Wise (1983), Smith, et al. (1986), Kung, et al. (1988), and Kung and Lee (1992).

Such electrostatic devices are also capable of being used as actuators, but are very nonlinear in this mode, as discussed below.

5.1.5 TUNNELING SENSING

Tunneling transduction, as discussed in the Optical Transducers chapter, is extremely sensitive due to the exponential relationship of tunneling current, I, to the tip/surface separation,

\[ I = I_0 e^{(-\beta \sqrt{z})} \]

where,
- \( I_0 \) = scaling factor, dependent on materials, tip shape, etc.
- \( \beta \) = conversion factor, typical value = 10.25 eV\(-1/2\)/nm
- \( \phi \) = tunnel barrier height in electronvolts (eV), typical value = 0.5 eV
- \( z \) = tip/surface separation in nanometers (nm), typical value = 1 nm
6. MECHANICAL ACTUATORS

6.1 ACTUATION MECHANISMS

By definition, mechanical actuators convert electrical (or other) energy to mechanical energy. Unfortunately, there is no “perfect” actuator technology, rather a series of trade-offs in terms of fabrication complexity, environment robustness, range of motion, available force, etc. The “ideal” actuator would have little power, a high mechanical efficiency, be robust to mechanical/environmental conditions, be capable of fast motion if necessary, have a high power-to-ratio, and have a linear proportionality between force/torque/speed, etc., control signal. In practice, not all of these issues would matter (e.g., one would require a slow, but strong actuator that is always shielded from temperature variation such as some types of muscle fibers).

There have recently been several interesting surveys of actuator technology and the reader is referred to Fujita and Gabriel (1991), Hollerbach, et al. (and Hunter and Lafontaine (1992) for useful comparisons and a multitude of literature references on the subject. Two general comparison tables based on the first third of these references are shown below.

<table>
<thead>
<tr>
<th>Type of Motor</th>
<th>Torque/Mass (N*m/kg)</th>
<th>Power/Mass (W/kg)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sarcos Dextrous Arm (electrohydraulic)</td>
<td>120</td>
<td>600</td>
</tr>
<tr>
<td>McGill/MIT Electromagnetic Motor</td>
<td>15</td>
<td>200</td>
</tr>
<tr>
<td>Polyacrylic Acid/Polyvinyl Alcohol Polymeric Actuator</td>
<td>17</td>
<td>6</td>
</tr>
<tr>
<td>NiTi Shape Memory Alloy</td>
<td>20</td>
<td>6</td>
</tr>
<tr>
<td>Human Biceps Muscle</td>
<td>20</td>
<td>50</td>
</tr>
<tr>
<td>Burleigh Piezoelectric Inchworm</td>
<td>3</td>
<td>0.1</td>
</tr>
</tbody>
</table>

*Table of general comparisons of (macroscopic) robotics motors and human skeletal muscle, in terms of torque/mass and power/mass ratios. After Hollerbach, (1991).*
<table>
<thead>
<tr>
<th>Type of Actuator</th>
<th>Stress (MPa)</th>
<th>Strain (%)</th>
<th>Strain Rate (Hz)</th>
<th>Power Density (W/kg)</th>
<th>Efficiency %</th>
</tr>
</thead>
<tbody>
<tr>
<td>Electrostatic (macroscopic composite)</td>
<td>0.04</td>
<td>&gt; 10</td>
<td>&gt; 1</td>
<td>&gt; 10</td>
<td>&gt; 20</td>
</tr>
<tr>
<td>Cardiac Muscle (human)</td>
<td>0.1</td>
<td>&gt; 40</td>
<td>4</td>
<td>&gt; 100</td>
<td>&gt; 35</td>
</tr>
<tr>
<td>Polymer (polyacrylic acid/polyvinyl alcohol)</td>
<td>0.3</td>
<td>&gt; 40</td>
<td>0.1</td>
<td>&gt; 5</td>
<td>30</td>
</tr>
<tr>
<td>Skeletal Muscle (human)</td>
<td>0.35</td>
<td>&gt; 40</td>
<td>5</td>
<td>&gt; 100</td>
<td>&gt; 35</td>
</tr>
<tr>
<td>Polymer (polyaniline)</td>
<td>180</td>
<td>&gt; 2</td>
<td>&gt; 1</td>
<td>&gt; 1,000</td>
<td>&gt; 30</td>
</tr>
<tr>
<td>Piezoelectric Polymer (PVDF)</td>
<td>3</td>
<td>0.1</td>
<td>&gt; 1</td>
<td>&gt; 100</td>
<td>&lt; 1</td>
</tr>
<tr>
<td>Piezoelectric Ceramic</td>
<td>35</td>
<td>0.09</td>
<td>&gt; 10</td>
<td>&gt; 1,000</td>
<td>&gt; 30</td>
</tr>
<tr>
<td>Magnetostrictive (Terfenol-D)</td>
<td>70</td>
<td>0.2</td>
<td>1</td>
<td>&gt; 1,000</td>
<td>&lt; 30</td>
</tr>
<tr>
<td>Shape Memory Alloy (NiTi bulk fiber)</td>
<td>&gt; 200</td>
<td>&gt; 5</td>
<td>3</td>
<td>&gt; 1,000</td>
<td>&gt; 3</td>
</tr>
</tbody>
</table>

Table of linear actuator materials. After Hunter and Lafontaine (1992). The authors noted that the values provided did not always represent optimal materials, as development is active in many of these categories and that the power needed for accessory systems, such as cooling, were not included in the calculations.

This section is organized by actuation mechanism, not by specific actuator function, since relatively few useful stand-alone actuators have been demonstrated. In general, the actuators form part of an overall functional unit, and these are described throughout the chapters in the appropriate domains of operation.

### 6.1.1 ELECTROSTATIC ACTUATION

The fundamental actuation principle behind electrostatic actuators is the attraction of two oppositely charged plates, as was quite familiar to Benjamin Franklin. In principle, despite the nonlinear force-to-voltage relationship in such actuators, they are very low power and simple to fabricate. They have been used extensively in micromachined devices, since it is relatively simple to fabricate closely spaced gaps with conductive “plates” on opposite sides (see Muller (1990)). In many cases, the delivered output power and efficiency are far less than those predicted by
theory (often due to fringing fields, surface leakage, etc.), but electrostatic actuators are still extremely important in many applications (particularly when the actuator merely needs to move itself, as opposed to other objects). An example of this was seen above in the case of the ADXL-50 micromachined accelerometer, where electrostatic actuation was used to keep a movable polysilicon proof mass centered between sensing capacitor plates.

In estimating the force generated by an electrostatic actuator, one can begin with Coulomb's law, which gives the force between two point charges,

$$F_{\text{elec}} = \frac{1}{4\pi\varepsilon_r\varepsilon_0} \frac{q_1q_2}{x^2}$$

where $q_1$ and $q_2$ are the two charges in coulombs and $x$ is the distance separating them. If there are more than two charges, it is necessary to determine the force between each charge pair and to superimpose these vectors to find the resultant force. For most realistic electrostatic actuators, this unfortunately becomes quite complex, although suitable finite element and other computational methods exist.

For first-order approximations in many simple geometries, one can sometimes start with a parallel-plate capacitor approximation, but for most actuator shapes (i.e., cantilevers), this only holds for very small angles. For a parallel-plate capacitor with plate area, $A$ (neglecting fringing fields), the energy stored at a given voltage $V$, is given by,

$$W = \frac{1}{2} CV^2 = \frac{1}{2} \frac{\varepsilon_r\varepsilon_0 AV^2}{x}$$

and the force between the plates is,

$$F = \frac{dW}{dx} = \frac{1}{2} \frac{\varepsilon_r\varepsilon_0 AV^2}{x^2}$$

(It is interesting to note that in all capacitors, the voltage applied tends to crush the inter-plate dielectric.) From this, it is clear that the force versus distance and force versus voltage relationships are nonlinear, although in some cases they can be linearized through closed-loop control.

**ELECTROSTATIC CANTILEVER ACTUATORS**

A useful analysis of the relationship of applied drive voltage and deflection in a micromachined cantilever beam (intended as an optical modulator in a display system, as discussed in the Optical Transducers chapter) can be found in Petersen (1978a). The basic structure of the actuator is illustrated below.
Illustration of an electrostatically deflected cantilever structure showing variable definitions for analysis. After Petersen (1978a).

Following Petersen's derivation, the goal is to obtain a relationship between applied drive voltage and tip deflection. From the mechanical engineering theory, it is known that a concentrated load at a position, \( x \), from the fixed end of a cantilever beam of width, \( w \), results in a tip deflection, \( \delta_T \), given by,

\[
(d\delta)_T = \frac{x^2}{6EI} (3L-x)wq(x) \, dx
\]

where the electrostatic force, \( q(x) \), at a distance \( x \), is,

\[
q(x) = \frac{\varepsilon_0}{2} \left( \frac{V}{d-d(x)} \right)^2
\]

and where,
- \( E \) = Young's modulus of the cantilever
- \( I \) = moment of inertia of the cantilever
- \( L \) = beam length
- \( x \) = distance of force (load) from the fixed end of the beam
- \( d \) = gap between cantilever and deflection electrode

The total tip deflection can be found by integrating the above equation from the fixed end to the tip of the beam (\( x = L \)),

\[
\delta_T = w \int_0^L \frac{(3L-x)}{6EI} x^2 q(x) \, dx
\]
To make the solution of the integral possible, one can assume a square-like curvature of the beam at any point along its length,

\[ \delta(x) = \left( \frac{x}{L} \right)^2 \delta_t \]

This in turn yields a normalized load, \( F \), required to produce a specified tip deflection

\[ F = \frac{\varepsilon_v w L^4 V^2}{2 E I d^3} = 4 \Delta^2 \left( \frac{2}{3(1 - \Delta)} \frac{\tanh^{-1} \sqrt{\Delta}}{\sqrt{\Delta}} - \frac{\ln(1 - \Delta)}{3 \Delta} \right)^{-1} \]

where \( \Delta \) is the deflection at the tip (\( \delta_t/d \)).

This equation for normalized load is plotted below versus normalized deflection, and shows that the relationship between tip deflection and applied voltage is indeed extremely nonlinear, and that once deflection exceeds a threshold voltage, the position of the tip is unstable, and the beam spontaneously deflects all the way down.

