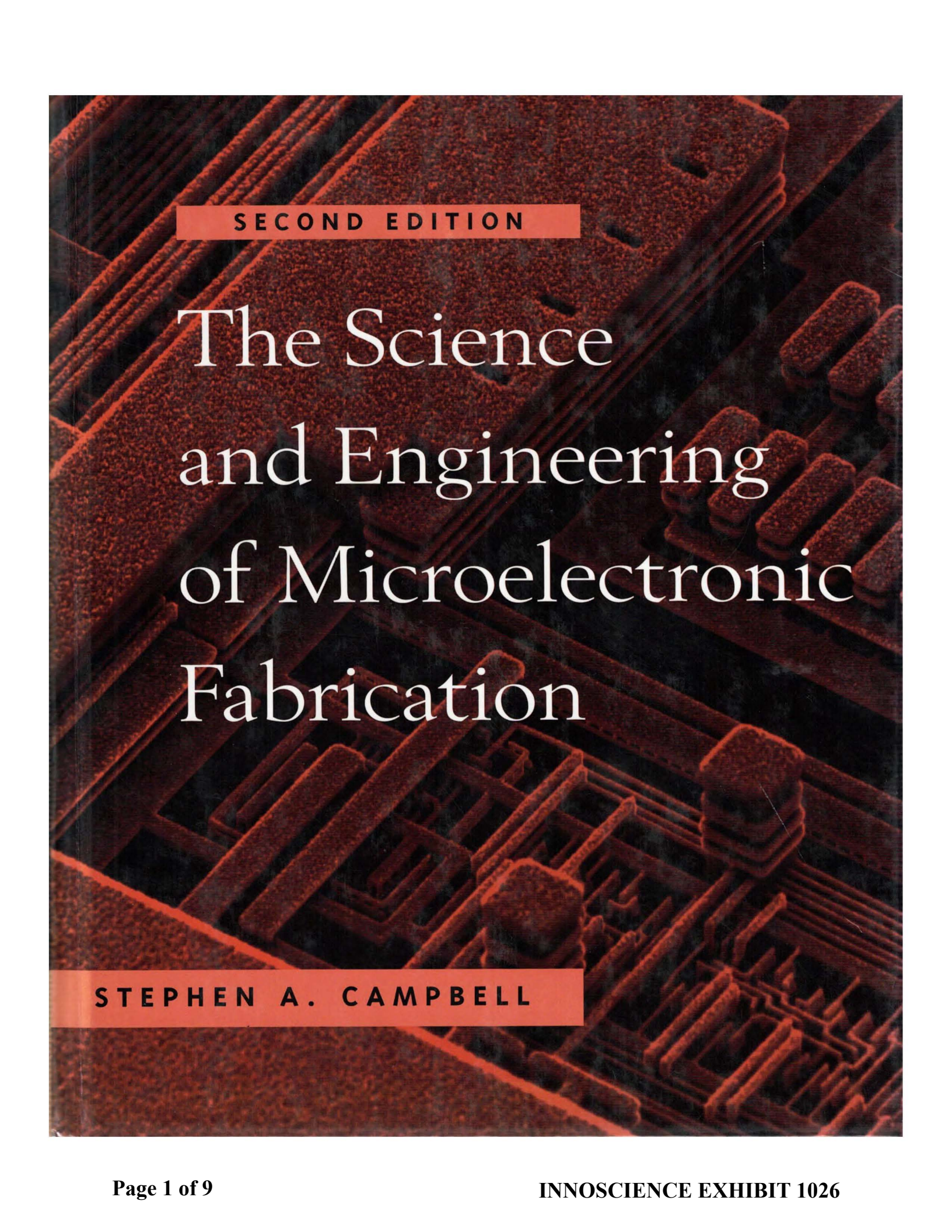


The background of the cover is a high-contrast, red-toned micrograph of a microelectronic circuit. It shows a complex pattern of rectangular and linear structures, likely representing various components and interconnects on a semiconductor chip. The lighting creates a sense of depth and texture.

SECOND EDITION

The Science and Engineering of Microelectronic Fabrication

STEPHEN A. CAMPBELL

The Science and Engineering of Microelectronic Fabrication

SECOND EDITION

Stephen A. Campbell
University of Minnesota

New York Oxford
OXFORD UNIVERSITY PRESS
2001

At first glance, these incredible densities and the associated design complexity would seem extremely daunting. This book, however, will focus on how these circuits are built rather than how they are designed or how the transistors operate. The fabrication process is similar no matter how many transistors are on the chip. The first half of the book will cover the basic operations required to build an IC. Using mechanical construction as an analogy, these would include steps such as forging, cutting, bending, drilling, and welding. These steps will be called unit processes in this text. If one knows how to do each of these steps for a certain material (e.g., steel), and if the machines and material required are available, they could be used to make a ladder, or a high-pressure cylinder, or a small ship. The required number and order of the steps will clearly depend on what is being built, but the basic unit processes remain the same. Furthermore, once a sequence that produces a good ship has been worked out, other ships of similar design could probably be built with the same process. The design of the ship, that is, what goes where, is a separate task. The shipbuilder is handed a set of blueprints to which he or she must build.

The collection and ordering of these unit processes for making a useful product will be called a *technology*. Part V of the book will cover some of the basic fabrication technologies. Whether the technology is used to make microprocessors, I/O controllers, or any other digital function is largely immaterial to the fabrication process. Even many analog designs can be built using a technology very similar to that used to build most digital circuits. An IC, then, starts with a need for some sort of electronic device. A designer or group of designers translates the requirements into a circuit design; that is, how many transistors, resistors, and capacitors must be used, what values they must have, and how they must be interconnected. The designer must have some input from the fabricator. In the shipbuilding example, the blueprints must somehow reflect the limitation that the shipbuilder cannot put rivets over weld joints or use small rivets and still expect them to hold very high pressures. The builder must therefore give the designer a document that says what can and cannot be done. In microelectronics this document is called the *design rules* or *layout rules*. They specify how small or large some features can be or how close two different features can be. If the design conforms to these rules, the chip can be built with the given technology.

1.1 Microelectronic Technologies: A Simple Example

Instead of blueprints, the circuit designer hands the IC fabricator a set of photomasks. The photomasks are a physical representation of the design that has been produced in accordance with the layout rules. As an example of this interface, assume that a need exists for an IC consisting of a simple voltage divider as shown in Figure 1.3. The technology to build this design is shown in Figure 1.4. Silicon wafers will be used as the substrate since they are flat, reasonably inexpensive, and most IC processing equipment is set up to handle them. The production of these substrates will be discussed in Chapter 2. Since the wafer is at least somewhat conductive, an insulating layer must first be deposited to prevent leakage between adjacent resistors. Alternatively, a thermal oxide of silicon could be grown, since it is an excellent insulator. The thermal oxidation of silicon is covered in Chapter 4. Next a conducting layer is deposited that will be used for the resistors. Several techniques for depositing both insulating and conducting layers will be discussed in Chapters 12–14.

This conducting layer must be divided up into individual resistors. This can be done by removing portions of the conducting layer, leaving rectangles of the film that are isolated from each other. The resistor value is given by

$$R = \rho \frac{L}{W \cdot t}$$

where ρ is the material resistivity, L is the resistor length, W is the resistor width, and t is the thickness of the layer. The designer can therefore select different values of resistors by choosing the width to