Plot showing the normalized load, \( F \), versus the normalized deflection, \( \Delta \), for a electrostatically deflected cantilever. Adapted from Petersen (1978a). Beyond certain threshold deflection, spontaneous collapse of the cantilever occurs.
The threshold voltage is approximately given by,

\[ V_{th} = \frac{\sqrt{18EI/d^3}}{\sqrt{5\varepsilon_oL^4w}} \]

and appeared to be fairly accurate in correlating with Petersen’s experimental results. Petersen also noted that the effective Young’s modulus/moment of inertial product for such a composite structure (his were a layer of metal above an insulating layer of silicon dioxide) could be approximated to first order (and used for the EI product in the above equations) by,

\[ (EI)_{eff} \approx \left( \frac{wt_1^3}{12} \right) \left( \frac{4 + 6 \frac{t_2}{t_1} + \frac{E_1t_1}{E_2t_2}}{1 + \frac{E_1t_1}{E_2t_2}} \right) \]

where, \( t_1, t_2, E_1, \) and \( E_2 \) are the thicknesses and Young’s moduli, respectively, of the lower (1) and upper (2) layers of the beam. This illustrates a point made above: most micromachined structures need to be treated as composite materials. Further, for such a cantilever beam structure, the first resonant frequency (bending) can be estimated using,

\[ f_{RI} = \frac{3.52}{2\pi} \frac{\sqrt{EI}}{\sqrt{wL^4}} \]

As can be seen from this reasonably tractable example (keeping in mind that fringing fields, which can be quite important at these aspect ratios, were ignored), the analyses of electrostatic structures can be rather involved. Analysis of more complex structures, such as the torsional or comb-drive electrostatic actuators discussed below, is often done using numerical approaches.

**TORSIONAL ELECTROSTATIC ACTUATORS**

Torsional electrostatic actuators have, in some cases, advantages over cantilever designs, particularly since they are (if supplied with dual deflection electrodes) able to deflect in two directions, rather than only one. As for the cantilever designs, they can be fabricated using a number of processes, including polysilicon with sacrificial oxide, electroplated metal with sacrificial organic layer, sputtered aluminum with sacrificial organic layer, etc.

As well as the Texas Instruments torsional electrostatic actuators discussed in Hornbeck (1995), a similar all-aluminum electrostatic actuator technology was
ELECTROSTATIC RELAYS AND SWITCHES

Applications of micromachining to the fabrication of relays and RF switches are discussed in detail below in Section 7.2. Electrostatic actuation has been one of the more successful means for driving these devices due to its extremely low power requirements, although most current devices still require relatively large voltages.

6.1.2 THERMAL ACTUATION

As discussed in the Thermal Transducers chapter, there are a number of possible thermal actuation means based on the expansion of solids or fluids. A large number of them have been exploited in micromachined devices.

THERMAL EXPANSION OF SOLIDS

Thermal expansion of materials can be readily applied to the actuation of microstructures. An interesting example is the thermal-expansion-driven gripper design demonstrated by Keller and Howe (1997). The gripper was fabricated using the HEXsil silicon mold replication process (described in the Micromachining Techniques chapter). The $8 \times 1.5$ mm $\times 40$ µm, normally closed gripper could be opened 35 µm with a 75 mW power input, and silicon dioxide pegs as small as $1 \times 4 \times 40$ µm could be manipulated and placed in $4 \times 4$ µm holes. In addition, an example of the use of single-crystal silicon beams as thermal expansion actuators can be found in Klaassen, et al. (1995). As for all thermal expansion driven actuators, careful design to maximize the generated strain and the thermal isolation of the expanding region are key to achieving realistic operating power levels.

BIMORPH THERMAL ACTUATORS

One of the basic thermal actuation schemes is to use the difference in thermal coefficients of expansion of two bonded materials, referred to as thermal bimorph actuation. A heater is typically sandwiched between the two “active” materials and, when electrically driven, causes them to expand differentially. Advantages of this approach include nearly linear deflection-versus-power relationships and environmental ruggedness (e.g., these actuators can be run in liquids of sufficiently low thermal conductivities). Disadvantages include high power, low bandwidth (determined by thermal time constants), and more complex construction than simple electrostatic actuators. Despite their disadvantages, these actuators have been used extensively. The basic principles of thermal bimorph actuators (differential thermal expansion) are discussed in the Thermal Transducers chapter. In addition, the underlying theory is discussed in a classic paper, Timoshenko (1925). It should be noted that micromachined bimorphs can be constructed from a wide variety of materials, both organic and inorganic.
Cross-sectional illustration of a thermal bimorph actuator where electroplated gold and a layer of epitaxial silicon form the main bimorph structure (with polysilicon and silicon nitride layers interspersed). After Riehmüller and Benecke (1988). (Not to scale.)


Reinhmuller and Benecke (1988) demonstrated a thermal bimorph actuator based on the differential thermal expansion of gold and silicon. Their actuators used up to 200 mW per actuator to achieve deflections up to 100 μm.
fabrication process used to realize these devices is illustrated below. They began
with the epitaxial growth of a 4 μm thick p+ silicon layer (1.3 × 10^{20} cm^{-3}) on
p-type (100) wafers. A thin LPCVD silicon nitride electrical insulating layer was
then deposited, followed by the deposition of an 0.5 μm polysilicon layer and its
doping via ion implantation. The polysilicon was then patterned using dry etching
to form the heaters, followed by the deposition of another LPCVD silicon nitride
layer, which was patterned to form contact vias and to define the geometry of the
actuators. A Cr seed layer was deposited, followed by the electroplating of a 1.8
μm gold layer to serve as the top layer of the bimorph and the electrical interconnects.
Finally, the Cr/Au layer was ion-milled, and the silicon etched in EDP to undercut
the bimorph cantilevers.

Yang and Kim (1995) demonstrated a multilayer thermal cantilever actuator
(actually a trilayer structure rather than a bimorph) that made use of two independent
polysilicon heater layers with a central silicon dioxide layer. Such an actuator
could be deflected up or down, based on which of the polysilicon layers was heated
by passing a current through it. Two of these actuators were combined with a
tension band of silicon nitride to form a bistable cantilever structure that could be
changed from an “up” to a “down” state with thermal drive, but with no holding
power. Electrical heating pulse times as low as 2 μs were used, for a maximum
mechanical switching frequency greater than 200 Hz.

THERMAL ARRAY ACTUATORS

Ataka, et al. (1993a, 1993b) presented an interesting paper demonstrating the
use of an array of thermal bimorph actuators to move objects by ciliary motion.
An array of 256 (total) actuators (128 facing each direction) were actuated with a
total average power of 1 W. A 2.6 × 1.5 × 0.26 mm piece of silicon weighing 2.4
mg was moved using this array. This early ciliary actuator array was limited such
that objects could only be moved along a line. Fujita, et al. (1996) described such
thermal actuators, as well as pneumatic devices (discussed below) and the strategies
required to operate them to obtain useful macroscopic motions of carried objects.

Suh, et al. (1996) demonstrated omnidirectional ciliary actuators using a
similar thermal bimorph actuation scheme, coupled with electrostatic actuators.
The actuators, as illustrated below, were fabricated using two layers of different
polyimides, Hitachi PIQ-L200 and PIQ-3200 (with coefficients of linear thermal
expansion of 2.0 and 54 ppm/°C, respectively). PECVD silicon nitride stiffening
layers, TiW heating resistors, and aluminum electrostatic plates were embedded
between the two polyimide layers. The devices were fabricated atop a sacrificial
aluminum layer that, in other regions where it was protected with a PECVD silicon
nitride layer removed after release, also served as the bond-pad and interconnect
layer. When the actuators were released using a standard aluminum wet etchant,
they deflected upward from the substrate (since they were cured at a higher temper-
ature, and the upper polyimide layer had the larger thermal expansion coefficient), aiding the release process.

The actuators, in groups of four (1 × 1 mm in size for each set of four) to allow four different basic motion directions, were fabricated in an 8 × 8 element array, for a total of 256 actuators. Each actuator element dissipated 16.7 mW for a 5 V drive signal, with a theoretical lifting capacity of 76 μN. Through proper sequencing, small objects could be moved across the array in arbitrary directions. Objects tested included a 3 × 3 mm silicon chip (8.6 mg), a #4-40 stainless steel nut (0.16 g), and an 8-pin plastic dual in-line packaged IC (0.5 g). Step sizes were as small as ≈ 3 μm. Electrostatic hold-down of the actuators after thermal actuation was also demonstrated with voltages as low as 100 V (but 500 V was required for purely electrostatic pull-in). Demonstrations of omnidirectional, vectored motions of components, as well as strategies for operation of such arrays, were presented by Konishi and Fujita (1995) and Bohringer, et al. (1997).

Illustration of one-quarter of an omnidirectional, thermal/electrostatic ciliary actuator. An embedded TiW heating resistor drives the device in thermal mode, and an aluminum electrostatic plate in its tip can be used to hold it down via electrostatic force relative to the substrate. Courtesy J. W. Suh, from Suh, et al. (1996).

Suh, et al. (1995) also demonstrated multi-segmented thermal bimorph actuators fabricated in the same manner. By thermally isolating regions of a bimorph via removal of selected regions of material, and using one TiW heater per segment, it was shown that individual segments could be actuated with only moderate cross-talk. The use of the resistance of the TiW heaters to provide temperature (and, indirectly,
position) feedback during heating was demonstrated (the TCR of the TiW heaters was measured to be 361 ppm/°C). Using analog computation circuits, the resistance of a given heater could be determined dynamically during heating. This approach could not only serve to provide the aforementioned position feedback, but also to potentially indicate the additional heat loss from an actuator contacting an object.

**DIELECTRIC LOSS HEATING OF THERMAL BIMORPHS**

Most dielectrics, when exposed to RF energy, exhibit losses that convert some of that energy into heat. This can serve as an indirect way of heating thermal bimorph structures. Dielectric losses are proportional to the square of the applied RF electric field (directly proportional to applied RF power). The thermal energy generated is given by,

$$P_{\text{thermal}} = \pi f \varepsilon_r \varepsilon_0 |E_0|^2 \text{ in W/m}^3$$

where,
- $f$ = excitation frequency, in Hz
- $\varepsilon_r$ = imaginary part of relative permittivity (may be a function of frequency)
- $E_0$ = electric field, in V/m

Rashidian and Allen (1993) demonstrated such thermal bimorph actuators based on a layer of the copolymer of vinylidene fluoride and trifluoroethylene (PVDF-TrFE, which is also a piezoelectric material) on top of a layer of polyimide. This combination of materials exhibits one of the largest differences in thermal expansion coefficient reported for a micromachined structure (PVDF-TrFE has an $\alpha_\text{T}$ of $140 \times 10^{-6}$°C$^{-1}$ and Dupont polyimide PI2611D has an $\alpha_\text{T}$ of $3 \times 10^{-6}$°C$^{-1}$). As illustrated below, they fabricated thermal bimorphs with a lower layer of polyimide and a layer of PVDF-TrFE sandwiched between two aluminum electrodes across which RF power could be applied.