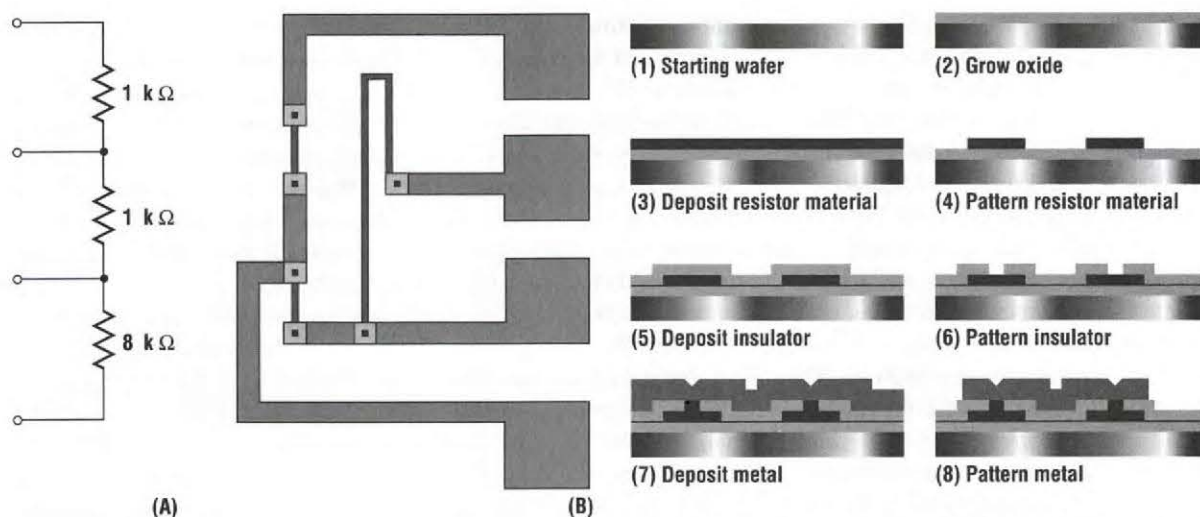


Figure 1.3 A simple resistor voltage divider. At left is a circuit representation; at right is a physical layout. The layers shown at right are resistor, contact, and low-resistance metal.

Figure 1.4 The technology flow for fabricating the resistor IC shown in Figure 1.3.

length ratio, subject to the limits specified by the layout rules. The technologist chooses the film thickness and the material (and therefore ρ) to give the designer an appropriate range of resistivities without forcing him to resort to extreme geometries. Since ρ and t are determined during the fabrication and are approximately constant across a wafer, the ratio ρ/t is more often specified than ρ or t individually. This ratio is called the *sheet resistance*, ρ_s . It has units of $\Omega/\square$ where the number of squares is the ratio of the length to width of the resistor line.

The resistor information from the design, namely L and W for each resistor, must be transferred from the photomask to the wafer. This is done using a process called *photolithography*. The most commonly used type of photolithography is *optical lithography*. In this process, a photosensitive layer called *photoresist* is first spread on the wafer (Figure 1.5). Light shining through the mask exposes the resist in the regions of the wafer where some of the metal resistor layer must be removed. In these exposed regions, a photochemical reaction occurs in the resist that causes it to be easily dissolved in a developer solution. After the develop step, the photoresist remains only in the areas where a resistor is desired. The wafer is then immersed in an acid that dissolves the exposed metal layer but does not significantly attack the resist. When the etch is complete, the wafers are removed from the acid bath, rinsed, and the photoresist is removed. The photolithographic process will be covered in Chapters 7 through 9. Chapter 11 will cover etching.

Although the resistors have now been formed, they still need to be interconnected and metal lines must be brought to the edge of the chip, where they can later be attached to metal wires for contact to the external world. This latter operation, called *packaging*, will not be covered in this text. If the metal lines have to cross over the resistors, another insulating layer must be deposited. To make electrical contact to the resistors, one can open up holes in the insulating layer using the same photolithographic and etch processes we had used for patterning the resistors, although the composition of the acid bath may be different. Finally, the fabrication sequence can be completed by depositing a

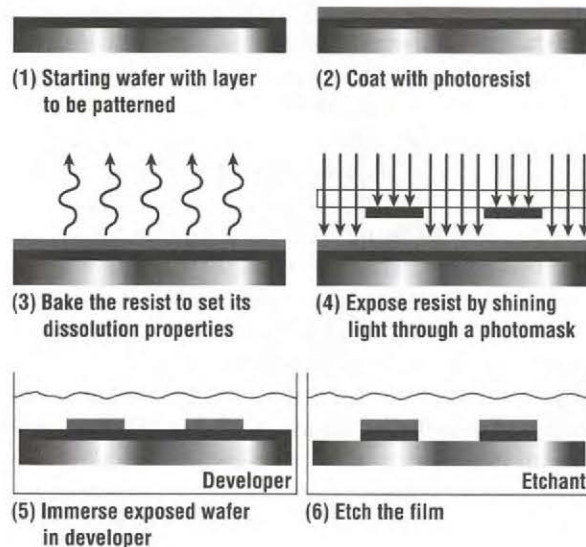


Figure 1.5 Steps required for a pattern transfer using optical lithography.

highly conductive metal layer, applying a third mask, and etching this metal interconnect layer.

The technology consists of four layers: the lower insulator, the resistor film, the upper insulator, and the interconnect metal. Photolithography is used to selectively remove some of these layers in certain regions, but any point on the wafer must be made of some subset of these layers in the same order that they were built up. Except for the edges of the patterns, the thickness of these films is constant. The technology uses three photolithography steps, three etch steps, and four thin film deposition (including oxide growth) steps. A very similar set of steps could be used to fabricate a capacitor. With only a few more steps, simple transistors can be built. Notice that the comparison with ship-building breaks down in one critical respect. The effort required to build the IC is independent of the number of resistors. The photomask might define one resistor or 1,000,000 resistors (assuming a 1,000,000 resistor circuit could be useful). In fact, two different sets of photomasks, one that defines only a few resistors and

one that defines thousands of resistors, could be used interchangeably with exactly the same technology and would require the same amount of work. This is because most of the unit processes described in this book operate in the whole wafer at the same time instead of one rivet at a time.

1.2 Unit Processes and Technologies

Some of the basic steps used in building an IC have already been discussed: photolithography, thin film deposition, and etching. Unit processes for thin film deposition include the processes of sputtering and evaporation. These are physical processes in that they do not generally depend on a chemical reaction. Sputtering is done by using charged atoms of argon called ions (Ar^+) to bombard a target containing the deposition material. The target erodes under this bombardment, and some of the material falls onto the wafers, coating them with a thin film of material. Evaporation involves heating the material to be deposited to a high temperature so that a vapor stream is created. The wafers are placed in this stream for coating. The third thin film deposition process that will be discussed is chemical vapor deposition. In this technique one or more gasses are made to flow into a chamber that contain the wafers to be coated. In many cases the wafers are also heated. A chemical reaction occurs that leaves the desired solid product on the surface of the wafer.

A resistor was chosen to simplify the first example. Most semiconductor devices require the formation of doped regions. For example, consider the n-channel MOSFET shown in Figure 1.6. Some familiar layers from the resistor example are recognizable: a blanket insulator and a patterned metal layer. One can also selectively dope the source and drain regions, putting together a technology to make the transistor. Dopants are either donors (n-type) or acceptors (p-type). For silicon, the most common n-type dopants are arsenic, phosphorus, and antimony, and the most common p-type dopant is boron. For gallium arsenide the most common n-type dopants are silicon, sulfur, and selenium, and the most common p-type dopants are carbon, beryllium, and zinc. In early semiconductor technologies, impurities were introduced by exposing heated wafers to a dopant containing gas. For example,

Chapter 6

Rapid Thermal Processing

The last few chapters have discussed the redistribution of impurities at high temperature. For small devices this redistribution is often very undesirable and so there has been a great deal of emphasis in recent years on low-temperature processes that minimize diffusion. Some processes, such as implant annealing, however, are not as effective at low temperature. Certain types of implant damage cannot be annealed out unless high temperatures are achieved. Furthermore, some dopants require anneal temperatures of at least 1000°C to be completely activated. The other avenue to reducing diffusion is to reduce the time at temperature. Standard furnace annealing is ill suited to short time anneals. The wafers in such a system heat from the edges inward. To avoid excessive temperature gradients that lead to warpage, the wafers must be heated and cooled slowly [1]. As a result, even though the annealing time could be short, the long temperature ramps result in significant diffusion. At the same time, there has been an increasing emphasis on processes that operate on one wafer at a time rather than large batches. These single wafer processes provide the best uniformity and reproducibility, particularly for large wafer sizes. Rapid thermal processing (RTP) describes a family of single wafer hot processes that has been developed to minimize the thermal budget of a process by reducing the time at temperature in addition to, or instead of, reducing the temperature.