RF power (1 to 100 MHz, 0 to 16 V peak) was applied and large deflections (250 to 900 μm) were achieved over millimeter-scale (length) cantilevers. The authors reported relatively large deflections using RF power inputs on the order of 1 mW (as compared to hundreds of milliwatts for directly heated devices). However, since most previously published thermal bimorph actuators used materials with much lower expansion coefficient mismatches, a fair comparison would have to take that into account. These devices could be remotely actuated by RF power and have been shown to work under water (using a wire-wound inductor connected to the beam drive electrodes and a second one nearby to couple power in, similar to a transformer with physically separated primary and secondary windings).
Dielectric loss heating of thermal bimorphs could turn out to be a very effective way of generating mechanical force in micromachined structures, particularly when they are relatively isolated from each other.

**VOLUME EXPANSION AND PHASE-CHANGE ACTUATORS**

Rather than making use of linear expansion of solids, it is possible to construct micromachined actuators that take advantage of volume expansion. A typical approach is to form a cavity with a sealed fluid (e.g., air, water vapor, liquid, etc.) that can be heated and thus expanded. If part of the cavity (i.e., one wall) is compliant, it will deform under pressure and generate mechanical force. This approach can generate large deflections and forces, but, like most thermal actuator schemes, uses a great deal of power and has low bandwidth due to thermal lag and constant. In addition, the formation of sealed, fluid-containing chambers can complicate a fabrication process. Most of these actuators are more properly described as phase-change devices, discussed below.

Phase-change thermal actuators involve thermally changing the phase of a substance to expand its volume and to create pressure (and hence mechanical force). For example, one can change liquids such as water from the liquid to the vapor phase by heating, generating bubbles that can be used to power mechanisms. Zdeblick et al. (1991) presented some such “bubble-powered” devices.

As discussed in the Microfluidic Devices chapter, an example of a thermal actuator is the so-called “fluistor,” which is commercially available from Redwood Microsystems (Menlo Park, CA) and is described in Zdeblick (1994). This device makes use of a resistive heater to cause a trapped fluid...
Week 15

Nanofabrication
Review

Yong Chen
Anne Pépin

Laboratoire de Microstructures
et de Microélectronique (L2M),
CNRS
Bagneux, France

Nanofabrication: Conventional and nonconventional methods

Nanofabrication is playing an ever increasing role in science and technology on the nanometer scale and will soon allow us to build systems of the same complexity as found in nature. Conventional methods that emerged from microelectronics are now used for the fabrication of structures for integrated circuits, microelectro-mechanical systems, microoptics and microanalytical devices. Nonconventional or alternative approaches have changed the way we pattern very fine structures and have brought about a new appreciation of simple and low-cost techniques. We present an overview of some of these methods, paying particular attention to those which enable large-scale production of lithographic patterns. We preface the review with a brief primer on lithography and pattern transfer concepts. After reviewing the various patterning techniques, we discuss some recent application issues in the fields of microelectronics, optoelectronics, magnetism as well as in biology and biochemistry.

Keywords: Nanofabrication / Lithography / Biosystems / Review

Contents

1 Introduction ........................................... 187
2 Preliminary concepts ................................. 188
3 Conventional lithographic methods .................. 189
3.1 Electron beam lithography ......................... 189
3.2 Focused ion beam lithography ..................... 190
3.3 Optical projection lithography ..................... 191
3.4 Extreme UV lithography ........................... 191
3.5 X-ray lithography ................................ 192
3.6 Electron and ion projection lithography .......... 192
4 Nonconventional methods ........................... 193
4.1 Nanoimprint lithography .......................... 193
4.2 Soft-lithography .................................. 194
4.3 Near-field optical lithography .................... 195
4.4 Proximity probe lithography ...................... 195
4.5 Other lithographic methods ....................... 196
4.6 Chemical and biological approaches ............. 196
5 Other nanofabrication issues ....................... 197
5.1 Thin-film deposition .............................. 197
5.2 Wet and dry etching techniques .................. 197
5.3 Micromachining .................................. 198
6 Applications in information technology .......... 199
6.1 Microelectronics ................................ 199
6.2 Optoelectronics and optics ...................... 199
6.3 Magnetism ....................................... 200
7 Applications in biology .............................. 200
7.1 Microfluidics and analysis systems .............. 200
7.2 Microarrays and gene chips ..................... 201
7.3 Nano-MEMS ...................................... 201
7.4 Nanobiotechnology ............................... 201
7.5 Nanomaterials for biological applications ...... 202
7.6 Bioelectronics .................................. 202
8 Conclusions ....................................... 202
9 References ........................................ 202

1 Introduction

Nanofabrication commonly refers to an ensemble of technologies used to fabricate very small structures (at scales ranging from 1 to 100 nm) which can be integrated into complex hierarchic systems. Sophisticated lithographic methods have origins in the microelectronics world [1]. For several decades, the semiconductor industry has been able to maintain a continuous increase of the integration rate with the number of transistors on a chip being doubled every 18 months [2]. Today, it is possible to integrate more than 100 million transistors on one square inch of a silicon wafer, which corresponds to a miniaturi-
zation level of making 100 nm structures and integrating them into complex circuits. To extend the current CMOS (complementary metal oxide semiconductor field effect transistor) technology to the generation of 30 nm and below, new lithographic techniques are under extensive investigation [2, 3]. Beyond this limit, quantum effects will most likely be used, and ultimately, electronic circuits working at molecular levels will be possible [3].

Nanofabrication is also a technological driver for other research and industry segments. Modern telecommunication systems, for example, employ enormous optoelectronic devices based on semiconductor processing. Even though these devices rely on similar technologies, they do not follow the integrated-circuit paradigm of size downscaling. Chemical and biological chips are other examples of areas which are in rapid progress toward a high degree of functionality and integration [4–7]. The challenges of applying nanofabrication technology to these and other new areas are formidable from the point of view of both basic and applied researches.

The standard fabrication process used by the semiconductor industry involves electron beam lithography for the formation of designed patterns on a set of masks and next optical projection lithography for the reproduction of the mask patterns at a high throughput level, combined to various material pattern transfer techniques. For applications in biology, new technological ingredients must be added to meet requirements such as size, power and bio-compatibility. The design and engineering of micro- and nanostructures are specific to a given application, and we will concentrate in this review article on more generic issues such as pattern formation. We provide an outlook on various lithography methods, including both conventional and nonconventional techniques as well as nonlithographic pattern formation. Pattern transfer techniques are briefly discussed, together with a number of microsystem issues. Finally, application examples and current trends in information technology and biosystems are presented.

2 Preliminary concepts

A typical lithographic process consists of three successive steps: (i) coating a substrate with irradiation-sensitive polymer layer (resist), (ii) exposing the resist with light, electron or ion beams, (iii) developing the resist image with a suitable chemical. Exposures can be done by either scanning a focused beam pixel by pixel from a designed pattern, or exposing through a mask for parallel replication. Serial beam scanning is used for mask fabrication and single component fabrication, but it does not supply adequate throughput for manufacturing. Industrial tech-

iques must be fast, reliable and cost-effective. Optical projection lithography with deep UV light is now used for large-scale fabrication. As next generation lithographies, several nonoptical techniques have been developed, including extreme UV lithography, X-ray lithography and projection lithography with either electrons or ions [2, 8, 9]. These methods, generally referred to as top-down methods, are very expensive but insufficiently flexible for new development out of the traditional microelectronic industry. In contrast, a number of new and nonconventional lithographic techniques such as nanoimprint, soft-lithography, near-field optical lithography and a number of proximity probe techniques are low cost and much more accessible [3]. In particular, these techniques not only conserve the lithographic printing strategy but also allow pattern processing at molecular levels. More generally, the so-called bottom up approaches involve self-assembly and self-organization in which molecular building blocks will automatically link together to form desired nanostructures. Much effort is currently being devoted to these new and exciting areas because they are very promising for the fabrication of future electronic devices [10, 11].

Conventional lithographic exposures involve the interaction of an incident beam with a solid substrate. Absorption of light or inelastic scattering of particles can affect the chemical structure of the resist by changing its solubility. The response of the resist to image-wise exposure can be either positive or negative, depending on whether the exposed or unexposed portions will be removed from the substrate after development. The next step after lithography is the pattern transfer from the resist to the substrate (Fig. 1). There are a number of pattern transfer techniques: selective growth of materials in the trenches of the resist, etching of the unprotected areas, and doping through the open areas of the resist by diffusion or implantation [12]. Both wet chemical etching and dry plasma etching can be used. For a high-resolution pattern transfer, dry etching is more suitable, and often requires a

![Figure 1. Schematic representation of lithography and pattern transfer techniques.](image-url)
metallic layer as a mask. This metal mask is obtained by lift-off, i.e., by first depositing a thin metallic layer over the developed resist pattern and then dissolving the resist in order to leave only the metal portions that are directly in contact with the substrate [1].

For multilevel fabrication, lithography and pattern transfer are used for each level. To increase productivity, optical projection lithography is performed on a large wafer in a step-and-repeat fashion by accurately controlling the sample stage with laser interferometry. The performance of lithography for production thus depends on a number of factors: resolution, throughput, pattern placement, overlay alignment accuracy, etc. However, the control of critical dimensions and, in turn, the microscopic properties of individual nanostructures, are the most important issues. Fabrication of semiconductor integrated circuits is a well-established technology. More specialized topics on nanofabrication are collected in recent books [13–15] and some dedicated issues [16, 17]. Discussion on general concepts of nanotechnology and molecular engineering can be found in [18–20]. Moreover, a number of conferences and workshops are dedicated every year to lithography and nanofabrication technologies [21–23].