RTP was originally developed for implant annealing. Although this application is still common, the use of rapid thermal approaches to processing has spread to oxidation, chemical vapor deposition, and epitaxial growth. Central to all rapid thermal processes is a set of common problems: heating and cooling the wafer uniformly, being able to maintain a uniform temperature during the process, and measuring the wafer temperature. The next few sections will describe the nature of these problems and the approaches that rapid thermal systems have adopted to address them. The chapter will then review some applications of the technology starting from the traditional implant annealing and the more recently developed oxidation and nitridization and concluding with the formation of silicides. The use of RTP for chemical vapor deposition (RTCVD) and epitaxial growth will be delayed until later in the book, since these topics also require an understanding of general chemical vapor deposition techniques.

Rapid thermal processes can be divided into three broad classes by the type of heating that is carried out: adiabatic, thermal flux, and isothermal [2]. The first demonstration of RTP used an adiabatic heat source [3]. In this approach, fast pulses of light in a broad beam heat only the front surface (of order a few microns) of the wafer as long as the pulse length is short compared with the thermal time constant of the substrate. Adiabatic systems are often powered by broad-beam coherent sources

such as excimer lasers. Although this type of annealing system allows the shortest time at temperature, it has several important drawbacks. These include difficulties in controlling temperatures and anneal times, large vertical temperature gradients, and large capital equipment costs. Thermal flux systems use an intense spot source such as an electron beam or a focused laser that is scanned across the wafer. The scan period must be short compared to the thermal time constant or large lateral thermal gradients will result. Although this type of system has been used for research, the defects caused by lateral thermal nonuniformity are usually large enough to prevent their use for IC fabrication. Isothermal heating uses a broad beam of radiation to heat the wafer for many seconds. These systems may have minimal temperature gradients across and through the wafer. Typically they are powered by incoherent sources such as an array of tungsten-halogen lamps. The wafer rests on quartz pins in isothermal systems. Quartz is selected because of its chemical stability and its low thermal conductivity. This arrangement is sometimes called *thermal isolation*. This chapter will concentrate on isothermal RTP systems since almost all current generation commercial systems use this design.

6.1 Gray Body Radiation, Heat Exchange, and Optical Absorption°

There are four types of heat transfer that may be of interest to semiconductor processing: conduction, convection, forced flow, and radiation. Thermal conduction is the diffusion of heat through a solid or gas. The heat flow through a cross-sectional area A of a solid or an immobile gas or liquid is given by

$$\dot{q}(T) = k_{th}(T)A\nabla T \quad (6.1)$$

where $k_{th}(T)$ is the thermal conductivity of the material. Dividing both sides by the area produces Fick's first law in terms of heat transfer. Since most of the optical energy in rapid thermal processing is absorbed in the first few microns of the wafer, thermal conduction through the wafer plays an important role in the final temperature distribution. When considering thermal conduction in a gas, however, one must also take into account gas flow since it can alter the rate of heat transfer. If the flow is caused by an externally applied pressure gradient it is said to be forced flow. Examples include flows caused by gas injection or pumping. Flows that are in response to temperature gradients in an otherwise closed system are called *natural flows*. The movement of water in a heated pot is a natural flow. An effective heat transfer can be defined as

$$\dot{q} = h(T - T_{\infty}) \quad (6.2)$$

where T_{∞} is the temperature of the gas far from the wafer and h is an effective heat transfer coefficient that depends on both free and forced flow. For most geometries h is a function of temperature and the position on the wafer.