3 Conventional lithographic methods

3.1 Electron beam lithography

Electron beam (e-beam) lithography is used for primary patterning directly from a computer-designed pattern [1, 24]. It is the essential basis for nanofabrication in addition to mask and prototype device manufacturing. The scanning technique is, however, not suitable for mass production because of the limited writing speed. A typical e-beam system consists of a column of electron optics for forming and controlling the electron beam, a sample stage, and control electronics. The column includes essentially an electron source, magnetic lenses, a beam blanker and a mechanism for deflecting the beam [24]. The electron source can be either a thermoionic emitter or thermal field emission. Depending on the design, e-beam energy varies in the range of 1–200 keV with a spot size down to a few nanometers. The beam current and the scanning field size are determined according to the experimental requirements. The sample position is precisely controlled with a laser interferometer feedback and thus a large area pattern can be exposed by decomposing the whole pattern into a number of scanning fields.

The resolution of e-beam lithography depends on the beam size and several factors related to the electron-solid interaction [24, 25]. In a resist, electrons undergo small angle forward scattering and some back-scattering events coming from the substrate. The forward scattering tends to broaden the initial beam diameter, whereas the back-scattering can spread over a large volume (proximity effects). During this process, the electrons are continuously slowing down, providing a cascade of low-energy electrons (secondary electrons) which are responsible for the resist reaction. Finally, the dose distribution looks like a sharp peak of the forward contribution superimposed with a back-scattered fog. The secondary electron-induced resist reaction can be polymerization, polymer cross-linking or chain scission as well as more complex processes such as chemical amplification involving acid-base reaction. The proximity effects can be corrected by dose variation and/or sophisticated software. In principle, low energy electrons (~1 keV) yield very limited proximity effects because of their short penetration depths. Correspondingly, thin resist layers or surface imaging techniques have to be used. Another advantage of using low energy electrons is the possibility of high-speed writing, because of the increased resist sensitivity.

The resolution of electron beam lithography is essentially defined by the e-beam spot size and the forward-scattering range of the electrons. The higher the energies, the smaller the minimum feature size obtainable. Commercial systems generally work at 50–100 keV energy, and are routinely used to produce sub-100 nm features [26]. Smaller feature sizes can be obtained with even higher electron energies [27]. Figure 2 shows scanning electron micrographs (SEM) of patterns defined by (a) 50 keV and (b) 200 keV electrons, followed by lift-off and subsequent reactive ion etching of SiO₂. Figure 3 shows two other examples of patterns equally defined by (a) 50 keV and (b) 200 keV electrons, but followed by reactive ion etching of tungsten (50 nm linewidth) or electroplating of gold (30 nm linewidth), respectively. These features were fabricated as nanoimprint molds (Fig. 2) and X-ray masks (Fig. 3) for one-by-one pattern replications (see below). Under optimal conditions, 10 nm [28] and sub-10 nm features [29, 30] can be obtained but particular attention has to be paid to the resist development and the pattern transfer processes (Fig. 4). Moreover, it is not possible to produce continuous metal lines at sub-10 nm scale by lift-off due to grain formation [30].

Electron beam lithography has also been used for hole drilling [31], contamination-induced growth [32], surface modification of inorganic materials (SiO₂, AlF₃, etc.) as well as Langmuir-Blodgett or self-assembled films [33, 34]. Finally, miniaturized electron beam systems have been prepared by microfabrication. Here, electrostatic lenses constructed from a series of parallel planar electrodes, each with a precision circular bore in the center, are used for the electron beam manipulation. The system
Figure 2. SEM of imprint molds obtained by e-beam lithography and reactive ion etching of SiO₂; minimum pitch size in (a) 120 nm; (b) 40 nm, respectively.

also includes an electron source consisting of a cathode with one or more electrodes to extract and accelerate the emitted electrons to the desired energy (1 keV), an objective lens, and a deflector unit for beam scanning. Large arrays of such microcolumns were fabricated, showing a parallel writing feasibility [35–37].

3.2 Focused ion beam lithography

Focused ion beam lithography is analogous to e-beam lithography [38, 39], but here magnetic lenses are replaced by electrostatic lenses because of the much heavier ion masses. The ion source is typically made of a sharp W-tip coated with liquid metal, Ga for example. In addition, a mass separator is used as an aperture to select one type of ion. Focused ion beams operate in the range of 10–200 keV. As the ions penetrate the material, they lose their energy at a rate several orders of magni-

Figure 3. Absorber features for X-ray lithography. (a) Reactive ion etched tungsten (50 nm linewidth); (b) electroplated gold (50 nm linewidth).

Figure 4. Typical Coulomb blockade device made by liftoff of a monogranular line between two metallic electrodes [30].

tude higher than that of electrons because of their masses. As a consequence, the penetration of ions, and thus the back-scattering-caused proximity effects, are much reduced compared to that of electrons. The heavier the ions, the smaller the penetration depth. Very thin layer resists are generally required for heavy ions. In practice,
light ions such as H⁺, He⁺⁺, Li⁺ and Be⁺⁺ are employed for lithography. The drawback of the focused ion beam lithography is its limited writing speed, due to the fact that the ion beam current density is 1 or 2 orders of magnitude smaller than that of the electron beams, although resists are generally more sensitive to ions.

On the other hand, focused ion beams have proven to be an essential tool for highly localized implantation doping, mixing, micromachining, controlled damage as well as ion-induced deposition [38]. Figure 5 displays two examples of ion beam micromachined structures on (a) GaAs and (b) a multilayer sample of 50 nm thick AlF₃ on GaAs [40, 41]. A large number of experiments have been performed to demonstrate the viability of the technique, including optical and X-ray mask repairing, quantum devices fabrication, scanning probe tip modification and, more recently, biomolecule microsurgery and analysis [42].

### 3.3 Optical projection lithography

Optical projection lithography is used for mass production of integrated circuits [43, 44]. Deep UV light is applied to expose a photosensitive polymer or its precursor on a glass substrate. Current projection systems used by the semiconductor industry work with a glass lens system with a demagnification factor of 5 or 6 and a step-and-repeat exposure capability over 8 inch silicon wafers. The resolution of optical projection lithography is defined by the Rayleigh criterion $R = k_1 \lambda / NA$, where $\lambda$ is the wavelength, $NA$ the numerical aperture of the optical system, and $k_1$ an empirical factor depending on the details of the experimental conditions. In practice, the image contrast goes to zero rapidly for $k_1 < 0.5$ [1, 44]. The depth of focus, $DOF = k_2 \lambda / NA^2$, is an important parameter to define the process latitude in optical lithography. A large $NA$ provides a better resolution but the corresponding $DOF$ decreases rapidly as $NA$ increases. Sophisticated masks (e.g., phase-shifting, optical proximity correction), off-axis illumination and top surface image techniques are used but reducing the exposure wavelength appears to be the main issue for the improvement of the resolution [2, 43, 44]. The previous projection systems worked with mercury-rare gas discharge lamps with radiation between 350- and 450 nm. Now, the production systems employ the 248 or 193 nm radiation produced by krypton fluoride (KrF) or argon fluoride (ArF) excimer lasers, respectively. The research efforts are being shifted to the 157 nm radiation produced by F₂ laser [45]. For this short wavelength, the quartz optics now in use are no longer transparent so that new optical materials such as calcium fluoride (CaF₂) have to be used for the lens and mask fabrication [45].

**Figure 5.** (a) Micromachined GaAs membrane and (b) fine lines on a multilayer AlF₃ (50 nm thick)/GaAs, obtained by direct focused ion beam lithography [40, 41].

Other issues including mask fabrication and repair, sensitive high resolution resists, overlay, etc., are also under investigation.

### 3.4 Extreme UV lithography

Extreme UV lithography refers to the exposure technique developed with 13.4 nm radiation and a reflective reduction system [46]. For this selective wavelength, the radiation is obtained from laser-induced plasmas or synchrotron radiation. For instance, it can be produced with a supersonic jet of xenon gas as the target, converting a few percent of laser energy into required radiation. The projection system can operate with a 4 × reduction by means of a set of high quality multilayer mirrors and a reflective mask. The radiation is first projected on the reflective mask with a couple of mirrors acting as the condenser. Then, it is focused with another mirror system onto the wafer. All mirrors are made of alternating molybdenum and silicon or beryllium multilayers that provide constructive interference in the direction of reflection.
The advantage of extreme UV lithography relies on its projection configuration and its high potentiality for several generation manufacturing. Fabrication of 100 nm line-and-space and 70 nm isolated lines has already been demonstrated [47]. In principle, the reflective masks are robust and the mask pattern can be made easily because the minimum feature size required is four times larger than that on a chip. However, there are still many technological challenges to be overcome, including mask optimization, fabrication of high precision optical elements as well as engineering issues such as alignment, system metrology and feedback control [2, 46]. In particular, the roughness of the surface and the interfaces of the mirrors and the reflective masks has to be minimized with an accuracy better than 0.25 nm. They must also be zero-stress and show robustness to thermal effects.

3.5 X-ray lithography

For smaller wavelengths ranging from 0.5 to 4 nm (soft-X rays), no material can be used for the construction of a projection system but a shadow technique is relevant [48–50]. A typical X-ray mask consists of 2 μm thick membrane of silicon carbide and absorber features of heavy metals such as Au, W, or Ta. Exposures can be done at a mask-to-wafer distance of ~10 μm with synchrotron radiation or laser-induced plasma source. The resolution of proximity X-ray lithography is defined by the Fresnel diffraction and the diffusion of photoelectrons in the resist. Increasing the X-ray wavelength decreases the diffusion range of the photoelectrons but increases the diffraction limited linewidth, estimated by \( W_0 = k_x \sqrt{\lambda g} \), \( g \) being the mask-to-wafer gap. It has been shown that the photoelectron scattering in polymethylmethacrylate (PMMA) is limited to a few nanometers [49] and the \( k_x \) factor is approximately 0.6 [51]. Experimentally, 50 nm lines were reproduced at a typical gap of 5 μm [51], and sub-30 nm structures were obtained at smaller gaps [52]. Figure 6 shows examples of the replicated resist patterns at very small gaps.

The X-ray lithography is nearly ready for industrial use [53–55], but some technical issues remain open, such as the mechanical and radiation stability of the masks, the availability of high throughput e-beam systems for mask making, and the reliability of high accuracy alignment. Soft X-rays can be also used to expose relatively thick resist. High aspect ratio features were thus obtained. Figure 7 shows one example of the fabrication of a 1.3 μm period and 6 μm thick photonic crystal template by three consecutive exposures of PMMA resist through a mask consisting of triangular lattice of holes [56]. Deep X-rays (\( \lambda < 0.1 \) nm) are used to expose much thicker resist for micromachining with the LIGA (lithography, electroforming and molding) process [57]. Figure 8 shows one example of the fabricated 100 μm thick PMMA resist structures.