The amount of power that can be delivered by gas flow is limited. As a result, most rapid thermal systems use radiative heat transfer as the primary method of heat exchange. One of the basic parameters of radiative heat transfer is the spectral radiant exitance $M_{\lambda}(\lambda, T)$, the amount of power radiated by a body into a perfectly absorbing environment (black box) per unit surface area of the emitting object and per unit wavelength of the radiation. Planck's radiation law gives the spectral radiant exitance as

$$M_{\lambda}(T) = \varepsilon(\lambda) \frac{c_1}{\lambda^5 (e^{c_2/\lambda T} - 1)} \quad (6.3)$$

been shown that using RTP to introduce small amounts (parts per million) of fluorine in an oxide produces excellent breakdown characteristics. This fluorine can be incorporated through the use of NF_3 in a rapid thermal processor [64].

It is also possible to produce nitrided oxides by growing the oxide directly in nitric (NO) or nitrous (N_2O) oxygen ambients. N_2O -grown films have better reliability and are more resistant to boron penetration than O_2 -grown films [65]. The peak nitrogen concentration occurs near the silicon interface and varies from about 1% (at 900°C) to about 1.7% (at 1100°C). Nitrogen incorporation in these films is known to slow the oxidation rate, presumably since it reduces the diffusivity of O_2 in the dielectric [66]. NO -grown oxynitrides are known to incorporate even more nitrogen. Once again, the concentration depends on the temperature of the growth. The profile of nitrogen in NO -grown films is strongly affected by processing conditions, with much higher nitrogen concentrations in the bulk of the films [67]. NO can also be used to incorporate nitrogen in SiO_2 layers grown in O_2 .

One significant potential advantage of RTP is its potential to be used as a hot wafer/cold wall process if the chamber walls are properly cooled. The importance of this observation is that multiple processes can be run in a single chamber since the walls do not contaminate sequential process steps. The wafer temperature can be used to start and stop these reactions. The gas flowing in the chamber can be changed while the wafer is cool, preventing any chemical reaction. The advantage of this technique, of course, is the minimization of wafer handling and resultant contamination. Multiple layers can be deposited in this way with well-controlled interfaces. This capability was exploited very early in the development of RTP [68] by thermally growing a thin oxide and then depositing a layer of polycrystalline silicon using chemical vapor deposition to fabricate MOS capacitors. The capacitors were found to have excellent interfacial properties. Later work by the same group demonstrated that the lower silicon electrode could be grown selectively in silicon dioxide windows followed by oxidation of the epitaxial layer and finally polysilicon deposition so that both interfaces of the dielectric were formed in the RTP reactor [69].

6.7 Silicidation and Contact Formation

As will be described in some detail in Chapters 15, 16, and 18, for some device applications the resistivity of even heavily doped silicon is too large. In those cases, it is common to form metal silicides on top of the exposed silicon to reduce the resistivity. Metal silicides are often formed by depositing a thin metal layer on top of the wafer and then heating it to drive a silicidation reaction. Common metals for silicide formation include Ti, Pt, W, Ta, Mo, and Co. RTP allows the careful control of the silicidation temperature and ambient to minimize impurities and to promote the growth of the most desirable stoichiometry and phase of the silicide. In some cases the wafer has a partial coverage of SiO_2 before metal deposition. The metal does not react with the oxide and so can be easily removed wet chemically after the silicide reaction.

TiSi_2 has been one of the most desirable films for many applications due to its low resistivity ($13\ \mu\Omega\text{-cm}$). During the silicide formation anneal, however, silicon can diffuse along the grain boundaries of the TiSi_2 , leading to an overgrowth of the silicide on top of the edges of the oxide (Figure 6.14) [70]. This overgrowth can be minimized by first annealing at a low temperature ($450\text{--}600^\circ\text{C}$) to form TiSi . The unreacted metal is then stripped and a high temperature RTA is done to form the desired phase (C54) of the silicide. This phase of TiSi_2 requires an anneal of at least 750°C [71, 72], with some reports indicating that as high as 900°C is required [73]. When the initial silicidation is carried out in a pure N_2 ambient, lateral growth of the silicide is further impeded [74]. The titanium silicidation reaction is sensitive to ppm levels of H_2O or O_2 in the ambient since oxidation and silicidation will compete at the interface. Oxygen will also dramatically increase the sheet

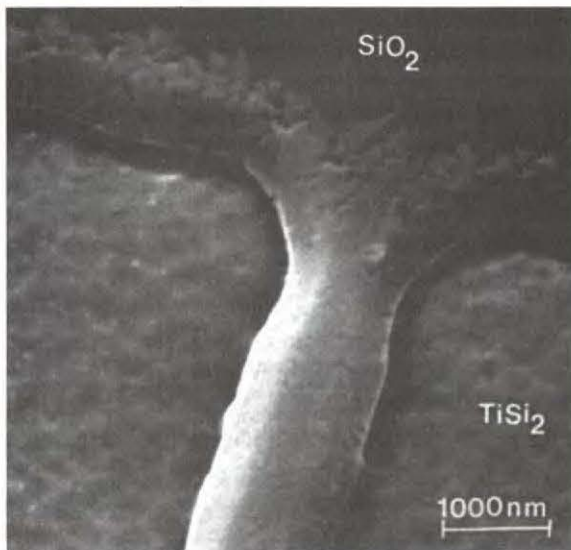


Figure 6.14 Results of a 900°C 30-sec formation of TiSi_2 in an oxide window. The ragged edges indicate silicide formed over the edges of the window (after Brat et al., used by permission, The Electrochemical Society).

resistance of the silicide. Silicides formed in N_2 containing 10 ppm of O_2 , for example, have nearly forty times the resistivity of silicides formed in 0.1 ppm O_2 [75]. RTP has a significant advantage over standard furnace anneals with regard to oxygen contamination, because the smaller RTP flow tubes reduce backstreaming from the air [76].

As will be discussed in a later chapter, cobalt silicide is very attractive for small MOS devices. The silicide forms as Co_2Si at 300–370°C, shifts toward CoSi at about 500°C, and forms CoSi_2 at 700°C or higher. At low temperatures Co is the dominant diffuser, however, in the CoSi_2 phase silicon diffusion dominates [77].

Another application of RTP for silicon technologies is the formation of barrier metals. These conducting films are used to prevent the interdiffusion of the silicon substrate and the aluminum-based alloys used for interconnecting devices. One common diffusion barrier is TiN. The film can be formed by reacting titanium with N_2 or NH_3 during a rapid thermal anneal, but this method produces high-resistance contacts when small geometries are used [78]. It is preferred instead to deposit the TiN with close to perfect stoichiometry

and then anneal the films in an RTA with a nitridizing ambient. This is sometimes called *grain boundary stuffing*, since the residual nitrogen is believed to occupy surface sites along the grain boundaries of the polycrystalline TiN. This nitrogen passivates the grain boundary and reduces the grain boundary diffusivity.

Rapid thermal processing has also been used for contact formation in GaAs technologies. As will be discussed in more detail in Chapter 15, low-resistance ohmic contacts to n-type GaAs are formed by depositing and thermally annealing a layer containing a mixture of gold and germanium. Although early technologies used conventional ovens for this annealing, it has recently been shown that very desirable contacts can be formed reproducibly using a short anneal at about 450°C [79]. The interfacial layer between the metal and semiconductor is considerably broader in conventional anneals [80], and the surface of the contact is more smooth, which improves circuit yield. Since temperature is a critical parameter in forming these contacts, a backside metallization on the GaAs wafer is sometimes used to ensure a uniform temperature across the wafer [81].

6.8 Alternative Rapid Thermal Processing Systems

As discussed earlier in the chapter, one of the principal problems with RTP is thermal uniformity. New RTP systems have recently been developed that use nontraditional chamber and/or heating designs to improve the thermal uniformity in RTP and to extend benefits of rapid thermal processing to small batch systems.

One of the more radical approaches to solving thermal uniformity in RTP is the resistively-heated or quartz lamp-heated bell jar (Figure 6.15). The wall is either made from SiC [82] or quartz [83]. The chamber walls are typically kept several hundred degrees above the maximum wafer temperature. The lower part of the chamber is kept cool. The wafer is rapidly heated in the chamber. The