3.6 Electron and ion projection lithography

Electron and ion projection lithography techniques are similar to optical projection lithography. A step-and-scan

![Figure 6. PMMA resist profiles of (a) a 20 nm linewidth and (b) 60 nm linewidth (aspect ratio >10), produced by proximity X-ray lithography.](image)

![Figure 7. Sub-μm resolution three-dimensional photonic crystal fabricated high-resolution X-ray exposures.](image)
exposing strategy is generally used for high-throughput production. A recent version of the electron projection lithography is the so-called scattering with angular limitation projection electron-beam lithography, referred to as Scalpel [58]. A comparable method called Prevail (projection reduction exposure with variable axis immersion lenses) has also been investigated, showing an 80 nm resolution ability [59]. By Scalpel, high-energy electrons (~100 keV) are projected onto a substrate passing successively through a scattering mask and two focusing lenses. A back focal plane aperture is placed between the two lenses, allowing only nonscattered electrons to pass through. The Scalpel mask consists of heavy metal structures (Cr/W) for strong scattering and a very thin SiN membrane for weak scattering, all supported by a framework of silicon struts. The advantage of Scalpel relies on its 4 × image reduction enabling an easy mask technology. The problem is that the energetic electrons significantly heat the wafer, possible leading to expansion and some distortion of the pattern. In addition, the interaction among the scattered electrons themselves can cause a blurring of the image. This limits the maximum achievable current and thus the rate of production.

Ion projection lithography uses accelerated hydrogen or helium ions at energies in the 70–150 keV range [60–62]. Now, ions are projected onto the wafer through a stencil mask and electrostatic lenses with a 4 × reduction. The ion stencil mask consists of an ion absorbing layer on a 2–3 μm thick silicon membrane in which ions are either stopped or pass through holes. Two complementary masks are generally required with extremely tight alignment to produce, for example, line arrays. Although a sub-100 nm resolution has been demonstrated [61], problems such as source uniformity, space charge, wafer heating, and energy deposition in stencil masks remain to be studied. Ion beam lithography can also be performed as for shadow printing. The approach is similar to proximity X-ray lithography except that ion source and stencil masks are now used as described above.

4 Nonconventional methods

4.1 Nanoimprint lithography

Nanoimprint is a nonconventional and low-cost technique for high resolution pattern replication [63, 64]. A rigid mold is used to physically deform a heated polymer layer coated on a substrate. The imprint mold can be made of metal or thermal silicon dioxide produced on a silicon substrate. The imprint can be done by applying a typical pressure of 50 bar in the temperature range of 100°C–200°C depending on the polymer in use. After imprinting, the resist pattern is followed by reactive ion etching in order to produce a useful profile for the subsequent hard material pattern transfer. The mold can be re-used many times without damage. The imprint temperature for PMMA is typically 170°C and the resist thickness can vary in the range of 50–200 nm. Structures as small as 6 nm could be produced [64]. For PMMA, the lower the molecular weight, the easier the polymer flow because of the reduced viscosity. Other types of polymers such as standard optical resist S1805 [65] and newly developed hybran, and hydrogen-bonded polymers were studied at lower imprint temperatures. In particular, the hybrane polymer shows a better etching resistance than PMMA and thus an easier process control [66].

Features with high aspect ratios can be obtained using trilayer techniques [67]. Now, the pattern is first imprinted into the top layer and then transferred to the middle and the bottom layers by sequential reactive ion etching with O2 and SF6 gases. Because of the high aspect ratio, lift-off and other techniques such as electroplating and direct etching can be applied for hard material pattern transfer. Trilayer nanoimprint lithography can have a large process latitude, an easy control of critical dimensions, and a good pattern homogeneity over large areas. Pattern placement accuracy is also acceptable, showing no significant thermal derivation. Until now, the problem of fine alignment has not yet been resolved. The fabrication of ultrahigh resolution molds is also a difficult task. Moreover, the whole process duration is still too long for mass production. Fabrication of reliable and cheap imprint devices is clearly another challenge. A wide range of applications has been studied, including GaAs photodetectors [68], waveguide polarizers [69], high-density magnetic structures [70, 71], silicon field-effect transistors [72], GaAs quantum devices [73], etc. Figure 9 shows SEM images of (a) a 60 nm pitch magnetic dot array and (b) a pattern of Fresnel zone plate obtained by a trilayer process and lift-off.
obtained by casting and curing of polydimethylsiloxane (PDMS) against a master patterned by electron beam lithography. Solutions of thiols in ethanol are used as inks for gold substrates [83, 84] and octadecyltrichlorosilane (OTS) in toluene for silicon substrates [85, 86]. A molecular transfer occurs when an inked stamp is in contact with the substrate, resulting in a self-assembled monolayer (SAM) deposited on the touched portions which then is used as a mask for wet etching. The resolution of soft-lithography is limited by the mold feature size and the diffusion of the ink molecules around the contact portions. In practice, the PDMS stamps provide a minimum feature size of ~200 nm. Smaller features can be obtained by using stamps of a higher stiffness such as a mixture of vinylmethyl-dimethylsiloxane copolymers, trimethylsiloxy terminated (VDT) and methylhydroxiloxane-dimethylsiloxy copolymers (HMS) polymer mixture [85, 86]. The commonly used thiol is hexadecanethiol (C_{16}H_{33}SH), but it can be replaced by eicosanethiol (C_{20}H_{41}SH) because of the reduced molecular diffusion [86].

A trilayer technique can be applied for an easier hard material pattern transfer including reactive ion etching, lift-off and ion implantation [87, 88]. A PMMA layer is first spin-coated on a substrate before gold deposition. Then, the etch-defined gold pattern serves as a mask for reactive ion etching of PMMA. Finally, the resulting PMMA pattern is used for lift-off or other hard material pattern transfer. Figure 10 shows patterns defined by (a) soft-lithography after wet etching of gold and (b) a trilayer process followed by lift-off of 40 nm thick Co.

Soft-lithography has been applied to the fabrication of field effect transistors [89, 90], magnetic structures [91] and optical devices [81]. Patterning on substrates such as SiO_{2}, glass, Ag, is also possible. More generally, SAMs were used as resists for lithography with UV light, X-rays,

Figure 9. Magnetic dot array of (a) 60 nm period and (b) X-ray Fresnel zone plate fabricated by nanoimprint and lift-off.

There are several related techniques involving thermodynamic molding of polymers. Injection molding, for example, is now capable of replicating 100 nm features [74, 75]. Metallic molds for both nanoimprint and mold injection can be fabricated by first sputtering a thin Ni film on a patterned resist master and then increasing the metal thickness by electroplating, followed by a master and mold separation [76]. When a heated polymer thin film is in close proximity of the mold (<1 μm), the electrostatic interaction between the two can result in a spatial redistribution of the quantity of the polymer. This phenomenon has been recently studied as a new patterning technique, referred to as lithographically induced self-construction of polymer [77, 78].

4.2 Soft-lithography

Soft-lithography, or microcontact printing, is based on the use of an elastomeric stamp to ink a solid substrate with the help of molecular self-assembly [79–82]. Stamps are

Figure 10. (a) Soft-lithography-defined 0.4 μm pitch dot array of gold after wet etching and (b) 1 μm pitch dot array of cobalt after a trilayer process followed by lift-off.
electron or ion beam, proximity probes, for passivation of surfaces and for fabrication of colloids, etc. [81]. Similarly, modification of the surface hydrophobicity can be done in a precise way. Finally, biological species such as proteins can be printed directly [92]. Based on the use of PDMS stamps with relief patterns on their surface, other techniques such as molding in capillaries [93], microtransfer molding [94] and replica molding [95] were developed. In particular, molding of organic polymer was obtained by casting and curing an UV-curable polymer against an elastomeric PDMS master. Polymer waveguide [96], blazed diffraction gratings [97], waveguide interferometers/coupler [98], suspended carbon microsensorator [99], and interdigitated carbon capacitors [100] have been fabricated as demonstration examples.

4.3 Near-field optical lithography

Near-field optical lithography is derived from conventional optical contact printing. The commonly used contact printing is performed with a mask made of chromium as an absorber on thick glass plates. For ideal optical contact printing, optical flat and dust-free mask and wafer have to be used. In practice, this is not possible because of the nonflatness of both mask and wafer and/or the presence of defects between the two. For gratings, the transferable period is limited by $p_{\text{min}} = \frac{\lambda}{g + h}$, where $g$ is the gap and $h$ is the resist layer thickness. This gives rise to a minimum feature size of the order of 1 μm. In intimate contact, periods of 0.6 μm were achievable [101]. With a deep-UV F$_2$ excimer laser and a trilayer resist stack on a substrate, a linewidth of 0.15 μm has been obtained [102].

Conformable masks or substrates can be used to obtain a better contact. One solution is to use a thin flex glass mask coupled with a vacuum exposure device [103]. More recently, a conformable embedded-amplitude mask together with a trilayer resist stack on the substrate and 200 nm deep-UV radiation was used to replicate 100 nm features [104]. With highly flexible polymeric masks such as PDMS, a conformable contact can be easily made. With a PDMS phase mask (relief pattern without absorber), for example, 100 nm features could be produced over a large area [105, 106]. When the noncontact portions of such a PDMS mask are covered with a thin film of metal, the mask becomes a light-coupling mask, in which light is differentially guided by the structures from its backside towards the substrate [107]. Pattern replication with 100 nm minimum feature size could be achieved with a relatively thick resist layer and 256 nm light irradiation. Amplitude masks were also made from SiN membranes, which resulted in a replication in vacuum of 140 nm period gratings with a broadband light source (365–600 nm) [108].

Similarly, near-field optical lithography has been studied with mold-assisted polymerization. In this method, an etched silica plate is pressed into a light curable monomer and then UV light is used for photopolymerization [109]. After curing, the liquid monomer is solidified which provides a copy of the mold relief. Alternatively, an improved mask configuration was proposed to create a high contrast aerial image at the downstream of the mold. Here, the top portions of the mold are coated with a thin layer of absorber, which is sufficient to stop the light but allows optical tunneling through a small opening. Features as small as 50 nm could be obtained in addition to the possibility of a better resist profile [110].

4.4 Proximity probe lithography

Proximal probe lithography refers to a new class of surface modification techniques, based on either scanning tunneling microscope (STM), atomic force microscope (AFM), or scanning near-field optical microscope (SNOM) [111–114]. All these techniques employ a sharp probe tip placed in the close proximity of the sample surface with a feedback loop for distance regulation. Tunneling current across the tip and the sample surface is used for STM whereas optical methods are employed to monitor the vertical movement of the AFM cantilever. For SNOM, either a tunneling current or a shear force monitor can be applied for the distance regulation.

A great diversity of material modification mechanisms can be involved in proximity probe lithography. In STM lithography, short current pulses are generated for atomic deposition, extraction or displacement. In AFM lithography, a metallized AFM tip can be used for current pulse generation, allowing selective oxidation, resist exposure, field ionization, field-induced evaporation, indentation, chemical etching, as well as precursor decomposition [114]. Individual atoms or molecules can also be displaced or repositioned [115, 116], thereby providing a way of building a true ‘atomic circuit’.

Implementation of the probe techniques in actual device fabrication has motivated considerable efforts on Si-based technology [117]. With a metallized AFM tip, for example, an H-passivated Si-surface can be anodically oxidized selectively with a high spatial resolution. Although these oxide patterns are a few monolayers thick, they form a robust mask for selective etching. Finally, an AFM tip can be used as a pen to plot molecular ink for self-assembling over a surface [118, 119]. For high-throughput patterning, many scanning probes must operate simultaneously over a sample surface. Fabrication of probe arrays containing up to 50 tips have been recently demonstrated [120], but their applicability as a true parallel lithography tool remains to be proven.
4.5 Other lithographic methods

A number of alternative methods can be used for the formation of more regular patterns. For example, holographic lithography produces periodic structures in a thin film resist. Typically, the optical standing waves used in holographic lithography are obtained by splitting a laser beam into two or more beams and then intersecting them via mirrors. With deep UV light, gratings of periods less than 200 nm were produced over large surface areas with a multilayer resist process [121]. More dense structures can be produced using an achromatic interferometric technique, with two-phase grating plates for the manipulation of the diffraction orders of a normal incident laser. By using two 200 nm period phase gratings for example, 100 nm period structures could be obtained [121, 122]. More recently, the interferometric technique has been used to pattern nonlinear optical polymers in order to create three-dimensional networks [123, 124].

Atom beam lithography is another example of the use of optical standing waves. When a coherent atom beam produced by laser cooling meets an optical standing wave just above a substrate, the atomic dipole interaction with the field gradient gives rise to a spatial redistribution of the atomic density, thereby producing a modulated atomic deposition on the substrate [125, 126]. Without using optical waves, atomic beam lithography can also be done through a mask [127].

In contrast, optical scanning lithography is a simple technique for direct patterning of a photo-resist film with a focused laser beam. Other kinds of material processing such as laser-induced polymerization [128] or etching and deposition [129] have also been demonstrated. The resolution of the optical scanning lithography is limited by diffraction [130]. If a laser beam is focused and a very short depth of focus is obtained, stereolithography can be performed [131, 132]. Moreover, by scanning an array of circular Fresnel zone plates (FZP) under UV light irradiation, parallel writing can be done [133]. An FZP is basically a diffractive grating consisting of concentric zones such that positive first order diffraction from all zones will add constructively at the focal point. The resolution of an FZP-produced light spot is determined by the smallest zone width of the FZP. Particularly interesting is that movable deflecting micromirrors can be used to address individual light beams [133]. A similar technique has also been proposed for X-rays, in which multiplexed micromechanical shutters should be used to turn individual X-ray beamlets on or off in response to commands from a control computer [134]. The first demonstration of this technique has been done with synchrotron radiation, without individual beam control, showing the possibility of parallel writing at large exposure gaps [135].

With more laser power, scanning optical lithography can be used as an ablation tool for micromachining with materials like polymer, glass, ceramics, and metals [136, 137]. Now hard and brittle materials can also be machined by ultrasonic impact grinding. Deep cavities with a feature size down to a hundred µm can be obtained [138]. Finally, for metal and alloys, electrochemical machining can be done with a fine electrolytic jet for scanning etching, a resist pattern for parallel dissolution, or ultrashort voltage pulses between a tool electrode and a work-piece in an electrochemical environment for three-dimensional fabrication [139, 140].

4.6 Chemical and biological approaches

Numerous possibilities of nanostructure formation exist based on chemical and biological approaches relying on self-assembling and self-organization. The advantage of these techniques is to realize specific nanostructures without having to construct them atom by atom "by hand". Chemical processing in which self-assembly takes place in organic or inorganic materials can be used for forming aggregates and colloidal assemblies. For example, high-quality semiconductor CdS and CdSe nanocrystals could be produced chemically with narrow size distribution [141]. Other nanocrystals such as ZnO, TiO2, and porous Si films were also obtained, offering electronic and optical properties not present in bulk crystalline semiconductors [142]. Because of their very small sizes, tunable emission spectrum and good photochemical stability, semiconductor nanocrystals can be used as fluorescent biological labels [143, 144]. Another example is the regular periodic structure formation from colloids that can be used as new materials or templates. Many recent research works on the fabrication of three-dimensional photonic crystals were based on templates made of monodisperse opals and infiltration with silicon, metal, polymer and other materials [145–149]. Now, it is also possible to create particular patterns of colloidal crystals with the help of focused UV light [150]. More general application of nanoparticles can be found in metallurgy, catalyst, magnetic storage, drug delivery as well as in quantum device fabrication [142]. Another example is the self-organized growth of nanostructures on a strained substrate. By molecule beam epitaxy [151], InAs quantum dots were produced by depositing a fraction of InAs monolayer on a GaAs substrate, followed by another GaAs layer [152]. Metallic quantum particles were also obtained by depositing a Cu bilayer on a strain-relief pattern of the (111) Pt surfaces [153].

Carbon nanotubes are new quantum materials discovered recently [154, 155]. The electronic properties of these tiny and hollow cylinders show particular electronic
properties which can be used for future nanoelectronics [156] or molecular computing [157]. Other physical properties of carbon nanotubes are also interesting because of their caged architecture which can receive various foreign atoms. The challenge is to integrate them into large structures in a controllable and efficient way. Organic nanostructure patterns can be obtained by interactive coupling of selectively activated monomers. Block copolymers, for example, were used to create honeycomb morphologies of star-polymer polystyrene films [158]. More recently, supramolecular networks could be created from nano-scale rods in binary, phase-separating mixtures [159]. Fabrication of more general organic nanostructures involves the association of individual molecular components into desired architectures by molecular and supramolecular assembly [160–164]. Many researchers are now turning to biocatalytic synthesis with template-directed polymerization or more general genetic engineering [165]. For example, this approach is typically employed for the synthesis of artificial proteins, in which the template is a DNA sequence and the polypeptide chain is obtained through the intermediacy of a complementary mRNA sequence [166].

The integration of functional organic building blocks with the outside world can be done by assembling them on patterned inorganic structures with or without biomolecules and their recognition properties [167]. Based on silica surfactant self-assembling [168], several lithographic techniques such as pen-lithography, ink-jet printing and dip coating have been used to create hierarchically organized structures [169]. The selective binding of peptides to metal, metal-oxide and/or semiconductor structures can also be used to link inorganic building blocks [170].

5 Other nanofabrication issues

Conventional lithography methods result in patterned resist profiles serving as masks for subsequent material processing. Selective etching, doping, ion implantation or regrowth can then be performed [12]. Semiconductors, metals and dielectric materials are commonly used in microelectronics. Other materials such as polymers, glasses, ceramics, magnetic and superconducting materials now are used in nanofabrication. Selected pattern transfer methods are discussed in this section.

5.1 Thin-film deposition

Basically, thin-film deposition is concerned with adhesion onto the substrate, stress, stoichiometry, defect, impurity, and homogeneity. A deposition method is chosen according to the material type and deposition characteristics [1].

Usually polymer thin films are deposited by spin-coating. Metallic thin films can be obtained by thermal deposition based on the Joule effect or electron-beam assisted evaporation, sputtering or electroplating. Dielectric layers are grown by gas or liquid phase epitaxy, chemical vapor deposition, molecular beam epitaxy, organometallic epitaxy, etc. In particular, thermal oxidation of silicon yields a thin film of silicon dioxide, which is used as an excellent dielectric layer.

Multilayers of alternative light (Si, C...) and heavy (W, Mo...) elements can be obtained for extreme UV and X-ray reflection. In semiconductor physics, the association of III–V, II–VI, and IV–IV compounds results in various heterostructures such as quantum wells and super-lattices [171] which are now widely used for rapid electronic circuits and optoelectronic devices (laser diodes, modulators and receivers). More recently, multilayers of alternating magnetic (Fe, Co, etc.) and nonmagnetic (Au, Pt, Cu, Al₂O₃, etc.) metals is making a new generation of magnetic head and storage media possible for high-density magnetic recording [172]. Further exploration of different combinations will provide new structures with designed properties [173, 174]. Lateral structures are obtained by lithography and one of the pattern transfer techniques. For high-resolution fabrication, lift-off of a thin metallic layer is often used to produce dry etching masks. The quality of a lift-off is affected by resist profile, angle of deposition, and resist removal.

5.2 Wet and dry etching techniques

Wet chemical etching is often used for cleaning, polishing, and resist stripping. The etching performance through a mask of patterned resist is characterized by resist adhesion, minor contamination, and etching uniformity control [170]. Wet etching involves three steps: the transportation of the reactants to the reacting surfaces, the chemical reaction at the surface and the transportation of the products away from the surface. The chemical reactivity, i.e., the etching rate, is affected by both agitation and the temperature of the etching solution [1, 175, 176]. In most cases, wet etch is isotropic, i.e., the etching rate is the same in all directions. A notable exception is silicon etching, carried out with a hot solution of KOH and water, ethylenediaminepyrocatechol, or tetramethylammonium hydroxide. Typically, the etching rate is 0.6 μm/min in the (100)-plane, 0.1 μm in the (110)-plane and 0.006 μm in the (111)-plane. This can be understood by considering that the (111)-plane is more closely packed than the other planes. The silicon wet-etching is widely used for micro-machining such objects as thin membranes as well as scanning proximity probes.
Dry etching utilizes fluxes of energetic particles to remove target materials [177, 178]. Ion beam etching (IBE), for example, is simply a sputtering etching technique in which atoms are ejected from a solid target as a result of ion bombardment. IBE can be applied to a large range of materials but the problems of surface damage and redeposition of the etched material on the resist side-walls make this technique less useful. IBE involving chemical species is referred to as reactive ion beam etching (RIBE) and chemical assisted ion-beam etching (CAIBE). Reactive ion etching is based on simultaneous exposure of the sample to chemical reaction species and fluxes of energetic particles. One or several types of reactive gases are introduced into an evacuated process chamber in which a radio-frequency-induced plasma is applied to create reactive ion species. Etching occurs by a chemical reaction between the substrate and atoms or radicals produced in the plasma, and etched materials are pumped away as volatile gaseous species. The dry etching performance is qualified by a number of factors: etch rate, etching directionality, etching selectivity, mask resistance, environment, cost, etc.

Particular attention also has to be paid to the damage at surfaces and side-walls, which may deteriorate the electrical and optical performances of the device. For polymers, silicon, silicon-based dielectric layers and some metals, SF₆, CHF₃ O₂ gas mixtures are commonly used. Figure 11 shows examples of reactive ion-etched SiO₂ structures. For compound semiconductors such as GaAs, GaN, InP, etc., Cl-based gases or a CH₃H₂ gas mixture can be used. Several types of etching reactors are commercially available. They consist of two parallel plates (anode and cathode), an radio-frequency (RF) generator, impedance regulators, a vacuum pump, and a water cooling system. A few techniques are used for increasing the plasma density. In magnetron ion etching (MIE) reactors, a magnetic field is applied for electron confinement. In electron cyclotron resonance (ECR) reactors, power is fed by microwaves at cyclotron resonance condition and an electromagnet is used for electron confinement. Finally, in inductively coupled plasma (ICP) reactors high-density plasmas are generated with an inductive coil, which provides an optimized etching speed for high-aspect ratio experiments.

5.3 Micromachining

Current efforts are devoted to the fabrication of so-called microelectro-mechanical systems (MEMS) [179–181]. The most essential elements of MEMS consist of miniaturized stationary and moving structures capable of realizing micromechanical actions with high precision and good repeatability. A capacitive microactuator, for example, is made of two chargeable plates, one of them being suspended. The electrostatic attractive force is used to move the structure by biasing one of the plates [180]. Various mechanical components such as micromotors, micro-pumps, microreactors, and microvalves can be produced on the same substrate. Micromachining techniques are currently used in the fabrication of accelerometers, ink jet printer heads, arrays of movable mirrors for color projection displays, and probes for atomic force imaging. Efforts are also made towards a full integration of microsensors, microactuators and microfluidic elements.

The fabrication of MEMS commonly involves bulk or surface machining. Bulk machining defines microstructures by etching directly into the bulk material such as single crystal silicon [182]. Both wet chemical etching or reactive ion etching can be applied to obtain suspended microstructures with a high-aspect ratio. The advantage of bulk machining is that it allows the integration of active devices and the use of integrated circuit technologies. Surface machining defines the release and movable structures in a polysilicon film on a sacrificial layer of silicon dioxide, both deposited on bulk silicon [183]. More complex microchips including multilayer interconnections can be obtained by bonding together and laser drilling of several layers of the components. Silicon, glasses and metals can be wet-etched. Oxides like alumina, quartz and rubidium molybdenum oxide can be machined by etching or ablation with electron, ion or laser beams. Polymers can be molded by imprint, injection molding or soft-lithography. LIGA process can be used for a large-scale production [184–186].

Elastomers can also be used in MEMS where rigidity is a drawback. The PDMS stamps are known for their low interfacial free energy, low reactivity and the possibility of
conformal contacting and easy releasing. Moreover, they are optically transparent and mechanically compressible, which makes the PDMS structures very attractive as building blocks for micro-optical components. Photothermal detectors [187], devices for measuring displacement, strain, stress, force, torque, and acceleration [188], optical modulators and display devices [189] were fabricated. In these devices, the active optical elements are blocks of PDMS with a relief of a binary diffraction grating on their surface. Mechanical compression/extension can be used for controlling the relative optical path of light passing through the grating. Fabrication of light valves [190] has also been achieved. Other applications include phase-shifting masks for near-field optics [105], and three-dimensional fabrication [191].

For biological applications, particular attention has to be paid to system stability, reliability, size, power, biocompatibility and also the functionality requirements. Microcomponents for biosystems are now mainly fabricated with the silicon technology. Remarkable efforts are also devoted to glass and polymers because of their optical and insulating properties. A typical microsystem for biology may contain a number of fluid control elements: microactuator, micropumps, microvalves and sensors [192, 193]. One may also include functional devices such as heat exchangers, mixers, separators and reaction units on the same chip [194]. Microactuators and active valves are based on electrostatic, electromagnetic, piezoelectric or thermopneumatic operation. Shape memory alloys and bias spring properties can also be used. Microfluidic pumping systems can be realized based on their electroosmotic, travelling wave (ultrasonic) and thermal capillary properties. Finally, microsensors are fabricated by miniaturizing mechanical, thermodynamic, electronic and optical devices. Electrochemical sensor arrays and gas sensor systems can also be included.

6 Applications in information technology

Information processing is based on electronic computing, optical communication and magnetic data storage. Therefore, the main concerns are the computing speed, the communication bandwidth and the storage capacity. Each of these subfields is undergoing rapid progress and all of them need nanofabrication for further development. Because of the growing interplay between these research and development areas, it should soon be possible to integrate electronic, optical as well as magnetic modules on the same chip and thus create optimal performance and functionality. In this section, the most important aspects of nanofabrication in information technology are presented.

6.1 Microelectronics

Microelectronics has been developed based on silicon technology. Since many years, the performance of electronic computing has been increased mainly by decreasing the size of circuit features. It is predicted that the minimum feature sizes of all involved critical components, including dynamic access memory, microprocessing units as well as application-specific integrated circuits will be continuously scaled down and that the current CMOS technology will be extended at least to a 50 nm generation by the year 2012 [2, 192–197]. New architectures such as massively parallel processing and three-dimensional transistor networks will probably further increase the CMOS capability. In the quantum regime, effects such as localization, electron wave-guide diffraction, coulomb blockade, resonant tunneling and many others can be used for data processing [3, 198, 199]. The fabrication efforts are now concentrated on single-electron transistors (SET) [200–202], resonant tunneling diodes (RTDs) [203, 204], and spin devices [205, 206]. Algorithms for quantum computer and artificial neural network computing are also under investigation [207–210]. Molecular electronics is still at a preliminary stage but has deserved vigorous investigation because of its huge potential. Molecular switching, for example, has recently been demonstrated, and may be used for memory elements [210, 211, 212]. The top down fabrication technologies described in previous sections can be equally applied to CMOS technology and quantum devices, whereas more elegant bottom-up approaches will be used for molecular electronics. It is clear that a total control over the emerging structures in terms of wiring and interconnections of the molecular devices will present enormous challenges.

6.2 Optoelectronics and optics

Optical communication is used for medium and long distance data transportation because of the ultimate light speed and the optical wavelength-broad bandwidths. Besides optical fibers, the most essential components for optical communication are optoelectronic and optical devices such as laser diodes, photodetectors, modulators, multiplexers, and demultiplexers. In addition to the medium and long-distance communication, optoelectronic devices will also be required for board-to-board or computer-to-computer connections. It is known that the performance of conventional laser diodes will be improved by using quantum wires and dots [213].

Semiconductor nanocrystals as well as organic materials can also be considered as the building blocks of new optoelectronic devices. More generally, nanoscale engineering of both optical and electronic waves will lead to
important improvements of the all optoelectronic components [213]. For example, photonic band gap structures [213–216] can be designed and fabricated to fully control the spontaneous emission, thereby providing a way of realizing the so-called threshold-less semiconductor lasers [213]. In addition, the engineered waveguide channels and specified point defect through which light can flow will be realized precisely. In optics, much has to be done for system integration. The development of new sources, binary diffractive elements, as well as new spectroscopy methods is now in rapid progress. Finally, more conventional devices such as compact disks for information storage, CCD cameras, thin-film transistor (TFT)-LCDs, as well as plasma display panels will benefit from the nanofabrication developments [217]. Nanofabrication technology is clearly a key for the improvement of existing devices and the implementation of new optoelectronic devices. This particularly includes nonconventional lithographic methods such as soft-lithography and micromolding in capillaries for the fabrication of flexible light circuits.

6.3 Magnetism

Micromagnetism is the main concern for high-density data storage. It is predicted that the bit size will be continuously scaled down from 1 to 0.1 μm and smaller during the next few years. Both the magnetic support (hard disk, magneto-optical bands and disks) and the recording head will be fabricated based on new technologies. In particular, ultrathin multilayers, in which the interface quality is controlled at the atomic level, and various lateral structures are fabricated for various applications [218–221]. Recording heads based on giant magneto-resistance and spin valves are used for compact disk recording. Proximity probe techniques are studied for ultimate high-density recording. In parallel, spin electronics are developed in such devices as magnetic random access memories (MRAM), etc.

7 Applications in biology

The fabrication technologies, initially developed for microelectronics and MEMS, are now penetrating the biotechnology and biomedical industry. Compared to silicon-based electronic devices, the fabrication challenges are much more diverse in biochips, because of the large variety of biomaterials, fluids, and chemicals. On the other hand, the reduction in sizes and in volumes in miniaturized biosystems such as “labs-on-a-chip” can lead to drastic improvements in performance, throughput, and cost. These would be greatly helpful for biomedical applications, such as diagnostics and drug screening, and research on genomics and proteomics. The question of interest is how to meld the well-established nanofabrication technology with the required biological functionalities.

The research frontiers in nanofabrication are on a scale that is much smaller than the size of a typical cell (~10 μm). And without using powerful nanofabrication tools, micromachining has already resulted in a very exciting area [4–7]. Improvements are expected in many devices by either further scaling down dimensions or including nanostructured elements. Moreover, by addressing the molecular level, novel biodevices will be fabricated. The contribution of nanofabrication to biology covers a wide range of applications. In this section, we give selected examples in which nanofabrication has already allowed important progress, or is currently undergoing active research.

7.1 Microfluidics and analysis systems

Microfluidics is a fundamental research area which can be applied to advanced chemical and biological analysis. Large efforts are being devoted to the development of individual components for fluid injection, pumping, controlled delivery of samples and reagents, mixing, reaction, separation, and detection. The system integration should result in the so-called miniaturized total analysis systems (μTAS) that are capable of performing simple or multiple biochemical tasks. A large number of research laboratories around the world are developing such strategic activities [4, 5] and several companies are involved in a commercial exploration (Caliper, Agilent, Microcosm Technologies, Gameta Bioscience, Cepheid, etc.). μTAS devices are typically fabricated in silicon, glass or plastic, and often comprise a large variety of components including optical, electrical, and mechanical elements. Some areas can particularly benefit from the use of nanofabrication. Integrated separation systems, for instance, such as capillary electrophoresis, dielectrophoresis, and chromatography devices, are under intensive study.

The first effort on μTAS appeared almost a decade ago [222]. Since then, the development of chip-integrated capillary electrophoresis has received increasing interest [223]. Recently, size-dependent separation of long DNA molecules in a nanofluidic channel device consisting of a series of entropic traps was achieved [224]. This device was fabricated using standard photolithography and etching of silicon to create alternating deep and shallow (75–100 nm) regions in a 30 μm wide channel fixed over a flat glass substrate to allow fluorescence microscopy observation. Patterning the surface of electrophoresis microchannels with an array of posts in order to create stationary obstacles for DNA molecules, and thus improve separation, has been investigated for some years [225], and new improvements are expected for small biomolecules from reducing the size of the pillars down to the
nano-scale range. Alternative approaches including soft material molding or porous composite materials are promising for large-scale manufacturing.

The detailed understanding of electrohydrodynamics phenomena on micro- and nano-structured surfaces of varying geometries and surface properties is important for bioanalysis and separation applications. A wide range of studies on electroosmotic flow have recently been launched. For example, measurement of electroosmotic flows in microchannels containing patterned surface charges was recently performed with the help of soft-lithography methods, i.e., laminar flow patterning and micromolding [226]. Superpositions of patterned surface charge and periodic asymmetrical patterns have also been considered [227] but need experimental observation. Most of these studies deal with patterns in the range of one to a hundred μm, and the promising sub-μm scale remains to be tackled.

In parallel, dielectrophoresis, or the use of AC electric fields via multiple electrodes is widely used. Separation of proteins was reported in electric fields created in micro-meter-sized interdigitated electrodes [228]. Sub-μm particles, such as viruses and protein-covered latex spheres, could be manipulated and separated in castellated micro-electrodes [229] and scaling down to achieve separation of smaller particles of biological relevance is now considered. The interdigitated micro- and nano-electrodes and asymmetrical patterns in sequential flow, in order to displace fluid droplets, is also very promising. The use of lithographically patterned stationary phase supports instead of bead-packed microchannels is also of great interest in chromatography for the separation of neutral analytes.

Of course, in all these integrated bioseparation techniques, proper detection systems are needed. Fluorescence is the main optical tool used by biologists and biochemists for characterization, and such techniques as confocal fluorescence detection now allow to probe volumes in the femtoliter range. Integration with optoelectronic or other detection components for high-throughput and high-resolution analysis is indeed highly desirable. Very recently, a microfluidic/integrated optics device was built in Cornell, allowing fluorescence correlation spectroscopy of sub-femtoliter volumes [230]. Other integrated detection systems such as laser-induced fluorescence detection with integrated vertical cavity surface emitting diode lasers as the light source are also considered [231]. Finally, a distinctive and expanding field, e.g., plastic CD-based centrifugal microfluidics, could also gain from the contribution of nanofabrication via such higher resolution and low cost techniques as nanoimprint.

7.2 Microarrays and gene chips

One of the main contributions of microtechnologies to biology is of course the development of gene chips, e.g., DNA microarray chips for genetic sequence analysis [232]. These two-dimensional arrays of small reaction cells (each on the order of 100 × 100 μm), each containing a different set of bound known DNA sequences, are microfabricated on either silicon wafers, thin sheets of glass, plastic or a nylon membrane and are now revolutionizing such applications as diagnosis of genetic diseases, drug design and toxicology. A number of industrial companies are now commercializing gene chips (Affymetrix, Nanogen, Packard Instruments, Synteni, Motorola, etc.). Given that proper readout tools and data processing via bioinformatics can be developed, further reducing the sizes of the reaction cells to the sub-μm range via nanofabrication techniques like soft-lithography would therefore increase array density and be very beneficial by allowing a larger variety of probes and thus increasing parallel hybridization and speed of identification. Biomolecular assembly at the nanoscale, through polymer and supramolecular chemistry including self-assembled monolayers with a variety of functional groups are now areas in full bloom [141–146]. In particular, direct microcontact printing of proteins on solid substrates was recently carried out [233].

7.3 Nano-MEMS

Another area where impressive progress had been made in the past few years is nano-electromechanical systems. An example of nanomechanical achievement is the recent realization of microfabricated silicon cantilevers functionalized with a selection of biomolecules in order to provide molecular recognition [234]. Studies on nanomachines powered by molecular motors is another fascinating subject currently under study in several laboratories [235]. Biocompatible implantable nanodevices for in vivo operation, such as neural probes or multifunctional catheters containing microsensors and microactuators of a variety of types, have also opened a new future of medical applications [236]. In another area, sensitive analysis of proteins could recently be obtained through a nanomachined nanoelectrospray tip for mass spectrometry coupled to an integrated microchannel system [237].

7.4 Nanobiochemistry

Nanobiochemistry in picoliter volumes obtained by nanofabrication allows high concentration and increased throughput for combinatorial chemistry and diagnostics. Polymerase chain reaction (PCR) amplification in nano-
fabricated nanovials was recently achieved [238]. Specific nanorobotics tools such as piezoelectric fluid dispensers are necessary and under development.

7.5 Nanomaterials for biological applications

A large effort is also being dedicated to the engineering of nanomaterials, especially nanoparticles, nanotubes and colloids suitable for biological purposes, with again a wide range of applications: bioencapsulation, bioanalytical chemistry, bioseparation, bioimaging, etc. As an example, arrays of magnetic nanoparticles could be used for the electrophoresis of large DNA molecules [239]. Specific gold nanoparticles can also induce colorimetric DNA detection [240] or be used for staining in biological electron microscopy. More striking is the very recent demonstration of direct observation of single-nucleotide polymorphism in kilobase-sized DNA achieved via single-wall carbon nanotube capping of a conventional silicon AFM tip [241]. The use of an array of such nanotube tip AFMs could indeed allow unprecedented ultrafast genetic haplotyping.

7.6 Bioelectronics

Finally, the contribution of biology to nanofabrication for novel electronics is altogether a different field. So-called bioelectronics and molecular electronics are receiving ever growing attention from the physics and electronics scientific communities. Promising results have been obtained recently with DNA molecules, in particular on DNA-template assembly and electrode attachment of a conducting silver wire [242]. Arrays of nanoelectrodes have also been developed to examine the molecular basis for electron transport processes between redox proteins and metal surface, and biological molecules are being incorporated into conducting polymers during the polymerization process to provide a possible mechanism for localizing proteins at a microelectrode. If the correct polymers are used, there is the possibility of creating a bioelectronic interface by providing molecular wires between the electrode and the biomolecular material [243]. Another example of melding between biology and electronics is the use of semiconductor field-effect transistors to sense a charge change during DNA hybridization [244].

8 Conclusions

We have attempted to review some of the current topics in nanofabrication. The starting point was the pattern formation technology, including the most expensive but reliable industrial approaches and a number of nonconventional methods. Other fabrication issues have also been briefly discussed. It appears that all lithography methods presented are useful in enabling certain applications. The semiconductor industry will continue to use optical projection lithography for one or more CMOS generations. Then, one of the next generation lithographies will be employed to scale down further the CMOS critical dimensions. Following this, quantum devices as well as molecular computing will most likely be developed based on new fabrication technologies with the help of molecular self-assembling and self-organization. In parallel, nanofabrication techniques will be involved in many other research and industry segments.

Information processing and biological microsystems are two well-identified areas of top research priority. In particular, the challenges of fabricating nano-scaled bio-devices are formidable from the point-of-views of system design, material choice, electric contacts, lithography and pattern transfer. An optimal strategy is to employ electrons, ions, X-rays or other nonconventional methods wherever they are most effective in solving the problem at hand. Nevertheless, nonconventional methods open more opportunities for engineering nanostructures and molecular systems. One should take care of the coming of new materials, algorithms, architectures and other innovative ways to overcome technological barriers in order to create designed functionalities. Due to the too many aspects of nanofabrication and the too rapid development of bioapplications today, it was not possible to cover the topic in an exhaustive manner, and a selection of the most important issues was thus chosen by the authors. Nevertheless, we hope that the stimulating conjectures contained in this article will encourage new exploration of this exciting area.

Received August 1, 2000

9 References


technology: Integrated Cooesecing Systems for Ultra-
precision and Ultrafine Products, Oxford University Press, 
48, 43–48.
(Ed.), Handbook of Microolithography, Microfabrication and 
470.
[222] Harrison, D. J., Manz, A., Fan, Z., Ludi, H., Widmer, H. M., 
[223] Ueda, M., Oana, H., Baba, Y., Doi, M., Yoshikawa, K., Bio-
phys. Chem. 1998, 71, 113–123; Nishimura, A., Tshako,
Ueda, M., Tabata, O., Baba, Y., J. IIEE Jap. 1999, 10,
460–463.
600–602.
[226] Stroock, A. D., Weck, M., Chiu, D. T., Huck, W. T. S.,
[228] Washizu, M., Suzuki, S., Kurosawa, O., Nishizaka, T., Shi-
[229] Morgan, H., Hughes, M. P., Green, M. G., Biophys. J.
[230] Foquet, M. E., Turner, S. W., Kottach, I., Webb, W. W., 
Craighead, H. G., Proceedings of Micro Total Analysis
[231] Arnold, D. W., Bailey, C. G., Gangulo, M. G., Matzke, 
C. M., Wondi, R., Sweatt, W. C., Kravitz, S. H., Warren, 
M. E., Inkestraw, D. J., Proceedings of Micro Total Analy-
[233] Bernard, A., Delamarche, E., Schmid, H., Michel, B., 
Bosshard, H. R., Biebuyck, H., Langmuir 1998, 14,
2225–2229.
[234] Fritz, J., Baller, M. K., Lang, H. P., Rothuizen, H., Vettiger,
P., Meyer, E., Guntherodt, H. J., Gerber, C., Gimzewski, 
[235] Montemagno, A., Bachand, G., Nanotechnology 1999, 10,
225–231.
[236] Proceedings of the 13th IEEE International MEMS Confer-
ence, Miyazaki, Japan, 23–27 January 2000.
[237] Lazar, I. M., Ramsey, R. S., Sundberg, S., Ramsey, J. M., 
[240] Eghbalian, R., Storhoff, J. J., Macic, R. C., Letsinger, 
D. E., Lieber, C. M., Nature Biotechnol. 2000, 18,
760–763.
[243] Jiang, L., Gildle, A., McNeil, C. J., Cooper, J. M., Biosen-